14 Introdução à Regressão Múltipla

Objetivos do Aprendizado

Neste capítulo, você aprenderá:

- Desenvolver o modelo de regressão múltipla
- Interpretar os coeficientes da regressão
- Determinar quais as variáveis independentes que devem ser incluídas no modelo de regressão
- Determinar quais variáveis independentes são mais importantes para prever uma variável dependente
- Utilizar variáveis categóricas em um modelo de regressão

UTILIZANDO A ESTATÍSTICA

@ OmniFoods

Você é o gerente de marketing de uma grande empresa de produtos alimentícios. A empresa está planejando o lançamento, em âmbito nacional, da OmniPower, uma nova barra energética. Embora originalmente comercializada para maratonistas, alpinistas e outros atletas, as barras energéticas são atualmente populares junto ao público em geral. A OmniFoods está ansiosa para capturar uma parcela desse vicejante mercado.

Uma vez que o mercado já possui diversas barras energéticas de sucesso, você precisa desenvolver uma estratégia de mercado efetiva. Em particular, você precisa determinar o efeito que o preço e as promoções internas da loja terão sobre as vendas de OmniPower. Antes de comercializá-la em âmbito nacional, você planeja conduzir um estudo baseado em um teste de mercado para vendas de OmniPower, utilizando uma amostra de 34 lojas em uma cadeia de supermercados. Como você pode estender os métodos de regressão linear, discutidos no Capítulo 13, no sentido de incorporar os efeitos decorrentes de preço *e* promoções, dentro do mesmo modelo? Como você pode utilizar esse modelo para aumentar o sucesso do lançamento da OmniPower em âmbito nacional?

O Capítulo 13 se concentrou em modelos de regressão linear simples que utilizam *uma* única variável numérica independente, X, para prever o valor de uma variável dependente, Y. De um modo geral, você consegue fazer previsões mais precisas utilizando *mais de uma* variável independente. Este capítulo apresenta **modelos de regressão múltipla** que utilizam duas ou mais variáveis independentes para prever o valor de uma variável dependente.

14.1 Desenvolvendo um Modelo de Regressão Múltipla

O objetivo estratégico com que se depara o gerente de marketing da OmniFoods é desenvolver um modelo para prever o volume de vendas mensais, por loja, de barras OmniPower e determinar quais variáveis influenciam as vendas. São consideradas duas variáveis independentes nesse caso: o preço de uma barra de OmniPower, medido em centavos de dólar (X_1), e o orçamento mensal para despesas com promoções internas da loja, medido em dólares (X_2). As despesas com promoções internas da loja geralmente incluem letreiros e cartazes, cupons de desconto com distribuição interna, além de amostras grátis. A variável dependente, Y, corresponde ao número de barras OmniPower vendidas ao longo de um mês. São coletados dados de uma amostra de 34 lojas integrantes de uma cadeia de supermercados selecionada, para fins de estudos sobre testes de mercado para as barras OmniPower. Todas as lojas selecionadas apresentam aproximadamente o mesmo volume de vendas mensais. Os dados estão organizados e armazenados no arquivo **OmniPower** e apresentados na Tabela 14.1.

TABELA 14.1
Vendas Mensais, Preço e Despesas com Promoções para a OmniPower

Loja	Vendas	Preço	Promoção	Loja	Vendas	Preço	Promoção
1	4.141	59	200	18	2.730	79	400
2	3.842	59	200	19	2.618	79	400
3	3.056	59	200	20	4.421	79	400
4	3.519	59	200	21	4.113	79	600
5	4.226	59	400	22	3.746	79	600
6	4.630	59	400	23	3.532	79	600
7	3.507	59	400	24	3.825	79	600
8	3.754	59	400	25	1.096	99	200
9	5.000	59	600	26	761	99	200
10	5.120	59	600	27	2.088	99	200
11	4.011	59	600	28	820	99	200
12	5.015	59	600	29	2.114	99	400
13	1.916	79	200	30	1.882	99	400
14	675	79	200	31	2.159	99	400
15	3.636	79	200	32	1.602	99	400
16	3.224	79	200	33	3.354	99	600
17	2.295	79	400	34	2.927	99	600

Interpretando os Coeficientes da Regressão

Quando existem diversas variáveis independentes, você pode estender o modelo de regressão linear simples da Equação (13.1), pressupondo uma relação linear entre cada uma das variáveis independentes e a variável dependente. Por exemplo, com k variáveis independentes, o modelo de regressão múltipla é expresso na Equação (14.1).

MODELO DE REGRESSÃO MÚLTIPLA COM k VARIÁVEIS INDEPENDENTES

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \qquad \textbf{(14.1)}$$

em que

β_0 = intercepto de Y

β_1 = inclinação de Y em relação à variável X_1, mantendo-se constantes as variáveis $X_2, X_3, \ldots, X_k$

β_2 = inclinação de Y em relação à variável X_2, mantendo-se constantes as variáveis $X_1, X_3, \ldots, X_k$

β_3 = inclinação de Y em relação à variável X_3, mantendo-se constantes as variáveis $X_1, X_2, X_4, \ldots, X_k$

β_k = inclinação de Y em relação à variável X_k, mantendo-se constantes as variáveis $X_1, X_2, X_3, \ldots, X_{k-1}$

ε_i = erro aleatório em Y para a observação i

A Equação (14.2) define o modelo de regressão múltipla com duas variáveis independentes.

MODELO DE REGRESSÃO MÚLTIPLA COM DUAS VARIÁVEIS INDEPENDENTES

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad \textbf{(14.2)}$$

em que

β_0 = intercepto de Y

β_1 = inclinação de Y em relação à variável X_1, mantendo-se constante a variável X_2

β_2 = inclinação de Y em relação à variável X_2, mantendo-se constante a variável X_1

ε_i = erro aleatório em Y para a observação i

Compare o modelo de regressão múltipla com o modelo de regressão linear simples [Equação (13.1)]:

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

No modelo de regressão linear simples, a inclinação, β_1, representa a alteração na média aritmética de Y para cada unidade de alteração em X, e não leva em consideração nenhuma outra variável. No modelo de regressão múltipla com duas variáveis independentes [Equação (14.2)], a inclinação, β_1, representa a alteração na média aritmética de Y para cada unidade de alteração em X_1, levando-se em consideração o efeito de X_2.

Do mesmo modo que no caso da regressão linear simples, você utiliza o método dos Mínimos Quadrados para calcular os coeficientes da regressão da amostra (b_0, b_1 e b_2) como estimadores para os parâmetros da população (β_0, β_1 e β_2). A Equação (14.3) define a equação da regressão para um modelo de regressão múltipla com duas variáveis independentes.

EQUAÇÃO DA REGRESSÃO MÚLTIPLA COM DUAS VARIÁVEIS INDEPENDENTES

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} \qquad \textbf{(14.3)}$$

A Figura 14.1 ilustra a planilha com os resultados da regressão para os dados sobre vendas de OmniPower que contém os valores correspondentes aos três coeficientes da regressão, nas células B17 a B19.

Com base na Figura 14.1, os valores calculados para os coeficientes de regressão são

$$b_0 = 5.837{,}5208 \quad b_1 = -53{,}2173 \quad b_2 = 3{,}6131$$

Portanto, a equação para a regressão múltipla é

$$\hat{Y}_i = 5.837{,}5208 - 53{,}2173 X_{1i} + 3{,}6131 X_{2i}$$

FIGURA 14.1
Planilha com resultados parciais da regressão múltipla para os dados sobre vendas de OmniPower

*A Figura 14.1 exibe a planilha **CÁLCULO** da pasta de trabalho **Regressão Múltipla**. Crie essa planilha utilizando as instruções na Seção GE14.1. Leia as instruções do Excel Avançado relativas a essa planilha para aprender sobre as fórmulas utilizadas ao longo de toda a planilha, incluindo a área de Cálculos nas colunas K a N (não ilustradas na Figura 14.1).*

	A	B	C	D	E	F	G	H	I
1	Regressão Múltipla								
2									
3		*Estatística de Regressão*							
4	R Múltiplo	0,8705							
5	R-Quadrado	0,7577							
6	R-Quadrado Ajustado	0,7421							
7	Erro-padrão	638,0653							
8	Observações	34							
9									
10	**ANOVA**								
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*			
12	Regressão	2	39472730,7730	19736365,3865	48,4771	0,0000			
13	Resíduo	31	12620946,6682	407127,3119					
14	Total	33	52093677,4412						
15									
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*	*95,0% inferiores*	*95,0% superiores*
17	Interseção	5837,5208	628,1502	9,2932	0,0000	4556,3999	7118,6416	4556,3999	7118,6416
18	Preço	-53,2173	6,8522	-7,7664	0,0000	-67,1925	-39,2421	-67,1925	-39,2421
19	Promoção	3,6131	0,6852	5,2728	0,0000	2,2155	5,0106	2,2155	5,0106

em que

$\hat{Y}_i$ = vendas mensais previstas das barras OmniPower para a loja i

X_{1i} = preço (em centavos de dólares) da barra OmniPower para a loja i

X_{2i} = gastos mensais (em dólares) com promoções internas nas lojas para a loja i

O intercepto de Y ($b_0 = 5.837,5208$) para a amostra estima o número de barras OmniPower vendidas durante um mês se o preço for \$0,00 e o montante total gasto com promoções for também igual a \$0,00. Uma vez que esses valores correspondentes a preço e promoções se encontram fora do intervalo de preço e promoções utilizados no estudo de teste de mercado, e como eles não fazem sentido no contexto do problema, o valor de b_0 apresenta pouca ou nenhuma interpretação em termos práticos.

A inclinação do preço em relação às vendas de OmniPower ($b_1 = -53,2173$) indica que, para um determinado montante correspondente a gastos mensais com promoções, estima-se que as vendas mensais previstas de OmniPower decresçam em 53,2173 barras por mês, para cada 1 centavo de dólar de aumento no preço. A inclinação de gastos mensais com promoções para as vendas de OmniPower ($b_2 = 3,6131$) indica que, para um determinado preço, se prevê que as vendas estimadas de OmniPower cresçam em 3,6131 barras para cada \$1 adicional gasto com promoções. Essas estimativas permitem que se compreenda melhor o possível efeito que decisões sobre preços e promoções terão no posicionamento de mercado. Por exemplo, estima-se que um decréscimo correspondente a 10 centavos no preço faça com que as vendas cresçam em 532,173 barras, com um montante fixo de gastos mensais com promoções. Estima-se que um crescimento de \$100 nos gastos com promoções faça com que as vendas cresçam em 361,31 barras, para um determinado preço.

Os coeficientes de regressão na regressão múltipla são conhecidos como **coeficientes líquidos da regressão**; eles estimam a média aritmética da variação prevista em Y para cada unidade de variação em um determinado X, *mantendo-se constante o efeito das outras variáveis X*. Por exemplo, no estudo sobre vendas de barras OmniPower para uma loja com um determinado montante de despesas com promoções, prevê-se que as vendas estimadas decresçam em 53,2173 barras por mês para cada 1 centavo de dólar de aumento no preço de uma barra OmniPower. Um outro modo de interpretar esse "efeito líquido" seria imaginar duas lojas com um igual montante em despesas com promoções. Se a primeira loja cobra 1 centavo a mais do que a outra loja, o efeito líquido dessa diferença é que se prevê que a primeira loja venda 53,2173 barras a menos por mês do que a outra loja. Para interpretar o efeito líquido de despesas com promoções, você pode considerar duas lojas que estejam cobrando o mesmo preço. Se a primeira loja gasta \$1 a mais em despesas com promoções, o efeito líquido dessa diferença é a previsão de que a primeira loja venda 3,6131 barras a mais por mês do que a segunda loja.

Prevendo a Variável Dependente Y

Você pode utilizar a equação da regressão múltipla, calculada pelo Microsoft Excel, para prever valores da variável dependente. Por exemplo, qual seria o valor das vendas previstas para uma loja que esteja cobrando 79 centavos de dólar durante um mês em que as despesas com promoções são de \$400? Utilizando a equação da regressão múltipla:

$$\hat{Y}_i = 5.837,5208 - 53,2173X_{1i} + 3,6131X_{2i}$$

com $X_{1i} = 79$ e $X_{2i} = 400$,

$$\hat{Y}_i = 5.837,5208 - 53,2173(79) + 3,6131(400)$$
$$= 3.078,57$$

Por conseguinte, você prevê que as lojas que estão cobrando 79 centavos de dólar e gastando \$400 em despesas com promoções venderão 3.078,57 barras OmniPower por mês.

Depois de ter desenvolvido a equação da regressão, de ter feito a análise de resíduos (veja a Seção 14.3) e determinado a significância do modelo geral ajustado (veja a Seção 14.2), você pode construir uma estimativa do intervalo de confiança para a média aritmética do valor e um intervalo de previsão para um valor individual. Você deve recorrer a softwares para realizar esses cálculos para você, dada a natureza complexa dos cálculos. A Figura 14.2 apresenta uma planilha com a estimativa do intervalo de confiança e o intervalo de previsão para os dados sobre vendas de barras OmniPower.

	A	B	C	D	
1	Estimativa do Intervalo de Confiança e Intervalo de Previsão				
2					
3	**Dados**				
4	Nível de Confiança	95%			
5		1			
6	Valor conhecido dp preço	79			
7	Valor conhecido da promoção	400			
8					
9	X'X	34	2646	13200	
10		2646	214674	1018800	
11		13200	1018800	6000000	
12					
13	Inverso de X'X	0,9692	-0,0094	-0,0005	
14		-0,0094	0,0001	0,0000	
15		-0,0005	0,0000	0,0000	
16					
17	G'X vezes o inverso de X'X	0,0121	0,0001	0,0000	
18					
19	[G'X vezes o inverso de X'X] vezes XG	0,0298			=MATRIZ.MULT(B17:D17, B5:B7)
20	Estatística *t*	2,0395			=INVT(1 - B4, CÁLCULO!B13)
21	Y Previsto (Ychapéu)	3078,57			{=MATRIZ.MULT(TRANSPOR(B5:B7), CÁLCULO!B17:B19)}
22					
23	**Para Previsão da Média de Y (Ychapéu)**				
24	Metade da Amplitude de Intervalo	224,50			=B20 * RAIZ(B19) * CÁLCULO!B7
25	Limite Inferior do Intervalo de Confiança	2854,07			=B21 - B24
26	Limite Superior do Intervalo de Confiança	3303,08			=B21 + B24
27					
28	**Para Y de Resposta Individual**				
29	Metade da Amplitude de Intervalo	1320,57			=B20 * RAIZ(1 + B19) * CÁLCULO!B7
30	Limite Inferior do Intervalo de Confiança	1758,01			=B21 - B29
31	Limite Superior do Intervalo de Confiança	4399,14			=B21 + B29

Também:
Intervalo de células B9:D11 =MATRIZ.MULT(TRANSPOR(MRDisposição!A2:C35), MRDisposição!A2:C35)
Intervalo de células B13:D15 =MATRIZ.INVERSO(B9:D11)
Intervalo de células B17:D17 =MATRIZ.MULTI(TRANSPOT(B5:B7), B13:D15)

FIGURA 14.2
Planilha com a estimativa do intervalo de confiança e o intervalo de previsão para os dados sobre vendas de OmniPower

*A Figura 14.2 exibe a planilha **CÁLCULO** da pasta de trabalho **Regressão Múltipla**. Crie essa planilha utilizando as instruções na Seção GE14.4.*

A estimativa do intervalo de confiança de 95% para a média aritmética das vendas de barras OmniPower em relação a todas as lojas que estejam cobrando 79 centavos de dólar e gastando \$400 em despesas com promoções é de 2.854,07 a 3.303,08 barras. O intervalo de previsão para uma loja individual é de 1.758,01 a 4.399,14 barras.

Problemas para a Seção 14.1

APRENDENDO O BÁSICO

14.1 Para este problema, utilize a seguinte equação da regressão múltipla:

$$\hat{Y}_i = 10 + 5X_{1i} + 3X_{2i}$$

a. Interprete o significado das inclinações.
b. Interprete o significado do intercepto de Y.

14.2 Para este problema, utilize a seguinte equação da regressão múltipla:

$$\hat{Y}_i = 50 - 2X_{1i} + 7X_{2i}$$

a. Interprete o significado das inclinações.
b. Interprete o significado do intercepto de Y.

APLICANDO OS CONCEITOS

14.3 Um fabricante de calçados está avaliando o desenvolvimento de uma nova marca de tênis de corrida. O problema estratégico com que o analista de mercado da empresa se depara é determinar quais variáveis utilizar para prever a durabilidade (ou seja, o efeito do impacto no longo prazo). Duas variáveis independentes a serem consideradas são: X_1 (IMPDIANT), uma unidade de medida para a capacidade de absorção de choque na parte anterior do pé; e X_2 (MEIOSOLA), uma unidade de medida para a alteração nas propriedades de impacto ao longo do tempo. A variável dependente Y é IMPLP, uma unidade de medida para a durabilidade do calçado depois de um teste de impacto repetido. Os dados são coletados a partir de uma amostra aleatória com 15 tipos de tênis de corrida fabricados atualmente, com os seguintes resultados:

Variável	Coeficientes	Erro-Padrão	Estatística *t*	Valor-*p*
INTERSEÇÃO	−0,02686	0,06905	−0,39	0,7034
IMPDIANT	0,79116	0,06295	12,57	0,0000
MEIOSOLA	0,60484	0,07174	8,43	0,0000

a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações, b_1 e b_2, neste problema.

AUTO-AVALIAÇÃO **14.4** Uma empresa de vendas por catálogo, e que vende componentes de informática, software e hardware, mantém um depósito centralizado. A direção da empresa está atualmente examinando o processo de distribuição do depósito. O problema estratégico da empresa com que se depara a gerência se relaciona aos fatores que afetam os custos de distribuição do depósito. Atualmente, uma pequena tarifa de frete está sendo acrescentada ao pedido, independentemente do valor da compra. Ao longo dos últimos 24 meses, foram coletados dados (armazenados em **CustoDeposito**) que indicam os custos de distribuição (em milhares de dólares), as vendas (em milhares de dólares) e a quantidade de pedidos recebidos.
a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações, b_1 e b_2, neste problema.
c. Explique a razão pela qual o coeficiente da regressão, b_0, não apresenta nenhum significado prático no contexto deste problema.
d. Faça a previsão para o custo de distribuição mensal do depósito, quando as vendas são iguais a $400.000 e a quantidade de pedidos é de 4.500.
e. Construa uma estimativa para o intervalo de confiança de 95% para a média aritmética do custo de distribuição mensal do depósito, quando as vendas são iguais a $400.000 e a quantidade de pedidos é de 4.500.
f. Construa um intervalo de previsão de 95% para o custo mensal de distribuição do depósito para um determinado mês em que as vendas são iguais a $400.000 e a quantidade de pedidos é de 4.500.
g. Explique a razão pela qual o intervalo em (e) é mais estreito do que o intervalo em (f).

