Resenha do Livro

Análise Multivariada de Dados, 6ª edição, Hair, Black, Babin, Anderson e Tatham[1]

Rodolfo Hoffmann[2]

Em 2009 a editora Prentice Hall publicou nos Estados Unidos a 7ª edição desse livro extraordinário. No mesmo ano a tradução para o português da 6ª edição foi publicada no Brasil. Trata-se de um livro didático apresentando enorme variedade de técnicas estatísticas multivariadas, com destaque para suas aplicações em problemas de administração de empresas, particularmente em *Marketing*, que é a área de experiência dos cinco autores.

O livro é extraordinariamente abrangente. Há capítulos sobre a análise preliminar dos dados, análise de regressão, análise fatorial, regressão logística, análise discriminante, análise multivariada de variância, análise conjunta, análise de agrupamentos, análise de correspondência e modelagem de equações estruturais.

Para quem já estudou estatística nos livros-texto usuais, chama a atenção o fato de Hair et al. (2009) praticamente não usarem expressões matemáticas. Isso é interessante para o leitor que nunca se acostumou com o uso da notação matemática, mas por vezes torna o texto desnecessariamente prolixo para quem sabe usar essa notação.

O livro nunca apresenta os cálculos necessários para obter os resultados estatísticos. Isso fica por conta dos pacotes estatísticos utilizados. A ideia é que o leitor aprenda quando deve usar determinado método, use o computador para obter os resultados estatísticos e saiba interpretá-los.

É claro que um livro didático dessa natureza poderia, em princípio, ser rigorosamente correto na apresentação dos pressupostos dos diferentes métodos e na interpretação dos resultados. Infelizmente, apesar de se tratar da 6ª edição americana, o livro apresenta algumas falhas conceituais graves. Alguns exemplos.

Os autores constatam que "A regressão múltipla é de longe a técnica multivariada mais utilizada entre aquelas examinadas neste texto" (Hair et al., 2009, p. 163). Usando o índice j para indicar as observações, o modelo de uma regressão linear múltipla de Y_j contra X_{1j} , X_{2j} , ... , X_{kj} é

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + u_j \ ,$$

com parâmetros α , β_1 , β_2 , ... , β_k e erro u_j .

O livro de Hair et al. (2009) ensina, erroneamente, que para usar a análise de regressão múltipla é necessário pressupor que todas as variáveis têm distribuição normal, como se pode constatar na figura 4-1 na p. 163 e na p. 197. Nas p. 82-91 afirmam, incorretamente, que em qualquer análise multivariada todas as variáveis devem ter distribuição normal e na tabela 2-11 (p. 90) e nas p. 210-211 indicam que devem ser feitas transformações das variáveis (incluindo variáveis explanatórias) para que sua distribuição se torne aproximadamente normal.

Na realidade, estimativas não tendenciosas dos parâmetros de uma equação de regressão podem ser obtidas pressupondo apenas a forma da relação e que o erro seja, em média, igual a zero. Apenas para usar os testes *t* e *F* é necessário pressupor que o erro (u_j) tem distribuição normal. Não é necessário pressupor que qualquer das variáveis explanatórias (X_{1j} , X_{2j} , ... , X_{kj}) tenha distribuição normal.

É comum, em análise de regressão, usar

[1]JEL Classification: C10. Registrado no CCTC, REA-18/2015.

[2]Engenheiro Agrônomo, Professor Sênior, Escola Superior de Agricultura "Luiz de Queiróz", Piracicaba, Estado de São Paulo, Brasil (e-mail: hoffmannr@usp.br).

variáveis explanatórias binárias, isto é, variáveis que só tem dois valores (geralmente 0 e 1). É obviamente absurdo pretender que tais variáveis tenham distribuição normal.

No livro de Hair et al. (2009) confunde-se o uso de transformações de variáveis para obter uma relação funcional que represente melhor a realidade com transformações destinadas a obter uma distribuição aproximadamente normal. Apenas a transformação da variável dependente pode ajudar a resolver simultaneamente as duas questões, como ocorre com o uso do logaritmo da renda (e não da própria renda) na estimação de equações de rendimento (equações com as quais se procura explicar como o rendimento de uma pessoa varia em função de suas características e do tipo de ocupação).

Nas páginas 192-193 (Hair et al., 2009, p. 192-193) comfunde-se o problema da multicolinearidade com a especificação correta do modelo de regressão. O grande mérito da técnica de regressão múltipla é possibilitar a estimação do efeito da variação de uma variável explanatória X_1 sobre a variável dependente (*Y*), controlando os efeitos das demais variáveis explanatórias (X_2 , X_3 , ...). Em uma ciência experimental esse efeito específico de X_1 pode ser examinado por meio de um experimento no qual se varia X_1 e se mantêm fixos os valores das demais variáveis que afetam *Y*. Nas ciências sociais, em geral, dispõe-se apenas de dados nos quais os valores de todas as variáveis estão mudando ao mesmo tempo e a regressão múltipla é uma ferramenta muito útil para tentar separar os efeitos específicos de cada variável explanatória sobre a variável dependente. Na p. 192, ao discutirem os exemplos numéricos apresentados na p. 193, Hair et al. (2009) em lugar de explicar os méritos da regressão múltipla, sugerem que se deve confiar mais nos resultados de regressões simples. Cabe ressaltar, ainda, que no Exemplo B (p. 193) deve haver erro nos valores de Z_1 apresentados na tabela A-9, pois não foi possível reproduzir os resultados logo abaixo, na mesma tabela, quando eles dependem dos valores dessa variável.

