# Noções sobre Regressão

7

(página deixada intencionalmente em branco)

O Capítulo 6 mostrou como se estuda a *relação entre duas variáveis*. Muitas vezes, porém, interessa estudar *como* uma variável varia em função da outra. Por exemplo, todos nós sabemos que as crianças crescem — as variáveis idade e altura têm correlação positiva — mas é preciso saber também *como* a altura de uma criança varia em função da idade. Todos nós sabemos que a população do Brasil aumentou nas últimas décadas. Mas como e quanto? Para dar uma primeira resposta a estas questões, é importante desenhar um gráfico de linhas.

## 7.1 – Gráfico de linhas

Para aprender como se faz um gráfico de linhas, vamos pensar em duas variáveis numéricas e — como fizemos no Capítulo 6 — chamar uma delas de $X$ e a outra de $Y$. Então cada unidade da amostra fornece dois valores, um para cada variável.

Quando se estuda a variação da variável $Y$ em função da variável $X$, diz-se que $Y$ é a *variável dependente* e que $X$ é a *variável explanatória*. Por exemplo, altura de criança varia em função da idade. Então *altura* é a *variável dependente* e *idade* é a *variável explanatória*.

Quem trabalha na área de saúde costuma observar *como* uma variável evolui ao longo do tempo. Com os dados observados de $Y$ ao longo do tempo $X$, é possível fazer um *gráfico de linhas*. Para fazer esse gráfico:

- Colete valores da variável $Y$ nos tempos que você quer estudar.
- Trace um sistema de eixos cartesianos; represente o tempo ($X$) no eixo das abscissas e a variável $Y$ no eixo das ordenadas.
- Estabeleça as escalas e faça, em cada eixo, as necessárias graduações.
- Escreva os nomes das variáveis nos respectivos eixos.
- Desenhe um ponto para representar cada par de valores ($X$, $Y$).
- Una os pontos por segmentos de reta.
- Escreva o título.

**Exemplo 7.1: Gráfico de linhas.**

Na Tabela 7.1 são dados pares de valores das variáveis $X$ e $Y$. A variável $X$ é o ano do Censo Demográfico do Brasil e a variável Y é a população residente. Veja a Figura 7.1: o gráfico de linhas mostra o crescimento no período de forma a complementar os dados da Tabela 7.1.

**TABELA 7.1**
**População residente no Brasil, segundo o ano do censo demográfico.**

| *Ano do censo* | *População* |
|---|---|
| 1940[1] | 41.236.315 |
| 1950[1] | 51.944.397 |
| 1960[1] | 70.191.370 |
| 1970 | 93.139.037 |
| 1980 | 119.002.706 |
| 1991 | 146.815.796 |
| 2000 | 169.799.170 |

Fonte: IBGE (2003)[1]