14.5 Uma organização de defesa do consumidor desejava desenvolver um modelo para prever a milhagem de consumo de gasolina (medida em milhas por galão), com base na potência do motor do automóvel, em cavalos-vapor, e o peso do automóvel (em libras). Foram coletados dados a partir de uma amostra com 50 modelos recentes de automóvel, e os dados organizados e armazenados em **Auto**.
a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações, b_1 e b_2, neste problema.
c. Explique a razão pela qual o coeficiente da equação, b_0, não apresenta nenhum significado prático no contexto deste problema.
d. Faça a previsão do consumo em milhas por galão para automóveis com 60 cavalos-vapor e 2.000 libras de peso.
e. Construa uma estimativa do intervalo de confiança de 95% para a média aritmética do consumo em milhas por galão para automóveis que possuam 60 cavalos-vapor e 2.000 libras de peso.
f. Construa um intervalo de previsão de 95% correspondente ao consumo em milhas por galão para um automóvel individual que tenha 60 cavalos-vapor e 2.000 libras de peso.

14.6 O problema estratégico enfrentado por uma empresa de produtos para consumo é medir a efetividade de diferentes meios de propaganda na promoção de seus produtos. Especificamente, a empresa está interessada na efetividade da propaganda no rádio e em jornais (incluindo o custo dos cupons de desconto). Dados são coletados de uma amostra de 22 cidades, com populações aproximadamente iguais, durante um período de teste de um mês. É alocado a cada cidade um patamar específico de despesas, tanto para propaganda em rádio quanto para propaganda em jornais. As vendas do produto (em milhares de dólares), bem como os patamares de despesas com os meios de propaganda, durante o mês de teste, foram registrados, com os resultados apresentados a seguir, e armazenados no arquivo **Propaganda**:

Cidade	Vendas ($1.000)	Propaganda em Rádio ($1.000)	Propaganda em Jornais ($1.000)
1	973	0	40
2	1.119	0	40
3	875	25	25
4	625	25	25
5	910	30	30
6	971	30	30
7	931	35	35
8	1.177	35	35
9	882	40	25
10	982	40	25
11	1.628	45	45
12	1.577	45	45
13	1.044	50	0
14	914	50	0
15	1.329	55	25
16	1.330	55	25
17	1.405	60	30
18	1.436	60	30
19	1.521	65	35
20	1.741	65	35
21	1.866	70	40
22	1.717	70	40

a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações, b_1 e b_2, neste problema.
c. Interprete o significado do coeficiente de regressão, b_0.
d. Qual tipo de propaganda é mais efetivo? Explique.

14.7 O problema estratégico com que se depara o diretor de operações de radiodifusão de uma emissora de televisão é avaliar a questão referente às horas de sobreaviso (isto é, as horas em que os artistas gráficos sindicalizados da emissora estão sendo remunerados mesmo não estando efetivamente envolvidos em nenhum tipo de atividade) e quais fatores estariam relacionados a essas horas de sobreaviso. O estudo inclui as seguintes variáveis:

Horas de sobreaviso (Y) — Número total de horas de sobreaviso em uma semana
Total da equipe presente (X_1) — Total semanal de dias de presença
Horas remotas (X_2) — Número total de horas trabalhadas por empregados, em locais fora do escritório central

Os dados foram coletados durante 26 semanas; esses dados estão organizados e armazenados em **Sobreaviso**:

a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações, b_1 e b_2, neste problema.
c. Explique a razão pela qual o coeficiente de regressão, b_0, não apresenta nenhum significado prático, no contexto deste problema.
d. Faça a previsão das horas de sobreaviso para uma semana em que o total da equipe presente corresponda a 310 dias e as horas remotas totalizem 400.
e. Construa uma estimativa para o intervalo de confiança de 95% da média aritmética das horas de sobreaviso para semanas nas quais o total da equipe presente tenha 310 dias e as horas remotas totalizem 400.
f. Construa um intervalo de previsão de 95% das horas de sobreaviso correspondentes a uma única semana na qual o total da equipe presente tenha 310 dias e as horas remotas totalizem 400.

14.8 O município de Nassau está localizado a aproximadamente 25 milhas a leste da cidade de Nova York. Os dados organizados e armazenados em **GlenCove** incluem o valor de avaliação, a área do terreno da propriedade em acres (1 acre = 4.046,84 m²) e a idade (tempo de construção), em anos, de uma amostra de 30 residências unifamiliares localizadas em Glen Cove, uma pequena cidade do município de Nassau. Desenvolva um modelo de regressão linear múltipla para prever o valor de avaliação, com base na área do terreno da propriedade e a idade (tempo de construção), em anos.
a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações, b_1 e b_2, neste problema.
c. Explique a razão pela qual o coeficiente da regressão, b_0, não apresenta nenhum significado prático no contexto deste problema.
d. Faça a previsão do valor de avaliação para uma residência que possua uma área de terreno de 0,25 acre e 45 anos de construção.
e. Construa uma estimativa do intervalo de confiança de 95% para a média aritmética do valor de avaliação para residências que possuam uma área de terreno de 0,25 acre e 45 anos de construção.
f. Construa uma estimativa para um intervalo de previsão de 95% do valor de avaliação de uma residência individual que possua uma área de terreno igual a 0,25 acre e 45 anos de construção.

14.2 r^2, r^2 Ajustado e o Teste *F* Geral

Esta seção discute três métodos que você pode utilizar para avaliar a utilidade geral do modelo de regressão múltipla: o coeficiente de determinação múltipla, r^2, o r^2 ajustado e o teste F geral.

Coeficiente de Determinação Múltipla

Lembre-se, com base na Seção 13.3, de que o coeficiente de determinação, r^2, mede a variação em Y que é explicada pela variável independente, X, no modelo de regressão múltipla. Na regressão múltipla, o **coeficiente de determinação múltipla** representa a proporção da variação em Y que é explicada pelo conjunto de variáveis independentes. A Equação (14.4) define o coeficiente de determinação múltipla para um modelo de regressão múltipla com duas ou mais variáveis independentes.

COEFICIENTE DE DETERMINAÇÃO MÚLTIPLA

O coeficiente de determinação múltipla é igual à soma dos quadrados da regressão (*SQReg*) dividida pela soma total dos quadrados (*STQ*).

$$r^2 = \frac{\text{Soma dos quadrados da regressão}}{\text{Soma total dos quadrados}} = \frac{SQReg}{STQ} \qquad \textbf{(14.4)}$$

em que

$SQReg$ = soma dos quadrados da regressão

STQ = soma total dos quadrados

No exemplo da OmniPower, da Figura 14.1, $SQReg = 39.472.730,77$ e $STQ = 52.093.677,44$. Consequentemente,

$$r^2 = \frac{SQReg}{STQ} = \frac{39.472,730,77}{52.093.677,44} = 0,7577$$

O coeficiente de determinação múltipla ($r^2 = 0,7577$) indica que 75,77% da variação nas vendas é explicada pela variação no preço e nas despesas com promoções. Você pode também encontrar o coeficiente de determinação múltipla diretamente a partir dos resultados do Microsoft Excel apresentados na Figura 14.1, ao lado da legenda R Quadrado.

r^2 Ajustado

Ao considerar modelos de regressão múltipla, alguns estatísticos sugerem que você utilize o **r^2 ajustado** para refletir tanto o número de variáveis independentes no modelo quanto o tamanho da amostra. Informar o r^2 ajustado é extremamente importante quando você está comparando dois ou mais modelos de regressão que estão prevendo a mesma variável dependente, embora tenham um número diferente de variáveis independentes. A Equação (14.5) define o r^2 ajustado.

r^2 AJUSTADO

$$r^2_{aj} = 1 - \left[(1 - r^2)\frac{n-1}{n-k-1}\right] \qquad \textbf{(14.5)}$$

em que k é o número de variáveis independentes na equação da regressão.

Por conseguinte, para os dados da OmniPower, uma vez que $r^2 = 0,7577$, $n = 34$ e $k = 2$,

$$\begin{aligned} r^2_{aj} &= 1 - \left[(1 - 0,7577)\frac{34-1}{34-2-1}\right] \\ &= 1 - \left[(0,2423)\frac{33}{31}\right] \\ &= 1 - 0,2579 \\ &= 0,7421 \end{aligned}$$

Portanto, 74,21% da variação nas vendas pode ser explicada pelo modelo de regressão múltipla – ajustado em função do número de variáveis independentes e do tamanho da amostra. Você pode também encontrar o r^2 ajustado diretamente dos resultados na planilha apresentada na Figura 14.1, ao lado da legenda R Quadrado Ajustado.

Teste da Significância do Modelo de Regressão Múltipla Geral

Você utiliza o **teste *F* geral** para determinar se existe uma relação significativa entre a variável dependente e o conjunto inteiro de variáveis independentes (o modelo de regressão múltipla geral). Tendo em vista que existe mais de uma variável independente, você utiliza a hipótese nula e a hipótese alternativa apresentadas a seguir:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (Não existe nenhuma relação linear entre a variável dependente e as variáveis independentes.)

H_1: Pelo menos uma $\beta_j \neq 0$, $j = 1, 2, \ldots, k$ (Existe uma relação linear entre a variável dependente e pelo menos uma das variáveis independentes.)

A Equação (14.6) define a estatística para o teste *F* geral. A Tabela 14.2 apresenta a tabela resumida de ANOVA.

TESTE *F* GERAL

A estatística do teste F_{ESTAT} é igual à média dos quadrados da regressão (*MQReg*) dividida pela média dos quadrados dos resíduos ou erros (*MQR*).

$$F_{ESTAT} = \frac{MQReg}{MQR} \qquad \textbf{(14.6)}$$

em que

F_{ESTAT} = estatística do teste, a partir de uma distribuição *F* com k e $n - k - 1$ graus de liberdade

k = número de variáveis independentes no modelo de regressão

TABELA 14.2
Tabela Resumida de ANOVA para o Teste *F* Geral

Fonte	Graus de Liberdade	Soma dos Quadrados	Média dos Quadrados (Variância)	*F*
Regressão	k	$SQReg$	$MQReg = \frac{SQReg}{k}$	$F_{ESTAT} = \frac{MQReg}{MQR}$
Erro	$n - k - 1$	SQR	$MQR = \frac{SQR}{n-k-1}$	
Total	$n - 1$	STQ		

A regra de decisão é

Rejeitar H_0, no nível de significância α, se $F_{ESTAT} > F_\alpha$;

caso contrário, não rejeitar H_0.

Utilizando um nível de significância de 0,05, o valor crítico da distribuição *F*, com 2 e 31 graus de liberdade, encontrado na Tabela E.5, é aproximadamente 3,32 (veja a Figura 14.3). Com base na Figura 14.1, a estatística do teste F_{ESTAT}, apresentada na tabela resumida de ANOVA, é 48,4771. Uma vez que $48,4771 > 3,32$, ou como o valor-$p = 0,000 < 0,05$, você rejeita H_0 e conclui que pelo menos uma das variáveis independentes (preço e/ou despesas com promoções) está relacionada a vendas.

FIGURA 14.3
Testando a significância de um conjunto de coeficientes de regressão, no nível de significância de 0,05, com 2 e 31 graus de liberdade

Problemas para a Seção 14.2

APRENDENDO O BÁSICO

14.9 A tabela resumida de ANOVA a seguir corresponde a um modelo de regressão múltipla com duas variáveis independentes:

Fonte	Graus de Liberdade	Soma dos Quadrados	Média dos Quadrados	*F*
Regressão	2	60		
Erro	18	120		
Total	20	180		

a. Determine a média dos quadrados da regressão (*MQReg*) e a média dos quadrados dos resíduos ou erros (*MQR*).
b. Calcule a estatística do teste geral, F_{ESTAT}.
c. Determine se existe uma relação significativa entre *Y* e as duas variáveis independentes, no nível de significância de 0,05.
d. Calcule o coeficiente de determinação, r^2, e interprete o seu respectivo significado.
e. Calcule o r^2 ajustado.

14.10 A tabela resumida de ANOVA a seguir corresponde a um modelo de regressão múltipla com duas variáveis independentes:

Fonte	Graus de Liberdade	Soma dos Quadrados	Média dos Quadrados	*F*
Regressão	2	30		
Erro	10	120		
Total	12	150		

a. Determine a média dos quadrados da regressão (*MQReg*) e a média dos quadrados dos resíduos ou erros (*MQR*).
b. Calcule a estatística do teste F_{ESTAT}-geral.
c. Determine se existe uma relação significativa entre *Y* e as duas variáveis independentes, no nível de significância de 0,05.
d. Calcule o coeficiente de determinação, r^2, e interprete o seu significado.
e. Calcule o r^2 ajustado.

APLICANDO OS CONCEITOS

14.11 Eileen M. Van Aken e Brian M. Kleiner, professores no Virginia Polytechnic Institute e na Virginia State University, investigaram os fatores que contribuem para a efetividade de equipes [dados extraídos de "Determinants of Effectiveness for Cross-Functional Organizational Design Teams", *Quality Management Journal*, 4 (1997), 51-79]. Os pesquisadores estudaram 34 variáveis independentes, tais como capacidades da equipe, diversidade da equipe, frequência de reuniões e clareza nas expectativas. Para cada uma das equipes estudadas, a cada uma das variáveis foi atribuído um valor de 1 a 100, com base nos resultados de entrevistas e dados da pesquisa, em que 100 representa a classificação mais alta. Foi também atribuído à variável dependente, desempenho da equipe, um valor de 1 a 100, com 100 representando a classificação mais elevada. Muitos modelos diferentes de regressão foram explorados, incluindo os seguintes:

Modelo 1

$$\text{Desempenho da equipe} = \beta_0 + \beta_1(\text{Capacidades da equipe}) + \varepsilon,$$

$$r^2_{aj} = 0{,}68$$

Modelo 2

$$\text{Desempenho da equipe} = \beta_0 + \beta_1(\text{Clareza nas expectativas}) + \varepsilon$$

$$r^2_{aj} = 0{,}78$$

Modelo 3

$$\text{Desempenho da equipe} = \beta_0 + \beta_1(\text{Capacidades da equipe}) + \beta_2(\text{Clareza nas expectativas}) + \varepsilon,$$

$$r^2_{aj} = 0{,}97$$

a. Interprete o r^2 ajustado para cada um dos três modelos.
b. Qual desses três modelos você acredita que seja o melhor estimador para o desempenho de equipe?

14.12 No Problema 14.3, em Problemas para a Seção 14.1, você previu a durabilidade de uma marca de tênis de corrida com base na capacidade de absorção de choques da parte dianteira do pé e na alteração nas propriedades de impacto ao longo do tempo. A análise da regressão resultou na seguinte tabela resumida de ANOVA:

Fonte	Graus de Liberdade	Soma dos Quadrados	Média dos Quadrados	*F*	Valor-*p*
Regressão	2	12,61020	6,30510	97,69	0,0001
Erro	12	0,77453	0,06454		
Total	14	13,38473			

a. Determine se existe uma relação significativa entre a durabilidade e as duas variáveis independentes, no nível de significância de 0,05.
b. Interprete o significado do valor-*p*.
c. Calcule o coeficiente de determinação, r^2, e interprete o seu significado.
d. Calcule o r^2 ajustado.

14.13 No Problema 14.5, em Problemas para a Seção 14.1, você utilizou a potência, em cavalos-vapor, e o peso do automóvel para prever o consumo de gasolina em milhas (dados em Auto). Utilizando os resultados daquele problema,
a. determine se existe uma relação significativa entre a milhagem e as duas variáveis independentes (potência, em cavalos-vapor, e peso), no nível de significância de 0,05.
b. interprete o significado do valor-*p*.
c. calcule o coeficiente de determinação, r^2, e interprete o seu significado.
d. calcule o r^2 ajustado.

AUTO-AVALIAÇÃO **14.14** No Problema 14.4, em Problemas para a Seção 14.1, você utilizou as vendas e a quantidade de pedidos para prever os custos de distribuição de uma empresa de venda por catálogo (dados no arquivo CustoDeposito). Utilizando os resultados daquele problema,
a. determine se existe uma relação significativa entre o custo de distribuição e as duas variáveis independentes (vendas e quantidade de pedidos), no nível de significância de 0,05.
b. interprete o significado do valor-*p*.
c. calcule o coeficiente de determinação, r^2, e interprete o seu significado.
d. calcule o r^2 ajustado.

14.15 No Problema 14.7, em Problemas para a Seção 14.1, você utilizou o total da equipe presente e as horas remotas para prever as horas de sobreaviso (dados no arquivo Sobreaviso). Utilizando os resultados daquele problema,
a. determine se existe uma relação significativa entre as horas de sobreaviso e as duas variáveis independentes (total da equipe presente e horas remotas), no nível de significância de 0,05.
b. interprete o significado do valor-*p*.
c. calcule o coeficiente de determinação, r^2, e interprete o seu significado.
d. calcule o r^2 ajustado.

14.16 No Problema 14.6, em Problemas para a Seção 14.1, você utilizou a verba de propaganda no rádio e em jornais para prever vendas (dados no arquivo Propaganda). Utilizando os resultados daquele problema,
a. determine se existe uma relação significativa entre vendas e as duas variáveis independentes (propaganda em rádio e propaganda em jornais), no nível de significância de 0,05.
b. interprete o significado do valor-*p*.
c. calcule o coeficiente de determinação, r^2, e interprete o seu significado.
d. calcule o r^2 ajustado.

14.17 No Problema 14.8, em Problemas para a Seção 14.1, você utilizou a área do terreno de uma propriedade residencial e a idade de um imóvel para prever o valor de avaliação (dados no arquivo GlenCove). Utilizando os resultados daquele problema,
a. determine se existe uma relação significativa entre o valor de avaliação e as duas variáveis independentes [área do terreno de uma propriedade residencial e a idade (tempo de construção) do imóvel], no nível de significância de 0,05.
b. interprete o significado do valor-*p*.
c. calcule o coeficiente de determinação, r^2, e interprete o seu significado.
d. calcule o r^2 ajustado.

