(a) Mostre que para uma tabela genérica com frequências (a, b) na linha 1 e (c, d) na linha 2, o número de pares concordantes é igual a ad, o número de pares discordantes é igual a bc e $Q = (ad - bc)/(ad + bc)$.

(b) Mostre que o valor absoluto de gama é igual a 1 para qualquer tabela 2×2 na qual uma das frequências da célula é 0.

***8.43** Construa uma tabela 3×3 para cada uma das seguintes condições:

(a) O gama é igual a 1. (*Dica*: não deve haver pares discordantes.)

(b) O gama é igual a -1.

(c) O gama é igual a 0.

***8.44** Uma variável qui-quadrado com graus de liberdade iguais a gl tem a representação $z_1^2 + \ldots + z_{gl}^2$, onde $z_1, \ldots, z_{gl}$ são variáveis normais padrão independentes.

(a) Se z é uma estatística-teste que tem uma distribuição normal padrão, que distribuição tem z^2?

(b) Explique como obter os valores do qui-quadrado para $gl = 1$ na Tabela C, dos escores-z, da tabela normal padrão (Tabela A). Ilustre com o valor do qui-quadrado de 6,63 que tem um valor-p de 0,01.

(c) A estatística qui-quadrado para testar H_0: independência entre a crença na vida após a morte (sim, não) e felicidade (pouco feliz, feliz, muito feliz) é χ_1^2 em uma tabela 2×3 para homens e χ_2^2 em uma tabela 2×3 para mulheres. Se H_0 é verdadeiro para cada gênero, então qual é a distribuição de probabilidade de $\chi_1^2 + \chi_2^2$?

***8.45** Para uma tabela 2×2 com frequências a, b, c, d, o logaritmo da razão de chances da amostra $\log\hat{\theta}$ têm uma distribuição amostral aproximadamente normal com erro padrão estimado igual a:

$$ep = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

Os anti-logaritmos dos extremos do intervalo de confiança para $\log(\theta)$ são os extremos do intervalo de confiança para θ. Para a Tabela 8.13 da página 267, mostre que $ep = 0{,}0833$ e um intervalo de 95% de confiança para a razão de chances é (67,3; 93,2). Interprete.

☑ NOTAS

[1] KOHUT, A., STOKES, B. *America Against the World: How We Are Different and Why We Are Disliked*, Times Books, 2006.

[2] CRYSDALE, S. *Intern. J. Compar. Social*, v. 16, p. 19-36, 1975.

[3] HOGE, C. et al. *New England J. Medic*, v.. 351, p. 13-21, 2004.

☑ 9 REGRESSÃO LINEAR E CORRELAÇÃO

O Capítulo 8 apresentou métodos para analisar a associação entre variáveis categóricas resposta e explicativa. Este capítulo apresenta métodos para analisar variáveis quantitativas resposta e explicativa.

A Tabela 9.1 mostra dados do *Statistical Abstract of the United States* (Resumo Estatístico dos Estados Unidos) para os 50 estados e o Distrito de Columbia (D.C.) no que segue:

- Taxa de assassinato: o número de assassinatos por 100000 habitantes.
- Taxa de crimes violentos: o número de assassinatos, estupros violentos, assaltos e agressão com circunstâncias agravantes por 100000 habitantes.
- Percentual da população com renda abaixo do nível de pobreza
- Percentual de famílias chefiadas por um único progenitor.

Para essas variáveis quantitativas, a taxa de crimes violentos e a taxa de assassinatos são variáveis respostas naturais. Trataremos a taxa de pobreza e o percentual de famílias com um único progenitor como variáveis explicativas para estas respostas à medida que estudamos os métodos para analisar os relacionamentos entre variáveis quantitativas neste capítulo e em alguns exercícios. O *site* do livro contém dois conjuntos de dados sobre estas e outras variáveis que iremos analisar em exercícios neste e em capítulos posteriores.

Analisamos três aspectos diferentes, porém relacionados, de tais relacionamentos:

1. Investigamos *se existe uma associação* entre as variáveis, testando a hipótese de independência estatística.
2. Estudamos a *força de sua associação* usando a medida de *correlação* da associação.
3. Estimamos a *equação da regressão* que prevê o valor da variável resposta a partir do valor da variável explicativa. Por exemplo, tal equação prevê a taxa de assassinatos do estado usando o percentual da população que está vivendo abaixo do nível de pobreza.

As análises são coletivamente chamadas de **análises de regressão**. A Seção 9.1 mostra como usar uma linha reta para a equação de regressão, e a Seção 9.2 mostra como usar os dados para estimar essa linha. A Seção 9.3 introduz o *modelo de regressão linear*, que leva em consideração a variabilidade dos dados em torno da linha de regressão. A Seção 9.4 usa a *correlação* e o seu quadrado para descrever a força da associação. A Seção 9.5 apresenta a inferência estatística para uma análise de regressão. A seção final faz um exame minucioso das associações e os riscos potenciais no uso da regressão.

☑ **Tabela 9.1** Dados de cada estado usados para ilustrar a análise de regressão

Estado	Crimes violentos	Taxa de assassinatos	Taxa de pobreza	Pai/mãe solteiro(a)	Estado	Crimes violentos	Taxa de assassinatos	Taxa de pobreza	Pai/mãe solteiro(a)
AK	761	9,0	9,1	14,3	MT	178	3,0	14,9	10,8
AL	780	11,6	17,4	11,5	NC	679	11,3	14,4	11,1
AR	593	10,2	20,0	10,7	ND	82	1,7	11,2	8,4
AZ	715	8,6	15,4	12,1	NE	339	3,9	10,3	9,4
CA	1078	13,1	18,2	12,5	NH	138	2,0	9,9	9,2
CO	567	5,8	9,9	12,1	NJ	627	5,3	10,9	9,6
CT	456	6,3	8,5	10,1	NM	930	8,0	17,4	13,8
DE	686	5,0	10,2	11,4	NV	875	10,4	9,8	12,4
FL	1206	8,9	17,8	10,6	NY	1074	13,38	16,4	12,7
GA	723	11,4	13,5	13,0	OH	504	6,0	13,0	11,4
HI	261	3,8	8,0	9,1	OK	635	8,4	19,9	11,1
IA	326	2,3	10,3	9,0	OR	503	4,6	11,8	11,3
ID	282	2,9	13,1	9,5	PA	418	6,8	13,2	9,6
IL	960	11,42	13,6	11,5	RI	402	3,9	11,2	10,8
IN	489	7,5	12,2	10,8	SC	1023	10,3	18,7	12,3
KS	496	6,4	13,1	9,9	SD	208	3,4	14,2	9,4
KY	463	6,6	20,4	10,6	TN	766	10,2	19,6	11,2
LA	1062	20,3	26,4	14,9	TX	762	11,9	17,4	11,8
MA	805	3,9	10,7	10,9	UT	301	3,1	10,7	10,0
MD	998	12,7	9,7	12,0	VA	372	8,3	9,7	10,3
ME	126	1,6	10,7	10,6	VT	114	3,6	10,0	11,0
MI	792	9,8	15,4	13,0	WA	515	5,2	12,1	11,7
MN	327	3,4	11,6	9,9	WI	264	4,4	12,6	10,4
MO	744	11,3	16,1	10,9	WV	208	6,9	22,2	9,4
MS	434	13,5	24,7	14,7	WY	286	3,4	13,3	10,8
					DC	2922	78,5	26,4	22,1

9.1 RELACIONAMENTOS LINEARES

☑ **Notação para variáveis respostas e explicativas**

Considere y a representação da variável *resposta* e considere x a representação da variável *explicativa*.

Analisaremos como os valores de y tendem a mudar de um subconjunto da população para outro, como definido pelos valores de x. Para variáveis categóricas, fizemos isso comparando as distribuições condicionais de y nas várias categorias de x, em uma tabela de contingência. Para variáveis quantitativas, uma fórmula matemática descreve como a distribuição condicional de y varia de acordo com o valor de x. Esta fórmula descreve como y = assassinatos estaduais, varia de acordo com o nível de x = percentual do nível de pobreza. A taxa de assassinatos tende a ser maior para estados que têm níveis maiores de pobreza?

Funções lineares

Qualquer fórmula em particular pode fornecer uma boa ou uma péssima descrição de como y se relaciona com x. Este capítulo introduz um tipo de fórmula mais simples – uma *linha reta*. Para ela, é dito que y é a **função linear** de x.

☑ **Função linear**

A fórmula $y = \alpha + \beta x$ expressa as observações em y como uma **função linear** das observações em x. A fórmula tem um gráfico de uma reta com **inclinação** β (beta) e **intercepto-y** α (alfa).

EXEMPLO 9.1 Exemplo de uma função linear

A fórmula $y = 3 + 2x$ é uma função linear. Ela tem a forma de $y = \alpha + \beta x$ com $\alpha = 3$ e $\beta = 2$. O intercepto-y é igual a 3 e a inclinação é igual a 2.

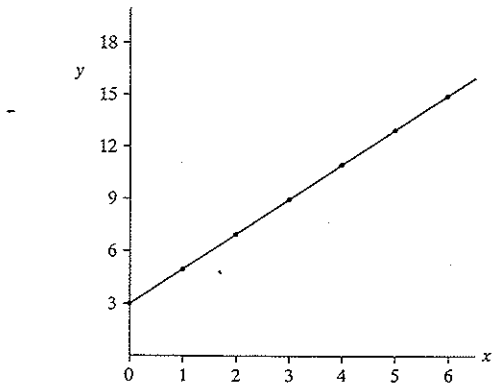
Cada número real x, quando substituído na fórmula $y = 3 + 2x$, gera um valor distinto de y. Por exemplo, $x = 0$ tem $y = 3 + 2(0) = 3$, e $x = 1$ tem $y = 3 + 2(1) = 5$. A Figura 9.1 representa graficamente essa função. O eixo horizontal, o **eixo x**, lista os valores possíveis de x. O eixo vertical, o **eixo y**, lista os valores possíveis de y. Os eixos se cruzam no ponto onde $x = 0$ e $y = 0$, chamado de *origem*. ■

Interpretando o intercepto e a inclinação

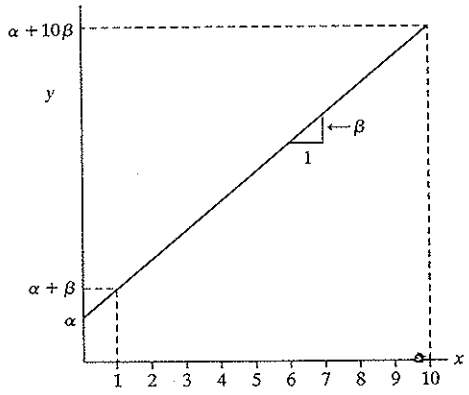
Em $x = 0$, a equação $y = \alpha + \beta x$ é simplificada a $y = \alpha + \beta x = \alpha + \beta(0) = \alpha$. Assim, a constante α nesta equação é o valor de y quando $x = 0$. Agora, os pontos no eixo y têm $x = 0$, assim a linha tem altura α no ponto da sua interseção com o eixo y. Por causa disto, α é chamada de **intercepto-y.** A linha $y = 3 + 2x$ intercepta o eixo y em $\alpha = 3$, como a Figura 9.1 mostra.

A **inclinação** β iguala a mudança em y para o aumento de uma unidade em x. Isto é, para dois valores-x que diferem por 1,0 (como $x = 0$ e $x = 1$), os valores-y diferem por β. Para a linha $y = 3 + 2x$, $y = 3$ em $x = 0$ e $y = 5$ em $x = 1$. Estes valores-y diferem por $\beta = 5 - 3 = 2$. Dois valores-x que estão 10 unidades à parte diferem por 10β nos seus valores-y. Por exemplo, quando $x = 0$, $y = 3$ e quando $x = 10$, $y = 3 + 2(10) = 23$ e $23 - 3 = 20 = 10\beta$. A Figura 9.2 mostra a intercepção do intercepto e da inclinação.

Para traçar a linha, encontramos quaisquer dois pares separados dos valores



☑ **Figura 9.1** Gráfico da reta y = 3 + 2x. O intercepto é 3 e a inclinação é 2.



☑ **Figura 9.2** Gráfico da reta y = α + βx. O intercepto é α e a inclinação é β.

(x, y) no gráfico e, então, traçamos a linha através dos pontos. Para ilustrar, vamos usar os pontos recém discutidos: ($x = 0$, $y = 3$) e ($x = 1$, $y = 5$). O ponto no gráfico com ($x = 0$, $y = 3$) está três unidades acima no eixo y. Para encontrar o ponto ($x = 1$, $y = 5$), iniciamos na origem ($x = 0$, $y = 0$) e movemos uma unidade à direita no eixo x e cinco unidades para cima no eixo y (veja Figura 9.1). Após assinalar os dois pontos traçamos a linha através dos dois pontos representando graficamente a função $y = 3 + 2x$.

EXEMPLO 9.2 Retas para prever crimes violentos

Para os 50 estados, considere as variáveis y = taxa de crimes violentos e x = taxa de pobreza. Veremos que a linha $y = 210 + 25x$ aproxima a relação. O intercepto-y é igual a 210. Isto representa a taxa de crimes violentos quando a taxa de pobreza é $x = 0$ (infelizmente, não existem tais estados). A inclinação é igual a 25. Quando o percentual com renda abaixo do nível de pobreza aumenta de uma unidade, a taxa de crimes violentos aumenta aproximadamente 25 crimes por ano para uma população de 100000 habitantes.

Por outro lado, se ao contrário x = percentual da população residindo em áreas urbanas, a linha aproximando o relacionamento é $y = 26 + 8x$. A inclinação de 8 é menor do que a inclinação de 25 quando a taxa de pobreza é o previsor. Um aumento de uma unidade no nível abaixo da pobreza corresponde a uma chance maior de crimes violentos do que um aumento de um no percentual urbano. A Figura 9.3 mostra as linhas relacionando a taxa de crimes violentos à taxa de pobreza e as residências urbanas. Geralmente, quanto maior o valor absoluto de β, mais inclinada será a linha. ■

Se β é positivo, então y *aumenta* à medida que x *aumenta* – a linha reta sobe, como as duas linhas recém mencionadas. Nesse caso, grandes valores de y ocorrem com grandes valores de x, e pequenos valores de y ocorrem com pequenos valores de x. Quando um relacionamento entre duas variáveis segue uma reta com $\beta > 0$, o relacionamento é dito **positivo**.

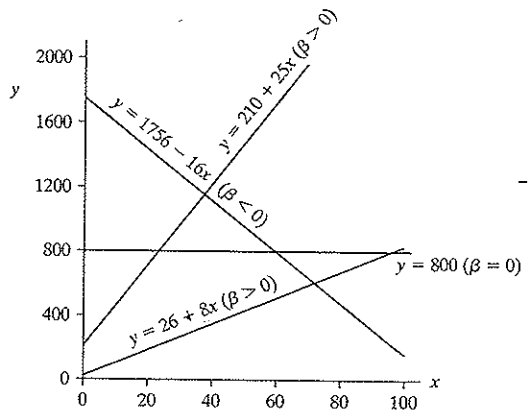
Se β é negativo, então y *diminui* à medida que x *aumenta*. A linha, então, desce e o relacionamento é dito **negativo**. Por exemplo, a equação $y = 1756 - 16x$, que tem inclinação igual a −16, aproxima o relacionamento entre y = taxa de crimes violentos e x = percentual de residentes que tem o ensino médio completo. Para cada aumento de um no percentual de quem tem o ensino médio completo, a taxa de crimes violentos diminui por aproximadamente 16. A Figura 9.3 também mostra esta linha.

Quando $\beta = 0$, o gráfico é uma linha horizontal. O valor de y é constante e não varia à medida que x varia. Se duas variáveis são independentes, com o valor de y não dependendo do valor de x, uma linha reta com $\beta = 0$ representa esse relacionamento. A linha $y = 800$ exibida na Figura 9.3 é um exemplo de uma linha com $\beta = 0$.

Modelos são aproximações simples para a realidade

Como a Seção 7.3 (página 219) explicou, um **modelo** é uma aproximação simples para o relacionamento entre as variáveis na população. A função linear é a função matemática mais simples. Ela fornece o modelo mais simples para o relacionamento entre duas variáveis quantitativas. Para um dado valor de x, o modelo $y = \alpha + \beta x$ prevê um valor para y. Quanto melhor forem essas previsões, melhor será o modelo.

Como mencionamos na Seção 3.4 (página 68) e iremos explicar mais tarde no início do Capítulo 10, uma *associação não implica causação*. Por exemplo, considere a interpretação da inclinação do Exemplo 9.2 de "quando o percentual com renda abaixo do nível de pobreza aumenta uma unidade, a taxa de crimes violentos aumenta aproximadamente 25 crimes por ano para uma população de 100000 habitantes." Isto não significa que se tivermos a habilidade de ir a um estado e aumentar o percentual de pessoas que vivem abaixo do nível de pobreza de 10% para 11%, poderíamos esperar que o número de crimes aumentasse no próximo ano em 25 crimes por 100000 habitantes. Simplesmente significa que, baseado em dados atuais, se um estado tinha uma taxa de pobreza de 10% e um tinha de 11%, iríamos prever que o estado com a taxa de pobreza mais alta teria 25 crimes a mais por ano por 100000 habitantes. Mas, como veremos na Seção 9.3, um modelo concreto é, na verdade, um pouco mais complexo do que o que apresentamos até agora.



☑ **Figura 9.3** Gráfico de linhas exibindo relacionamentos positivos (β > 0), um relacionamento negativo (β < 0) e independência (β = 0).

9.2 EQUAÇÃO DE PREVISÃO PELOS MÍNIMOS QUADRADOS

Usando dados amostrais, podemos estimar o modelo linear. O processo trata α e β na *equação* $y = \alpha + \beta x$ como parâmetros desconhecidos e os estima. A função linear estimada, então, fornece os valores-y previstos para valores fixos de x.

Um diagrama de dispersão exibe os dados

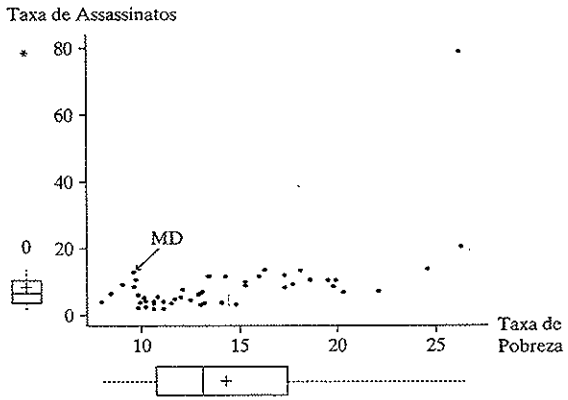
A primeira etapa no ajuste do modelo é fazer um gráfico dos dados para revelar se um modelo com uma tendência linear faz sentido. Os valores (x, y) para qualquer sujeito forma um ponto em relação aos eixos y e x. Um gráfico dos n pares de observações representados como n pontos é denominado de um **diagrama de dispersão**.

EXEMPLO 9.3 Diagrama de dispersão entre a taxa de assassinatos estadual e o nível de pobreza

Para a Tabela 9.1, considere x = taxa de pobreza e y = taxa de assassinatos. Para verificar se uma reta descreve bem o relacionamento, inicialmente construímos um diagrama de dispersão para as 51 observações. A Figura 9.4 mostra esse gráfico.

Cada ponto na Figura 9.4 exibe os valores da taxa de pobreza e de assassinatos para um dado estado. Para Maryland, por exemplo, a taxa de pobreza é $x = 9{,}7$ e a de assassinatos é $y = 12{,}7$. Seu ponto $(x, y) = (9{,}7; 12{,}7)$ tem a coordenada 9,7 para o eixo x e 12,7 para o eixo y. Este ponto é rotulado como MD na Figura 9.4.

A Figura 9.4 indica que a tendência dos pontos parece ser bem aproximada por uma linha. Observe, entretanto, que um ponto está bem longe do resto. Este é o ponto do Distrito de Columbia (D.C.). Para ele, a taxa de assassinatos é muito mais alta do que para qualquer outro estado. Este ponto está longe da tendência geral. A Figura 9.4, também, mostra diagramas de caixa e bigodes para as variáveis. Eles revelam que o D.C. é um *valor atípico* extremo da taxa de assassinatos. Na verdade, ele está 6,5 desvios padrão acima da média. Iremos ver que os valores atípicos podem causar um impacto sério no modelo de regressão. ■



☑ **Figura 9.4** Diagrama de dispersão para y = taxa de assassinatos e x = percentual de residentes abaixo do nível de pobreza, para os 50 estados mais o D.C., diagramas de caixa e bigodes são exibidos para a taxa de assassinatos à esquerda do diagrama de dispersão e para a taxa de pobreza abaixo do diagrama de dispersão.

