NARESH

MALHOTRA

6ª EDIÇÃO

PESQUISA de MARKETING

UMA ORIENTAÇÃO APLICADA

M249p Malhotra, Naresh K.

Pesquisa de marketing [recurso eletrônico] : uma orientação aplicada / Naresh K. Malhotra ; tradução: Leme Belon Ribeiro, Monica Stefani ; revisão técnica: Janaina de Moura Engracia Giraldi. – 6. ed. – Dados eletrônicos. – Porto Alegre : Bookman, 2012.

Editado também como livro impresso em 2012.
ISBN 978-85-407-0062-8

1. Marketing. 2. Pesquisa de marketing. I. Título.

CDU 658.8:005.52

Catalogação na publicação: Ana Paula M. Magnus – CRB 10/2052

Como fazer análise de regressão bivariada

As etapas a serem desenvolvidas na análise de regressão bivariada são descritas na Figura 17.2. Suponha que o pesquisador pretenda explicar atitudes em relação à cidade de residência em termos da duração da residência (ver Tabela 17.1). Para deduzir tais relacionamentos, é muitas vezes útil examinar, em primeiro lugar, um diagrama de dispersão.

Fazer o diagrama de dispersão

Um diagrama de dispersão é um gráfico dos valores de duas variáveis para todos os casos ou observações. É costume grafar a variável dependente no eixo vertical e a variável independente no eixo horizontal. O diagrama de dispersão serve para determinar a forma da relação entre as variáveis e pode alertar o pesquisador quanto a determinados padrões dos dados, ou possíveis problemas. Quaisquer combinações incomuns das duas variáveis podem ser facilmente identificadas. A Figura 17.3 mostra um gráfico de Y (atitude em relação à cidade) comparado com X (tempo de residência). Os pontos parecem dispor-se em uma faixa que vai da esquerda inferior para a direita superior. Pode-se ver logo o padrão: à medida que uma das variáveis aumenta, a outra também aumenta. Por esse gráfico, parece que a relação entre X e Y é linear, podendo ser descrita por uma linha reta. Como determinar a reta que melhor descreve os dados?

A técnica mais comum de ajuste de uma linha reta a um diagrama de dispersão é o **procedimento dos mínimos quadrados**. Essa técnica determina a reta de melhor ajuste minimizando o quadrado das distâncias verticais de todos os pontos a partir da reta e esse procedimento é chamado de regressão dos mínimos quadrados ordinários (MQO*). A reta de melhor ajuste é chamada de *reta de regressão.* Qualquer ponto que não esteja sobre a reta de regressão não é plenamente considerado. A distância vertical do ponto até a reta é o erro, e_j (ver Figura 17.5). Elevam-se ao quadrado as distâncias de todos os pontos até a reta e somam-se os resultados, obtendo-se a soma dos quadrados dos erros, que é a medida do total dos erros, Σe_j^2. Ao ajustar a reta, o procedimento de mínimos quadrados minimiza a soma dos quadrados dos erros. Colocando-se Y no eixo vertical e X no eixo horizontal, como na Figura 17.5, a reta de melhor ajuste é chamada de regressão de Y em função de X, pois as distâncias verticais são minimizadas. O diagrama de dispersão indica se a relação entre Y e X pode ser modelada como em uma linha reta e, consequentemente, se o modelo de regressão bivariada é apropriado.

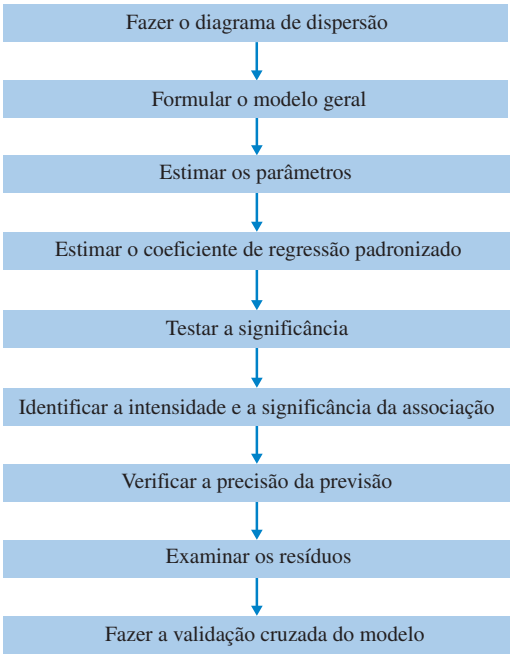

FIGURA 17.2 **Como fazer uma análise de regressão bivariada.**

Procedimento dos mínimos quadrados
Técnica de ajuste de uma linha reta a um diagrama de dispersão pela minimização do quadrado das distâncias verticais de todos os pontos a partir da reta. Tal procedimento é denominado regressão dos mínimos quadrados ordinários.

Formular o modelo de regressão bivariada

No modelo de regressão bivariada, a forma geral de uma reta é:

$$Y = \beta_0 + \beta_1 X$$

onde

Y = variável dependente ou de critério
X = variável independente ou previsora
β_0 = intercepto da reta
β_1 = coeficiente angular da reta

SPSS Arquivo de Saída

SAS Arquivo de Saída

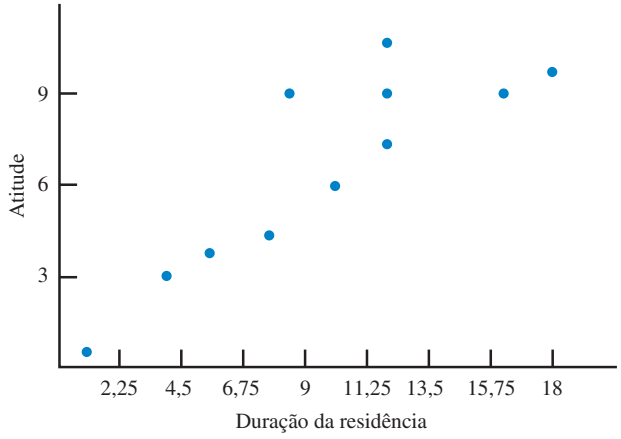

FIGURA 17.3 **Gráfico da atitude *versus* tempo de residência.**

* N de T.: Ordinary least-squares (OLS).

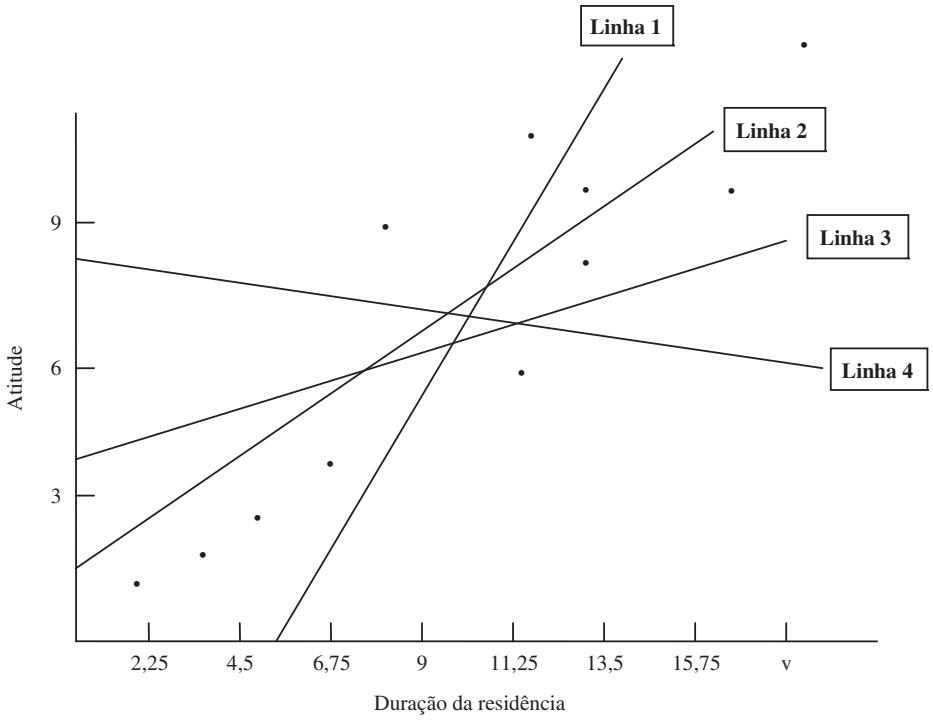

FIGURA 17.4 **Que linha reta é a melhor?**

Este modelo implica uma relação determinística, no sentido de que Y é completamente determinado por X. O valor de Y pode ser perfeitamente previsto desde que conheçamos β_0 e β_1. Em pesquisa de marketing, entretanto, poucas relações são determinísticas. Por isso, o processo de regressão acrescenta um termo de erro para responder pela natureza probabilística ou estocástica da relação. A equação básica da regressão se escreve:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Onde e_i é o termo de erro associado à i-ésima observação.[8] A estimação dos parâmetros de regressão, β_0 e β_1, é relativamente simples.