14.3 Análise de Resíduos para o Modelo de Regressão Múltipla

Na Seção 13.5, você utilizou a análise de resíduos para avaliar o ajuste do modelo de regressão linear simples. Para o modelo de regressão múltipla, com duas variáveis independentes, você precisa construir e analisar os gráficos de resíduos citados a seguir:

1. Resíduos *versus* $\hat{Y}_i$
2. Resíduos *versus* X_{1i}
3. Resíduos *versus* X_{2i}
4. Resíduos *versus* tempo

O primeiro gráfico de resíduos examina o padrão dos resíduos em relação aos valores previstos para *Y*. Caso os resíduos demonstrem um padrão para diferentes valores previstos de *Y*, existem evidências de um possível efeito curvilíneo (veja a Seção 15.1) em pelo menos uma variável independente, uma possível violação do pressuposto de igualdade de variâncias (veja a Figura 13.13) e/ou a necessidade de transformar a variável *Y*.

O segundo e terceiro gráficos de resíduos envolvem as variáveis independentes. Padrões no gráfico de resíduos em relação a uma variável independente podem indicar a existência de um efeito curvilíneo e, por conseguinte, a necessidade de ser acrescentada ao modelo de regressão múltipla uma variável independente curvilínea (ver Seção 15.1). O quarto gráfico é utilizado para investigar padrões nos resíduos, visando validar o pressuposto da independência, quando os dados são coletados em ordem cronológica. Em associação a esse gráfico de resíduos, assim como na Seção 13.6, você pode calcular a estatística de Durbin-Watson para determinar a existência de uma autocorrelação positiva entre os resíduos.

A Figura 14.4 ilustra os gráficos de resíduos para o exemplo das vendas de OmniPower. Existe muito pouco ou nenhum padrão na relação entre os resíduos e o valor previsto de *Y*, o valor de X_1 (preço) ou o valor de X_2 (despesas com promoções). Desse modo, você pode concluir que o modelo de regressão linear múltipla é apropriado para prever as vendas. Não há necessidade de fazer o gráfico dos resíduos em relação ao tempo, uma vez que os dados não foram coletados em ordem cronológica.

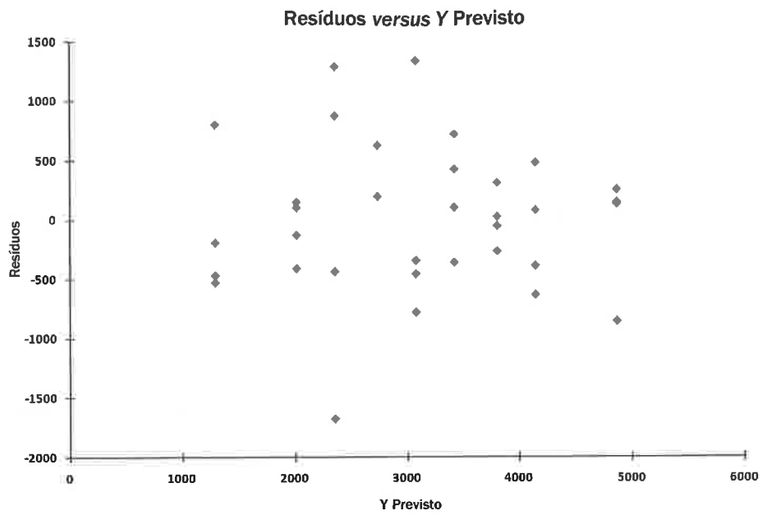
FIGURA 14.4
Gráficos de resíduos para os dados de vendas de OmniPower:

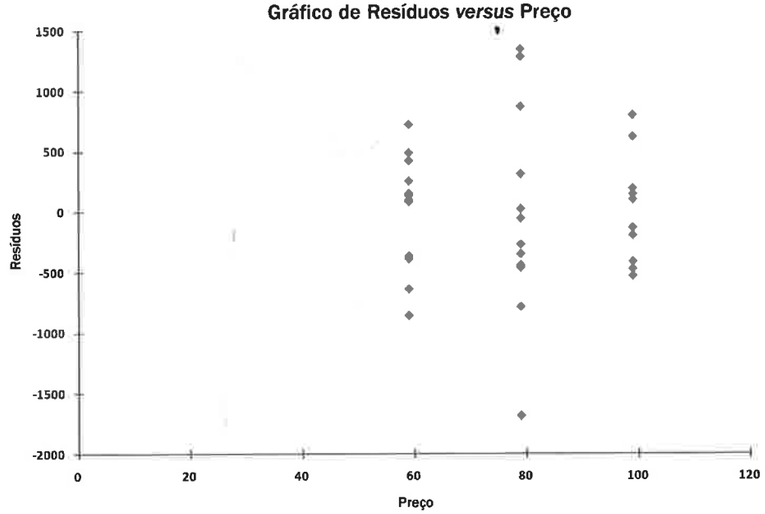
Painel A, resíduos *versus* Y previsto;

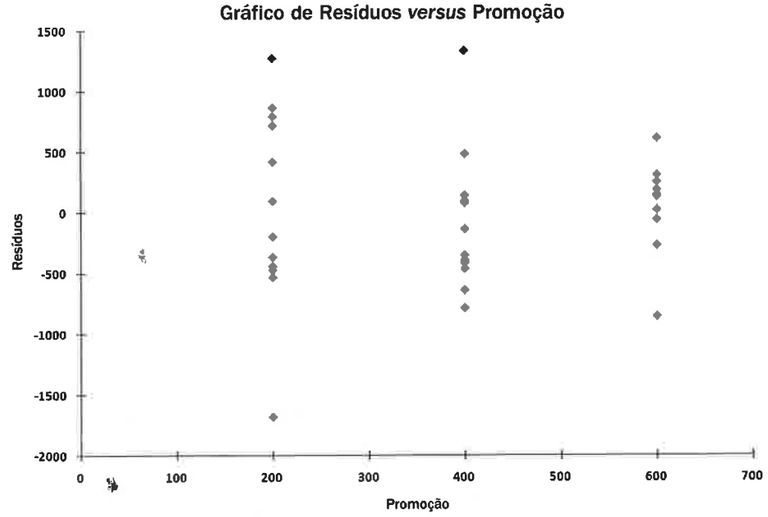
Painel B, resíduos *versus* preço;

Painel C, resíduos *versus* promoção

Crie gráficos de resíduos utilizando as instruções na Seção GE13.5.







Problemas para a Seção 14.3

APLICANDO OS CONCEITOS

14.18 No Problema 14.4, em Problemas para a Seção 14.1, você utilizou vendas e a quantidade de pedidos para prever os custos de distribuição em uma empresa de venda por catálogo (dados no arquivo **CustoDeposito**).

a. Faça um gráfico para os resíduos em relação a $\hat{Y}_i$.
b. Faça um gráfico para os resíduos em relação a X_{1i}.
c. Faça um gráfico para os resíduos em relação a X_{2i}.
d. Faça um gráfico para os resíduos em relação ao tempo.
e. No gráfico de resíduos criado em (a) até (d), existem evidências de uma violação do pressuposto da regressão? Explique.
f. Determine a estatística de Durbin-Watson.
g. No nível de significância de 0,05, existem evidências de uma autocorrelação positiva nos resíduos?

14.19 No Problema 14.5, em Problemas para a Seção 14.1, você utilizou a potência, em cavalos-vapor, e o peso para prever a milhagem do consumo de gasolina (dados no **Auto**).

a. Faça um gráfico para os resíduos em relação a $\hat{Y}_i$.
b. Faça um gráfico para os resíduos em relação a X_{1i}.
c. Faça um gráfico para os resíduos em relação a X_{2i}.
d. No gráfico de resíduos criado em (a) até (c), existem evidências de uma violação do pressuposto da regressão? Explique.
e. Você deveria calcular a estatística de Durbin-Watson para esses lados? Explique.

14.20 No Problema 14.6, em Problemas para a Seção 14.1, você utilizou a verba de propaganda em rádio e em jornais para prever vendas (dados no arquivo **Propaganda**).

a. Realize uma análise de resíduos em seus resultados.
b. Caso seja apropriado, realize o teste de Durbin-Watson, utilizando $\alpha = 0{,}05$.
c. Os pressupostos da regressão são válidos para esses dados?

14.21 No Problema 14.7, em Problemas para a Seção 14.1, você utilizou o total da equipe presente e as horas remotas para prever as horas de sobreaviso (veja o arquivo **Sobreaviso**).

a. Realize uma análise de resíduos em seus resultados.
b. Caso seja apropriado, realize o teste de Durbin-Watson, utilizando $\alpha = 0{,}05$.
c. Os pressupostos da regressão são válidos para esses dados?

14.22 No Problema 14.8, em Problemas para a Seção 14.1, você utilizou a área do terreno de uma propriedade e a idade (tempo de construção) de uma residência para prever o valor de avaliação (dados aramazenados no arquivo **GlenCove**).

a. Realize uma análise de resíduos em seus resultados.
b. Caso seja apropriado, realize o teste de Durbin-Watson, utilizando $\alpha = 0{,}05$.
c. Os pressupostos da regressão são válidos para esses dados?

14.4 Inferências Relacionadas aos Coeficientes de Regressão da População

Na Seção 13.7, você testou a inclinação em um modelo de regressão linear simples para determinar a significância da relação entre X e Y. Além disso, você construiu uma estimativa de intervalo de confiança para a inclinação da população. Esta seção estende esses procedimentos para a regressão múltipla.

Testes de Hipóteses

Em um modelo de regressão linear simples, para testar uma determinada hipótese com referência à inclinação da população, β_1, você utilizou a Equação (13.16):

$$t_{ESTAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

A Equação (14.7) generaliza essa equação para a regressão múltipla.

TESTANDO A INCLINAÇÃO NA REGRESSÃO MÚLTIPLA

$$t_{ESTAT} = \frac{b_j - \beta_j}{S_{b_j}} \qquad \textbf{(14.7)}$$

em que

b_j = inclinação da variável j com Y, mantendo-se constantes os efeitos de todas as outras variáveis independentes

S_{bj} = erro-padrão do coeficiente de regressão, b_j

t_{ESTAT} = estatística do teste para uma distribuição t, com $n - k - 1$ graus de liberdade

k = número de variáveis independentes na equação da regressão

β_j = valor citado na hipótese para a inclinação da população em relação à variável j, mantendo-se constantes os efeitos de todas as outras variáveis independentes

Para determinar se a variável X_2 (o montante de despesas com promoções) exerce um efeito significativo sobre vendas, levando em consideração os preços de barras OmniPower, a hipótese nula e a hipótese alternativa são:

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

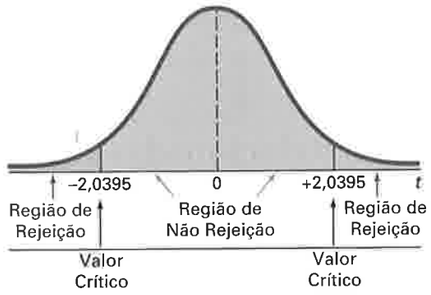
Partindo da Equação (14.7) e da Figura 14.1,

$$t_{ESTAT} = \frac{b_2 - \beta_2}{S_{b_2}}$$

$$= \frac{3{,}6131 - 0}{0{,}6852} = 5{,}2728$$

Caso você selecione um nível de significância de 0,05, os valores críticos de t, para 31 graus de liberdade, a partir da Tabela E.3, são −2,0395 e +2,0395 (veja a Figura 14.5).

FIGURA 14.5
Testando a significância de um coeficiente de regressão, no nível de significância de 0,05, com 31 graus de liberdade



Com base na Figura 14.1, observe que a estatística do teste t_{ESTAT} calculada é 5,2728. Uma vez que $t_{ESTAT} = 5{,}2728 > 2{,}0395$, ou tendo em vista que o valor-p é aproximadamente igual a zero, você rejeita H_0 e conclui que existe uma relação significativa entre a variável X_2 (despesas com promoções) e vendas, levando-se em consideração o preço, X_1. Esse valor-p extremamente pequeno permite que você rejeite veementemente a hipótese nula de que não existe nenhuma relação linear entre vendas e despesas com promoções. O Exemplo 14.1 apresenta o teste para a significância de β_1, a inclinação de vendas em relação a preço.

EXEMPLO 14.1

Testando a Significância da Inclinação de Vendas em Relação a Preço

No nível de significância de 0,05, existem evidências de que a inclinação de vendas em relação a preço seja diferente de zero?

SOLUÇÃO A partir da Figura 14.1, $t_{ESTAT} = -7{,}7664 < -2{,}0395$ (o valor crítico para $\alpha = 0{,}05$), ou o valor-$p = 0{,}0000 < 0{,}05$. Portanto, existe uma relação significativa entre preço, X_1, e vendas, levando em consideração as despesas com promoções, X_2.

Conforme observado em relação a cada uma dessas duas variáveis X independentes, o teste de significância para um determinado coeficiente de regressão na regressão múltipla é um teste para a significância de ser acrescentada uma determinada variável a um modelo de regressão, sabendo-se que outra variável está incluída. Em outras palavras, o teste t para o coeficiente da regressão é, na realidade, um teste para a contribuição de cada uma das variáveis independentes.

Estimativa do Intervalo de Confiança

Em vez de testar a significância da inclinação de uma população, pode ser que você queira estimar o valor da inclinação para uma população. A Equação (14.8) define a estimativa do intervalo de confiança para a inclinação de uma população na regressão múltipla.

ESTIMATIVA DO INTERVALO DE CONFIANÇA PARA A INCLINAÇÃO

$$b_j \pm t_{\alpha/2} S_{b_j} \qquad \textbf{(14.8)}$$

em que $t_{\alpha/2}$ é o valor crítico correspondente a uma probabilidade de cauda superior de $\alpha/2$ a partir da distribuição t com $n - k - 1$ graus de liberdade (ou seja, uma área acumulada de $1 - \alpha/2$, e k corresponde à quantidade de variáveis independentes.

Para construir uma estimativa do intervalo de confiança de 95% para a inclinação da população, β_1 (o efeito do preço, X_1, sobre vendas, Y, mantendo-se constante o efeito de despesas com promoções, X_2), o valor crítico de t, no nível de confiança de 95%, com 31 graus de liberdade, é igual a 2,0395 (veja a Tabela E.3). Assim, utilizando-se a Equação (14.8) e a Figura 14.1,

$$b_1 \pm t_{\alpha/2} S_{b_1}$$

$$-53{,}2173 \pm (2{,}0395)(6{,}8522)$$

$$-53{,}2173 \pm 13{,}9752$$

$$-67{,}1925 \leq \beta_1 \leq -39{,}2421$$

Levando-se em consideração o efeito de despesas com promoções, o efeito estimado de um aumento de 1 centavo de dólar no preço é a redução da média aritmética das vendas em aproximadamente 39,2 a 67,2 barras. Você tem 95% de confiança de que esse intervalo estima corretamente a relação entre essas variáveis. Do ponto de vista de um teste de hipóteses, uma vez que esse intervalo de confiança não inclui 0, você conclui que o coeficiente de regressão, β_1, possui um efeito significativo.

O Exemplo 14.2 constrói e interpreta uma estimativa do intervalo de confiança para a inclinação de vendas com despesas com promoções.

EXEMPLO 14.2

Construindo uma Estimativa do Intervalo de Confiança para a Inclinação de Vendas em Relação a Despesas com Promoções

Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população de vendas em relação a despesas com promoções.

SOLUÇÃO O valor crítico de t, no nível de confiança de 95%, com 31 graus de liberdade, é 2,0395 (veja a Tabela E.3). Utilizando a Equação (14.8) e a Figura 14.1,

$$b_2 \pm t_{\alpha/2} S_{b_2}$$

$$3{,}6131 \pm (2{,}0395)(0{,}6852)$$

$$3{,}6131 \pm 1{,}3975$$

$$2{,}2156 \leq \beta_2 \leq 5{,}0106$$

Por conseguinte, levando-se em consideração o efeito do preço, o efeito estimado de cada dólar adicional de despesas com promoções é o aumento da média aritmética das vendas em aproximadamente 2,22 a 5,01 barras. Você tem 95% de confiança de que esse intervalo estima corretamente a relação entre essas variáveis. Partindo do ponto de vista de um teste de hipóteses, uma vez que esse intervalo de confiança não inclui o 0, você pode concluir que o coeficiente de regressão, β_2, exerce um efeito significativo.

Problemas para a Seção 14.4

APRENDENDO O BÁSICO

14.23 Utilize as seguintes informações, extraídas de uma análise de regressão múltipla:

$$n = 25 \quad b_1 = 5 \quad b_2 = 10 \quad S_{b_1} = 2 \quad S_{b_2} = 8$$

a. Qual dentre as variáveis apresenta a maior inclinação, em unidades de uma estatística *t*?
b. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população, β_1.
c. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa em relação ao modelo de regressão. Com base nesses resultados, indique as variáveis independentes que devem ser incluídas nesse modelo.

14.24 Utilize as seguintes informações, extraídas de uma análise de regressão múltipla:

$$n = 20 \quad b_1 = 4 \quad b_2 = 3 \quad S_{b_1} = 1{,}2 \quad S_{b_2} = 0{,}8$$

a. Qual dentre as variáveis apresenta a maior inclinação, em unidades de uma estatística *t*?
b. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população, β_1.
c. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa em relação ao modelo de regressão. Com base nesses resultados, indique as variáveis independentes que devem ser incluídas nesse modelo.

APLICANDO OS CONCEITOS

14.25 No Problema 14.3, em Problemas para a Seção 14.1, você previu a durabilidade de uma marca de tênis de corrida, com base na capacidade de absorção de choques da parte dianteira do pé (IMPDIANT) e na alteração nas propriedades de impacto ao longo do tempo (MEIOSOLA), para uma amostra de 15 pares de calçados. Utilize os seguintes resultados:

Variável	Coeficiente	Erro-padrão	Estatística *t*	Valor-*p*
INTERSEÇÃO	−0,02686	0,06905	−0,39	0,7034
IMPDIANT	0,79116	0,06295	12,57	0,0000
MEIOSOLA	0,60484	0,07174	8,43	0,0000

a. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população entre durabilidade e capacidade de absorção de impacto na parte dianteira do pé.
b. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Com base nesses resultados, indique as variáveis independentes que devem ser incluídas nesse modelo.

AUTO-AVALIAÇÃO **14.26** No Problema 14.4, em Problemas para a Seção 14.1, você utilizou as vendas e a quantidade de pedidos para prever os custos de distribuição em uma empresa de venda por catálogo (dados no arquivo **CustoDeposito**). Utilizando o resultado daquele problema,

a. construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população entre custo de distribuição e vendas.
b. no nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa em relação ao modelo de regressão. Com base nesses resultados, indique as variáveis independentes que devem ser incluídas nesse modelo.