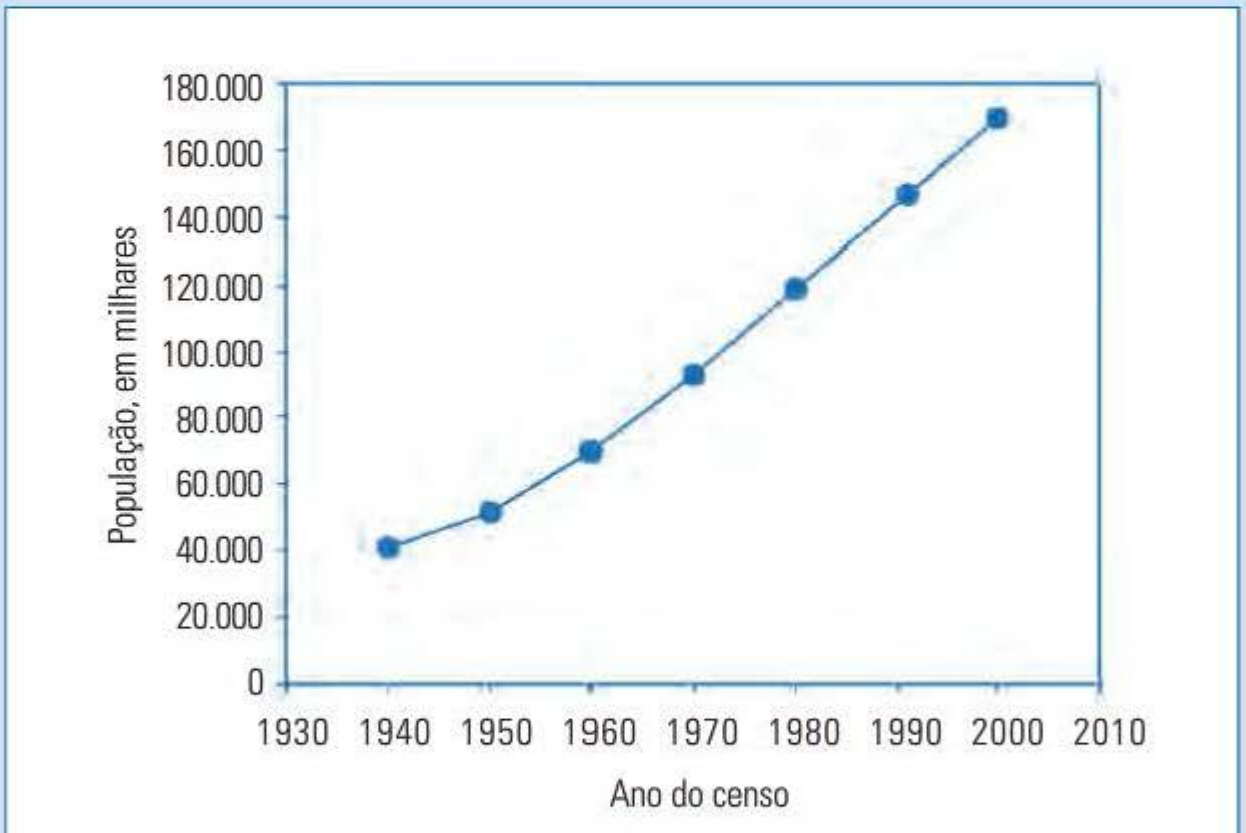


**FIGURA 7.1** População residente no Brasil, segundo o ano do censo demográfico.

[1]IBGE. Censo 2000: um retrato do Brasil na década de 90. Disponível em:< http: /www.ibge.gov.br>. Acesso em: abr. 2003.

## 7.2 – Reta de regressão

A variação de $Y$ em função de $X$ deve ser observada no gráfico de linhas. Se os pontos ficam dispersos em torno de uma reta, é razoável traçar uma reta no meio desses pontos. A *melhor* reta (melhor, no sentido que tem propriedades estatísticas desejáveis) recebe o nome de *reta de regressão*[2]. São dadas, nesta seção, as fórmulas para obter essa reta.

**Exemplo 7.2: A idéia de regressão.**

Observe os dados apresentados na Tabela 7.2. Foi colocada a mesma quantidade de plasma humano em oito tubos de ensaio e depois se ajuntou, em cada tubo, uma quantidade fixa de procaína (anestésico local). Mediu-se então, em tempos diferentes, a quantidade de procaína que já havia se hidrolisado. O diagrama de dispersão apresentado na Figura 7.2 mostra que a quantidade de procaína hidrolisada varia em função do tempo decorrido após sua administração.

**TABELA 7.2**
**Quantidade de procaína hidrolisada, em 10 moles/litro, no plasma humano, em função do tempo, em minutos, decorrido após sua administração.**

| *Tempo* | *Quantidade hidrolisada* |
|---|---|
| 2 | 3,5 |
| 3 | 5,7 |
| 5 | 9,9 |
| 8 | 16,3 |
| 10 | 19,3 |
| 12 | 25,7 |
| 14 | 28,2 |
| 15 | 32,6 |

[2]Muitos autores referem-se à reta de regressão como reta de mínimos quadrados porque esse é o método estatístico utilizado para chegar às fórmulas dadas nesta Seção.

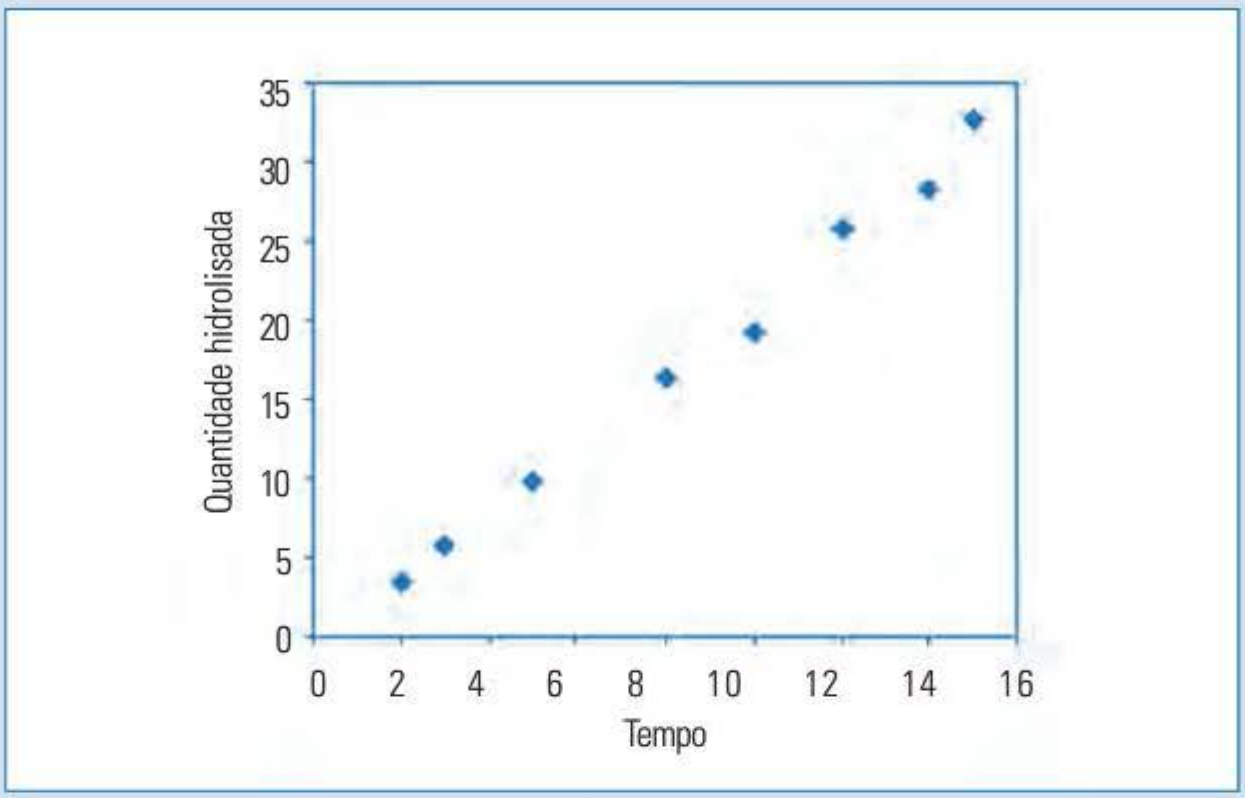


**FIGURA 7.2** Quantidade de procaína hidrolisada, em 10 moles/litro, no plasma humano, em função do tempo, em minutos, decorrido após sua administração.

