13 Regressão Linear Simples

Objetivos do Aprendizado

Neste capítulo, você aprenderá:

- A utilizar a análise da regressão para prever o valor de uma variável dependente com base em uma variável independente
- O significado dos coeficientes de regressão b_0 e b_1
- A avaliar o pressuposto da análise da regressão e saber o que fazer caso os pressupostos sejam violados
- A fazer inferências sobre a inclinação e o coeficiente de inclinação
- A estimar a média aritmética dos valores e prever valores individuais

UTILIZANDO A ESTATÍSTICA

@ Sunflowers Roupas

As vendas da Sunflowers Roupas, uma cadeia de lojas de roupas de primeira linha, vêm aumentando nos últimos 12 anos à medida que a cadeia expande o número de lojas abertas. Até agora, os gerentes da Sunflowers têm selecionado locais para instalação de lojas com base em fatores subjetivos, tais como a disponibilidade de um bom contrato de arrendamento ou a percepção de que um determinado local parece ideal para uma loja de roupas. No papel de novo diretor de planejamento, você precisa desenvolver uma abordagem sistemática que leve à tomada de decisões mais eficazes durante o processo de seleção de locais para instalação. Como ponto de partida, você acredita que o tamanho da loja contribui significativamente para as vendas e deseja utilizar essa relação no processo de tomada de decisão. De que modo você pode utilizar a estatística, de maneira a ser capaz de fazer prognósticos sobre as vendas anuais de uma loja sugerida, com base no seu respectivo tamanho?

Neste capítulo e nos próximos dois capítulos, você aprende como a **análise da regressão** possibilita que você desenvolva um modelo para prever os valores de uma variável numérica com base no valor de outras variáveis.

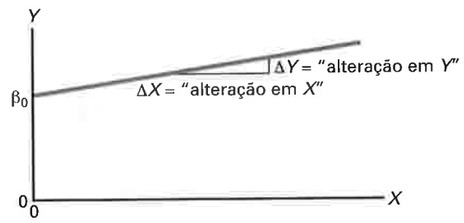
Na análise da regressão, a variável que você deseja prever é chamada de **variável dependente**. As variáveis utilizadas para fazer a previsão são chamadas de **variáveis independentes**. Além de prever valores para a variável dependente, a análise da regressão permite também que você identifique o tipo de relação matemática que existe entre uma variável dependente e uma variável independente; quantifique o efeito que alterações na variável independente exercem sobre a variável dependente; e identifique observações incomuns. Por exemplo, no papel de diretor de planejamento, é possível que você queira prever as vendas de uma determinada loja da Sunflowers com base no tamanho dessa loja. Outros exemplos incluem a previsão do aluguel mensal de um apartamento com base em seu respectivo tamanho e das vendas mensais de um determinado produto em um supermercado com base na quantidade de espaço na prateleira de que o produto dispõe.

Este capítulo discorre sobre a **regressão linear simples**, na qual uma *única* variável independente numérica, X, é utilizada para prever a variável dependente numérica, Y, como é o caso na utilização das dimensões físicas da loja para prever as vendas anuais dessa loja. Os Capítulos 14 e 15 discorrem sobre *modelos de regressão múltipla*, que utilizam *diversas* variáveis independentes para prever uma única variável dependente numérica, Y. Por exemplo, você pode utilizar o montante de gastos com propaganda, o preço e a quantidade de espaço disponível na prateleira para um determinado produto, com o objetivo de prever as suas respectivas vendas mensais.

13.1 Tipos de Modelos de Regressão

Na Seção 2.7, você utilizou um **gráfico de dispersão** (também chamado de **diagrama de dispersão**) para examinar a relação entre uma variável X no eixo horizontal e uma variável Y no eixo vertical. A natureza da relação entre duas variáveis pode assumir inúmeras formas, abrangendo desde funções matemáticas simples até funções matemáticas extremamente complicadas. A relação mais simples consiste em uma relação em forma de linha reta, ou **relação linear**. A Figura 13.1 ilustra uma relação em linha reta.

FIGURA 13.1
Uma relação em linha reta



A Equação (13.1) representa o modelo da linha reta (linear).

MODELO DE REGRESSÃO LINEAR SIMPLES

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \textbf{(13.1)}$$

em que

β_0 = intercepto de Y para a população

β_1 = inclinação da população

ε_i = erro aleatório em Y para a observação i

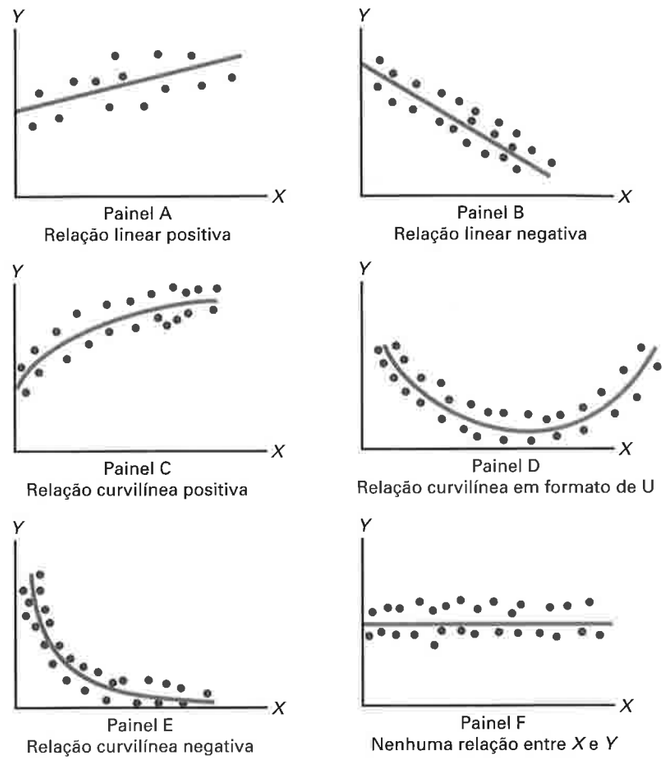
Y_i = variável dependente (algumas vezes denominada **variável de resposta**) para a observação i

X_i = variável independente (algumas vezes denominada **variável explanatória**) para a observação i

A parcela $Y_i = \beta_0 + \beta_1 X_i$ para o modelo de regressão linear simples, expressa na Equação (13.1), corresponde a uma linha reta. A **inclinação** da linha, β_1, representa a variação esperada em Y por uma unidade de variação em X. Representa a média aritmética da quantidade em que Y varia (positiva ou negativamente) para uma unidade de alteração em X. O **intercepto de Y**, β_0, representa a média aritmética do valor de Y quando X é igual a 0. O último componente do modelo, ε_i, representa o erro aleatório em Y para cada observação, i. Em outras palavras, ε_i corresponde à distância vertical do valor verdadeiro de Y_i acima ou abaixo do valor previsto para Y_i na linha.

A seleção do modelo matemático apropriado depende da distribuição dos valores de X e de Y no gráfico de dispersão. A Figura 13.2 ilustra seis diferentes tipos de relação.

FIGURA 13.2
Seis tipos de relações encontradas em gráficos de dispersão



No Painel A, os valores de Y estão, de um modo geral, crescendo linearmente à medida que X cresce. Esse painel é semelhante à Figura 13.3, que ilustra a relação positiva entre as dimensões da loja, medidas em unidades de pés quadrados, e as vendas anuais das filiais da Sunflowers Roupas, cadeia de lojas de roupas femininas.

O Painel B é um exemplo de uma relação linear negativa. À medida que X cresce, os valores de Y estão, de um modo geral, decrescendo. Um exemplo desse tipo de relação poderia ser o preço de um determinado produto e o montante de vendas.

Os dados no Painel C mostram uma relação curvilínea positiva entre X e Y. Os valores de Y crescem à medida que X cresce, mas esse crescimento vai se estabilizando quando se ultrapassam determinados valores de X. Um exemplo de uma relação curvilínea positiva poderia ser a idade (tempo de uso) e o custo de manutenção de um equipamento. À medida que o equipamento vai se tornando mais antigo, o custo de manutenção pode crescer rapidamente no princípio, mas, depois disso, acima de um determinado número de anos, ele vai se estabilizando.

O Painel D mostra uma relação em formato de U entre X e Y. À medida que X cresce, Y geralmente diminui no princípio, mas, à medida que X continua a crescer, Y não somente para de decrescer como, na realidade, cresce além de seu valor mínimo. Um exemplo desse tipo de relação poderia ser o número de erros por hora em uma determinada tarefa e o número de horas trabalhadas. O número de erros por hora decresce à medida que o indivíduo vai ficando mais especializado na tarefa, mas, depois de um tempo, passa a crescer acima de um certo patamar devido a fatores tais como fadiga e monotonia.

O Painel E indica uma relação exponencial entre X e Y. Nesse caso, Y decresce muito rapidamente à medida que X começa a crescer, embora, depois disso, passe a decrescer muito menos rapidamente conforme X continua a crescer. Um exemplo de uma relação exponencial poderia ser o valor de revenda de um automóvel e a idade desse bem. No primeiro ano, o valor de revenda cai drasticamente em relação ao seu preço original, mas esse valor de revenda decresce muito menos rapidamente nos anos subsequentes.

Por fim, o Painel F mostra um conjunto de dados nos quais existe muito pouca, ou nenhuma, relação entre X e Y. Valores altos e valores baixos de Y aparecem a cada valor de X.

Embora gráficos de dispersão sejam úteis para ilustrar visualmente a fórmula matemática de uma relação, procedimentos estatísticos mais sofisticados encontram-se disponíveis para determinar o modelo mais apropriado para um conjunto de variáveis. O restante deste capítulo discute o modelo utilizado quando existe uma relação *linear* entre variáveis.

13.2 Determinando a Equação da Regressão Linear Simples

No cenário que trata da Sunflowers Roupas, do início deste capítulo, o objetivo estratégico do diretor de planejamento era prever as vendas anuais para todas as novas lojas, com base no tamanho da loja. Para examinar a relação entre o tamanho de uma loja, medido em unidades de pés quadrados, e as suas respectivas vendas anuais, foram coletados os dados de uma amostra de 14 lojas. A Tabela 13.1 apresenta os dados organizados, armazenados no arquivo **Localização**.

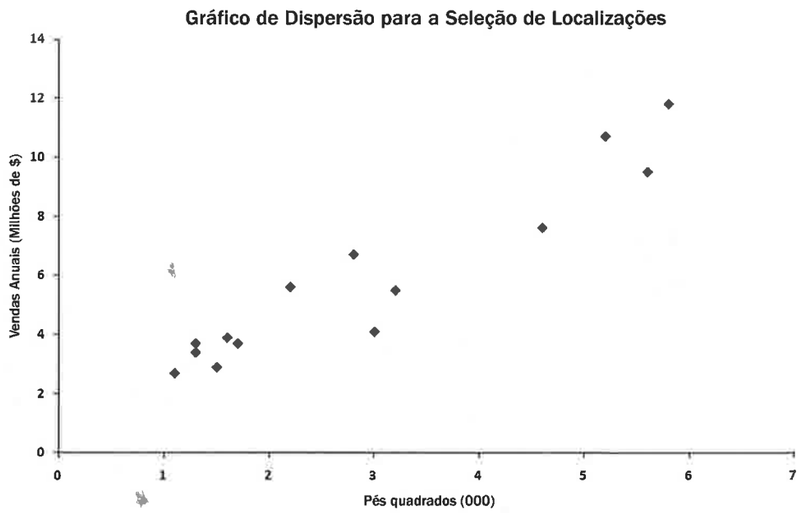
TABELA 13.1
Área em Pés Quadrados (em Milhares de Pés Quadrados) e Vendas Anuais (em Milhões de Dólares) para uma Amostra de 14 Filiais da Sunflowers Roupas

Loja	Área em Pés Quadrados (Milhares)	Vendas Anuais (em Milhões de Dólares)	Loja	Área em Pés Quadrados (Milhares)	Vendas Anuais (em Milhões de Dólares)
1	1,7	3,7	8	1,1	2,7
2	1,6	3,9	9	3,2	5,5
3	2,8	6,7	10	1,5	2,9
4	5,6	9,5	11	5,2	10,7
5	1,3	3,4	12	4,6	7,6
6	2,2	5,6	13	5,8	11,8
7	1,3	3,7	14	3,0	4,1

A Figura 13.3 ilustra o gráfico de dispersão para os dados na Tabela 13.1. Observe a relação crescente entre a área em pés quadrados (X) e as vendas anuais (Y). À medida que o tamanho da loja aumenta, as vendas anuais crescem aproximadamente em uma linha reta. Assim, você pode pressupor que uma linha reta proporciona um modelo matemático adequado para essa relação. Agora, você precisa determinar a linha reta específica que representa o *melhor* ajuste para esses dados.

FIGURA 13.3
Gráfico de dispersão para os dados relacionados à Sunflowers Roupas

Crie gráficos de dispersão utilizando as instruções na Seção GE2.7.



O Método dos Mínimos Quadrados

Na seção anterior, um modelo estatístico foi colocado a título de hipótese para representar a relação entre duas variáveis, área em pés quadrados e vendas, para toda a população de lojas da Sunflowers Roupas. No entanto, conforme ilustrado na Tabela 13.1, os dados são apenas de uma amostra aleatória de lojas. Caso determinados pressupostos sejam válidos (veja a Seção 13.4), você pode utilizar o intercepto de Y da amostra, b_0, e a inclinação da amostra, b_1, como estimadores para os respectivos parâmetros da população, β_0 e β_1. A Equação (13.2) utiliza esses estimadores para formar a **equação da regressão linear simples**. Essa linha reta é frequentemente denominada **linha de previsão**.

EQUAÇÃO DA REGRESSÃO LINEAR SIMPLES: A LINHA DE PREVISÃO

O valor previsto de Y é igual ao intercepto de Y somado à inclinação vezes o valor de X.

$$\hat{Y}_i = b_0 + b_1 X_i \quad \textbf{(13.2)}$$

em que

$\hat{Y}_i$ = valor previsto de Y para a observação i

X_i = valor de X para observação i

b_0 = intercepto da amostra, Y

b_1 = inclinação da amostra

A Equação (13.2) requer a determinação de dois **coeficientes da regressão** – b_0 (o intercepto de Y da amostra) e b_1 (a inclinação da amostra). A abordagem mais comum para encontrar b_0 e b_1 é o método dos mínimos quadrados. Esse método minimiza a soma das diferenças, elevadas ao quadrado, entre os valores verdadeiros (Y_i) e os valores previstos ($\hat{Y}_i$),utilizando a equação da regressão linear simples [ou seja, a linha de previsão; veja a Equação (13.2)]. Essa soma das diferenças elevadas ao quadrado é igual a

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

Uma vez que $\hat{Y}_i = b_0 + b_1 X_i$,

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i - (b_0 + b_1 X_i)]^2$$

Tendo em vista que essa equação possui duas incógnitas, b_0 e b_1, a soma das diferenças elevadas ao quadrado depende do intercepto de Y da amostra, b_0, e da inclinação da amostra, b_1. O **método dos mínimos quadrados** determina os valores de b_0 e b_1 que minimizam a soma das diferenças elevadas ao quadrado. Quaisquer valores para b_0 e b_1 outros que não sejam aqueles determinados pelo método dos mínimos quadrados resultam em maior soma das diferenças elevadas ao quadrado entre os valores reais (Y_i) e os valores previstos, ($\hat{Y}_i$). A Figura 13.4[1] apresenta em uma planilha os resultados para o modelo de regressão linear simples com os dados da Tabela 13.1.

Na Figura 13.4, observe que $b_0 = 0{,}9645$ e $b_1 = 1{,}6699$. Por conseguinte, a linha de previsão [veja a Equação (13.2)] para esses dados é:

$$\hat{Y}_i = 0{,}9645 + 1{,}6699 X_i$$

A inclinação, b_1, é igual a +1,6699. Isso significa que para cada crescimento equivalente correspondente a 1 unidade em X estima-se que o valor previsto para Y cresça em 1,6699 unidade. Em outras palavras, para cada crescimento de 1,0 mil pés quadrados no tamanho da loja estima-se que a previsão para as vendas anuais cresça em 1,6699 milhão de dólares. Por conseguinte, a inclinação representa a parcela das vendas anuais que se estima variar de acordo com o tamanho da loja.

O intercepto de Y, b_0, é igual a +0,9645. O intercepto de Y representa o valor previsto para Y quando X é igual a 0. Uma vez que a área da loja, em pés quadrados, não pode ser 0 (zero), esse

[1] As equações utilizadas para calcular esses resultados estão ilustradas nos Exemplos 13.3 e 13.4, nesta seção e na próxima, respectivamente. Para maiores conjuntos de dados, você deve fazer uso de softwares para realizar esses cálculos, dada a sua natureza complexa.

	A	B	C	D	E	F	G	H	I
1	Regressão Linear Simples								
2									
3	*Estatística de Regressão*								
4	R Múltiplo	0,9509							
5	R-quadrado	0,9042							
6	R-quadrado ajustado	0,8962							
7	Erro-padrão	0,9664							
8	Observações	14							
9									
10	ANOVA								
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*			
12	Regressão	1	105,7476	105,7476	113,2335	0,0000			
13	Resíduos	12	11,2067	0,9339					
14	Total	13	116,9543						
15									
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*	*95,0% inferiores*	*95,0% superiores*
17	Interseção	0,9645	0,5262	1,8329	0,0917	-0,1820	2,1110	-0,1820	2,11095
18	Área em Pés Quadrados	1,6699	0,1569	10,6411	0,0000	1,3280	2,0118	1,3280	2,01177

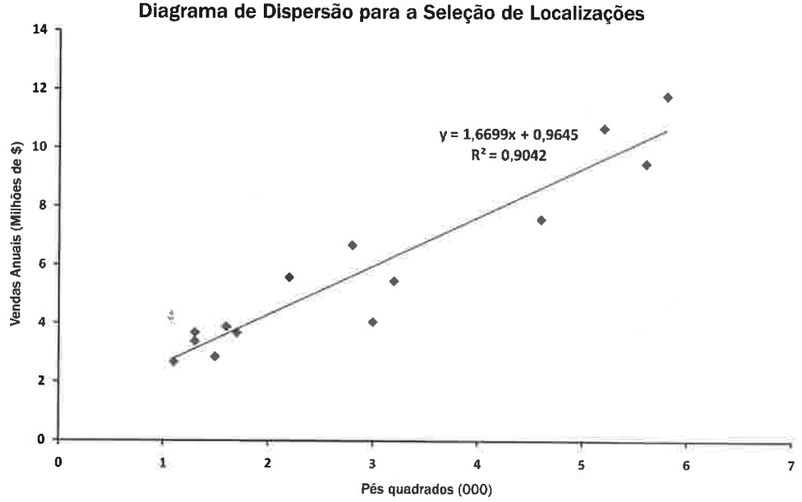
FIGURA 13.4
Resultados da planilha para o modelo de regressão linear simples da Tabela 13.1 para os dados relacionados à Sunflowers Roupas

A Figura 13.4 ilustra a ***planilha CÁLCULO*** *da* ***pasta de trabalho Regressão Linear Simples****. Crie essa planilha utilizando as instruções na Seção GE13.2. Leia as instruções do Excel Avançado referentes a essa planilha para aprender sobre as fórmulas utilizadas em toda a sua extensão, incluindo a área de Cálculos nas colunas de K a M (não ilustradas na Figura 13.4).*

intercepto de Y tem pouca ou nenhuma interpretação prática. Do mesmo modo, o intercepto de Y para este exemplo encontra-se fora do intervalo dos valores observados para a variável X, e, por conseguinte, as interpretações para o valor de b_0 devem ser realizadas com bastante cautela. A Figura 13.5 ilustra as observações reais e a linha de previsão. Para ilustrar uma situação em que existe uma interpretação direta para o intercepto de Y, b_0, examine o Exemplo 13.1.

FIGURA 13.5
Gráfico de dispersão e linha de regressão para os dados relacionados à Sunflowers Roupas

Acrescente uma linha de previsão e uma equação de regressão utilizando as instruções na Seção GE13.2.



EXEMPLO 13.1

Interpretando o Intercepto de Y, b_0, e a Inclinação, b_1

Um professor de estatística deseja utilizar o número de horas que um aluno estuda para uma prova final de estatística (X) para prever a nota da prova final (Y). Foi ajustado um modelo de regressão com base nos dados coletados de uma classe durante o semestre anterior, com os seguintes resultados:

$$\hat{Y}_i = 35{,}0 + 3X_i$$

Qual é a interpretação para o intercepto de Y, b_0, e para a inclinação, b_1?

SOLUÇÃO O intercepto de Y, $b_0 = 35{,}0$, indica que, quando o aluno não estuda para a prova final, o resultado previsto para a prova final é 35,0. A inclinação, $b_1 = 3$, indica que, para cada crescimento de uma hora no tempo dedicado ao estudo, a alteração prevista para o resultado da prova final é igual a +3,0. Em outras palavras, é previsto que o resultado para a prova final cresça em 3 pontos para cada hora de aumento no tempo de estudo.

Retorne ao cenário que trata da Sunflowers Roupas. O Exemplo 13.2 ilustra o modo como você utiliza a linha de previsão para prever as vendas anuais.

EXEMPLO 13.2

Prevendo as Vendas Anuais, com Base na Área em Pés Quadrados

Utilize a linha de previsão para prever a média aritmética das vendas anuais para uma loja com 4.000 pés quadrados.

SOLUÇÃO Você pode determinar o valor previsto substituindo X por 4 (mil pés quadrados), na equação para a regressão linear simples:

$$\hat{Y}_i = 0{,}9645 + 1{,}6699X_i$$

$$\hat{Y}_i = 0{,}9645 + 1{,}6699(4) = 7{,}644 \text{ ou } \$7.644.000$$

Por conseguinte, uma loja com uma área de 4.000 pés quadrados tem uma previsão de vendas anual de $7.644.000.

Previsões na Análise da Regressão: Interpolação *Versus* Extrapolação

Ao utilizar um modelo de regressão para fins de previsão, você precisa considerar somente o **intervalo relevante** da variável independente ao fazer previsões. Esse intervalo relevante inclui todos os valores, desde o menor X até o maior X, utilizados no desenvolvimento do modelo de regressão. Por conseguinte, ao prever Y para um determinado valor de X, você pode interpolar dentro dos limites desse intervalo relevante de valores de X, mas você não deve extrapolar além do intervalo dos valores de X. Quando você utiliza a área em pés quadrados para prever as vendas anuais, a área em pés quadrados (em milhares de pés quadrados) varia desde 1,1 até 5,8 (veja a Tabela 13.1). Por conseguinte, você deve prever as vendas anuais *somente* para lojas cujo tamanho esteja entre 1,1 e 5,8 mil pés quadrados. Qualquer previsão de vendas anuais para lojas cujo tamanho esteja fora desse intervalo pressupõe que a relação observada entre vendas e tamanho da loja, para lojas com tamanho entre 1,1 e 5,8 mil pés quadrados, é a mesma para lojas que estejam fora desse intervalo. Por exemplo, você não pode extrapolar a relação linear para além de 5.800 mil pés quadrados, no Exemplo 13.2. Seria impróprio utilizar a linha de previsão para fazer prognósticos para vendas de uma nova loja que tenha uma área de 8.000 pés quadrados uma vez que a relação entre vendas e tamanho da loja tem um ponto de retornos decrescentes. Caso isso seja verdadeiro, à medida que a área cresce além de 5.800 pés quadrados, o efeito sobre as vendas vai se tornando cada vez menor.