O diagrama de dispersão fornece um auxílio visual para verificarmos se um relacionamento é aproximadamente linear. Quando o relacionamento parece ser altamente não linear, não é lógico usar um modelo linear. A Figura 9.5 ilustra este caso. Esta figura mostra um relacionamento negativo sobre parte do intervalo de valores x e um relacionamento positivo sobre o restante. Eles cancelam um ao outro se for utilizado um modelo linear. Para tais dados, um modelo diferente, apresentado na Seção 14.5, é mais apropriado.

Equação de previsão

Quando o diagrama de dispersão sugere que o modelo $y = \alpha + \beta x$ é realístico, usamos os dados para estimar esta linha. A notação

$$\hat{y} = a + bx$$

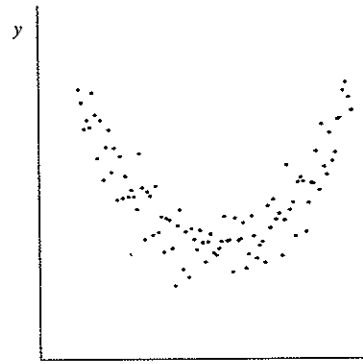
representa a equação *amostral* que estima o modelo linear populacional. Na equação amostral, o intercepto-y (a) estima o intercepto-y α do modelo e a inclinação (b) estima a inclinação β. Substituindo um valor-x em particular em $a + bx$, temos um valor, representado por $\hat{y}$, que prevê y para aquele valor de x. A equação amostral $\hat{y} = a + bx$ é chamada de **equação de previsão**, porque fornece uma previsão $\hat{y}$ para a variável resposta, dado qualquer valor de x.

A equação da previsão é a melhor linha reta, estando mais próxima dos pontos do diagrama de dispersão, em um sentido discutido mais tarde nesta seção. As fórmulas para a e b na equação de previsão $\hat{y} = a + bx$ são:

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}.$$

Se qualquer observação tem ambos os valores x e y acima das suas médias ou ambos os valores x e y abaixo das suas médias, então $(x - \bar{x})(y - \bar{y})$ é positivo. A estimativa da inclinação b tende a ser positiva quando a maioria das observações são como essa, isto é, quando pontos com valores-x grandes também tendem a ter valores-y grandes e pontos com valores-x pequenos tendem a ter valores-y pequenos. A Figura 9.4 é um exemplo de tal caso.

Não nos estenderemos nessas fórmulas ou até mesmo ilustraremos como usá-las, porque elas são complicadas para



☑ **Figura 9.5** Relacionamento não linear, para o qual é inapropriado o uso de um modelo linear.

cálculos manuais. Qualquer um que leva a sério a modelagem por regressão usa o computador ou uma calculadora que tem essas fórmulas programadas. Para usar um *software* estatístico, você fornece o arquivo de dados e geralmente seleciona o método de regressão a partir de um menu. O apêndice do final do livro fornece os detalhes.

EXEMPLO 9.4 Prevendo a taxa de assassinatos a partir da taxa de pobreza

Das 51 observações em y = taxa de assassinatos e x = taxa de pobreza da Tabela 9.1, o SPSS fornece os resultados mostrados na Tabela 9.2. A taxa de assassinatos tem $\bar{y} = 8{,}7$ e $s = 10{,}7$ indicando que é, provavelmente, altamente assimétrica à direita. O diagrama de caixa e bigodes para a taxa de assassinato na Figura 9.4 mostra que a observação atípica extrema para D.C. contribuiu para isso.

As estimativas de α e β estão listadas sob o título "B", o símbolo que o SPSS usa para representar um coeficiente de regressão estimado. O intercepto-y estimado é $\alpha = -10{,}14$, listado como "(Constante)". A estimativa da inclinação é $b = 1{,}32$, listada no nome da variável da qual ela é o coeficiente na equação de previsão, "POBREZA". Portanto, a equação de previsão é $\hat{y} = a + bx = -10{,}14 + 1{,}32x$.

A inclinação $b = 1{,}32$ é positiva. Assim, quanto maior a taxa de pobreza, maior é a taxa de assassinatos prevista. O valor 1,32 indica que um aumento de um no percentual dos que vivem abaixo do nível de pobreza corresponde a um aumento de 1,32 na taxa prevista de assassinatos.

De forma similar, um aumento de 10 na taxa de pobreza corresponde a 10(1,32) = 13,2 – unidades de aumento na taxa prevista de assassinatos. Se um estado tem 12% de taxa de pobreza e outro tem 22%, por exemplo, o número anual de assassinatos previstos para uma população de 100000 habitantes é de 13,2 a mais no segundo estado do que no primeiro estado. Visto que a taxa média de assassinatos é de 8,7, parece que a taxa de pobreza é um previsor importante da taxa de assassinatos. Este diferencial de 13 assassinatos por uma população de 100000 habitantes se traduz a 130 por milhão ou 1300 para uma população de 10 milhões. Se cada um dos dois estados tivesse uma população de 10 milhões, aquele com a taxa de pobreza mais alta teria a previsão de ter 1300 assassinatos a mais por ano. ■

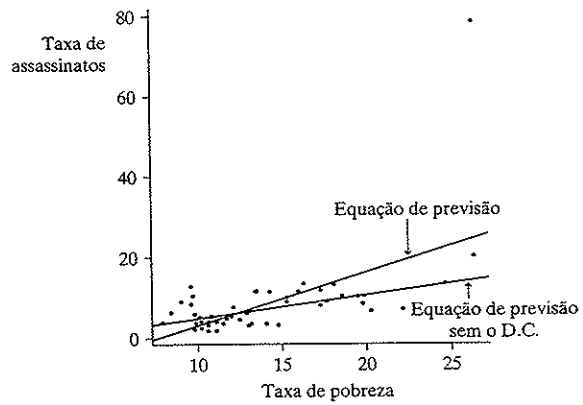
Efeito dos valores atípicos na equação de previsão

A Figura 9.6 traça a equação de previsão do Exemplo 9.4 sobre o diagrama de dispersão. O diagrama mostra que a observação para o D.C. é um **valor atípico da regressão** – ele está bem longe da tendência seguida pelo restante dos dados. Esta observação parece ter um efeito substancial. A linha parece ter sido puxada em sua direção e longe do centro da tendência geral dos pontos.

Vamos agora ajustar novamente a linha usando as observações para os 50 estados, mas deixando fora o D.C. A Tabela 9.3 mostra que a equação de previsão é igual a $\hat{y} = -0.86 + 0{,}58x$. A Figura 9.6 também mostra essa linha, que passa mais

☑ **Tabela 9.2** Parte da saída do SPSS para ajustar o modelo de regressão linear às observações dos 50 estados mais o D.C. em x = percentual de pobreza e y = taxa de assassinatos

Variável	Média	Desvio padrão		B	Erro padrão
ASSASSINATOS	8,727	10.718	(Constante)	-10,1364	4,1206
POBREZA	14,259	4.584	Pobreza	1,3230	0,2754



☑ **Figura 9.6** Equações de previsão relacionando a taxa de assassinatos e o percentual de pobreza, com e sem a observação para o D.C.

diretamente sobre os 50 pontos. A inclinação é de 0,58, comparada a 1,32 quando a observação para o D.C. é incluída. A observação atípica tem o impacto de mais do que dobrar a inclinação!

Uma observação é chamada de **influente** se sua remoção resulta em uma grande mudança na equação de previsão. A não ser que o tamanho da amostra seja grande, uma observação pode ter uma forte influência na inclinação se o seu valor-x for baixo ou alto comparado com o restante dos dados e se for um valor atípico da regressão.

Em resumo, a linha para o conjunto de dados incluindo o D.C. parece distorcer o relacionamento para os 50 estados. Parece ser mais prudente usar a equação baseada nos dados para somente os 50 estados do que usar uma única equação para os 50 estados mais o D.C. Esta linha para os 50 estados representa melhor a

☑ **Tabela 9.3** Parte da saída para ajustar modelos lineares para os 50 estados (sem o D.C.) em x = percentual de pobreza e y = taxa de assassinatos

	Soma dos quadrados	gl	Média dos quadrados	Coeficientes não padronizados
				B
Regressão	307,342	1	307,34	(Constante) −0,857
Resíduos	470,406	48	9,80	POBREZA 0,584
Total	777,749	49		

	ASSASSINATOS	PREVISÃO	RESÍDUO
1	9,0000	4,4599	4,5401
2	11,6000	9,3091	2,2909
3	10,2000	10,8281	−0,6281
4	8,6000	8,1406	0,4594

tendência geral. No relato desses resultados, veríamos que a taxa de assassinatos para o D.C. está fora desta tendência, sendo muito maior do que essa equação prevê.

Erros de previsão são chamados de resíduos

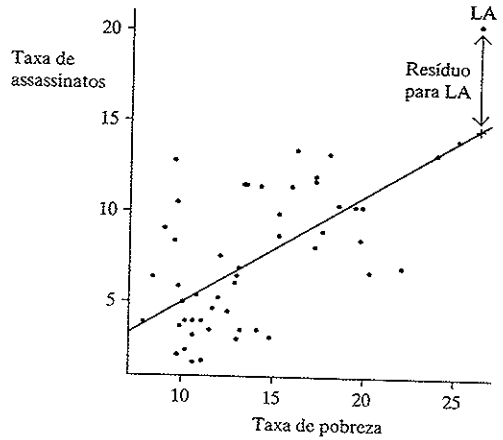
A equação de previsão $\hat{y} = -0{,}86 + 0{,}58x$ prevê as taxas de assassinato usando x = taxa de pobreza. Para os dados amostrais, uma comparação das taxas de assassinato *reais* com os valores *previstos* verifica a qualidade da equação de previsão.

Por exemplo, Massachusetts tinha $x = 10{,}7$ e $y = 3{,}9$. A taxa de assassinatos prevista ($\hat{y}$) para $x = 10{,}7$ é $\hat{y} = -0{,}86 + 0{,}58x = -0{,}86 + 0{,}58(10{,}7) = 5{,}4$. O erro de previsão é a diferença entre o valor-y real que é 3,9 e o valor previsto que é 5,4, ou $y - \hat{y} = 3{,}9 - 5{,}4 = -1{,}5$. A equação de previsão superestimou a taxa de assassinato por 1,5. De forma similar, para Louisiana, $x = 26{,}4$ e $\hat{y} = -0{,}86 + 0{,}58(26{,}4) = 14{,}6$. A taxa de assassinatos real é $y = 20{,}3$, portanto a previsão é muito baixa. O erro de previsão é $y - \hat{y} = 20{,}3 - 14{,}6 = 5{,}7$. Os erros de previsão são chamados de **resíduos**.

> ☑ **Resíduos**
>
> Para uma observação, a diferença entre um valor observado e o valor previsto da variável resposta, $y - \hat{y}$, é denominado **resíduo**.

A Tabela 9.3 mostra as taxas de assassinato, os valores previstos e os resíduos para os quatro primeiros estados do arquivo de dados. Um resíduo *positivo* resulta quando o valor observado y é *maior* do que o valor previsto $\hat{y}$, assim $y - \hat{y} > 0$. Um resíduo *negativo* resulta quando o valor observado é menor do que o valor previsto. Quanto menor o valor absoluto do resíduo, melhor é a previsão, visto que o valor previsto está mais próximo do valor observado.

Em um diagrama de dispersão, o resíduo para uma observação é a distância vertical entre seu ponto e a linha de previsão. A Figura 9.7 ilustra isso. Por exemplo, a observação para Louisiana é o ponto (x, y) com coordenadas (26,4; 20,3). A previsão é representada pelo ponto (26,4; 14,6) na linha da previsão obtida substituindo a diferença entre o ponto observado e o previsto, que é a distância vertical $y - \hat{y} = 20{,}3 - 14{,}6 = 5{,}7$.



☑ **Figura 9.7** Equações de previsão e resíduos. Um resíduo é uma distância vertical entre um ponto e a linha de previsão.

A equação de previsão tem a propriedade dos mínimos quadrados

Cada observação tem um resíduo. Se a linha de previsão está próxima dos pontos no diagrama de dispersão, os resíduos são pequenos. Resumimos o tamanho dos resíduos pela soma dos seus valores ao quadrado. Esta quantidade, representada por SQE, é igual a:

$$\text{SQE} = \sum(y - \hat{y})^2.$$

Em outras palavras, o resíduo é calculado para cada observação da amostra, cada resíduo é elevado ao quadrado e, então, SQE é a soma desses quadrados. O símbolo SQE é uma abreviação para **soma dos quadrados dos erros**. Esta terminologia se refere ao resíduo sendo uma medida de prever o erro usando $\hat{y}$ para prever y. Alguns *softwares* (como o SPSS) chamam o SQE de **soma dos quadrados dos resíduos**. Ela é a variação dos dados em torno da linha de previsão.

Quanto melhor a equação de previsão, menores tendem a ser os resíduos e, portanto, menor tende a ser a SQE. Qualquer equação em particular tem resíduos correspondentes e um valor da SQE. A equação de previsão especificada pelas fórmulas para as estimativas a e b de α e β tem o *menor* valor da SQE de todas as equações possíveis de previsão lineares.

> ☑ **Estimativas dos mínimos quadrados**
>
> As **estimativas dos mínimos quadrados** a e b são os valores que fornecem a equação de previsão $\hat{y} = a + bx$ para a qual a soma dos quadrados dos resíduos (erros), $\text{SQE} = \sum(y - \hat{y})^2$, é um mínimo.

A linha de previsão $\hat{y} = a + bx$ é chamada de **linha dos mínimos quadrados**, porque ela é a linha com a menor soma dos quadrados dos resíduos. Se elevarmos os resíduos ao quadrado (como os da Tabela 9.3) para a linha dos mínimos quadrados $\hat{y} = -0{,}86 + 0{,}58x$ e, então os somamos, obtemos:

$$\text{SQE} = \sum(y - \hat{y})^2 = (4{,}54)^2 + (2{,}29)^2 + \cdots = 470{,}4.$$

Este valor é menor do que o valor do SQE para *qualquer* outra linha reta prevista, como $\hat{y} = -0{,}88 + 0{,}60x$. Neste sentido, os dados estão mais próximos desta linha do que de *qualquer* outra linha.

O *software* para a regressão lista o valor do SQE. A Tabela 9.3 relata-o na coluna *Sum of Squares* (Soma dos Quadrados), na linha rotulada de *Residual* (Resíduos). Em alguns *softwares*, como o SAS, isto é rotulado como *Error* na coluna da soma dos quadrados.

Além de tornar os erros tão pequenos quanto possível no sentido de uma medida resumo, a linha dos mínimos quadrados:

- Tem alguns resíduos positivos e alguns negativos, mas a soma (e a média) dos resíduos é igual a 0.
- Passa pelo ponto $(\bar{x}, \bar{y})$.

A primeira propriedade nos diz que as previsões muito baixas são contrabalançadas pelas previsões muito altas. Assim como os desvios das observações da média $\bar{y}$ satisfazem $\sum(y - \bar{y}) = 0$, assim é a equação de previsão definida para que $\sum(y - \hat{y}) = 0$. A segunda propriedade nos diz que a linha passa pelo centro dos dados.

9.3 O MODELO DE REGRESSÃO LINEAR

Para o modelo $y = \alpha + \beta x$, a cada valor de x corresponde um único valor de y. Tal modelo é dito ser **determinista**. Ele não é re-

alista na pesquisa em ciências sociais porque não esperamos que todos os sujeitos que têm o mesmo valor-x tenham o mesmo valor-y. Ao contrário, os valores-y variam.

Por exemplo, considere x = número de anos de escolaridade e y = renda anual. Os sujeitos que têm $x = 12$ anos de escolaridade não têm a mesma renda porque a renda não depende apenas da educação. Ao contrário, a distribuição de probabilidade descreve a renda anual para indivíduos com $x = 12$. A distribuição se refere à variabilidade nos valores y para um valor *fixo* de x, assim ela é uma **distribuição condicional.** Uma distribuição condicional separada se aplica para aquelas com outros valores de x. Cada nível de escolaridade tem a sua própria distribuição de renda. Por exemplo, a média da distribuição condicional da renda seria provavelmente mais alta nos níveis mais altos de escolaridade.

Um modelo **probabilístico** para o relacionamento leva em consideração a variabilidade de y para cada valor de x. Mostraremos, agora, como uma função linear é a base para um modelo probabilístico.

Função de regressão linear

Um modelo probabilístico usa $\alpha + \beta x$ para representar a *média* dos valores-y, em vez do próprio y, como uma função de x. Para um valor dado de x, $\alpha + \beta x$ representa a média da distribuição condicional de y para sujeitos tendo aquele valor de x.

> ☑ **Valor esperado de *y***
>
> Considere $E(y)$ a representação da média da distribuição condicional de y. O símbolo E representa o *valor esperado*, que é outro termo para a *média*.

Agora, usamos a equação

$$E(y) = \alpha + \beta x$$

para modelar o relacionamento entre x e a média da distribuição condicional de y. Para y = renda anual, em dólares, e x = número de anos de escolaridade, suponha que $E(y) = -5000 + 3000x$. Por exemplo, aqueles com ensino médio completo ($x = 12$) têm uma renda média de $E(y) = -5000 + 3000(12) = 31000$ dólares. O modelo determina que a renda *média* é de 31000, em vez de determinar que *cada* sujeito com $x = 12$ tem uma renda de 31000 dólares. O modelo permite que diferentes sujeitos com 12 anos de escolaridade ($x = 12$) tenham diferentes níveis de renda.

Uma equação da forma $E(y) = \alpha + \beta x$ que relaciona os valores de x à média da distribuição condicional de y é chamada de *função da regressão.*

> ☑ **Função da regressão**
>
> Uma **função da regressão** descreve como a média da variável resposta muda de acordo com o valor da variável explicativa.

A função $E(y) = \alpha + \beta x$ é chamada de função de regressão *linear* porque usa uma linha reta para relacionar a média de y aos valores de x. O intercepto-y α e a inclinação β são chamados de **coeficientes da regressão** para a função de regressão linear.

Na prática, os parâmetros da função de regressão linear são desconhecidos. Os mínimos quadrados fornecem a equação de previsão com base na amostra $\hat{y} = a + bx$. Para um valor fixo de x, $\hat{y} = a + bx$ *estima* a média de y para todos os sujeitos na população tendo aquele valor de x.

Descrevendo a variação sobre a linha da regressão

O modelo de regressão linear tem um parâmetro adicional σ descrevendo o desvio padrão de cada distribuição condicional. Isto é, σ mensura a variabilidade dos valores y para todos os sujeitos tendo o mesmo valor-x. Nós nos referimos a σ como o **desvio padrão condicional.**

Um modelo também assume uma distribuição da probabilidade particular para a distribuição condicional de y. Isto é necessário para fazer inferência sobre os parâmetros. Para variáveis quantitativas, a suposição mais comum é que a distribuição condicional de y é normal para cada valor fixo de x.

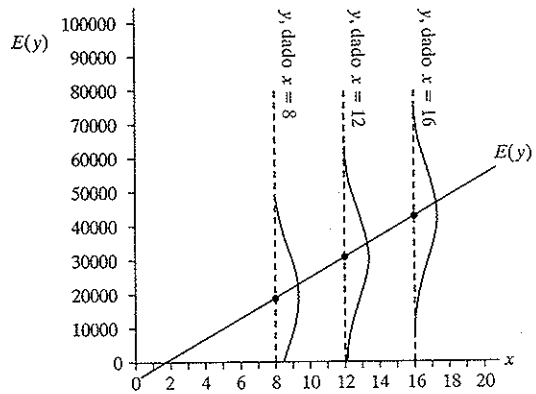
EXEMPLO 9.5 Descrevendo como a renda varia para um dado nível de escolaridade

Novamente, suponha que $E(y) = -5000 + 3000x$ descreve o relacionamento entre a renda média anual e o número de anos de escolaridade. Suponha, também, que a distribuição condicional da renda é normal, com $\sigma = 13000$. De acordo com este modelo, para indivíduos com x anos de escolaridade, suas rendas têm uma distribuição normal com uma média de $E(y) = -5000 + 3000x$ e um desvio padrão de 13000.

Aqueles tendo o ensino médio completo ($x = 12$) têm uma renda média de $E(y) = -5000 + 3000(12) = 31000$ dólares e um desvio padrão de 13000 dólares. Assim, aproximadamente 95% dos rendimentos estão dentro de dois desvios padrão da média, isto é, entre $31000 - 2(13000) = 5000$ e $31000 + 2(13000) = 57000$ dólares. Aqueles com ensino superior completo ($x = 16$) têm uma renda média anual de $E(y) = -5000 + 3000(16) = 43000$ dólares, com aproximadamente 95% das rendas estando entre \$17000 e \$69000.

