ANÁLISE MULTIVARIADA DE DADOS: ESTUDOS PRELIMINARES À ANÁLISE FATORIAL CONFIRMATÓRIA (AFC)

Débora Fernanda Santos Dantas (1); Mylena Baia de Sousa (2); Gilberto da Silva Matos (3)

(1) / (2) Universidade Federal de Campina Grande, CCT/UAEP – dfsdantas1@gmail.com / mylenabaiasousa@gmail.com
(3) Orientador. Universidade Federal de Campina Grande, CCT/UAEst – gsmatos@gmail.com

1.0 INTRODUÇÃO

A compreensão e aplicação de modelos e técnicas estatísticas multivariadas de dados são de grande importância para a análise e tomada de decisões. Atualmente é fato que qualquer problema que envolve grandes conjuntos de dados é facilmente analisado por vários programas estatísticos em microcomputadores. A pesquisa em questão procura estudar e explicar duas técnicas para análise de dados: Modelo de Regressão Linear Múltipla (MRLM) e Análise de Componentes Principais (ACP). Este estudo visa não somente o aspecto teórico mas também o aspecto prático e tem como objetivo principal a preparação para a pesquisa científica na área de análise multivariada de dados cujo intuito específico é testar hipóteses sobre estruturas de variáveis latentes (variáveis não observáveis) e seus relacionamentos que frequentemente são discutidas na área de psicologia, administração, ciências sociais, marketing e pesquisa de mercado. De fato, o presente trabalho pode ser considerado como estudos preliminares ao estudo de uma técnica estatística mais avançada conhecida como Análise Fatorial Confirmatória (AFC).

2.0 METODOLOGIA

Para que os objetivos deste estudo fossem alcançados, reuniões semanais foram sendo realizadas onde foram discutidas algumas técnicas de análise multivariada de dados, tais como: Regressão Linear Múltipla (MRLM) e Análise de Componentes Principais (ACP). Tais técnicas podem ser consideradas como sendo técnicas preliminares ao estudo de uma técnica multivariada mais avançada conhecida por Análise Fatorial Confirmatória (AFC).

Para efeito de aplicação das técnicas multivariadas estudadas, dados do livro “Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada” de Sueli Aparecida Mingoti foram utilizados, bem como a análise de dados obtidos por simulações computacionais do ambiente computacional e estatístico R o qual é de distribuição gratuita e de código aberto.

2.1 Modelo de Regressão Linear Múltipla (MRLM)

O modelo de regressão linear múltipla (MRLM) relaciona uma variável resposta Y com p variáveis explicativas X_j, j=1,2,...,p. Por ser uma técnica simples e direta, fornece ao pesquisador previsão e explicação. O uso da regressão linear múltipla é permitido em quase toda relação de dependência, uma vez que é uma técnica bastante flexível e adaptável.

2.1.2 O modelo de regressão linear múltipla (MRLM) em forma matricial

Em termos matriciais, define-se

$$\underset{n\,x1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} \qquad \underset{n\,xp}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & . & . & X_{1,p-1} \\ 1 & X_{21} & . & . & X_{2,p-1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{n1} & . & . & X_{n,p-1} \end{bmatrix} \qquad \underset{p\,x1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\,x1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Assim: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Os estimadores de mínimos quadrados ou de máxima verossimilhança são não tendenciosos. O parâmetro β_0 é o intercepto do plano de regressão. O parâmetro β_j indica a mudança na resposta média *E(Y)* por unidade de acréscimo em X_j quando as demais variáveis explicativas (preditoras) X_k são mantidas constantes. **ε** é um vetor de variáveis aleatórias independentes e normalmente distribuídas com **E(ε)=0** e matriz de variância-covariância dada por $\boldsymbol{\sigma}^2(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, onde $\mathbf{I}$ é a matriz identidade de ordem *n*.

Segue assim que o vetor das observações **Y** tem esperança e variância dadas por:

$$\underset{n\,x1}{\mathbf{E}(\mathbf{Y})} = \mathbf{X}\boldsymbol{\beta} \qquad \underset{n\,xn}{\boldsymbol{\sigma}^2(\mathbf{Y})} = \sigma^2\mathbf{I}$$

O sistema de equações normais para o modelo: $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$

E os estimadores de mínimos quadrados são dados por: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

2.1.3 Valores estimados e resíduos

Os valores estimados são obtidos por: $\underset{n\,x1}{\hat{\mathbf{Y}}} = \mathbf{X}\mathbf{b}$

Os resíduos são obtidos através da expressão matricial:

$$\underset{n\mathbf{x}\mathbf{1}}{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb}$$

Além disso, é calculada a análise de variância, feito um gráfico de dispersão para analisar visualmente se o modelo parece ser realmente linear e/ou se existem valores extremos (outliers), é calculado o teste F, valor p, o coeficiente de determinação ($0 \leq R^2 \leq 1$) e vários outros suportes para realizar uma análise eficiente de regressão linear múltipla.