Calculando o Intercepto de Y, b_0, e a Inclinação, b_1

Para pequenos conjuntos de dados, você pode utilizar uma calculadora de mão para calcular os coeficientes da regressão dos mínimos quadrados. As Equações (13.3) e (13.4) fornecem os valores de b_0 e b_1, que minimizam

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i - (b_0 + b_1X_i)]^2$$

FÓRMULA DE CÁLCULO PARA A INCLINAÇÃO, b_1

$$b_1 = \frac{SQXY}{SQX} \tag{13.3}$$

em que

$$SQXY = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n}X_iY_i - \frac{\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{n}$$

$$SQX = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}X_i^2 - \frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}$$

FÓRMULA DE CÁLCULO PARA O INTERCEPTO DE Y, b_0

$$b_0 = \overline{Y} - b_1\overline{X} \tag{13.4}$$

em que

$$\overline{Y} = \frac{\sum_{i=1}^{n}Y_i}{n}$$

$$\overline{X} = \frac{\sum_{i=1}^{n}X_i}{n}$$

EXEMPLO 13.3

Calculando o Intercepto de Y, b_0, e a Inclinação, b_1

Calcule o intercepto de Y, b_0, e a inclinação, b_1, para os dados relativos à Sunflowers Roupas.

SOLUÇÃO Examinando as Equações (13.3) e (13.4), cinco valores precisam ser calculados para determinar b_0 e b_1. Esses valores são n, o tamanho da amostra; $\sum_{i=1}^{n}X_i$, a soma dos valores de X; $\sum_{i=1}^{n}Y_i$, a soma dos valores de Y; $\sum_{i=1}^{n}X_i^2$, a soma dos valores de X elevados ao quadrado; e $\sum_{i=1}^{n}X_iY_i$, a soma do produto dos valores de X e Y. Para os dados da Sunflowers Roupas, o número de pés quadrados (X) é utilizado para prever as vendas anuais em uma loja (Y). A Tabela 13.2 apresenta os cálculos das somas necessárias para o problema que trata da seleção de locais para lojas. A

TABELA 13.2
Cálculos para os Dados da Sunflowers Roupas

Loja	Área em Pés Quadrados (X)	Vendas Anuais (Y)	X^2	Y^2	XY
1	1,7	3,7	2,89	13,69	6,29
2	1,6	3,9	2,56	15,21	6,24
3	2,8	6,7	7,84	44,89	18,76
4	5,6	9,5	31,36	90,25	53,20
5	1,3	3,4	1,69	11,56	4,42
6	2,2	5,6	4,84	31,36	12,32
7	1,3	3,7	1,69	13,69	4,81
8	1,1	2,7	1,21	7,29	2,97
9	3,2	5,5	10,24	30,25	17,60
10	1,5	2,9	2,25	8,41	4,35
11	5,2	10,7	27,04	114,49	55,64
12	4,6	7,6	21,16	57,76	34,96
13	5,8	11,8	33,64	139,24	68,44
14	3,0	4,1	9,00	16,81	12,30
Totais	40,9	81,8	157,41	594,90	302,30

tabela inclui também, $\sum_{i=1}^{n}Y_i^2$, a soma dos valores de Y elevados ao quadrado que serão utilizados para calcular STQ na Seção 13.3.

Utilizando as Equações (13.3) e (13.4), você pode calcular os valores para b_0 e b_1:

$$SQXY = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n}X_iY_i - \frac{\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{n}$$

$$SQXY = 302,3 - \frac{(40,9)(81,8)}{14}$$
$$= 302,3 - 238,97285$$
$$= 63,32715$$

$$SQX = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}X_i^2 - \frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}$$

$$= 157,41 - \frac{(40,9)^2}{14}$$
$$= 157,41 - 119,48642$$
$$= 37,92358$$

Portanto,

$$b_1 = \frac{SQXY}{SQX}$$
$$= \frac{63,32715}{37,92358}$$
$$= 1,6699$$

E,

$$\overline{Y} = \frac{\sum_{i=1}^{n}Y_i}{n} = \frac{81,8}{14} = 5,842857$$

$$\overline{X} = \frac{\sum_{i=1}^{n}X_i}{n} = \frac{40,9}{14} = 2,92143$$

Portanto,

$$b_0 = \overline{Y} - b_1\overline{X}$$
$$= 5,842857 - (1,6699)(2,92143)$$
$$= 0,9645$$

EXPLORAÇÕES VISUAIS — Explorando Coeficientes da Regressão Linear Simples

Utilize o procedimento Simple Linear Regression (Regressão Linear Simples) em Visual Explorations (Explorações Visuais) para criar uma linha de previsão que seja a mais próxima possível da linha de previsão definida pela solução dos mínimos quadrados. Abra a pasta de trabalho do suplemento **Visual Explorations (Explorações Visuais)** (veja a Seção D.7 do Apêndice D deste livro) e:

1. No Excel 2007, selecione **Suplementos → VisualExplorations (ExploraçõesVisuais) → Simple Linear Regres-**

sion (Regressão Linear Simples). No Excel 2003, selecione **VisualExplorations (ExploraçõesVisuais) → Simple Linear Regression (Regressão Linear Simples)**

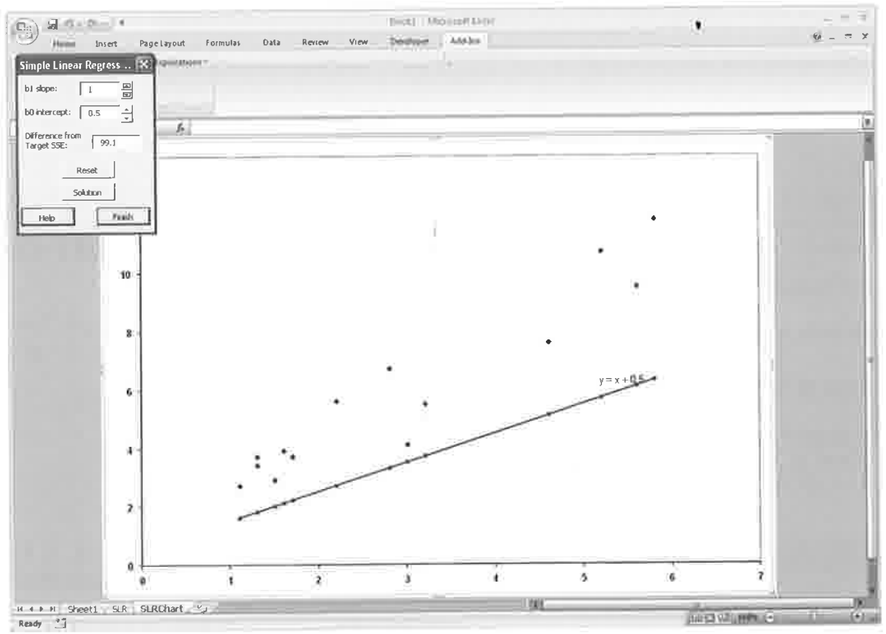
Na caixa de diálogo Simple Linear Regression (Regressão Linear Simples), ilustrada a seguir:

2. Clique nos botões giratórios correspondentes à inclinação da linha de previsão **b1 slope (inclinação b1)** e ao intercepto de *Y* da linha de previsão, **b0 intercept (intercepto b0)** para modificar a linha de previsão.
3. Utilizando a resposta visual do gráfico, tente criar uma linha de previsão ajustada que seja a mais próxima possível da linha de previsão definida pelas estimativas dos mínimos quadrados. Em outras palavras, tente tornar o valor para **Difference from Target SSE (Diferença em relação a SQR desejado)** o menor possível (veja a Seção 13.3 para uma explicação sobre *SQR*).

A qualquer momento, clique no botão **Reset (Limpar)** para limpar os valores de b_0 e b_1 ou em **Solution (Solução)** para revelar a linha de previsão definida pelo método dos mínimos quadrados. Clique em **Finish (Concluir)** assim que tiver terminado de fazer este exercício.

Utilizando Seus Próprios Dados de Regressão

Selecione **Simple Linear Regression with your worksheet data (Regressão Linear Simples com os dados de sua planilha)** do menu **VisualExplorarions (ExploraçõesVisuais)** para explorar os coeficientes da regressão linear simples utilizando dados que você forneça a partir de uma planilha. Na caixa de diálogo do procedimento, insira um intervalo de células para a sua variável *Y* na caixa de edição **Y Variable Cell Range (Intervalo de Células da Variável Y)** e o seu intervalo de células da variável *X* na caixa de edição **X Variable Cell Range (Intervalo de Células da Variável X)**. Clique em **First cell in both ranges contain a label (Primeira célula em ambos os intervalos contém uma legenda)**, insira um título na caixa **Title** e clique em **OK**. Tão logo apareça um gráfico de dispersão com uma linha de previsão inicial, utilize as instruções apresentadas anteriormente nas etapas 2 e 3.



Problemas para a Seção 13.2

APRENDENDO O BÁSICO

13.1 O ajuste de uma linha reta a um conjunto de dados produz a seguinte linha de previsão:

$$\hat{Y}_i = 2 + 5X_i$$

a. Interprete o significado do intercepto de *Y*, b_0.
b. Interprete o significado da inclinação, b_1.
c. Faça a previsão do valor de *Y* para $X = 3$.

13.2 Se os valores de *X* no Problema 13.1 se estendem desde 2 até 25, você deve utilizar esse modelo para prever a média aritmética do valor de *Y* quando *X* é igual a
a. 3? **b.** –3? **c.** 0? **d.** 24?

13.3 O ajuste de uma linha reta a um conjunto de dados produz a seguinte linha de previsão:

$$\hat{Y}_i = 16 - 0{,}5X_i$$

a. Interprete o significado do intercepto de *Y*, b_0.
b. Interprete o significado da inclinação, b_1.
c. Faça a previsão do valor de *Y* para $X = 6$.

APLICANDO OS CONCEITOS

AUTO-AVALIAÇÃO

13.4 O gerente de marketing de uma grande cadeia de supermercados gostaria de utilizar o espaço disponível em prateleiras de supermercado para prever as vendas de rações para animais de estimação. Uma amostra aleatória de 12 lojas de igual tamanho foi selecionada, com os seguintes resultados (armazenados no arquivo **Ração**):

Loja	Espaço de Prateleira, (*X*) (Pés)	Vendas Semanais, (*Y*)
1	5	160
2	5	220
3	5	140
4	10	190
5	10	240
6	10	260
7	15	230
8	15	270
9	15	280
10	20	260
11	20	290
12	20	310

a. Construa um gráfico de dispersão.

Para esses dados, $b_0 = 145$ e $b_1 = 7{,}4$.

b. Interprete o significado da inclinação, b_1, neste problema.
c. Faça a previsão das vendas semanais para rações para animais de estimação para lojas com 8 pés de espaço de prateleira destinado ao produto.

13.5 A circulação é a energia vital do negócio das editoras de revistas. Quanto maior o volume de vendas de uma revista, mais ela consegue cobrar de seus anunciantes. No entanto, foi detectada uma defasagem de circulação entre os dados informados pelos editores de revistas sobre as vendas em bancas e as subsequentes auditorias feitas pelo Audit Bureau of Circulations (Departamento de Auditoria de Circulação), nos Estados Unidos. Os dados no arquivo **Circulação** representam o volume de vendas de revistas informado e o volume auditado (em milhares) para as 10 revistas a seguir:

Revista	Informado (*X*)	Auditado (*Y*)
YM	621,0	299,6
CosmoGirl	359,7	207,7
Rosie	530,0	325,0
Playboy	492,1	336,3
Esquire	70,5	48,6
TeenPeople	567,0	400,3
More	125,5	91,2
Spin	50,6	39,1
Vogue	353,3	268,6
Elle	263,6	214,3

Fonte: *Extraído de M. Rose, "In Fight for Ads, Publishers Often Overstate Their Sales"*, The Wall Street Journal, *6 de agosto de 2003, pp. A1, A10.*

a. Construa um gráfico de dispersão.

Para esses dados, $b_0 = 26{,}724$ e $b_1 = 0{,}5719$.

b. Interprete o significado da inclinação, b_1, neste problema.
c. Faça a previsão para o volume de vendas de revistas auditado para um editor de revistas que informa suas vendas em bancas como sendo de 400.000.

13.6 O proprietário de uma empresa de mudanças geralmente faz com que o seu gerente mais experiente faça a previsão do número de horas de trabalho que serão necessárias para realizar uma mudança que esteja por ocorrer. Esse método mostrou-se útil no passado, mas o proprietário tem como objetivo estratégico da empresa desenvolver um método mais preciso para prever a quantidade de horas de trabalho. Em um esforço preliminar para proporcionar um método mais acurado, o proprietário decidiu utilizar a quantidade de pés cúbicos a serem transportados na mudança como a variável independente e coletou dados gerados por 36 mudanças cuja origem e destino estavam dentro dos limites de Manhattan, na cidade de Nova York, e nas quais o tempo de transporte representava uma parcela insignificante em relação à quantidade de horas trabalhadas. Os dados encontram-se armazenados no arquivo **Mudança**.
a. Construa um gráfico de dispersão.
b. Pressupondo uma relação linear, utilize o método dos mínimos quadrados para determinar os coeficientes de regressão, b_0 e b_1.
c. Interprete o significado da inclinação, b_1, neste problema.
d. Faça a previsão para a quantidade de horas trabalhadas necessárias para uma mudança com um volume de 500 pés cúbicos.

13.7 Um aspecto crucialmente importante no atendimento a clientes em um supermercado é o tempo de espera na hora de pagar pelas compras na saída do supermercado (definido como o intervalo de tempo desde o momento em que o cliente entra na fila até o momento em que é atendido). Foram coletados dados durante períodos do dia nos quais uma quantidade constante de caixas para pagamento estava aberta. O número total de clientes na loja e os tempos de espera (em minutos) foram registrados. Os resultados estão armazenados no arquivo **Supermercado**.
a. Construa um gráfico de dispersão.
b. Pressupondo uma relação linear, utilize o método dos mínimos quadrados para determinar os coeficientes de regressão, b_0 e b_1.
c. Interprete o significado da inclinação, b_1, neste problema.
d. Faça uma previsão para o tempo de espera quando existem 20 clientes dentro da loja.

13.8 O valor de uma franquia no setor de esportes está diretamente relacionado ao volume de receitas que essa franquia pode gerar. Os dados no arquivo **BBReceita** representam o valor em 2009 (em milhões de dólares) e a receita anual (em milhões de dólares) para as 30 franquias da principal liga de beisebol nos EUA. Suponha que você queira desenvolver um modelo de regressão linear simples para prever o valor da franquia, com base na receita anual gerada.
a. Construa um gráfico de dispersão.
b. Utilize o método dos mínimos quadrados para encontrar os coeficientes de regressão, b_0 e b_1.
c. Interprete o significado de b_0 e b_1 neste problema.
d. Faça a previsão para o valor de uma franquia de beisebol que gere $200 milhões em receitas anuais.

13.9 Um corretor de uma imobiliária em uma grande cidade gostaria de poder prever o custo mensal para aluguéis de apar-

tamentos com base no tamanho do imóvel, definido pela área em pés quadrados. O corretor seleciona uma amostra com 25 apartamentos em um determinado bairro residencial e coleta os dados a seguir (armazenados em **Aluguel**).

a. Construa um gráfico de dispersão.
b. Utilize o método dos mínimos quadrados para determinar os coeficientes de regressão, b_0 e b_1.
c. Interprete o significado de b_0 e b_1 neste problema.
d. Faça a previsão para o custo mensal do aluguel para um apartamento com uma área de 1.000 pés quadrados.
e. Por que não seria apropriado utilizar o modelo para prever o aluguel mensal para apartamentos de 500 pés quadrados?
f. Seus amigos Jim e Jennifer estão avaliando a possibilidade de assinar um contrato de aluguel para um apartamento nesse mesmo bairro residencial. Eles estão tentando decidir entre dois apartamentos, um com área de 1.000 pés quadrados e aluguel mensal de \$1.275 e outro com área de 1.200 pés quadrados e aluguel mensal de \$1.425. Com base nos itens (a) até (d), qual apartamento você acha que seria um melhor negócio?

13.10 Uma empresa que administra os direitos de distribuição de DVD para filmes anteriormente liberados para exibição exclusiva em salas de cinema deseja estimar as vendas de DVDs com base no respectivo sucesso de bilheteria. O arquivo **Cinema** apresenta uma lista com a receita bruta de bilheteria (em milhões de dólares) para cada um de 30 filmes e a quantidade de DVDs vendidos (em milhares). Para esses dados,

a. construa um gráfico de dispersão.
b. pressupondo uma relação linear, utilize o método dos mínimos quadrados para determinar os coeficientes da regressão, b_0 e b_1.
c. interprete o significado da inclinação, b_1, neste problema.
d. faça a previsão para as vendas do DVD de um filme que tenha auferido uma receita bruta de bilheteria de \$75 milhões.

Tabela para o Problema 13.9

Apartamento	Aluguel Mensal (\$)	Tamanho (Pés Quadrados)	Apartamento	Aluguel Mensal (\$)	Tamanho (Pés Quadrados)
1	950	850	14	1.800	1.369
2	1.600	1.450	15	1.400	1.175
3	1.200	1.085	16	1.450	1.225
4	1.500	1.232	17	1.100	1.245
5	950	718	18	1.700	1.259
6	1.700	1.485	19	1.200	1.150
7	1.650	1.136	20	1.150	896
8	935	726	21	1.600	1.361
9	875	700	22	1.650	1.040
10	1.150	956	23	1.200	755
11	1.400	1.100	24	800	1.000
12	1.650	1.285	25	1.750	1.200
13	2.300	1.985			

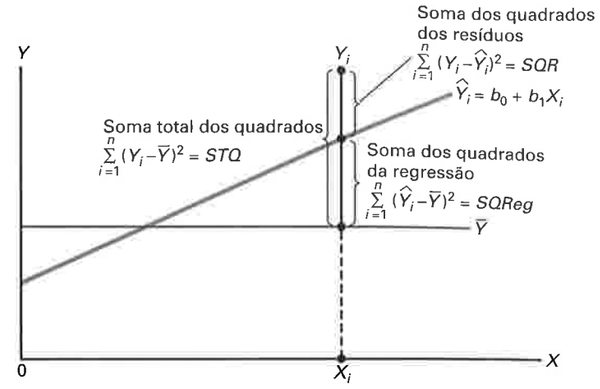
13.3 Medidas de Variação

Ao utilizar o método dos mínimos quadrados para determinar os coeficientes de regressão para um conjunto de dados, você precisa calcular três importantes medidas de variação. A primeira medida, a **soma total dos quadrados (*STQ*)**, é a medida de variação dos valores de Y_i em torno de sua média aritmética, $\bar{Y}$. A **variação total**, ou soma total dos quadrados, é subdividida entre **variação explicada** e **variação não explicada**. A variação explicada, ou **soma dos quadrados da regressão (*SQReg*)**, representa a variação decorrente da relação entre X e Y, enquanto a variação não explicada, ou **soma dos quadrados dos resíduos (erros) (*SQR*)**, representa a variação decorrente de fatores outros que não a relação entre X e Y. A Figura 13.6 mostra essas diferentes medidas de variação.

Calculando a Soma dos Quadrados

A soma dos quadrados da regressão (*SQReg*) é baseada na diferença entre $\hat{Y}_i$ (o valor previsto de Y com base na linha de previsão) e $\bar{Y}$ (a média aritmética para o valor de Y). A soma dos quadrados dos resíduos (erros) (*SQR*) representa a parcela da variação em Y que não é explicada pela regressão. Ela é baseada na diferença entre Y_i e $\hat{Y}_i$. As Equações (13.5), (13.6), (13.7) e (13.8) definem essas medidas de variação.

FIGURA 13.6
Medidas de variação



MEDIDAS DE VARIAÇÃO NA REGRESSÃO

A soma total dos quadrados é igual à soma dos quadrados da regressão (*SQReg*) acrescida da soma dos quadrados dos resíduos ou erros (*SQReg*).

$$STQ = SQReg + SQR \qquad \textbf{(13.5)}$$

SOMA TOTAL DOS QUADRADOS (*STQ*)

A soma total dos quadrados (*STQ*) é igual à soma das diferenças elevadas ao quadrado entre cada um dos valores observados de Y e a média aritmética para o valor de Y.

$$STQ = \text{Soma total dos quadrados}$$
$$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad \textbf{(13.6)}$$

SOMA DOS QUADRADOS DA REGRESSÃO (*SQReg*)

A soma dos quadrados da regressão (*SQReg*) é igual à soma das diferenças elevadas ao quadrado entre cada um dos valores previstos para Y e a média aritmética para o valor de Y.

$$SQReg = \text{Variação explicada ou soma da regressão dos quadrados}$$
$$= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \qquad \textbf{(13.7)}$$

SOMA DOS QUADRADOS DOS RESÍDUOS OU ERROS (*SQR*)

A soma dos quadrados dos resíduos ou erros (*SQR*) é igual à soma das diferenças elevadas ao quadrado entre o valor observado de Y e o valor previsto para Y.

$$SQR = \text{Variação não explicada ou soma dos quadrados dos resíduos (erros)}$$
$$= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad \textbf{(13.8)}$$

A Figura 13.7 apresenta a parte da planilha correspondente à área da soma dos quadrados para os dados relacionados à Sunflowers Roupas. A variação total, *STQ*, é igual a 116,9543. Esse valor é subdividido entre a soma dos quadrados explicada pela regressão (*SQReg*), igual a 105,7476, e a soma dos quadrados que não é explicada pela regressão (*SQR*), igual a 11,2067. Com base na Equação (13.5),

$$STQ = SQReg + SQR$$
$$116{,}9543 = 105{,}7476 + 11{,}2067$$

FIGURA 13.7
Resultado da planilha com a soma dos quadrados, para os dados relacionados à Sunflowers Roupas

Na planilha, SQReg é representada pela célula (C12) que tem a legenda ***SQ Regressão****, enquanto SQR é representada pela célula (C13), que tem a legenda* ***SQ Resíduo****.*

	A	B	C	D	E	F
10	**ANOVA**					
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*
12	Regressão	1	105,7476	105,7476	113,2335	0,0000
13	Resíduos	12	11,2067	0,9339		
14	Total	13	116,9543			

O Coeficiente de Determinação

Por si sós, *SQReg*, *SQR* e *STQ* oferecem poucas informações. No entanto, o quociente entre a soma dos quadrados da regressão (*SQReg*) e a soma total dos quadrados (*STQ*) mede a proporção da variação em *Y* que é explicada pela variável independente *X* no modelo de regressão. Esse quociente, conhecido como **coeficiente de determinação**, r^2, é definido na Equação (13.9).

COEFICIENTE DE DETERMINAÇÃO

O coeficiente de determinação é igual à soma dos quadrados da regressão (ou seja, a variação explicada) dividida pela soma total dos quadrados (ou seja, a variação total).

$$r^2 = \frac{\text{Soma dos quadrados da regressão}}{\text{Soma total dos quadrados}} = \frac{SQReg}{STQ} \qquad \textbf{(13.9)}$$

O **coeficiente de determinação** mede a proporção da variação em *Y* que é explicada pela variável independente *X*, no modelo de regressão.