Vamos discutir um pouco mais o Exemplo 7.2. Parece razoável concluir, observando a Figura 7.2, que a variação da quantidade de procaína hidrolisada no plasma humano em função do tempo decorrido após sua administração pode ser descrita por meio de *uma reta de regressão*.

Para *ajustar uma reta de regressão* (isto é, estabelecer a equação da reta) aos dados apresentados na Tabela 7.2, é preciso obter o coeficiente linear e o coeficiente angular da reta, também chamados *coeficientes de regressão*. Convém lembrar o que são esses coeficientes.

No sistema de eixos cartesianos, a equação $Y = a + bX$ é uma reta. O *coeficiente linear* da reta, indicado neste livro por $a$, dá a *altura* em que a reta corta o eixo das ordenadas. Se $a$ for um número:

- *positivo*, a reta corta o eixo das ordenadas *acima* da origem;
- *negativo*, a reta corta o eixo das ordenadas *abaixo* da origem.
- *zero*, a reta passa na origem do sistema de eixos cartesianos.

**Exemplo 7.3: Equação da reta: coeficientes lineares diferentes.**

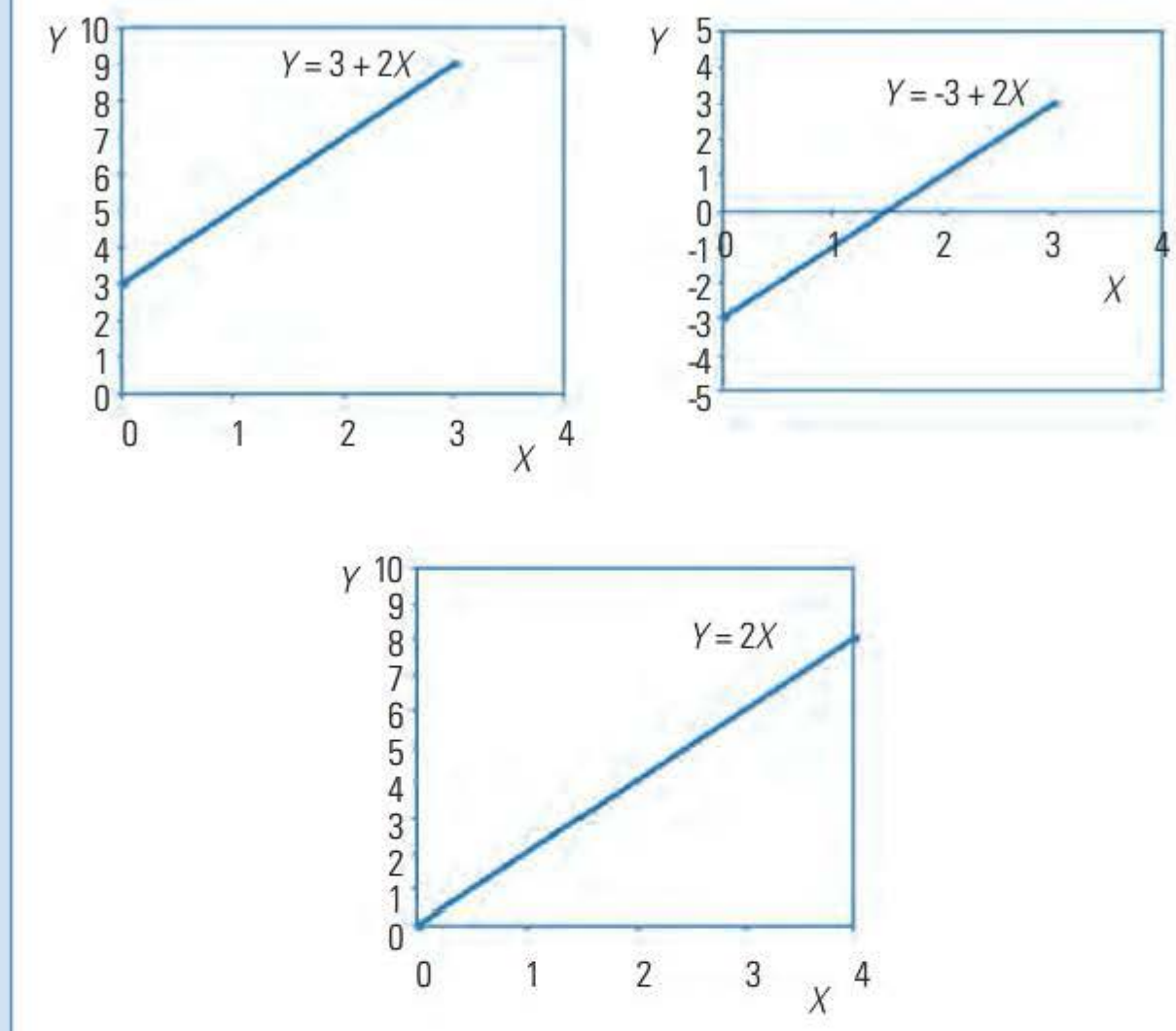


**FIGURA 7.3** Apresentação gráfica de retas com diferentes coeficientes lineares.

O *coeficiente angular da reta*, indicado neste livro por *b*, dá a inclinação da reta[3]. Se *b* for um número:

- *positivo*, a reta é ascendente;
- *negativo*, a reta é descendente;
- *zero*, a reta é paralela aos eixos das abscissas.