A inclinação $\beta = 3000$ implica que a renda média aumenta \$3000 para cada ano de aumento da escolaridade. A Figura 9.8 mostra este modelo de regressão. Ela mostra as distribuições condicionais da renda para $x = 8$, 12 e 16 anos. ■

Na Figura 9.8, cada distribuição condicional é normal e cada uma tem o mesmo desvio padrão $\sigma = 13000$. Na prática, as distribuições não seriam exatamente normais e o desvio padrão não precisa ser o mesmo para cada uma. *Nenhum modelo se mantém o mesmo na prática.* Ele é meramente uma simples aproximação da realidade. Para dados amostrais, iremos aprender sobre as formas para verificar se um modelo em particular é realístico. A suposição mais importante é que a equação de regressão é linear. O diagrama de dispersão nos ajuda a verificar se esta suposição é fortemente violada, como iremos discutir mais tarde no capítulo.



☑ **Figura 9.8** O modelo de regressão $E(y) = -5000 + 3000x$, com $\sigma = 13000$, relacionando y = renda (em dólares) e x = escolaridade (em anos).

Erro quadrático médio: estimando a variação condicional

O modelo de regressão linear ordinário assume que o desvio padrão σ da distribuição condicional y é idêntico para os vários valores de x. A estimativa de σ usa o valor numérico da SQE $= \sum(y - \hat{y})^2$, que mensura a variabilidade amostral sobre a linha dos mínimos quadrados. A estimativa é:

$$s = \sqrt{\frac{\text{SQE}}{n-2}} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}.$$

Se a suposição da variação constante não é válida, então s fornece uma medida da variabilidade *média* sobre a linha.

EXEMPLO 9.6 Assistir à televisão e rendimento acadêmico (GPA – *Grade Point Averages*)

Um levantamento de dados[1] com 50 estudantes universitários em uma classe de psicologia introdutória observou autorrelatos de y = GPA do ensino médio e x = número semanal de horas assistindo à televisão. O estudo relatou $\hat{y} = 3{,}44 - 0{,}03x$. Para esses dados, um *software* informa o seguinte:

	Soma dos quadrados	gl	Média dos quadrados
Regressão	3,63	1	3,63
Resíduos	11,66	48	0,24
Total	15,29	49	

A soma dos quadrados dos resíduos ao se utilizar x para prever y foi SQE = 11,66. O desvio padrão condicional estimado é:

$$s = \sqrt{\frac{\text{SQE}}{n-2}} = \sqrt{\frac{11{,}66}{50-2}} = 0{,}49.$$

Para qualquer valor fixo de assistir à televisão (x), o modelo prevê que os GPAs variam em torno da média de $3{,}44 - 0{,}03x$ com um desvio padrão de 0,49. Para x = 20, por exemplo, a distribuição condicional do GPA é estimada como tendo uma média de $3{,}44 - 0{,}03(20) = 2{,}83$ e um desvio padrão de 0,49. ■

O termo $(n - 2)$ no denominador de s são os **graus de liberdade** (gl) para a estimativa. Em geral, quando uma equação de regressão tem p parâmetros desconhecidos, então $gl = n - p$. A equação $E(y) = \alpha + \beta x$ tem dois parâmetros (α e β), assim $gl = n - 2$. A tabela no exemplo anterior lista SQE = 11,66 e seu $gl = n - 2 = 50 - 2 = 48$. A razão destes, $s^2 = 0{,}24$, está listada na saída na coluna denominada "Média dos quadrados". Alguns *softwares* chamam isto de EQM, abreviação para *erro quadrático médio*. Sua raiz quadrada é a estimativa do desvio padrão condicional de y, a saber, $s = \sqrt{0{,}24} = 0{,}49$. (O SPSS lista isto sob um título um tanto equivocado, *Std. Error of the Estimate* [Erro Padrão da Estimativa].)

A variação condicional tende a ser menor do que a variação marginal

Das Seções 3.3 (página 64) e 5.1 (página 131), uma estimativa por ponto do desvio padrão da população de uma variável y é:

$$\sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}.$$

Este é o desvio padrão da distribuição *marginal* de y porque ele usa somente os valores-y. Ele ignora os valores de x. Para enfatizar isto, este desvio padrão depende somente dos valores de y, no restante do livro ele será representado por s_y para uma amostra e σ_y para a população. Ele difere do desvio padrão da distribuição *condicional* de y, para um valor fixo de x.

A soma dos quadrados $\sum(y - \bar{y})^2$ no numerador de s_y é chamada de **soma dos quadrados total (SQT)**. Na tabela anterior para os GPAs dos 50 estudantes, ela é 15,29. Assim, o desvio padrão marginal do GPA é $s_y = \sqrt{15{,}29/(50-1)} = 0{,}56$. O Exemplo 9.6 mostrou que o desvio padrão condicional é 0,49.

Normalmente, uma menor dispersão dos valores-y ocorre em um valor fixo de x do que sobre a totalidade dos valores. Veremos que quanto mais forte a associação entre x e y, menor a variabilidade condicional tende a ser em relação à variabilidade marginal.

Por exemplo, a distribuição *marginal* dos GPAs da universidade (y) na sua faculdade pode, primeiramente, estar entre 1,0 e 4,0. Talvez uma amostra tenha um desvio padrão de $s_y = 0{,}60$. Suponha que você poderia prever *perfeitamente* o GPA da universidade usando x = GPA do ensino médio, com a equação de previsão $\hat{y} = 0{,}40 + 0{,}90x$. Então, SQE seria 0 e o desvio padrão condicional seria $s = 0$. Na prática, uma previsão perfeita não acontece. Entretanto, quanto mais forte a associação em termos de um menor erro de previsão, menor será a variabilidade condicional. Veja a Figura 9.9 que exibe uma distribuição marginal que é muito mais dispersa do que cada distribuição condicional.

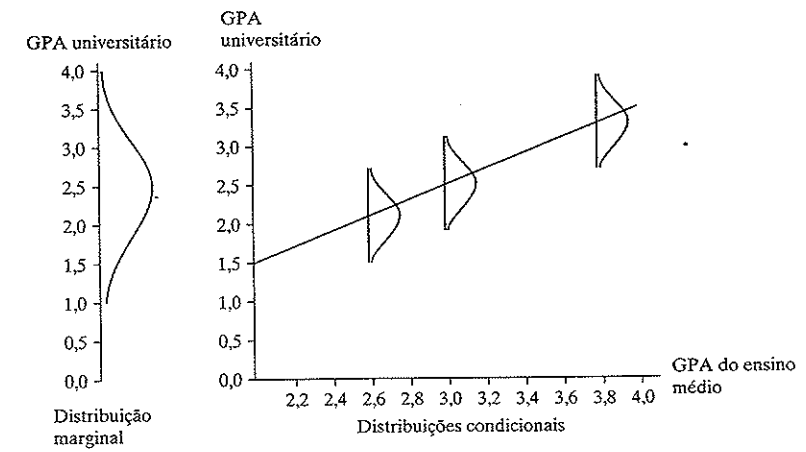
9.4 MENSURANDO A ASSOCIAÇÃO LINEAR: A CORRELAÇÃO

O modelo de regressão linear usa uma linha reta para descrever um relacionamento. Esta seção introduz duas medidas da força da associação entre as variáveis.

A inclinação e a força da associação

A inclinação b da equação de previsão nos diz a *direção* da associação. Seu sinal indica se a linha de previsão está inclinada para cima ou para baixo à medida que x aumenta, isto é, se a associação é positiva ou negativa. A inclinação, entretanto, não nos diz diretamente a força da associação. A razão para isto é que seu valor numérico está intrinsecamente ligado às unidades de mensuração.

Por exemplo, considere a equação de previsão $\hat{y} = 0{,}86 + 0{,}58x$ para y = taxa de assassinatos e x = percentual vivendo abaixo do nível de pobreza. Um aumento de uma unidade em x corresponde a um aumento de $b = 0{,}58$ no número de assassinatos previstos por 1000000 habitantes. Isto é equivalente a um aumento de 5,8



☑ **Figura 9.9** Distribuição marginal e condicional. A distribuição marginal mostra a variabilidade geral nos valores y, enquanto a distribuição condicional mostra como y varia para um valor fixo de x.

no número de assassinatos previstos por 1000000 habitantes. Assim, se a taxa de assassinatos é o número de assassinatos por 1000000 habitantes em vez de por 100000, a inclinação é 5,8 em vez de 0,58. A força da associação é a mesma em cada caso, visto que as variáveis e os dados são os mesmos. Somente as unidades de mensuração para y diferem. Em resumo, a inclinação b não indica diretamente se a associação é forte ou fraca porque podemos tornar b tão grande ou tão pequeno quanto quisermos fazendo uma escolha apropriada de unidades.

A inclinação *é* útil para comparar efeitos de dois previsores tendo as mesmas unidades. Por exemplo, a equação de previsão relacionando a taxa de assassinatos ao percentual vivendo em áreas urbanas é $3{,}28 + 0{,}06x$. Um aumento de uma unidade no percentual vivendo em áreas urbanas corresponde a um aumento previsto de 0,06 na taxa de assassinatos, enquanto um aumento de uma unidade no percentual abaixo do nível de pobreza corresponde a um aumento previsto de 0,58 na taxa de assassinatos. Um aumento de um ponto percentual abaixo do nível de pobreza tem um efeito muito maior na taxa de assassinatos do que um aumento de um ponto no percentual urbano.

As medidas de mensuração que estudamos agora não dependem das unidades de mensuração. Assim como as medidas de mensuração que foram apresentadas para dados categóricos no Capítulo 8, suas magnitudes indicam a força da associação.

A correlação

A Seção 3.5 (página 73) introduziu a **correlação** entre variáveis quantitativas. Esta é uma versão *padronizada* da inclinação. Seu valor, diferente daquele da inclinação ordinária b, não depende das unidades de mensuração. A padronização ajusta a inclinação b para o fato de que os desvios padrão de x e y dependam das suas unidades de mensuração. A correlação é o valor que a inclinação assumiria para unidades tais que as variáveis tenham desvios padrão iguais.

Considere s_x e s_y a representação dos desvios padrão amostrais marginais de x e y:

$$s_x = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \quad \text{e} \quad s_y = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}}.$$

☑ Correlação

A **correlação**, representada por r, se relaciona com a inclinação b da equação de previsão $\hat{y} = a + bx$ por:

$$r = \left(\frac{s_x}{s_y}\right) b.$$

Quando as dispersões da amostra são iguais ($s_x = s_y$), $r = b$. Por exemplo, quando as variáveis são padronizadas convertendo seus valores em escores-z, ambas as variáveis padronizadas têm desvios padrão de 1,0. Por causa do relacionamento entre r e b, a correlação é também chamada de **coeficiente padronizado da regressão** para o modelo $E(y) = \alpha + \beta x$. Na prática, não é necessário padronizar as variáveis, mas geralmente é útil interpretar a correlação como sendo igual ao valor da inclinação se as variáveis fossem igualmente dispersas.

A estimativa por ponto da correlação r foi proposta pelo estatístico britânico Karl Pearson em 1896. Apenas quatro anos antes ele desenvolveu o teste qui-quadrado de independência para tabelas de contingência. Na verdade, esta estimativa é, algumas vezes, chamada de **correlação de Pearson**.

EXEMPLO 9.7 A correlação entre as taxas de assassinatos e de pobreza

Para os dados dos 50 estados na Tabela 9.1, a equação de previsão relacionando y = taxa de assassinatos com x = taxa de pobreza é $\hat{y} = -0{,}86 + 0{,}58x$. O *software* nos diz que $s_x = 4{,}29$ para a taxa de pobreza e $s_y = 3{,}98$ para a taxa de assassinatos. A correlação é igual a:

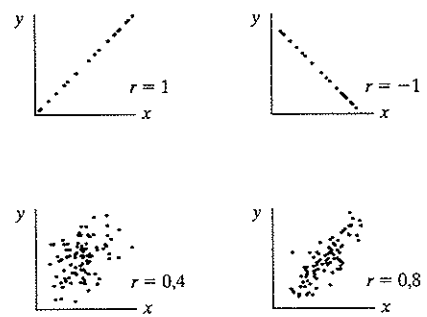
$$r = \left(\frac{s_x}{s_y}\right) b = \left(\frac{4{,}29}{3{,}98}\right)(0{,}58) = 0{,}63.$$

Interpretaremos este valor após estudar as propriedades da correlação. ■

Propriedades da correlação

- A correlação é somente válida quando uma linha reta é um modelo lógico para o relacionamento. Visto que r é proporcional à inclinação de uma equação de previsão linear, ele mensura a *força da associação linear* entre x e y.
- $-1 \leq r \leq 1$. A correlação, diferente da inclinação b, deve estar entre -1 e $+1$. A razão será vista mais tarde na seção.
- O r tem o mesmo sinal da inclinação b. Visto que r é igual a b multiplicado pela razão de dois desvios padrão (positivos), o sinal é preservado. Portanto, $r > 0$ quando as variáveis estão relacionadas positivamente e $r < 0$ quando as variáveis estão relacionadas negativamente.
- $r = 0$ para aquelas linhas tendo $b = 0$. Quando $r = 0$, não existe uma tendência de um aumento ou diminuição linear dos dados.
- $r = \pm 1$ quando todos os pontos da amostra estão exatamente sobre a linha de previsão. Eles correspondem a associações lineares perfeitas tanto positivas quanto negativas. Não existe, então, erro de previsão quando a equação de previsão $\hat{y} = a + bx$ prevê y.
- Quanto maior o valor absoluto de r, mais forte a associação linear. As variáveis com uma correlação de $-0{,}80$ estão mais forte e linearmente associadas do que as variáveis com a correlação de 0,40. A Figura 9.10 mostra diagramas de dispersão tendo vários valores de associação.
- A correlação, diferente da inclinação b, trata x e y simetricamente. A equação de previsão usando y para prever x tem a mesma correlação daquela usando x para prever y.
- O valor de r não depende das unidades das variáveis.

Por exemplo, se y é o número de assassinatos por 1000000 de habitantes, em vez de uma população de 100000, obtemos o mesmo valor $r = 0{,}63$. Da mesma forma, quando a taxa de assassinatos prevê a taxa de pobreza, a correlação é a mesma quando a taxa de pobreza prevê a taxa de assassinatos, $r = 0{,}63$ em ambos os casos.



☑ **Figura 9.10** Diagramas de dispersão para diferentes correlações.

A correlação é útil para comparar associações para variáveis com diferentes unidades. Outro previsor potencial para a taxa de assassinatos é o número médio de anos de escolaridade para adultos residentes no estado. A taxa de pobreza e de escolaridade tem unidades diferentes, assim uma mudança de uma unidade na taxa de pobreza não é comparável a uma mudança de uma unidade da escolaridade. Suas inclinações, das equações de previsão separadas, não são comparáveis. As correlações são comparáveis. Suponha que a correlação da taxa de assassinatos com a escolaridade é de $-0{,}30$. Visto que a correlação da taxa de assassinatos com a taxa de pobreza é 0,63 e visto que $0{,}63 > |-0{,}30|$, a taxa de assassinatos está mais fortemente associada à taxa de pobreza do que à de escolaridade.

Muitas propriedades da correlação são similares às da medida ordinal de associação *gama* (Seção 8.5). Ela está entre -1 e $+1$, é simétrica, e valores absolutos maiores indicam associações mais fortes.

Enfatizamos que a correlação descreve relacionamentos *lineares*. Para relacionamentos curvilíneos, a linha de previsão do melhor ajuste pode ser completamente ou aproximadamente horizontal e $r = 0$ quando $b = 0$. Veja a Figura 9.11. Um valor absoluto menor para r não indica, então, que as variáveis não estão associadas, mas que a associação não é linear.

Correlação implica regressão em direção à média

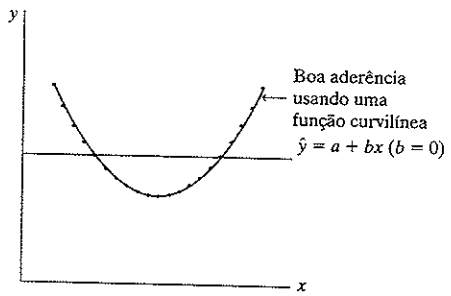
Outra interpretação da correlação se relaciona à sua propriedade da inclinação padronizada. Podemos reescrever a igualdade

$$r = (s_x/s_y)b \text{ como } s_x b = rs_y.$$

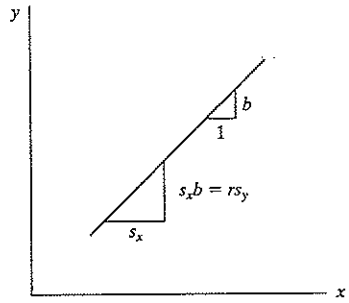
Agora, a inclinação b é a mudança em $\hat{y}$ para uma unidade de aumento em x. Um aumento em x de s_x unidades tem uma mudança prevista de $s_x b$ unidades. (Por exemplo, se $s_x = 10$, um aumento de 10 unidades em x corresponde a uma mudança em $\hat{y}$ de $10b$.) Veja a Figura 9.12. Visto que $s_x b = rs_y$, um aumento de s_x em x corresponde a uma mudança prevista de r desvios padrão nos valores y. Quanto maior for o valor absoluto de r, mais forte será a associação, no sentido de que uma mudança no desvio padrão em x corresponde a uma proporção maior de mudança no desvio padrão de y.

EXEMPLO 9.8 A altura da criança regride em direção da média

O cientista britânico Sir Francis Galton descobriu as ideias básicas da regressão e da correlação em 1880. Após multiplicar a altura de cada mulher por 1,08 para explicar diferenças de gênero, ele notou que a correlação entre x = altura dos pais (a média da altura do pai e da mãe) e y = altura da criança é aproximadamente 0,5.



☑ **Figura 9.11** Diagrama de dispersão para o qual r = 0, embora exista um forte relacionamento curvilíneo.



☑ **Figura 9.12** Um aumento de s_x unidades em x corresponde a uma mudança prevista de rs_y unidades em y.

Da propriedade recém discutida, uma mudança de um desvio padrão na altura dos pais corresponde a uma mudança de meio desvio padrão na altura dos filhos.

Para pais com altura média, a altura das crianças é prevista, também, na média. Se, por outro lado, a altura dos pais está um desvio padrão acima da média, é previsto que a criança esteja apenas meio desvio padrão acima da média. Se a altura dos pais está dois desvios padrão abaixo da média é previsto que a criança esteja apenas um desvio padrão abaixo da média (porque a correlação é 0,5).

Visto que r é menor do que 1, a previsão é de que um valor-y esteja a menos desvios padrão da sua média do que x está. Pais altos tendem a ter filhos altos, mas, na média, não tão altos. Por exemplo, se você considerar todos os pais com 7 pés de altura, talvez seus filhos tenham uma média de 6 pés e 5 polegadas – mais altos do que a média, mas não extremamente altos; se você considerar todos os pais com a altura de 5 pés, talvez seus filhos tenham uma altura média de 5 pés e 5 polegadas – mais baixos do que a média, mas não extremamente baixos. Em cada caso, Galton destacou a *regressão em direção à média*. Essa é a origem do nome da análise de regressão. ■

Para x = taxa de pobreza e y = taxa de assassinatos para os 50 estados, a correlação é de 0,63. Assim um aumento no desvio padrão da taxa de pobreza corresponde a um aumento previsto de 0,63 no desvio padrão da taxa de assassinato. Em contraste, $r = 0{,}37$ entre a taxa de pobreza e a taxa de crimes violentos. Essa associação é fraca. O aumento no desvio padrão da taxa de pobreza corresponde a uma mudança menor na taxa prevista de crimes violentos do que na taxa prevista de assassinatos (em unidades de desvio padrão).

r ao quadrado: redução proporcional no erro de previsão

Uma medida relacionada de associação resume quão bem x pode prever y. Se pudermos prever y muito melhor pela substituição dos valores-x na equação de previsão $\hat{y} = a + bx$ do que sem o conhecimento dos valores-x, as variáveis estão fortemente associadas. Esta medida de associação tem quatro elementos:

- Regra 1 para prever y sem usar x.
- Regra 2 para prever y usando a informação em x.
- Uma medida de resumo para o erro de previsão para cada regra, E_1 para erros pela regra 1 e E_2 para erros pela regra 2.
- A diferença no valor do erro com as duas regras é $E_1 - E_2$. Convertendo

esta redução no erro para uma proporção, temos a definição:

$$\text{Redução proporcional no erro} = \frac{E_1 - E_2}{E_1}.$$

Regra 1 (Prevendo y sem usar x): O melhor previsor é $\bar{y}$, a média amostral.

Regra 2 (Prevendo y usando x): Quando o relacionamento entre x e y é linear, a equação de previsão $\hat{y} = a + bx$ fornece o melhor previsor de y. Para cada sujeito, substituir o valor-x nessa equação fornece o valor previsto de y.