14.27 No Problema 14.5, em Problemas para a Seção 14.1, você utilizou a potência, em cavalos-vapor, e o peso para prever a milhagem do consumo de gasolina (dados no arquivo **Auto**). Utilizando os resultados daquele problema,
a. construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população entre milhagem do consumo de gasolina e potência, em cavalos-vapor.
b. no nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Com base nesses resultados, indique as variáveis independentes que devem ser incluídas nesse modelo.

14.28 No Problema 14.6, em Problemas para a Seção 14.1, você utilizou verba de propaganda no rádio e em jornais para prever vendas (dados no arquivo **Propaganda**). Utilizando o resultado daquele problema,
a. construa uma estimativa do intervalo de confiança de 95% para a inclinação da população entre vendas e a propaganda em rádio.
b. no nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Com base nesses resultados, indique as variáveis independentes que devem ser incluídas nesse modelo.

14.29 No Problema 14.7, em Problemas para a Seção 14.1, você utilizou a quantidade total da equipe presente e as horas remotas para prever as horas de sobreaviso (dados no arquivo **Sobreaviso**). Utilizando os resultados daquele problema,
a. construa uma estimativa do intervalo de confiança de 95% para a inclinação da população entre horas de sobreaviso e a quantidade total da equipe presente.
b. no nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Com base nesses resultados, indique as variáveis independentes que devem ser incluídas nesse modelo.

14.30 No Problema 14.8, em Problemas para a Seção 14.1, você utilizou a área do terreno de uma propriedade residencial e a idade (tempo de construção) do imóvel para prever o valor de avaliação (dados no arquivo **GlenCove**). Utilizando os resultados daquele problema,
a. construa uma estimativa do intervalo de confiança de 95% para a inclinação da população entre valor de avaliação e a área do terreno de uma propriedade residencial.
b. no nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Com base nesses resultados, indique as variáveis independentes que devem ser incluídas nesse modelo.

14.5 Testando Partes do Modelo de Regressão Múltipla

Ao desenvolver um modelo de regressão múltipla, você deseja utilizar somente aquelas variáveis independentes que reduzam significativamente o erro ao prever o valor de uma variável dependente. Se uma variável independente não melhorar essa previsão, você pode excluí-la do modelo de regressão múltipla e utilizar um modelo com um menor número de variáveis independentes.

O **teste *F* parcial** é um método alternativo ao teste *t* discutido na Seção 14.4 para determinar a contribuição de uma variável independente. Utilizando esse método, você determina a contribuição oferecida por cada uma das variáveis independentes em relação à soma dos quadrados da regressão, depois que todas as outras variáveis independentes foram incluídas no modelo. A nova variável independente é incluída unicamente se vier a tornar o modelo significativamente melhor.

Para conduzir testes *F* parciais para o exemplo que trata das vendas de OmniPower, você precisa avaliar a contribuição de despesas com promoções (X_2), depois que preço (X_1) tiver sido incluído no modelo, e, também, avaliar a contribuição de preço (X_1) depois que despesas com promoções (X_2) tiverem sido incluídas no modelo.

Em geral, se existem diversas variáveis independentes, você determina a contribuição de cada uma das variáveis independentes a ser incluída no modelo levando em conta a soma dos quadrados da regressão de um modelo que inclui todas as variáveis independentes, exceto a variável de interesse, *j*. A soma dos quadrados da regressão é representada por *SQReg* (todas as variáveis *X*, exceto *j*). A Equação (14.9) determina a contribuição da variável *j*, pressupondo que todas as outras variáveis já estão incluídas.

DETERMINANDO A CONTRIBUIÇÃO DE UMA VARIÁVEL INDEPENDENTE PARA O MODELO DE REGRESSÃO

$$SQReg\,(X_j/\text{ Todas as variáveis } X, \text{ exceto } j) = SQReg\,(\text{Todas as variáveis } X) - SQReg\,(\text{Todas as variáveis } X, \text{ exceto } j) \qquad \textbf{(14.9)}$$

Caso existam duas variáveis independentes, você utiliza as Equações (14.10a) e (14.10b) para determinar a contribuição de cada uma.

CONTRIBUIÇÃO DA VARIÁVEL X_1, SABENDO-SE QUE X_2 FOI INCLUÍDA

$$SQReg(X_1 \mid X_2) = SQReg(X_1 \text{ e } X_2) - SQReg(X_2) \qquad \textbf{(14.10a)}$$

CONTRIBUIÇÃO DA VARIÁVEL X_2, SABENDO-SE QUE X_1 FOI INCLUÍDA

$$SQReg(X_2 \mid X_1) = SQReg(X_1 \text{ e } X_2) - SQReg(X_1) \qquad \textbf{(14.10b)}$$

O termo $SQReg(X_2)$ representa a soma dos quadrados que é decorrente da regressão para um modelo que inclui somente a variável independente X_2 (despesas com promoções). De maneira análoga, $SQReg(X_1)$ representa a soma dos quadrados que é decorrente da regressão para um modelo que inclui somente a variável independente X_1 (preço). As Figuras 14.6 e 14.7 apresentam os resultados, sob a forma de planilha, para esses dois modelos.

Com base na Figura 14.6, $SQReg(X_2) = 14.915.814{,}10$, e com base na Figura 14.1, $SQReg(X_1 \text{ e } X_2) = 39.472.730{,}77$. Então, utilizando a Equação (14.10a),

$$\begin{aligned} SQReg\,(X_1 \mid X_2) &= SQReg\,(X_1 \text{ e } X_2) - SQReg(X_2) \\ &= 39.472.730{,}77 - 14.915.814{,}10 \\ &= 24.556.916{,}67 \end{aligned}$$

FIGURA 14.6
Planilha de resultados da regressão para uma análise da regressão linear simples para vendas e despesas com promoções, $SQReg(X_2)$

	A	B	C	D	E	F	G
1	Análise de Vendas & Despesas com Promoções						
2							
3	*Estatística de Regressão*						
4	R Múltiplo	0,5351					
5	R-Quadrado	0,2863					
6	R-Quadrado Ajustado	0,2640					
7	Erro-padrão	1077,8721					
8	Observações	34					
9							
10	ANOVA						
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*	
12	Regressão	1	14915814,1025	14915814,1025	12,8384	0,0011	
13	Resíduo	32	37177863,3387	1161808,2293			
14	Total	33	52093677,4412				
15							
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*
17	Interseção	1496,0161	483,9789	3,0911	0,0041	510,1843	2481,8480
18	Preço	4,1281	1,1521	3,5831	0,0011	1,7813	6,4748

FIGURA 14.7
Planilha de resultados da regressão para um modelo de regressão linear simples para vendas venda e preço, $SQReg(X_1)$

Crie as planilhas ilustradas na Figura 14.6 e na Figura 14.7 utilizando as instruções na Seção GE13.2.

	A	B	C	D	E	F	G
1	Análise de Vendas & Preço						
2							
3	*Estatística de Regressão*						
4	R Múltiplo	0,7351					
5	R-Quadrado	0,5404					
6	R-Quadrado Ajustado	0,5261					
7	Erro-padrão	864,9457					
8	Observações	34					
9							
10	ANOVA						
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*	
12	Regressão	1	28153486,1482	28153486,1482	37,6318	0,0000	
13	Resíduo	32	23940191,2930	748130,9779			
14	Total	33	52093677,4412				
15							
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*
17	Interseção	7512,3480	734,6189	10,2262	0,0000	6015,9796	9008,7164
18	Preço	-56,7138	9,2451	-6,1345	0,0000	-75,5455	-37,8822

Para determinar se X_1 aperfeiçoa significativamente um modelo, depois que X_2 foi incluída, você divide a soma dos quadrados da regressão em duas partes componentes, conforme ilustrado na Tabela 14.3.

TABELA 14.3
Tabela de ANOVA Dividindo a Soma dos Quadrados da Regressão em Componentes para Determinar a Contribuição da Variável X_1

Fonte	Graus de Liberdade	Soma dos Quadrados	Média dos Quadrados (Variância)	*F*
Regressão	2	39.472.730,77	19.736.365,39	
$\begin{Bmatrix} X_2 \\ X_1 \mid X_2 \end{Bmatrix}$	$\begin{Bmatrix} 1 \\ 1 \end{Bmatrix}$	$\begin{Bmatrix} 14.915.814,10 \\ 24.556.916,67 \end{Bmatrix}$	24.556.916,67	60,32
Erro	31	12.620.946,67	407.127,31	
Total	33	52.093.677,44		

A hipótese nula e a hipótese alternativa para testar a contribuição de X_1 para o modelo são:

H_0: A variável X_1 não aperfeiçoa significativamente o modelo depois que a variável X_2 foi incluída.

H_1: A variável X_1 aperfeiçoa significativamente o modelo depois que a variável X_2 foi incluída.

A Equação (14.11) define o teste F parcial para testar a contribuição de uma variável independente.

ESTATÍSTICA DO TESTE *F* PARCIAL

$$F_{ESTAT} = \frac{SQReg\,(X_j \mid \text{Todas as variáveis } X \text{ exceto } j)}{MQReg} \qquad \textbf{(14.11)}$$

A estatística do teste F parcial segue uma distribuição F, com 1 e $n - k - 1$ graus de liberdade.

Com base na Tabela 14.3,

$$F_{ESTAT} = \frac{24.556.916,67}{407.127,31} = 60,32$$

A estatística do teste F_{ESTAT} parcial possui 1 e $n - k - 1 = 34 - 2 - 1 = 31$ graus de liberdade. Utilizando um nível de significância de 0,05, o valor crítico, com base na Tabela E.5, é aproximadamente 4,17 (veja a Figura 14.8).

FIGURA 14.8
Testando a contribuição de um coeficiente de regressão para um modelo de regressão múltipla, no nível de significância de 0,05, com 1 e 31 graus de liberdade

Uma vez que a estatística do teste F_{ESTAT} parcial calculada é maior do que esse valor crítico de F (60,32 > 4,17), rejeite H_0. Você consegue concluir que o acréscimo da variável X_1 (preço) melhora significativamente um modelo de regressão que já contém a variável X_2 (despesas com promoções).

Para avaliar a contribuição da variável X_2 (despesas com promoções) para um modelo no qual a variável X_1 (preço) foi incluída, você precisa utilizar a Equação (14.10b). Em primeiro lugar, com base na Figura 14.7, observe que $SQReg(X_1 \text{ e } X_2) = 28.153.486,15$. Em segundo lugar, a partir da Tabela 14.3, observe que $SQReg(X_1 \text{ e } X_2) = 39.472.730,77$. Então, utilizando a Equação (14.10b),

$$SQReg(X_2 \mid X_1) = 39.472.730,77 - 28.153.486,15 = 11.319.244,62$$

Para determinar se X_2 melhora significativamente um modelo depois que X_1 foi incluído, você pode dividir a soma dos quadrados da regressão em duas partes componentes, conforme mostrado na Tabela 14.4.

TABELA 14.4
Tabela de ANOVA Dividindo a Soma dos Quadrados da Regressão em Componentes para Determinar a Contribuição da Variável X_2

Fonte	Graus de Liberdade	Soma dos Quadrados	Média dos Quadrados (Variância)	*F*
Regressão	2	39.472.730,77	19.736.365,39	
$\begin{Bmatrix} X_1 \\ X_2 \mid X_1 \end{Bmatrix}$	$\begin{Bmatrix} 1 \\ 1 \end{Bmatrix}$	$\begin{Bmatrix} 28.153.486,15 \\ 11.319.244,62 \end{Bmatrix}$	11.319.244,62	27,80
Erro	31	12.620.946,67	407.127,31	
Total	33	52.093.677,44		

A hipótese nula e a hipótese alternativa para testar a contribuição de X_2 para o modelo são

H_0: A variável X_2 não melhora significativamente o modelo depois que a variável X_1 foi incluída.

H_1: A variável X_2 melhora significativamente o modelo depois que a variável X_1 foi incluída.

Utilizando a Equação (14.11) e a Tabela 14.4,

$$F_{ESTAT} = \frac{11.319.244,62}{407.127,31} = 27,80$$

Na Figura 14.8, você pode verificar que, utilizando um nível de significância de 0,05, o valor crítico de F, com 1 e 31 graus de liberdade, é aproximadamente 4,17. Uma vez que a estatística do teste F_{ESTAT} parcial calculada é maior do que esse valor crítico (27,80 > 4,17), rejeite H_0. Você consegue concluir que o acréscimo da variável X_2 (despesas com promoções) melhora significativamente o modelo de regressão múltipla que já contém X_1 (preço).

Por conseguinte, ao testar a contribuição de cada uma das variáveis independentes depois que a outra foi incluída no modelo, você determina que cada uma das duas variáveis independentes melhora significativamente o modelo. Portanto, o modelo de regressão múltipla deve incluir tanto o preço, X_1, quanto as despesas com promoções, X_2.

A estatística do teste F parcial desenvolvida nesta seção e a estatística do teste t da Equação (14.7) são, ambas, utilizadas para determinar a contribuição de uma variável independente a um modelo de regressão múltipla. Os testes de hipóteses associados a essas duas estatísticas sempre resultam na mesma decisão (ou seja, os valores-p são idênticos). A estatística do teste t_{ESTAT} para o modelo de regressão da OmniPower são −7,7664 e +5,2728, e as estatísticas correspondentes do teste F_{ESTAT} são 60,32 e 27,80. A Equação (14.12) expressa essa relação entre t e F.[1]

[1]Essa relação se mantém somente quando a estatística F_{ESTAT} possui 1 grau de liberdade no numerador.

RELAÇÃO ENTRE UMA ESTATÍSTICA t E UMA ESTATÍSTICA F

$$t^2_{ESTAT} = F_{ESTAT} \qquad \textbf{(14.12)}$$

Coeficientes de Determinação Parcial

Lembre-se, com base na Seção 14.2, de que o coeficiente de determinação múltipla, r^2, mede a proporção da variação em Y que é explicada pela variação nas variáveis independentes. Os **coeficientes de determinação parcial** ($r^2_{Y1.2}$ e $r^2_{Y2.1}$) medem a proporção da variação na variável dependente que é explicada por cada uma das variáveis independentes, ao mesmo tempo em que a outra variável independente é controlada ou mantida constante. A Equação (14.13) define os coeficientes de determinação parcial para um modelo de regressão múltipla com duas variáveis independentes.

COEFICIENTES DE DETERMINAÇÃO PARCIAL PARA UM MODELO DE REGRESSÃO MÚLTIPLA QUE CONTÉM DUAS VARIÁVEIS INDEPENDENTES

$$r^2_{Y1.2} = \frac{SQReg(X_1 \mid X_2)}{STQ - SQReg(X_1 \text{ e } X_2) + SQReg(X_1 \mid X_2)} \qquad \textbf{(14.13a)}$$

e

$$r^2_{Y2.1} = \frac{SQReg(X_2 \mid X_1)}{STQ - SQReg(X_1 \text{ e } X_2) + SQReg(X_2 \mid X_1)} \qquad \textbf{(14.13b)}$$

em que

$SQReg(X_1 \mid X_2)$ = soma dos quadrados da contribuição da variável X_1 ao modelo de regressão, considerando-se que a variável X_2 foi incluída no modelo

STQ = soma total dos quadrados para Y

$SQReg(X_1 \text{ e } X_2)$ = soma dos quadrados da regressão quando as variáveis X_1 e X_2 estão, ambas, incluídas no modelo de regressão múltipla

$SQReg(X_2 \mid X_1)$ = soma dos quadrados da contribuição da variável X_2 para o modelo de regressão, considerando-se que a variável X_1 foi incluída no modelo

Para o exemplo de vendas de OmniPower,

$$r^2_{Y1.2} = \frac{24.556.916,67}{52.093.677,44 - 39.472.730,77 + 24.556.916,67} = 0,6605$$

$$r^2_{Y2.1} = \frac{11.319.244,62}{52.093.677,44 - 39.472.730,77 + 11.319.244,62} = 0,4728$$

O coeficiente de determinação parcial, $r^2_{Y1.2}$, da variável Y com X_1, mantendo X_2 constante, é 0,6605. Por conseguinte, para um determinado montante (constante) de despesas com promoções, 66,05% da variação nas vendas de OmniPower é explicada pela variação no preço. O coeficiente de determinação parcial, $r^2_{Y2.1}$, da variável Y com X_2, mantendo X_1 constante, é 0,4728. Por conseguinte, para um determinado preço (constante), 47,28% da variação nas vendas de barras OmniPower pode ser explicada pela variação no montante de despesas com promoções.

A Equação (14.4) define o coeficiente de determinação parcial para a j-ésima variável em um modelo de regressão múltipla que contém diversas variáveis independentes (k).

COEFICIENTE DE DETERMINAÇÃO PARCIAL PARA UM MODELO DE REGRESSÃO MÚLTIPLA QUE CONTÉM k VARIÁVEIS INDEPENDENTES

$$r^2_{Y_j.(\text{Todas as variáveis } exceto\ j)} = \frac{SQReg\,(X_j \mid \text{Todas as variáveis } X \text{ } exceto\ j)}{STQ - SQReg(\text{Todas as variáveis } X) + SQReg(X_j \mid \text{Todas as variáveis } X \text{ } exceto\ j)} \qquad \textbf{(14.14)}$$

Problemas para a Seção 14.5

APRENDENDO O BÁSICO

14.31 É apresentada a seguir a tabela resumida de ANOVA para um modelo de regressão múltipla com duas variáveis independentes:

Fonte	Graus de Liberdade	Soma dos Quadrados	Média dos Quadrados	F
Regressão	2	60		
Erro	18	120		
Total	20	180		

Se $SQReg(X_1) = 45$ e $SQReg(X_2) = 25$,

a. determine se existe uma relação significativa entre Y e cada uma das variáveis independentes, no nível de significância de 0,05.

b. calcule os coeficientes de determinação parcial, $r^2_{Y1.2}$ e $r^2_{Y2.1}$, e interprete seus respectivos significados.

14.32 É apresentada a seguir a tabela resumida de ANOVA para um modelo de regressão múltipla com duas variáveis independentes:

Fonte	Graus de Liberdade	Soma dos Quadrados	Média dos Quadrados	F
Regressão	2	30		
Erro	10	120		
Total	12	150		

Se $SQReg(X_1) = 20$ e $SQReg(X_2) = 15$,

a. determine se existe uma relação significativa entre Y e cada uma das variáveis independentes, no nível de significância de 0,05.

b. calcule os coeficientes de determinação parcial, $r^2_{Y1.2}$ e $r^2_{Y2.1}$, e interprete seus respectivos significados.