Para os dados relacionados à Sunflowers Roupas, com $SQReg = 105{,}7476$, $SQR = 11{,}2067$ e $STQ = 116{,}9543$,

$$r^2 = \frac{105{,}7476}{116{,}9543} = 0{,}9042$$

Por conseguinte, 90,42% da variação nas vendas anuais é explicada pela variabilidade no tamanho da loja, medida com base na área em pés quadrados. Esse alto valor de r^2 indica uma forte relação linear positiva entre duas variáveis, uma vez que a utilização de um modelo de regressão reduziu em 90,42% a variabilidade na previsão de vendas anuais. Somente 9,58% da variabilidade da amostra, em termos de vendas anuais, é decorrente de fatores outros que não aqueles que são considerados pelo modelo de regressão linear que utiliza a área em pés quadrados.

A Figura 13.8 apresenta a parte da planilha de resultados da regressão que corresponde à tabela de Estatísticas da Regressão para os dados da Sunflowers Roupas. Essa tabela contém o coeficiente de determinação na célula B5 (com a legenda R Quadrado).

FIGURA 13.8
Parte da planilha contendo os resultados das Estatísticas da Regressão para os dados relacionados à Sunflowers Roupas

	A	B
3	***Estatística de Regressão***	
4	R Múltiplo	0,9509
5	R-quadrado	0,9042
6	R-quadrado ajustado	0,8962
7	Erro-padrão	0,9664
8	Observações	14

EXEMPLO 13.4
Calculando o Coeficiente de Determinação

Calcule o coeficiente de determinação, r^2, para os dados relacionados à Sunflowers Roupas.

SOLUÇÃO Você pode calcular *STQ*, *SQReg* e *SQR*, que estão definidas nas Equações (13.6), (13.7) e (13.8), fazendo uso das Equações (13.10), (13.11) e (13.12).

FÓRMULA DE CÁLCULO PARA *STQ*

$$STQ = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}Y_i^2 - \frac{\left(\sum_{i=1}^{n}Y_i\right)^2}{n} \qquad \textbf{(13.10)}$$

FÓRMULA DE CÁLCULO PARA *SQReg*

$$SQReg = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

$$= b_0\sum_{i=1}^{n}Y_i + b_1\sum_{i=1}^{n}X_iY_i - \frac{\left(\sum_{i=1}^{n}Y_i\right)^2}{n} \qquad \textbf{(13.11)}$$

FÓRMULA DE CÁLCULO PARA *SQR*

$$SQR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}Y_i^2 - b_0\sum_{i=1}^{n}Y_i - b_1\sum_{i=1}^{n}X_iY_i \qquad \textbf{(13.12)}$$

Utilizando os resultados resumidos da Tabela 13.2,

$$STQ = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}Y_i^2 - \frac{\left(\sum_{i=1}^{n}Y_i\right)^2}{n}$$

$$= 594{,}9 - \frac{(81{,}8)^2}{14}$$

$$= 594{,}9 - 477{,}94571$$

$$= 116{,}95429$$

$$SQReg = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

$$= b_0\sum_{i=1}^{n}Y_i + b_1\sum_{i=1}^{n}X_iY_i - \frac{\left(\sum_{i=1}^{n}Y_i\right)^2}{n}$$

$$= (0{,}9645)(81{,}8) + (1{,}6699)(302{,}3) - \frac{(81{,}8)^2}{14}$$

$$= 105{,}74726$$

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

$$= \sum_{i=1}^{n}Y_i^2 - b_0\sum_{i=1}^{n}Y_i - b_1\sum_{i=1}^{n}X_iY_i$$

$$= 594{,}9 - (0{,}9645)(81{,}8) - (1{,}6699)(302{,}3)$$

$$= 11{,}2067$$

Portanto,

$$r^2 = \frac{105{,}74726}{116{,}95429} = 0{,}9042$$

Erro-Padrão da Estimativa

Embora o método dos mínimos quadrados resulte na linha que ajusta os dados com a quantidade mínima de erro, a menos que todos os pontos de dados observados se posicionem em uma linha reta, a linha de previsão não se configura como um mecanismo perfeito de previsão. Assim como não se pode esperar que todos os valores dos dados sejam exatamente iguais à sua respectiva média aritmética, tampouco se pode esperar que todos os valores em uma análise da regressão se posicionem exatamente na linha de previsão. A Figura 13.5, ilustra a variabilidade em torno da linha de previsão para os dados relativos à Sunflowers Roupas. Observe que, embora muitos dos valores verdadeiros de Y se posicionem perto da linha de previsão, nenhum desses valores se posiciona exatamente sobre a linha.

O **erro-padrão da estimativa** mede a variabilidade dos valores verdadeiros de Y, a partir dos valores previstos para Y, do mesmo modo que o desvio-padrão, desenvolvido no Capítulo 3, mede a variabilidade para cada um dos valores em torno da média aritmética da amostra. Em outras palavras, o erro-padrão da estimativa corresponde ao desvio-padrão *em torno* da linha de previsão, enquanto o desvio-padrão, apresentado no Capítulo 3, corresponde ao desvio-padrão *em torno* da média aritmética da amostra. A Equação (13.13) define o erro-padrão da estimativa, representado pelo símbolo S_{YX}.

ERRO-PADRÃO DA ESTIMATIVA

$$S_{YX} = \sqrt{\frac{SQR}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} \qquad \textbf{(13.13)}$$

em que

Y_i = valor real de Y para um determinado X_i

$\hat{Y}_i$ = valor previsto de Y para um determinado X_i

SQR = soma dos quadrados dos resíduos (erros)

Com base na Equação (13.8) e na Figura 13.4 ou na Figura 13.7, $SQR = 11{,}2067$. Consequentemente,

$$S_{YX} = \sqrt{\frac{11{,}2067}{14-2}} = 0{,}9664$$

Esse erro-padrão da estimativa, igual a 0,9664 milhão de dólar (ou seja, \$966.400), tem como título Erro-padrão nos resultados apresentados na planilha da Figura 13.8. O erro-padrão da estimativa representa um indicador da variação em torno da linha de previsão. Ele é medido nas mesmas unidades utilizadas pela variável dependente Y. A interpretação do erro-padrão da estimativa é semelhante àquela para o desvio-padrão. Exatamente do mesmo modo que o desvio-padrão mede a variabilidade em torno da média aritmética, o erro-padrão da estimativa mede a variabilidade em torno da linha de previsão. No que diz respeito à Sunflowers Roupas, a diferença típica entre as vendas anuais reais de uma loja e as vendas anuais previstas utilizando-se a equação da regressão é de aproximadamente \$966.400.

Problemas para a Seção 13.3

APRENDENDO O BÁSICO

13.11 De que maneira você interpreta um coeficiente de determinação, r^2, igual a 0,80?

13.12 Se $SQReg = 36$ e $SQR = 4$, determine STQ e, depois, calcule o coeficiente de determinação, r^2, e interprete o seu significado.

13.13 Se $SQReg = 66$ e $STQ = 88$, calcule o coeficiente de determinação, r^2, e interprete o seu significado.

13.14 Se $SQR = 10$ e $SQReg = 30$, calcule o coeficiente de determinação, r^2, e interprete o seu significado.

13.15 Se $SQReg = 120$, por que é impossível que STQ seja igual a 110?

APLICANDO OS CONCEITOS

AUTO-AVALIAÇÃO **13.16** No Problema 13.4, em Problemas para a Seção 13.2, o gerente de marketing utilizou o espaço em prateleiras disponibilizado para rações para animais de estimação para prever as vendas semanais (dados armazenados em Ração). Para esses dados, $SQReg = 20{,}535$ e $STQ = 30{,}025$.

a. Determine o coeficiente de determinação, r^2, e interprete o seu significado.
b. Determine o erro-padrão da estimativa.
c. Qual o grau de utilidade desse modelo de regressão para fins de previsão sobre vendas?

13.17 No Problema 13.5, em Problemas para a Seção 13.2, você utilizou as vendas informadas de revistas em bancas para prever as vendas auditadas (dados armazenados no arquivo Circulação). Para esses dados, $SQReg = 130.301{,}41$ e $STQ = 144.538{,}64$.

a. Determine o coeficiente de determinação, r^2, e interprete o seu significado.
b. Determine o erro-padrão da estimativa.
c. Qual a utilidade desse modelo de regressão para fins de previsão das vendas auditadas?

13.18 No Problema 13.6, em Problemas para a Seção 13.2, um proprietário de uma empresa de mudanças desejava prever as horas trabalhadas, com base no volume, em pés cúbicos, de material transportado (dados armazenados em Mudança). Utilizando os resultados para aquele problema,

a. determine o coeficiente de determinação, r^2, e interprete o seu significado.
b. determine o erro-padrão da estimativa.
c. Qual o grau de utilidade desse modelo de regressão para fins de previsão de horas de trabalho?

13.19 No Problema 13.7, em Problemas para a Seção 13.2, você utilizou o número total de clientes para o tempo de espera na fila do caixa para pagamento das compras em um supermercado (dados armazenados no arquivo Supermercado). Utilizando os resultados daquele problema,

a. determine o coeficiente de determinação, r^2, e interprete o seu significado.
b. determine o erro-padrão da estimativa.
c. Qual o grau de utilidade desse modelo de regressão para fins de previsão para o tempo de espera na fila do caixa para pagamento das compras na saída de um supermercado?

13.20 No Problema 13.8, em Problemas para a Seção 13.2, você utilizou as receitas anuais para prever o valor de uma franquia de beisebol (dados armazenados no arquivo BBReceita). Utilizando os resultados daquele problema,

a. determine o coeficiente de determinação, r^2, e interprete o seu significado.
b. determine o erro-padrão da estimativa.
c. Qual o grau de utilidade desse modelo de regressão para fins de previsão do valor de uma franquia de beisebol?

13.21 No Problema 13.9, em Problemas para a Seção 13.2, um corretor de uma imobiliária desejava prever o aluguel mensal para os apartamentos com base no tamanho de sua área (dados armazenados no arquivo Aluguel). Utilizando os resultados daquele problema,

a. determine o coeficiente de determinação, r^2, e interprete o seu significado.
b. determine o erro-padrão da estimativa.
c. Qual é a utilidade desse modelo de regressão para fins de previsão do aluguel mensal?
d. Você consegue imaginar outras variáveis que poderiam explicar a variação no aluguel mensal?

13.22 No Problema 13.10, em Problemas para a Seção 13.2, você utilizou a receita bruta de bilheteria para prever as vendas de DVDs (dados armazenados no arquivo Cinema). Utilizando os resultados daquele problema,

a. determine o coeficiente de determinação, r^2, e interprete o seu significado.
b. determine o erro-padrão da estimativa.
c. Qual é a utilidade desse modelo de regressão para fins de previsão para a venda de DVDs?
d. Você consegue imaginar outras variáveis que poderiam explicar a variação nas vendas de DVD?

13.4 Pressupostos

Quando testes de hipóteses e análise da variância foram discutidos ao longo dos Capítulos 9 a 12, foi enfatizada a importância dos pressupostos para a validação de quaisquer conclusões tiradas. Os pressupostos necessários para a regressão são semelhantes àqueles correspondentes à análise da variância, uma vez que ambos os tópicos fazem parte da categoria geral de *modelos lineares* (referência 4).

Os quatro **pressupostos da regressão** (conhecidos como LINI) são os seguintes:

- **L**inearidade
- **I**ndependência de erros
- **N**ormalidade de erros
- **I**gualdade de variâncias

O primeiro pressuposto, **linearidade**, afirma que a relação entre as variáveis é linear. As relações entre variáveis que não sejam lineares são discutidas no Capítulo 15.

O segundo pressuposto, **independência de erros**, requer que os erros (ε_i) sejam independentes entre si. Esse pressuposto é particularmente importante quando os dados são coletados ao longo de um período de tempo. Em tais tipos de situação, os erros para um determinado período de tempo são, algumas vezes, correlacionados aos erros do período de tempo anterior.

O terceiro pressuposto, **normalidade**, requer que os erros (ε_i) sejam normalmente distribuídos para cada um dos valores de X. Do mesmo modo que o teste t e o teste F de ANOVA, a análise da regressão é relativamente robusta em relação a afastamentos do pressuposto da normalidade. Desde que a distribuição dos erros em cada um dos níveis de X não seja extremamente diferente de uma distribuição normal, inferências em relação a β_0 e β_1 não serão seriamente afetadas.

O quarto pressuposto, **igualdade de variâncias**, ou **homoscedasticidade**, requer que a variância dos erros (ε_i) seja constante em relação a todos os valores de X. Em outras palavras, a variabilidade dos valores de Y é a mesma quando X é um valor baixo ou quando X é um valor elevado. O pressuposto da igualdade de variâncias é importante para se realizar inferências em relação a β_0 e β. Caso existam sérios afastamentos em relação a esse pressuposto, você pode utilizar tanto as transformações de dados quanto os métodos dos mínimos quadrados ponderados (veja a referência 4).

13.5 Análise de Resíduos

As Seções 13.2 e 13.3 desenvolveram um modelo de regressão utilizando a abordagem dos mínimos quadrados para os dados relacionados à Sunflowers Roupas. Seria esse o modelo correto para os dados em questão? São válidos os pressupostos introduzidos na Seção 13.4? A **análise de resíduos** avalia visualmente esses pressupostos e ajuda você a determinar se o modelo de regressão que está sendo selecionado é apropriado.

O **resíduo**, ou valor do erro estimado, e_i, corresponde à diferença entre os valores observados (Y_i) e os valores previstos ($\hat{Y}_i$) da variável dependente para um determinado valor de X_i. Um resíduo aparece em um gráfico de dispersão como a distância vertical entre um valor observado de Y e a linha de previsão. A Equação (13.14) define o resíduo.

RESÍDUO

O resíduo corresponde à diferença entre o valor observado para Y e o valor previsto para Y.

$$e_i = Y_i - \hat{Y}_i \qquad \textbf{(13.14)}$$

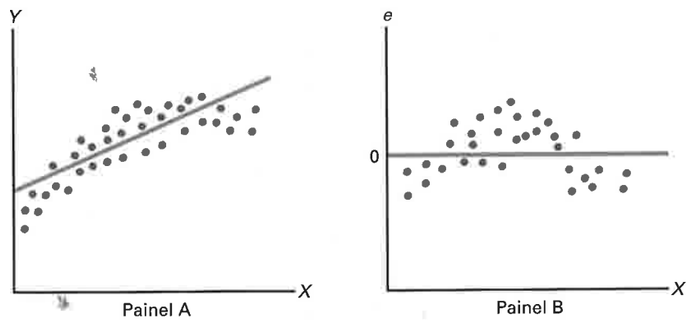
Avaliando os Pressupostos

Lembre-se, com base na Seção 13.4, de que os quatro pressupostos da regressão (conhecidos pelo acrônimo LINI) são linearidade, independência, normalidade e igualdade de variâncias.

Linearidade Para avaliar a linearidade, você insere os resíduos no eixo vertical de um gráfico em contraposição aos valores correspondentes da variável independente, X_i, no eixo horizontal. Caso o modelo linear seja apropriado para os dados, você não verificará nenhum padrão aparente nesse gráfico. No entanto, caso o modelo linear não seja apropriado, existirá, no gráfico de resíduos, uma relação entre os valores de X_i e os resíduos, e_i.

Você pode observar esse tipo de padrão na Figura 13.9. O Painel A ilustra uma situação na qual, embora exista uma tendência crescente em Y à medida que X cresce, a relação aparenta ser curvilínea porque a tendência ascendente diminui para valores crescentes de X. Esse efeito quadrático é realçado no Painel B, no qual existe uma relação clara entre X_i e e_i. Ao se inserirem os resíduos em um gráfico, a tendência linear de X com Y é removida, expondo, consequentemente, a falta de ajuste ao modelo linear simples. Por conseguinte, um modelo quadrático representa um melhor ajuste e deve ser utilizado no lugar do modelo linear simples. (Veja a Seção 15.1 para uma discussão mais detalhada sobre o ajuste de modelos curvilíneos.)

FIGURA 13.9
Estudando a conveniência do modelo de regressão linear simples



Para determinar se o modelo de regressão linear simples é apropriado, retorne à avaliação dos dados relacionados à Sunflowers Roupas. A Figura 13.10 apresenta os valores previstos para as vendas anuais e os respectivos resíduos.

FIGURA 13.10
Tabela de resíduos para os dados relacionados à Sunflowers Roupas

*A Figura 13.10 ilustra a planilha **RESÍDUOS** da pasta de trabalho **Regressão Linear Simples**. Crie essa planilha utilizando as instruções na Seção GE13.5. Se você utilizar o procedimento **Regressão** no suplemento **Ferramentas de Análise**, a **tabela de resíduos** não incluirá a segunda e a quarta colunas e aparecerá em uma área com o nome de RESULTADOS DE RESÍDUOS na planilha principal de resultados da regressão*

	A	B	C	D	E
1	Observação	Área em Pés Quadrados	Vendas Anuais Previstas	Vendas Anuais	Resíduos
2	1	1,7	3,803239598	3,7	0,103239598
3	2	1,6	3,636253367	3,9	-0,263746633
4	3	2,8	5,640088147	6,7	-1,059911853
5	4	5,6	10,31570263	9,5	0,815702635
6	5	1,3	3,135294672	3,4	-0,264705328
7	6	2,2	4,638170757	5,6	-0,961829243
8	7	1,3	3,135294672	3,7	-0,564705328
9	8	1,1	2,801322208	2,7	0,101322208
10	9	3,2	6,308033074	5,5	0,808033074
11	10	1,5	3,469267135	2,9	0,569267135
12	11	5,2	9,647757708	10,7	-1,052242292
13	12	4,6	8,645840318	7,6	1,045840318
14	13	5,8	10,6496751	11,8	-1,150324902
15	14	3,0	5,974060611	4,1	1,874060611

Para avaliar a linearidade, os resíduos são inseridos em um gráfico, em contraposição à variável independente (tamanho da loja, em milhares de pés quadrados), na Figura 13.11. Embora exista uma ampla dispersão no gráfico de resíduos, não existe nenhum padrão ou relação aparente entre os resíduos e X_i. Os resíduos aparentam estar uniformemente dispersos acima e abaixo de 0, para os diferentes valores de X. Você pode concluir que o modelo linear é apropriado para os dados da Sunflowers Roupas.

FIGURA 13.11
Gráfico dos resíduos em contraposição à área em pés quadrados de uma loja para os dados relacionados à Sunflowers Roupas

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

Independência Você pode avaliar o pressuposto da independência de erros desenhando um gráfico de resíduos na ordem ou na sequência em que os dados foram coletados. Caso os valores de Y façam parte de uma série temporal (veja a Seção 2.7), pode ser que um resíduo, algumas vezes, esteja relacionado ao resíduo anterior. Caso exista esse tipo de relação entre resíduos consecutivos (o que viola o pressuposto da independência), o gráfico dos resíduos em contraposição ao período de tempo no qual foram coletados os dados geralmente exibe um padrão cíclico. Uma vez que os dados da Sunflowers Roupas foram coletados durante o mesmo período de tempo, você não precisa avaliar o pressuposto da independência para esses dados.

Normalidade Você pode avaliar o pressuposto da normalidade nos erros organizando os resíduos em uma distribuição de frequências, conforme ilustrado na Tabela 13.3. Você não pode construir

um histograma que tenha algum significado uma vez que o tamanho da amostra é demasiadamente pequeno. E com um tamanho de amostra assim tão pequeno ($n = 14$) pode ser difícil avaliar o pressuposto da normalidade utilizando uma disposição ramo e folha (veja a Seção 2.6), um boxplot (veja a Seção 3.3) ou um gráfico da probabilidade normal (veja a Seção 6.3).

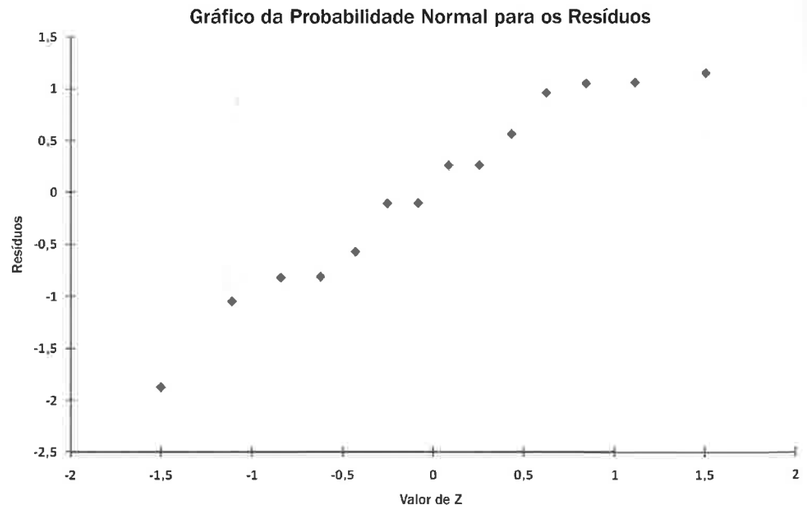
TABELA 13.3
Distribuição de Frequências de 14 Valores de Resíduos para os Dados da Sunflowers Roupas

Resíduos	Frequência
−2,25 porém menos que −1,75	1
−1,75 porém menos que −1,25	0
−1,25 porém menos que −0,75	3
−0,75 porém menos que −0,25	1
−0,25 porém menos que +0,25	2
+0,25 porém menos que +0,75	3
+0,75 porém menos que +1,25	4
	14

Diante do exposto, com base no gráfico da probabilidade normal para os resíduos na Figura 13.12, os dados não aparentam se distanciar substancialmente de uma distribuição normal. A robustez da análise de regressão, no que diz respeito a afastamentos moderados em relação à normalidade, possibilita que você conclua que não há necessidade de se preocupar demasiadamente com afastamentos do pressuposto da normalidade no que concerne aos dados relacionados à Sunflowers Roupas.

FIGURA 13.12
Gráfico da probabilidade normal para os resíduos dos dados relacionados à Sunflowers Roupas

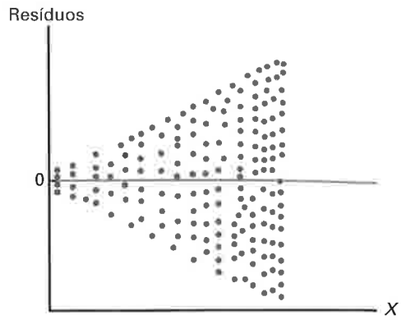
Crie gráficos da probabilidade normal utilizando as instruções na Seção GE6.3.



Igualdade de Variâncias Você pode avaliar o pressuposto da igualdade de variâncias a partir de um gráfico de resíduos em relação a X_i. Para os dados da Sunflowers Roupas, na Figura 13.11, não parecem existir diferenças significativas em termos da variabilidade nos resíduos para diferentes valores de X_i. Consequentemente, você pode concluir que não existe nenhuma violação aparente no pressuposto da igualdade de variâncias em relação a cada um dos níveis de X.

Para examinar um caso em que o pressuposto da igualdade de variâncias tenha sido violado, observe a Figura 13.13, que corresponde a um gráfico de resíduos em relação a X_i para um conjunto hipotético de dados. Nesse gráfico, a variabilidade dos resíduos aumenta drasticamente à medida que X cresce, demonstrando a falta de homogeneidade nas variâncias de Y_i em cada um dos níveis de X. No que concerne a esses dados, o pressuposto de igualdade de variâncias é inválido.