Erros Previstos: O erro previsto para cada sujeito é a diferença entre os valores observados e previstos de y. O erro previsto usando a Regra 1 é $y - \bar{y}$, e o erro previsto usando a regra 2 é $y - \hat{y}$, o resíduo. Para cada previsor, alguns erros previstos são positivos, alguns são negativos e a soma dos erros é igual a 0. Resumimos os erros previstos pelas somas dos quadrados dos seus valores:

$$E = \sum(\text{valor observado de } y - \text{valor previsto de } y)^2.$$

Para a Regra 1, os valores previstos são iguais a $\bar{y}$. O erro de previsão total é igual a:

$$E_1 = \sum(y - \bar{y})^2.$$

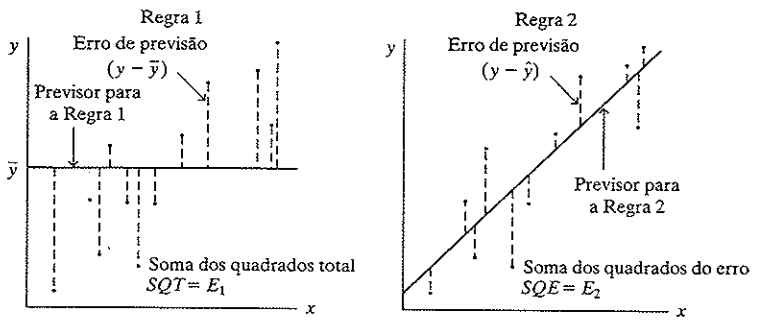
Esta é *a soma dos quadrados total* dos valores-y em relação a sua média. Representamos esse valor por SQT. Para a regra 2, os valores previstos são os valores-$\hat{y}$ da equação de previsão. O erro previsto total é igual a:

$$E_2 = \sum(y - \hat{y})^2.$$

Representamos este valor por SQE e denominamos de **soma dos quadrados do erro** ou a **soma dos quadrados dos resíduos**.

Quando x e y têm uma associação linear forte, a equação de previsão fornece previsões ($\hat{y}$) que são muito melhores do que $\bar{y}$, no sentido de que a soma dos quadrados dos erros previstos é substancialmente menor. A Figura 9.3 mostra representações gráficas dos dois previsores e seus erros de previsão. Para a Regra 1, o mesmo previsor ($\bar{y}$) se ajusta ao valor de y independentemente do valor de x. Para a Regra 2, a previsão muda à medida que x muda e os erros de previsão tendem a ser menores.

Definição da Medida: A redução proporcional no erro pelo uso da equação de previsão linear em vez de $\bar{y}$ para prever y é:



☑ **Figura 9.13** Representação gráfica da Regra 1 e soma dos quadrados total $E_1 = SQT = \Sigma(y - \bar{y})^2$, Regra 2 e soma dos quadrados dos resíduos $E_2 = SQE = \Sigma(y - \hat{y})^2$.

$$r^2 = \frac{E_1 - E_2}{E_1} = \frac{\text{SQT} - \text{SQE}}{\text{SQT}} = \frac{\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}.$$

Ela é denominada de ***r* ao quadrado** ou, algumas vezes, de **coeficiente de determinação**.

A notação r^2 é usada para esta medida porque, na verdade, a redução proporcional no erro é igual ao quadrado da correlação r. Não precisamos usar as somas dos quadrados da sua definição para encontrar r^2, porque podemos elevar a correlação ao quadrado. Sua definição é útil para interpretar o r^2, mas não é necessária para a sua determinação.

EXEMPLO 9.9 r^2 para a taxa de assassinatos e a taxa de pobreza

A correlação entre a taxa de pobreza e a taxa de assassinatos para os 50 estados é $r = 0{,}629$. Portanto, $r^2 = (0{,}629)^2 = 0{,}395$. Para prever a taxa de assassinatos, a equação de previsão linear $\hat{y} = -0{,}86 + 0{,}58x$ tem 39,5% menos erro do que $\bar{y}$.

Um *software* para determinar a regressão rotineiramente fornece tabelas que contêm a soma dos quadrados que compõe r^2. Por exemplo, parte da Tabela 9.3 é:

	Soma dos quadrados
Regressão	307,342
Resíduos	470,406
Total	777,749

A soma dos quadrados do erro usando a equação da previsão é SQE = $\Sigma(y - \hat{y})^2$ = 470,4, e a soma dos quadrados total é STQ = $\Sigma(y - \bar{y})^2$ = 777,7. Portanto,

$$r^2 = \frac{\text{SQT} - \text{SQE}}{\text{SQT}} = \frac{777{,}7 - 470{,}4}{777{,}7} = \frac{307{,}3}{777{,}7} = 0{,}395.$$

Na prática, não será necessário executar esse cálculo, visto que o *software* informa r ou r^2 ou ambos. ■

Propriedades do *r* ao quadrado

As propriedades do r^2 seguem diretamente daquelas da correlação r ou da sua definição em termos das somas dos quadrados.

- Visto que $-1 \leq r \leq 1$, r^2 está entre 0 e 1.
- O valor mínimo possível para SQE é 0, nesse caso r^2 = SQT/SQT = 1. Para SQE = 0, todos os pontos da amostra devem estar exatamente sobre a linha de previsão. Neste caso, não existe erro de previsão usando x para prever y. Esta condição corresponde a $r = \pm 1$.
- Quando a inclinação da linha dos mínimos quadrados é $b = 0$, o intercepto-y a é igual a $\bar{y}$ (porque $a = \bar{y} - b\bar{x}$, que é igual a $\bar{y}$ quando $b = 0$). Então $\hat{y} = \bar{y}$ para todo x. As duas regras de previsão são, então, idênticas, assim $SQE = SQT$ e $r^2 = 0$.
- Como a correlação, r^2 mensura a força da associação *linear*. Quanto mais próximo r^2 está de 1, mais forte a associação linear, no sentido do quanto mais efetiva é a linha dos mínimos quadrados $\hat{y} = a + bx$ comparada a $\bar{y}$ na previsão de y.
- O r^2 não depende das unidades de mensuração e ele assume o mesmo valor tanto quando x prevê y como quando y prevê x.

As somas dos quadrados descrevem a variabilidade condicional e marginal

Para resumir, a correlação r está entre -1 e $+1$. Ela indica a direção da associação, positiva ou negativa, por meio do seu sinal. É uma inclinação padronizada e será igual

à inclinação (b) quando x e y estiverem igualmente dispersos. Uma mudança de um desvio padrão em x corresponde a uma mudança prevista de r desvios padrão em y. O quadrado da correlação é interpretado como a redução proporcional do erro relacionado à previsão de y usando $\hat{y} = a + bx$ em vez de $\bar{y}$.

A soma dos quadrados total, $SQT = \sum(y - \bar{y})^2$, resume a variabilidade das observações em y, visto que esta quantidade dividida por $n - 1$ é a variância amostral (s_y^2) dos valores-y. Da mesma forma o $SQE = \sum(y - \hat{y})^2$ resume a variabilidade em torno da equação de previsão, que se refere à variabilidade para as distribuições condicionais. Quando $r^2 = 0{,}39$, a variabilidade de y usando x para fazer previsões (via a equação de previsão) é 39% menor do que a variabilidade geral dos valores y. Portanto, o resultado do r^2 é geralmente expresso como "a taxa de pobreza explica 39% da variabilidade da taxa de assassinatos" ou "39% da variância da taxa de assassinatos é explicada pelo seu relacionamento linear com a taxa da pobreza". Falando *grosso modo*, a variância da distribuição condicional da taxa de assassinatos para uma dada taxa de pobreza é 39% menor do que a variância da distribuição marginal da taxa de assassinatos.

Esta interpretação tem a fragilidade, entretanto, de que a variabilidade é resumida pela *variância*. Muitos estatísticos acham o r^2 menos útil do que o r, porque (sendo baseado nas somas dos quadrados) usa o quadrado da escala original de mensuração. É mais fácil interpretar a escala original do que a escala ao quadrado. Esta também é uma vantagem do desvio padrão sobre a variância.

Quando duas variáveis estão fortemente associadas, a variação nas distribuições condicionais é consideravelmente menor do que a variação na distribuição marginal. A Figura 9.9 ilustra isso.

9.5 INFERÊNCIAS PARA A INCLINAÇÃO E A CORRELAÇÃO

As Seções 9.1 a 9.3 mostraram como o modelo de regressão linear pode representar a *forma* do relacionamento entre variáveis quantitativas. A Seção 9.4 usou a correlação e seu quadrado para descrever a *força* da associação. Estas partes de uma análise de regressão são descritivas. Agora apresentaremos métodos inferenciais para o modelo de regressão.

Um teste para verificar se duas variáveis quantitativas são estatisticamente independentes tem o mesmo objetivo do teste qui-quadrado para variáveis categóricas. Um intervalo de confiança para a inclinação da equação de regressão ou da correlação nos informa sobre o tamanho do efeito. Essas inferências nos permitem julgar se as variáveis estão associadas e estimar a direção e a força da associação.

Suposições para a inferência estatística

As inferências estatísticas para a regressão fazem as seguintes suposições:

- O estudo usou a aleatorização, como uma amostra aleatória simples.
- A média de y está relacionada a x pela equação linear $E(y) = \alpha + \beta x$.
- O desvio padrão condicional σ é o mesmo para cada valor-x.
- A distribuição condicional de y para cada valor de x é normal.

A segunda suposição afirma que a função da regressão linear é válida. A suposição sobre um σ comum significa que as estimativas dos mínimos quadrados são as melhores estimativas dos coeficientes da regressão.[2] A suposição sobre a normalidade assegura que a estatística-teste para um teste de independência tem uma distribuição amostral t. Na prática, nenhuma destas suposições é exatamente satisfeita. Na seção final do capítulo veremos que as suposições mais importantes são as duas primeiras.

Teste de independência

Sob as suposições acima, suponha que a média populacional de y é idêntica em cada valor-x. Em outras palavras, a distribuição condicional normal de y é a mesma em cada valor-x. Então, as duas variáveis quantitativas são estatisticamente independentes. Para a função de regressão linear $E(y) = \alpha + \beta x$, isto significa que a inclinação $\beta = 0$ (veja Figura 9.14). A hipótese nula de que as variáveis são estatisticamente independentes é H_0: $\beta = 0$.

Podemos testar a independência contra H_a: $\beta \neq 0$, ou uma alternativa unilateral, H_a: $\beta > 0$ ou H_a: $\beta < 0$, para prever a direção da associação. A estatística-teste é igual a:

$$t = \frac{b}{ep},$$

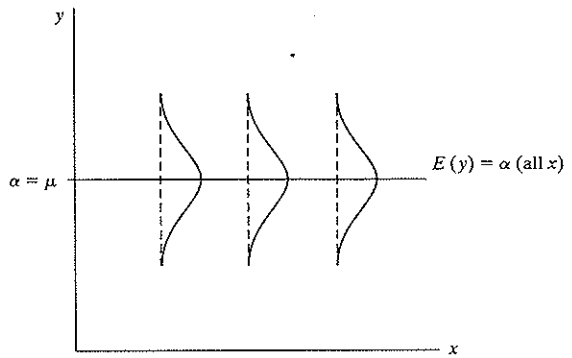
onde ep é o erro padrão da inclinação amostral b. A forma da estatística-teste é a usual para um teste t ou z. Tomamos a estimativa b do parâmetro β, subtraímos o valor da hipótese nula ($\beta = 0$) e dividimos pelo erro padrão da estimativa b. Sob as suposições feitas, essa estatística-teste tem uma distribuição amostral t com $gl = n - 2$. Os graus de liberdade são os mesmos dos gl da estimativa do desvio padrão condicional s.

A fórmula para o erro padrão de b é:

$$ep = \frac{s}{\sqrt{\sum(x - \bar{x})^2}}, \text{ onde } s = \sqrt{\frac{\text{SQE}}{n - 2}}.$$

Isto depende da estimativa por pontos, s, do desvio padrão da distribuição condicional de y. Quanto menor for s, mais precisamente b estima β. Um s pequeno ocorre quando os pontos dos dados mostram pouca variabilidade em relação à equação de previsão. Da mesma forma, o erro padrão de b está inversamente relacionado a $\sum(x - \bar{x})^2$, a soma dos quadrados dos valores observados de x em relação à sua média. Esta soma aumenta, e por conseguinte b estima β mais precisamente, à medida que o tamanho da amostra n aumenta. (O ep também diminui quando os valores-x estão mais dispersos, mas o pesquisador geralmente não tem controle sobre isto, exceto em experimentos projetados.)

O valor-p para H_a: $\beta \neq 0$ é a probabilidade das duas caudas da distribuição t. O *software* fornece o valor-p. Para gl grandes, lembre que a distribuição t é similar



☑ **Figura 9.14** x e y são estatisticamente independentes quando a inclinação $\beta = 0$ no modelo de regressão $E(y) = \alpha + \beta x$.

à normal padrão, assim o valor-p pode ser aproximado usando a tabela normal de probabilidades.

EXEMPLO 9.10 Regressão para os preços de vendas de casas

O que afeta o preço de venda de uma casa? A Tabela 9.4 mostra observações das vendas de casas em Gainesville, Flórida, no outono de 2006. Esta tabela mostra dados para 8 casas. Todo o arquivo incluindo as vendas de 100 casas é o *House selling price* (preço de venda de casas) disponível no *site* do livro. As variáveis listadas são o preço de venda (em dólares), o tamanho da casa (em pés quadrados), as taxas anuais (em dólares), o número de quartos, o número de banheiros e se a casa foi recém construída. Por agora, usaremos somente os dados de y = preço de venda e x = tamanho da casa.

Visto que estas 100 observações vêm de uma só cidade, não podemos usá-las para fazer inferências sobre o relacionamento entre x e y em geral. Vamos tratá-las como uma amostra aleatória de uma população conceitual de vendas de casas nesse mercado, em particular, para analisar como estas variáveis parecem estar relacionadas.

A Figura 9.15 mostra um diagrama de dispersão que exibe uma forte tendência positiva. O modelo $E(y) = \alpha + \beta x$ parece apropriado. Alguns dos pontos nos níveis altos do tamanho da casa são valores atípicos da regressão e um ponto está bem abaixo da tendência geral. Discutiremos esta anormalidade na Seção 14.5, que introduz um modelo alternativo que não assume variabilidade constante em torno da linha de regressão.

A Tabela 9.5 mostra parte de uma saída do SPSS para a análise de regressão. A equação de previsão é $\hat{y} = -50926 + 126{,}6x$. O preço de venda previsto aumenta em $b = 126{,}6$ dólares para um aumento de um pé quadrado no tamanho da casa. A Figura 9.15 apresenta, também a equação de previsão sobre o diagrama de dispersão. No SPSS, "Beta" representa a estimativa do coeficiente de regressão padronizado. Para o modelo de regressão desse capítulo, isso é a correlação e não deve ser confundida com a inclinação na população, β, que é desconhecida.

A Tabela 9.5 informa que o erro padrão da estimativa da inclinação é $ep = 8{,}47$. Isto está listado sob *Std. Error* (Erro Padrão) para o previsor do tamanho. Este valor estima a variabilidade nos valores da inclinação amostral que resultaria da seleção repetida de amostras aleatórias de 100 vendas de casas em Gainesville e do cálculo de equações de previsão para cada amostra.

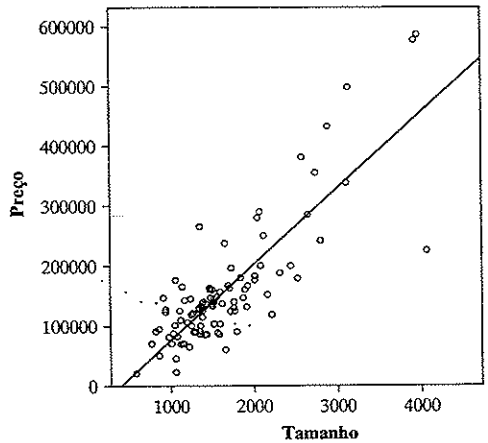
Para testar independência, H_0: $\beta = 0$, a estatística-teste é:

$$t = \frac{b}{ep} = \frac{126{,}6}{8{,}47} = 14{,}95,$$

☑ **Tabela 9.4** Preços de venda e fatores relacionados para uma amostra de vendas de casas em Gainesville, Flórida

Casa	Preço de venda	Tamanho	Taxas	Quartos	Banheiros	Nova
1	279900	2048	3104	4	2	não
2	146500	912	1173	2	1	não
3	237700	1654	3076	4	2	não
4	200000	2068	1608	3	2	não
5	159900	1477	1454	3	3	não
6	499900	3153	2997	3	2	sim
7	265500	1355	4054	3	2	não
8	289900	2075	3002	3	2	sim

Nota: Para o arquivo completo das 100 casas consulte o *site* do livro em www.grupoa.com.br.



☑ **Figura 9.15** Diagrama de dispersão e equação de previsão para y = preço de venda (em dólares) e x = tamanho da casa (em pés quadrados).

mostrada na última linha da Tabela 9.5. Visto que $n = 100$, seus graus de liberdade são $gl = n - 2 = 98$. Esta é uma estatística-teste extremamente grande. O valor-p, listado na Tabela 9.5 sob o título "Sig." é 0,000 com três casas decimais. Isso se refere a uma alternativa bilateral H_a: $\beta \neq 0$. Ela é a probabilidade bilateral de uma estatística t pelo menos tão grande em valor absoluto quanto o valor absoluto do t ob-

☑ **Tabela 9.5** Informação da saída do SPSS para a análise de regressão de y = preço de venda e x = tamanho da casa

	N	Média	Desvio padrão
preço	100	155331,00	101262,21
tamanho	100	1629,28	666,94

	Soma dos quadrados	gl	Média dos quadrados
Regressão	7,057E+11	1	7,057E+11
Resíduos	3,094E+11	98	3157352537
Total	10,15E+11	99	

R Quadrado	Erro padrão da estimativa
0,695	56190,324

	Coeficientes não padronizados		Coeficientes padronizados		
	B	Erro padrão	Beta	t	Sig.
(Constante)	-50926,3	14896,373		-3,42	0,001
Tamanho	126,594	8,468	0,834	14,95	0,000

servado, $|t| = 14{,}95$, presumindo que H_0 é verdadeiro.

A Tabela 9.6 mostra parte de uma saída do SAS para a mesma análise. O valor-p bilateral, listado sob o título "Pr > |t|", é < 0,0001 para quatro casas decimais. (Ele é, na verdade, 0,0000000... com um número enorme de casas decimais, mas o SAS informa-o desta forma em vez de 0,0000, assim não pense que o valor-p é *exatamente* 0.)

Tanto a saída do SAS quanto a do SPSS contém um erro padrão e um teste t para o intercepto-y. Não usamos essa informação, visto que raramente existe uma razão para testar a hipótese de que um intercepto-y é igual a 0. Para esse exemplo, o intercepto-y não tem nenhuma interpretação, visto que casas do tamanho $x = 0$ não existem.

Em resumo, existe uma evidência extremamente forte de que o tamanho da casa tem um efeito positivo no preço de venda. Em média, o preço da casa aumenta à medida que o tamanho aumenta. Isto não é surpresa. Na verdade, ficaríamos chocados se estas variáveis fossem independentes. Para estes dados, estimar o tamanho do efeito é mais relevante do que testar se ele existe. ■

Intervalo de confiança para a inclinação

Um valor-p pequeno para H_0: $\beta = 0$ sugere que a linha de regressão tem uma inclinação que não é zero. Deveríamos estar mais preocupados com o tamanho da inclinação β do que saber meramente que ela não é 0. Um intervalo de confiança para β tem a fórmula:

$$b \pm t(ep).$$

O escore-t é o valor da Tabela B, com $gl = n - 2$, para o nível de confiança desejado. A forma do intervalo é similar ao intervalo de confiança para uma média (Seção 5.3, na página 140), ou seja, tome a estimativa do parâmetro e adicione e subtraia um t múltiplicado pelo erro padrão. O ep é o mesmo do ep do teste para β.

EXEMPLO 9.11 Estimando a inclinação para os preços de venda de casas

Para os dados em x = tamanho da casa e y = preço de venda, $b = 126{,}6$ e $ep = 8{,}47$. O parâmetro β se refere à mudança no preço médio de venda (em dólares) para cada aumento de um pé quadrado no tamanho da casa. Para um intervalo de 95% de confiança, usamos o valor $t_{0{,}025}$ para um $gl = n - 2 = 98$, que é $t_{0{,}025} = 1{,}984$. (A Tabela B, na página 651, mostra $t_{0{,}025} = 1{,}984$ para $gl = 100$.) O intervalo é

$$b \pm t_{0{,}025}(ep) = 126{,}6 \pm 1{,}984(8{,}47) = 126{,}6 \pm 16{,}8 \text{ ou } (110, 143).$$

Podemos estar 95% confiantes de que β está entre 110 e 143. O preço médio de venda aumenta entre \$110 e \$143 para um aumento de um pé quadrado no tamanho da casa. ■

Na prática, fazemos inferências sobre a mudança em $E(y)$ para um aumento em x que é uma porção relevante do intervalo real dos valores-x. Se o aumento de uma unidade em x é muito pequeno ou muito grande em termos práticos, o intervalo de confiança para β pode ser ajustado para se referir a uma mudança diferente em x. Para obter um intervalo de confiança para um múltiplo da inclinação (como 100β, a mudança na média de y para um aumento de 100 unidades em x), multiplique os pontos extremos do intervalo de confiança para β pela mesma constante.