APLICANDO OS CONCEITOS

14.33 No Problema 14.5, em Problemas para a Seção 14.1, você utilizou a potência, em cavalos-vapor, e o peso do automóvel para prever a milhagem em relação ao consumo de gasolina (dados no arquivo **Auto**). Utilizando os resultados daquele problema,

a. no nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Com base nesses resultados, indique o modelo de regressão mais apropriado para esse conjunto de dados.

b. calcule os coeficientes de determinação parcial, $r^2_{Y1.2}$ e $r^2_{Y2.1}$, e interprete seus respectivos significados.

AUTO-AVALIAÇÃO **14.34** No Problema 14.4, em Problemas para a Seção 14.1, você utilizou as vendas e a quantidade de pedidos para prever os custos de distribuição de uma empresa de venda por catálogo (dados no arquivo **CustoDeposito**). Utilizando os resultados daquele problema,

a. no nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Com base nesses resultados, indique o modelo de regressão mais apropriado para esse conjunto de dados.

b. calcule os coeficientes de determinação parcial, $r^2_{Y1.2}$ e $r^2_{Y2.1}$, e interprete seus respectivos significados.

14.35 No Problema 14.7, em Problemas para a Seção 14.1, você utilizou o total da equipe presente e as horas remotas para prever as horas de sobreaviso (dados no arquivo **Sobreaviso**). Utilizando os resultados daquele problema,

a. no nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Com base nesses resultados, indique o modelo de regressão mais apropriado para esse conjunto de dados.

b. calcule os coeficientes de determinação parcial, $r^2_{Y1.2}$ e $r^2_{Y2.1}$, e interprete seus respectivos significados.

14.36 No Problema 14.6, em Problemas para a Seção 14.1, você utilizou a verba gasta em propaganda no rádio e em jornais para prever vendas (dados no arquivo **Propaganda**). Utilizando os resultados daquele problema,

a. no nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Com base nesses resultados, indique o modelo de regressão mais apropriado para esse conjunto de dados.

b. calcule os coeficientes de determinação parcial, $r^2_{Y1.2}$ e $r^2_{Y2.1}$, e interprete seus respectivos significados.

14.37 No Problema 14.8, em Problemas para a Seção 14.1, você utilizou a área do terreno de uma propriedade residencial e a idade (tempo de construção) de um imóvel para prever o valor de avaliação (dados no arquivo **GlenCove**). Utilizando os resultados daquele problema,

a. no nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Com base nesses resultados, indique o modelo de regressão mais apropriado para esse conjunto de dados.

b. calcule os coeficientes de determinação parcial, $r^2_{Y1.2}$ e $r^2_{Y2.1}$, e interprete seus respectivos significados.

14.6 Utilizando Variáveis Binárias (*Dummy*) e Termos de Interação em Modelos de Regressão

Os modelos de regressão múltipla, discutidos nas Seções 14.1 a 14.5, adotavam o pressuposto de que cada uma das variáveis independentes é numérica. Por exemplo, na Seção 14.1, você utilizou preço e despesas com promoções para prever as vendas mensais de barras energéticas OmniPower. No entanto, para alguns modelos, pode ser desejável incluir o efeito de uma variável categórica independente. Por exemplo, para prever as vendas mensais de barras OmniPower, pode ser desejável incluir no modelo a variável categórica local de exposição na prateleira (exposição fora da ponta de corredor ou em ponta de corredor).

Para incluir uma variável independente categórica em um modelo de regressão, utilize uma **variável binária (*dummy*)**. Uma variável binária (*dummy*) recodifica as categorias de uma variável categórica utilizando os valores numéricos 0 e 1. Se uma determinada variável independente categórica possui somente duas categorias, assim como local de exposição na prateleira no exemplo anterior, você pode, então, definir uma variável binária (*dummy*), X_d, de modo a representar as duas categorias, como

$X_d = 0$, se a observação estiver na categoria 1 (exposição fora da ponta de corredor, no exemplo)

$X_d = 1$, se a observação estiver na categoria 2 (exposição em ponta de corredor, no exemplo)

Para ilustrar a aplicação de variáveis *dummy* (binárias) à regressão, considere o problema estratégico que envolve o desenvolvimento de um modelo para prever o valor de avaliação de casas ($1.000) com base no tamanho do imóvel (em milhares de pés quadrados) e no fato de a casa possuir ou não uma lareira. Para incluir a variável categórica que se refere à presença de uma lareira, a variável binária (*dummy*), X_2, é definida como

$X_2 = 0$, se a casa não possui uma lareira

$X_2 = 1$, se a casa possui uma lareira

Dados coletados de uma amostra de 15 casas estão organizados e armazenados no arquivo **Casa3**. A Tabela 14.5 apresenta os dados. Na última coluna da Tabela 14.5, você pode verificar como os dados categóricos são convertidos em valores numéricos.

Pressupondo que a inclinação do valor de avaliação em relação ao tamanho do imóvel seja a mesma para as casas com e sem lareira, o modelo de regressão múltipla é

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

TABELA 14.5
Prevendo o Valor de Avaliação, com Base no Tamanho da Casa e na Presença de uma Lareira

Crie termos de variáveis binárias (dummy) utilizando as instruções na Seção GE14.6.

Valor de Avaliação	Tamanho	Lareira	Lareira Codificada
234,4	2,00	Sim	1
227,4	1,71	Não	0
225,7	1,45	Não	0
235,9	1,76	Sim	1
229,1	1,93	Não	0
220,4	1,20	Sim	1
225,8	1,55	Sim	1
235,9	1,93	Sim	1
228,5	1,59	Sim	1
229,2	1,50	Sim	1
236,7	1,90	Sim	1
229,3	1,39	Sim	1
224,5	1,54	Não	0
233,8	1,89	Sim	1
226,8	1,59	Não	0

em que

Y_i = valor de avaliação, em milhares de dólares, para a casa i

β_0 = intercepto de Y

X_{1i} = tamanho do imóvel, em milhares de pés quadrados, para a casa i

β_1 = inclinação do valor de avaliação em relação ao tamanho do imóvel, mantendo-se constante a presença ou a ausência de uma lareira

X_{2i} = variável binária (*dummy*) representando a presença ou a ausência de uma lareira para a casa i

β_2 = efeito incremental líquido decorrente da presença de uma lareira em relação ao valor de avaliação, mantendo-se constante o tamanho do imóvel

ε_i = erro aleatório em Y para a casa i

A Figura 14.9 apresenta a planilha com os resultados da regressão para esse modelo.

FIGURA 14.9
Planilha com os resultados da regressão para o modelo de regressão que inclui tamanho da casa e presença de lareira

Crie termos de variáveis binárias (dummy) utilizando as instruções na Seção GE14.6.

	A	B	C	D	E	F	G
1	Análise do Valor de Avaliação						
2							
3	*Estatística de Regressão*						
4	R Múltiplo	0,9006					
5	R-Quadrado	0,8111					
6	R-Quadrado Ajustado	0,7796					
7	Erro-padrão	2,2626					
8	Observações	15					
9							
10	ANOVA						
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*	
12	Regressão	2	263,7039	131,8520	25,7557	0,0000	
13	Resíduo	12	61,4321	5,1193			
14	Total	14	325,1360				
15							
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*
17	Interseção	200,0905	4,3517	45,9803	0,0000	190,6090	209,5719
18	Tamanho	16,1858	2,5744	6,2871	0,0000	10,5766	21,7951
19	Lareira Codificada	3,8530	1,2412	3,1042	0,0091	1,1486	6,5574

Com base na Figura 14.9, a equação da regressão é:

$$\hat{Y}_i = 200{,}0905 + 16{,}1858X_{1i} + 3{,}8530X_{2i}$$

Para casas sem lareira, você substitui X_2 por 0 na equação da regressão:

$$\begin{aligned}\hat{Y}_i &= 200{,}0905 + 16{,}1858X_{1i} + 3{,}8530X_{2i}\\ &= 200{,}0905 + 16{,}1858X_{1i} + 3{,}8530(0)\\ &= 200{,}0905 + 16{,}1858X_{1i}\end{aligned}$$

Para casas com lareira, você substitui X_2 por 1 na equação da regressão:

$$\begin{aligned}\hat{Y}_i &= 200{,}0905 + 16{,}1858X_{1i} + 3{,}8530X_{2i}\\ &= 200{,}0905 + 16{,}1858X_{1i} + 3{,}8530(1)\\ &= 203{,}9435 + 16{,}1858X_{1i}\end{aligned}$$

Nesse modelo, os coeficientes da regressão são interpretados da seguinte maneira:

1. Mantendo-se constante o fato de uma casa ter, ou não, uma lareira, para cada acréscimo correspondente a 1.000 pés quadrados no tamanho do imóvel estima-se que o valor de avaliação previsto aumente em 16,1858 milhares de dólares (ou seja, $16.185,80).
2. Mantendo-se constante o tamanho do imóvel, estima-se que a presença de uma lareira faça crescer em 3,8530 mil dólares (ou $3.853) o valor de avaliação previsto para o imóvel.

Na Figura 14.9, a estatística do teste t_{ESTAT} para a inclinação do tamanho do imóvel em relação ao valor de avaliação é igual a 6,2871, e o valor-p é aproximadamente 0,000; a estatística do teste t_{ESTAT} para a presença de uma lareira é igual a 3,1042, e o valor-p é 0,0091. Por conseguinte, cada uma das duas variáveis oferece uma contribuição significativa ao modelo no nível de significância de 0,01. Além disso, o coeficiente de determinação múltipla indica que 81,11% da variação no valor de avaliação são explicados pela variação no tamanho do imóvel e pelo fato de a casa ter ou não uma lareira.

EXEMPLO 14.3

Modelando uma Variável Categórica de Três Níveis

Defina um modelo de regressão múltipla utilizando vendas como a variável dependente e modelo da embalagem e preço como variáveis independentes. Modelo da embalagem é uma variável categórica contendo três níveis, com os modelos A, B ou C.

SOLUÇÃO Para modelar uma variável categórica de três níveis, modelo de embalagem, são necessárias duas variáveis binárias (dummy), X_1 e X_2:

$X_{1i} = 1$, se for utilizado o modelo de embalagem A na observação i; caso contrário, 0.

$X_{2i} = 1$, se for utilizado o modelo de embalagem B na observação i; caso contrário, 0.

Por conseguinte, se a observação i utiliza o modelo de embalagem A, então $X_{1i} = 1$ e $X_{2i} = 0$; se a observação i utiliza o modelo de embalagem B, então $X_{1i} = 0$ e $X_{2i} = 1$; e se a observação i utiliza o modelo de embalagem C, então $X_{1i} = X_{2i} = 0$. Uma terceira variável independente é utilizada para preço:

$$X_{3i} = \text{preço para a observação } i$$

Portanto, o modelo de regressão para este exemplo é

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

em que

Y_i = vendas, para a observação i

β_0 = intercepto de Y

β_1 = diferença entre as vendas previstas para o modelo A e as vendas previstas para o modelo C, mantendo-se constante o preço

β_2 = diferença entre as vendas previstas para o modelo B e as vendas previstas para o modelo C, mantendo-se constante o preço

β_3 = inclinação de vendas em relação a preço, mantendo-se constante o modelo da embalagem

ε_i = erro aleatório em Y para a observação i

Interações

Em todos os modelos de regressão discutidos até agora, foi pressuposto que o efeito que uma variável independente exerce sobre a variável dependente é estatisticamente independente das outras variáveis independentes no modelo. Ocorre uma **interação** se o efeito de uma variável independente sobre a variável dependente se modifica de acordo com o *valor* de uma segunda variável independente. Por exemplo, é possível que a propaganda exerça um efeito significativo sobre as vendas de um produto quando o preço desse produto é baixo. Entretanto, se o preço do produto for demasiadamente alto, incrementos na propaganda não modificarão radicalmente as vendas. Nesse caso, afirma-se que preço e propaganda interagem. Em outras palavras, não se pode fazer afirmações generalizadas no tocante ao efeito da propaganda sobre as vendas. O efeito que a propaganda exerce sobre as vendas é *dependente* do preço. Você utiliza um **termo de interação** (algumas vezes conhecido como **termo de multiplicação**) para modelar um efeito de interação em um modelo de regressão.

Para ilustrar o conceito de interação e o uso de um termo de interação, retorne ao exemplo que trata dos valores de avaliação de imóveis residenciais, no início desta seção. No modelo de regressão, você adotou o pressuposto de que o efeito que o tamanho do imóvel exerce sobre o valor de avaliação é dependente do fato de a casa ter ou não uma lareira. Em outras palavras, você partiu do pressuposto de que a inclinação do valor de avaliação em relação ao tamanho do imóvel era a mesma para casas com lareira e para casas sem lareira. Caso essas duas inclinações sejam diferentes, existe uma interação entre o tamanho do imóvel e a existência de uma lareira.

Para avaliar a possibilidade de existência de uma interação, você define, inicialmente, um termo de interação que corresponda ao produto entre a variável independente X_1 (tamanho da casa) e a variável binária (*dummy*) X_2 (lareira). Depois disso, você testa se essa variável de interação oferece uma contribuição significativa para o modelo de regressão. Caso a interação seja significativa, você não pode utilizar o modelo original para fins de previsão. Para os dados da Tabela 14.5, você define o seguinte:

$$X_3 = X_1 \times X_2$$

A Figura 14.10 apresenta a planilha com os resultados da regressão para esse modelo de regressão que inclui o tamanho do imóvel, X_1, a presença de uma lareira, X_2, e a interação entre X_1 e X_2 (definida como X_3).

FIGURA 14.10
Planilha de resultados da regressão para um modelo que inclui tamanho da casa, presença de lareira e interação entre tamanho e lareira

Crie termos de interação utilizando as instruções na Seção GE14.6.

	A	B	C	D	E	F	G
1	Análise do Valor de Avaliação						
2							
3	*Estatística de Regressão*						
4	R Múltiplo	0,9179					
5	R-Quadrado	0,8426					
6	R-Quadrado Ajustado	0,7996					
7	Erro-padrão	2,1573					
8	Observações	15					
9							
10	ANOVA						
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*	
12	Regressão	3	273,9441	91,3147	19,6215	0,0001	
13	Resíduo	11	51,1919	4,6538			
14	Total	14	325,1360				
15							
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*
17	Interseção	212,9522	9,6122	22,1544	0,0000	191,7959	234,1084
18	Tamanho	8,3624	5,8173	1,4375	0,1784	-4,4414	21,1662
19	Lareira Codificada	-11,8404	10,6455	-1,1122	0,2898	-35,2710	11,5902
20	Tamanho * Lareira Codificada	9,5180	6,4165	1,4834	0,1661	-4,6046	23,6406

Para testar a existência de uma interação, você utiliza a hipótese nula

$$H_0: \beta_3 = 0$$

contra a hipótese alternativa

$$H_1: \beta_3 \neq 0.$$

Na Figura 14.10, a estatística do teste t_{ESTAT} para a interação entre tamanho do imóvel e a presença de uma lareira é 1,4834. Uma vez que $t_{ESTAT} = 1{,}4834 < 2{,}201$ ou valor-$p = 0{,}1661 > 0{,}05$, você não rejeita a hipótese nula. Portanto, a interação não oferece uma contribuição significativa para o modelo, dado que tamanho do imóvel e presença de uma lareira já estão incluídos. Você pode concluir que a inclinação do valor de avaliação em relação ao tamanho do imóvel é a mesma para casas com lareira e sem lareira.

Modelos de regressão podem conter diversas variáveis numéricas independentes. O Exemplo 14.4 ilustra um modelo de regressão no qual existem duas variáveis numéricas independentes, assim como uma variável categórica independente.

EXEMPLO 14.4

Estudando um Modelo de Regressão que Contém uma Variável Binária (*Dummy*)

O problema estratégico com o qual se depara um corretor imobiliário envolve a previsão do consumo de óleo para calefação em residências unifamiliares. As variáveis independentes consideradas são a temperatura atmosférica, X_1, e a quantidade de isolamento no sótão, X_2. São coletados dados de uma amostra de 15 residências unifamiliares. Dentre as 15 casas selecionadas, as casas 1, 4, 6, 7, 8, 10 e 12 são casas no estilo colonial. Os dados estão organizados e armazenados no arquivo **Calefação**. Desenvolva e analise um modelo de regressão apropriado utilizando essas três variáveis independentes, X_1, X_2 e X_3 (em que X_3 é a variável binária para casas em estilo colonial).

SOLUÇÃO Defina X_3, uma variável binária (*dummy*) para casa em estilo colonial, como se segue:

$X_3 = 0$, se o estilo não for colonial

$X_3 = 1$, se o estilo for colonial

Pressupondo que a inclinação entre consumo de óleo para calefação residencial e temperatura atmosférica, X_1, e entre consumo de óleo para calefação residencial e quantidade de isolamento térmico no sótão, X_2, seja a mesma para ambos os estilos de casas, o modelo de regressão é

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

em que

$Y_i =$ consumo mensal de óleo para calefação, em galões, para a casa i

$\beta_0 =$ intercepto de Y

$\beta_1 =$ inclinação do consumo de óleo para calefação com temperatura atmosférica, mantendo-se constantes o efeito do isolamento no sótão e o estilo da casa

$\beta_2 =$ inclinação do consumo de óleo para calefação com isolamento no sótão, mantendo-se constantes o efeito da temperatura atmosférica e o estilo da casa

$\beta_3 =$ efeito incremental da presença de uma casa em estilo colonial, mantendo-se constantes os efeitos da temperatura atmosférica e do isolamento no sótão

$\varepsilon_i =$ erro aleatório em Y para a casa i

A Figura 14.11 ilustra os resultados de regressão no Microsoft Excel para esse modelo.