Estimar os parâmetros

Na maioria dos casos, β_0 e β_1 são desconhecidos e devem ser estimados com base nas observações amostrais, mediante a equação

$$\hat{Y}_i = a + bx_i$$

onde $\hat{Y}_i$ é o valor estimado, ou previsto, de Y_i, e a e b estimam β_0 e β_1, respectivamente. A constante b costuma ser chamada de coeficiente de regressão não padronizado. É o coeficiente angular da reta de regressão e indica a variação esperada em Y quando X varia de uma unidade. As fórmulas para o cálculo de a e b são simples.[9]

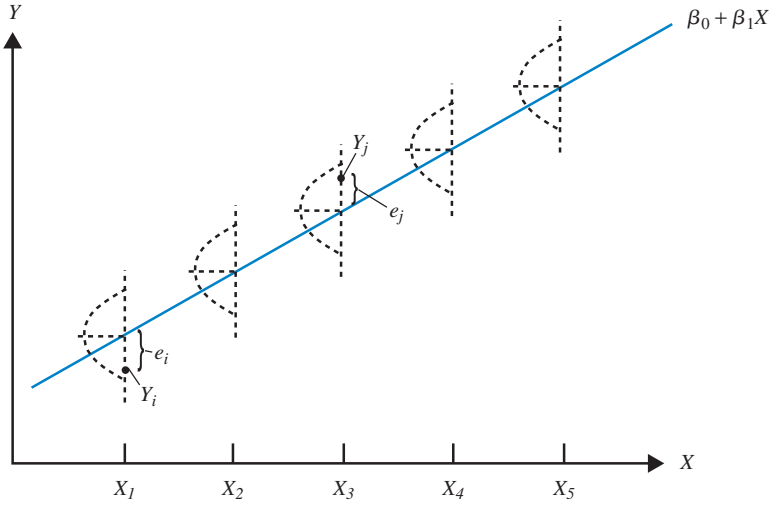

FIGURA 17.5 **Regressão bivariada.**

O coeficiente angular, *b*, pode ser calculado em termos da covariância entre *X* e Y *(COV_{xy})* e da variância de *X* como:

$$b = \frac{COV_{xy}}{s_x^2}$$
$$= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^{n}X_iY_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n}X_i^2 - n\overline{X}^2}$$

Pode-se então calcular o intercepto *a* como:

$$a = \overline{Y} - b\overline{X}$$

Para os dados da Tabela 17.1, pode-se ilustrar a estimação dos parâmetros como segue:

$$\sum_{i=1}^{12}X_iY_i = (10)(6) + (12)(9) + (12)(8) + (4)(3) + (12)(10) + (6)(4) + (8)(5) + (2)(2) + (18)(11) + (9)(9) + (17)(10) + (2)(2)$$
$$= 917$$
$$\sum_{i=1}^{12}X_i^2 = 10^2 + 12^2 + 12^2 + 4^2 + 12^2 + 6^2 + 8^2 + 2^2 + 18^2 + 9^2 + 17^2 + 2^2$$
$$= 1.350$$

Convém lembrar, de cálculos anteriores da correlação simples, que

$$\overline{X} = 9{,}333$$
$$\overline{Y} = 6{,}583$$

Dado $n = 12$, pode-se calcular *b* como segue:

$$b = \frac{917 - (12)(9{,}333)(6{,}583)}{1350 - (12)(9{,}333)^2}$$
$$= 0{,}5897$$

$$a = \overline{Y} - b\overline{X}$$
$$= 6{,}583 - (0{,}5897)(9{,}333)$$
$$= 1{,}0793$$

Observe que esses coeficientes foram estimados com base nos dados brutos (não transformados). Se a padronização dos dados for considerada desejável, o cálculo dos coeficientes padronizados também pode ser feito de imediato.

Estimar o coeficiente de regressão padronizado

Padronização é o procedimento pelo qual os dados brutos são transformados em novas variáveis, com média 0 e variância 1 (Capítulo 14). Quando os dados são padronizados, o intercepto toma o valor 0. Usa-se a expressão *coeficiente beta* ou *peso beta* para denotar o coeficiente de regressão padronizado. Neste caso, o coeficiente angular obtido pela regressão de *Y* sobre *X*, B_{yx}, é o mesmo que o coeficiente angular obtido pela regressão de *X* sobre *Y*, B_{xy}. Além disso, cada um desses coeficientes de regressão é igual à correlação simples entre *X* e *Y*.

$$B_{yx} = B_{xy} = r_{xy}$$

Há uma relação simples entre os coeficientes de regressão padronizados e não padronizados:

$$B_{yx} = b_{yx}(s_x/s_y)$$

Para os resultados de regressão dados na Tabela 17.2, o valor do coeficiente beta é estimado em 0,9361. Observe que esse também é o valor de *r* calculado anteriormente neste capítulo.

Uma vez estimados, os parâmetros podem ser testados quanto à sua significância.

Testar a significância

Podemos testar a significância estatística da relação linear entre *X* e Y examinando as hipóteses:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

A hipótese nula implica que não há qualquer relação linear entre *X* e Y. A hipótese alternativa é que há alguma relação, positiva ou

SPSS Arquivo de Saída

SAS Arquivo de Saída

TABELA 17.2

Regressão bivariada

R múltiplo	0,93608
R^2	0,87624
R^2 ajustado	0,86387
Erro padrão	1,22329

	gl	Análise da variância Soma de quadrados	Quadrado médio
Regressão	1	105,95222	105,95222
Residual	10	14,96444	1,49644

$F = 70{,}80266$ Significância de $F = 0{,}0000$

Variáveis na equação

Variável	*b*	EP_B	Beta (*B*)	*t*	Significância de *t*
Tempo de residência	0,58972	0,07008	0,93608	8,414	0,0000
(constante)	1,07932	0,74335		1,452	0,1772

negativa, entre *X* e Y. Em geral, faz-se um teste bicaudal. Pode-se utilizar uma estatística *t* com *n* – 2 graus de liberdade, onde

$$t = \frac{b}{EP_b}$$

EP_b denota o desvio-padrão de b e é chamado de *erro padrão*.[10] A distribuição t foi estudada no Capítulo 15.

