UM CASO DE USO ERRÔNEO DA ANÁLISE FATORIAL

RODOLFO HOFFMANN[1]

Recentemente a Revista de Economia Rural publicou um artigo sobre o desenvolvimento rural da República da Coréia (Lemos, 1987). Vou limitar meus comentários à análise fatorial feita pelo autor.

A análise fatorial é uma técnica estatística usada quando se tem um número relativamente grande de variáveis e se deseja um número menor de variáveis, denominadas fatores, que expliquem (estatisticamente) as correlações entre as variáveis originais. No modelo clássico de análise fatorial, admite-se que cada um das **n** variáveis originais é uma combinação linear de **k** fatores comuns e mais um componente específico de cada variável. Normalmente o número de fatores comuns (**k**) é bem menor do que o número de variáveis (**n**). A fração da variância de cada variável que é explicada (estatisticamente) pelos fatores comuns é denominada **comunalidade** dessa variável.

Harman (1976), em um livro já clássico sobre o assunto, mostra que há várias maneiras de se fazer uma análise fatorial. Há o método da máxima verossimilhança, mas são mais comumente utilizados métodos baseados na extração de componentes principais de uma modificação da matriz de correlações das variáveis originais. A modificação consiste em substituir os valores (todos iguais a 1) da diagonal principal da matriz de correlações por estimativas preliminares da comunalidade de cada variável. É comum adotar, como estimativa preliminar da comunalidade, o coeficiente de determinação de uma regressão múltipla da variável contra as demais n-1 variáveis.

Há também vários métodos de fazer a rotação dos fatores, que facilita a sua caracterização. No caso de uma rotação ortogonal podem ser adotados, entre outros, os critérios VARIMAX ou QUARTIMAX.

Devido à multiplicidade de métodos que podem ser utilizados, um trabalho sobre o assunto deve, necessariamente, explicitar o método utilizado. Isso não é feito no artigo comentado.

O autor afirma que iniciou a análise com 20 variáveis mas que "apenas nove foram selecionadas". Afirma, ainda, que

> "Isto, provavelmente prendeu-se ao fato do número de observações ser bastante reduzido. Apenas dispunhamos de informações para um período de cinco anos. Verificou-se que quando as vinte variáveis foram colocadas simultaneamente no modelo de análise fatorial, houve forte colinearidade entre elas, o que impossibilitou a inversão da matriz de correlação entre as variáveis, que é fundamental para a estimação..." (Lemos, 1987, p.260).

---

1 Professor da ESALQ/USP – C.P. 9 – 13.400 Piracicaba, SP.

Considero estranho que um pesquisador selecione 9 de 20 variáveis e confesse que não sabe com certeza porque fez isso.

Em análise fatorial variáveis altamente correlacionadas entre si ficarão fortemente associadas com um mesmo fator. A finalidade da análise fatorial é, exatamente, explicar a estrutura de correlação entre as variáveis. É um absurdo, portanto, eliminar variáveis por causa da "forte colinearidade entre elas".

O curioso é que o autor afirma que o problema residia no fato de a "forte colinearidade" tornar impossível a inversão da matriz de correlações e que teria sido essa a razão para reduzir o número de variáveis de 20 para 9. Acontece que ele afirma que só dispunha de 5 observações[2]. Então a sua matriz de correlações pode ter no máximo 5 linhas ou colunas linearmente independentes, isto é, a matriz de correlações tem característica menor ou igual a 5. Conclui-se que a matriz de correlações de 9 variáveis é singular, da mesma maneira que a matriz de correlações com as 20 variáveis. Como foi possível inverter uma e não a outra?

A singularidade da matriz de correlações não impede que se faça uma análise fatorial. Entretanto, se o número de variáveis é igual ou maior do que o número de observações, não tem sentido estimar a comunalidade de uma variável através de uma regressão múltipla dessa variável contra todas as outras variáveis. Essa regressão múltipla não é definida quando o número de observações é menor do que o número de variáveis. Uma possibilidade bastante simples é não modificar a matriz de correlações (deixar os valores 1 na diagonal principal). Neste caso estaria sendo utilizado o método de componentes principais (que, a rigor, se distingue da análise fatorial).

Talvez seja oportuno ressaltar que alguns autores, como Chatfield e Collins (1980), consideram que na maioria dos casos (mesmo quando o número de observações é maior do que o número de variáveis), deveria ser utilizado o método de componentes principais, que é mais simples, e não a análise fatorial clássica.

REFERÊNCIAS

CHATFIELD, & COLLINS, J. **Introduction to Multivariate Analysis.** New York, Chapman and Hall, 1980.

HARMAN, Harri H. **Modern factor analysis.** 3ª ed. Chicago, University of Chicago Press, 1976.

LEMOS, J. J. S. Análise do Desenvolvimento Rural da República da Coréia: um Estudo Exploratório. **R. Econ. Rural 25**(2):251-162, abr./jun. 1987.

2 Na página anterior o autor afirma que "as informações relevantes cobrem o período que vai de 1979 até 1984". Incluindo os extremos, tem-se 6 observações, e não 5, mas isso não afeta, essencialmente, o que se segue.