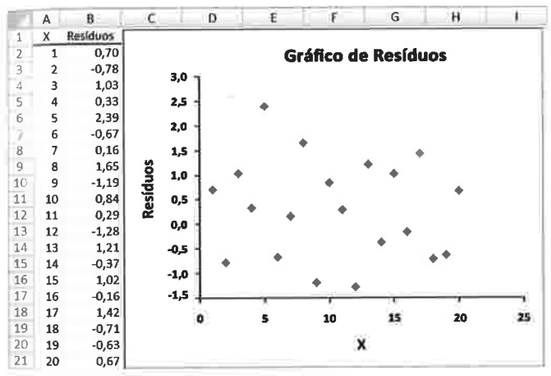
FIGURA 13.13
Violação do pressuposto da igualdade de variâncias



Problemas para a Seção 13.5

APRENDENDO O BÁSICO

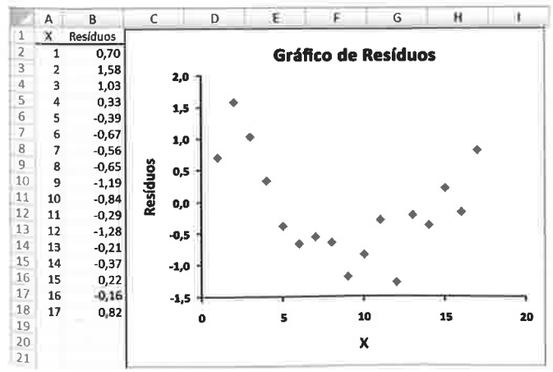
13.23 Os resultados a seguir apresentam os valores de X, os resíduos e um gráfico de resíduos a partir de uma análise de regressão:



	X	Resíduos
1	1	0,70
2	2	-0,78
3	3	1,03
4	4	0,33
5	5	2,39
6	6	-0,67
7	7	0,16
8	8	1,65
9	9	-1,19
10	10	0,84
11	11	0,29
12	12	-1,28
13	13	1,21
14	14	-0,37
15	15	1,02
16	16	-0,16
17	17	1,42
18	18	-0,71
19	19	-0,63
20	20	0,67

Existe alguma evidência de um padrão nos resíduos? Explique.

13.24 Os resultados a seguir apresentam os valores de X, os resíduos e um gráfico de resíduos a partir de uma análise de regressão:



	X	Resíduos
1	1	0,70
2	2	1,58
3	3	1,03
4	4	0,33
5	5	-0,39
6	6	-0,67
7	7	-0,56
8	8	-0,65
9	9	-1,19
10	10	-0,84
11	11	-0,29
12	12	-1,28
13	13	-0,21
14	14	-0,37
15	15	0,22
16	16	-0,16
17	17	0,82

Existe alguma evidência de um padrão nos resíduos? Explique.

APLICANDO OS CONCEITOS

13.25 No Problema 13.5, em Problemas para a Seção 13.2, você utilizou as vendas informadas de revistas em bancas no intuito de prever as vendas auditadas. Realize uma análise dos resíduos em relação a esses dados (armazenados no arquivo **Circulação**). Avalie se os pressupostos para a regressão foram, ou não, seriamente violados.

AUTO-AVALIAÇÃO **13.26** No Problema 13.4, em Problemas para a Seção 13.2, o gerente de marketing utilizou o espaço na prateleira destinado a rações para animais domésticos para prever as vendas semanais. Realize uma análise dos resíduos em relação a esses dados (armazenados no arquivo **Ração**). Avalie se os pressupostos para a regressão foram, ou não, seriamente violados.

13.27 No Problema 13.7, em Problemas para a Seção 13.2, você utilizou o número total de clientes para prever o tempo de espera na fila do caixa de pagamento de compras em um supermercado. Realize uma análise dos resíduos para esses dados (armazenados no arquivo **Supermercado**). Com base nesses resultados, avalie se os pressupostos para a regressão foram, ou não, seriamente violados.

13.28 No Problema 13.6, em Problemas para a Seção 13.2, o proprietário de uma empresa de mudanças desejava prever as horas trabalhadas com base no volume de pés cúbicos a serem transportados. Realize uma análise dos resíduos para esses dados (armazenados no arquivo **Mudança**). Com base nesses resultados, avalie se os pressupostos para a regressão foram, ou não, seriamente violados.

13.29 No Problema 13.9, em Problemas para a Seção 13.2, um corretor de uma imobiliária desejava prever o aluguel mensal para apartamentos com base no tamanho desses imóveis. Realize uma análise de resíduos para esses dados (armazenados no arquivo **Aluguel**). Com base nesses resultados, avalie se os pressupostos para a regressão foram, ou não, seriamente violados.

13.30 No Problema 13.8, em Problemas para a Seção 13.2, você utilizou as receitas anuais para prever o valor de uma franquia de beisebol. Realize uma análise de resíduos para esses dados (armazenados no arquivo **BBReceita**). Com base nesses resultados, avalie se os pressupostos para a regressão foram, ou não, seriamente violados.

13.31 No Problema 13.10, em Problemas para a Seção 13.2, você utilizou a receita bruta auferida na bilheteria dos cinemas para prever as vendas de DVDs. Realize uma análise de resíduos para esses dados (armazenados no arquivo **Cinema**). Com base nesses resultados, avalie se os pressupostos para a regressão foram, ou não, seriamente violados.

13.6 Medindo a Autocorrelação: A Estatística de Durbin-Watson

Um dos pressupostos básicos do modelo de regressão é a independência dos erros. Esse pressuposto é às vezes violado quando os dados são coletados ao longo de períodos sequenciais de tempo, uma vez que um resíduo, em qualquer ponto individual no tempo, pode tender a ser semelhante a resíduos em pontos adjacentes no tempo. Esse tipo de padrão nos resíduos é conhecido como **autocorrelação**. Quando um determinado conjunto de dados apresenta um volume substancial de autocorrelação, a validação de um modelo de regressão ajustado passa a ficar sob sérias dúvidas.

Gráficos de Resíduos para Detectar Autocorrelação

Conforme mencionado na Seção 13.5, um dos meios de detectar a autocorrelação é colocar os resíduos em um gráfico na ordem sequencial de tempo. Caso esteja presente um efeito de uma autocorrelação positiva, existirão concentrações de resíduos com o mesmo sinal, e você prontamente detectará um padrão aparente. Caso exista uma autocorrelação negativa, os resíduos tenderão a pular para a frente e para trás, de positivo para negativo e novamente para positivo e assim sucessivamente. Esse tipo de padrão é observado muito raramente na análise da regressão. Por conseguinte, o foco desta seção será a autocorrelação positiva. Para ilustrar a autocorrelação positiva, considere o exemplo apresentado a seguir.

A gerente de uma loja de remessas de encomendas se depara com o problema estratégico de prever as vendas semanais. Ao abordar esse problema, ela decidiu desenvolver um modelo de regressão para utilizar o número de clientes que fizeram compras como a variável independente. Foram coletados dados ao longo de um período de 15 semanas. A Tabela 13.4 organiza os dados (armazenados em **ClienteVendas**).

TABELA 13.4
Clientes e Vendas para um Período de 15 Semanas Consecutivas

Semana	Clientes	Vendas (Milhares de Dólares)	Semana	Clientes	Vendas (Milhares de Dólares)
1	794	9,33	9	880	12,07
2	799	8,26	10	905	12,55
3	837	7,48	11	886	11,92
4	855	9,08	12	843	10,27
5	845	9,83	13	904	11,80
6	844	10,09	14	950	12,15
7	863	11,01	15	841	9,64
8	875	11,49			

Uma vez que os dados foram coletados ao longo de um período de 15 semanas consecutivas na mesma loja, você precisa determinar se está presente uma autocorrelação. A Figura 13.14 ilustra a planilha com os resultados para esses dados.

FIGURA 13.14
Planilha de resultados da regressão para os dados da Tabela 13.4 sobre a loja de remessa de encomendas

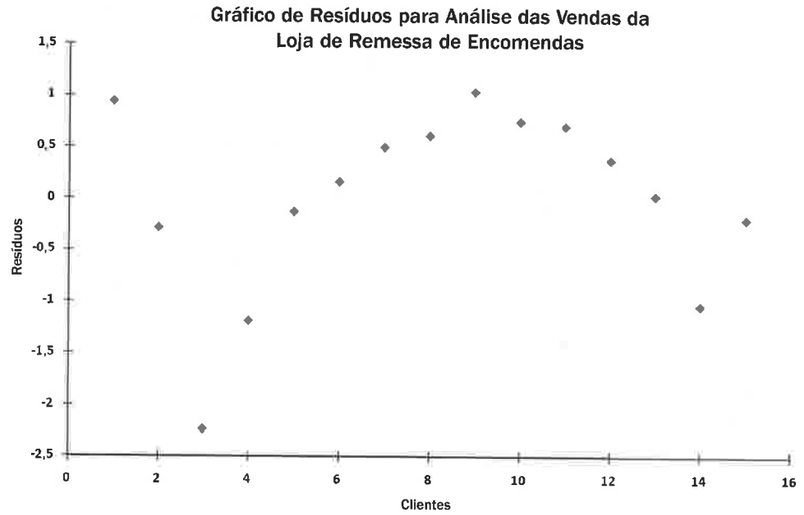
Crie modelos lineares simples utilizando as instruções da Seção GE13.2.

	A	B	C	D	E	F	G
1	Análise das Vendas da Loja de Remessa de Encomendas						
2							
3	*Estatística de Regressão*						
4	R Múltiplo	0,8108					
5	R-quadrado	0,6574					
6	R-quadrado ajustado	0,6311					
7	Erro-padrão	0,9360					
8	Observações	15					
9							
10	ANOVA						
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*	
12	Regressão	1	21,8604	21,8604	24,9501	0,0002	
13	Resíduos	13	11,3901	0,8762			
14	Total	14	33,2506				
15							
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*
17	Interseção	-16,0322	5,3102	-3,0192	0,0099	-27,5041	-4,5603
18	Área em Pés Quadrados	0,0308	0,0062	4,9950	0,0002	0,0175	0,0441

Com base na Figura 13.14, observe que r^2 é igual a 0,6574, indicando que 65,74% da variação nas vendas pode ser explicada pela variação no número de clientes. Além disso, o intercepto de Y, b_0, é –16,0322, e a inclinação, b_1, é 0,0308. No entanto, antes de utilizar esse modelo para fins de previsão, você deve realizar uma análise nos resíduos. Uma vez que os dados foram coletados ao longo de um período consecutivo de 15 semanas, além de verificar os pressupostos de linearidade, normalidade e igualdade de variâncias, você deve necessariamente investigar o pressuposto da independência de erros. Para fazer isso, você insere em um gráfico os resíduos em relação ao tempo, como na Figura 13.15, para que isso o ajude a verificar se existe um padrão. Na Figura 13.15, você pode verificar que os resíduos tendem a flutuar para cima e para baixo, em um padrão cíclico. Esse padrão cíclico representa uma forte causa de preocupação no que concerne à autocorrelação nos resíduos e, consequentemente, uma violação do pressuposto da independência dos erros.

FIGURA 13.15
Gráfico de resíduos para os dados da Tabela 13.4 sobre a loja de remessa de encomendas

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



A Estatística de Durbin-Watson

A **estatística de Durbin-Watson** é utilizada para medir a autocorrelação. Essa estatística mede a correlação entre cada um dos resíduos e o resíduo correspondente ao período de tempo imediatamente anterior. A Equação (13.15) define a estatística de Durbin-Watson.

ESTATÍSTICA DE DURBIN-WATSON

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad \textbf{(13.15)}$$

em que

e_i = resíduo para o período de tempo i

Para melhor compreender a estatística de Durbin-Watson, D, você pode examinar a Equação (13.15). O numerador, $\sum_{i=2}^{n}(e_i - e_{i-1})^2$, representa a diferença, elevada ao quadrado, entre dois resíduos sucessivos, somados desde o segundo valor até o n-ésimo valor. O denominador, $\sum_{i=1}^{n} e_i^2$, representa a soma dos resíduos (erros) elevados ao quadrado. Quando resíduos sucessivos são positivamente autocorrelacionados, o valor de D se aproxima de 0. Se os resíduos não forem cor-

relacionados, o valor de D estará próximo de 2. (Se existir uma autocorrelação negativa, D será maior do que 2 e pode, até mesmo, se aproximar de seu valor máximo, que é 4.) Para os dados sobre a loja de entrega de encomendas, os resultados da planilha da Figura 13.16 mostram que a estatística de Durbin-Watson, D, é igual a 0,8830.

A Figura 13.16 ilustra uma planilha semelhante à ***planilha DURBIN_WATSON*** *da pasta* ***de trabalho Regressão Linear Simples****. Crie planilhas que calculam a estatística de Durbin-Watson utilizando as instruções na Seção GE13.6.*

FIGURA 13.16
Planilha da estatística de Durbin-Watson para os dados sobre a loja de remessa de encomendas

	A	B	
1	Estatística de Durbin-Watson		
2			
3	Soma das Diferenças ao Quadrado dos Resíduos	10,0575	=SOMAXMY2(RESÍDUOS!E3:E16,RESÍDUOS!E2:E15)
4	Soma dos Quadrados dos Resíduos	11,3901	=SOMAQUAD(RESÍDUOS!E2:E16)
5			
6	Estatística de Durbin-Watson	0,8830	=B3/B4

Você precisa determinar as situações nas quais a autocorrelação é grande o suficiente para que se conclua que existe uma correlação positiva. Depois de calcular D, você compara esse valor aos valores críticos da estatística de Durbin-Watson encontrados na Tabela E.8, parte da qual é apresentada na Tabela 13.15. Os valores críticos dependem de α, o nível de significância escolhido, e de n, o tamanho da amostra, e de k, o número de variáveis independentes no modelo (na regressão linear simples $k = 1$).

TABELA 13.5
Encontrando Valores Críticos da Estatística de Durbin-Watson

	$\alpha = ,05$									
	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	d_I	d_S	d_I	d_S	d_I	d_S	d_I	d_S	d_I	d_S
15	1,08	1,36	,95	1,54	,82	1,75	,69	1,97	,56	2,21
16	1,10	1,37	,98	1,54	,86	1,73	,74	1,93	,62	2,15
17	1,13	1,38	1,02	1,54	,90	1,71	,78	1,90	,67	2,10
18	1,16	1,39	1,05	1,53	,93	1,69	,82	1,87	,71	2,06

Na Tabela 13.5, dois valores são exibidos para cada uma das combinações entre α (nível de significância), n (tamanho da amostra) e k (número de variáveis independentes no modelo). O primeiro valor, d_I, representa o valor crítico inferior. Se D estiver abaixo de d_I, você conclui que existem evidências de autocorrelação positiva entre os resíduos. Caso isso ocorra, o método dos mínimos quadrados, utilizado neste capítulo, não é apropriado, e você deve utilizar métodos alternativos (veja a referência 4). O segundo valor, d_S, representa o valor crítico superior de D, acima do qual você concluiria que não existe nenhuma evidência de autocorrelação positiva entre os resíduos. Caso D se posicione entre d_I e d_S, você fica impossibilitado de chegar a uma conclusão definitiva.

No que diz respeito aos dados relacionados à loja de remessa de encomendas, com uma variável independente ($k = 1$) e 15 valores ($n = 15$), $d_I = 1{,}08$ e $d_S = 1{,}36$. Uma vez que $D = 0{,}8830 < 1{,}08$, você conclui que existe autocorrelação positiva entre os resíduos. A análise de regressão dos mínimos quadrados para esses dados não é apropriada em razão da presença de uma autocorrelação positiva significativa entre os resíduos. Em outras palavras, o pressuposto da independência de erros não é válido. Você precisa utilizar as abordagens alternativas discutidas na referência 4.

Problemas para a Seção 13.6

APRENDENDO O BÁSICO

13.32 Os resíduos correspondentes a 10 períodos de tempo consecutivos são os seguintes:

Período de Tempo	Resíduo	Período de Tempo	Resíduo
1	−5	6	+1
2	−4	7	+2
3	−3	8	+3
4	−2	9	+4
5	−1	10	+5

a. Elabore um gráfico para os resíduos em relação ao tempo. A que conclusões você consegue chegar quanto ao padrão para os resíduos em relação ao tempo?
b. Com base em (a), a que conclusões você consegue chegar sobre a autocorrelação dos resíduos?

13.33 Os resíduos correspondentes a 15 períodos de tempo consecutivos são os seguintes:

Período de Tempo	Resíduo	Período de Tempo	Resíduo
1	+4	9	+6
2	−6	10	−3
3	−1	11	+1
4	−5	12	+3
5	+2	13	0
6	+5	14	−4
7	−2	15	−7
8	+7		

a. Elabore um gráfico para os resíduos em relação ao tempo. A que conclusões você consegue chegar quanto ao padrão para os resíduos em relação ao tempo?
b. Calcule a estatística de Durbin-Watson. No nível de significância de 0,05, existem evidências de autocorrelação positiva entre os resíduos?
c. Com base em (a) e (b), a que conclusões você consegue chegar sobre a autocorrelação dos resíduos?

APLICANDO OS CONCEITOS

13.34 No Problema 13.4, em Problemas para a Seção 13.2, que trata da venda de ração para animais de estimação, o gerente de marketing utilizou o espaço na prateleira destinado a rações para animais de estimação com o objetivo de prever as vendas semanais.
a. É necessário calcular a estatística de Durbin-Watson no presente caso? Explique.
b. Sob quais circunstâncias seria necessário calcular a estatística de Durbin-Watson antes de dar prosseguimento ao método dos mínimos quadrados da análise da regressão?

13.35 Qual é a relação entre o preço do óleo bruto e o preço que você paga pela gasolina na bomba? O arquivo **Petróleo & Gasolina** contém o preço correspondente a um barril de petróleo cru e um galão de gasolina durante 100 semanas, encerrando em 1.º de junho de 2009 (dados extraídos de Energy Information Administration, U.S. Department of Energy, **www.eia.doe.gov**).
a. Construa um gráfico de dispersão com petróleo no eixo horizontal e gasolina no eixo vertical.
b. Utilize o método dos mínimos quadrados para desenvolver uma equação para a regressão linear simples, de modo a prever o preço de um galão de gasolina, utilizando o preço de um barril de petróleo cru como a variável independente.
c. Interprete o significado da inclinação, b_1, neste problema.
d. Desenhe o gráfico dos resíduos em relação ao período de tempo.
e. Calcule a estatística de Durbin-Watson.
f. No nível de significância de 0,05, existem evidências de autocorrelação positiva entre os resíduos?
g. Com base nos resultados de (d) a (f), existe alguma razão para questionar a validade desse modelo?

AUTO-AVALIAÇÃO **13.36** Uma empresa de venda por pedidos de catálogo, e que vende componentes de informática, software e hardware, mantém um depósito centralizado para a distribuição dos produtos encomendados. A direção da empresa está atualmente examinando o processo de distribuição a partir do depósito, e está interessada em estudar os fatores que afetam os custos de distribuição do depósito. Atualmente, uma pequena tarifa de frete está sendo acrescentada ao pedido, independentemente do valor do pedido de compra. Ao longo dos últimos 24 meses, foram coletados dados que indicam os custos de distribuição para o depósito e o número de pedidos de compra recebidos. Esses dados estão armazenados no arquivo **CustoDepósito**. Os resultados estão ilustrados na tabela a seguir:

Meses	Causa de Distribuição (Milhares de Dólares)	Número de Pedidos de Compra
1	52,95	4.015
2	71,66	3.806
3	85,58	5.309
4	63,69	4.262
5	72,81	4.296
6	68,44	4.097
7	52,46	3.213
8	70,77	4.809
9	82,03	5.237
10	74,39	4.732
11	70,84	4.413
12	54,08	2.921
13	62,98	3.977
14	72,30	4.428
15	58,99	3.964
16	79,38	4.582
17	94,44	5.582
18	59,74	3.450
19	90,50	5.079
20	93,24	5.735
21	69,33	4.269
22	53,71	3.708
23	89,18	5.387
24	66,80	4.161

a. Pressupondo uma relação linear, utilize o método dos mínimos quadrados para encontrar os coeficientes de regressão, b_0 e b_1.
b. Faça a previsão dos custos mensais de distribuição do depósito quando o número de pedidos é de 4.500.
c. Faça o gráfico dos resíduos em relação ao período de tempo.
d. Calcule a estatística de Durbin-Watson. No nível de significância de 0,05, existem evidências de autocorrelação positiva entre os resíduos?
e. Com base nos resultados de (c) e (d), existe alguma razão para questionar a validade do modelo?

13.37 Uma dose de café expresso feito na hora possui três componentes distintos: o núcleo, o corpo e a espuma. A separação desses três componentes geralmente dura somente de 10 a 20 segundos. Para utilizar a dose de expresso para fazer um café com leite, um *cappuccino* ou alguma outra bebida, a dose deve ser despejada na bebida durante o processo de separação do núcleo, do corpo e da espuma. Caso a dose seja utilizada depois de ocorrida a separação, a bebida torna-se excessivamente amarga e ácida, estragando o sabor final. Consequentemente, um maior tempo de separação proporciona à pessoa que está preparando a bebida maior quantidade de tempo para despejar a dose e garantir que a bebida atenderá às expectativas. Um empregado em uma cafeteria levantou a hipótese de que, quanto maior a pressão dos grãos do café expresso no porta-filtro antes de serem fervidos, maior seria o tempo de separação. Foi conduzido um experimento utilizando 24 observações para testar essa relação. A variável independente, Compressão, mede a distância, em polegadas, entre os grãos do café expresso e o topo do porta-filtro (ou seja, quanto maior a compressão, maior a distância). A variável dependente, Tempo, corresponde ao número de segundos em que o núcleo, o corpo e a espuma são separados (ou seja, a quantidade de tempo depois que a dose é despejada antes que seja utilizada para a bebida do cliente). Os dados estão armazenados no arquivo **Expresso**, e estão ilustrados na tabela a seguir:

Dose	Compressão (Polegadas)	Tempo (Segundos)	Dose	Compressão (Polegadas)	Tempo (Segundos)
1	0,20	14	13	0,50	18
2	0,50	14	14	0,50	13
3	0,50	18	15	0,35	19
4	0,20	16	16	0,35	19
5	0,20	16	17	0,20	17
6	0,50	13	18	0,20	18
7	0,20	12	19	0,20	15
8	0,35	15	20	0,20	16
9	0,50	9	21	0,35	18
10	0,35	15	22	0,35	16
11	0,50	11	23	0,35	14
12	0,50	16	24	0,35	16

a. Utilize o método dos mínimos quadrados para desenvolver uma equação para a regressão linear simples, utilizando Tempo como a variável dependente e Compressão como a variável independente.
b. Faça a previsão para o tempo de separação, para uma distância de Compressão de 0,50 polegada.
c. Faça o gráfico dos resíduos em relação à ordem de tempo do experimento. Existe algum padrão que possa ser observado?
d. Calcule a estatística de Durbin-Watson. No nível de significância de 0,05, existem evidências de autocorrelação positiva entre os resíduos?
e. Com base nos resultados de (c) e (d), existe alguma razão para questionar a validade do modelo?

13.38 O proprietário de uma cadeia de sorveterias tem como objetivo estratégico da empresa aperfeiçoar o mecanismo de previsão de vendas diárias de modo tal que possa vir a ser minimizada a escassez de mão de obra durante a temporada do verão. O proprietário decidiu começar desenvolvendo um modelo de regressão linear para prever as vendas com base na temperatura atmosférica. É selecionada uma amostra com 21 dias consecutivos, e os resultados estão armazenados no arquivo **Sorvete**.