Com um programa de computador, a regressão da atitude sobre o tempo de residência, utilizando-se os dados da Tabela 17.1, apresentou os resultados da Tabela 17.2. O intercepto *a* é 1,0793 e o coeficiente angular *b* é 0,5897. Portanto, a equação estimada é:

$$\text{Atitude } (\hat{Y}) = 1{,}0793 + 0{,}5897 \text{ (tempo de residência)}$$

O erro padrão ou o desvio-padrão de *b* é estimado em 0,07008, e o valor da estatística *t* é t = 0,5897/0,07008 = 8,414, com *n* – 2 = 10 graus de liberdade. Na Tabela 4 dos Apêndices Estatísticos, vemos que o valor crítico de *t* com 10 graus de liberdade e $\alpha = 0{,}05$ é 2,228 para um teste bicaudal. Como o valor calculado de *t* é maior do que o valor crítico, rejeitamos a hipótese nula. Logo, existe uma relação linear significativa entre a atitude em relação à cidade e o tempo de residência nela. O sinal positivo do coeficiente angular indica que essa relação é positiva. Em outras palavras, os que residem há mais tempo na cidade têm atitude mais favorável em relação a ela. A implicação para gerentes, autoridades municipais e políticos é a mesma que a discutida para a correlação simples, sujeita à representatividade da amostra.

Determinar a intensidade e a significância da associação

Uma inferência relacionada envolve a determinação da intensidade e da significância da associação entre *Y* e *X*. A intensidade da relação é medida pelo coeficiente de determinação, r^2. Na regressão bivariada, r^2 é o quadrado do coeficiente de correlação simples obtido ao correlacionar as duas variáveis. O coeficiente r^2 varia entre 0 e 1 e indica a proporção da variação total em *Y* que é ocasionada pela variação em *X*. A decomposição da variação total em *Y* é análoga à da análise da variância (Capítulo 16). Conforme mostra a Figura 17.6, a variação total, SQ_y, pode ser decomposta na variação proporcionada pela reta de regressão, SQ_{reg}, e o erro ou a variação residual, SQ_{erro} ou SQ_{res}, como segue:

$$SQ_y = SQ_{reg} + SQ_{res}$$

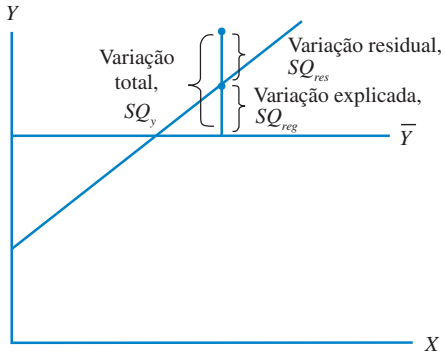

FIGURA 17.6 **Decomposição da variação total na regressão bivariada.**

Onde

$$SQ_y = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$SQ_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

$$SQ_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Pode-se então calcular a intensidade da associação:

$$r^2 = \frac{SQ_{reg}}{SQ_y}$$

$$= \frac{SQ_y - SQ_{res}}{SQ_y}$$

Para ilustrar os cálculos de r^2, consideremos novamente o efeito do tempo de residência sobre a atitude em relação à cidade. Pelos cálculos anteriores do coeficiente de correlação simples, sabemos que:

$$SQ_y = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$= 120{,}9168$$

Os valores previstos $(\hat{Y})$ podem ser calculados com auxilio da equação de regressão:

$$\text{Atitude } (\hat{Y}) = 1{,}0793 + 0{,}5897 \text{ (tempo de residência)}$$

Para a primeira observação da Tabela 17.1, esse valor é:

$$(\hat{Y}) = 1{,}0793 + 0{,}5897 \text{ x } 10 = 6{,}9763$$

Para cada observação sucessiva, os valores previstos são, pela ordem: 8,1557, 8,1557, 3,4381, 8,1557, 4,6175, 5,7969, 2,2587, 11,6939, 6,3866, 11,1042, 2,2587. Portanto,

$$\begin{aligned} SQ_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 &= (6{,}9763 - 6{,}5833)^2 + (8{,}1557 - 6{,}5833)^2 \\ &+ (8{,}1557 - 6{,}5833)^2 + (3{,}4381 - 6{,}5833)^2 \\ &+ (8{,}1557 - 6{,}5833)^2 + (4{,}6175 - 6{,}5833)^2 \\ &+ (5{,}7969 - 6{,}5833)^2 + (2{,}2587 - 6{,}5833)^2 \\ &+ (11{,}6939 - 6{,}5833)^2 + (6{,}3866 - 6{,}5833)^2 \\ &+ (11{,}1042 - 6{,}5833)^2 + (2{,}2587 - 6{,}5833)^2 \\ &= 0{,}1544 + 2{,}4724 + 2{,}4724 + 9{,}8922 + 2{,}4724 \\ &+ 3{,}8643 + 0{,}6184 + 18{,}7021 + 26{,}1182 \\ &+ 0{,}0387 + 20{,}4385 + 18{,}7021 \\ &= 105{,}9524 \end{aligned}$$

$$\begin{aligned} SQ_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 &= (6 - 6{,}9763)^2 + (9 - 8{,}1557)^2 \\ &+ (8 - 8{,}1557)^2 + (3 - 3{,}4381)^2 \\ &+ (10 - 8{,}1557)^2 + (4 - 4{,}6175)^2 \\ &+ (5 - 5{,}7969)^2 + (2 - 2{,}2587)^2 \\ &+ (11 - 11{,}6939)^2 + (9 - 6{,}3866)^2 \\ &+ (10 - 11{,}1042)^2 + (2 - 2{,}2587)^2 \\ &= 14{,}9644 \end{aligned}$$

Pode-se ver que $SQ_y = SQ_{reg} + SQ_s$. Além disso,

$$r^2 = \frac{SQ_{reg}}{SQ_y} = \frac{105{,}9524}{120{,}9168} = 0{,}8762$$

Outro teste equivalente para examinar a significância da relação linear entre X e Y (significância de b) é o teste da significância do coeficiente de determinação. As hipóteses, neste caso, são:

$$H_0: R^2_{pop} = 0$$
$$H_1: R^2_{pop} > 0$$

A estatística de teste apropriada é a estatística F:

$$F = \frac{SQ_{reg}}{SQ_{res}/(n-2)}$$

que tem distribuição F com 1 e $n-2$ graus de liberdade. O teste F é uma forma generalizada do teste t (ver Capítulo 15). Se uma variável aleatória tiver distribuição t com n graus de liberdade, então t^2 tem distribuição F com 1 e n graus de liberdade. Logo, o teste F para testar a significância do coeficiente de determinação é equivalente a testar as seguintes hipóteses:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Ou

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

Pela Tabela 17.2, pode-se ver que:

$$r^2 = \frac{105{,}9524}{(105{,}9524 + 14{,}9644)} = 0{,}8762$$

que é o mesmo valor já calculado anteriormente. O valor da estatística F é:

$$F = \frac{105{,}9524}{(14{,}9644/10)} = 70{,}8027$$

com 1 e 10 graus de liberdade. A estatística F calculada excede o valor crítico de 4,96 obtido na Tabela 5 dos Apêndices Estatísticos. Portanto, a relação é significativa ao nível $\alpha = 0{,}05$, corroborando os resultados do teste t. Se a relação entre X e Y for significativa, faz sentido prever os valores de Y com base nos valores de X e estimar a precisão da predição.

Verificar a precisão da previsão

Para estimar a precisão dos valores previstos, $\hat{Y}$, convém calcular o erro padrão da estimativa, *EPE*. Essa estatística é o desvio-padrão dos valores reais de Y em relação aos valores $\hat{Y}$ previstos.

$$EPE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2}}$$

Ou

$$EPE = \sqrt{\frac{SQ_{res}}{n-2}}$$

ou, de forma mais geral, se há k variáveis independentes,

$$EPE = \sqrt{\frac{SQ_{res}}{n-k-1}}$$

O *EPE* pode ser interpretado como uma espécie de resíduo médio ou erro médio na predição de Y com base na equação de regressão.[11]

Podem surgir dois casos de previsão. O pesquisador pode querer predizer o valor médio de Y para todos os casos com um determinado valor de X, digamos X_0, ou prever o valor de Y para um único caso. Em ambas as situações, o valor previsto é o mesmo e é dado por $\hat{Y}$, onde:

$$\hat{Y} = a + bX_0$$

Entretanto, o erro padrão é diferente nas duas situações, embora em ambas seja uma função do *EPE*. Para grandes amostras, o erro padrão na predição do valor médio de Y é $EPE/\sqrt{n}$, e para predizer valores individuais de Y é *EPE*. Logo, a construção de intervalos de confiança (ver Capítulo 12) para os valores previstos varia, conforme estejamos prevendo o valor médio ou o valor para uma única observação.