[3] O coeficiente angular, chamado neste livro de *b*, é a tangente trigonométrica do ângulo formado pelo eixo das abscissas e pela reta de equação $Y = a + bX$.

**Exemplo 7.4: Equação da reta: coeficientes angulares diferentes.**

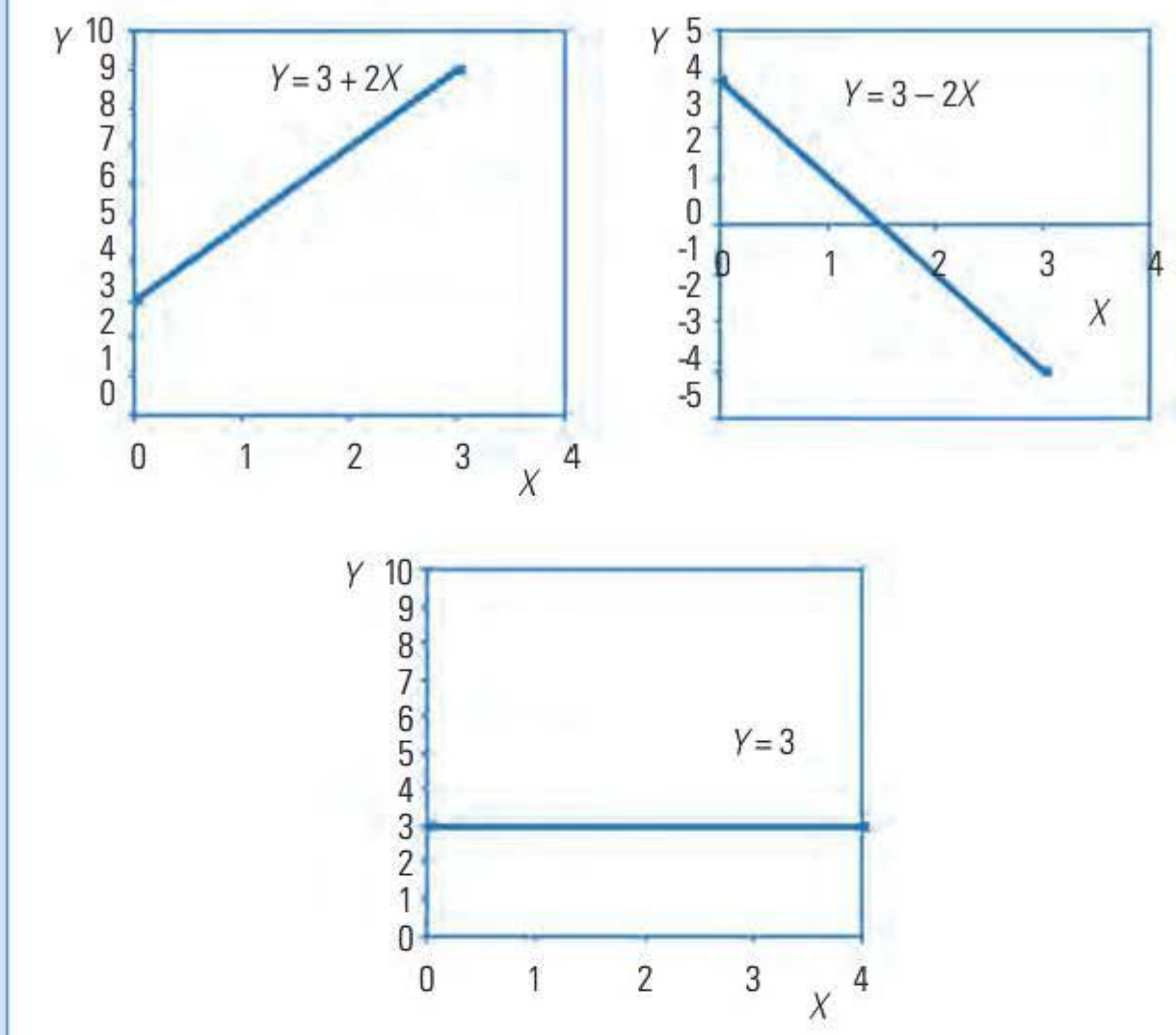


**FIGURA 7.4** Apresentação gráfica de retas com diferentes coeficientes angulares.

Em Estatística, o coeficiente angular da reta é obtido por meio da fórmula:

$$b = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X - \frac{(\sum X)^2}{n}}$$

e o coeficiente linear é obtido por meio da fórmula:

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

em que $\overline{Y}$ e $\overline{X}$ são as médias de $Y$ e $X$, respectivamente. Veja o Exemplo 7.5.

**Exemplo 7.5: Cálculo dos coeficientes de regressão.**

Calcule a reta de regressão para o problema apresentado no Exemplo 7.2.