FIGURA 14.11
Planilha de resultados da regressão para um modelo de regressão que inclui temperatura, isolamento e estilo para os dados sobre óleo para calefação

	A	B	C	D	E	F	G
1	**Análise do Consumo de Óleo para Calefação**						
2							
3	*Estatística de Regressão*						
4	R Múltiplo	0,9942					
5	R-Quadrado	0,9884					
6	R-Quadrado Ajustado	0,9853					
7	Erro-padrão	15,7489					
8	Observações	15					
9							
10	ANOVA						
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*	
12	Regressão	3	233406,9094	77802,3031	313,6822	0,0000	
13	Resíduo	11	2728,3200	248,0291			
14	Total	14	236135,2293				
15							
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*
17	Interseção	592,5401	14,3370	41,3295	0,0000	560,9846	624,0956
18	Temperatura	-5,5251	0,2044	-27,0267	0,0000	-5,9751	-5,0752
19	Isolamento	-21,3761	1,4480	-14,7623	0,0000	-24,5632	-18,1891
20	Estilo colonial	-38,9727	8,3584	-4,6627	0,0007	-57,3695	-20,5759

Partindo dos resultados na Figura 14.11, a equação da regressão é:

$$\hat{Y}_i = 592{,}5401 - 5{,}5251X_{1i} - 21{,}3761X_{2i} - 38{,}9727X_{3i}$$

Para casas que não são em estilo colonial, uma vez que $X_3 = 0$, a equação da regressão se reduz a

$$\hat{Y}_i = 592{,}5401 - 5{,}5251X_{1i} - 21{,}3761X_{2i}$$

Para casas que são em estilo colonial, uma vez que $X_3 = 0$, a equação da regressão se reduz a

$$\hat{Y}_i = 553{,}5674 - 5{,}5251X_{1i} - 21{,}3761X_{2i}$$

Os coeficientes da regressão são interpretados como se segue:

1. Mantendo-se constantes o isolamento no sótão e o estilo da casa, para cada 1°F (grau Fahrenheit) adicional de aumento na temperatura atmosférica, você estima que o consumo previsto de óleo para calefação decresça em 5,5251 galões.
2. Mantendo-se constantes a temperatura atmosférica e o estilo da casa, para cada 1 polegada adicional de aumento no isolamento do sótão você estima que o consumo previsto de óleo para calefação decresça em 21,3761 galões.
3. b_3 mede o efeito no consumo de óleo decorrente de a casa ter estilo colonial ($X_3 = 1$), comparado ao fato de ter uma casa que não seja em estilo colonial ($X_3 = 0$). Por conseguinte, mantendo-se constantes a temperatura atmosférica e o isolamento no sótão, você estima que o consumo previsto de óleo para calefação seja 38,9727 galões a menos para uma casa em estilo colonial do que para uma casa que não seja em estilo colonial.

As três estatísticas t_{ESTAT} que representam as inclinações para temperatura, isolamento e estilo colonial são −27,0267, −14,7623 e −4,6627. Cada um dos valores-p correspondentes é extremamente pequeno (inferiores a 0,001). Assim, cada uma das três variáveis fornece uma contribuição significativa para o modelo. Além disso, o coeficiente de determinação múltipla indica que 98,84% da variação no consumo de óleo é explicada pela variação na temperatura, pelo isolamento e pelo fato de a casa ser ou não em estilo colonial.

Antes que possa utilizar o modelo no Exemplo 14.4, você precisa determinar se as variáveis independentes interagem umas com as outras. No Exemplo 14.5, são acrescentados ao modelo três termos de interação.

EXEMPLO 14.5

Avaliando um Modelo de Regressão com Diversas Interações

Para os dados do Exemplo 14.4, determine se o acréscimo de termos de interação oferece uma contribuição significativa para o modelo de regressão.

SOLUÇÃO Para avaliar possíveis interações entre as variáveis independentes, são construídos três termos de interação, do seguinte modo: $X_4 = X_1 \times X_2$; $X_5 = X_1 \times X_3$ e $X_6 = X_2 \times X_3$.

O modelo de regressão é agora

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i$$

em que X_1 é a temperatura, X_2 é o isolamento; X_3 é a variável binária para estilo colonial, X_4 é a interação entre temperatura e isolamento, X_5 é a interação entre temperatura e estilo colonial e X_6 é a interação entre isolamento e estilo colonial. A Figura 14.12 apresenta a planilha com os resultados da regressão para o modelo em questão.

Para testar se as três interações aperfeiçoam significativamente o modelo de regressão, você utiliza o teste F parcial. A hipótese nula e a hipótese alternativa são

H_0: $\beta_4 = \beta_5 = \beta_6 = 0$ (Não existe nenhuma interação entre X_1, X_2 e X_3.)

H_1: $\beta_4 \neq 0$ e/ou $\beta_5 \neq 0$ e/ou $\beta_6 \neq 0$ (X_1 interage com X_2 e/ou X_1 interage com X_3 e/ou X_2 interage com X_3.)

Com base na Figura 14.12,

$$SQReg(X_1, X_2, X_3, X_4, X_5, X_6) = 234.510{,}5818 \text{ com 6 graus de liberdade}$$

e com base na Figura 14.11, $SQReg(X_1, X_2, X_3) = 233.406{,}9094$ com 3 graus de liberdade.

Por conseguinte, $SQReg(X_1, X_2, X_3, X_4, X_5, X_6) - SQReg(X_1, X_2, X_3) = 234.510{,}5818 - 233.406{,}9094 = 1.103{,}6724$. A diferença, em graus de liberdade, é $6 - 3 = 3$.

Para utilizar o teste F parcial para a contribuição simultânea de três variáveis para um modelo, você utiliza uma extensão da Equação (14.11).[2] A estatística do teste F_{ESTAT} parcial é

[2]Em geral, se um modelo possui diversas variáveis independentes e você deseja testar se um conjunto adicional de variáveis independentes contribui para o modelo, o numerador do teste F é $SQReg$ (para todas as variáveis independentes) − $SQReg$ (para o conjunto inicial de variáveis), dividido pelo número de variáveis independentes cuja contribuição está sendo testada.

FIGURA 14.12
Planilha de resultados da regressão para um modelo de regressão que inclui temperatura, X_1; isolamento, X_2; a variável binária (*dummy*) estilo colonial, X_3; a interação entre temperatura e isolamento, X_4; a interação entre temperatura e estilo colonial, X_5; e a interação entre isolamento e estilo colonial, X_6.

	A	B	C	D	E	F	G
1	Análise do Consumo de Óleo para Calefação						
2							
3	*Estatística de Regressão*						
4	R Múltiplo	0,9966					
5	R-Quadrado	0,9931					
6	R-Quadrado Ajustado	0,9880					
7	Erro-padrão	14,2506					
8	Observações	15					
9							
10	ANOVA						
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*	
12	Regressão	6	234510,5818	39085,0970	192,4607	0,0000	
13	Resíduo	8	1624,6475	203,0809			
14	Total	14	236135,2293				
15							
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*
17	Interseção	642,8867	26,7059	24,0728	0,0000	581,3027	704,4707
18	Temperatura	-6,9263	0,7531	-9,1969	0,0000	-8,6629	-5,1896
19	Isolamento	-27,8825	3,5801	-7,7882	0,0001	-36,1383	-19,6268
20	Estilo	-84,6088	29,9956	-2,8207	0,0225	-153,7788	-15,4389
21	Temperatura * Isolamento	0,1702	0,0886	1,9204	0,0911	-0,0342	0,3746
22	Temperatura * Estilo colonial	0,6596	0,4617	1,4286	0,1910	-0,4051	1,7242
23	Isolamento * Estilo colonial	4,9870	3,5137	1,4193	0,1936	-3,1156	13,0895

$$F = \frac{[SQReg(X_1, X_2, X_3, X_4, X_5, X_6) - SQReg(X_1, X_2, X_3)]/3}{MQR(X_1, X_2, X_3, X_4, X_5, X_6)} = \frac{1.103,6724/3}{203,0809} = 1,8115$$

Você compara a estatística calculada do teste F_{ESTAT}, com o valor crítico de F para 3 e 8 graus de liberdade. Utilizando um nível de significância de 0,05, o valor crítico de F, com base na Tabela E.5, é 4,07. Uma vez que $F_{ESTAT} = 1,8115 < 4,07$, você conclui que as interações não oferecem uma contribuição significativa ao modelo, sabendo-se que o modelo já inclui temperatura, X_1; isolamento, X_2; e o fato de a casa ser ou não em estilo colonial, X_3. Portanto, o modelo de regressão múltipla que utiliza X_1, X_2 e X_3, mas nenhum termo de interação, é o melhor modelo. Se tivesse rejeitado essa hipótese nula, você testaria, então, a contribuição de cada uma das interações, separadamente, para determinar quais termos de interação deveriam ser incluídos no modelo.

Problemas para a Seção 14.6

APRENDENDO O BÁSICO

14.38 Suponha que X_1 seja uma variável numérica, X_2 seja uma variável binária e a equação da regressão para uma amostra de $n = 20$ seja

$$\hat{Y}_i = 6 + 4X_{1i} + 2X_{2i}$$

a. Interprete o coeficiente da regressão associado à variável X_1.
b. Interprete o coeficiente da regressão associado à variável X_2.
c. Suponha que a estatística t_{ESTAT} para testar a contribuição da variável X_2 seja 3,27. No nível de significância de 0,05, existem evidências de que a variável X_2 oferece uma contribuição significativa para o modelo?

APLICANDO OS CONCEITOS

14.39 O decano do departamento de contabilidade de uma universidade deseja desenvolver um modelo de regressão para prever a média de pontos de conceito em contabilidade de alunos que estão se formando em contabilidade, com base no total dos resultados do SAT (Scholastic Aptitude Test – Teste de Aptidão Escolar) do aluno e no fato de o aluno ter recebido, ou não, um conceito B ou maior que B no curso de introdução à estatística (0 = não e 1 = sim).

a. Explique as etapas envolvidas no desenvolvimento de um modelo de regressão para esses dados. Não deixe de indicar os modelos específicos que você precisa avaliar e comparar.
b. Suponha que o coeficiente de regressão para a variável que se refere a o aluno ter, ou não, recebido um conceito B ou maior que B, no curso de introdução à estatística, seja +0,30. Como você interpreta esse resultado?

14.40 Uma imobiliária em uma comunidade no subúrbio dos Estados Unidos gostaria de estudar a relação entre o tamanho de um imóvel unifamiliar (medido em termos do número de cômodos) e o preço de venda do imóvel (em milhares de dólares). Dois diferentes bairros estão incluídos no estudo, um deles na região leste da comunidade (= 0) e o outro na região oeste (= 1). Foi selecionada uma amostra aleatória de 20 casas, com os resultados fornecidos no arquivo **Vizinhança**. Para os itens de (a) a (k), não inclua um termo de interação.

a. Expresse a equação da regressão múltipla que preveja o preço de venda, com base no número de cômodos e no bairro.
b. Interprete os coeficientes da regressão em (a).
c. Faça a previsão para o preço de venda de uma casa com nove cômodos localizada em um bairro na região leste. Construa uma estimativa do intervalo de confiança de 95% e um intervalo de previsão de 95%.
d. Realize uma análise de resíduos nos resultados e determine se os pressupostos da regressão são válidos.
e. Existe uma relação significativa entre o preço de venda e as duas variáveis independentes (cômodos e bairro), no nível de significância de 0,05?
f. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.
g. Construa e interprete uma estimativa para o intervalo de confiança de 95% para a inclinação da população correspondente à relação entre preço de venda e número de cômodos.
h. Construa e interprete uma estimativa para o intervalo de confiança de 95% para a inclinação da população correspondente à relação entre preço de venda e bairro.
i. Calcule e interprete o r^2 ajustado.
j. Calcule os coeficientes de determinação parcial e interprete os seus respectivos significados.
k. Que pressuposto você precisa adotar em relação à inclinação do preço de venda com o número de cômodos?
l. Acrescente um termo de interação ao modelo e, no nível de significância de 0,05, determine se ele oferece uma contribuição significativa ao modelo.
m. Com base nos resultados para (f) e (l), qual dos modelos é o mais apropriado? Explique.

14.41 O gerente de marketing de uma grande cadeia de supermercados se viu diante do problema estratégico de determinar o efeito sobre as vendas de ração para animais de estimação exercido pelo espaço disponível em prateleiras, e se o produto estava posicionado na parte da frente (=1) ou no fundo (=0) do corredor. Foram coletados dados de uma amostra aleatória de 12 lojas de igual tamanho. Os resultados estão ilustrados na tabela a seguir (e organizados e contidos no arquivo **Ração**):

Loja	Espaço na Prateleira (Área em Pés)	Localização	Vendas Semanais (Dólares)
1	5	Fundos	160
2	5	Frente	220
3	5	Fundos	140
4	10	Fundos	190
5	10	Fundos	240
6	10	Frente	260
7	15	Fundos	230
8	15	Fundos	270
9	15	Frente	280
10	20	Fundos	260
11	20	Fundos	290
12	20	Frente	310

Para os itens de (a) a (m), não inclua um termo de interação.

a. Expresse a equação da regressão múltipla que possa prever vendas com base no espaço disponível na prateleira e na localização.
b. Interprete os coeficientes da regressão em (a).
c. Faça a previsão para as vendas semanais de ração para animais domésticos, para uma loja com uma área de 8 pés de espaço disponível na prateleira e localização do produto no final do corredor. Construa uma estimativa do intervalo de confiança de 95% e um intervalo de previsão de 95%.
d. Realize uma análise de resíduos nos resultados e determine se os pressupostos da regressão são válidos.
e. Existe uma relação significativa entre vendas e as duas variáveis independentes (espaço disponível na prateleira e localização no corredor), no nível de significância de 0,05?
f. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.
g. Construa e interprete estimativas de intervalos de confiança de 95% para a inclinação da população correspondente à relação entre vendas e espaço disponível na prateleira e entre vendas e localização no corredor.
h. Compare a inclinação em (b) com a inclinação para o modelo de regressão linear simples no Problema 13.4, em Problemas para a Seção 13.2. Explique a diferença nos resultados.
i. Calcule e interprete o significado do coeficiente de determinação múltipla, r^2.
j. Calcule e interprete o r^2 ajustado.
k. Compare r^2 com o valor de r^2 calculado no Problema 13.16(a), em Problemas para a Seção 13.3.
l. Calcule os coeficientes de determinação parcial e interprete os seus respectivos significados.
m. Qual pressuposto sobre a inclinação do espaço na prateleira com relação a vendas você precisa adotar neste problema?
n. Acrescente um termo de interação ao modelo e, no nível de significância de 0,05, determine se ele oferece uma contribuição significativa ao modelo.
o. Com base nos resultados de (f) e (n), qual dos modelos é o mais apropriado? Explique.

14.42 Na engenharia de mineração, perfurações são feitas geralmente através das pedras, com o uso de sondas. À medida que a perfuração vai se tornando mais profunda, outros tubos são acrescentados à sonda, de modo a permitir a continuidade da perfuração. Espera-se que o tempo de perfuração aumente em função da profundidade. Esse aumento no tempo de perfuração pode ser causado por diversos fatores, incluindo o peso da coluna de tubos conectados. O problema estratégico está relacionado ao fato de a perfuração ser mais rápida com o uso de brocas secas ou de brocas úmidas. O uso de brocas secas envolve a injeção de ar comprimido através dos tubos, com o intuito de remover os fragmentos de solo e movimentar a broca. O uso de brocas úmidas envolve a injeção de água, em vez de ar, através da coluna de tubos. Foram coletados dados de uma amostra de 50 orifícios, que contém medições correspondentes ao tempo necessário para perfurar cada 5 pés adicionais (em minutos), à profundidade (em pés) e ao fato de a perfuração ser seca ou úmida. Os dados estão organizados e armazenados em **Sonda**. Desenvolva um modelo para prever o tempo adicional de perfuração, com base na profundidade e no tipo de perfuração (seca ou úmida). Para os itens de (a) a (k), não inclua um termo de interação.

Fonte: *Dados extraídos de R. Penner e D. G. Watts, "Mining Information"*, The American Statistician *45, 1991, pp. 4-9.*

a. Expresse a equação da regressão múltipla.
b. Interprete os coeficientes da regressão em (a).
c. Faça a previsão para o tempo adicional de perfuração em relação a um orifício a uma profundidade de 100 pés. Construa

uma estimativa para o intervalo de confiança de 95% e um intervalo de previsão de 95%.

d. Realize uma análise de resíduos nos resultados e determine se os pressupostos da regressão são válidos.

e. Existe uma relação significativa entre o tempo adicional de perfuração e as duas variáveis independentes (profundidade e tipo de perfuração), no nível de significância de 0,05?

f. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.

g. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população correspondente à relação entre o tempo adicional de perfuração e a profundidade.

h. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população correspondente à relação entre o tempo adicional de perfuração e o tipo de perfuração.

i. Calcule e interprete o r^2 ajustado.

j. Calcule os coeficientes de determinação parcial e interprete os seus respectivos significados.

k. Qual pressuposto você precisa adotar no que diz respeito à inclinação do tempo adicional de perfuração com relação à profundidade?

l. Acrescente um termo de interação ao modelo e, no nível de significância de 0,05, determine se ele oferece uma contribuição significativa ao modelo.

m. Com base nos resultados para (f) e (l), qual modelo é o mais apropriado? Explique.

14.43 O proprietário de uma empresa de mudanças geralmente faz com que o seu gerente mais experiente faça a previsão do número total de horas de trabalho que serão necessárias para realizar uma mudança que está por ocorrer. Esse método se mostrou útil no passado, mas o proprietário tem o objetivo estratégico de desenvolver um método mais preciso para prever a quantidade de horas de trabalho. Em um esforço preliminar para fornecer um método mais preciso, o proprietário decidiu utilizar a quantidade de pés cúbicos a serem transportados na mudança e o fato de existir ou não um elevador no prédio de apartamento como as variáveis independentes e coletou dados correspondentes a 36 mudanças, cuja origem e destino estavam dentro dos limites de Manhattan, em Nova York, e cujo tempo de transporte representou uma parcela insignificante da quantidade de horas trabalhadas. Os dados estão organizados e armazenados no arquivo **Mudança**. Para os itens de (a) a (k), não inclua um termo de interação.

a. Expresse a equação da regressão múltipla para prever o total de horas trabalhadas, utilizando a quantidade de pés cúbicos a serem transportados na mudança e o fato de existir ou não um elevador.

b. Interprete os coeficientes da regressão em (a).

c. Faça a previsão da média aritmética para as horas trabalhadas para uma mudança com 500 pés cúbicos em um prédio de apartamentos que tenha um elevador, e construa uma estimativa para o intervalo de confiança de 95% e um intervalo de previsão de 95%.

d. Realize uma análise de resíduos nos resultados e determine se os pressupostos da regressão são válidos.

e. Existe uma relação significativa entre o total de horas trabalhadas e as duas variáveis independentes (quantidade de pés cúbicos transportados na mudança e o fato de existir ou não um elevador no prédio de apartamentos), no nível de significância de 0,05?

f. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.

g. Construa uma estimativa para o intervalo de confiança de 95% da inclinação da população para a relação entre horas trabalhadas e a quantidade de pés cúbicos transportados.

h. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população correspondente à relação entre horas trabalhadas e a presença de um elevador.

i. Calcule e interprete o r^2 ajustado.

j. Calcule os coeficientes de determinação parcial e interprete os seus respectivos significados.

k. Qual pressuposto você precisa adotar quanto à inclinação das horas trabalhadas com relação à quantidade de pés cúbicos transportados?

l. Acrescente um termo de interação ao modelo e, no nível de significância de 0,05, determine se ele oferece uma contribuição significativa ao modelo.

m. Com base nos resultados para (f) e (l), qual modelo é o mais apropriado? Explique.