Para os dados da Tabela 17.2, o *EPE* é estimado conforme a seguir:

$$EPE = \sqrt{\frac{14{,}9644}{(12-2)}} = 1{,}22329$$

As duas etapas finais da regressão bivariada, a saber, o exame dos resíduos e a validação cruzada do modelo, serão consideradas mais adiante.

Suposições

O modelo de regressão exige várias suposições na estimativa dos parâmetros e no teste de significância, conforme mostra a Figura 17.5:

1. O termo de erro tem distribuição normal. Para cada valor fixo de X, a distribuição de Y é normal.[12]
2. As médias de todas essas distribuições normais de Y, dado X, situam-se em uma reta de coeficiente angular b.
3. A média do termo de erro é 0.
4. A variância do termo de erro é constante. Essa variância não depende dos valores que X toma.
5. Os termos de erro não são correlacionados. Em outras palavras, as observações são extraídas independentemente umas das outras.

Mediante exame dos resíduos, pode-se obter uma visualização do alcance dessas suposições. Esse assunto é abordado na próxima seção sobre regressão múltipla.[13]

PESQUISA ATIVA

Associando a propaganda e as vendas da Ford

Acesse www.ford.com e pesquise na Internet (utilizando um dispositivo de busca) e no banco de dados on-line de sua biblioteca informações sobre as relações entre propaganda e vendas para fabricantes de automóveis.

Formule um modelo de regressão bivariada explicando a relação entre propaganda e vendas na indústria automobilística.

Como diretor de marketing da Ford Motor Company, como você determinaria suas despesas com propaganda?

Regressão múltipla

A **regressão múltipla** envolve uma única variável dependente e duas ou mais variáveis independentes. As questões suscitadas no contexto da regressão bivariada também podem ser resolvidas via regressão múltipla, com a consideração de variáveis independentes adicionais:

- A variação nas vendas pode ser explicada em termos da variação nas despesas de propaganda, nos preços e no nível de distribuição?
- A variação na participação de mercado pode ser decorrência do tamanho da equipe de vendas, das despesas de propaganda e dos orçamentos de promoção de vendas?
- A conscientização dos consumidores quanto à qualidade é determinada pela sua percepção quanto a preços, imagem e atributos da marca?

regressão múltipla
Técnica estatística que desenvolve simultaneamente uma relação matemática entre duas ou mais variáveis independentes e uma variável dependente intervalar.

A regressão múltipla pode também responder a outras questões:

- Quanto da variação nas vendas pode ser explicado pelas despesas de propaganda, pelos preços e pelo nível de distribuição?
- Qual é a contribuição das despesas de propaganda para explicar a variação nas vendas, quando os níveis de preços e de distribuição são controlados?
- Que níveis de venda podemos esperar, dados os níveis de despesas, de preços e de distribuição?

Pesquisa real

Marcas globais – anúncios locais

Os europeus são receptivos a produtos de outros países, mas quando se trata de propaganda, preferem a "prata da casa". Em uma pesquisa feita por Yankelovich and Partners (www.yankelovich.com) e suas afiliadas, constatou-se que os comerciais favoritos da maioria dos europeus se referiam a marcas locais, embora eles não hesitem em comprar produtos de marcas estrangeiras. Respondentes na França, na Alemanha e no Reino Unido indicaram a Coca-Cola como o refrigerante mais comprado. Entretanto, os franceses escolheram como favorito o anúncio da afamada e premiada água Perrier. Na Alemanha, o anúncio preferido foi o de uma marca alemã de cerveja sem álcool, Clausthaler. No Reino Unido, porém, a Coca-Cola foi não apenas a bebida favorita como a preferida na propaganda. À luz desses resultados, a questão importante é: a propaganda ajuda? Ela contribui para aumentar a probabilidade de venda da marca, ou apenas mantém em alta o conhecimento da marca? Uma forma de resolver esse problema consiste em fazer uma regressão na qual a variável dependente é a probabilidade de compra da marca e as variáveis independentes são as avaliações das qualidades do produto e avaliações da propaganda. Podem ser elaborados modelos separados, com propaganda e sem ela, para avaliar qualquer diferença significativa na contribuição. Podem também ser analisados testes t para verificar a contribuição significativa tanto dos atributos da marca como da propaganda. Os resultados indicam até que ponto a propaganda desempenha um papel importante nas decisões de compra da marca. Junto a esses resultados, um estudo realizado recentemente revelou que a tentativa de construir fidelidade na compra de uma marca por meio de promoções de vendas não é uma forma desejável de alcançar esse objetivo. Segundo o estudo, as promoções de vendas apenas incentivam uma troca momentânea de marca e simplesmente melhoram o desempenho a curto prazo para as empresas. Além disso, no longo prazo, uma promoção de vendas pode implicar uma baixa qualidade, ou imagem de marca instável frente aos consumidores, ou pode inclusive confundi-los, o que poderia também levar a um declínio na fidelidade à marca. Os resultados desse estudo mostram que sacrificar a propaganda e confiar nas promoções de vendas reduz as associações de marca, o que finalmente acabará levando a uma diminuição nas compras por fidelidade à marca.[14] ■

A forma geral do **modelo de regressão múltipla** é:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + e$$

que é estimado pela seguinte equação:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

modelo de regressão múltipla
Equação usada para explicar os resultados da análise de regressão múltipla.

Como anteriormente, o coeficiente a representa o intercepto, mas os bs são agora coeficientes de regressão parcial. O critério de mínimos quadrados estima os parâmetros de forma a minimizar o erro total, SQ_{res}. Esse processo também maximiza a correlação entre os valores reais de Y e os valores previstos, $\hat{Y}$. Todas as suposições feitas na regressão bivariada aplicam-se também à regressão múltipla. A seguir definimos algumas estatísticas associadas e, posteriormente, descrevemos o processo da análise de regressão múltipla.[15]

Estatísticas associadas à regressão múltipla

A maioria das estatísticas e dos termos estatísticos utilizados na regressão bivariada também se aplicam à regressão múltipla. Além disso, são empregadas as seguintes estatísticas:

R^2 *ajustado*: R^2, coeficiente de determinação múltipla, é ajustado para o número de variáveis independentes e para o tamanho da amostra levando em conta os retornos decrescentes. Após as primeiras variáveis, as variáveis independentes adicionais não oferecem grande contribuição.

Coeficiente de determinação múltipla: a intensidade de associação em regressão múltipla é medida pelo quadrado do coeficiente de correlação múltipla, R^2, que é chamado também de *coeficiente de determinação múltipla*.

***Teste* F:** o teste F é usado para testar a hipótese nula de que o coeficiente de determinação múltipla na população, R^2_{pop}, é zero. Isso equivale a testar a hipótese nula H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$. A estatística de teste tem distribuição F com k e $(n - k - 1)$ graus de liberdade.

Teste F parcial: pode-se testar a significância de um coeficiente de regressão parcial, β_i, de X_i, com auxílio de uma estatística F incremental. A estatística F incremental se baseia no incremento da soma explicada de quadrados resultante da adição da variável independente X_i à equação de regressão após terem sido incluídas todas as outras variáveis independentes.

Coeficiente de regressão parcial: o coeficiente de regressão parcial, b_1, denota a variação no valor previsto, $\hat{Y}$, por unidade de variação em X_1 quando as outras variáveis independentes, X_2 a X_k, são mantidas constantes.

Como fazer análise de regressão múltipla

Os passos para a elaboração de uma análise de regressão múltipla são similares aos adotados na regressão bivariada. O foco da discussão reside nos coeficientes de regressão parcial, intensidade de associação, teste de significância e avaliação de resíduos.