Para a Tabela 9.4, x = tamanho da casa tem $\bar{x} = 1629$ e $s_x = 669$. Uma mudança de um pé quadrado no tamanho é pequena. Vamos estimar o efeito de um aumento em tamanho de 100 pés quadrados. A mudança na média de y é 100β. O intervalo de 95% de confiança para β é (110, 143), assim o intervalo de 95% de confiança para 100β tem pontos extremos 100(110) = 11100 e 100(143) = 14300. Assim, inferimos que o preço médio de venda aumenta por, pelo menos, \$11100 e no máximo \$14300, para um aumento no tamanho da casa de 100 pés quadrados. Por exemplo, assumindo que o modelo de regressão linear é válido, concluímos que o preço médio está entre \$11100 e \$14300 acima para casas com 1700 pés quadrados do que para casa com 1600 pés quadrados.

☑ **Tabela 9.6** Parte da saída do SAS para a análise de regressão do preço de venda pelo tamanho da casa

Fonte	gl	Soma dos quadrados	Média dos quadrados	
Modelo	1	7,05729E11	7,05729E11	
Erro	98	3,094205E11	3157352537	
Total corrigido	99	1,01515E12		
Raiz do EQM (Erro Quadrático Médio) 56190,3				
Variável	gl	Estimativa do parâmetro	Erro padrão	Valor t	Pr > |t|
Intercepto	1	-50926	14896	-3,42	0,0009
Tamanho	1	126,59411	8,46752	14,95	< 0,0001

Lendo a saída fornecida pelo computador

Vamos examinar a saída nas Tabelas 9.5 e 9.6. Elas contêm algumas informações que ainda não discutimos. Por exemplo, na tabela da soma dos quadrado, a soma dos quadrados do erro (SQE) é 3,094 vezes 10^{11}. Este é um número enorme porque os valores-y são muito grandes e seus desvios estão elevados ao quadrado. O desvio padrão condicional estimado de y é

$$s = \sqrt{\text{SQE}/(n-2)} = 56{,}190.$$

O SAS rotula isto como *Root MSE* para a raiz quadrada do erro quadrático médio. O SPSS equivocadamente rotula-o com *Std. Error of the Estimate* (Erro Padrão da Estimativa). Esse é um nome pouco apropriado porque s se refere a um desvio padrão condicional de preços de venda (para casas de um determinado tamanho) e não ao erro padrão de uma estatística.

A tabela da soma dos quadrados também informa a soma dos quadrados total, $\text{SQT} = \sum(y - \bar{y})^2 = 10{,}15 \times 10^{11}$. Deste valor e SQE:

$$r^2 = \frac{\text{SQT} - \text{SQE}}{\text{SQT}} = 0{,}695.$$

Esse valor é a redução proporcional no erro ao se usar o tamanho da casa para prever o preço de venda. Visto que a inclinação da equação de previsão é positiva, a correlação é uma raiz quadrada positiva deste valor ou 0,834. Uma associação positiva forte existe entre essas variáveis.

A soma dos quadrados total (SQT) pode ser dividida em duas partes: a soma dos quadrados do erro, $\text{SQE} = 3{,}094 \times 10^{11}$ e a diferença entre SQT e a SQE, $\text{SQT} - \text{SQE} = 7{,}057 \times 10^{11}$. Esta diferença é o numerador da medida r^2. O SPSS chama isso de **soma dos quadrados da regressão** e o SAS de **soma dos quadrados do modelo**. Ela representa o montante da variação total SQT em y que é explicado por x usando a

linha dos mínimos quadrados. A razão dessa soma de quadrados para a SQT é igual ao r^2.

A tabela da soma dos quadrados tem uma lista associada de valores de graus de liberdade. Os graus de liberdade para a soma dos quadrados total, SQT = $\Sigma(y - \bar{y})^2$ é $n - 1 = 99$, visto que SQT se refere à variabilidade da distribuição *marginal* de y, que tem uma variância estimada $s_y^2 = \text{SQT}/(n - 2)$. Os graus de liberdade para SQE são iguais a $n - 2 = 98$, visto que SQE se refere à variabilidade na distribuição *condicional* de y, que tem uma variância estimada de $s^2 = \text{SQE}/(n - 2)$ para um modelo com dois parâmetros. A soma dos quadrados da regresssão (ou modelo) tem *gl* igual ao número de variáveis explicativa do modelo de regressão que, nesse caso, é 1. A soma dos graus de liberdade da soma dos quadrados da regressão com a soma dos quadrados dos resíduos ou erros (SQE) é igual a $n - 1$ que é igual a soma dos quadrados total que, nesse caso, é 1 + 98 = 99.

Inferência para a correlação*

A correlação r é igual a zero nas mesmas situações em que a inclinação b é igual a zero. Considere ρ (letra grega rho) como a representação do valor de correlação na população. Então, $\rho = 0$ precisamente quando $\beta = 0$. Na verdade, um teste de H_0: $\rho = 0$ usando o valor amostral r é equivalente ao teste t de H_0: $\beta = 0$ usando o valor amostral b.

A estatística-teste para testar H_0: $\rho = 0$ é:

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$

Isso fornece o mesmo valor que a estatística-teste t = b/ep. Utilize qualquer uma dessas estatísticas para testar H_0: independência, uma vez que cada uma delas tem a mesma distribuição amostral t com gl = n − 2 e produzem o mesmo valor-p. Por exemplo, a correlação de r = 0,834 para os dados dos preços de venda de casas fornece:

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{0{,}834}{\sqrt{(1 - 0{,}695)/98}} = 14{,}95.$$

Isto é a mesma estatística-teste t que o Exemplo 9.10 (página 310) tinha para testar H_0: $\beta = 0$.

Para um conjunto de variáveis, um *software* pode determinar a correlação para cada par em uma **matriz de correlações**. Esta matriz é uma tabela quadrada listando as variáveis como linhas e, novamente, como colunas. A Tabela 9.7 mostra a forma como um *software* informa a matriz de correlações para as variáveis preço de venda, tamanho da casa, taxas e número de quartos. A correlação entre cada par de variáveis aparece duas vezes. Por exemplo, a correlação de 0,834 entre preço de venda e tamanho da casa ocorre tanto na linha para preço quanto na coluna para tamanho e na linha do tamanho com a coluna do preço. O valor-p para testar H_0: $\rho = 0$ contra H_a: $\rho \neq 0$ está listada sob o valor da correlação.

As correlações na diagonal indo do canto superior esquerdo para o canto inferior direito de uma matriz de correlações, todas são iguais a 1,000. Isto meramente indica que a correlação entre uma variável e ela mesma é 1,0. Por exemplo, se conhecemos o valor de y, então podemos prever o valor de y perfeitamente.

Construir um intervalo de confiança para a correlação ρ é mais complicado do que para a inclinação β. A razão é que a distribuição amostral de r não é simétrica, exceto quando $\rho = 0$. A falta de simetria é causada pelo intervalo restrito $[-1, 1]$ para os valores r. Se ρ está próximo de 1,0, por exemplo, a amostra r não pode estar muito acima de ρ, mas ela pode estar bem abaixo de ρ. A distribuição amostral de r é, então, assimétrica à esquerda. O Exercício 9.64 mostra como construir intervalos de confiança para correlações.

Dados que faltam (*missing data*)

Em uma análise de correlação, alguns sujeitos podem não ter observações para uma ou mais variáveis. Por exemplo, a Tabela 9.13 dos exercícios lista 10 variáveis para 40 países. Observações em algumas das variáveis, tal como a taxa de analfabetismo, estão faltando para vários países.

Para análises estatísticas, alguns *softwares* eliminam todos os sujeitos para os quais os dados estão faltando em pelo menos uma variável. Isto é chamado de **eliminação em lista** (*listwise deletion*). Outros *softwares* eliminam somente o sujeito para a análise no qual aquela observação é necessária. Por exemplo, esta abordagem usa um sujeito para encontrar a correlação para duas variáveis se aquele sujeito fornecer observações para ambas as variáveis, sem levar em consideração se o sujeito fornece observações para as demais variáveis. Esta abordagem é chamada de **eliminação em pares** (*pairwise deletion*). Com esta abordagem, o tamanho da amostra pode ser maior em cada análise.

Nos dias de hoje, existem estratégias melhores e mais sofisticadas do que estas. Elas ainda não estão disponíveis na maioria dos *softwares* e elas estão além do alcance deste livro. Para detalhes, veja Allison (2002).

9.6 SUPOSIÇÕES DO MODELO E VIOLAÇÕES

Terminamos este capítulo reconsiderando as suposições que fundamentam a análise de regressão linear. Discutiremos os efeitos da violação destas suposições e os efeitos das observações *influentes*. Finalmente, mostraremos uma forma alternativa para expressar o modelo.

Quais suposições são importantes?

O modelo de regressão linear assume que o relacionamento entre x e a média de y segue uma linha reta. A forma real é desconhecida. Ela certamente não é *exatamente* linear. No entanto, uma função linear geralmente fornece uma aproximação decente para o relacionamento real. A Figura 9.16 ilustra uma linha reta estando próxima a um relacionando curvilíneo real.

As inferências discutidas na seção anterior são apropriadas para detectar associações lineares positivas ou negativas. Suponha que, ao contrário, o relacionamento

☑ **Tabela 9.7** Matriz de correlação para os dados dos preços de venda das casas. O valor sob a correlação é um valor-p bilateral para testar H_0: $\rho = 0$

	Correlações / valor-p para H_0: $\rho = 0$			
	preço	tamanho	taxas	quartos
preço	1,00000	0,83378	0,84198	0,39396
		< 0,0001	< 0,0001	< 0,0001
tamanho	0,83378	1,00000	0,81880	0,54478
	< 0,0001		< 0,0001	< 0,0001
taxas	0,84198	0,81880	1,00000	0,47393
	< 0,0001	< 0,0001		< 0,0001
quartos	0,39396	0,54478	0,47393	1,00000
	< 0,0001	< 0,0001	< 0,0001	

verdadeiro tivesse a forma de U, como na Figura 9.5. Então, as variáveis seriam estatisticamente dependentes, visto que a média de y iria mudar à medida que o valor de x mudar. O teste t de H_0: $\beta = 0$ pode não detectá-lo, no entanto, porque a inclinação b da linha dos mínimos quadrados estaria próxima a 0. Em outras palavras, um valor-p pequeno não iria provavelmente ocorrer mesmo que uma associação exista. Em resumo, $\beta = 0$ não precisa corresponder à independência se a suposição de um modelo de regressão linear é violada. Por esta razão, você deve sempre construir um diagrama de dispersão para verificar essa suposição fundamental.

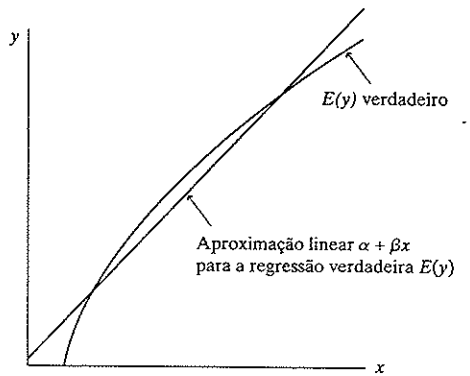
Tanto a linha dos mínimos quadrados quanto r e r^2 são estatísticas descritivas válidas não importando a forma da distribuição condicional dos valores-y para cada valor-x. Entretanto, as inferências estatísticas na Seção 9.5 também assumem que as distribuições condicionais de y são (1) normais, com (2) desvios padrão σ idênticos para cada valor-x. Estas suposições não são, também, *exatamente* satisfeitas na prática. Para amostras grandes, a suposição de normalidade é relativamente pouco importante porque um Teorema Central do Limite estendido implica que as inclinações da amostra e as correlações têm distribuições amostrais aproximadamente normais. Se a suposição sobre o σ idêntico é violada, outras estimativas podem ser mais eficientes do que os mínimos quadrados (isto é, tendo valores *ep* menores), mas os métodos de inferência usuais são ainda aproximadamente válidos.

A aleatorização da amostra e a suposição de linearidade são muito importantes. Se o relacionamento verdadeiro se desvia muito de uma linha reta, por exemplo, não faz sentido usar uma inclinação ou uma correlação para descrevê-lo. O Capítulo 14 discute formas de verificar as suposições e fazer modificações na análise, se for necessário.

A extrapolação é perigosa

É perigoso aplicar uma equação de previsão aos valores de x fora do intervalo dos valores observados. O relacionamento pode não ser linear fora desse intervalo. Podemos obter previsões ineficientes ou até mesmo absurdas extrapolando além do intervalo observado dos valores de x.

Para ilustrar, a equação de previsão $\hat{y} = -0{,}86 + 0{,}58x$ da Seção 9.2 relacionando x = taxa de pobreza a y = taxa de assassinatos era baseada nas taxas de pobreza observadas entre 8,0 e 26,4. Não é válido extrapolar muito abaixo ou acima desse intervalo. A taxa de assassinatos prevista para uma taxa de pobreza de $x = 0\%$ é $\hat{y} = -0{,}86$. Este é um valor impossível para a taxa de assassinatos, que não pode ser negativa.



☑ **Figura 9.16** Uma equação de regressão linear como uma aproximação para um relacionamento não linear.

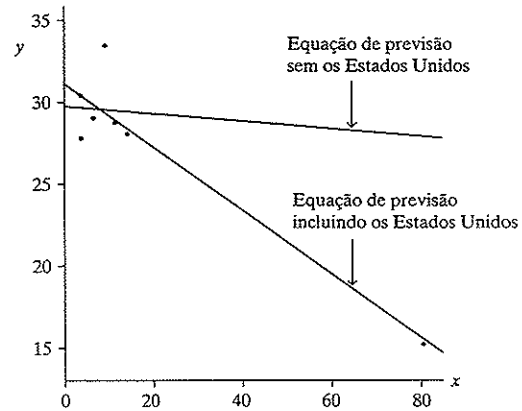
Informações influentes

O método dos mínimos quadrados tem uma longa história e é a forma padrão de ajustar equações de previsões a dados. Uma desvantagem dos mínimos quadrados, entretanto, é que observações individuais podem indevidamente influenciar os resultados. Uma única observação pode ter um grande efeito se ela é um *valor atípico da regressão* – tendo um valor-x relativamente grande ou relativamente pequeno e ficando distante da tendência que o restante dos dados segue.

A Figura 9.17 ilustra isso. Ela faz uma representação gráfica das observações para várias nações africanas e asiáticas utilizando y = taxa bruta de natalidade (número de nascimentos por cada 1000 da população) e x = número de televisores por 100 habitantes. Adicionamos aos dados a observação Estados Unidos, que é um valor atípico com uma taxa de natalidade muito mais baixa do que os demais países e com um valor muito mais alto do número de televisores. A Figura 9.17 mostra as equações de previsão com e sem os Estados Unidos. A equação de previsão muda de $\hat{y} = 29{,}8 - 0{,}024x$ para $\hat{y} = 31{,}2 - 0{,}195x$. Acrescentar apenas um ponto ao conjunto de dados acarretou uma dramática inclinação, para baixo, na linha de previsão.

A Seção 9.2 mostrou uma versão não tão extrema disto. A inclinação da equação de previsão mais do que dobrou quando incluímos a observação para o D.C. no conjunto de dados sobre as taxas de assassinatos por todo o estado.

Quando um diagrama de dispersão mostra um valor atípico severo da regressão, você deve investigar as razões dessa ocorrência. A observação pode ter sido registrada incorretamente. Se a observação for correta, talvez aquela observação seja fundamentalmente diferente das outras de alguma forma, como a observação dos Estados Unidos na Figura 9.17. Ela pode sugerir um previsor adicional para



☑ **Figura 9.17** Equações de previsão para y = taxa de natalidade e x = número de televisores, com e sem a observação dos Estados Unidos.

o modelo usando métodos do Capítulo 11. Geralmente vale a pena ajustar novamente o modelo sem um ou dois valores atípicos extremos da regressão para ver se essas observações têm um efeito grande no ajuste, como fizemos seguindo o Exemplo 9.4 (página 294) com a observação do D.C. para as taxas de assassinatos.

As observações que têm uma grande influência nas estimativas dos parâmetros do modelo podem também ter um grande impacto sobre a correlação. Por exemplo, para os dados na Figura 9.17, a correlação é −0,935 quando os Estados Unidos estão incluídos e −0,051 quando eles são retirados do conjunto de dados. Um ponto pode fazer grande diferença, especialmente quando o tamanho da amostra é pequeno.

Fatores que influenciam a correlação

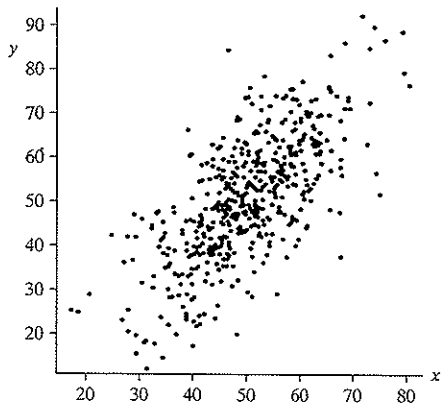
Além de ser influenciada por valores atípicos, a correlação depende do intervalo dos valores-x amostrados. Quando uma amostra tem um intervalo bem menor de variação em x do que a população, a correlação da amostra tende a subestimar drasticamente (em valor absoluto) a correlação da população.

A Figura 9.18 mostra um diagrama de dispersão de 500 pontos que é regular e tem uma correlação de $r = 0{,}71$. Suponha que, ao contrário, tivéssemos amostrado apenas a metade central dos pontos, aproximadamente entre os valores x de 43 a 57. Então, a correlação seria igual a $r = 0{,}33$, consideravelmente mais baixa. Para a relação entre o preço da casa e o tamanho da mesma, apresentado na Figura 9.15, $r = 0{,}834$. Se amostrarmos somente as vendas nas quais o tamanho da casa está entre 1300 e 2000 pés quadrados, que inclui 44 das 100 observações, r diminuiria para 0,254.

A correlação é mais apropriada como medida de associação se os valores (x, y) forem uma amostra aleatória da população. Nessa situação, existe uma amostra representativa tanto da variação de x como da variação de y.

EXEMPLO 9.12 O SAT prevê o GPA da faculdade?

Considere a associação entre x = escore do SAT para admissão na universidade e y = GPA (rendimento) da faculdade ao final do segundo ano. A força da correlação depende da variabilidade nos escores



☑ **Figura 9.18** A correlação é afetada pelo intervalo de valores-x. A correlação diminui de 0,71 para 0,33 quando apenas os pontos x entre 43 e 57 forem utilizados.

do SAT na amostra. Se estudarmos a associação somente para os estudantes da Universidade de Harvard, a correlação será, provavelmente, fraca, porque os escores do SAT da amostra estarão bem concentrados na parte superior da escala. Ao contrário, se amostrarmos aleatoriamente a população de *todos* os estudantes do ensino médio que fazem o SAT e colocarmos estes estudantes no ambiente de Harvard, estudantes com péssimos escores do SAT tenderiam a ter baixos GPAs em Harvard. Observaríamos, então, uma correlação muito mais forte. ■

Outros aspectos da regressão, como ajustar uma equação de previsão aos dados e fazer inferências sobre a inclinação, permanecem válidos quando amostramos aleatoriamente y dentro de um intervalo restrito de valores-x. Simplesmente limitamos nossas previsões para aquele intervalo. A inclinação da equação de previsão não é afetada por uma restrição no intervalo de x. Para a Figura 9.18, por exemplo, a inclinação da amostra é igual a 0,97 para todos os dados e 0,96 para o conjunto restrito aos valores centrais. A correlação faz mais sentido, entretanto, quando ambos x e y são aleatórios, em vez de somente y.

Modelo de regressão com termos erro*

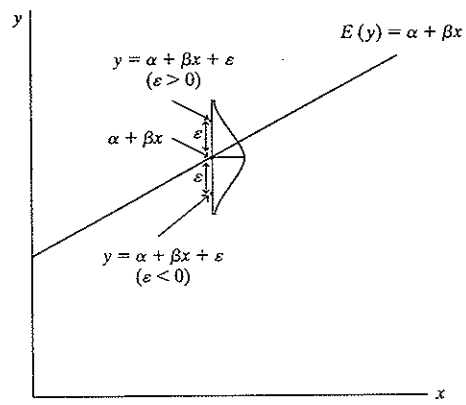
Lembre que para cada valor fixo de x, o modelo de regressão permite aos valores de y flutuar em torno da sua média, $E(y) = \alpha + \beta x$. Qualquer observação pode estar acima da média (isto é, acima da linha de regressão) ou abaixo (abaixo da linha de regressão). O desvio padrão σ resume o tamanho típico dos desvios da média.