**TABELA 7.3**
**Cálculos intermediários para a obtenção de *a* e de *b*.**

| X | Y | XY | $X^2$ |
|---|---|---|---|
| 2 | 3,5 | 7 | 4 |
| 3 | 5,7 | 17,1 | 9 |
| 5 | 9,9 | 49,5 | 25 |
| 8 | 16,3 | 130,4 | 64 |
| 10 | 19,3 | 193 | 100 |
| 12 | 25,7 | 308,4 | 144 |
| 14 | 28,2 | 394,8 | 196 |
| 15 | 32,6 | 489 | 225 |
| 69 | 141,2 | 1.589,2 | 767 |

Aplicando as fórmulas, obtém-se:

$$b = \frac{1589{,}2 - \frac{69 \times 141{,}2}{8}}{767 - \frac{69^2}{8}} = \frac{371{,}35}{171{,}875} = 2{,}16$$

$$a = \frac{141{,}2}{8} - 2{,}16 \times \frac{69}{8} = -0{,}98$$

Para traçar a *reta de regressão* é preciso dar valores arbitrários para $X$ e depois calcular os valores de $Y$. Indicam-se os valores calculados de $Y$ por $\hat{Y}$.

Fazendo $X = 5$, tem-se que:

$$\hat{Y} = -0{,}98 + 2{,}16 \times 5 = 9{,}82$$

e fazendo $X = 15$, tem-se que:

$$\hat{Y} = -0{,}98 + 2{,}16 \times 15 = 31{,}42.$$

Os dois pares de valores ($X = 5$ e $\hat{Y} = 9{,}82$) e ($X = 15$ e $\hat{Y} = 31{,}42$) permitem traçar a reta de regressão.

**Exemplo 7.6: Traçado da reta de regressão.**

Apresente, no diagrama de dispersão da Figura 7.2, a reta de *equação*

$$\hat{Y} = -0{,}98 + 2{,}16\,X.$$

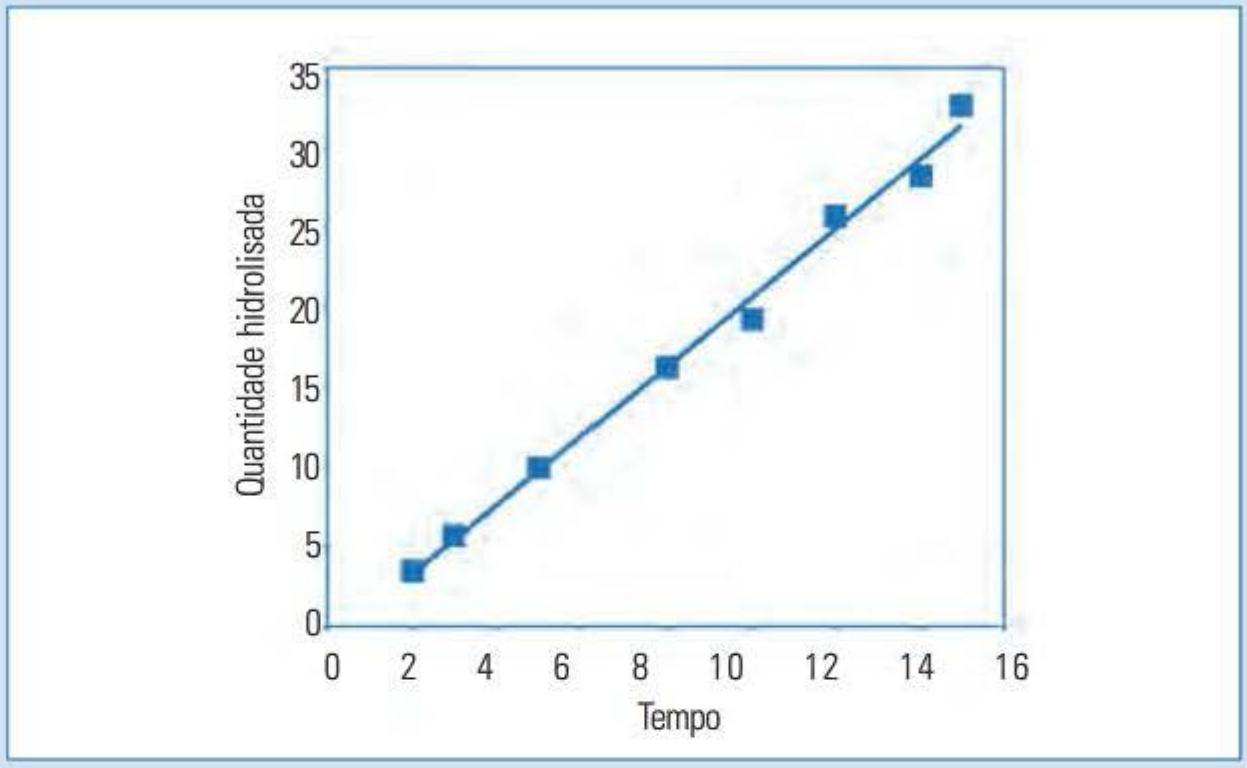


**FIGURA 7.5** Reta de regressão: quantidade de procaína hidrolisada, em 10 moles/litro, no plasma humano, em função do tempo, em minutos, decorrido após sua administração.

A equação da reta de regressão permite *estimar* valores de $Y$ para quaisquer valores de $X$ dentro do intervalo estudado, mesmo que tais valores não existam na amostra. Observe os dados apresentados na Tabela 7.2. Não existe o valor $X = 13$, mas é possível estimar o valor de $Y$ para $X = 13$. Basta fazer:

$$\hat{Y} = -0{,}98 + 2{,}16 \times 13 = 27{,}10$$

O valor $\hat{Y} = 27{,}10$ é uma *previsão*, feita com base na equação da reta de regressão, para a quantidade de procaína que deve estar hidrolisada 13 minutos após sua administração.