Coeficientes de regressão parcial

Para entender o significado de um coeficiente de regressão parcial, consideremos um caso em que há duas variáveis independentes, de forma que

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

Observemos inicialmente que a magnitude relativa do coeficiente de regressão parcial de uma variável independente costuma ser diferente daquela do seu coeficiente de regressão bivariada. Em outras palavras, o coeficiente de regressão parcial, b_1, será diferente do coeficiente de regressão, b, obtido ao fazermos Y regredir sobre X_1 somente. Isso ocorre porque X_1 e X_2 são geralmente correlacionadas. Na regressão bivariada, não consideramos X_2, e qualquer variação em Y compartilhada por X_1 e X_2 foi atribuída a X_1. Entretanto, no caso de variáveis independentes múltiplas, isso não se justifica mais.

A interpretação do coeficiente de regressão parcial, b_1, é que ele representa a variação esperada em Y quando X_1 varia de uma unidade mas X_2 é mantida constante ou controlada de outra forma. De maneira semelhante, b_2 representa a variação esperada em Y para uma variação unitária em X_2, quando X_1 é mantida constante. É, pois, adequada a designação de coeficientes de regressão parcial para b_1 e b_2. Pode-se ver também que os efeitos combinados de X_1 e X_2 sobre Y são aditivos. Em outras palavras, se X_1 e X_2 variam cada um de uma unidade, a variação esperada em Y será $(b_1 + b_2)$.

Conceitualmente, pode-se ilustrar como segue a relação entre o coeficiente de regressão bivariada e o coeficiente de regressão parcial. Suponhamos que se deva remover de X_1 o efeito de X_2. Para tanto, fazemos uma regressão de X_1 sobre X_2. Em outras palavras, estimaríamos a equação $\hat{X}_1 = a + bX_2$ e calcularíamos o resíduo X_r. $= (X_1 - \hat{X}_1)$. O coeficiente de regressão parcial, b_1, é igual ao coeficiente de regressão bivariada, b_r, obtido da equação $\hat{Y} = a + b_rX_r$. Em outras palavras, o coeficiente de regressão parcial, b_1, é igual ao coeficiente de regressão, b_r, entre Y e os resíduos de X_1 dos quais foi removido o efeito de X_2. Pode-se dar interpretação análoga ao coeficiente parcial b_2.

A extensão ao caso de k variáveis é imediata. O coeficiente de regressão parcial, b_1, representa a variação esperada em Y quando X_1 varia de uma unidade e X_2 a X_k são mantidas constantes. Pode ser interpretado também como o coeficiente de regressão bivariada, b, para a regressão de Y sobre os resíduos de X_1 quando o efeito de X_2 a X_k foi removido de X_1.

Os coeficientes beta são os coeficientes de regressão parcial obtidos quando todas as variáveis $(Y,X_1,X_2,\ldots X_k)$ foram padronizadas com média 0 e variância 1 antes de estimar a equação de regressão. A relação dos coeficientes padronizados para os não padronizados é a mesma que a anterior:

$$B_1 = b_1\left(\frac{s_{x_1}}{s_y}\right)$$

.

.

$$B_k = b_k\left(\frac{s_{x_k}}{s_y}\right)$$

O intercepto e os coeficientes de regressão parcial são estimados ao resolver um sistema de equações simultâneas obtido ao diferenciar e igualar a 0 as derivadas parciais. Como esses coeficientes são estimados automaticamente por vários programas de computador, não vamos apresentar os detalhes. Cabe notar, entretanto, que as equações não podem ser resolvidas se (1) o tamanho da amostra, n, não superar o número de variáveis independentes, k, ou (2) uma variável independente tiver correlação perfeita com outra.

Suponha que, ao explicar a atitude em relação à cidade, introduzamos uma segunda variável – a importância atribuída ao clima. A Tabela 17.1 apresenta os dados dos 12 entrevistados em um teste preliminar sobre atitude em relação à cidade, tempo de residência e importância atribuída ao clima. A Tabela 17.3 exibe os resultados da análise de regressão múltipla. O coeficiente de regressão parcial

SPSS Arquivo de Saída

SAS Arquivo de Saída

TABELA 17.3

Regressão múltipla

R múltiplo	0,97210
R^2	0,94498
R^2 ajustado	0,93276
Erro padrão	0,85974

	gl	Análise da variância Soma de quadrados	Quadrado médio
Regressão	2	114,26425	57,13213
Resíduo	9	6,65241	0,73916

$F = 77,29364$ Significância de $F = 0,0000$

Variáveis na equação

Variável	*b*	SE_B	Beta (*B*)	*t*	Significância de *t*
Importância	0,28865	0,08608	0,31382	3,353	0,0085
Tempo	0,48108	0,05895	0,76363	8,160	0,0000
(Constante)	0,33732	0,56736		0,595	0,5668

para o tempo de residência (X_1) agora é 0,48108, diferente do que era no caso bivariado. O coeficiente beta correspondente é 0,7636. O coeficiente de regressão parcial para a importância atribuída ao clima (X_2) é 0,28865, com um coeficiente beta de 0,3138. A equação estimada de regressão é:

$$(\hat{Y}) = 0{,}33732 + 0{,}48108X_1 + 0{,}28865X_2$$

Ou

$$\text{Atitude} = 0{,}33732 + 0{,}48108\ (\text{Tempo}) + 0{,}28865\ (\text{Importância})$$

Essa equação pode ser utilizada para vários fins, inclusive a previsão de atitudes em relação à cidade com base no conhecimento do tempo de residência dos entrevistados na cidade e a importância que eles atribuem ao clima.

Intensidade de associação

Pode-se determinar a intensidade da relação estipulada pela equação de regressão utilizando medidas adequadas de associação. A variação total se decompõe como no caso bivariado:

$$SQ_y = SQ_{reg} + SQ_{res}$$

Onde:

$$SQ_y = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

$$SQ_{reg} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

$$SQ_{res} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

A intensidade da associação é medida pelo quadrado do coeficiente de correlação múltipla, R^2, também chamado de *coeficiente de determinação múltipla.*

$$R^2 = \frac{SQ_{reg}}{SQ_y}$$

O coeficiente de correlação múltipla, R, também pode ser visto como o coeficiente de correlação simples, r, entre Y e $\overline{Y}$. São dignos de nota vários pontos sobre as características de R^2. O coeficiente de determinação múltipla, R^2, não pode ser menor que o maior bivariado, r^2, de qualquer variável independente individual com a variável dependente. R^2 será maior quando as correlações entre as variáveis independentes forem baixas. Se as variáveis independentes forem estatisticamente independentes (não correlacionadas), então R^2 será a soma dos r^2 bivariados de cada variável independente com a variável dependente. R^2 não pode decrescer quando se acrescentam mais variáveis independentes à equação de regressão. Entretanto, em virtude dos retornos decrescentes, as variáveis adicionais não dão qualquer contribuição sensível.[16] Por essa razão, R^2 é ajustado para o número de variáveis independentes e o tamanho da amostra pela fórmula:

$$R^2 \text{ ajustado} = R^2 - \frac{k(1 - R^2)}{n - k - 1}$$

Para os resultados de regressão dados na Tabela 17.3, o valor de R^2 é:

$$R^2 = \frac{114{,}2643}{(114{,}2643 + 6{,}6524)} = 0{,}9450$$

Esse valor é maior do que o valor de r^2, 0,8762, obtido no caso bivariado, que é o quadrado da correlação simples (momento-produto) entre atitude em relação à cidade e tempo de residência. O R^2 obtido na regressão múltipla também é maior do que o quadrado da correlação simples entre atitude e importância atribuída ao clima (que pode ser estimada em 0,5379). O R^2 ajustado é estimado em:

$$R^2 \text{ ajustado} = 0{,}9450 - \frac{2(1{,}0 - 0{,}9450)}{(12 - 2 - 1)} = 0{,}9328$$

Observe que o valor de R^2 ajustado está próximo de R^2 e ambos são maiores do que r^2 para o caso bivariado. Isso sugere que o acréscimo da segunda variável independente, importância atribuída ao clima, dá uma contribuição para explicar a variação da atitude em relação à cidade.