Uma formulação alternativa para o modelo expressa cada observação em y, em vez da média $E(y)$ dos valores, em termos de x. Vimos que o *modelo determinista* $y = \alpha + \beta x$ não é realista porque não permite a variabilidade dos valores-y. Para permitir essa variabilidade, incluímos um termo para o desvio da observação y da média:

$$y = \alpha + \beta x + \varepsilon.$$

O termo representado por ε (a letra grega epsilon) representa o desvio entre y e sua média, $\alpha + \beta x$. Cada observação tem seu próprio valor de ε.

Se ε é positivo, então $\alpha + \beta x + \varepsilon$ é maior do que $\alpha + \beta x$ e a observação está acima da média. Veja a Figura 9.19. Se ε é negativo, a observação está abaixo da média. Quando ε



☑ **Figura 9.19** Valores-ε positivos e negativos correspondem a observações acima e abaixo da média da distribuição condicional.

$= 0$, a observação está exatamente na média. A média dos valores-ε é 0.

Para cada x, a variabilidade nos valores-y corresponde à variabilidade em ε. O termo ε é denominado de **termo erro**, visto que ele representa o erro que resulta do uso do valor médio $(\alpha + \beta x)$ de y, para determinado valor de x, para prever um observação individual.

Na prática, não conhecemos os n valores de ε, assim como não conhecemos os valores dos parâmetros e a média verdadeira $\alpha + \beta x$. Para dados amostrais e sua equação de previsão, seja e tal que:

$$y = a + bx + e.$$

Isto é, $y = \hat{y} + e$, tal que $e = y - \hat{y}$. Então, e é o *resíduo*, a diferença entre os valores observados e os previstos de y. Visto que $y = \alpha + \beta x + \varepsilon$, o resíduo e estima ε. Podemos interpretar ε como um **resíduo populacional**. Portanto, ε é a diferença entre a observação y e a média $\alpha + \beta x$ de todas as observações possíveis em y para aquele valor de x. Graficamente, ε é a distância vertical entre o ponto observado e a verdadeira linha de regressão.

Em resumo, podemos expressar o modelo de regressão como:

$$E(y) = \alpha + \beta x \text{ ou como } y = \alpha + \beta x + \varepsilon.$$

Usaremos a primeira equação nos últimos capítulos porque ela se conecta melhor com modelos de regressão para variáveis respostas que têm outras distribuições além da normal. Os modelos para as variáveis quantitativas discretas e para variáveis categóricas são expressos em termos das médias, não em termos dos valores de y.

Modelos e realidade

Novamente, enfatizamos que o modelo de regressão *aproxima* o relacionamento verdadeiro. Nenhum pesquisador consciente espera que um relacionamento seja exatamente linear, com distribuições condicionais exatamente normais em cada x e com exatamente o mesmo desvio padrão dos valores-y em cada valor-x. Por definição, os modelos meramente se aproximam da realidade.

Se o modelo parece muito simples para ser adequado, o diagrama de dispersão ou outros diagnósticos podem sugerir um aperfeiçoamento usando modelos que serão introduzidos mais tarde no livro. Tais modelos podem ser ajustados, revisados e talvez modificados mais adiante. A construção do modelo é um processo repetitivo. Seus objetivos são encontrar um modelo realista que é adequado para descrever o relacionamento e fazer previsões, mas que ainda é simples o suficiente para ser interpretado facilmente. Os Capítulos 11 a 15 estendem o modelo para que ele se aplique a situações nas quais as suposições deste capítulo são muito simplistas.

9.7 RESUMO DO CAPÍTULO

Os Capítulos 7 a 9 trataram da descoberta e descrição da *associação entre duas variáveis*. O Capítulo 7 mostrou como comparar médias ou proporções para dois grupos. O Capítulo 8 tratou da *associação entre duas variáveis categóricas*. As medidas de associação como a diferença das proporções, a razão das chances e gama descrevem a força da associação. A estatística qui-quadrado para dados nominais ou uma estatística z baseada no gama amostral para dados ordinais testa a hipótese de independência.

Este capítulo tratou da *associação entre variáveis quantitativas*. Um elemento novo estudado aqui foi o modelo de regressão para descrever a *forma* do relacionamento entre a variável explicativa x e a média $E(y)$ da variável resposta. Os aspectos principais da análise são os seguintes:

- A **equação de regressão linear** $E(y) = \alpha + \beta x$ descreve a *forma* do relacionamento. Esta equação é apropriada quando o relacionamento entre x e a média de y é aproximadamente linear.

- Um **diagrama de dispersão** permite a visualização dos dados e, portanto, verificar se o relacionamento é aproximadamente linear. Se for, as estimativas dos **mínimos quadrados** do intercepto-y α e da inclinação β fornecem a equação de previsão $\hat{y} = a + bx$ mais próxima dos dados em termos da soma dos resíduos ao quadrado.
- A **correlação** r e seu quadrado descrevem a *força* da associação linear. A correlação é a inclinação padronizada tendo o mesmo sinal do coeficiente b, mas variando no intervalo entre -1 e $+1$. Seu quadrado, r^2, informa a redução proporcional na variabilidade em relação à equação de previsão comparada com a variabilidade em relação a $\bar{y}$.
- Para uma inferência sobre o relacionamento, um teste t usando a inclinação ou a correlação testa a **hipótese nula de independência**, a saber, que a inclinação da população e a correlação são iguais a 0. Um intervalo de confiança para a inclinação estima o tamanho do efeito.

A Tabela 9.8 resume o método estudado nos últimos três capítulos.

O Capítulo 11 introduz o modelo de **regressão múltipla**, uma generalização que permite *muitas* variáveis explicativas no modelo. O Capítulo 12 mostra como incluir previsores categóricos em um modelo de regressão. O Capítulo 13 inclui tanto previsores categóricos quanto quantitativos. O Capítulo 14 introduz modelos para relacionamentos mais complexos, como os não lineares. Finalmente, o Capítulo 15 apresenta modelos de regressão para variáveis respostas categóricas. Antes de discutirmos estes modelos multivariados, entretanto, introduziremos no próximo capítulo alguns conceitos novos que nos auxiliam a entender e interpretar os relacionamentos multivariados.

☑ **Tabela 9.8** Resumo dos testes de independência e medidas de associação

	Níveis de mensuração das variáveis		
	Nominal	**Ordinal**	**Intervalar**
Hipótese nula	H_0: Independência	H_0: Independência	H_0: Independência $(\beta = 0)$
Estatística-teste	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$	$z = \frac{\hat{\gamma}}{ep}$	$t = \frac{b}{ep}, gl = n - 2$
Medida de associação	$\hat{\pi}_2 - \hat{\pi}_1$	$\hat{\gamma} = \frac{C-D}{C+D}$	$r = b\left(\frac{s_x}{s_y}\right)$
	Razão de chances		$r^2 = \frac{SQT - SQE}{SQT}$

EXERCÍCIOS

Praticando o básico

9.1 Para as seguintes variáveis, em uma análise de regressão, qual variável faz o papel de x mais naturalmente (variável explicativa) e qual faz o papel de y (variável resposta)?

(a) Média geral das notas na universidade (GPA) e média geral no ensino médio (GPA)?
(b) Número de filhos e nível de escolaridade da mãe?
(c) Renda anual e número de anos de escolaridade?
(d) Renda anual e avaliação da residência?

9.2 Esboce os diagramas das seguintes equações de previsão, para valores de x entre 0 e 10:
(a) $\hat{y} = 7 + 0{,}5x$
(b) $\hat{y} = 7 + x$
(c) $\hat{y} = 7 - x$
(d) $\hat{y} = 7 - 0{,}5x$
(e) $\hat{y} = 7$
(f) $\hat{y} = x$

9.3 Os antropólogos geralmente tentam reconstruir informações usando partes de cadáveres humanos em cemitérios. Por exemplo, após encontrar um fêmur, eles podem querer prever qual era a altura do indivíduo. Uma equação que eles usam para fazer isto é $\hat{y} = 61{,}4 + 2{,}4x$, onde $\hat{y}$ é a altura prevista e x é o comprimento do fêmur, ambos em centímetros.[3]
(a) Identifique o intercepto-y e a inclinação da equação. Interprete a inclinação.
(b) Um fêmur encontrado em um sítio particular tem um comprimento de 50 cm. Qual é a altura prevista da pessoa que tinha aquele fêmur?

9.4 A OECD ou Organization for Economic Cooperation and Development (Organização para o Desenvolvimento e Cooperação Econômica) consiste em 20 países industrialmente avançados. Para essas nações[4], a equação de previsão relacionando y = taxa de pobreza infantil em 2000 e x = gasto social como um percentual do produto interno bruto é $\hat{y} = 22 - 1{,}3x$. Os valores-y variam de 2,8% (Finlândia) a 21,9% (Estados Unidos). Os valores-x variam de 2% (Estados Unidos) a 16% (Dinamarca).
(a) Interprete o intercepto-y e a inclinação.
(b) Encontre as taxas de pobreza previstas para os Estados Unidos e para a Dinamarca.
(c) A correlação é −0,79. Interprete.

9.5 Olhe para a Figura 2, do artigo disponível em http://ajph.aphapublications.org/cgi/reprint/93/4/652?ck=nck, um diagrama de dispersão para os estados norte-americanos com correlação 0,53 entre x = taxa de pobreza infantil e y = taxa de mortalidade infantil. Aproxime o intercepto-y e a inclinação da equação de previsão mostrada ali.

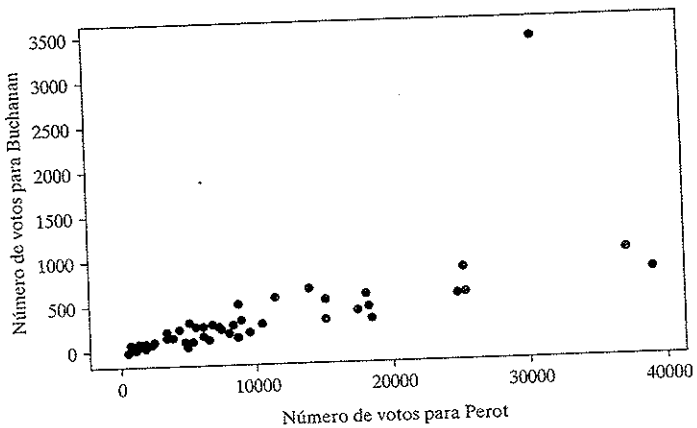
9.6 Um estudo[5] dos padrões da taxa de resposta de um levantamento de dados realizado pelo correio encontrou uma equação de previsão, relacionando x = idade (entre aproximadamente 60 e 90 anos) e y = percentual de sujeitos respondendo, de $\hat{y} = 90{,}2 - 0{,}6x$.
(a) Interprete a inclinação.
(b) Encontre a taxa de resposta prevista para pessoas com (i) 60 anos, (ii) 90 anos.

9.7 Para dados recentes das Nações Unidas (NU) de 39 países em y = emissões de dióxido de carbono *per capita* (toneladas métricas *per capita*) e x = produto interno bruto *per capita* (PIB, em dólares), a equação de previsão foi $\hat{y} = 1{,}26 + 0{,}346x$.
(a) Faça uma previsão de y para o (i) valor mínimo $x = 0{,}8$, (ii) valor máximo $x = 34{,}3$.
(b) Para os Estados Unidos, $x = 34{,}3$ e $y = 19{,}7$. Encontre a resposta prevista para o dióxido de carbono. Encontre o resíduo e interprete.
(c) Para a Suíça, $x = 28{,}1$ e $y = 5{,}7$. Encontre a resposta prevista para o dióxido de carbono e o resíduo. Interprete.

9.8 Um funcionário do setor de ingressos de uma universidade usa a regressão para aproximar o relacionamento entre y = GPA na universidade e x = GPA no ensino médio (ambos mensurados em uma escala de quatro pontos) para estudantes da universidade.
(a) Qual equação é mais realista: $\hat{y} = 0{,}5 + 7{,}0x$ ou $\hat{y} = 0{,}5 + 0{,}7x$? Por quê?
(b) Suponha que a equação de previsão seja $\hat{y} = x$. Identifique o intercepto e a inclinação e interprete a inclinação.

9.9 Para os dados da Tabela 9.1 onde y = taxa de crimes violentos e x = taxa de pobreza, a equação de previsão é $\hat{y} = 209{,}9 + 25{,}5x$.
(a) Interprete o intercepto-y e a inclinação.
(b) Encontre a taxa de crimes violentos prevista e o resíduo para Massachusetts, que tinha $x = 10{,}7$ e $y = 805$. Interprete.
(c) Dois estados diferem por 10,0 nas suas taxas de pobreza. Encontre a diferença nas suas taxas previstas de crimes violentos.
(d) Qual é o sinal da correlação entre essas variáveis? Por quê?

9.10 Na eleição presidencial norte-americana de 2000 o candidato Democrata era Al Gore e o candidato Republicano era George W. Bush. No condado de Palm Beach, Flórida, os resultados iniciais da eleição informaram 3407 votos para o candidato do partido Reformista, Pat Buchanan. Alguns analistas políticos acharam que a maioria destes votos poderia ter sido destinada, na verdade, a Gore (cujo nome estava próximo ao de Buchanan na cédula de voto), mas erroneamente destinados a Buchanan por causa do formato da "célula borboleta" usada naquele condado, que alguns eleitores acharam confusa. Para os 67 condados na Flórida, a Figura 9.20 é um diagrama de dispersão dos votos do condado para os candidatos do partido Reformista em 2000 (Buchanan) e em 1996 (Perot).
(a) O ponto mais alto é para o condado de Palm Beach. O que ele sugere?
(b) A equação de previsão ajustada para todas as observações exceto para o condado de Palm Beach é $\hat{y} = 45{,}7 + 0{,}02414x$. No condado de Palm Beach, $x = 30{,}739$. Encontre os votos previstos para Buchanan, o resíduo e interprete.
(c) Por que o ponto mais alto, mas não os pontos mais à direita são considerados um valor atípico da regressão? (*Nota:* análises estatísticas previram que menos do que 900 dos 3407 votos foram realmente destinados a Buchanan. Bush venceu no estado por 537 votos e, com eles, o Colégio Eleitoral e a eleição. Outros fatores que influenciaram foram 110000 células desclassificadas com "votos a mais" nas quais as pessoas erroneamente votaram em mais do que um candidato a Presidente – com Gore marcado em 84197 células e Bush em 37731 – geralmente por causa da confusão dos nomes estarem listados em mais do que uma página da célula e 61000 "votos a menos" ocasionados por fatores como a "perfuração pendente" das máquinas manuais de perfuração.)



☑ **Figura 9.20** Diagrama de dispersão dos votos dos condados da Flórida para os candidatos do partido reformista Pat Buchanan em 2000 e Ross Perot em 1996.

9.11 A Figura 9.21 é um diagrama de dispersão relacionando y = percentual de pessoas usando telefones celulares e x = produto interno bruto *per capita* (PIB) para países listados no *Human Development Report* (Relatório do Desenvolvimento Humano).

(a) Dê as coordenadas aproximadas x e y para o país que tem o maior (i) uso de telefone celular, (ii) PIB.

(b) A equação de previsão dos mínimos quadrados é $\hat{y} = -0{,}13 + 2{,}62x$. Para os Estados Unidos, $x = 34{,}3$ e $y = 45{,}1$. Encontre o uso previsto do telefone celular e o resíduo. Interprete o resíduo.

(c) A correlação é positiva ou negativa? Explique o significado do sinal da correlação.

9.12 Para os países listados no *Human Development Report* (Relatório do Desenvolvimento Humano), a correlação do percentual de pessoas usando a internet é de 0,888 com o PIB *per capita* (PIB, um resumo da descrição da riqueza de um país), 0,818 com o percentual usando telefones celulares, 0,669 com a taxa de analfabetismo, −0,551 com a taxa de fertilidade (o número médio de filhos por mulher adulta) e 0,680 com as emissões de dióxido de carbono *per capita*.

(a) Explique como interpretar o sinal da correlação entre o uso da internet e (i) o PIB, (ii) a taxa de fertilidade.

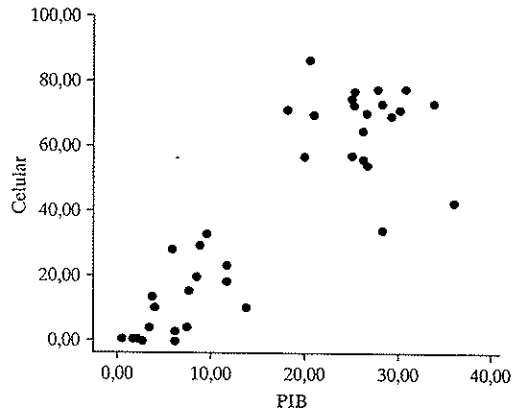
(b) Qual a variável que tem a associação (i) mais forte, (ii) mais fraca com o uso da internet?

9.13 Um relatório resumindo os resultados de um estudo sobre o relacionamento entre um teste x de aptidão verbal e um teste y de aptidão matemática afirma que $\bar{x} = 480$, $\bar{y} = 500$, $s_x = 80$, $s_y = 120$ e $r = 0{,}60$. Usando as fórmulas para a correlação e para as estimativas dos mínimos quadrados, encontre a equação de previsão.

9.14 A Tabela 9.16 no Exercício 9.39 mostra os dados nacionais para muitas variáveis na Flórida. Para esses condados, a Tabela 9.9 mostra parte da saída da análise de regressão relacionando y = renda mediana (em milhares de dólares) a x = percentual de residentes com pelo menos o ensino médio completo.

(a) Determine a equação de previsão e interprete a inclinação.

(b) O condado A tem 10% a mais dos seus residentes com pelo menos o ensino médio completo do que o condado B. Encontre a sua diferença nas rendas medianas previstas.



☑ **Figura 9.21** Diagrama de despersão relacionando y = percentual de pessoas usando celulares e x = produto interno bruto *per capita*.

☑ **TABELA 9.9**

Variável	Média	Desvio Padrão
RENDA	24,51	4,69
EDUC	69,49	8,86

Variável	Parâmetro Estimativa
(Constante)	-4,63
EDUC	0,42

(c) Encontre a correlação. Interprete-a usando (i) o sinal, (ii) a magnitude, (iii) a inclinação padronizada.

(d) Encontre r^2. Explique como interpretá-lo.

9.15 Uma estudante do ensino médio analisa se existe um relacionamento entre x = número de livros lidos por prazer no ano anterior e y = média de horas diárias gastas assistindo à televisão. Para os seus três melhores amigos, a Tabela 9.10 mostra as observações.

(a) Construa um diagrama de dispersão. Pelo diagrama determine a equação de previsão e intreprete-a. (*Nota:* você pode fazer isto sem usar as fórmulas dos mínimos quadrados.)

(b) Por inspeção, determine a correlação amostral entre x e y e interprete-a.

☑ **Tabela 9.10**

x	y
0	5
5	3
10	1

9.16 Para o conjunto de dados dos estudantes no Exercício 1.11 (página 25), a correlação amostral entre y = ideologia política (com escores 1 a 7, com os valores mais altos representando mais conservadorismo) e x = número de vezes por semana que lê um jornal é $r = -0{,}066$.

(a) Você concluiria que a associação amostral é forte ou fraca?

(b) Interprete o quadrado da correlação.

(c) Quando y é previsto usando x = religiosidade (com que frequência comparece em eventos religiosos, com escores 0, 1, 2, 3), a correlação amostral é $r = 0{,}580$. Quais destas duas variáveis explicativas parecem ter um relacionamento linear mais forte com y? Explique.

9.17 Para o estudo do Exemplo 9.6 (página 300) de y = GPA do ensino médio e x = horas semanais assistindo à televisão, $\hat{y} = 3{,}44 - 0{,}03x$.

(a) O estudo encontrou um r ao quadrado = 0,237. Interprete.

(b) Determine e interprete a correlação.

(c) Suponha que você encontrou a correlação somente para aqueles estudantes que assistiram à televisão não mais do que 3 horas por semana. Você esperaria que a correlação fosse mais forte ou mais fraca do que para todos os estudantes? Por quê?

9.18 Para os estudantes que cursam Estatística 101 na Faculdade de Lake Wobegon em Minnesota, tanto x = escore da prova do meio do semestre quanto y = escore do exame final têm média = 75 e desvio padrão = 10.

(a) A equação de previsão $\hat{y} = 30 + 0{,}60x$. Encontre o escore previsto do exame final para um estudante que tem um (i) escore do meio do semestre = 100, (ii) escore do semestre = 50. Observe que o escore previsto do exame final regride do escore do meio do semestre em direção à média.

(b) Mostre que a correlação é igual a 0,60.

(c) Se, em vez disso, tivéssemos $\hat{y} = x$, mostre que $r = 1{,}0$.