(Dica: Determine quais são as variáveis independente e dependente.)

a. Pressupondo uma relação linear, utilize o método dos mínimos quadrados para encontrar os coeficientes de regressão, b_0 e b_1.
b. Faça a previsão para as vendas em um dia no qual a temperatura corresponda a 83°F.
c. Faça o gráfico dos resíduos em relação ao período de tempo.
d. Calcule a estatística de Durbin-Watson. No nível de significância de 0,05, existem evidências de autocorrelação positiva entre os resíduos?
e. Com base nos resultados de (c) e (d), existe alguma razão para se questionar a validade desse modelo?

13.7 Inferências sobre a Inclinação e o Coeficiente de Correlação

Nas Seções 13.1 até 13.3, a regressão foi utilizada exclusivamente para fins de descrição. Você aprendeu como o método dos mínimos quadrados determina os coeficientes da regressão e aprendeu a prever o valor de Y para um determinado valor de X. Além disso, você aprendeu a calcular e interpretar o erro-padrão da estimativa e o coeficiente de determinação.

Quando a análise dos resíduos, conforme discutido na Seção 13.5, indica que os pressupostos de um modelo de regressão dos mínimos quadrados não estão sendo seriamente violados e que o modelo correspondente à linha reta é apropriado, você pode realizar inferências sobre a relação linear entre as variáveis na população.

Teste *t* para a Inclinação

Para determinar a existência de uma relação linear significativa entre as variáveis X e Y, você testa se β_1 (a inclinação da população) é igual a 0. A hipótese nula e a hipótese alternativa se apresentam como:

H_0: $\beta_1 = 0$ [Não existe nenhuma relação linear (a inclinação é zero).]

H_1: $\beta_1 \neq 0$ [Existe uma relação linear (a inclinação não é igual a zero).]

Caso rejeite a hipótese nula, você conclui que existem evidências de uma relação linear. A Equação (13.16) define a estatística do teste.

TESTANDO UMA HIPÓTESE PARA A INCLINAÇÃO DE UMA POPULAÇÃO, β_1, UTILIZANDO O TESTE *t*

A estatística do teste t_{ESTAT} é igual à diferença entre a inclinação da amostra e o valor citado na hipótese para a inclinação da população, dividida pelo erro-padrão da inclinação.

$$t_{ESTAT} = \frac{b_1 - \beta_1}{S_{b_1}} \qquad \textbf{(13.16)}$$

em que

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SQX}}$$

$$SQX = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

A estatística do teste t_{ESTAT} segue uma distribuição t com $n - 2$ graus de liberdade.

Retorne ao cenário que trata da Sunflowers Roupas, apresentado no início do capítulo. Para testar se existe uma relação linear significativa entre o tamanho da loja e as vendas anuais, no nível de significância de 0,05, reporte-se aos resultados para o teste t apresentados na planilha da Figura 13.17.

FIGURA 13.17
Planilha com os resultados do teste *t* para a inclinação dos dados relacionados à Sunflowers Roupas

Na Figura 13.17, a estatística do teste t_{ESTAT} aparece na célula D18 e está com a legenda Stat t.

	A	B	C	D	E	F	G	H	I
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*	*95,0% inferiores*	*95,0% superiores*
17	Interseção	0,9645	0,5262	1,8329	0,0917	-0,1820	2,1110	-0,1820	2,11095
18	Área em Pés Quadrados	1,6699	0,1569	10,6411	0,0000	1,3280	2,0118	1,3280	2,01177

Com base na Figura 13.17,

$$b_1 = +1{,}6699 \quad n = 14 \quad S_{b_1} = 0{,}1569$$

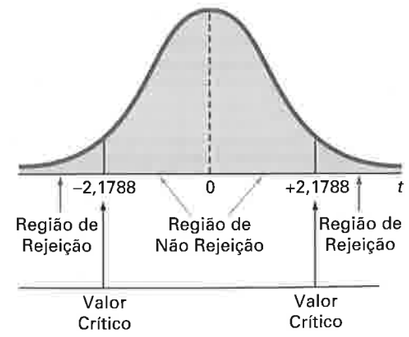
e

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

$$= \frac{1{,}6699 - 0}{0{,}1569} = 10{,}6411$$

Utilizando o nível de significância de 0,05, o valor crítico de t, com $n - 2 = 12$ graus de liberdade, é 2,1788. Uma vez que $t_{ESTAT} = 10{,}6411 > 2{,}1788$, ou porque o valor-$p$ é aproximadamente igual a 0, que é menor do que $\alpha = 0{,}05$ é aproximadamente 0, você rejeita H_0 (veja a Figura 13.18). Consequentemente, você pode concluir que existe uma relação linear significativa entre a média aritmética das vendas anuais e o tamanho da loja.

FIGURA 13.18
Testando uma hipótese sobre a inclinação da população, no nível de significância de 0,05, com 12 graus de liberdade



Teste *F* para a Inclinação

Como uma alternativa ao teste *t*, você pode utilizar um teste *F* para determinar se a inclinação na regressão linear simples é estatisticamente significativa. Na Seção 10.4, você utilizou a distribuição *F* para testar a proporcionalidade entre duas variâncias. A Equação (13.17) define o teste *F* para a inclinação como a proporcionalidade entre a variância que é devida à regressão (*MQReg*) dividida pela variância do erro ($MQR = S_{YX}^2$).

TESTANDO UMA HIPÓTESE PARA A INCLINAÇÃO DE UMA POPULAÇÃO, β_1, UTILIZANDO O TESTE *F*

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*).

$$F_{ESTAT} = \frac{MQReg}{MQR} \qquad \textbf{(13.17)}$$

em que

$$MQReg = \frac{SQReg}{1} = SQReg$$

$$MQR = \frac{SQR}{n-2}$$

A estatística do teste F_{ESTAT} segue uma distribuição *F*, com 1 e $n - 2$ graus de liberdade.

Utilizando o nível de significância α, a regra de decisão é

Rejeitar H_0, se $F > F\alpha$;

caso contrário, não rejeitar H_0.

A Tabela 13.6 organiza o conjunto completo de resultados em uma tabela de análise da variância (ANOVA).

TABELA 13.6
Tabela de ANOVA para Testar a Significância de um Coeficiente de Regressão

Fonte	*gl*	Soma dos Quadrados	Média dos Quadrados (Variância)	*F*
Regressão	1	*SQReg*	$MQReg = \frac{SQReg}{1} = SQReg$	$F_{ESTAT} = \frac{MQReg}{MQR}$
Erro	$n - 2$	*SQR*	$MQR = \frac{SQR}{n-2}$	
Total	$n - 1$	*STQ*		

A Figura 13.19, uma tabela completa de ANOVA para os dados da Sunflowers, mostra que a estatística F_{ESTAT} calculada é 113,2335, e o valor-*p* é aproximadamente igual a 0.

FIGURA 13.19
Planilha com os resultados do teste *F* para os dados relacionados à Sunflowers Roupas

	A	B	C	D	E	F	G	H	I
10	**ANOVA**								
11		*gl*	*SQ*	*MQ*	*F*	*F de significação*			
12	Regressão	1	105,7476	105,748	113,234	0,0000			
13	Resíduos	12	11,2067	0,9339					
14	Total	13	116,9543						
15									
16		*Coeficientes*	*Erro-padrão*	*Stat t*	*Valor-p*	*95% inferiores*	*95% superiores*	*95,0% inferiores*	*95,0% superiores*
17	Interseção	0,9645	0,5262	1,8329	0,0917	-0,1820	2,1110	-0,1820	2,11095
18	Área em Pés Quadrados	1,6699	0,1569	10,6411	0,0000	1,3280	2,0118	1,3280	2,01177

Utilizando um nível de significância de 0,05, com base na Tabela E.5, o valor crítico para a distribuição *F*, com 1 e 12 graus de liberdade, é 4,75 (veja a Figura 13.20). Uma vez que $F_{ESTAT} = 113,2335 > 4,75$, ou pelo fato de que o valor-$p = 0,0000 < 0,05$, você rejeita H_0 e conclui que o

tamanho da loja é significativamente relacionado às vendas anuais. Tendo em vista que o teste *F*, na Equação (13.17), é equivalente ao teste *t* na Equação (13.16), você chega à mesma conclusão.

FIGURA 13.20
Regiões de rejeição e de não rejeição ao testar a significância da inclinação, no nível de significância de 0,05, com 1 e 12 graus de liberdade

Estimativa do Intervalo de Confiança para a Inclinação

Como uma alternativa para testar a existência de uma relação linear entre as variáveis, você pode construir uma estimativa do intervalo de confiança para β_1 utilizando a Equação (13.18).

ESTIMATIVA DO INTERVALO DE CONFIANÇA PARA A INCLINAÇÃO, β_1

A estimativa do intervalo de confiança para a inclinação pode ser construída tomando-se a inclinação da amostra, b_1, e adicionando-se e subtraindo-se o valor crítico de *t*, multiplicado pelo erro-padrão da inclinação.

$$b_1 \pm t_{\alpha/2} S_{b_1} \qquad \textbf{(13.18)}$$

em que

$t_{\alpha/2}$ é o valor crítico correspondente a uma probabilidade de cauda superior igual a $\alpha/2$ a partir da distribuição *t* com $n - 2$ graus de liberdade (ou seja, uma área acumulada de $1 - \alpha/2$).

Com base os resultados da planilha apresentada na Figura 13.17,

$$b_1 = 1,6699 \quad n = 14 \quad S_{b_1} = 0,1569$$

Para construir uma estimativa do intervalo de confiança de 95%, $\alpha/2 = 0,025$, e, com base na Tabela E.3, $t_{12} = 2,1788$. Portanto,

$$b_1 \pm t_{\alpha/2} S_{b_1} = 1,6699 \pm (2,1788)(0,1569)$$

$$= 1,6699 \pm 0,3419$$

$$1,3280 \leq \beta_1 \leq 2,0118$$

Por conseguinte, você estima, com 95% de confiança, que a inclinação se posiciona entre 1,3280 e 2,0118. Uma vez que esses valores estão acima de 0, você conclui que existe uma relação linear significativa entre as vendas anuais e o tamanho da loja. Caso o intervalo tivesse incluído o valor de 0, você teria concluído que não existe nenhuma relação significativa entre as variáveis. O intervalo de confiança indica que, para cada crescimento equivalente a 1.000 pés quadrados, estima-se que as vendas anuais aumentem em pelo menos \$1.328.000, porém não mais do que \$2.011.800.

Teste *t* para o Coeficiente de Correlação

Na Seção 3.5, no Capítulo 3, foi medida a força da relação entre duas variáveis numéricas utilizando-se o **coeficiente de correlação**, *r*. Os valores para o coeficiente de correlação se estendem desde –1 para uma correlação negativa perfeita até +1 para uma correlação positiva perfeita. Você pode utilizar o coeficiente de correlação para determinar se existe uma relação linear estatisticamente significativa entre *X* e *Y*. Para isso, você formula a hipótese de que o coeficiente de correlação da população, ρ, é igual a 0. Consequentemente, a hipótese nula e a hipótese alternativa são:

$$H_0: \rho = 0 \text{ (nenhuma correlação)}$$

$$H_1: \rho \neq 0 \text{ (correlação)}$$

A Equação (13.19) define a estatística de teste para determinar a existência de uma correlação significativa.

TESTANDO A EXISTÊNCIA DE CORRELAÇÃO

$$t_{ESTAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} \quad \textbf{(13.19a)}$$

em que

$$r = +\sqrt{r^2} \quad \text{se} \quad b_1 > 0$$
$$r = -\sqrt{r^2} \quad \text{se} \quad b_1 < 0$$

A estatística do teste t_{ESTAT} segue uma distribuição t com $n - 2$ graus de liberdade. r é calculado do seguinte modo:

$$r = \frac{\text{cov}(X,Y)}{S_X S_Y} \quad \textbf{(13.19b)}$$

em que

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1}}$$

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

No problema que trata da Sunflowers Roupas, $r^2 = 0{,}9042$ e $b_1 = +1{,}6699$ (veja a Figura 13.4). Uma vez que $b_1 > 0$, o coeficiente de correlação para vendas anuais e tamanho da loja corresponde à raiz quadrada positiva de r^2, ou seja, $r = +\sqrt{0{,}9042} = +0{,}9509$. O uso da Equação (13.19a), para testar a hipótese nula de que não existe nenhuma correlação entre essas duas variáveis resulta na seguinte estatística t observada:

$$t_{ESTAT} = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

$$= \frac{0{,}9509 - 0}{\sqrt{\frac{1 - (0{,}9509)^2}{14 - 2}}} = 10{,}6411$$

Utilizando um nível de significância de 0,05, pelo fato de que $t_{ESTAT} = 10{,}6411 > 2{,}1788$, você rejeita H_0. Você conclui que existe uma associação significativa entre vendas anuais e tamanho da loja. Essa estatística do teste t_{ESTAT} é equivalente à estatística do teste t_{ESTAT} encontrada ao testar se a inclinação da população, β_1, é igual a zero.

Problemas para a Seção 13.7

APRENDENDO O BÁSICO

13.39 Você está testando a hipótese nula de que não existe nenhuma relação linear entre duas variáveis, X e Y. Com base em sua amostra de $n = 10$, você determina que $r = 0{,}80$.
a. Qual é o valor da estatística do teste, t_{ESTAT}?
b. No nível de significância $\alpha = 0{,}05$, quais são os valores críticos?
c. Com base em suas respostas para (a) e (b), qual decisão estatística você deve tomar?

13.40 Você está testando a hipótese nula de que não existe nenhuma relação entre duas variáveis, X e Y. Com base em sua amostra de $n = 18$, você determina que $b_1 = +4{,}5$ e $S_{b1} = 1{,}5$.
a. Qual é o valor de t_{ESTAT}?
b. No nível de significância $\alpha = 0{,}05$, quais são os valores críticos?
c. Com base em suas respostas para (a) e (b), qual decisão estatística você deve tomar?
d. Construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população, β_1.

13.41 Você está testando a hipótese nula de que não existe nenhuma relação entre duas variáveis, X e Y. Com base em sua amostra de $n = 20$, você determina que $SQReg = 60$ e $SQR = 40$.
a. Qual é o valor de F_{ESTAT}?
b. No nível de significância $\alpha = 0{,}05$, qual é o valor crítico?
c. Com base em suas respostas para (a) e (b), qual decisão estatística você deve tomar?
d. Calcule o coeficiente de correlação calculando inicialmente r^2 e pressupondo que b_1 seja negativa.
e. No nível de significância de 0,05, existe uma correlação significativa entre X e Y?

APLICANDO OS CONCEITOS

AUTO-AVALIAÇÃO

13.42 No Problema 13.4, em Problemas para a Seção 13.2, o gerente de marketing utilizou o espaço em prateleiras de supermercado destinado a rações para animais de estimação visando prever as vendas semanais. Os dados estão armazenados no arquivo **Ração**. Com base nos resultados para aquele problema, $b_1 = 7{,}4$ e $S_{b1} = 1{,}59$.
a. No nível de significância de 0,05, existem evidências de uma relação linear entre o espaço na prateleira e as vendas?
b. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população, β_1.

13.43 No Problema 13.5, em Problemas para a Seção 13.2, você relatou as vendas informadas de uma revista em bancas, com o objetivo de prever as vendas auditadas. Os dados estão armazenados no arquivo **Circulação**. Utilizando os resultados para aquele problema, $b_1 = 0{,}5719$ e $S_{b1} = 0{,}0668$.
a. No nível de significância de 0,05, existem evidências de uma relação linear entre vendas informadas e vendas auditadas?
b. Construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população, β_1.

13.44 No Problema 13.6, em Problemas para a Seção 13.2, o proprietário de uma empresa de mudanças queria prever as horas trabalhadas, com base no número de pés cúbicos transportados. Os dados estão armazenados no arquivo **Mudança**. Utilizando os resultados para aquele problema,
a. no nível de significância de 0,05, existem evidências de uma relação linear entre o número de pés cúbicos transportados e as horas trabalhadas?
b. construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população, β_1.

13.45 No Problema 13.7, em Problemas para a Seção 13.2, você utilizou o número de clientes para prever o tempo de espera na fila do caixa para pagamento das compras. Os dados estão armazenados no arquivo **Supermercado**. Utilizando os resultados correspondentes àquele problema,
a. no nível de significância de 0,05, existem evidências de uma relação linear entre o número de clientes e o tempo de espera na fila do caixa para o pagamento das compras?
b. construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população, β_1.

13.46 No Problema 13.8, em Problemas para a Seção 13.2, você utilizou as receitas anuais para prever o valor de uma franquia de beisebol. Os dados estão armazenados no arquivo **BBReceita**. Utilizando os resultados para aquele problema,
a. no nível de significância de 0,05, existem evidências de uma relação linear entre receita anual e valor da franquia?
b. construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população, β_1.

13.47 No Problema 13.9, em Problemas para a Seção 13.2, um corretor de uma imobiliária queria prever o aluguel mensal para apartamentos com base no tamanho do imóvel. Os dados estão armazenados no arquivo **Aluguel**. Utilizando os resultados para aquele problema,
a. no nível de significância de 0,05, existem evidências de uma relação linear entre o tamanho do apartamento e o aluguel mensal?
b. construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população, β_1.

13.48 No Problema 13.10, em Problemas para a Seção 13.2, você utilizou a receita bruta auferida em bilheterias de cinemas para prever as vendas de DVDs. Os dados estão armazenados no arquivo **Cinema**. Utilizando os resultados para aquele problema,
a. no nível de significância de 0,05, existem evidências de uma relação linear entre a receita bruta auferida em bilheterias de cinemas e as vendas de DVDs?
b. construa uma estimativa para o intervalo de confiança de 95% para a inclinação da população, β_1.

13.49 A volatilidade de uma ação em bolsa é geralmente medida por seu valor beta. Você pode estimar o valor beta de uma ação desenvolvendo um modelo de regressão linear simples, utilizando a variação percentual semanal do valor da ação como a variável dependente e a variação percentual semanal em um índice de mercado como a variável independente. O Índice S&P 500 é um índice habitualmente empregado. Por exemplo, caso desejasse estimar o valor de beta para a Disney, você poderia utilizar o modelo a seguir, que, algumas, vezes é denominado *modelo de mercado*:

$$\text{(variação \% semanal na Disney)} = \beta_0$$
$$+ \beta_1 \text{ (variação \% semanal no Índice S\&P 500)} + \varepsilon$$

A estimativa da regressão dos mínimos quadrados para a inclinação, b_1, é a estimativa do valor de beta para a Disney. Uma

ação com um valor de beta igual a 1,0 tende a variar na mesma intensidade e na mesma direção do mercado global. Uma ação com um valor de beta igual a 1,5 tende a variar 50% a mais do que o mercado global, e uma ação com um valor de beta igual a 0,6 tende a variar somente 60% em relação ao mercado global. Ações com valores de beta negativos tendem a variar de maneira oposta à do mercado global. A tabela a seguir fornece alguns valores de beta para algumas ações amplamente negociadas, utilizando os dados equivalentes a um ano, terminando em maio de 2009. Observe que nos primeiros 10 meses dessa grade, o S&P 500 perdeu aproximadamente 40% de seu valor e, depois disso, teve uma recuperação de cerca de 10% nos últimos 2 meses.

Empresa	Sigla	Beta
Procter & Gamble	PG	0,54
AT&T	T	0,73
Disney	DIS	1,10
Apple	AAPL	1,52
eBay	EBAY	1,69
Ford	F	2,86

Fonte: *Dados extraídos de* **finance.yahoo.com**, 27 de maio de 2009.

a. Para cada uma das seis empresas, interprete o valor de beta.
b. De que modo os investidores utilizam o valor de beta como um guia para investimentos?

13.50 Fundos indexados são fundos mútuos que tentam imitar o movimento dos principais índices, tais como o Índice S&P 500 ou o Índice Russell 2000. Os valores de beta para esses fundos (conforme descrito no Problema 13.49) são, por conseguinte, aproximadamente iguais a 1,0. Os modelos de mercado estimados para esses fundos correspondem a aproximadamente:

(variação % semanal no fundo indexado) =
0,0 + 1,0 (variação % semanal no índice)

Fundos indexados alavancados são projetados de modo a ampliar o movimento dos principais índices. A Direxion Funds é um dos principais provedores de índices alavancados e outros produtos de fundos mútuos de classe alternativa para consultores de investimentos e investidores sofisticados. Dois dos fundos mais populares da empresa são ilustrados na tabela a seguir (extraída de **www.direxionfunds.com**, 7 de janeiro de 2009).

Nome	Sigla	Descrição
S&P 500 Bull 2,5× Fundo	DXSLX	250% do Índice S&P 500
China Bull 2× Fundo	DXHLX	200% do Índice Xinhua China 25

Os modelos de mercado estimados para esses fundos correspondem a aproximadamente:

(variação % semanal no DXSLX) = 0,0 + 2,5
(variação % semanal no S&P500)

(variação % semanal no DXHLX) = 0,0 + 2
(variação % semanal no Xinhua China 25)

Por conseguinte, se o Índice S&P500 obtiver um ganho de 10% ao longo de um determinado período de tempo, o fundo mútuo alavancado DXSLX obtém um ganho aproximado de 25%. No caso de uma queda, se o mesmo índice tiver uma perda de 20%, o DXSLX perde aproximadamente 50%.

a. O objetivo do fundo Direxion Funds Small Capp Bull 2,5 ×, DXRLX, corresponde a 250% do desempenho do Índice Russell 2000. Qual é o modelo de mercado aproximado?
b. Se o Índice Russell 2000 obtiver um ganho de 10% em um ano, que retorno você esperaria que o DXRLX tivesse?
c. Se o Índice Russell 2000 tiver uma perda de 20% em um ano, que retorno você esperaria que o DXRLX tivesse?
d. Que tipo de investidores seriam atraídos para os fundos indexados alavancados? Que tipo de investidores devem se manter afastados desses fundos?

13.51 Os dados no arquivo **BebidasCafé** representam as calorias e a gordura (em gramas) de 16 onças de bebidas geladas à base de café na Dunkin' Donuts e na Starbucks:

Produto	Calorias	Gordura
Iced Mocha Swirl (leite integral), da Dunkin' Donuts	240	8,0
Coffee Frappuccino, blended coffee, da Starbucks	260	3,5
Coffee Coolatta (creme), da Dunkin' Donuts	350	22,0
Mocha Expresso Café Gelado (leite integral e creme chantilly), da Starbucks	350	20,0
Mocha Frappuccino blended coffee (creme chantilly), da Starbucks	420	16,0
Chocolate Brownie Frappuccino blended coffee (creme chantilly), da Starbucks	510	22,0
Chocolate Frappuccino Blended Crème (creme chantilly), da Starbucks	530	19,0

Fonte: *Dados extraídos de "Coffee as Candy at Dunkin' Donuts and Starbucks"*, Consumer Reports, *junho de 2004, p. 9.*

a. Calcule e interprete o coeficiente de correlação, *r*.
b. No nível de significância de 0,05, existe uma relação linear significativa entre calorias e gordura?

13.52 Existem diversos métodos para se calcular a economia de combustível. O arquivo **Milhagem** (dados ilustrados na tabela a seguir) contém a milhagem com base no cálculo feito por proprietários de veículos e pelos padrões correntes do governo norte-americano:

Veículo	Proprietário	Padrões do Governo
Ford F-150 2005	14,3	16,8
Chevrolet Silverado 2005	15,0	17,8
Honda Accord LX 2002	27,8	26,2
Honda Civic 2002	27,9	34,2
Honda Civic Hybrid 2004	48,8	47,6
Ford Explorer 2002	16,8	18,3
Toyota Camry 2005	23,7	28,5
Toyota Corolla 2003	32,8	33,1
Toyota Prius 2005	37,3	56,0

a. Calcule e interprete o coeficiente de correlação, *r*.
b. No nível de significância de 0,05, existe uma relação linear significativa entre a milhagem calculada pelos proprietários de veículos e a milhagem calculada nos padrões governamentais correntes?