2.2 Análise de Componentes Principais (ACP)

A análise de componentes principais é uma técnica da estatística multivariada que consiste na transformação de um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão denominadas de componentes principais. As propriedades dos componentes principais são: cada componente principal é uma combinação linear de todas as variáveis originais, são independentes entre si e estimados com a intenção de conservar o máximo de informação em termos da variação total comportada nos dados.

A matriz de dados é de ordem ‘n x p’ e normalmente denominada de matriz ‘X’ quando considera-se a situação em que se observa ‘p’ características de ‘n’ indivíduos de uma população π. As características observadas são representadas pelas variáveis X1, X2, X3, ..., Xp.

A estrutura de interdependência entre as variáveis da matriz de dados é evidenciada pela matriz de covariância ‘S’ ou pela matriz de correlação ‘R’. O objetivo da análise de componentes principais é transformar a estrutura representada pelas variáveis X1, X2, X3, ..., Xp, em uma outra estrutura representada pelas variáveis Y1, Y2, Y3, ..., Yp não correlacionadas e com variâncias ordenadas, para que seja possível comparar os indivíduos usando apenas as variáveis Yis que apresentam maior variância. A solução é dada a partir da matriz de covariância S ou da matriz de correlação R.

A matriz S é simétrica e de ordem ‘p x p’. Podem-se fazer uma estimativa da matriz de covariância Σ da população π que representa-se por S a partir da matriz X de dados de ordem ‘n x p’.

Geralmente, as características são observadas em unidades de medidas diferentes entre si, e neste caso, quando houver discrepância nos dados, é conveniente padronizar as variáveis Xj (j=1, 2, 3, 5,..., p). A padronização pode ser feita com média zero e variância 1, ou com variância 1 e média qualquer. Após a padronização obtemos uma nova matriz de dados Z.

Para determinar os componentes principais normalmente partimos da matriz de correlação R. A matriz Z das variáveis padronizadas zj é igual a matriz de correlação da matriz de dados X. Vale salientar que o resultado encontrado para a análise a partir da matriz S pode ser diferente do resultado encontrado a partir da matriz R.

Os componentes principais são definidos resolvendo-se a equação característica da matriz S ou R, isto é:

$$\det\left[R - \lambda I\right] = 0 \; ou \; \left|R - \lambda I\right| = 0$$

Se a matriz R não apresentar nenhuma coluna, que seja combinação linear de outra, a equação $|R - \lambda I| = 0$ terá 'p' raízes chamadas de autovalores ou raízes características da matriz R. Na montagem da matriz de dados X é importante observar que o valor de 'n' dever ser pelo menos igual a 'p+1', ou seja, é aconselhável que o delineamento estatístico apresente pelo menos 'p+1' tratamentos.

Sejam λ1, λ2, λ3, ..., λp as raízes da equação característica da matriz R ou S, então para cada autovalor □i é gerado um autovetor $\tilde{a}_i$ correspondente.

$$\tilde{a}_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ip} \end{bmatrix}, i = 1,..., p$$

Os autovetores são normalizados e ortogonais entre si.

A contribuição Ci de cada componente principal Yi é expressa em porcentagem. A importância de um componente principal é classificada por meio de sua contribuição, isto é, pela proporção de variância total explicada pelo componente. A soma dos primeiros *k* autovalores representa a proporção de informação retida na redução de *p* para *k* dimensões. Com isso, pode-se decidir quantos componentes irá se usar na análise, isto é, quantos componentes serão utilizados para diferenciar os indivíduos. Não existe um modelo estatístico que auxilie nesta decisão. Na maioria dos casos, o número de componentes utilizados tem sido aquele que acumula70% ou mais de proporção da variância total. Para fazer a comparação da influência de X1, X2, ..., Xp sobre Y1 analisamos o peso ou "loading" de cada variável sobre o componente Y1.

Como objetivo da análise é comparar ou agrupar indivíduos, a análise continua e se faz necessário calcular os escores para cada componente principal que será utilizado na análise.

Os escores são os valores numéricos dos componentes principais. Após a redução de p para k dimensões, os k componentes principais serão os novos indivíduos e toda análise é realizada utilizando-se os escores desses componentes. Logo abaixo é exemplificada a organização de um conjunto de dados composto por *n* tratamentos, *p* variáveis e *k* componentes principais.