(d) Se, em vez disso, tivéssemos $\hat{y} = 75$ (isto é, inclinação = 0), mostre que $r = 0{,}0$

9.19 A equação de previsão relacionando x = anos de escolaridade e y = renda anual (em dólares) é $\hat{y} = -20000 + 4000x$ e a correlação é igual a 0,50. Os desvios padrão são 2,0 para x e 16000 para y.

(a) Mostre como encontrar a correlação a partir da inclinação.

(b) Os resultados foram traduzidos em unidades de euros, em um período em que a taxa de câmbio era de $1,25 por euro. Encontre a equação de previsão e a correlação.

9.20 Um artigo na edição de 16 de setembro de 2006 da *The Economist* mostrou um diagrama de dispersão para vários países relacionando y = consumo anual de gasolina por pessoa (em barris) e x = PIB por pessoa (em milhares de dólares). Os valores (x, y) mostrados no diagrama foram aproximadamente (3, 1) Índia, (8, 2) China, (9, 4) Brasil, (10, 7) México, (11, 8) Rússia, (20, 18) Coreia do Sul, (29, 12) Itália, (30, 13) França, (31, 11) Grã-Bretanha, (31, 12) Alemanha, (31, 16) Japão, (34, 26) Canadá e (41, 26) Estados Unidos. Usando um *software*:

(a) Determine e interprete a equação de previsão.

(b) Determine e interprete a correlação.

(c) Determine e interprete o valor previsto e o resíduo para o Canadá.

9.21 Para os dados da PSG de 2004, a matriz de correlações para a escolaridade do sujeito "EDUC", escolaridade da mãe "MAEDUC" e escolaridade do pai "PAEDUC" é:

	EDUC	PAEDUC	MAEDUC
EDUC	1,00	0,40	0,38
PAEDUC	0,40	1,00	0,65
MAEDUC	0,38	0,65	1,00

Interprete essa matriz identificando o par de variáveis com a associação mais forte e dê a implicação do sinal de cada correlação.

9.22 No *Human Development Report* (Relatório do Desenvolvimento Humano) das Nações Unidas uma variável mensurada foi x = percentual de adultos que usam métodos contraceptivos. A Tabela 9.11 mostra parte de uma saída para uma análise de regressão usando y = fertilidade (número médio de filhos por mulher adulta) para 22 países presentes naquele relatório. Para esses países, x tinha uma média de 60,0 e um desvio padrão de 20,6.

(a) Formule uma questão de pesquisa que poderia ser feita com essa saída.

(b) Identifique a equação de previsão e encontre a fertilidade prevista quando (i) $x = 0$, (ii) $x = 100$. Mostre como a diferença entre eles pode ser obtida usando a inclinação.

(c) Encontre e interprete r e r^2.

(d) O que a sua análise diz sobre a questão em (a)?

☑ **TABELA 9.11**

Previsor	Coef	Erro padrão do Coef	t	Sig.
Constante	6,6633	0,4771	13,97	0,000
CONTRA	-0,064843	0,007539	-8,60	0,000

Fonte	Soma dos quadrados	gl
Regressão	37,505	1
Resíduos (Erro)	10,138	20
Total	47,644	21

9.23 Considerando os dados de vários países, queremos descrever se o percentual de pessoas usando a internet está mais fortemente relacionado com o PIB *per capita* ou com a taxa de fertilidade.

(a) Podemos comparar as inclinações das equações de regressão quando o PIB e a fertilidade prevêm separadamente o uso da internet? Por que sim ou por que não?

(b) Seja x = PIB (milhares de dólares *per capita*). Para dados recentes de 39 países das Nações Unidas onde y = percentual que usa telefones celulares, $\hat{y} = -0{,}13 + 2{,}62x$, enquanto y = percentual que usa a internet, tem-se $\hat{y} = -3{,}61 + 1{,}55x$. Por que faz sentido comparar estas inclinações, deste modo concluindo que um aumento de uma unidade no PIB tem um impacto levemente maior no percentual dos que usam telefones celulares do que no percentual dos que usam a internet?

9.24 Para os dados de vendas de casas na Tabela 9.4, a Tabela 9.12 mostra uma análise de regressão relacionando o preço de venda ao número de quartos.

(a) Identifique a equação de previsão e interprete a inclinação.

(b) Usando a inclinação amostral e os desvios padrão, encontre a correlação. Interprete o seu valor.

(c) Determine r^2 e interprete o seu valor.

(d) Construa um intervalo de 95% de confiança para β e interprete.

(e) Interprete o valor rotulado de "Raiz do EQM".

9.25 Considere o conjunto de dados *2005 statewide crime* no *site* do livro. Para todas as 51 observações, use um *software* para analisar o relacionamento entre y = taxa de assassinatos e x = taxa de pobreza.

(a) Construa um diagrama de dispersão. O relacionamento parece ser positivo ou negativo?

(b) Determine a equação de previsão e encontre a taxa de assassinatos prevista e o resíduo para o D.C. Interprete.

(c) Baseado no diagrama de dispersão, você consideraria o D.C. como um valor atípico da regressão? Ajuste novamente o modelo sem ele e observe o efeito na inclinação.

9.26 Considere os dados do *site* do livro mostrados na Tabela 9.16 do Exercício 9.39, apresentando valores para todos os condados de muitas variáveis na Flórida. Para estes dados use um *software* para analisar y = taxa de crimes e x = percentual que vive em um ambiente urbano.

(a) Construa um diagrama de caule e folhas e um de caixa e bigodes para y. Interprete.

(b) Mostre que $\hat{y} = 24{,}5 + 0{,}56x$. Interprete o intercepto-y e a inclinação.

(c) Encontre a taxa de crimes prevista e o resíduo para o Condado de Alachua. Interprete.

(d) Usando a inclinação, encontre as diferenças nas taxas de crime previstas entre condados que são 100% urbanos e condados que são 0% urbanos. Interprete.

(e) Relate e interprete a correlação e o r^2.

☑ **TABELA 9.12**

Variável	N	Média	Desvio padrão
preço	100	155331	101262,2
quartos	100	3,000	0,651339

Fonte	gl	Soma dos quadrados	Média dos quadrados
Modelo	1	1,575534E+11	1,575534E+11
Erro	98	8,575962E+11	8750981211
Total	99	1,01515E+12	

Raiz do EQM 93547 R Quadrado 0,1552

Variável	Estimativa do parâmetro	Erro padrão	t	Sig.
Intercepto	-28412	44303	-0,64	0,5228
Quartos	61248	14435	4,24	< 0,0001

9.27 A Tabela 9.13, que é o arquivo de dados das Nações Unidas (*UN data*) no *site* do livro, mostra valores de 2005 de vários países sobre o IDH (Índice de Desenvolvimento Humano), que tem componentes que se referem à expectativa de vida ao nascer, nível educacional e rendimento *per capita*), taxa de fertilidade (nascimentos por mulher), percentual de mulheres que usam contraceptivos, percentual que usa telefones celulares, percentual que usa a internet, produto interno bruto *per capita* (PIB em dólares), emissões de dióxido de carbono *per capita* (toneladas métricas), expectativa de vida das mulheres, taxa de alfabetização feminina e taxa de atividade econômica feminina (número de mulheres na força de trabalho por 100 homens na força de trabalho). Este exercício usa y = taxa de fertilidade e x = atividade econômica feminina. A Tabela 9.14 mostra parte de uma saída do SPSS para uma análise de regressão.

(a) Formule uma questão de pesquisa que poderia ser respondida com esta saída.

(b) Identifique a equação de previsão e interprete.

(c) Identifique r e r^2 e interprete.

(d) O que a sua análise sugere sobre a questão proposta em (a)?

9.28 Considere o exercício anterior e agora o percentual que usa contraceptivos como variável explicativa para prever fertilidade. Usando um *software* com os dados no *site* do livro:

(a) Construa um diagrama de caule e folhas ou um de caixa e bigodes para a fertilidade e descreva sua distribuição.

(b) Construa um diagrama de dispersão. Um modelo linear parece ser apropriado?

(c) Ajuste o modelo e interprete as estimativas do parâmetro.

(d) Você pode comparar as inclinações das equações de previsão com os dois previsores para determinar qual tem um efeito mais forte? Explique.

(e) Qual das variáveis, o uso do contraceptivo ou a atividade econômica das mulheres, parece ter a associação mais forte com a fertilidade? Explique.

☑ **Tabela 9.13** Dados das Nações Unidas de vários países, disponíveis no arquivo *UN Data* no *site* do livro

País	IDH	Fert	Cont	Cell	Inter	PIB	CO_2	Life	Liter	FemEc
África do Sul	0,66	2,8	56	36,4	..	3489	7,4	50,2	80,9	59
Alemanha	0,93	1,3	75	78,5	47,3	29115	9,8	81,5	..	71
Algéria	0,72	2,5	64	4,5	..	2090	2,9	72,4	60,1	41
Arábia Saudita	0,77	4,1	32	32,1	6,7	9532	15,0	73,9	69,3	29
Argentina	0,86	2,4	..	..	..	3524	3,5	78,2	97,2	48
Austrália	0,96	1,7	76	71,9	56,7	26275	18,3	82,8	..	79
Áustria	0,94	1,4	51	87,9	46,2	31289	7,8	81,8	..	66
Bélgica	0,95	1,7	78	79,3	38,6	29096	6,8	82,0	..	67
Brasil	0,79	2,3	77	26,4	..	2788	1,8	74,6	88,6	52
Canadá	0,95	1,5	75	41,9	..	27079	16,5	82,4	..	83
Chile	0,85	2,0	..	51,1	27,2	4591	3,6	80,9	95,6	50
China	0,76	1,7	84	21,5	6,3	1100	2,7	73,5	86,5	86
Dinamarca	0,94	1,8	78	88,3	54,1	39332	8,9	79,4	..	85
Egito	0,66	3,3	60	8,4	4,4	1220	2,1	72,1	43,6	46
Espanha	0,93	1,3	81	91,6	23,9	20404	7,3	83,2	..	58
Estados Unidos	0,94	2,0	76	54,6	55,6	37648	20,1	80,0	..	83
Filipinas	0,76	3,2	49	27,0	..	989	0,9	72,5	92,7	62
Finlândia	0,94	1,7	77	91,0	53,4	31058	12,0	81,7	..	87
França	0,94	1,9	75	69,6	36,6	29410	6,2	83,0	..	78
Grã-Bretanha	0,94	1,7	84	91,2	..	30253	9,2	80,6	..	76
Grécia	0,91	1,3	..	90,2	15,0	15608	8,5	80,9	88,3	60
Holanda	0,94	1,7	79	76,8	52,2	31532	9,4	81,1	..	68
Iêmem	0,49	6,2	21	3,5	..	565	0,7	61,9	28,5	37
Índia	0,60	3,1	48	2,5	1,7	564	1,2	65,0	47,8	50
Irã	0,74	2,1	73	5,1	7,2	2066	5,3	71,9	70,4	39
Irlanda	0,95	1,9	..	88,0	31,7	38487	11,0	80,3	..	54
Israel	0,92	2,9	68	96,1	..	16481	11,0	81,7	95,6	69
Japão	0,94	1,3	56	67,9	48,3	33713	9,4	85,4	..	68
Malásia	0,80	2,9	55	44,2	34,4	4187	6,3	75,6	85,4	62
México	0,81	2,4	68	29,5	12,0	6121	3,7	77,5	88,7	49
Nigéria	0,45	5,8	13	2,6	0,6	428	0,4	43,6	59,4	56
Noruega	0,96	1,8	74	90,9	34,6	48412	12,2	81,9	..	86
Nova Zelândia	0,93	2	75	64,8	52,6	19847	8,7	81,3	..	81
Paquistão	0,53	4,3	28	1,8	..	555	0,7	63,2	35,2	44
Rússia	0,80	1,3	73	24,9	..	3018	9,9	72,1	99,2	83
Suécia	0,95	1,6	78	98,0	..	33676	5,8	82,4	..	90
Suíça	0,95	1,4	82	84,3	39,8	43553	5,7	83,2	..	67
Turquia	0,75	2,5	64	39,4	8,5	3399	3,0	71,1	81,1	63
Vietnã	0,70	2,3	79	3,4	4,3	482	0,8	72,6	86,9	91

Fonte: Human Development Report, 2005. disponível em hdr.undp.org/statistic/data.

9.29 Para as 2428 observações da PSG de 2004 em y = número de anos de escolaridade (EDUC) e x = número de anos da escolaridade da mãe (MAEDUC), $\hat{y} = 10{,}5 + 0{,}294x$, com $ep = 0{,}0149$.

(a) Teste a hipótese nula de que estas variáveis são independentes e interprete.

(b) Encontre um intervalo de 95% de confiança para a inclinação da população. Interprete.

(c) A correlação é igual a 0,37. Explique "regressão em direção à média" em termos dessas variáveis.

9.30 Um estudo foi conduzido uilizando 49 mulheres católicas estudantes da Universidade Texas A&M. A variável mensurada se refere aos pais destas estudantes. A variável resposta é o número de filhos que os pais têm. Uma das variáveis explicativas é o nível educacional da mãe, mensurado como o número de anos de educação formal. Para estes dados, $\bar{x} = 9{,}88$, $s_x = 3{,}77$, $\bar{y} = 3{,}35$, $s_y = 2{,}19$, a equação de previsão é $\hat{y} = 5{,}40 - 0{,}207x$, o erro padrão da inclinação estimada é 0,079, e a SQE = 201,95.

(a) Encontre o número previsto de filhos para mulheres com (i) 8, (ii) 16 anos de educação.

(b) Encontre a correlação e interprete o seu valor.

(c) Teste a hipótese nula de que o número médio de filhos é independen-

☑ **Tabela 9.14** Regressão da taxa de fertilidade sobre a atividade econômica feminina

R Quadrado		0,3207		
	B	Erro padrão	t	Sig.
(Constante)	4,845	0,619	7,82	0,000
FEMEC	-0,039	0,00928	-4,18	0,000

te do nível educacional da mãe e determine e interprete o valor-p.

(d) Esboce um diagrama de dispersão potencial tal que a análise que você conduziu acima seria inapropriada.

9.31 A ideologia política está associada à renda? Quando os dados da PSG para 779 casos em 2004 foram usados para fazer a regressão de y = opinião política ("POLVIEWS", usando escores de 1 a 7 com 1 = extremamente liberal e 7 = extremamente conservador) em x = renda do respondente ("RINCOME", usando escores de 1 a 12 para as 12 categorias de renda), obtivemos os resultados mostrados na Tabela 9.15.

(a) Mostre todas as etapas do teste de hipóteses de que a opinião política é independente da renda. Interprete.

(b) O que o SPSS informa sob "Beta" nesta saída? Como você interpretaria este valor?

☑ **Tabela 9.15** Regressão da opinião política sobre a renda

R Quadrado		0,00024			
	B	Erro padrão	Beta	t	Sig.
Constante	4,12668	0,18271		22,58	0,000
RINCOME	0,00739	0,01706	0,01554	0,43	0,665

9.32 Considere o exercício anterior. Quando é feita a regressão da ideologia política em x = número de horas gastas em casa em atividades religiosas no mês anterior ("RELHRS1"), obtemos:

	B	EP de B	Beta	t	Sig.
Constante	4,0115	0,0422		95,10	0,0000
RELHRS1	0,0064	0,0020	0,087	3,20	0,0015

(a) Relate e interprete o valor-p para testar a hipótese de que essas variáveis são independentes.

(b) Use estes resultados para ilustrar que a significância estatística não implica significância prática.

9.33 Para os dados dos países da OCDE na Tabela 3.11 da página 80, use um *software* para construir um diagrama de dispersão relacionando y = emissões de dióxido de carbono e x = PIB.

(a) Baseado neste diagrama, identifique um ponto que pode ter grande influência na determinação da correlação. Mostre que a correlação cai de 0,64 para 0,41 se você remover esta observação do conjunto de dados.

(b) Suponha que você construiu este diagrama usando todos os dados das nações, em vez de somente as nações economicamente avançadas que formam a OCDE. Você esperaria que a correlação fosse mais fraca, aproximadamente do mesmo valor ou mais forte? Por quê?

Conceitos e aplicações

9.34 Para o arquivo *Student Survey* (Levantamento de dados sobre os estudantes) (Exercício 1.11 da página 25), execute uma análise de regressão relacionando (i) y = ideologia política e x = religiosidade, (ii) y = GPA do ensino médio e x = horas que assiste à televisão. Prepare um relatório:

(a) Usando formas gráficas de exibir as variáveis individuais e seu relacionamento.

(b) Interpretando estatísticas descritivas para resumir as variáveis individuais e seu relacionamento.

(c) Resumindo e interpretando resultados de análises inferenciais.

9.35 Considere o arquivo de dados que você criou no Exercício 1.12 (página 26). Para as variáveis escolhidas pelo seu professor, proponha uma questão de pesquisa e execute uma análise de regressão e correlação. Relate tanto uma análise descritiva quanto uma inferencial interpretando e resumindo as suas descobertas.

9.36 Proponha uma questão de pesquisa sobre a satisfação no emprego e nível educacional. Use os dados mais recentes da PSG em "SATJOB" e "EDUC" com a opção da regressão múltipla em sda.berkeley.edu/GSS, com escores (1, 2, 3, 4) para (muito satisfeito, ... , muito insatisfeito), realize uma análise descritiva e inferencial para responder essa questão. Prepare um relatório de uma página resumindo as suas análises.

9.37 Considere o Exercício 3.6 da página 80. Proponha uma questão de pesquisa relacionada à associação entre o percentual de cadeiras no parlamento mantidas por mulheres e a atividade econômica feminina. Usando um *software*, analise os dados da Tabela 3.11 (página 80) para responder esta questão e resuma suas análises.

9.38 Os guias de restaurantes Zagat avaliam cada restaurante em uma escala de 30 pontos para comida, decoração, serviço e preço. O arquivo de dados *Zagat restaurant ratings* no *site* do livro mostra as avaliações de 2007 para restaurantes italianos de Boston, Londres e Nova Iorque. Execute uma análise de correlação para descrever as associações para restaurantes em Boston entre qualidade da comida com as avaliações da decoração, do serviço e do preço.

9.39 A Tabela 9.16 mostra dados de todos os 67 condados da Flórida sobre a taxa de criminalidade (por 1000 residentes), rendimento mediano (em milhares de dólares), percentual de residentes com pelo menos o ensino médio (daqueles com idade acima de 25 anos) e o percentual de residentes morando em áreas urbanas. Usando a taxa de criminalidade como um previsor, analise estes dados (disponíveis no *site* do livro). No seu relatório, forneça interpretações de todas as análises.

9.40 Considere a Tabela 9.1 (página 288), disponível no conjunto de dados *Statewide crime 2* no *site* do livro. Proponha uma questão de pesquisa sobre o relacionamento entre a taxa de assassinatos e o percentual de famílias com pai/mãe solteiro(a). Usando um *software*, realize análises para responder essa questão. Escreva um relatório mostrando as suas análises e forneça interpretações.

9.41 Considere os dados das Nações Unidas sobre vários países, apresentados na Tabela 9.13 (página 328) e disponíveis no *site* do livro. Usando um *software*, obtenha a matriz de correlações. Quais pares de variáveis estão altamente correlacionadas? Descreva a natureza dessas correlações e explique como o *software* usado lidou com os valores que estão faltando.

9.42 Um estudo recente[6], depois de ressaltar que as dietas com muita gordura e açúcares (ruins para a saúde) têm o preço mais acessível do que dietas com muitas frutas e vegetais (boas para a saúde), relatou: "Cada 100g extra de gorduras e doces consumidos diminui os custos da dieta de 0,05 para 0,4 euros, enquanto cada 100g extras de frutas e vegetais consumidos aumentam o custo da dieta de 0,18 para 0,29 euros". Indique os parâmetros aos quais essas interpretações se referem e a inferência estatística que foi executada para dar este resumo.

9.43 O título de um artigo no *Gainesville Sun* (de 17 de outubro de 2003) declarou: "Altura pode gerar um salário maior". Ele descreve uma análise de quatro grandes estudos nos Estados Unidos e Grã-Bretanha de um professor da Universidade da Flórida sobre a altura dos sujeitos e seus salários. O artigo constatou que, para cada gênero, "cada polegada vale aproximadamente $789 por ano em salário. Assim, uma pessoa que tem 6 pés de altura irá ganhar aproximadamente $5523 a mais por ano do que uma pessoa com 5 pés e 5 polegadas".

(a) Para a interpretação entre aspas, identifique a variável resposta e a variável explicativa e determine a inclinação da equação de previsão quando a altura é medida em polegadas e o salário em dólares.

(b) Explique como o valor de $5523 se relaciona à inclinação.

9.44 Em 2002, um levantamento de dados da Census Bureau (Agência do Censo) verificou que o salário médio total que um trabalhador norte-americano de tempo integral espera ganhar dos 25 aos 64 anos de idade é de $1,2 milhões para os que possuem apenas o ensino médio e $2,1 milhões para aqueles com nível universitário.