13.53 O basquete em faculdades é um grande negócio, com salários de treinadores, receitas e despesas em milhões de dólares. O arquivo **Basquete-Faculdades** representa os salários dos treinadores e as receitas correspondentes ao basquete de faculdades, em escolas selecionadas, em um ano recente (dados extraídos de R. Adams, "Pay for Playoffs", *The Wall Street Journal*, 11-12 de março de 2006, pp. P1, P8).

a. Calcule e interprete o coeficiente de correlação, *r*.
b. No nível de significância de 0,05, existe uma relação linear significativa entre o salário de um treinador e as receitas?

13.54 Jogadores de futebol americano que estão tentando se habilitar para a NFL estão sendo submetidos ao teste de inteligência padronizado, o Wonderlic. O arquivo de dados **Wonderlic** contém a média dos resultados do Wonderlic para jogadores de futebol que estão tentando se habilitar para a NFL e o histórico da graduação de jogadores de futebol nas escolas que eles frequentaram (dados extraídos de S. Walker, "The NFL's Smartest Team", *The Wall Street Journal*, 30 de setembro de 2005, pp. W1, W10).

a. Calcule e interprete o coeficiente de correlação, *r*.
b. No nível de significância de 0,05, existe uma relação linear significativa entre a média dos resultados do Wonderlic para os jogadores de futebol que estão tentando se habilitar para a NFL e o histórico da graduação de jogadores de futebol em escolas selecionadas?
c. A que conclusões você consegue chegar sobre a relação entre a média dos resultados do Wonderlic para os jogadores de futebol que estão tentando se habilitar para a NFL e o histórico da graduação de jogadores de futebol em escolas selecionadas?

13.8 Estimativa da Média Aritmética dos Valores e Previsão de Valores Individuais

Esta seção apresenta métodos para estimar a média aritmética de *Y* e prever valores individuais de *Y*.

A Estimativa do Intervalo de Confiança

No Exemplo 13.2, você utilizou a linha de previsão para prever o valor de *Y* para um determinado *X*. As vendas anuais para lojas com área de 4.000 pés quadrados foi prevista como sendo de 7.644 milhões de dólares ($7.644.000). Essa estimativa, no entanto, é uma *estimativa de ponto* para a média aritmética da população. No Capítulo 8, você estudou o conceito de estimativa de intervalo de confiança para a média aritmética da população. De uma maneira semelhante, a Equação (13.20) define a **estimativa do intervalo de confiança para a média aritmética da resposta** para um determinado valor de *X*.

ESTIMATIVA DO INTERVALO DE CONFIANÇA PARA A MÉDIA ARITMÉTICA DE Y

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

$$\hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{h_i} \leq \mu_{Y|X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{h_i} \qquad \textbf{(13.20)}$$

em que

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{SQX}$$

$\hat{Y}_i$ = valor previsto de Y; $\hat{Y}_i = b_0 + b_1 X_i$

S_{YX} = erro-padrão da estimativa

n = tamanho da amostra

X_i = valor determinado de X

$\mu_{Y|X=X_i}$ = valor médio de Y quando $X = X_i$

$$SQX = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$t_{\alpha/2}$ é o valor crítico correspondente a uma probabilidade de cauda superior de $\alpha/2$, a partir da distribuição t com $n - 2$ graus de liberdade (ou seja, uma área acumulada de $1 - \alpha/2$).

A amplitude do intervalo de confiança na Equação (13.20) depende de vários fatores. Variações crescentes em torno da linha de previsão, medidas com base no erro-padrão da estimativa, resultam em um intervalo mais amplo. Como seria de se esperar, um tamanho crescente de amostra reduz a amplitude do intervalo. Além disso, a amplitude do intervalo também varia em relação a valores diferentes de X. Quando você prevê Y para valores de X próximos a $\overline{X}$, o intervalo é mais estreito do que para previsões de valores de X que estejam mais distantes de $\overline{X}$.

No exemplo da Sunflowers Roupas, suponha que você queira construir uma estimativa do intervalo de confiança de 95% para a média aritmética das vendas anuais em relação à população inteira de lojas que contenham uma área de 4.000 pés quadrados ($X = 4$). Utilizando a equação da regressão linear simples,

$$\begin{aligned}\hat{Y}_i &= 0{,}9645 + 1{,}6699X_i \\ &= 0{,}9645 + 1{,}6699(4) = 7{,}6439 \text{ (milhões de dólares)}\end{aligned}$$

Do mesmo modo, conhecendo-se:

$$\overline{X} = 2{,}9214 \quad S_{YX} = 0{,}9664$$

$$SQX = \sum_{i=1}^{n}(X_i - \overline{X})^2 = 37{,}9236$$

A partir da Tabela E.3, $t_{\alpha/2} = 2{,}1788$. Assim,

$$\hat{Y}_i \pm t_{\alpha/2}S_{YX}\sqrt{h_i}$$

em que

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{SQX}$$

de modo tal que:

$$\begin{aligned}&\hat{Y}_i \pm t_{\alpha/2}S_{YX}\sqrt{\frac{1}{n} + \frac{(X_i - \overline{X})^2}{SQX}} \\ &= 7{,}6439 \pm (2{,}1788)(0{,}9664)\sqrt{\frac{1}{14} + \frac{(4 - 2{,}9214)^2}{37{,}9236}} \\ &= 7{,}6439 \pm 0{,}6728\end{aligned}$$

então

$$6{,}9711 \leq \mu_{Y|X=4} \leq 8{,}3167$$

Portanto, a estimativa do intervalo de confiança de 95% é de que a média aritmética das vendas anuais esteja entre \$6.971.100 e \$8.316.700 para a população de lojas com área de 4.000 pés quadrados.

O Intervalo de Previsão

Além de construir um intervalo de confiança para a média aritmética de um valor de Y, você pode também construir um intervalo de previsão para um valor individual de Y. Embora a fórmula para o intervalo de previsão seja semelhante à fórmula para a estimativa do intervalo de confiança da Equação (13.20), o intervalo de previsão está prevendo um valor individual, e não estimando um parâmetro. A Equação (13.21) define o **intervalo de previsão para uma resposta individual, Y**, em um determinado valor, X_i, representado por $Y_{X=Xi}$.

INTERVALO DE PREVISÃO PARA UMA RESPOSTA INDIVIDUAL, Y

$$\hat{Y}_i \pm t_{\alpha/2}S_{YX}\sqrt{1 + h_i} \qquad \textbf{(13.21)}$$

$$\hat{Y}_i - t_{\alpha/2}S_{YX}\sqrt{1 + h_i} \leq Y_{X=X_i} \leq \hat{Y}_i + t_{\alpha/2}S_{YX}\sqrt{1 + h_i}$$

em que

h_i, $\hat{Y}_i$, S_{YX}, n e X_i são definidos na forma da Equação (13.20) e $Y_{X=Xi}$ é um valor futuro de Y quando $X = X_i$.

$t_{\alpha/2}$ é o valor crítico correspondente a uma probabilidade de cauda superior de $\alpha/2$, a partir da distribuição t, com $n - 2$ graus de liberdade (ou seja, uma área acumulada de $1 - \alpha/2$).

Para construir um intervalo de previsão de 95% para as vendas anuais de uma loja individual que tenha uma área de 4.000 pés quadrados ($X = 4$), você calcula, inicialmente, $\hat{Y}_i$. Utilizando a linha de previsão:

$$\begin{aligned}\hat{Y}_i &= 0{,}9645 + 1{,}6699X_i \\ &= 0{,}9645 + 1{,}6699(4) \\ &= 7{,}6439 \text{ (milhões de dólares)}\end{aligned}$$

Do mesmo modo, conhecendo-se:

$$\overline{X} = 2{,}9214 \quad S_{YX} = 0{,}9664$$

$$SQX = \sum_{i=1}^{n}(X_i - \overline{X})^2 = 37{,}9236$$

A partir da Tabela E.3, $t_{\alpha/2} = 2{,}1788$. Consequentemente,

$$\hat{Y}_i \pm t_{\alpha/2}S_{YX}\sqrt{1 + h_i}$$

em que

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

de modo que

$$\begin{aligned}&\hat{Y}_i \pm t_{\alpha/2}S_{YX}\sqrt{1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{SQX}} \\ &= 7{,}6439 \pm (2{,}1788)(0{,}9664)\sqrt{1 + \frac{1}{14} + \frac{(4 - 2{,}9214)^2}{37{,}9236}} \\ &= 7{,}6439 \pm 2{,}2104\end{aligned}$$

então

$$5{,}4335 \leq Y_{X=4} \leq 9{,}8543$$

Por conseguinte, com 95% de confiança, você prevê que as vendas anuais para uma loja individual com uma área de 4.000 pés quadrados estão entre \$5.433.500 e \$9.854.300.

A Figura 13.21 apresenta uma planilha que calcula a estimativa do intervalo de confiança e do intervalo de previsão para os dados relacionados à Sunflowers Roupas. Se comparar os resultados da estimativa do intervalo de confiança com o intervalo de previsão, você vai verificar que a amplitude do intervalo de previsão para uma loja individual é muito maior do que a amplitude da estimativa do intervalo de confiança para a média aritmética. Lembre-se de que existe uma quantidade bem maior de variação ao se prever um valor individual do que ao se estimar a média aritmética de um valor.

FIGURA 13.21
Planilha para estimativa do intervalo de confiança e intervalo de previsão para os dados relacionados à Sunflowers Roupas

A Figura 13.21 ilustra a planilha EICeIP da pasta de trabalho Regressão Linear Simples. Crie essa planilha utilizando as instruções na Seção GE13.8.

	A	B	C
1	Estimativa do Intervalo de Confiança e Previsão de Intervalo		
2			
3	**Dados**		
4	Valor de X	4	
5	Nível de Confiança	95%	
6			
7	**Cálculos Intermediários**		
8	Tamanho da Amostra	14	=CONT.NÚM(DadosRLS!A:A)
9	Graus de Liberdade	12	=B8 - 2
10	Valor t	2,1788	=INVT(1 - B5, B9)
11	Média da Amostra	2,9214	=DESVQ(DadosRLS!A:A)
12	Soma dos Quadrados das Diferenças	37,9236	=MÉDIA(DadosRLS!A:A)
13	Erro-padrão da Estimativa	0,9664	=CÁLCULO!B7
14	Estatística h	0,1021	=1/B8 + (B4 - B11)^2/B12
15	Média de Y Previsto (YChapéu)	7,6439	=TENDÊNCIA(DadosRLS!B2:B15, DadosRLS!A2:A15, B4)
16			
17	**Para a Média de Y**		
18	Metade da Amplitude do Intervalo	0,6728	=B10 * B13 * RAIZ(B14)
19	Limite Inferior do Intervalo de Confiança	6,9711	=B15 - B18
20	Limite Superior do Intervalo de Confiança	8,3167	=B15 + B18
21			
22	**Para Y da Resposta Individual**		
23	Metade da Amplitude do Intervalo	2,2104	=B10 * B13 * RAIZ(1 + B14)
24	Limite Inferior do Intervalo de Confiança	5,4335	=B15 - B23
25	Limite Superior do Intervalo de Confiança	9,8544	=B15 + B23

Problemas para a Seção 13.8

APRENDENDO O BÁSICO

13.55 Com base em uma amostra de tamanho $n = 20$, o método dos mínimos quadrados foi utilizado para desenvolver a seguinte linha de previsão: $\hat{Y}_i = 5 + 3X_i$. Além disso,

$$S_{YX} = 1{,}0 \quad \overline{X} = 2 \quad \sum_{i=1}^{n}(X_i - \overline{X})^2 = 20$$

a. Construa uma estimativa do intervalo de confiança de 95% para a média aritmética da resposta da população para $X = 2$.
b. Construa um intervalo de previsão de 95% para uma resposta individual para $X = 2$.

13.56 Com base em uma amostra de tamanho $n = 20$, o método dos mínimos quadrados foi utilizado para desenvolver a seguinte linha de previsão: $\hat{Y}_i = 5 + 3X_i$. Além disso,

$$S_{YX} = 1{,}0 \quad \overline{X} = 2 \quad \sum_{i=1}^{n}(X_i - \overline{X})^2 = 20$$

a. Construa uma estimativa do intervalo de confiança de 95% para a média aritmética da resposta da população para $X = 4$.
b. Construa um intervalo de previsão de 95% para uma resposta individual para $X = 4$.
c. Compare os resultados de (a) e (b) com os de (a) e (b) para o Problema 13.55. Qual dos intervalos é o mais amplo? Por quê?

APLICANDO OS CONCEITOS

13.57 No Problema 13.5, em Problemas para a Seção 13.2, você utilizou as vendas informadas de revistas em bancas com o objetivo de prever as vendas auditadas. Os dados estão armazenados no arquivo **Circulação**. Para esses dados, $S_{YX} = 42{,}186$ e $h_i = 0{,}108$ quando $X = 400$.
a. Construa uma estimativa para o intervalo de confiança de 95% para a média aritmética das vendas auditadas para revistas que informaram vendas de banca correspondentes a 400.000 exemplares.
b. Construa um intervalo de previsão de 95% para as vendas auditadas de uma revista individual que informe vendas de banca de 400.000 exemplares.
c. Explique a diferença nos resultados em (a) e (b).

AUTO-AVALIAÇÃO **13.58** No Problema 13.4, em Problemas para a Seção 13.2, o gerente de marketing utilizou o espaço em prateleiras de supermercado destinado a rações para animais de estimação para prever as vendas semanais. Os dados estão armazenados no arquivo **Ração**. Para esses dados, $S_{YX} = 30{,}81$ e $h_i = 0{,}1373$, quando $X = 8$.
a. Construa uma estimativa de intervalo de confiança de 95% para a média aritmética das vendas semanais de todas as lojas que possuem 8 pés de espaço de prateleira para rações de animais de estimação.
b. Construa um intervalo de previsão de 95% para as vendas semanais de uma loja que possui 8 pés de espaço de prateleira para rações de animais de estimação.
c. Explique a diferença nos resultados em (a) e (b).

13.59 No Problema 13.7, em Problemas para a Seção 13.2, você utilizou o número total de clientes dentro da loja para prever o tempo de espera na fila do caixa para pagamento das compras na saída do supermercado. Os dados estão armazenados no arquivo **Supermercado**.
a. Construa uma estimativa do intervalo de confiança de 95% para a média aritmética do tempo de espera na fila para todos os clientes, quando existem 20 clientes dentro da loja.
b. Construa um intervalo de previsão de 95% para o tempo de espera na fila para um cliente individual, quando existem 20 clientes dentro da loja.
c. Por que o intervalo em (a) é mais estreito do que o intervalo em (b)?

13.60 No Problema 13.6, em Problemas para a Seção 13.2, o proprietário de uma empresa de mudanças desejava prever as horas trabalhadas com base no número de pés cúbicos transportados. Os dados estão armazenados no arquivo **Mudança**.
a. Construa uma estimativa do intervalo de confiança de 95% para a média aritmética das horas trabalhadas para todas as mudanças com 500 pés cúbicos.
b. Construa um intervalo de previsão de 95% das horas trabalhadas para uma mudança individual com 500 pés cúbicos.
c. Por que o intervalo em (a) é mais estreito do que o intervalo em (b)?

13.61 No Problema 13.9, em Problemas para a Seção 13.2, um corretor de uma imobiliária queria prever o aluguel mensal para apartamentos com base no tamanho do imóvel. Os dados estão armazenados no arquivo **Aluguel**.
a. Construa uma estimativa para o intervalo de confiança de 95% para a média aritmética do aluguel mensal para todos os apartamentos com área de 1.000 pés quadrados.
b. Construa um intervalo de previsão de 95% para o aluguel mensal de um apartamento individual com área de 1.000 pés quadrados.
c. Explique a diferença nos resultados obtidos em (a) e (b).

13.62 No Problema 13.8, em Problemas para a Seção 13.2, você previu o valor de uma franquia de beisebol com base na receita corrente. Os dados estão armazenados no arquivo **BBReceita**.
a. Construa uma estimativa para o intervalo de confiança de 95% para a média aritmética do valor de todas as franquias de beisebol que geram \$150 milhões em receitas anuais.
b. Construa um intervalo de previsão de 95% para o valor de uma franquia individual de beisebol que gera \$200 milhões em receitas anuais.
c. Explique a diferença nos resultados obtidos em (a) e (b).

13.63 No Problema 13.10, em Problemas para a Seção 13.2, você utilizou a receita bruta auferida em bilheterias de cinema para prever a quantidade de DVDs vendidos. Os dados estão armazenados no arquivo **Cinema**. A empresa está em vias de lançar em DVD um filme que obteve uma receita bruta de bilheteria de \$30 milhões.
a. Qual é a quantidade prevista de DVDs que a empresa venderá?
b. Qual intervalo é mais útil no presente caso; a estimativa para o intervalo de confiança da média aritmética ou o intervalo de previsão para uma resposta individual? Explique.
c. Construa e interprete o intervalo que você selecionou em (b).

13.9 Armadilhas na Regressão

Algumas das armadilhas envolvidas na utilização da análise de regressão são as seguintes:

- Deixar de atentar para os pressupostos da regressão dos mínimos quadrados
- Não saber como avaliar os pressupostos da regressão dos mínimos quadrados
- Não saber quais são as alternativas à regressão dos mínimos quadrados caso um determinado pressuposto seja violado
- Utilizar um modelo de regressão sem conhecimento do assunto
- Extrapolar além do intervalo relevante
- Concluir que uma relação significativa identificada em um estudo observacional é decorrente de uma relação do tipo causa e efeito

A ampla disponibilidade de softwares estatísticos e de planilhas eletrônicas tornou bem mais fácil a análise da regressão. Entretanto, para muitos usuários, essa maior disponibilidade de softwares não foi acompanhada pelo entendimento de como utilizar a análise de regressão de maneira apropriada. Não se pode esperar que alguém que não esteja familiarizado com os pressupostos da regressão ou com o modo de avaliar esses pressupostos conheça quais são as alternativas à regressão dos mínimos quadrados caso um determinado pressuposto seja violado.

Os dados na Tabela 13.7 (armazenados no arquivo **Anscombe**) ilustram a importância de se utilizar gráficos de dispersão e análises de resíduos para ir além da simples manipulação de números envolvida no cálculo do intercepto de Y, da inclinação e de r^2.

Anscombe (referência 1) mostrou que todos os quatro conjuntos de dados ilustrados na Tabela 13.7 apresentam os seguintes resultados idênticos:

$$\hat{Y}_i = 3{,}0 + 0{,}5X_i$$
$$S_{YX} = 1{,}237$$
$$S_{b_1} = 0{,}118$$
$$r^2 = 0{,}667$$

$$SQReg = \text{Variação explicada} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = 27{,}51$$

$$SQR = \text{Variação não explicada} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 13{,}76$$

TABELA 13.7
Quatro Conjuntos de Dados Artificiais

Conjunto de Dados A		Conjunto de Dados B		Conjunto de Dados C		Conjunto de Dados D	
X_i	Y_i	X_i	Y_i	X_i	Y_i	X_i	Y_i
10	8,04	10	9,14	10	7,46	8	6,58
14	9,96	14	8,10	14	8,84	8	5,76
5	5,68	5	4,74	5	5,73	8	7,71
8	6,95	8	8,14	8	6,77	8	8,84
9	8,81	9	8,77	9	7,11	8	8,47
12	10,84	12	9,13	12	8,15	8	7,04
4	4,26	4	3,10	4	5,39	8	5,25
7	4,82	7	7,26	7	6,42	19	12,50
11	8,33	11	9,26	11	7,81	8	5,56
13	7,58	13	8,74	13	12,74	8	7,91
6	7,24	6	6,13	6	6,08	8	6,89

Fonte: *Dados extraídos de F. J. Anscombe, "Graphs in Statistical Analysis"*, American Statistician, *27 (1973), 17–21.*

$$STQ = \text{Variação total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 41{,}27$$

Caso tivesse que interromper a análise neste ponto, você deixaria de observar as importantes diferenças entre os quatro conjuntos de dados.

Com base nos gráficos de dispersão na Figura 13.22 e seus respectivos gráficos de resíduos na Figura 13.23, você verifica o quão diferentes são os conjuntos de dados. Cada um deles tem uma relação diferente entre X e Y. O único conjunto de dados que aparenta seguir aproximadamente uma linha reta é o conjunto de dados A. O gráfico de resíduos para o conjunto de dados A não demonstra nenhum padrão óbvio ou resíduo extremo (*outlier*). Isso certamente não é verdadeiro para os conjuntos de dados B, C e D. O gráfico de dispersão para o conjunto de dados B mostra que um modelo curvilíneo de regressão é mais apropriado. Essa conclusão é reforçada pelo gráfico de resíduos para o conjunto de dados B. O diagrama de dispersão e o gráfico de resíduos para o conjunto de dados C demonstram claramente uma observação extrema (*outlier*). Nesse caso, é utilizado um método para remover o valor extremo e estimar novamente o modelo de regressão (veja a referência 4). O diagrama de dispersão para o conjunto de dados D representa uma situação em que o modelo é fortemente dependente do resultado de um único ponto de dado ($X_8 = 19$ e $Y_8 = 12{,}50$). Qualquer modelo de regressão com essa característica deve ser utilizado com cautela.

FIGURA 13.22
Gráficos de dispersão para quatro conjuntos de dados

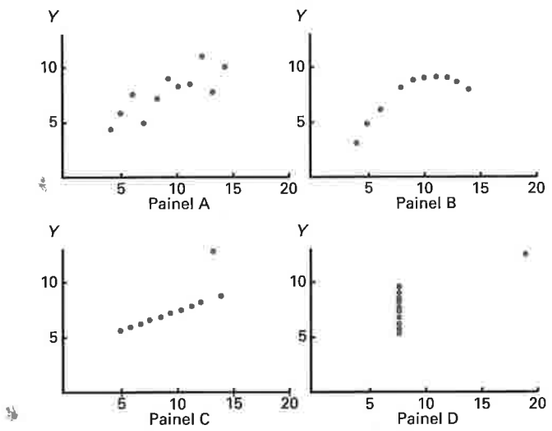
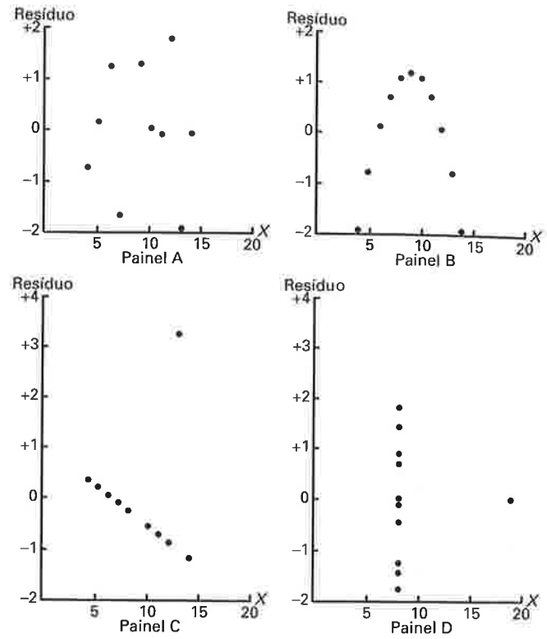


FIGURA 13.23
Gráficos dos resíduos para quatro conjuntos de dados



Em resumo, gráficos de dispersão e gráficos de resíduos são de vital importância em uma análise de regressão completa. As informações que eles fornecem são tão fundamentais para uma análise fidedigna, que você deve sempre incluir esses métodos gráficos como parte de uma análise de regressão. Portanto, uma estratégia que você pode empregar para ajudar a evitar as armadilhas da regressão é a seguinte:

1. Inicie com um gráfico de dispersão para observar a possível relação entre X e Y.
2. Verifique os pressupostos para a regressão (**l**inearidade, **i**ndependência, **n**ormalidade, **i**gualdade de variâncias) realizando uma análise de resíduos que inclua o seguinte:
   - **a.** Faça o gráfico dos resíduos em relação à variável independente para determinar se o modelo linear é apropriado e verificar o pressuposto da igualdade entre variâncias.
   - **b.** Construa um histograma, uma disposição ramo e folha, um box-plot ou um gráfico da probabilidade normal para os resíduos, com o objetivo de verificar o pressuposto da normalidade (esta etapa é necessária apenas se os dados são coletados ao longo do tempo).
   - **c.** Construa um gráfico para os resíduos em relação ao tempo e verifique o pressuposto da independência (esta etapa é necessária apenas se os dados são coletados ao longo do tempo).
3. Caso existam violações em relação aos pressupostos, utilize métodos alternativos para a regressão dos mínimos quadrados ou modelos alternativos para os mínimos quadrados (veja a referência 4).
4. Caso não existam violações dos pressupostos, realize os testes para a significância dos coeficientes da regressão e desenvolva intervalos de confiança e de previsão.
5. Evite fazer previsões e prognósticos que extrapolem o intervalo relevante da variável independente.
6. Tenha em mente que as relações identificadas em estudos observacionais podem ou não ser decorrentes de relações do tipo causa e efeito. Lembre-se de que, embora causa implique correlação, correlação não implica causa.