Teste da significância

Esse teste envolve o teste da significância não só da equação de regressão global como dos coeficientes específicos de regressão parcial. A hipótese nula para o teste global é que o coeficiente de determinação múltipla na população, $R^2_{pop,}$ é zero.

$$H_0: R^2_{pop} = 0$$

Isso equivale à seguinte hipótese nula:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

O teste global pode ser feito com uma estatística F:

$$F = \frac{SQ_{reg}/k}{SQ_{res}/(n - k - 1)} = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

que tem distribuição F com k e $(n - k - 1)$ graus de liberdade.[17] Para os resultados de regressão múltipla da Tabela 17.3,

$$F = \frac{114{,}2643/2}{6{,}6524/9} = 77{,}2936$$

significativo ao nível $\alpha = 0{,}05$.

Se a hipótese nula for rejeitada, pelo menos um coeficiente de regressão parcial da população é diferente de zero. Para determinar que coeficientes específicos (β'_is) são diferentes de zero, são necessários testes adicionais. O teste da significância dos β'_is pode ser feito da maneira análoga ao do caso bivariado, utilizando testes t. A significância do coeficiente parcial da importância atribuída ao clima pode ser testada pela seguinte equação:

$$t = \frac{b}{EP_b} = \frac{0{,}2887}{0{,}08608} = 3{,}353$$

que tem distribuição t com $n - k - 1$ graus de liberdade. Esse coeficiente é significativo ao nível $\alpha = 0{,}05$. Testa-se de maneira análoga a significância do coeficiente do tempo de residência, que constatamos ser significativa. Logo, tanto o tempo de residência como a importância atribuída ao clima são importantes para explicar a atitude em relação à cidade.

Alguns programas de computador contêm um teste *F* equivalente, geralmente chamado de *teste F parcial*. Esse teste envolve uma decomposição da soma de quadrados de regressão, SQ_{reg}, em componentes relativos a cada variável independente. Na abordagem padrão, isso se faz supondo que cada variável independente tenha sido acrescentada à equação de regressão após terem sido incluídas todas as outras variáveis independentes. O incremento na soma de quadrados explicada, resultante da adição de uma variável independente, é o componente da variação atribuída àquela variável, e se denota por SQ_{xi}.[18] Testa-se a significância do coeficiente de regressão parcial para esta variável, com auxílio de uma estatística *F* incremental:

$$F = \frac{SQ_{x_i}/1}{SQ_{res}/(n - k - 1)}$$

que tem distribuição *F* com 1 e $(n - k - 1)$ graus de liberdade.

Embora um valor alto de R^2 e coeficientes significativos de regressão parcial sejam satisfatórios, a eficácia do modelo de regressão deve ser avaliada mais cuidadosamente mediante o exame dos resíduos.

Exame dos resíduos

Um *resíduo* é a diferença entre o valor observado de Y_i, e o valor previsto pela equação de regressão, $\hat{Y}_i$. Os resíduos são utilizados no cálculo de várias estatísticas associadas à regressão. Além disso, os diagramas de dispersão, em que são diagramados os resíduos *versus* os valores previstos, $\hat{Y}_i$, tempo, ou variáveis previsoras, permitem uma visão adequada das suposições fundamentais e da validade do modelo ajustado.[19]

resíduo
Diferença entre o valor observado de *Y*, e o valor previsto pela equação de regressão, $\hat{Y}_i$.

A suposição de um termo de erro distribuído normalmente pode ser avaliada construindo um histograma dos resíduos padronizados. Uma verificação visual revela se a distribuição é normal. Também é útil examinar o gráfico de probabilidade de normalidade dos resíduos padronizados, que mostra os resíduos padronizados comparados a resíduos padronizados esperados de uma distribuição normal. Se os resíduos observados forem normalmente distribuídos, eles ficarão em uma reta de 45°. Além disso, dê uma olhada na tabela de estatísticas residuais e identifique quaisquer valores padronizados previstos ou resíduos padronizados que são maiores do que ± um ou dois desvios-padrão. Essas porcentagens podem ser comparadas com o que se poderia esperar com a distribuição normal (68 e 95%, respectivamente). Com o teste K-S de uma amostra, fazemos uma avaliação mais formal.

A suposição de variância constante do termo de erro pode ser examinada diagramando os resíduos *versus* os valores previstos da variável dependente, $\hat{Y}_i$. Se o padrão não for aleatório, a variância do termo de erro não é constante. A Figura 17.7 mostra um padrão cuja variância depende dos valores de $\hat{Y}_i$.

Um gráfico dos resíduos ao longo do tempo, ou da sequência de observações, lançará alguma luz sobre a suposição de que os termos de erro não são correlacionados. Se essa suposição for verdadeira, deve-se observar um padrão aleatório. Um gráfico como o da Figura 17.8 indica uma relação linear entre os resíduos e o tempo. O teste de Durbin-Watson é um procedimento mais formal para estudar as correlações entre os termos de erro[20].

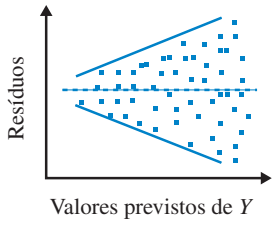

FIGURA 17.7 Gráfico dos resíduos, indicando que a variância não é constante.

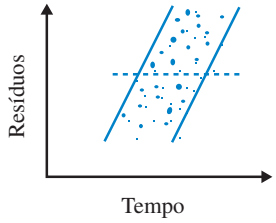

FIGURA 17.8 Gráfico indicando uma relação linear entre resíduos e tempo.

O gráfico de resíduos *versus* variáveis independentes evidencia se um modelo linear é adequado ou não. Mais uma vez, o gráfico deve apresentar um padrão aleatório. Os resíduos dispõem-se aleatoriamente, com dispersão relativamente igual em torno de 0, e não devem apresentar qualquer tendência, seja positiva ou negativa.

Para verificar se devemos incluir quaisquer variáveis adicionais na equação de regressão, podemos fazer uma regressão dos resíduos sobre as variáveis propostas. Se qualquer variável explica uma proporção significativa da variação residual, ela deve ser incluída. A inclusão de variáveis na equação de regressão deve ser fortemente orientada pela teoria do pesquisador. Assim, um estudo dos resíduos proporciona uma visualização valiosa da adequação das suposições básicas e do modelo que é ajustado. A Figura 17.9 exibe um gráfico que indica que as suposições básicas são satisfeitas e que o modelo linear é adequado. Se o exame dos resíduos indicar que as suposições básicas da regressão linear não são satisfeitas, o pesquisador pode transformar as variáveis, em uma tentativa de satisfazer as suposições. Transformações, como extrair logaritmos, ou raízes quadradas ou recíprocas, podem estabilizar a variância, normalizar a distribuição ou tornar linear a relação.

Os gráficos e as tabelas residuais podem ser pedidos quando a regressão é feita, por exemplo, ao usar o SPSS. Você deve realizar essas análises para a regressão múltipla dos dados da Tabela 17.1. A partir do histograma, pode-se ver que cinco resíduos são positivos, enquanto sete são negativos. Ao comparar a distribuição de

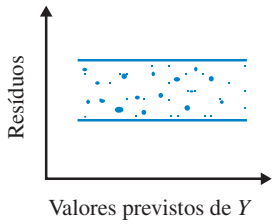

FIGURA 17.9 Gráfico de resíduos indicando que o modelo ajustado é adequado.

frequência com a distribuição normal mostrada no mesmo resultado, visualizamos que a suposição de normalidade provavelmente não é satisfeita, mas que o desvio da normalidade pode não ser significativo. Certamente, podemos fazer um teste estatístico mais formal para a normalidade se isso for garantido. Todos os resíduos estatísticos estão dentro de ± dois desvios-padrão. Além disso, muitos dos resíduos são relativamente pequenos, o que indica que a maioria dos modelos de previsão são bons.