(a) Assumindo quatro anos para a graduação universitária e uma regressão linear de y = salário médio total em x = número de anos de escolaridade, qual é a inclinação?

(b) Se y for o salário anual (em vez do total em 40 anos), então, qual será a inclinação?

9.45 Explique por que a variabilidade condicional pode ser bem menor do que a variabilidade marginal, usando o relacionamento entre y = peso e x = idade, para uma amostra de meninos entre 2 e 12 anos, para os quais talvez $\sigma_y = 30$, mas o condicional é $\sigma = 10$.

9.46 Descreva uma situação na qual é inapropriado o uso da correlação para medir a associação entre duas variáveis quantitativas.

9.47 A renda anual, em dólares, é uma variável explicativa em uma análise de regressão. Para a versão britânica do relatório da análise, todas as respostas estão convertidas para libras esterlinas britânicas (1 libra era aproximadamente igual a 2,0 dólares, em 2007).

(a) Como muda, se realmente muda, a inclinação da equação de previsão?

(b) Como muda, se realmente muda, a correlação?

9.48 Declare as suposições (a) na utilização da regressão $E(y) = \alpha + \beta x$ para representar o relacionamento entre duas variáveis e (b) para fazer inferências sobre a equação utilizando o método dos mínimos quadrados. Quais as suposições mais críticas?

9.49 Considere o exercício anterior. Tendo em vista estas suposições, indique por que um modelo seria ou não seria bom nas seguintes situações.

(a) x = tempo, y = percentual de trabalhadores norte-americanos desempregados. (*Dica*: isto tende continuamente a aumentar ou diminuir?)

(b) x = renda, y = contribuições beneficentes no ano anterior. (*Dica*: as pessoas pobres mostrariam tanta variação quanto as pessoas mais ricas?)

(c) x = idade, y = despesas médicas anuais. (*Dica*: suponha que as despesas tendem a ser relativamente altas para recém-nascidos e para idosos.)

(d) x = renda *per capita*, y = expectativa de vida dos países. (*Dica*: a tendência crescente tende a se estabilizar.)

9.50 Para uma turma de 100 estudantes, o professor pega os 10 alunos que têm o pior desempenho nos exames do meio do semestre e os inscreve em um programa especial de monitoramento. A média geral da turma é de 70 tanto no exame do meio do semestre quanto no do final do semestre, mas a média para os estudantes com o monitoramento especial aumenta de 50 para 60. Você pode concluir que o programa de monitoramento foi bem-sucedido? Explique.

9.51 Um estudo do Readership Institute[7] da Northwestern University usou dados de um inquérito para analisar como o comportamento do leitor era influenciado pela Guerra do Iraque. A variável resposta foi o Reader Behavior Score (Escore do Comportamento do Leitor, ECL), uma medida combinada resumindo a frequência de utilização do jornal, tempo gasto com ele e o quanto ele era lido. Comparando os escores pré-guerra do ECL, o estudo observou que houve um aumento significativo de leitura por leitores moderados (a média do ECL mudando de 2,05 para 2,32, $p < 0{,}001$), mas uma diminuição significativa nos leitores assíduos (a média do ECL mudou de 5,87 para 5,66, $p < 0{,}001$). Você concluiria que a Guerra do Iraque causou uma mudança no comportamento do leitor ou poderia haver outra explicação?

9.52 Considere o Exercício 9.39. Para estes países, a correlação entre a taxa de escolaridade de ensino médio e o rendimento mediano é igual a 0,79. Suponha que também temos dados tanto a nível individual como agregado para um país. Esboce um diagrama de dispersão para mostrar que, ao nível individual, a correlação poderia ser mais fraca. (*Dica:* mostre que pode existir muita variabilidade para os indivíduos, mas o resumo dos valores dos países poderia estar próximo a uma linha reta. Portanto, é equivocado estender os resultados de valores agregados para indivíduos. Fazer previsões sobre indivíduos baseado no comportamento de agregados é conhecido como **falácia ecológica**. Veja ROBINSON, W. S. *American Sociological Review*. v. 15, p. 351, 1950.)

☑ **Tabela 9.16**

Condado	Taxa de criminalidade	Renda mediana	Ensino médio	Percentual urbano	Condado	Taxa de criminalidade	Renda mediana	Ensino médio	Percentual urbano
ALACHUA	104	22,1	82,7	73,2	LAFAYETTE	0	20,7	58,2	0,0
BAKER	20	25,8	64,1	21,5	LAKE	42	23,4	70,6	43,2
BAY	64	24,7	74,7	85,0	LEE	59	28,4	76,9	86,1
BRADFORD	50	24,6	65,0	23,2	LEON	107	27,3	84,9	82,5
BREVARD	64	30,5	82,3	91,9	LEVY	45	18,8	62,8	0,0
BROWARD	94	30,6	76,8	98,9	LIBERTY	8	22,3	56,7	0,0
CALHOUN	8	18,6	55,9	0,0	MADISON	26	18,2	56,5	20,3
CHARLOTTE	35	25,7	75,7	80,2	MANATEE	79	26,0	75,6	88,7
CITRUS	27	21,3	68,6	31,0	MARION	64	22,5	69,6	39,6
CLAY	41	34,9	81,2	65,8	MARTIN	53	31,8	79,7	83,2
COLLIER	55	34,0	79,0	77,6	MONROE	89	29,4	79,7	73,2
COLUMBIA	69	22,0	69,0	31,1	NASSAU	42	30,2	71,2	44,9
DADE	128	26,9	65,0	98,8	OKALOOSA	37	27,9	83,8	84,0
DESOTO	69	21,0	54,5	44,6	OKEECH.	51	21,4	59,1	30,1
DIXIE	49	15,4	57,7	0,0	ORANGE	93	30,3	78,8	93,1
DUVAL	97	28,5	76,9	98,8	OSCEOLA	78	27,3	73,7	66,4
ESCAMBIA	70	25,2	76,2	85,9	PALM B.	90	32,5	78,8	94,7
FLAGLER	34	28,6	78,7	63,1	PASCO	42	21,5	66,9	67,4
FRANKLIN	37	17,2	59,5	30,2	PINELLAS	70	26,3	78,1	99,6
GADSDEN	52	20,0	59,9	28,8	POLK	84	25,2	68,0	70,3
GILCHRIST	15	20,6	63,0	0,0	PUTNAM	83	20,2	64,3	15,7
GLADES	62	20,7	57,4	0,0	SANTA R.	43	27,6	79,9	57,2
GULF	19	21,9	66,4	35,2	SARASOTA	58	29,9	71,7	92,1
HAMILTON	6	18,7	58,4	0,0	SEMINOLE	56	35,6	78,5	44,4
HARDEE	57	22,1	54,8	16,7	ST JOHNS	54	29,9	81,3	93,2
HENDRY	47	24,9	56,6	44,7	ST LUCIE	58	27,7	84,6	92,8
HERNANDO	44	22,7	70,5	61,3	SUMTER	37	19,6	64,3	19,3
HIGHLANDS	56	21,1	68,2	24,8	SUWANEE	37	19,8	63,8	23,6
HILLSBOR.	110	28,5	75,6	89,2	TAYLOR	76	21,4	62,1	41,8
HOLMES	5	17,2	57,1	16,8	UNION	6	22,8	67,7	0,0
INDIAN R.	58	29,0	76,5	83,0	VOLUSIA	62	24,8	75,4	83,9
JACKSON	32	19,5	61,6	21,7	WAKULLA	29	25,0	71,6	0,0
JEFFERSON	36	21,8	64,1	22,3	WALTON	18	21,9	66,5	20,9
					WASHINGTON	21	18,3	60,9	22,9

Fonte: Dr. Larry Winner, Universidade da Flórida.

9.53 Para qual corpo discente você acha que a correlação entre o GPA do ensino médio e o GPA da universidade seria maior? Yale University ou a University of Bridgeport, em Connecticut? Explique por quê.

9.54 Explique por que a correlação entre x = número de anos de escolaridade e y = renda anual é provavelmente menor se

usarmos uma amostra aleatória de adultos que têm educação universitária em vez de usarmos uma amostra aleatória de todos os adultos.

9.55 Explique cuidadosamente as interpretações dos desvios padrão (a) s_y, (b) s_x, (c) s = raiz quadrada do EQM, (d) ep para b.

9.56 Podemos considerar o problema, estudado no Capítulo 5, de estimar uma única média μ como estimar o parâmetro no modelo simples $E(y) = \mu$, com um único parâmetro. Use este fato para explicar por que a estimativa s_y do desvio padrão de uma distribuição marginal tem $gl = n - 1$.

9.57 O estatístico George Box, que teve uma carreira acadêmica reconhecida na University of Wisconsin, geralmente é citado como tendo dito: "Todos os modelos estão errados, mas alguns modelos são úteis". Por que você acha que, na prática, (a) todos os modelos estão errados, (b) mas alguns modelos *não* são úteis?

9.58 As variáveis y = renda anual (milhares de dólares), x_1 = número de anos de escolaridade e x_2 = número de anos de experiência no emprego, são mensuradas para todos os empregados tendo empregos públicos, em Knoxville, Tennessee. As seguintes equações de previsão e correlações se aplicam.

(i) $\hat{y} = 10 + 1{,}0x_1$, $\quad r = 0{,}30$.

(ii) $\hat{y} = 14 + 0{,}4x_2$, $\quad r = 0{,}60$.

A correlação é $-0{,}40$ entre x_1 e x_2. Quais das seguintes afirmações são verdadeiras? (*Dica*: sete afirmações são verdadeiras.)

(a) A associação amostral mais forte é entre y e x_2.

(b) A associação amostral mais fraca é entre x_1 e x_2.

(c) A equação de previsão usando x_2 para prever x_1 tem uma inclinação negativa.

(d) O aumento do desvio padrão na escolaridade corresponde a um aumento previsto de 0,3 desvios padrão na renda.

(e) Existe uma redução de 30% no erro pelo uso da escolaridade em vez de $\bar{y}$, para prever a renda.

(f) Cada ano adicional no emprego corresponde a um aumento de \$400 na renda prevista.

(g) Quando x_1 é o previsor de y, a soma quadrática dos resíduos (SQE) é maior do que quando x_2 é o previsor de y.

(h) A renda média prevista para empregados com 20 anos de experiência é \$4000 mais alta do que a dos empregados tendo 10 anos de experiência.

(i) Se $s = 8$ para o modelo usando x_1 para prever y, então não seria incomum observar uma renda de \$70000 para um empregado que tem 10 anos de escolaridade.

(j) É possível que $s_y = 12{,}0$ e $s_{x_1} = 3{,}6$.

(k) É possível que $\bar{y} = 20$ e $\bar{x}_1 = 13$.

Selecione a(s) melhor(es) resposta(s) nos Exercícios 9.59 a 9.61. (Mais de uma resposta pode estar correta.)

9.59 Podemos interpretar $r = 0{,}30$ como segue:

(a) Uma redução de 30% no erro ocorre usando x para prever y.

(b) Uma redução de 9% no erro ocorre usando x para prever y comparado ao uso de $\bar{y}$ para prever y.

(c) Em 9% do tempo $\hat{y} = y$.

(d) y muda 0,30 unidades para cada aumento de uma unidade em x.

(e) Quando x prevê y, o resíduo médio é 0,3.

(f) x muda exatamente 0,30 desvios padrão quando y muda um desvio padrão.

9.60 A correlação é inapropriada como uma medida de associação entre duas variáveis quantitativas:

(a) Quando pessoas diferentes mensuram as variáveis usando unidades diferentes.

(b) Quando o relacionamento é altamente não linear.

(c) Quando os pontos de dados estão exatamente em uma linha reta.

(d) Quando a inclinação da equação prevista é 0 usando aproximadamente todos os dados, mas existem dois valores atípicos extremamente altos em y na parte superior da escala x.

(e) Quando y tende a diminuir à medida que x aumenta.

(f) Quando temos dados para toda a população em vez de uma amostra.

(g) Quando a amostra tem um intervalo bem menor para os valores-x do que a população.

9.61 A inclinação da equação de previsão e a correlação são similares no sentido de que:

(a) Elas não dependem das unidades.

(b) Ambas devem estar entre -1 e $+1$.

(c) Ambas têm o mesmo sinal.

(d) Ambas são iguais a 1 quando existe uma associação mais forte.

(e) Os seus quadrados têm redução proporcional nas interpretações do erro.

(f) Elas têm o mesmo valor da estatística t para testar H_0: Independência.

(g) Ambas podem ser fortemente afetadas por valores atípicos severos.

***9.62** Um estudo realizado em 2000 pela National Highway Traffic Safety Administration (Administração Nacional para a Segurança do Trânsito nas Rodovias) estimou que 73% das pessoas usam o cinto de segurança, que o não uso do cinto causou 9200 mortes no ano anterior e que aquele valor iria diminuir para 270 para cada 1 ponto percentual ganho no uso do cinto de segurança. Considere $\hat{y}$ = número previsto de mortes em um ano e x = percentual de pessoas que usam o cinto de segurança. Encontre a equação de previsão que gere esses resultados.

***9.63** As observações em x e y são padronizadas tendo médias estimadas de 0 e desvios padrão de 1 (veja a Seção 4.3, na página 99). Mostre que a equação de previsão tem a forma de $\hat{y} = rx$, onde r é a correlação amostral entre x e y. Isto é, para variáveis padronizadas, o intercepto-y é igual a 0 e a inclinação é igual à correlação.

***9.64** Um intervalo de confiança para a correlação populacional ρ requer uma transformação matemática em r para a qual a sua distribuição amostral é aproximadamente normal. Esta transformação é $T(r) = (1/2)\log_e[(1 + r)/(1 - r)]$, onde $\log_e$ representa o logaritmo natural (base-e). A transformação do valor populacional ρ é representada por $T(\rho)$. A variável $T(r)$ tem distribuição aproximadamente normal sobre $T(\rho)$ com desvio padrão $\sigma_T = 1/\sqrt{n - 3}$. Um intervalo de confiança para $T(\rho)$ é $T(r) \pm z\sigma_T$. Uma vez obtidos os pontos finais do intervalo para $T(\rho)$, substituímos cada ponto final por T na transformação inversa $\rho = (e^{2T} - 1)/(e^{2T} + 1)$, onde e representa a função exponencial (o inverso da função logarítmo natural). Estes dois valores formam os pontos finais do intervalo de confiança para ρ.

(a) Para a correlação de 0,8338 e com os dados dos preços de venda e tamanho das casas apresentados parcialmente na Tabela 9.4, mostre que $T(r) = 1{,}20$. Mostre que o erro padrão de $T(r)$ é 0,1015.

(b) Mostre que um intervalo de 95% de confiança para $T(\rho)$ é (1,00; 1,40).

(c) Mostre que o intervalo de confiança correspondente para ρ é (0,76; 0,89). (A não ser que $r = 0$, o intervalo de confiança para ρ não é simétrico em relação à estimativa por ponto r, em virtude da não simetria da distribuição amostral de r.)

(d) Um intervalo de confiança para o valor populacional ρ^2 a partir de r^2 segue diretamente pela elevação ao quadrado dos limites do intervalo de confiança para ρ. Encontre e interprete este intervalo. (*Nota*: quando um intervalo de confiança para ρ contém 0, o ponto mais inferior do intervalo de confiança para ρ^2 é 0 e o ponto superior é o maior valor dos extremos do intervalo para ρ elevado ao quadrado.)

***9.65** Considere o exercício anterior. Sejam ρ_1 e ρ_2 as corelações populacinais entre duas variáveis para duas populações diferentes. Sejam r_1 e r_2 correlações para amostras aleatórias independentes dessas populações. Para testar H_0: $\rho_1 = \rho_2$, a estatística-teste é:

$$z = \frac{T_2 - T_1}{s_{T_2 - T_1}} \text{ com } s_{T_2 - T_1}$$

$$= \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}},$$

onde T_1 e T_2 são valores transformados de r_1 e r_2. Se H_0 é verdadeira, esta estatística-teste tem uma distribuição aproximadamente normal padrão. Na Tabela 9.4, a correlação entre o preço das casas e o tamanho da casa é $r_1 = 0{,}96$ para as 11 casas novas e $r_2 = 0{,}76$ para as 89 casas mais velhas. Encontre o valor-p para testar H_0: $\rho_1 = \rho_2$ contra H_a: $\rho_1 \neq \rho_2$. Interprete.

***9.66** Considere a fórmula $a = \bar{y} - b\bar{x}$ para o intercepto-y.

(a) Mostre que substituindo $x = \bar{x}$ na equação de previsão $\hat{y} = a + bx$ gera o valor previsto $\hat{y} = \bar{y}$. Mostre que isso significa que a previsão pela equação dos mínimos quadrados passa pelo ponto com coordenadas $(\bar{x}, \bar{y})$, o centro de gravidade dos dados.

(b) Mostre que uma forma alternativa de expressar o modelo de regressão é $\hat{y} - \bar{y} = b(x - \bar{x})$.

(c) Considere y = o escore do exame do final do semestre e x = o escore do exame do meio do semestre. Suponha que a correlação é de 0,70 e o desvio padrão é o mesmo para cada conjunto de escores. Mostre que $(\hat{y} - \bar{y}) = 0{,}70(x - \bar{x})$; isto é, a diferença prevista entre a sua nota do exame do final do semestre e a média da turma é 70% da diferença entre o seu escore do exame do meio do semestre e a média da turma, assim o seu escore está previsto regredir em direção à média.

***9.67** Fórmulas alternativas para definir a correlação usam termos similares àqueles na equação para b:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\left[\sum(x - \bar{x})^2\right]\left[\sum(y - \bar{y})^2\right]}}$$

$$= \frac{1}{n - 1}\sum\left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right).$$

Aproximadamente, a correlação é o produto cruzado médio do escore-z para x vezes do escore-z para y. Usando esta fórmula, explique por que (a) a correlação tem o mesmo valor quando x prevê y do que quando y prevê x, (b) a correlação não depende das unidades de mensuração. (*Nota:* para a população, a correlação é geralmente definida como:

$$\frac{\text{Covariância entre } x \text{ e } y}{(\text{Desvio padrão de } x)(\text{Desvio padrão de } y)}.$$

A **covariância** entre x e y é a média dos produtos cruzados $(x - \mu_x)(y - \mu_y)$ em relação às médias populacionais.)

***9.68** Os valores de y são multiplicados por uma constante c. Das fórmulas, mostre que o desvio padrão s_y e a inclinação dos mínimos quadrados b também são multiplicados por c. Mostre, também, que $r = bs_x/s_y$ permanece o mesmo, isto é, que r não depende das unidades de mensuração.

***9.69** Suponha que o modelo de regressão linear $E(y) = \alpha + \beta x$ com normalidade e desvio padrão constante σ é apropriado. Então, o intervalo:

$$\hat{y} \pm t_{0{,}025}s\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

prevê onde a nova observação de y vai estar para o valor de x dado. Esse intervalo, que para n grande é aproximadamente $\hat{y} \pm 2s$, é um **intervalo de previsão** de 95% para y. Para fazer uma inferência para a *média* de y (em vez de um valor de y individual) para um dado valor de x, podemos usar o **intervalo de confiança**:

$$\hat{y} \pm t_{0{,}025}s\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}.$$

Para n grande e próximo a $\bar{x}$ este intervalo é aproximadamente $\hat{y} \pm 2s\sqrt{1/n}$. O valor-t nestes intervalos é baseado em $gl = n - 2$. A maioria dos *softwares* tem opções para calcular estas fórmulas.

Considere o arquivo de dados *house selling price* (preço de vendas de casas) no *site* do livro.

(a) Usando um *software*, encontre um intervalo de 95% de confiança para o preço de uma casa com um tamanho $x = 2000$.

(b) Usando um *software*, encontre um intervalo de 95% de confiança para o preço médio das casas com um tamanho $x = 2000$.

(c) Explique intuitivamente por que um intervalo de previsão para uma única observação é muito maior do que um intervalo de confiança para a média.

(d) Explique por que os intervalos de previsão seriam provavelmente errados se, de fato, (i) a variabilidade nos preços das casas tendesse a aumentar à medida que o tamanho da casa aumentasse, (ii) a variável resposta é discreta, como y = número de filhos no Exercício 9.30.

☑ NOTAS

[1] www.iusb.edu/~journal/2002/hershberger/hershberger.html.

[2] Sob a suposição de normalidade com um σ comum, as estimativas dos mínimos quadrados são casos especiais das estimativas de *máxima verossimilhança*, introduzidas na Seção 5.1 (página 131).

[3] JUNGER, S. *Vanity Fair*. Outubro de 1999.

[4] *Fonte:* Figura 8H em www.stateofworkingamerica.org.

[5] KALDENBERG, D. et al. *Public Opinion Quarterly*, v. 58, p. 68, 1994.

[6] FRAZAO, E., GOLAN, E. *Evidence-Based Healthcare e Public Health*, v. 9, p. 104-7, 2005.

[7] http://www.readership.org/consumers/data/FINAL_war_study.pdf.