PENSE SOBRE ISSO As *Top Models* Norte-Americanas

Talvez você esteja familiarizado com o *America's Next Top Model*, programa de TV criado pela modelo Tyra Banks para encontrar a próxima *top model* americana. Pode ser que você esteja ainda menos familiarizado com um outro conjunto de *top models* que está emergindo do mundo empresarial.

Em um artigo da *Business Week*, edição de 23 de janeiro de 2006 (S. Baker, "Why Math Will Rock Your World: More Math Geeks Are Calling the Shots in Business. Is Your Industry Next?" *Business Week*, pp. 54-62), Stephen Baker fala sobre como os "*quants*" viraram as finanças de cabeça para baixo e estão se deslocando para outras áreas de negócios. O nome *quants* deriva dos "métodos quantitativos" que os "gênios da matemática" utilizam para desenvolver modelos e prognósticos. Esses métodos se fundamentam nos princípios da análise da regressão, discutidos neste capítulo, embora os modelos reais sejam bem mais complicados do que os modelos lineares simples discutidos aqui. Um outro artigo (S. Lohr, "For Today's Graduate, Just One Word: Statistics." *The New York Times*, 6 de agosto de 2009, p. A1, A3) discute sobre como a estatística está sendo utilizada para "minerar" grandes conjuntos de dados para descobrir padrões (frequentemente utilizando modelos de regressão).

Os modelos baseados na regressão passaram a ser os "*top models*" para muitos tipos de análises de negócios. Alguns exemplos incluem o seguinte:

- **Propaganda e marketing** Gerentes dessas áreas utilizam modelos econométricos (em outras palavras, modelos de regressão) para determinar o efeito de uma propaganda sobre as vendas, com base em um conjunto de fatores. Do mesmo modo, esses gerentes utilizam a mineração de dados (*data mining*) para prever padrões de comportamento em relação àquilo que os clientes comprarão no futuro, com base em informações históricas sobre o consumidor.
- **Finanças** A qualquer momento que leia sobre um "modelo" financeiro, você deve pressupor que algum tipo de modelo de regressão está sendo utilizado. Por exemplo, um artigo do *New York Times* de 18 de junho de 2006, intitulado "An Old Formula That Points to New Worry, de Mark Hulbert (p. BU8), discorre sobre um modelo de oportunidade de marketing que prevê o rendimento de ações nos próximos três a cinco anos, com base na geração de dividendos do mercado de ações e na taxa de juros de 90 dias dos títulos do Tesouro norte-americano.
- **Alimentos e bebidas** A Enologix, uma empresa de consultoria da Califórnia, desenvolveu uma "fórmula" (um modelo de regressão) que prevê o índice da qualidade de um vinho com base em um conjunto de componentes químicos encontrados nele (veja D. Darlington, "The Chemistry of a 90+ Wine", *The New York Times Magazine*, 7 de agosto de 2005, p. 36-39).
- **Livros** Um estudo do efeito de alterações nos preços realizado sobre as vendas na Amazon.com e na BN.com (novamente, análises de regressão) descobriu que um crescimento de preços de 1% na BN.com pressionava as vendas 4% para baixo, mas o mesmo crescimento de preços na Amazon.com pressionava as vendas para baixo em somente 0,5% (veja V. Postrel, "Economic Scene: When It Comes to Books, Internet Selling Has not Led to Uniformly Low Prices", *The New York Times*, 11 de setembro de 2003, p. C2).
- **Transportes** A Farecast.com utiliza a mineração de dados (*data mining*) e tecnologias de previsão para prever, objetivamente, o preço de passagens aéreas (veja D. Darlin, "Airfares Made Easy (Or Easier)", *The New York Times Magazine*, 1º de julho de 2006, p. C1, C6).
- **Imóveis** A Zillow.com utiliza informações sobre as características de um imóvel, bem como a sua localização, para desenvolver estimativas sobre o valor de mercado do imóvel, utilizando uma "fórmula" elaborada com um modelo proprietário.

No artigo da *Business Week*, Baker declara que a estatística e a probabilidade passarão a ser competências essenciais para empresários e consumidores. Aqueles que forem bem-sucedidos saberão como utilizar a estatística, estejam eles construindo modelos financeiros ou elaborando planos de marketing. Ele enfatiza também veementemente a necessidade de que todas as pessoas no mundo empresarial saibam utilizar o Microsoft Excel para que possam ser capazes de produzir análises e relatórios estatísticos. No artigo do *New York Times*, Lohr cita Hal Varian, principal economista do Google, ao dizer: "Continuo afirmando que o emprego atraente nos próximos dez anos será dos estatísticos."

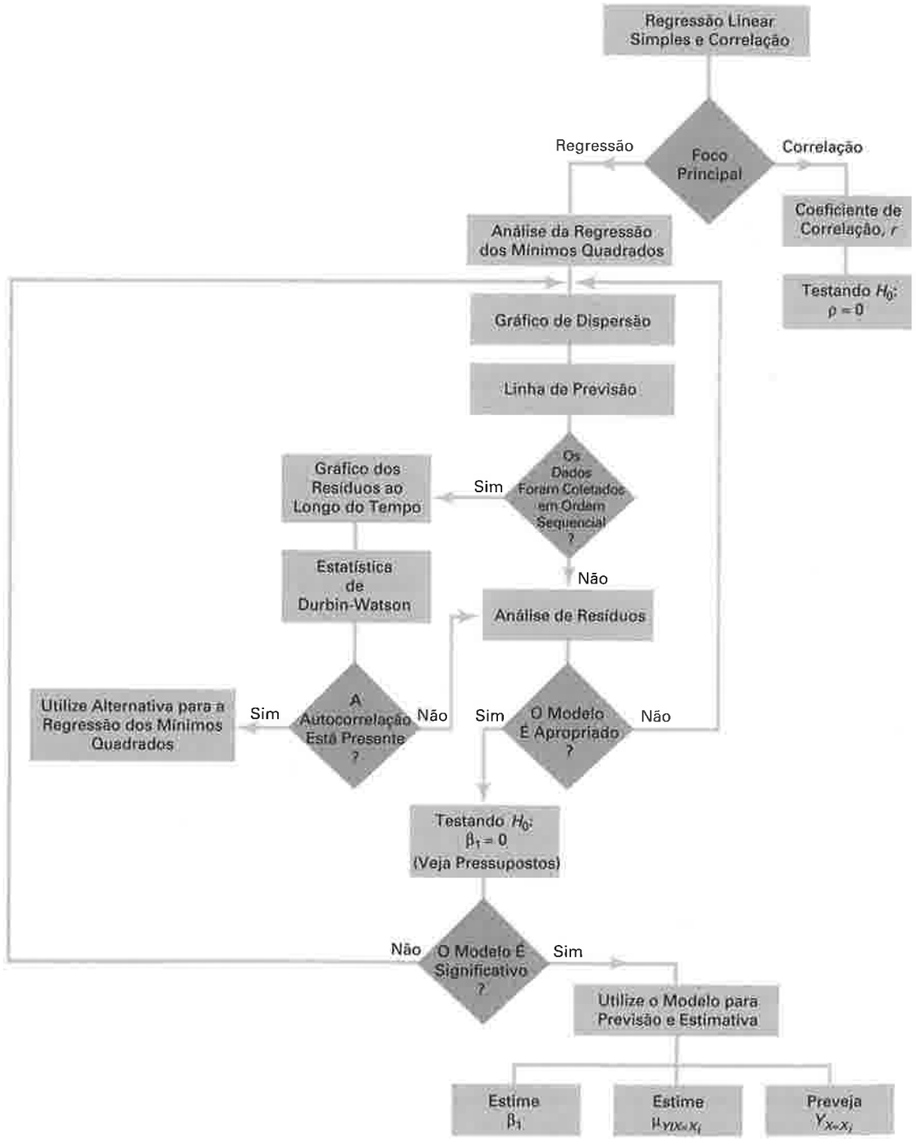
UTILIZANDO A ESTATÍSTICA @ Sunflowers Roupas Revisitada

No cenário da Sunflowers Roupas, você era o diretor de planejamento de uma cadeia de lojas de roupas femininas de primeira linha. Até agora, os gerentes da Sunflowers vêm selecionando locais para instalação de lojas com base em fatores tais como a disponibilidade de um bom contrato de arrendamento ou uma opinião subjetiva de que um determinado local parece ideal para uma loja de roupas. Para tomar decisões mais objetivas, você desenvolveu um modelo de regressão para analisar a relação entre o tamanho de uma loja e suas respectivas vendas anuais. O modelo indicou que aproximadamente 90,4% da variação nas vendas foi explicada pelo tamanho da loja. Além disso, para cada crescimento de 1.000 pés quadrados, foi estimado que a média aritmética das vendas anuais crescesse em $1,67 milhão. Você pode agora utilizar seu modelo para ajudá-lo a tomar decisões mais bem fundamentadas ao selecionar novos locais para instalação de lojas e também para prever vendas para as lojas já existentes.

RESUMO

Como você pode verificar no roteiro da Figura 13.24, este capítulo desenvolve o modelo de regressão linear simples, discute os pressupostos do modelo e mostra como avaliá-los. Uma vez certo de que o modelo é apropriado, você pode prever valores utilizando a linha de previsão e testar a significância da inclinação. No Capítulo 14, a análise da regressão é estendida para situações nas quais mais de uma variável numérica é utilizada para prever o valor de uma variável dependente.

FIGURA 13.24
Roteiro para a regressão linear simples



EQUAÇÕES-CHAVE

Modelo de Regressão Linear Simples

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (13.1)$$

Equação da Regressão Linear Simples: A Linha de Previsão

$$\hat{Y}_i = b_0 + b_1 X_i \quad (13.2)$$

Fórmula de Cálculo para a Inclinação, b_1

$$b_1 = \frac{SQXY}{SQX} \quad (13.3)$$

Fórmula de Cálculo para o Intercepto de *Y*, b_0

$$b_0 = \overline{Y} - b_1 \overline{X} \quad (13.4)$$

Medidas de Variação na Regressão

$$STQ = SQReg + SQR \quad (13.5)$$

Soma Total dos Quadrados (*STQ*)

$$STQ = \text{Soma total dos quadrados} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \quad (13.6)$$

Soma dos Quadrados da Regressão (*SQReg*)

$$SQReg = \text{Variação explicada ou soma dos quadrados da regressão} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 \quad (13.7)$$

Soma dos Quadrados dos Resíduos (*SQR*)

$$SQR = \text{Variação não explicada ou soma dos quadrados dos resíduos (erros)} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad (13.8)$$

Coeficiente de Determinação

$$r^2 = \frac{\text{Soma da regressão dos quadrados}}{\text{Soma total dos quadrados}} = \frac{SQReg}{STQ} \quad (13.9)$$

Fórmula de Cálculo para *STQ*

$$STQ = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \quad (13.10)$$

Fórmula de Cálculo para *SQReg*

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

$$= b_0 \sum_{i=1}^{n} Y_i + b_1 \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \quad (13.11)$$

Fórmula de Cálculo para *SQR*

$$SQR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i Y_i \quad (13.12)$$

Erro-Padrão da Estimativa

$$S_{YX} = \sqrt{\frac{SQR}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}} \quad (13.13)$$

Resíduos

$$e_i = Y_i - \hat{Y}_i \quad (13.14)$$

Estatística de Durbin-Watson

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}, \quad (13.15)$$

Testando uma Hipótese para a Inclinação da População, β_1, Utilizando o Teste *t*

$$t_{ESTAT} = \frac{b_1 - \beta_1}{S_{b_1}} \quad (13.16)$$

Testando uma Hipótese para a Inclinação da População, β_1, Utilizando o Teste *F*

$$F_{ESTAT} = \frac{MQReg}{MQE} \quad (13.17)$$

Estimativa do Intervalo de Confiança para a Inclinação, β_1

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

$$b_1 - t_{\alpha/2} S_{b_1} \leq \beta_1 \leq b_1 + t_{\alpha/2} S_{b_1} \quad (13.18)$$

Testando a Existência de Correlação

$$t_{ESTAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n-2}}} \quad (13.19a)$$

$$r = \frac{cov(X,Y)}{S_X S_Y} \quad (13.19b)$$

Estimativa do Intervalo de Confiança para a Média Aritmética de *Y*

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

$$\hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{h_i} \leq \mu_{Y|X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{h_i} \quad (13.20)$$

Intervalo de Previsão para uma Resposta Individual, *Y*

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

$$\hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{1 + h_i} \leq Y_{X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{1 + h_i} \quad (13.21)$$

TERMOS-CHAVE

análise da regressão
análise de resíduos
autocorrelação
coeficiente de correlação
coeficiente de determinação
coeficiente de regressão
diagrama de dispersão
equação da regressão linear simples
erro-padrão da estimativa
estatística de Durbin-Watson
estimativa do intervalo de confiança para a média aritmética da resposta
gráfico de dispersão
homoscedasticidade
igualdade de variâncias
inclinação
independência de erros
intercepto de *Y*
intervalo de previsão para uma resposta individual, *Y*
intervalo relevante
linearidade
linha de previsão
método dos mínimos quadrados
normalidade
pressupostos da regressão
regressão linear simples
relação linear
resíduos
soma dos quadrados da regressão (*SQReg*)
soma dos quadrados dos erros ou resíduos (*SQR*)
soma total dos quadrados (*STQ*)
variação explicada
variação não explicada
variação total
variável de resposta
variável dependente
variável explanatória
variável independente

PROBLEMAS DE REVISÃO DO CAPÍTULO 13

AVALIANDO O SEU ENTENDIMENTO

13.64 Qual é a interpretação para o intercepto de *Y* e para a inclinação na equação da regressão linear simples?

13.65 Qual é a interpretação para o coeficiente de determinação?

13.66 Em que situação a variação não explicada (ou seja, a soma dos quadrados dos resíduos) será igual a 0?

13.67 Em que situação a variação explicada (ou seja, a soma dos quadrados da regressão) será igual a 0?

13.68 Por que você deve sempre realizar uma análise de resíduos como parte de um modelo de regressão?

13.69 Quais são os pressupostos da análise da regressão?

13.70 De que modo você avalia os pressupostos da análise da regressão?

13.71 Quando e como você utiliza a estatística de Durbin-Watson?

13.72 Qual é a diferença entre a estimativa de um intervalo de confiança para a média aritmética da resposta, $\mu_{Y|X=X_i}$, e o intervalo de previsão de $Y_{X=X_i}$?

APLICANDO OS CONCEITOS

13.73 Pesquisadores da Lubin School of Business, na Pace University, conduziram um estudo sobre cursos com suporte na Internet. Em uma parte do estudo, foram coletadas quatro variáveis numéricas relativas a 108 alunos em um curso de Introdução à Administração que tinha encontros presenciais uma vez por semana, ao longo de todo um semestre. Uma das variáveis coletadas foi a *constância de acessos*. Para mensurar a constância de acessos, os pesquisadores procederam do seguinte modo: caso um determinado aluno não visitasse o endereço da Internet no intervalo entre as aulas, esse aluno recebia um 0 para esse período de tempo. Caso um determinado aluno visitasse o endereço da Internet uma vez ou mais no intervalo entre as aulas, recebia 1 para esse período de tempo. Uma vez que existiam 13 períodos de tempo, a pontuação de um aluno em relação à constância de acesso poderia ir de 0 a 13.

As outras três variáveis incluíam a média do aluno no curso, a média acumulada de notas do aluno (GPA – *grade point average*) e o número total de acessos que o aluno teve ao site da Internet de suporte ao curso. A tabela a seguir fornece o coeficiente de correlação para todos os pares de variáveis. Observe que as correlações marcadas com um * são estatisticamente significativas. Utilizando $\alpha = 0{,}001$:

Variável	Correlação
Média no Curso, GPA Acumulado	0,72*
Média no Curso, Total de Acessos	0,08
Média no Curso, Constância de Acessos	0,37*
GPA Acumulado, Total de Acessos	0,12
GPA Acumulado, Constância de Acessos	0,32*
Total de Acessos, Constância de Acessos	0,64*

Fonte: *Dados extraídos de D. Baugher, A. Varanelli and E. Weisbord, "Student Hits in an Internet-Supported Course: How Can Instructors Use Them and What Do They Mean?"* Decision Sciences Journal of Innovative Education, *outono de 2003, 1(2), pp.159–179.*

a. A que conclusões você pode chegar tendo como base essa análise da correlação?
b. Você está surpreso com os resultados, ou eles são coerentes com as suas próprias observações e experiências?

13.74 A gerência de uma empresa de envasamento de refrigerantes tem como objetivo estratégico desenvolver um método para transferir para os clientes os custos de entrega. Embora um dos custos esteja claramente relacionado ao tempo de transporte dentro de um determinado roteiro, uma outra variável de custo reflete o tempo necessário para descarregar as caixas de refrigerante no ponto de entrega. Para começar, a gerência decidiu desenvolver um modelo de regressão para prever o tempo de entrega com base no número de caixas entregues. Foi selecionada uma amostra de 20 entregas dentro de um determinado roteiro. Os tempos de entrega e o número de caixas entregues foram organizados na tabela a seguir (e armazenados no arquivo **Entrega**):

Cliente	Número de Caixas	Tempo de Entrega (Minutos)	Cliente	Número de Caixas	Tempo de Entrega (Minutos)
1	52	32,1	11	161	43,0
2	64	34,8	12	184	49,4
3	73	36,2	13	202	57,2
4	85	37,8	14	218	56,8
5	95	37,8	15	243	60,6
6	103	39,7	16	254	61,2
7	116	38,5	17	267	58,2
8	121	41,9	18	275	63,1
9	143	44,2	19	287	65,6
10	157	47,1	20	298	67,3

a. Utilize o método dos mínimos quadrados para calcular os coeficientes de regressão, b_0 e b_1.
b. Interprete o significado de b_0 e b_1 neste problema.
c. Faça a previsão para o tempo de entrega de 150 caixas de refrigerantes.
d. Será que você deveria utilizar o modelo para prever o tempo de entrega em relação a um cliente que esteja recebendo 500 caixas de refrigerantes? Por que sim ou por que não?
e. Determine o coeficiente de determinação, r^2, e explique seu significado neste problema.
f. Realize uma análise dos resíduos. Existe alguma evidência de um padrão nos resíduos? Explique.
g. No nível de significância de 0,05, existem evidências de uma relação linear entre o tempo de entrega e o número de caixas entregues?
h. Construa uma estimativa para o intervalo de confiança de 95% para a média aritmética do tempo de entrega de 150 caixas de refrigerantes e um intervalo de previsão de 95% para o tempo de entrega de uma única entrega de 150 caixas de refrigerantes.

13.75 Custos mistos são bastante comuns nas empresas e consistem em um elemento de custo fixo e um elemento de custo variável. Custos fixos correspondem a custos constantes, recorrentes, e que não variam quando varia a atividade das empresas. Custos variáveis são custos adicionais associados a cada uma das atividades desempenhadas pela organização. A relação pode ser caracterizada pela seguinte equação:

$$\text{Custos Totais} = \text{Custo fixo} + (\text{Custo por unidade}) \times (\text{Número de unidades de atividade da empresa})$$

Em um dos mais importantes livros didáticos que trata de contabilidade gerencial, os autores discorrem sobre os custos totais de manutenção de um hospital e utilizam a análise da regressão para estimar o elemento relativo ao custo fixo de manutenção e o custo variável associado ao número de dias relativos à permanência de pacientes. Os custos totais de manutenção e o número de dias de permanência de pacientes durante sete meses estão listados na tabela a seguir e armazenados no arquivo **CustosMistos**.

Custos Totais de Manutenção	Dias de Permanência de Pacientes
$7.900	5.600
$8.500	7.100
$7.400	5.000
$8.200	6.500
$9.100	7.300
$9.800	8.000
$7.800	6.200

Fonte: *Dados extraídos de P. C. Brewer, R. H. Garrison, and E. W. Noreen* Introduction to Managerial Accounting, *4th ed. (Boston: McGraw-Hill Irwin, 2008).*

a. Utilizando os custos totais de manutenção como a variável dependente e o número de dias de permanência de pacientes como a variável independente, utilize o método dos mínimos quadrados para encontrar os coeficientes de regressão, b_0 e b_1.
b. Qual coeficiente da regressão representa o custo fixo?
c. Qual coeficiente da regressão representa o custo variável para cada dia de permanência de pacientes?
d. Faça a previsão para o custo total de manutenção para um mês com 7.500 dias de permanência de pacientes.

13.76 Você deseja desenvolver um modelo para prever o preço para a venda de casas com base no valor de avaliação. É selecionada uma amostra aleatória de 30 residências unifamiliares vendidas recentemente em uma pequena cidade para estudar a relação entre o preço de venda (em milhares de dólares) e o valor de avaliação (em milhares de dólares). As casas na cidade haviam sido reavaliadas com base em seu valor máximo um ano antes desse estudo. Os resultados estão no arquivo **Casa1**.

(Dica: Primeiramente, determine qual é a variável independente e qual é a variável dependente.)

a. Construa um gráfico de dispersão e, pressupondo uma relação linear, utilize o método dos mínimos quadrados para calcular os coeficientes da regressão, b_0 e b_1.
b. Interprete o significado para o intercepto de Y, b_0, e para a inclinação, b_1, neste problema.
c. Utilize a linha de previsão desenvolvida em (a) para prever o preço de venda para uma casa cujo valor de avaliação seja de $170.000.
d. Determine o coeficiente de determinação, r^2, e interprete o seu significado neste problema.
e. Realize uma análise de resíduos em seus resultados e avalie os pressupostos da regressão.
f. No nível de significância de 0,05, existem evidências de uma relação linear entre o preço de venda e o valor de avaliação?
g. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população.