3.0 RESULTADOS E DISCUSSÃO

Para exemplificar e discutir os assuntos abordados neste trabalho utilizamos um exemplo didático. Neste exemplo, ilustramos a aplicação da técnica de componentes principais com uma posterior aplicação de análise de regressão linear múltipla. Foram considerados os dados relativos às notas de dezenove estudantes (dados obtidos de [4]). Cada aluno têm notas correspondentes a três provas: X_1, X_2, X_3 e assim aplicamos a essas notas a análise de componentes principais. Então construímos um índice global de desempenho dos estudantes, para relacioná-lo com as variáveis: tempo de estudo dedicado à disciplina, X_4, e gênero do estudante, X_5. A análise de componentes principais foi feita pela matriz de covariâncias e, como resultado, observamos que a primeira componente representava 87,9% da variância total do conjunto de dados relativos

às notas das três provas. Sendo assim, esta componente foi utilizada para construir um índice de desempenho global do estudante na disciplina, o $\hat{ID}$:

$$\hat{ID} = 0{,}514(nota\_1ª\ prova) + 0{,}671(nota\_2ª\ prova) + 0{,}535(nota\_3ª\ prova)$$

O desempenho do estudante é diretamente proporcional ao índice $\hat{ID}$, ou seja, quando maior o índice melhor será o desempenho global. Os escores observados de $\hat{ID}$ foram utilizados para ajustar um modelo de regressão linear que terá como variável resposta esses valores observados, e as variáveis X_4 (tempo de estudo) e X_5 (gênero) como explicativas. O modelo de regressão linear obtido foi:

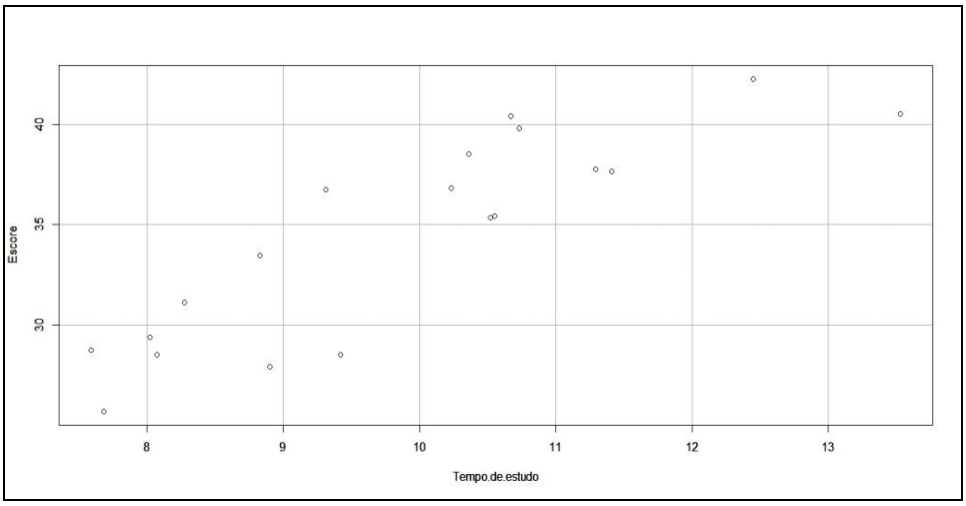
$$\hat{ID} = 8{,}663 + 2{,}657X_4 - 0{,}848X_5$$

A análise de significância, baseada no *valor "p"* nos revelou que os parâmetros do modelo de regressão referentes à constante e X_4 são significativos, pois os valores "p" de ambas foram menores do que 0,05. Em contrapartida a variável X_5 (gênero), que teve esse valor superior a 0,05, foi não significativa. Resumidamente, o modelo de regressão estimado nos revela que o desempenho do estudante é influenciado somente pelo tempo de estudo dedicado à disciplina (X_4). Podemos então excluir a variável sexo do modelo de regressão linear múltipla e encontrar um novo modelo, o modelo de regressão linear simples (ver Gráfico 1). O novo modelo estimado foi:

$$\hat{ID} = 7{,}63729 + 2{,}71154\,X_4$$

Este modelo ajustado diz que, para cada unidade a mais no tempo de estudo, o desempenho global do estudante é acrescido de 2,712 unidades.

Gráfico 1 – Modelo de regressão ajustado



4.0 CONCLUSÃO

Neste trabalho podemos constatar como a Análise de Componentes Principais (ACP) pode reduzir a quantidade de variáveis separando a informação relevante da que não tem grande contribuição para o problema analisado. Empregando-se essa análise, tornamos mais rápida e eficiente a visualização e o uso de uma grande quantidade de variáveis de um conjunto de dados. Com a análise de regressão avaliamos os efeitos da variável tempo de estudo e gênero sobre o índice global de desempenho dos alunos, obtido pela ACP. Desta forma, foi possível constatar que apenas a variável tempo de estudo foi significativa no sentido de explicar índice global de desempenho dos alunos.

5.0 REFERÊNCIAS BIBLIOGRÁFICAS

[1] Freire, Clarice Azevedo de Luna; Charnet, Eugênia M. Reginato; Bonvino, Heloísa; Charnet, Reginaldo; *Análise de Modelos de Regressão Linear com Aplicações*. Campinas: Editora da Unicamp, 1999.

[2] J. F. Hair Jr., W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham. *Análise Multivariada de Dados*. Bookman, 2009.

[3] Varella, Carlos Alberto Alves. *Análise de Componentes Principais*. Rio de Janeiro: Seropédica, 2008.

[4] Sueli Aparecida Mingoti. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Editora UFMG, 2005.