Dada a reta de regressão, fica fácil calcular o valor de $Y$ para qualquer valor de $X$. No entanto, o bom senso deve fazer com que você *não* estime valores de $Y$ para valores de $X$ muito além do intervalo estudado: a *extrapolação* pode levar ao absurdo, porque a relação entre $X$ e $Y$, linear no intervalo estudado, pode não ser linear fora desse intervalo.

É verdade que as pessoas tendem a prever, com base no que se observou em determinado período, o que acontecerá em outro período, próximo ou longínquo. A *extrapolação* é, geralmente, incorreta ou até desastrosa. Por exemplo, por volta dos 6 anos começam a irromper dentes permanen-

tes em crianças, mas isso só acontece até certa idade. Ninguém espera, pelo fato de terem irrompido quatro dentes numa criança entre os 7 e os 8 anos, que isso ocorra entre 30 e 31 anos de idade.

**Exemplo 7.7: A extrapolação indevida.**

A Tabela 7.4 apresenta as temperaturas médias mensais, nos primeiros sete meses do ano, de uma cidade do sul do Brasil. Esses dados estão no diagrama de dispersão da Figura 7.6. Se alguém ajustar uma reta como a mostrada no diagrama e quiser usar essa reta para "prever" a temperatura na cidade em dezembro (mês 12), chegará a um valor absurdo, menor do que 2 graus negativos. A razão disso é óbvia: o fenômeno não é linear além do período estudado.

**TABELA 7.4**
**Temperaturas médias segundo o mês, de uma cidade do sul do Brasil.**

| *Mês* | *Número do mês* | *Temperatura média no mês* |
|---|---|---|
| Janeiro | 1 | 23 |
| Fevereiro | 2 | 22 |
| Março | 3 | 20 |
| Abril | 4 | 18 |
| Maio | 5 | 15 |
| Junho | 6 | 12 |
| Julho | 7 | 9 |

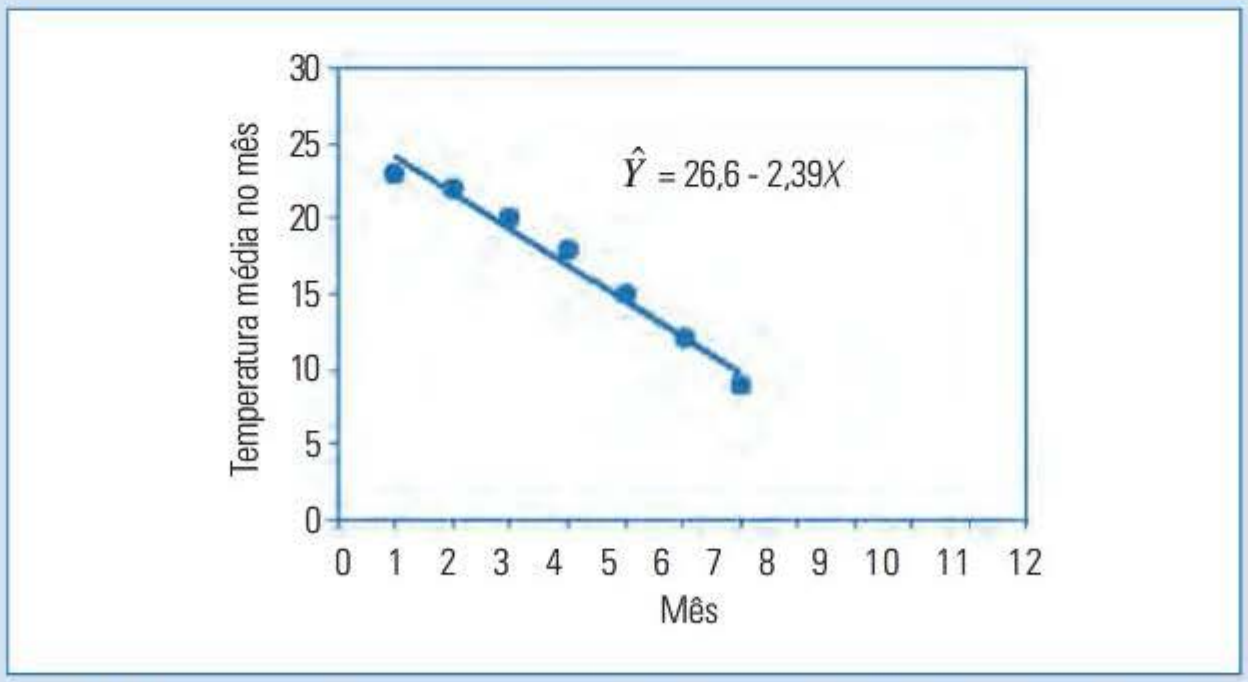


**FIGURA 7.6** Reta ajustada às temperaturas médias de uma cidade do sul do Brasil, segundo o mês.

## 7.3 – Escolha da variável explanatória

Quando os valores de $X$ são fixados antes do início da coleta dos dados, ajusta-se a regressão de $Y$ contra $X$. No Exemplo 7.2, o pesquisador fixou os tempos em que iria observar a quantidade de procaína que estaria hidrolisada no plasma, antes de iniciar a pesquisa. Então, a quantidade de procaína hidrolisada *depende* do tempo em que foi medida — não o contrário.