13.77 Você deseja desenvolver um modelo para prever o valor de avaliação de casas, com base na área aquecida. Foi selecionada uma amostra de 15 residências unifamiliares em uma determinada cidade. O valor de avaliação (em milhares de dólares) e a área aquecida das casas (em milhares de pés quadrados) são registrados, e os seguintes resultados são armazenados em **Casa2**.

(Dica: Primeiramente, determine qual é a variável independente e qual é a variável dependente.)

a. Construa um gráfico de dispersão e, pressupondo uma relação linear, utilize o método dos mínimos quadrados para calcular os coeficientes da regressão, b_0 e b_1.
b. Interprete o significado para o intercepto de Y, b_0, e para a inclinação, b_1, neste problema.
c. Utilize a linha de previsão desenvolvida em (a) para prever o valor de avaliação para uma casa cuja área aquecida é igual a 1.750 pés quadrados.

d. Determine o coeficiente de determinação, r^2, e interprete o seu significado neste problema.
e. Realize uma análise de resíduos em seus resultados e avalie os pressupostos da regressão.
f. No nível de significância de 0,05, existem evidências de uma relação linear entre o valor de avaliação e a área aquecida?

13.78 O diretor da graduação de uma grande faculdade de negócios gostaria de prever a média geral no período da graduação (GPA – *grade point average*) de alunos de um programa de MBA, com base nos resultados para o GMAT (Graduate Management Admission Test). Foi selecionada uma amostra de 20 alunos que haviam completado 2 anos no programa. Os resultados estão armazenados no arquivo **GPIGMAT**.

(Dica: Primeiramente, determine qual é a variável independente e qual é a variável dependente.)

a. Construa um gráfico de dispersão e, pressupondo uma relação linear, utilize o método dos mínimos quadrados para calcular os coeficientes da regressão, b_0 e b_1.
b. Interprete o significado para o intercepto de Y, b_0, e para a inclinação, b_1, neste problema.
c. Utilize a linha de previsão desenvolvida em (a) para prever o GPA para um aluno com um resultado de GMAT igual a 600.
d. Determine o coeficiente de determinação, r^2, e interprete o seu significado neste problema.
e. Realize uma análise de resíduos em seus resultados e avalie os pressupostos da regressão.
f. No nível de significância de 0,05, existem evidências de uma relação linear entre o resultado do GMAT e do GPA?
g. Construa uma estimativa para o intervalo de confiança de 95% da média aritmética do GPA de alunos com um resultado de GMAT igual a 600 e um intervalo de previsão de 95% para o GPA de um determinado aluno com um resultado de GMAT igual a 600.
h. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população.

13.79 O chefe da contabilidade de uma grande loja de departamentos gostaria de desenvolver um modelo para prever a quantidade de tempo necessária para processar faturas. Os dados foram coletados a partir dos últimos 32 dias de trabalho, e o número de faturas processadas e o tempo para seu respectivo preenchimento (em horas) estão armazenados em **Fatura**.

(Dica: Determine, inicialmente, qual é a variável independente e qual é a variável dependente.)

a. Pressupondo uma relação linear, utilize o método dos mínimos quadrados para calcular os coeficientes da regressão, b_0 e b_1.
b. Interprete o significado para o intercepto de Y, b_0, e para a inclinação, b_1, neste problema.
c. Utilize a linha de previsão desenvolvida em (a) para prever a quantidade de tempo necessária para processar 150 faturas.
d. Determine o coeficiente de determinação, r^2, e interprete o seu significado.
e. Elabore um gráfico dos resíduos em relação ao número de faturas processadas e também em relação ao tempo.
f. Tomando como base os gráficos em (e), o modelo parece apropriado?
g. Tomando como base os resultados de (e) até (f), a que conclusões você consegue chegar sobre a validade da previsão feita em (c)?

13.80 Em 28 de janeiro de 1986, o ônibus espacial *Challenger* explodiu, matando os sete astronautas a bordo. Antes do lançamento, a temperatura atmosférica prevista para o local de lançamento indicava congelamento. Engenheiros da Morton Thiokol (fabricante do motor do foguete) prepararam gráficos com o objetivo de defender que o lançamento não fosse realizado devido ao tempo frio. Esses argumentos foram rejeitados, e o lançamento tragicamente veio a ocorrer. Diante de investigações posteriores à tragédia, especialistas concordaram que o desastre ocorrera em razão de um vazamento nos anéis retentores de borracha que não teriam realizado a vedação apropriadamente em virtude da baixa temperatura. Os dados que indicam a temperatura atmosférica no momento de 23 lançamentos anteriores e o índice correspondente a danos nos anéis retentores encontram-se armazenados no arquivo **Retentor**.

Observação: Os dados do voo 4 foram omitidos devido ao desconhecimento das condições do retentor.

Fonte: *Dados extraídos de* Report of the Presidential Commission on the Space Shuttle Challenger Accident, *Washington, DC, 1986, Vol. II (H1-H3) e Vol. IV (664), e* Post Challenger Evaluation of Space Shuttle Risk Assessment and Management, Washington, DC, 1988, pp. 135-136.

a. Construa um gráfico de dispersão para os sete voos nos quais havia um dano no retentor (índice de dano no retentor $\neq 0$). A que conclusões, se houver alguma, você pode chegar sobre a relação entre a temperatura atmosférica e os danos nos retentores?
b. Construa um diagrama de dispersão para todos os 23 voos.
c. Explique quaisquer diferenças na interpretação da relação entre temperatura atmosférica e danos no retentor em (a) e (b).
d. Com base no gráfico de dispersão em (b), apresente razões pelas quais não deveria ser feita uma previsão para uma temperatura atmosférica de 31°F, a temperatura na manhã do lançamento da *Challenger*.
e. Embora o pressuposto de uma relação linear com a temperatura atmosférica possa não ser válido para o conjunto de 23 voos, ajuste um modelo de regressão linear simples de modo a prever os danos nos retentores, com base na temperatura atmosférica.
f. Inclua a linha de previsão encontrada em (e) no gráfico de dispersão desenvolvido em (b).
g. Com base nos resultados de (f), você acredita que um modelo linear é apropriado para esses dados? Explique.
h. Realize uma análise nos resíduos. A que conclusões você chega?

13.81 Crazy Dave, um analista de beisebol bastante conhecido, gostaria de estudar várias estatísticas de times para a temporada de beisebol de 2008, de modo a determinar quais variáveis poderiam ser úteis para prever o número de vitórias alcançadas pelos times durante a temporada. Ele decidiu começar utilizando a média de voltas percorridas (ERA – *earned run average*), um indicador para o desempenho de respostas a arremessos, de modo a prever o número de vitórias. Os dados relativos aos 30 times da liga principal de beisebol estão armazenados no arquivo **BB2008**.

(Dica: Primeiramente, determine qual é a variável independente e qual é a variável dependente.)

a. Pressupondo uma relação linear, utilize o método dos mínimos quadrados para calcular os coeficientes da regressão, b_0 e b_1.

b. Interprete o significado para o intercepto de Y, b_0, e para a inclinação, b_1, neste problema.
c. Utilize a linha de previsão desenvolvida em (a) para prever o número de vitórias para um time com um ERA de 4,50.
d. Calcule o coeficiente de determinação, r^2, e interprete o seu significado.
e. Realize uma análise nos resíduos de seus resultados e determine a adequação do ajuste do modelo.
f. No nível de significância de 0,05, existem evidências de uma relação linear entre o número de vitórias e o ERA?
g. Construa uma estimativa do intervalo de confiança de 95% para a média aritmética do número de vitórias esperadas para times com um ERA de 4,50.
h. Construa um intervalo de previsão de 95% para o número de vitórias de um determinado time com um ERA de 4,50.
i. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população.
j. Os 30 times constituem uma população. Para que se possa utilizar a inferência estatística, como é o caso em (f) até (i), deve-se pressupor que os dados representem uma amostra aleatória. Sobre qual "população" essa amostra estaria tirando conclusões?
k. Que outras variáveis independentes você deveria considerar para fins de inclusão no modelo?

13.82 Você seria capaz de utilizar as receitas anuais geradas pelas franquias da NBA (National Basketball Association) para prever o valor das franquias? A Figura 2.18 mostra um gráfico de dispersão para receitas em relação ao valor da franquia, e a Figura 3.9 mostra o coeficiente de correlação. Agora, você quer desenvolver um modelo de regressão linear simples para prever valores de franquias com base em receitas. (Valores de franquias e receitas estão armazenados em **ValoresNBA**.)
a. Pressupondo uma relação linear, utilize o método dos mínimos quadrados para encontrar os coeficientes da regressão, b_0 e b_1.
b. Interprete o significado para o intercepto de Y, b_0, e para a inclinação, b_1, neste problema.
c. Faça a previsão do valor de uma franquia da NBA que gere $200 milhões em receitas anuais.
d. Calcule o coeficiente de determinação, r^2, e interprete o seu significado.
e. Realize uma análise nos resíduos de seus resultados e avalie os pressupostos da regressão.
f. No nível de significância de 0,05, existem evidências de uma relação linear entre as receitas geradas e o valor de uma franquia da NBA?
g. Construa uma estimativa para o intervalo de confiança de 95% para a média aritmética do valor de todas as franquias da NBA que geraram $150 milhões em receitas anuais.
h. Construa um intervalo de previsão de 95% para o valor de uma determinada franquia da NBA que gere $150 milhões em receitas anuais.
i. Compare os resultados de (a) até (h) com os resultados das franquias de beisebol nos Problemas 13.18, 13.20, 13.30, 13.46 e 13.62 e das franquias da National Football League no Problema 13.83.

13.83 No Problema 13.82, você utilizou receitas anuais para desenvolver um modelo para prever o valor de franquias de times da National Basketball Association (NBA). Você seria também capaz de utilizar as receitas anuais geradas pelas franquias da NFL (National Football League) para prever os valores para as franquias? (Os valores das franquias da NFL e as respectivas receitas estão armazenados no arquivo **ValoresNFL**.)
a. Repita o Problema 13.82 de (a) a (h) para as franquias da NFL.
b. Compare os resultados de (a) com os resultados das franquias de beisebol nos Problemas 13.18, 13.20, 13.30, 13.46 e 13.62 e das franquias da NBA no Problema 13.82.

13.84 Durante a safra de outono nos Estados Unidos, abóboras são vendidas em grande quantidade em barracas nas fazendas. Frequentemente, em vez de pesar as abóboras antes da venda, o fazendeiro da barraca simplesmente coloca a abóbora no cortador circular apropriado em cima do balcão. Indagado sobre o porquê disso, um fazendeiro respondeu: "Eu consigo afirmar o peso da abóbora com base em sua circunferência." Para determinar se isso é realmente verdadeiro, uma amostra com 23 abóboras teve suas circunferências e pesos aferidos, e os resultados estão armazenados em **Abóbora**.
a. Pressupondo uma relação linear, utilize o método dos mínimos quadrados para encontrar os coeficientes da regressão, b_0 e b_1.
b. Interprete o significado para a inclinação, b_1, neste problema.
c. Faça uma previsão do peso de uma abóbora que tenha 60 centímetros de circunferência.
d. Você acredita que seja uma boa ideia para o fazendeiro vender as abóboras com base na circunferência e não no peso? Explique.
e. Determine o coeficiente de determinação, r^2, e interprete seu significado.
f. Realize uma análise dos resíduos para esses dados e avalie os pressupostos da regressão.
g. No nível de significância de 0,05, existem evidências de uma relação linear entre a circunferência e o peso de uma abóbora?
h. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população, β_1.

13.85 Dados demográficos podem ser úteis para prever as vendas de lojas de artigos esportivos? O arquivo **Esportes** contém os totais de vendas mensais de uma amostra aleatória de 38 lojas de uma grande cadeia nacional de lojas de artigos esportivos. Todas as lojas na franquia, e, consequentemente, dentro da amostra, são aproximadamente do mesmo tamanho e têm as mesmas mercadorias. A cidade ou, em alguns casos, as cidades onde as lojas possuem a maioria de seus clientes são aqui definidas como base da clientela. São disponibilizados os dados demográficos da base da clientela de cada uma das 38 lojas. Os dados são verdadeiros, mas o nome da franquia não é utilizado, por solicitação da empresa. O conjunto de dados contém as seguintes variáveis:

Vendas – Total de vendas no último mês (dólares)
Idade – Mediana da idade da base da clientela (anos)
SG – Percentual da base da clientela com um diploma de segundo grau
Faculdade – Percentagem da base da clientela com um diploma de terceiro grau
Crescimento – Taxa de crescimento populacional anual da base da clientela ao longo dos últimos 10 anos
Renda – Mediana da renda familiar da base da clientela (dólares)

a. Construa um gráfico de dispersão utilizando vendas como a variável dependente e a mediana da renda familiar como a variável independente. Discuta o diagrama de dispersão.

b. Pressupondo uma relação linear, utilize o método dos mínimos quadrados para calcular os coeficientes da regressão, b_0 e b_1.
c. Interprete o significado para o intercepto de Y, b_0, e para a inclinação, b_1, neste problema.
d. Calcule o coeficiente de determinação, r^2, e interprete seu significado.
e. Faça uma análise dos resíduos nos seus resultados e determine a adequação do ajuste do modelo.
f. No nível de significância de 0,05, existem evidências de uma relação linear entre a variável dependente e a variável independente?
g. Construa uma estimativa do intervalo de confiança de 95% para a inclinação da população e interprete seu significado.

13.86 Para os dados do Problema 13.85, repita (a) até (g), utilizando a mediana da idade como a variável independente.

13.87 Para os dados do Problema 13.85, repita (a) até (g), utilizando o índice de graduação no segundo grau como a variável independente.

13.88 Para os dados do Problema 13.85, repita (a) até (g), utilizando o índice de graduação no terceiro grau como a variável independente.

13.89 Para os dados do Problema 13.85, repita (a) até (g), utilizando o crescimento populacional como a variável independente.

13.90 O Zagat's publica cotações de restaurantes de várias localidades nos Estados Unidos. O arquivo de dados **CustoRest** contém a cotação do Zagat para comida, decoração, serviço e preço por pessoa para uma amostra de 100 restaurantes localizados na área urbana (cidade de Nova York) e em um subúrbio da cidade de Nova York. Desenvolva um modelo de regressão para prever o preço por pessoa, com base na variável que representa a soma entre as cotações para comida, decoração e serviço.

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

a. Pressupondo uma relação linear, utilize o método dos mínimos quadrados para calcular os coeficientes da regressão, b_0 e b_1.
b. Interprete o significado para o intercepto de Y, b_0, e para a inclinação, b_1, neste problema.
c. Utilize a linha de previsão desenvolvida em (a) para prever o preço por pessoa de um restaurante cuja soma das cotações seja 50.
d. Calcule o coeficiente de determinação, r^2, e interprete seu significado.
e. Faça uma análise dos resíduos nos seus resultados e avalie os pressupostos da regressão.
f. No nível de significância de 0,05, existem evidências de uma relação linear entre o preço por pessoa e a soma das cotações?
g. Qual a utilidade da soma das cotações como um mecanismo de previsão de preço? Explique.

13.91 Reporte-se à discussão sobre valores beta e modelos de mercado do Problema 13.49, em Problemas para a Seção 13.7. O Índice S&P 500 acompanha o movimento geral do mercado de ações considerando os preços de ações de 500 grandes corporações. O arquivo **Ações 2009** contém os dados semanais do S&P 500 e de ações de três empresas para 100 semanas que finalizam em 1.º de junho de 2009. Estão incluídas as seguintes variáveis:

SEMANA – Semana corrente
S&P – Valor do fechamento semanal para o Índice S&P 500
GE – Valor do fechamento semanal para ações da General Electric
IBM – Valor do fechamento semanal para ações da IBM
XOM – Valor do fechamento semanal para ações da Exxon Mobil

Fonte: *Dados extraídos de* ***finance.yahoo.com****, 3 de junho de 2009.*

a. Faça a estimativa para o modelo de mercado da GE. (Dica: Utilize a variação percentual no Índice S&P 500 como a variável independente e a variação percentual no preço das ações da GE como a variável dependente.)
b. Interprete o valor beta para a GE.
c. Repita (a) e (b) para a IBM.
d. Repita (a) e (b) para a Exxon Mobil.
e. Escreva um relatório sucinto sobre suas descobertas.

13.92 Você acredita que os bônus e os pacotes totais de remuneração para os executivos-chefes (CEO) de grandes empresas estejam correlacionados ao desempenho das ações dessas empresas? Em 2009, as ações incluídas no S&P 500 tiveram uma baixa de 38,5%, mas qual foi o ganho real dos executivos-chefes? As variáveis no arquivo **Remuneração** incluem os seguintes dados para 385 executivos-chefes de empresas cotadas no S&P 500 que registraram seus representantes entre 1º de janeiro e 20 de abril de 2009.

EMPRESA – Nome de uma amostra de 385 das 500 empresas listadas no S&P 500
AÇÕES – Desempenho das ações em 2009 (crescimento ou diminuição percentual)
BÔNUS – Bônus pagos aos executivos-chefes em 2009
REMUNERAÇÃO – Total da remuneração financeira (salários, bônus, opções de ações etc.) paga aos executivos-chefes em 2009

Fonte: *Extraído de Dow Jones e B. Hansen "CEO Pay Dives in a Rough 2008",* ***usatoday.com****, 1.º de maio de 2009.*

a. Calcule os coeficientes de correlação para AÇÃO e BÔNUS, AÇÃO e REMUNERAÇÃO e BÔNUS e REMUNERAÇÃO.
b. No nível de significância de 0,05, alguma das correlações é estatisticamente significativa?
c. Redija um breve resumo de suas descobertas em (a) e (b). Os resultados o surpreendem?

13.93 O arquivo **Invest 2009** contém o preço de fechamento diário para ouro, prata, o S&P 500 e o Nasdaq para os 68 primeiros dias de negociações de 2009 (dados extraídos de **finance.yahoo.com**, 10 de abril de 2009).
a. Calcule os coeficientes de correlação entre os quatro diferentes investimentos. (Existem seis deles.)
b. No nível de significância de 0,05, alguma das correlações entre os diferentes investimentos é estatisticamente significativa?
c. Redija um breve resumo de suas descobertas.

EXERCÍCIOS DE REDAÇÃO DE RELATÓRIOS

13.94 Nos Problemas 13.85 a 13.89, você desenvolveu modelos de regressão para prever vendas mensais de uma loja de artigos esportivos. Agora, escreva um relatório baseado nos modelos que você desenvolveu. Anexe ao seu relatório todos os gráficos e informações estatísticas apropriados.

ADMINISTRANDO O *SPRINGVILLE HERALD*

Para garantir que o máximo possível de assinaturas a título de promoção seja convertido em assinaturas regulares, o departamento de marketing do *Herald* trabalha estreitamente interligado com o departamento de distribuição no sentido de conseguir um processo tranquilo de remessa inicial para os clientes que estão no período de promoção. Para auxiliar nesse esforço, o departamento de marketing precisa prever acuradamente o número de novas assinaturas regulares para os meses subsequentes.

Uma equipe formada dos gerentes dos departamentos de marketing e de distribuição foi chamada para desenvolver um método mais eficiente para prever novas assinaturas. Antes disso, depois de examinar os dados sobre novas assinaturas relativos a três meses anteriores, um grupo de três gerentes desenvolveria um prognóstico subjetivo para o número de novas assinaturas. Lauren Hall, contratada recentemente pela empresa para aplicar sua especialidade em métodos quantitativos de previsão, sugeriu que o departamento buscasse fatores que pudessem ajudar a prever as novas assinaturas.

Os membros da equipe descobriram que as previsões do ano anterior haviam sido particularmente imprecisas em razão de, em alguns meses, ter sido gasto um montante de tempo bem maior com telemarketing do que em outros meses. De um modo particular, no mês anterior, somente 1.055 horas foram completadas, porque os atendentes estavam ocupados durante a primeira semana do mês frequentando sessões de treinamento sobre o estilo de abordagem pessoal porém formal e sobre o novo guia de apresentações padronizadas (veja "Administrando o *Springville Herald*" no Capítulo 11). Lauren coletou dados (armazenados em **SH13**) correspondentes ao número de novas assinaturas e às horas gastas em telemarketing para cada um dos meses ao longo dos últimos dois anos.

EXERCÍCIOS

SH13.1 Que tipo de crítica você pode fazer em relação ao método de previsão que envolvia a adoção de dados de novas assinaturas nos três meses anteriores como a base para projeções futuras?

SH13.2 Que fatores outros que não o número de horas gastas com telemarketing poderiam ser úteis para a previsão do número de novas assinaturas? Explique.

SH13.3
- **a.** Analise os dados e desenvolva um modelo de regressão para prever o número de novas assinaturas para um mês, com base no número de horas gastas com telemarketing em busca de novas assinaturas.
- **b.** Se você espera gastar 1.200 horas por mês com telemarketing, faça a estimativa para o número de novas assinaturas para o mês. Indique os pressupostos nos quais essa previsão se baseou. Você acredita que esses pressupostos sejam válidos? Explique.
- **c.** Qual seria o perigo inerente ao fato de prever o número de novas assinaturas para um mês no qual 2.000 horas tenham sido gastas com telemarketing?

CASO DE INTERNET

Aplique os seus conhecimentos sobre regressão linear simples neste Caso de Internet, que representa uma extensão do cenário Utilizando a Estatística deste capítulo, que trata da Sunflowers Roupas.

Corretores imobiliários da Triangle Mall Management Corporation sugeriram que a Sunflowers considerasse várias locações em alguns centros comerciais de conveniência recentemente reformados que têm um apelo para consumidores com uma renda disponível acima da média. Embora os imóveis sejam menores do que o tipo de imóvel habitual da Sunflowers Roupas, os corretores argumentam que a renda disponível acima da média na comunidade da vizinhança é um melhor prognóstico para um alto volume de vendas do que o tamanho do imóvel. Os corretores imobiliários defendem que dados de amostras oriundos de 14 lojas da Sunflowers provam que isso é verdadeiro.

Utilizando um navegador para a Web, abra o Caso de Internet do Capítulo 13 no site da LTC Editora para este livro, ou abra diretamente o arquivo **Triangle_Sunflowers.htm** caso já tenha baixado para seu computador os arquivos com os Casos de Internet, para examinar a proposta dos corretores e os documentos que respaldam essas afirmativas. Depois disso, responda ao seguinte:

1. A média aritmética da renda disponível deve ser utilizada para prever vendas com base em uma amostra de 14 lojas da rede Sunflowers?
2. A administração da Sunflowers deveria aceitar as declarações dos corretores imobiliários da Triangle? Por que sim ou por que não?
3. É possível que a média aritmética da renda disponível da vizinhança não seja um fator importante para fins de locação de novos imóveis para instalação de lojas? Explique.
4. Existem outros fatores, não mencionados pelos corretores imobiliários, que possam ser relevantes para a decisão da loja em relação à locação?

REFERÊNCIAS

1. Anscombe, F. J., "Graphs in Statistical Analysis," *The American Statistician* 27 (1973): 17-21.
2. Hoaglin, D. C., and R. Welsch, "The Hat Matrix in Regression and ANOVA," *The American Statistician* 32 (1978): 17-22.
3. Hocking, R. R., "Developments in Linear Regression Methodology: 1959-1982," *Technometrics* 25 (1983): 219-250.
4. Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. (New York: McGraw-Hill/Irwin, 2005).
5. *Microsoft Excel 2007* (Redmond, WA: Microsoft Corp., 2007).