O gráfico da probabilidade de normalidade mostra que os resíduos estão bem próximos da reta de 45° apresentada. Quando comparamos o gráfico dos resíduos padronizados com os valores previstos, nenhum padrão sistemático pode ser visto na disposição dos resíduos. Finalmente, a tabela de estatísticas residuais indica que todos os valores previstos e todos os resíduos estão dentro de ± dois desvios-padrão. Assim, concluímos que a regressão múltipla dos dados da Tabela 17.1 não parece resultar em violações inaceitáveis das suposições. Isso sugere que a relação que estamos tentando prever é linear e que os termos de erro são mais ou menos distribuídos normalmente.

Pesquisa real

O que influencia os preços dos ingressos? Um novo estádio!

Uma das principais fontes de receita para qualquer time profissional é a venda de ingressos, especialmente a venda para os sócios da temporada. Um estudo fez uma análise de regressão para identificar que fatores causavam a variação dos preços dos ingressos entre os times na mesma liga em um determinado ano. A equação de regressão empregada foi a seguinte:

$$LNPMI = a_0 + a_1NVIT + a_2RENDA + a_3PAG + a_4POP + a_5TEND + a_6CAP + a_7EST$$

Onde:

LNPMI = logaritmo neperiano do preço médio dos ingressos
PMI = preço médio dos ingressos
NVIT = número médio de vitórias do time nas últimas três temporadas
RENDA = nível médio de renda da população da cidade
PAG = folha de pagamento do time
POP = tamanho da população da cidade
TEND = tendências no setor
CAP = público como porcentagem da capacidade
EST = se o time está jogando em um estádio novo

A pesquisa reuniu dados cobrindo um período de 7 anos (1996-2002). Os dados financeiros foram obtidos do Team Marketing Reports e os outros dados foram coletados utilizando fontes disponíveis publicamente, como reportagens esportivas. Os resultados das análises de regressão podem ser vistos na tabela no pé da página.

Os resultados sugerem que diversos fatores influenciam os preços dos ingressos, e o principal deles foi o fato de o time estar jogando em um estádio novo. [21] ■

Como no exemplo anterior, algumas variáveis independentes consideradas em um estudo muitas vezes se mostram insignificantes. Quando há muitas variáveis independentes e o pesquisador suspeita que nem todas elas são significantes, a regressão passo a passo deve ser usada.

Regressão passo a passo

O objetivo da **regressão passo a passo** é selecionar, entre inúmeras variáveis previsoras, um pequeno subconjunto de variáveis que respondam pela maior parte da variação na variável dependente. Nesse procedimento, as variáveis previsoras entram na equação de regressão, ou saem dela, uma de cada vez.[22] Há várias abordagens para a regressão passo a passo.

regressão passo a passo
Procedimento de regressão em que as variáveis previsoras entram na equação de regressão, ou saem dela, uma de cada vez.

1. ***Inclusão avançada***. Inicialmente, não há variáveis previsoras na equação de regressão. Elas são introduzidas uma de cada vez somente se satisfizerem certos critérios definidos em ter-

Resultados da regressão

	MLB			NBA			NFL			NHL		
Variável	Coeficiente	Estatística *t*	Valor *p*	Coeficiente	Estatística *t*	Valor *p*	Coeficiente	Estatística *t*	Valor *p*	Coeficiente	Estatística *t*	Valor *p*
Constante	1,521	12,012	0,000	2,965	20,749	0,000	2,886	18,890	0,000	3,172	16,410	0,000
POP	0,000	5,404	0,000	0,000	5,036	0,000	0,000	–2,287	0,023	0,000	2,246	0,026
RENDA	0,000	3,991	0,000	0,000	0,208	0,836	0,000	3,645	0,000	0,000	0,669	0,504
EST	0,337	5,356	0,000	0,108	3,180	0,002	0,226	3,357	0,001	0,321	4,087	0,000
NVIT	0,000	0,091	0,927	0,004	3,459	0,001	0,013	2,190	0,030	0,001	0,369	0,713
CAP	0,006	8,210	0,000	0,000	2,968	0,003	0,002	1,325	0,187	0,005	3,951	0,000
PAG	0,004	4,192	0,000	0,008	5,341	0,000	0,001	0,607	0,545	0,002	1,099	0,273
TEND	0,047	6,803	0,000	0,016	1,616	0,100	0,058	6,735	0,000	0,009	0,718	0,474
CAN (Canadá)										–0,146	–3,167	0,002
R^2 Ajustado			0,778			0,488			0,443			0,292
Estatística *F*			98,366			28,227			24,763			9,545
Significância de *F*			0,000			0,000			0,000			

mos da razão *F*. A ordem em que as variáveis são incluídas se baseia na contribuição para a variância explicada.

2. ***Eliminação para trás***. Inicialmente, todas as variáveis previsoras são incluídas na equação de regressão. Removem-se então as variáveis previsoras uma de cada vez, com base na razão *F*.
3. ***Solução passo a passo***. Combina-se a inclusão antecipada com a remoção das variáveis previsoras que não mais satisfazem o critério especificado em cada passo.

Os procedimentos da regressão passo a passo não resultam em equações ótimas de regressão, no sentido de gerar o maior R^2 para um número determinado de previsores. Em razão das correlações entre previsores, pode ocorrer que uma variável importante nunca venha a ser incluída, enquanto variáveis menos importantes podem ser introduzidas na equação. Para identificar uma equação ótima de regressão, teríamos de calcular soluções combinatórias em que se examinem todas as combinações possíveis. Ainda assim, a regressão passo a passo é útil quando o tamanho da amostra for grande em relação ao número de variáveis previsoras, conforme mostra o exemplo a seguir.

Pesquisa real

Saindo... para o shopping center

Até mesmo no século XXI olhar e comparar é uma parte fundamental das compras – seja on-line ou no shopping. Os clientes gostam de analisar suas decisões de compra antes de realizá-las. Muitos consideram que os varejistas de lojas físicas têm uma vantagem sobre os varejistas da Internet quando se trata de comparar, porque os primeiros são maiores em tamanho e ofertas de produtos. Embora a Web seja mais atraente para os compradores mais jovens, o *shopping* continuará muito à frente nessa corrida, especialmente com tantas opções de entretenimento sendo construídas dentro dele atualmente. Elaborou-se um perfil dos clientes comparadores em *shopping centers* regionais utilizando três conjuntos de variáveis independentes: demográficas, comportamentais e variáveis psicológicas de atitude. A variável dependente consistiu em um índice de curiosidade/comparação. Em uma regressão passo a passo incluindo os três conjuntos de variáveis, constatou-se que o aspecto demográfico era o previsor mais poderoso do comportamento de comparação. A equação final de regressão, que continha 20 das 36 variáveis possíveis, incluía todas as características demográficas. A tabela a seguir apresenta os coeficientes de regressão, erros padrão dos coeficientes e seus níveis de significância.

Ao interpretar os coeficientes, deve-se ter em mente que quanto menor for o índice de curiosidade/comparação (a variável dependente), maior a tendência de apresentar um comportamento associado à comparação. Os dois previsores com maiores coeficientes são gênero e situação de emprego. Os comparadores tendem a ser mulheres empregadas. Tendem também a se situar em posição ligeiramente inferior em comparação com outros clientes do *shopping center,* apresentando níveis mais baixos de instrução e de renda, após levar em conta os efeitos do gênero e da situação de emprego. Embora os comparadores tendam a ser um pouco mais jovens que os não comparadores, não são necessariamente solteiros; os que relatam tamanhos maiores de família tendem a se associar a menores valores do índice de curiosidade/comparação.