Nem sempre os valores de $X$ são fixados *antes* do início dos trabalhos. Nesses casos, tanto se pode ajustar a regressão de $Y$ contra $X$, como a regressão de $X$ contra $Y$, mas recomenda-se identificar a variável que *deve ser prevista*, conhecido o valor da outra variável e ajustar a regressão de $Y$ contra $X$ toda vez que se pretende estudar a variação de $Y$ (prever $Y$) em função da variação de $X$.

**Exemplo 7.8: A escolha da variável explanatória.**

Calcule a reta de regressão para os dados apresentados na Tabela 7.5.

É razoável estudar a variação da pressão arterial ($Y$) em função do peso ($X$), porque é o peso que pode explicar (explanar) a pressão arterial — e não o contrário. Então se deve ajustar uma regressão da pressão arterial ($Y$) contra o peso ($X$).

**TABELA 7.5**
**Pressão arterial (PA), em milímetros de mercúrio e peso de homens adultos, em quilogramas.**

| *Peso* | *PA* | *Peso* | *PA* | *Peso* | *PA* |
|---|---|---|---|---|---|
| 14 | 105 | 18 | 113 | 21 | 127 |
| 14 | 102 | 19 | 107 | 22 | 125 |
| 15 | 111 | 19 | 125 | 22 | 116 |
| 15 | 104 | 19 | 130 | 23 | 130 |
| 15 | 107 | 19 | 110 | 23 | 107 |
| 16 | 90 | 19 | 107 | 23 | 103 |
| 16 | 105 | 20 | 102 | 24 | 135 |
| 16 | 102 | 20 | 116 | 24 | 143 |
| 16 | 126 | 21 | 135 | 28 | 121 |
| 17 | 134 | 21 | 100 | 28 | 135 |

Foram calculados:

$$b = \frac{271159 - \dfrac{3624 \times 2238}{30}}{167386 - \dfrac{2238^2}{30}} = 1{,}88$$

$$a = \frac{3624}{30} - 1{,}88 \times \frac{2238}{30} = -19{,}1$$

A reta de regressão

$$\hat{Y} = -19{,}1 + 1{,}88X$$

apresentada na Figura 7.7 mostra a *tendência* de ocorrer aumento de pressão arterial quando aumenta o peso, mas convém observar que os pontos estão *muito dispersos* em torno da reta. Isso significa que a *previsão* da pressão arterial de um homem adulto em função de seu peso tem grande margem de erro.

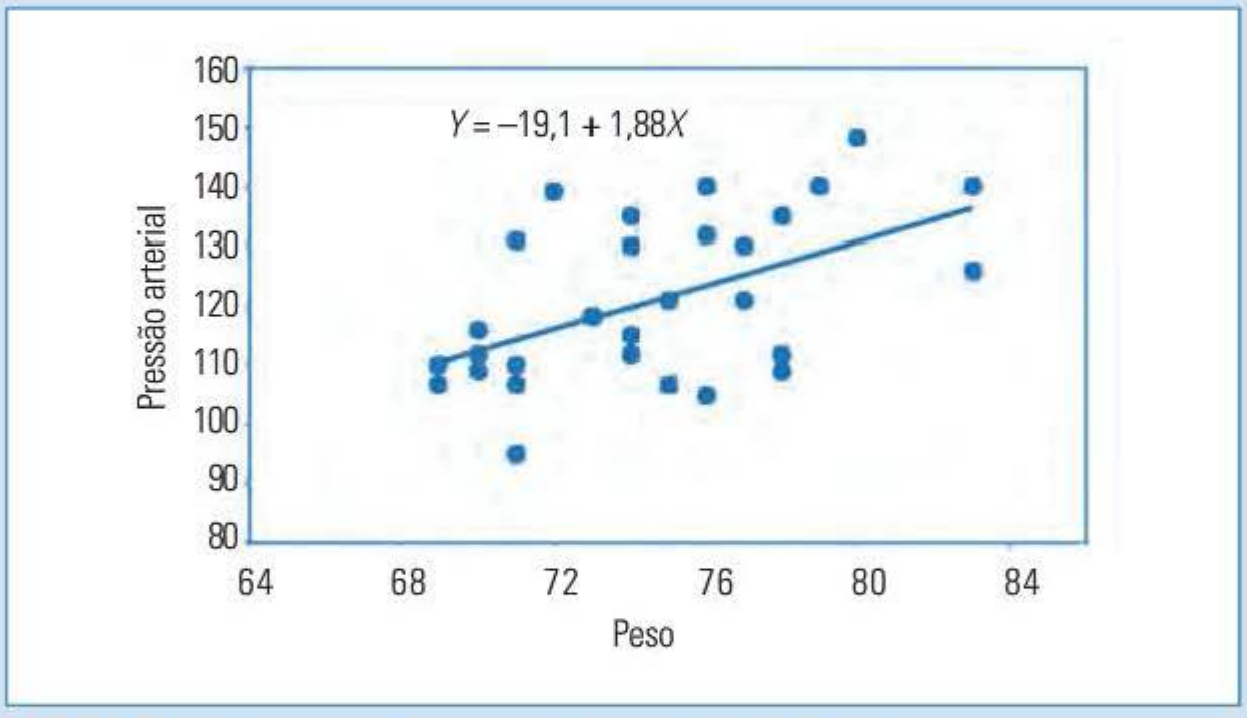


**FIGURA 7.7** Reta de regressão para pressão arterial em função do peso.

## 7.4 – Coeficiente de determinação

Antes de aprender o que é coeficiente de determinação, vamos entender o que é uma relação matemática e o que é uma relação estatística. Se você aumentar o lado de um quadrado em 1 cm, a área aumenta. E se você continuar aumentando o lado do quadrado de 1 cm em 1 cm, a área continuará aumentando. Você sabe dizer *exatamente* a área do quadrado para cada tamanho de lado porque a relação entre a área de um quadrado e seus lados é matemática: área = lado × lado.