AUTO-AVALIAÇÃO **14.44** No Problema 14.4, em Problemas para a Seção 14.1, você utilizou vendas e pedidos para prever o custo de distribuição (veja o arquivo **CustoDepósito**). Desenvolva um modelo de regressão para prever o custo de distribuição que inclua vendas, pedidos e a interação entre vendas e pedidos.

a. No nível de significância de 0,05, existem evidências de que o termo de interação oferece uma contribuição significativa ao modelo?

b. Qual modelo de regressão é o mais apropriado: o modelo utilizado no item (a) ou aquele utilizado no Problema 14.4? Explique.

14.45 A Zagat's publica avaliações de restaurantes de várias localidades dos Estados Unidos. O arquivo **Restaurantes** contém a classificação da Zagat para comida, decoração, serviço e preço por pessoa, para uma amostra de 50 restaurantes localizados em uma área urbana e 50 restaurantes localizados em uma área do subúrbio dos Estados Unidos. Desenvolva um modelo de regressão para prever o custo, por pessoa, com base na variável que representa a soma das classificações para comida, decoração e serviço, e uma variável binária (*dummy*) que represente a localização (urbana *versus* suburbana). Para os itens de (a) a (m), não inclua um termo de interação.

Fonte: *Extraído de* Zagat Survey 2008 New York City Restaurants *e* Zagat Survey 2007-2008 Long Island Restaurants.

a. Expresse a equação da regressão múltipla.

b. Interprete os coeficientes da regressão em (a).

c. Faça a previsão do custo para um restaurante com uma soma de classificações totalizando 60 que esteja localizado na área urbana e construa uma estimativa para o intervalo de confiança de 95% e um intervalo de previsão de 95%.

d. Realize uma análise de resíduos nos resultados e determine se os pressupostos da regressão foram satisfeitos.

e. Existe uma relação significativa entre preço e as duas variáveis independentes (soma das classificações e localização), no nível de significância de 0,05?

f. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.

g. Construa uma estimativa para o intervalo de confiança de 95% da inclinação da população para a relação entre custo e soma das classificações.

h. Compare a inclinação em (b) com a inclinação para o modelo de regressão linear simples do Problema 13.90 em Problemas de Revisão do Capítulo, no Capítulo 13. Explique a diferença nos resultados.

i. Calcule e interprete o significado do coeficiente de determinação múltipla.

j. Calcule e interprete o r^2 ajustado.

k. Compare r^2 com o valor de r^2 calculado no Problema 13.90(d), em Problemas de Revisão do Capítulo, no final do Capítulo 13.

l. Calcule os coeficientes de determinação parcial e interprete os seus respectivos significados.

m. Qual pressuposto sobre a inclinação do custo com a soma das classificações você precisa adotar neste problema?

n. Acrescente um termo de interação ao modelo e, no nível de significância de 0,05, determine se ele oferece uma contribuição significativa ao modelo.

o. Com base nos resultados para (f) e (n), qual dos modelos é o mais apropriado? Explique.

14.46 No Problema 14.6, em Problemas para a Seção 14.1, você utilizou a propaganda no rádio e a propaganda em jornais para prever vendas (dados no arquivo **Propaganda**). Desenvolva um modelo de regressão para prever vendas que inclua a propaganda em rádio, a propaganda em jornais e a interação entre propaganda em rádio e propaganda em jornais.

a. No nível de significância de 0,05, existem evidências de que o termo de interação oferece uma contribuição significativa ao modelo?

b. Qual modelo de regressão é o mais apropriado: o modelo utilizado neste problema ou aquele utilizado no Problema 14.6? Explique.

14.47 No Problema 14.5, em Problemas para a Seção 14.1, foram utilizados a potência, em cavalos-vapor, e o peso para prever o consumo de combustível, em milhas por galão (dados no arquivo **Auto**). Desenvolva um modelo de regressão que inclua a potência, em cavalos-vapor, o peso e a interação entre potência, em cavalos-vapor, e peso para prever a milhagem, em milhas por galão.

a. No nível de significância de 0,05, existem evidências de que o termo de interação oferece uma contribuição significativa ao modelo?

b. Qual modelo de regressão é o mais apropriado: o modelo utilizado neste problema ou aquele utilizado no Problema 14.5? Explique.

14.48 No Problema 14.7, em Problemas para a Seção 14.1, você utilizou o total da equipe presente e as horas remotas para prever as horas de sobreaviso (veja o arquivo **Sobreaviso**). Desenvolva um modelo de regressão para prever horas de sobreaviso que inclua o total da equipe presente, horas remotas e a interação entre total da equipe presente e horas remotas.

a. No nível de significância de 0,05, existem evidências de que o termo de interação oferece uma contribuição significativa ao modelo?

b. Qual modelo de regressão é o mais apropriado: o modelo utilizado neste problema ou aquele utilizado no Problema 14.7? Explique.

14.49 A diretora de um programa de treinamento de uma grande companhia de seguros tem o objetivo estratégico de determinar qual método de treinamento é o melhor para treinar seus corretores. Os três métodos a serem avaliados são: tradicional, baseado em CD-ROM e baseado na Internet. Os 30 treinandos são divididos em três grupos de 10 designados aleatoriamente. Antes do início do treinamento, é aplicada a cada um dos treinandos uma prova de proficiência, que mede competências em matemática e informática. Ao final do treinamento, todos os alunos fazem a mesma prova de encerramento do treinamento. Os resultados estão organizados e armazenados em **Corretores**.

Desenvolva um modelo de regressão múltipla para prever o resultado da prova de encerramento do treinamento, com base no resultado da prova de proficiência e no método de treinamento utilizado. Para os itens de (a) a (k), não inclua um termo de interação.

a. Expresse a equação da regressão múltipla.

b. Interprete os coeficientes da regressão em (a).

c. Faça a previsão para o resultado da prova de final de treinamento para um aluno com um resultado de prova de proficiência igual a 100 e cujo treinamento tenha sido baseado na Internet.

d. Realize uma análise de resíduos nos resultados e determine se os pressupostos da regressão são válidos.

e. Existe uma relação significativa entre o resultado da prova de final de treinamento e as variáveis independentes (resultado da prova de proficiência e método de treinamento), no nível de significância de 0,05?

f. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.

g. Construa e interprete uma estimativa do intervalo de confiança de 95% para a inclinação da população correspondente à relação entre o resultado da prova de final de treinamento e o resultado da prova de proficiência.

h. Construa e interprete uma estimativa do intervalo de confiança de 95% para a inclinação da população correspondente à relação entre o resultado da prova de final de treinamento e o tipo de método de treinamento.

i. Calcule e interprete o r^2 ajustado.

j. Calcule os coeficientes de determinação parcial e interprete os seus respectivos significados.

k. Qual pressuposto sobre a inclinação do resultado da prova de proficiência com relação ao resultado da prova de final de treinamento você precisa adotar neste problema?

l. Acrescente termos de interação ao modelo e, no nível de significância de 0,05, determine se algum um dos termos oferece uma contribuição significativa ao modelo.

m. Com base nos resultados para (f) e (l), qual dos modelos é o mais apropriado? Explique.

@ OmniFoods Revisitada

No cenário Utilizando a Estatística, você era o gerente de marketing da OmniFoods, uma grande empresa de produtos alimentícios que está planejando o lançamento, em âmbito nacional, de uma nova barra energética, a OmniPower. Você precisava determinar o efeito que o preço e as promoções internas da loja teriam sobre as vendas de OmniPower a fim de desenvolver uma estratégia de marketing efetiva. Foi selecionada uma amostra de 34 lojas em uma cadeia de supermercados para fins de um estudo de teste de mercado. As lojas cobravam entre 59 e 99 centavos de dólar por barra, e foi concedida a elas uma verba da ordem de $200 a $600 para promoções internas nas lojas.

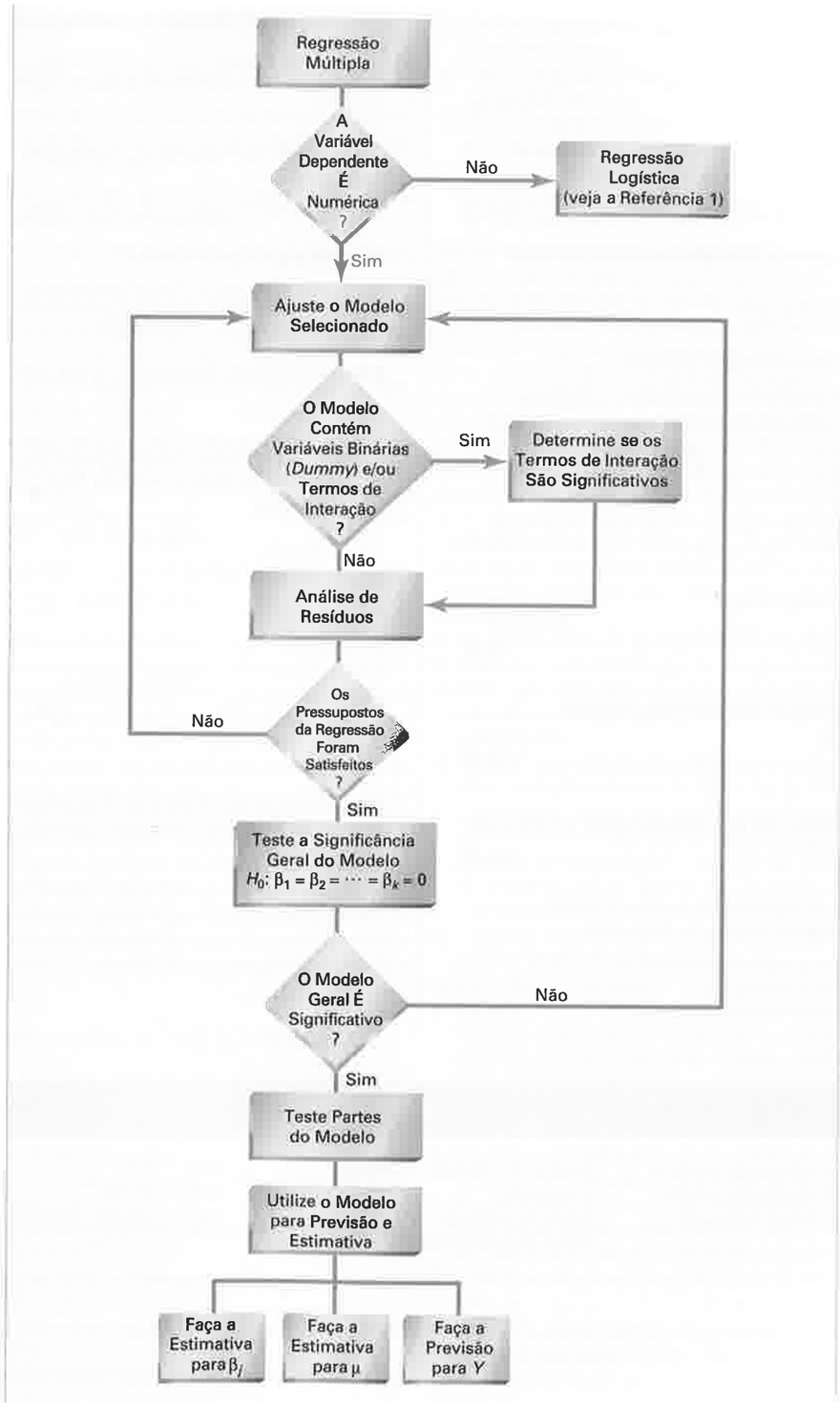
Ao final do estudo de teste de mercado, com duração de um mês, você realizou uma análise de regressão múltipla nos dados. Duas variáveis independentes foram consideradas: o preço de uma barra OmniPower e o orçamento mensal para gastos com promoções internas na loja. A variável dependente era o número de barras OmniPower vendidas em um mês. O coeficiente de determinação indicou que 75,8% da variação nas vendas era explicada pelo conhecimento do preço cobrado e pelo montante gasto com promoções internas da loja. O modelo indicou que pode ser estimado que as vendas previstas de OmniPower decresçam em 532 barras por mês para cada 10 centavos de dólar de aumento no preço e que pode ser estimado que as vendas previstas de OmniPower cresçam em 361 barras por mês para cada $100 gastos com promoções.

Depois de estudar os efeitos relativos de preço e promoção, a OmniFoods precisa estabelecer padrões para preço e promoções para um lançamento em âmbito nacional (obviamente, preços mais baixos e verbas maiores para promoções acarretam maior volume de vendas, embora isso ocorra à custa de uma margem de lucro mais baixa). Você determinou que, se as lojas gastarem $400 por mês com promoções internas na loja e cobrarem 79 centavos por barra, a estimativa para o intervalo de confiança de 95% para a média aritmética das vendas mensais é de 2.854 a 3.303 barras. A OmniFoods pode multiplicar os limites inferior e superior desse intervalo de confiança pelo número de lojas incluídas em um lançamento de âmbito nacional para estimar o total de vendas mensais. Por exemplo, se 1.000 lojas fizerem parte do lançamento nacional, então o total de vendas mensais deveria estar entre 2.854 milhões e 3,303 milhões de barras.

RESUMO

Neste capítulo, você aprendeu como modelos de regressão múltipla permitem que você utilize duas ou mais variáveis independentes para prever o valor de uma variável dependente. Você aprendeu, também, a incluir variáveis categóricas independentes e termos de interação em modelos de regressão. A Figura 14.13 apresenta um roteiro do capítulo.

FIGURA 14.13
Roteiro para a regressão múltipla



EQUAÇÕES-CHAVE

Modelo de Regressão Múltipla com k Variáveis Independentes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (14.1)$$

Modelo de Regressão Múltipla com Duas Variáveis Independentes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (14.2)$$

Equação da Regressão Múltipla com Duas Variáveis Independentes

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} \quad (14.3)$$

Coeficiente de Determinação Múltipla

$$r^2 = \frac{\text{Soma dos quadrados da regressão}}{\text{Soma total dos quadrados}} = \frac{SQReg}{STQ} \quad (14.4)$$

r^2 Ajustado

$$r^2_{aj} = 1 - \left[(1 - r^2)\frac{n-1}{n-k-1}\right] \quad (14.5)$$

Estatística do Teste F Geral

$$F_{ESTAT} = \frac{MQReg}{MQR} \quad (14.6)$$

Testando a Inclinação na Regressão Múltipla

$$t_{ESTAT} = \frac{b_j - \beta_j}{S_{b_j}} \quad (14.7)$$

Estimativa do Intervalo de Confiança para a Inclinação

$$b_j \pm t_{\alpha/2} S_{b_j} \quad (14.8)$$

Determinando a Contribuição de uma Variável Independente para o Modelo de Regressão

$$SQReg(X_j \mid (Todas \text{ as variáveis } X \text{ exceto } j) = SQReg(Todas \text{ as variáveis } X) - SQReg(Todas \text{ as variáveis } X \text{ exceto } j) \quad (14.9)$$

Contribuição da Variável X_1, Sabendo-se que X_2 Foi Incluída

$$SQReg(X_1 \mid X_2) = SQReg(X_1 \text{ e } X_2) - SQReg(X_2) \quad (14.10a)$$

Contribuição da Variável X_2, Sabendo-se que X_1 Foi Incluída

$$SQReg(X_2 \mid X_1) = SQReg(X_1 \text{ e } X_2) - SQReg(X_1) \quad (14.10b)$$

Estatística do Teste F Parcial

$$F_{ESTAT} = \frac{SQReg(X_j \mid (Todas \text{ as variáveis } X \text{ exceto } j)}{MQR} \quad (14.11)$$

Relação entre uma Estatística t e uma Estatística F

$$t^2_{ESTAT} = F_{ESTAT} \quad (14.12)$$

Coeficientes de Determinação Parcial para um Modelo de Regressão Múltipla Contendo Duas Variáveis Independentes

$$r^2_{Y1.2} = \frac{SQReg(X_1 \mid X_2)}{STQ - SQReg(X_1 \text{ e } X_2) + SQReg(X_1 \mid X_2)} \quad (14.13a)$$

e

$$r^2_{Y2.1} = \frac{SQReg(X_2 \mid X_1)}{STQ - SQReg(X_1 \text{ e } X_2) + SQReg(X_2 \mid X_1)} \quad (14.13b)$$

Coeficientes de Determinação Parcial para um Modelo de Regressão Múltipla Contendo k Variáveis Independentes

$$r^2_{Y_j(Todas \text{ as variáveis } exceto\ j)} = \frac{SQReg(X_j \mid \text{Todas as variáveis } X \text{ exceto } j)}{STQ - SQReg(\text{Todas as variáveis } X) + SQReg(X_j \mid \text{Todas as variáveis } X \text{ exceto } j)} \quad (14.14)$$

TERMOS-CHAVE

coeficiente de determinação múltipla
coeficiente de determinação parcial
coeficientes líquidos de regressão
interação
modelo de regressão múltipla
r^2 ajustado
termo de interação
termo de multiplicação
teste F geral
teste F parcial
variável binária (*dummy*)

PROBLEMAS DE REVISÃO DO CAPÍTULO 14

AVALIANDO SEU ENTENDIMENTO

14.50 Como a interpretação dos coeficientes de regressão difere entre a regressão múltipla e a regressão linear simples?

14.51 Como o teste para a significância do modelo de regressão múltipla completo difere do teste para a contribuição de cada uma das variáveis independentes?

14.52 Como os coeficientes de determinação parcial diferem do coeficiente de determinação múltipla?

14.53 Por que e como são utilizadas as variáveis binárias (*dummy*)?

14.54 De que modo você consegue avaliar se a inclinação da variável dependente com uma variável independente é a mesma para cada um dos níveis da variável binária (*dummy*)?