São bem conhecidos os exemplos de uma correlação simples espúria, devido ao fato de as duas variáveis (X_1 e *Y*) estarem associadas a uma terceira variável (X_2). Assim, se o objetivo é estimar o efeito direto de X_1 sobre *Y*, é necessário introduzir, como controles, na regressão múltipla, todas as demais variáveis exógenas que afetam *Y*. Mas há, também, o perigo de incluir, no modelo, controles inapropriados (*bad controls*). Esse problema é analisado na seção sobre "*Bad control*" do livro de Angrist e Pischke (2009, p. 64-68) e também é sumariamente abordado nas p. 168-171 de Hoffmann (2015). Tudo isso se refere à correta especificação do modelo de regressão múltipla. A multicolinearidade, por outro lado, é um problema que depende da amostra que será utilizada e apenas em casos especiais ela levaria a modificar a especificação do modelo de regressão. O texto de Hair et al. (2009) sobre esses temas, na p. 192, é confuso e inapropriado.

A ideia errônea de que todas as variáveis devem ter distribuição normal leva a uma análise inapropriada das observações discrepantes (denominadas de observações atípicas no livro, nas p. 79-82). Ao fazer uma regressão linear simples de *Y* comtra *X*, por exemplo, não há necessidade de pressupor que essas duas variáveis tenham distribuição conjunta normal e, portanto, não cabe analisar se os dados da amostra são ou não compatíveis com tal pressuposição. A análise da existência de observações discrepantes deve ser baseada nos resíduos da regressão, particularmente no resíduo estudentizado externamente, que permite avaliar se uma observação é discrepante usando uma estimativa da variância do erro que não seja contaminada pela própria observação discrepante. Além das observações discrepantes, é interessante detectar e analisar as observações muito influentes, isto é, as observações cuja exclusão da amostra afetam muito as estimativas dos parâmetros. O texto clássico sobre detecção de observações discrepantes (*outliers*) e observações muito influentes é o livro de Belsley, Kuh e Welsch (1980). Uma exposição didática pode ser encontrada em Hoffmann (2011).

Teoricamente, uma variável com distribuição *t* de *Student* é, por definição, a raiz quadrada de uma variável com distribuição *F* com 1 grau de liberdade

no numerador. A mesma relação ($t^2 = F$) vale para valores calculados dos testes quando eles são corretamente utilizados. Como se trata de uma relação matemática exata, é estranho ler que esses testes "são diretamente comparáveis, pois o valor *t* é aproximadamente a raiz quadrada do valor *F*" (Hair et al., 2009, p. 201).

A rigor, o nível de significância de um teste de hipóteses não deve ser confundido com a probabilidade caudal associada ao valor calculado do teste, comumente denominado **valor p** do teste, como é feito no segundo parágrafo da p. 185. O nível de significância é a probabilidade de rejeitar a hipótese de nulidade se ela for verdadeira, estabelecido ao se adotar determinada maneira de fazer o teste. A probabilidade caudal do teste só pode ser determinada após se calcular a estatística de teste (*t* ou *F*, por exemplo). O resultado do teste é denominado significativo se a probabilidade caudal for menor ou igual ao nível de significância adotado previamente.

Não faz sentido a afirmativa de que "a avaliação da significância de um termo polinomial ou de interação se consegue com a avaliação do R^2 incremental, e não a significância de coeficientes individuais, devido à alta multicolinearidade" (Hair et al., 2009, p. 173). A contribuição de um termo adicional na equação para a soma de quadrados de regressão pode ser testada por meio de um teste *F*, mas ele é perfeitamente equivalente ao teste *t* da nulidade do parâmetro desse termo adicional.

Deve-se distinguir claramente o erro, como um dos termos do modelo de regressão, do desvio (ou resíduo), que é a diferença entre o valor estimado e o valor observado da variável dependente. O que se pode calcular é a soma de quadrados dos desvios (ou soma de quadrados residual), e não a soma de quadrados dos erros (Hair et al., 2009, p. 159).

Tudo indica que os autores do livro têm vasta experiência de aplicação dos procedimentos estatísticos descritos, mas que, infelizmente, nenhum deles conhece, em profundidade, a base estatística desses procedimentos. Finalmente, cabe assinalar um erro de tradução comum em textos de econometria. A palavra "*assuming*" em inglês deveria, em geral, ser traduzida por "pressupondo" ou "admitindo", e não pela palavra "assumindo", apenas porque ela é a palavra mais semelhante em português. Trata-se de exemplo típico de falso cognato.

LITERATURA CITADA

Angrist, J. D.; Pischke, J. -S. **Mostly harmless econometrics**. London: Princeton University Press, 2009.

Belsley, D. A. Kuh, E.; Welsch, R. E. **Regression diagnostics**: identifying influential data and sources of collinearity. New Jersey: John Wiley e Sons, 1980. 287 p.

Hair, J. F. et al. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009. 688 p.

Hoffmann, R. **Análise de regressão**: uma introdução à econometria. São Paulo: Hucitec, 2015. 378 p.

_____. **Análise estatística de relações lineares e não-lineares**. São Paulo: LP-Books, 2011. 272 p.

Recebido em 29/11/2015. Liberado para publicação em 12/02/2016.