O perfil menos afluente dos curiosos em relação a outros clientes indica que as lojas especializadas nos *shopping centers* devem dar ênfase a produtos de preço moderado. Isso pode explicar a taxa historicamente baixa de falência em *shopping centers* de tais lojas e a tendência das lojas especializadas, com preços elevados, a se localizarem apenas em galerias de prestígio ou em *shopping centers* mais qualificados.[23] ■

Regressão do índice de curiosidade/comparação sobre as variáveis descritivas e de atitude por ordem de entrada na regressão passo a passo

Descrição da variável	Coeficiente	EP	Significância
Gênero (0 = masc., 1= fem.)	– 0,485	0,164	0,001
Situação de emprego (0 = empregado)	0,391	0,182	0,003
Autoconfiança	– 0,152	0,128	0,234
Instrução	0,079	0,072	0,271
Intenção quanto à marca	– 0,063	0,028	0,024
Vê TV durante o dia? (0 = sim)	0,232	0,144	0,107
Tensão	– 0,182	0,069	0,008
Renda	0,089	0,061	0,144
Frequência das visitas ao shopping	– 0,130	0,059	0,028
Menos amigos que a maioria	0,162	0,084	0,054
Bom comprador	– 0,122	0,090	0,174
As opiniões de outros são importantes	– 0,147	0,065	0,024
Controle sobre a vida	– 0,069	0,069	0,317
Tamanho da família	– 0,086	0,062	0,165
Pessoa entusiasta	– 0,143	0,099	0,150
Idade	0,036	0,069	0,603
Número de compras feitas	– 0,068	0,043	0,150
Compras por estabelecimento	0,209	0,152	0,167
Compra com economia	– 0,055	0,067	0,412
Excelente avaliador de qualidade	– 0,070	0,089	0,435
CONSTANTE	3,250		
R^2 global = 0,477			

Multicolinearidade

A regressão passo a passo e a regressão múltipla são dificultadas pela presença da multicolinearidade. Praticamente todas as análises de regressão múltipla feitas em pesquisa de marketing envolvem previsores ou variáveis independentes que são correlacionados. Entretanto, surge a **multicolinearidade** quando as intercorrelações entre os previsores são muito altas. A multicolinearidade pode originar vários problemas, incluindo:

multicolinearidade
Situação de intercorrelações muito altas entre variáveis independentes.

1. Os coeficientes de regressão parcial podem não ser estimados com precisão. Os erros padrão tendem a ser muito altos.
2. As magnitudes e os sinais dos coeficientes de regressão parcial podem variar de uma amostra para outra.
3. Torna-se difícil avaliar a importância relativa das variáveis independentes ao explicar a variação na variável dependente.
4. Algumas variáveis previsoras podem ser incluídas ou removidas incorretamente na regressão passo a passo.

Nem sempre fica claro o que constitui uma multicolinearidade grave, embora tenham sido sugeridas várias regras e processos empíricos, bem como processos de maior ou menor complexidade para enfrentar o problema.[24] Um processo simples consiste em utilizar apenas uma das variáveis em um conjunto de variáveis altamente correlacionadas. Alternativamente, pode-se transformar o conjunto de variáveis independentes em um novo conjunto de previsoras mutuamente independentes, recorrendo-se a técnicas como análise dos componentes principais (ver Capítulo 19). Podem ser utilizadas também técnicas mais especializadas, como regressão em crista e regressão de raízes latentes.[25]

PESQUISA ATIVA

Avaliação da marca e preferência pelos laptops Lenovo

Visite www.lenovo.com e pesquise na Internet (utilizando um dispositivo de busca) e no banco de dados *on-line* de sua biblioteca informações sobre os fatores que os consumidores utilizam ao avaliar as marcas concorrentes de laptops.

Como diretor de marketing da Lenovo Computers, como você melhoraria a imagem e o posicionamento competitivo de sua marca?

Formule um modelo de regressão múltipla explicando as preferências do consumidor por marcas de laptop como uma função das avaliações da marca nos fatores de critérios de escolha dos consumidores para avaliar marcas concorrentes.

Importância relativa dos previsores

Na presença da multicolinearidade, exige-se um cuidado especial na avaliação da importância relativa de variáveis independentes. Na pesquisa de marketing aplicada, é conveniente determinar a *importância relativa dos previsores.* Melhor dizendo: qual é a importância das variáveis independentes na justificativa para a variação na variável dependente?[26] Infelizmente, como os previsores são correlacionados, não existe uma medida não ambígua da importância relativa dos previsores na análise de regressão.[27] Não obstante, há várias abordagens para avaliar a importância relativa das variáveis previsoras.

1. ***Significância estatística***. Se o coeficiente de regressão parcial de uma variável não for significativo, conforme determinado por um teste incremental *F,* essa variável é considerada como não importante. Ocorre uma exceção a essa regra se houver fortes razões teóricas para crer que a variável seja importante.
2. ***Quadrado do coeficiente de correlação simples***. Essa medida, r^2, representa a proporção da variação na variável dependente explicada pela variável independente em uma relação bivariada.
3. ***Quadrado do coeficiente de correlação parcial***. Essa medida, $R^2_{yxi.xj\ xk}$ é o coeficiente de determinação entre a variável dependente e a variável independente, controlando os efeitos das outras variáveis independentes.
4. ***Quadrado do coeficiente de correlação de partes***. Este coeficiente representa um aumento em R^2 quando se introduz uma variável em uma equação de regressão que já contém as outras variáveis independentes.
5. ***Medidas baseadas em coeficientes padronizados ou pesos beta***. As medidas mais usadas são os valores absolutos dos pesos beta, $|B_i|$, ou seus quadrados B_i^2. Como são coeficientes parciais, os pesos beta levam em conta o efeito das outras variáveis independentes. Essas medidas vão se tornando menos confiáveis conforme aumentam as correlações entre as variáveis previsoras (a multicolinearidade aumenta).
6. ***Regressão passo a passo***. Utiliza-se a ordem em que os previsores entram em uma equação de regressão ou saem dela para inferir sua importância relativa.

Como os previsores são correlacionados, ao menos até certo ponto, em praticamente todas as situações de regressão, nenhuma dessas medidas é satisfatória. É possível também que as diferentes medidas indiquem uma ordem diferente de importância dos previsores.[28] Todavia, se todas as medidas forem examinadas coletivamente, pode-se obter uma visualização conveniente da importância relativa dos previsores.

Pesquisa de decisão

West Michigan Whitecaps: estimulando a fidelidade dos torcedores

A situação

O West Michigan Whitecaps (www.whitecaps-baseball.com), um time da liga nacional de beisebol de Gram Rapids, Estados Unidos, queria saber o que eles deveriam fazer para desenvolver a fidelidade dos torcedores. Como eles poderiam mantê-la, fazê-la crescer e aproveitá-la? O diretor geral Scott Lane contratou a empresa de pesquisa Message Factors (www.messagefactors.com), com base em Memphis, Tennessee, para ajudar a identificar maneiras de manter com eficácia a fidelidade dos torcedores com um orçamento limitado. A Message Factors desenvolveu um estudo que usou uma técnica proprietária de análise de valor que examinava a relação entre o valor geral percebido e os atributos de satisfação específicos a fim de identificar os elementos que impulsionam a fidelidade. Ela ajudou a determinar as quatro coisas que os clientes querem lhe dizer, que são os elementos básicos – o que os clientes esperam da empresa; questões de valor – o que os clientes valorizam na empresa; irritações – o que os clientes não gostam na empresa; e sem importância – com o que os clientes não se importam na empresa.

Pesquisas qualitativas foram feitas para identificar um conjunto de 71 atributos que influenciaram a fidelidade dos torcedores. Em seguida, um questionário elaborado para incorporar os 71 atributos foi aplicado