14.55 Sob quais circunstâncias você inclui um termo de interação em um modelo de regressão?

14.56 Quando uma variável binária (*dummy*) é incluída em um modelo de regressão que tem uma variável numérica independente, qual pressuposto você precisa adotar no que diz respeito à inclinação entre a variável de resposta, Y, e a variável independente numérica, X?

APLICANDO OS CONCEITOS

14.57 Um aumento na satisfação do consumidor geralmente resulta em um comportamento de maior tendência para compras. Para muitos produtos, existe mais de um indicador da satisfação do consumidor. Em muitos desses casos, o comportamento de tendência para compras pode crescer drasticamente em razão de um crescimento em qualquer um dos indicadores de satisfação do consumidor, não necessariamente todos eles ao mesmo tempo. Gunst e Barry ("One Way to Moderate Ceiling Effects," *Quality Progress*, outubro de 2003, pp. 83-85) consideram um produto com dois indicadores de satisfação, X_1 e X_2, que variam desde o menor nível de satisfação, 1, até o mais alto nível de satisfação, 7. A variável dependente, Y, é um indicador do comportamento de tendência para compras, com o valor mais alto gerando a maior quantidade de vendas. É apresentada a equação da regressão a seguir:

$$\hat{Y}_i = -3{,}888 + 1{,}449 X_{1i} + 1{,}462 X_{2i} - 0{,}190 X_{1i} X_{2i}$$

Suponha que X_1 seja a qualidade percebida do produto e X_2 seja o valor percebido para o produto. (Observação: Se o consumidor imagina que o produto está acima do preço, ele percebe o produto como sendo de baixo valor, e vice-versa.)

a. Qual é o comportamento previsto para tendência para compras quando $X_1 = 2$ e $X_2 = 2$?
b. Qual é o comportamento previsto para tendência para compras quando $X_1 = 2$ e $X_2 = 7$?
c. Qual é o comportamento previsto para tendência para compras quando $X_1 = 7$ e $X_2 = 2$?
d. Qual é o comportamento previsto para tendência para compras quando $X_1 = 7$ e $X_2 = 7$?
e. Qual é a equação da regressão quando $X_2 = 2$? Qual é, agora, a inclinação para X_1?
f. Qual é a equação da regressão quando $X_2 = 7$? Qual é, agora, a inclinação para X_1?
g. Qual é a equação da regressão quando $X_1 = 2$? Qual é, agora, a inclinação para X_2?
h. Qual é a equação da regressão quando $X_1 = 7$? Qual é, agora, a inclinação para X_2?
i. Discuta as implicações dos itens (a) a (h), dentro do contexto de aumentar as vendas para esse produto, com dois indicadores de satisfação do consumidor.

14.58 O proprietário de uma empresa de mudanças geralmente faz com que seu gerente mais experiente faça a previsão do número total de horas de trabalho que serão necessárias para realizar uma determinada mudança que está por ocorrer. Esse método se mostrou útil no passado, porém o proprietário tem como objetivo estratégico da empresa desenvolver um método mais preciso para prever a quantidade de horas de trabalho. Em um estudo preliminar para oferecer um método mais preciso, o proprietário decidiu utilizar a quantidade de pés cúbicos a serem transportados na mudança e o número de peças de mobiliário de grande porte como as variáveis independentes, e coletou dados de 36 mudanças cuja origem e destino estavam dentro dos limites de Manhattan, em Nova York, e cujo tempo de transporte representava uma parcela insignificante da quantidade de horas trabalhadas. Os dados estão organizados e armazenados no arquivo **Mudança**.

a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações nessa equação.
c. Faça a previsão para o total de horas trabalhadas, para 500 pés cúbicos transportados, com duas peças de mobiliário de grande porte.
d. Realize uma análise de resíduos em seus resultados e determine se os pressupostos da regressão são válidos.
e. Determine se existe uma relação significativa entre as horas trabalhadas e as duas variáveis independentes (quantidade de pés cúbicos transportados e número de peças de mobiliário de grande porte), no nível de significância de 0,05.
f. Determine o valor-p em (e) e interprete o seu significado.
g. Interprete o significado do coeficiente de determinação múltipla no âmbito deste problema.
h. Determine o r^2 ajustado.
i. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa para o modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.
j. Determine os valores-p em (i) e interprete os seus respectivos significados.
k. Construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população entre horas trabalhadas e a quantidade de pés cúbicos transportados. Como a interpretação da inclinação, neste problema, difere da interpretação do Problema 13.44, em Problemas para a Seção 13.7?
l. Calcule e interprete os coeficientes de determinação parcial.

14.59 O basquete profissional se tornar verdadeiramente um esporte que gera interesse entre fãs em todo o mundo. Um número cada vez maior de jogadores vem de fora dos Estados Unidos para jogar na National Basketball Association (NBA).

Você deseja desenvolver um modelo de regressão para prever o número de vitórias conquistadas por cada time da NBA, com base na percentagem de cestas de campo (arremessos feitos) para um time e para o time adversário. Os dados estão armazenados em **NBA2009**.

a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações nessa equação.
c. Faça a previsão do número de vitórias para um time que tenha uma percentagem de cestas de campo igual a 45% e uma percentagem de cestas de campo do time oponente igual a 44%.
d. Realize uma análise de resíduos em seus resultados e determine se os pressupostos da regressão são válidos.
e. Existe uma relação significativa entre o número de vitórias e as duas variáveis independentes (percentagem de cestas de campo para o time e para o time adversário), no nível de significância de 0,05?
f. Determine o valor-p em (e) e interprete o seu significado.
g. Interprete o significado do coeficiente de determinação múltipla neste problema.
h. Determine o r^2 ajustado.
i. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.
j. Determine os valores-p em (i) e interprete os seus respectivos significados.
k. Calcule e interprete os coeficientes de determinação parcial.
l. Calcule e interprete os coeficientes de determinação parcial.

14.60 Foi selecionada uma amostra de 30 casas unifamiliares recentemente vendidas em uma pequena cidade. Desenvolva um modelo para prever o preço de venda (em milhares de dólares), utilizando o valor de avaliação (em milhares de dólares), bem como o período de tempo (em meses desde a reavaliação). As casas na cidade haviam sido reavaliadas em seu valor pleno um ano antes do estudo. Os resultados estão armazenados no arquivo **Casa1**:

a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações nessa equação.
c. Faça a previsão para o preço de venda de uma casa que tenha um valor de avaliação de US$170.000 e tenha sido vendida no período de tempo 12.
d. Realize uma análise de resíduos em seus resultados e determine se os pressupostos da regressão são válidos.
e. Determine se existe uma relação significativa entre o preço de venda e as duas variáveis independentes (valor de avaliação e período de tempo), no nível de significância de 0,05.
f. Determine o valor-p em (e) interprete o seu significado.
g. Interprete o significado do coeficiente de determinação múltipla no contexto deste problema.
h. Determine o r^2 ajustado.
i. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.
j. Determine os valores-p em (i) e interprete os seus respectivos significados.
k. Construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população entre preço de venda e valor de avaliação. Como a interpretação da inclinação, no presente problema, difere da interpretação para o Problema 13.76 nos Problemas de Revisão do Capítulo 13?
l. Calcule e interprete os coeficientes de determinação parcial.

14.61 Medir a altura de uma árvore do tipo sequoia da Califórnia é um empreendimento bastante difícil, uma vez que essas árvores alcança alturas superiores a 300 pés. As pessoas familiarizadas com essas árvores entendem que a altura de uma sequoia da Califórnia está relacionada a outras características da árvore, incluindo o diâmetro da árvore na altura do peito de uma pessoa e a espessura do córtex da árvore. O arquivo **Sequoia** contém a altura, o diâmetro na altura do peito de uma pessoa e a espessura do córtex, de uma amostra de 21 árvores do tipo sequoia da Califórnia.

a. Expresse a equação da regressão múltipla que preveja a altura de uma árvore, com base no diâmetro dessa árvore na altura do peito de uma pessoa e na espessura do córtex.
b. Interprete o significado das inclinações nessa equação.
c. Faça a previsão da altura para uma árvore que apresente um diâmetro de 25 polegadas na altura do peito de uma pessoa e uma espessura de córtex de 2 polegadas.
d. Interprete o significado do coeficiente de determinação múltipla no contexto deste problema.
e. Realize uma análise de resíduos nos resultados e determine se os pressupostos da regressão são válidos.
f. Determine se existe uma relação significativa entre a altura das sequoias e as duas variáveis independentes (diâmetro da árvore na altura do peito e espessura do córtex), no nível de significância de 0,05.
g. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população entre a altura das sequoias e o diâmetro na altura do peito de uma pessoa e entre a altura das sequoias e a espessura do córtex.
h. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Indique as variáveis independentes que devem ser incluídas nesse modelo.
i. Construa uma estimativa do intervalo de confiança de 95% para a média aritmética da altura das árvores que tenham um diâmetro de 25 polegadas e uma espessura de córtex de 2 polegadas, juntamente com um intervalo de previsão para uma árvore individual.
j. Calcule e interprete os coeficientes de determinação parcial.

14.62 Desenvolva um modelo para prever o valor de avaliação (em milhares de dólares), utilizando o tamanho de imóveis (em milhares de pés quadrados) e a idade ou tempo de construção dessas casas (em anos), com base na tabela a seguir (dados armazenados no arquivo **Casa2**):

a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações nessa equação.
c. Faça a previsão do valor de avaliação para uma casa que tenha como tamanho 1.750 pés quadrados e 10 anos de construída.
d. Realize uma análise de resíduos nos resultados e determine se os pressupostos da regressão são válidos.
e. Determine se existe uma relação significativa entre o valor de avaliação e as duas variáveis independentes (tamanho e tempo de construção), no nível de significância de 0,05.
f. Determine o valor-p em (e) e interprete o seu respectivo significado.
g. Interprete o significado do coeficiente de determinação múltipla no contexto deste problema.

Imóvel	Valor de Avaliação ($1.000)	Tamanho do Imóvel (Milhares de Pés Quadrados)	Idade (Anos)
1	184,4	2,00	3,42
2	177,4	1,71	11,50
3	175,7	1,45	8,33
4	185,9	1,76	0,00
5	179,1	1,93	7,42
6	170,4	1,20	32,00
7	175,8	1,55	16,00
8	185,9	1,93	2,00
9	178,5	1,59	1,75
10	179,2	1,50	2,75
11	186,7	1,90	0,00
12	179,3	1,39	0,00
13	174,5	1,54	12,58
14	183,8	1,89	2,75
15	176,8	1,59	7,17

h. Determine o r^2 ajustado.
i. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.
j. Determine os valores-p em (i) e interprete seus respectivos significados.
k. Construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população entre valor de avaliação e tamanho. Como a interpretação da inclinação, nesse caso, difere daquela do Problema 13.77, nos Problemas de Revisão do Capítulo 13?
l. Calcule e interprete os coeficientes de determinação parcial.
m. A assessoria da imobiliária declarou publicamente que a idade do imóvel não exerce nenhum tipo de influência sobre o valor de avaliação. Com base em suas respostas para os itens de (a) até (l), você concorda com essa afirmativa? Explique.

14.63 Crazy Dave, um conhecido comentarista de beisebol, deseja determinar quais variáveis são importantes na previsão de vitórias para um time em uma determinada temporada. Ele coletou dados relacionados a vitórias, à média conquistada de voltas (ERA – *earned run average*) e ao número de voltas percorridas para a temporada de 2008 (dados aramazenados no arquivo **BB2008**). Desenvolva um modelo para prever o número de vitórias, com base na média conquistada de voltas (ERA) e nas voltas percorridas.

a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações nessa equação.
c. Faça a previsão da média aritmética do número de vitórias para um time com uma ERA igual a 4,50 e que tenha pontuado 750 voltas.
d. Realize uma análise de resíduos nos resultados e determine se os pressupostos da regressão são válidos.
e. Existe uma relação significativa entre o número de vitórias e as duas variáveis independentes (ERA e quantidade de voltas percorridas), no nível de significância de 0,05?
f. Determine o valor-p em (e) e interprete o seu significado.
g. Interprete o significado do coeficiente de determinação múltipla no contexto deste problema.
h. Determine o r^2 ajustado.
i. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.
j. Determine os valores-p em (i) e interprete seus respectivos significados.
k. Construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população entre vitórias e ERA.
l. Calcule os coeficientes de determinação parcial.
m. O que é mais importante para a previsão de vitórias – arremessos, medidos com base na ERA, ou ataques, medidos com base nas voltas pontuadas? Explique.

14.64 Fazendo referência ao Problema 14.63, suponha que, além de utilizar a ERA para prever o número de vitórias, Crazy Dave deseje incluir a Liga (Americana *versus* Nacional) como uma variável independente. Desenvolva um modelo para prever vitórias com base na ERA e na Liga. Para os itens de (a) a (k), não inclua um termo de interação.

a. Expresse a equação da regressão múltipla.
b. Interprete o significado das inclinações neste problema.
c. Faça uma previsão da média aritmética do número de vitórias para um time com uma ERA igual a 4,50 na Liga Americana. Construa uma estimativa para o intervalo de confiança de 95% para todos os times e um intervalo de previsão de 95% para um time individual.
d. Realize uma análise de resíduos nos resultados e determine se os pressupostos da regressão são válidos.
e. Existe uma relação significativa entre vitórias e as duas variáveis independentes (ERA e Liga), no nível de significância de 0,05?
f. No nível de significância de 0,05, determine se cada uma das variáveis independentes oferece uma contribuição significativa ao modelo de regressão. Indique o modelo de regressão mais apropriado para esse conjunto de dados.
g. Construa uma estimativa para o intervalo de confiança de 95% da inclinação da população para a relação entre vitórias e ERA.
h. Construa uma estimativa para o intervalo de confiança de 95% da inclinação da população para a relação entre vitórias e a Liga.
i. Calcule e interprete o r^2 ajustado.
j. Calcule e interprete os coeficientes de determinação parcial.
k. Que pressuposto você precisa adotar em relação à inclinação das vitórias com ERA?
l. Acrescente um termo de interação ao modelo e, no nível de significância de 0,05, determine se esse termo oferece uma contribuição significativa ao modelo.
m. Com base nos resultados para os itens (f) e (l), qual modelo é mais apropriado? Explique.

14.65 Você é um corretor imobiliário que deseja comparar valores de propriedades em Glen Cove e Roslyn (que estão localizadas a aproximadamente 8 milhas de distância uma da outra). Para que possa fazer isso, você analisará os dados em **GCRoslyn**, um arquivo que inclui amostras de casas em Glen Cove e Roslyn. Não deixando de incluir a variável binária (*dummy*) para localização (Glen Cove ou Roslyn), desenvolva um modelo de regressão para prever o valor de avaliação, com base na área do terreno de uma propriedade, a idade do imóvel e a localização. Não deixe de determinar se precisam ser incluídos no modelo quaisquer termos de interação.

14.66 Um artigo recente discute o processo de deposição do metal no qual uma peça metálica é colocada em um banho de ácido e uma liga metálica é disposta sobre ela. O objetivo do serviço dos engenheiros que trabalham no processo era reduzir a variação na espessura da cobertura da liga metálica. Para começar, a temperatura e a pressão no tanque que contém o banho de ácido devem ficar como variáveis independentes. Os dados são coletados de 50 amostras. Os resultados estão organizados e registrados em **Espessura** (dados extraídos de J. Conklin, "It's a Marathon, Not a Sprint", *Quality Progress*, June 2009, PP. 46-49).

Desenvolva um modelo de regressão múltipla que use a temperatura e a pressão do tanque que contém o banho de ácido para prever a espessura da cobertura da liga metálica. Certifique-se de fazer uma análise residual. O artigo sugere que existe uma interação significativa entre a pressão e a temperatura no tanque. Você concorda?

ADMINISTRANDO O *SPRINGVILLE HERALD*

Em seu estudo continuado sobre o processo de solicitação de assinaturas com entrega domiciliar, uma equipe do departamento de marketing deseja testar os efeitos de dois tipos de apresentações estruturadas de vendas (pessoal formal e pessoal informal) e o número de horas gastas com telemarketing com relação ao número de novas assinaturas. A equipe registrou esses dados no arquivo **SH14** ao longo das últimas 24 semanas.

Analise esses dados e desenvolva um modelo de regressão múltipla para prever o número de novas assinaturas, por um período de uma semana, com base no número de horas gastas com telemarketing e no tipo de apresentação de vendas. Redija um relatório, fornecendo descobertas detalhadas em relação ao modelo de regressão utilizado.

CASO DE INTERNET

Aplique os seus conhecimentos sobre modelos de regressão múltipla neste Caso de Internet, que estende o cenário Utilizando a Estatística deste capítulo, que trata da OmniFoods.

Para garantir um teste de mercado bem-sucedido para suas barras energéticas OmniPower, o departamento de marketing da OmniFoods contratou junto a uma empresa especialista em escolha de local de exposição para produtos, a In-Store Placements Group (ISPG), uma consultoria de estudo de mercado. A ISPG irá trabalhar junto à cadeia de supermercados que está conduzindo o estudo de teste de mercado. Utilizando a mesma amostra de 34 lojas de supermercado utilizada no estudo de teste de mercado, a ISPG afirma que a escolha da localização na prateleira e a presença de pessoas dentro da loja distribuindo cupons de desconto para a OmniPower farão crescer as vendas das barras energéticas.

Utilizando seu navegador para a Web, abra o arquivo na pasta Caso de Internet para o Capítulo 14, no site da LTC Editora para este livro, ou abra diretamente o arquivo **Omni_ISPGMemo.htm** caso já tenha baixado para seu computador os arquivos dos Casos de Internet, para examinar as declarações da ISPG e os dados que as respaldam. Depois, responda ao seguinte:

1. Os dados que respaldam as afirmativas são coerentes com as declarações da ISPG? Realize uma análise estatística apropriada para confirmar (ou negar) a relação declarada entre vendas e as duas variáveis independentes que tratam da localização do produto na prateleira e da distribuição de cupons de desconto para a OmniPower dentro da loja.
2. Se estivesse prestando consultoria à OmniFoods, você recomendaria uma localização específica na prateleira e pessoas dentro da loja distribuindo cupons de desconto para a venda de barras OmniPower?
3. Que tipo de dados adicionais você aconselharia que fossem coletados, para determinar a efetividade das técnicas de promoção de vendas utilizadas pela ISPG?

REFERÊNCIAS

1. Hosmer, D. W., and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. (New York: Wiley, 2001).
2. Kutner, M., C. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. (New York: McGraw-Hill/Irwin, 2005).
3. *Microsoft Excel 2007* (Redmond, WA: Microsoft Corp., 2007).