Pense agora em uma pessoa que quer diminuir o peso porque — seu médico lhe disse — os gordos têm tendência a ter pressão arterial alta. Sabe-se, portanto, que o aumento da pressão arterial é função do aumento de peso. Será que existe uma *relação exata* entre essas duas variáveis, isto é, para cada quilo a mais haverá um aumento fixo na pressão arterial? *Não* é assim. Existe tendência de a pressão arterial aumentar com o aumento de peso, mas a pressão arterial também aumenta em função de outros fatores como idade, vida sedentária, hereditariedade e certos hábitos, como o hábito de fumar e o consumo excessivo de sal. E mesmo que conhecêssemos muitas das causas que explicam o aumento da pressão arterial, ainda assim não saberíamos prever *exatamente* a pressão arterial de uma pessoa. A relação entre pressão arterial e peso é probabilística e, portanto, sujeita a erro.

Com estes exemplos queremos lembrar a você que existem *relações determinísticas* — como é a relação entre lado e área de um quadrado — e *relações probabilísticas* — como é a relação entre peso e pressão arterial. No primeiro caso, não existe erro na previsão, isto é, dado o lado de um quadrado você pode dizer *exatamente* qual é a área: está determinado. No segundo caso, a previsão é possível, mas dentro de certas margens de erro. Neste ponto, a pergunta é inevitável: qual é o "tamanho" desse erro?

Existe uma estatística chamada *coeficiente de determinação*, indicada por $R^2$, que mede a *contribuição* de uma variável na *previsão* de outra. Parece complicado, mas tente entender este exemplo: imagine que você quer comprar uma camiseta para uma criança. Você chega na loja e pede ajuda à vendedora. O que primeiro ela pergunta? A idade da criança, claro. Por quê? Porque o tamanho de uma criança é função da idade. Boa parte da variação do tamanho das crianças é explicada pela variação de suas idades — o que é medido pelo $R^2$. Portanto, saber a idade da criança ajuda na *previsão* do tamanho da sua camiseta[4].

O *coeficiente de determinação* é a proporção da variação de *Y* explicada pela variação de *X*.

O *coeficiente de determinação* é dado pelo quadrado do coeficiente de correlação. *Não* pode, portanto, ser negativo. Varia entre zero e 1, inclusive. Para interpretar o coeficiente de determinação, é melhor transformá-lo em porcentagem, multiplicando o resultado obtido em seu cálculo por 100. Veja o Exemplo 7.9.

[4]A vendedora também pergunta se o presente é para menino ou menina. Essa informação também contribui, embora menos do que idade, para a escolha do tamanho (na primeira infância os meninos são maiores), mas ajuda na escolha do modelo.

**Exemplo 7.9. Coeficiente de determinação.**

Calcule o coeficiente de determinação para os dados apresentados na Tabela 7.2 e na Tabela 7.5 e discuta cada um deles.

Usando os cálculos intermediários já apresentados na Tabela 7.3, é possível obter $R^2 = 0{,}994$. Isto significa que 99,4% da variação da quantidade de procaína hidrolisada no plasma se explica pelo tempo decorrido após sua administração. Em outras palavras, se você souber o tempo que decorreu depois que a procaína foi colocada no plasma, poderá justificar 99,4% da variação de procaína que hidrolisou.

Para os dados da Tabela 7.5, com a ajuda de um computador (ou de seu professor) é possível obter, $R^2 = 0{,}282$, um valor baixo. Se fosse alto, a explicação seria de que, dado o peso de um homem, a pressão arterial seria altamente previsível. No entanto, fatores como idade, vida sedentária, hereditariedade e certos hábitos, como o hábito de fumar e consumo abusivo de sal devem ser, também, importantes.

## 7.5 – Uma pressuposição básica

Para ajustar uma regressão linear simples de $X$ contra $Y$, é preciso que os dados de $X$ e $Y$ tenham sidos *obtidos independentemente*. Então, quando você for interpretar os resultados do ajuste de uma regressão, verifique como foram obtidos os dados de $X$ e $Y$. Veja o Exemplo 7.7: a regressão obtida é uma *falácia* porque não se pode fazer uma regressão da diferença das variáveis contra o valor inicial.

**Exemplo 7.10: Uma falácia.**

Observe os dados da Tabela 7.6, que estão no diagrama de dispersão da Figura 7.8: *os pontos não sugerem correlação entre as variáveis*. O coeficiente de determinação é $R^2 = 0{,}030$. No entanto, se você fizer a diferença $Y$-$X$ e colocar a diferença como função do valor inicial ($X$), obterá o diagrama de dispersão da Figura 7.9, com $R^2 = 0{,}582$. Só que isso *não* pode ser feito: a regressão obtida é uma falácia.

**TABELA 7.6**
**Notas de 10 alunos em duas provas.**

| *1ª prova* | *2ª prova* | *Diferença = 2ª prova –1ª prova* |
|---|---|---|
| 7 | 7 | 0 |
| 5 | 5 | 0 |
| 4 | 8 | 4 |
| 9 | 9 | 0 |
| 2 | 10 | 8 |
| 4 | 3 | –1 |
| 8 | 4 | –4 |
| 10 | 6 | –4 |
| 6 | 4 | –2 |
| 7 | 3 | –4 |

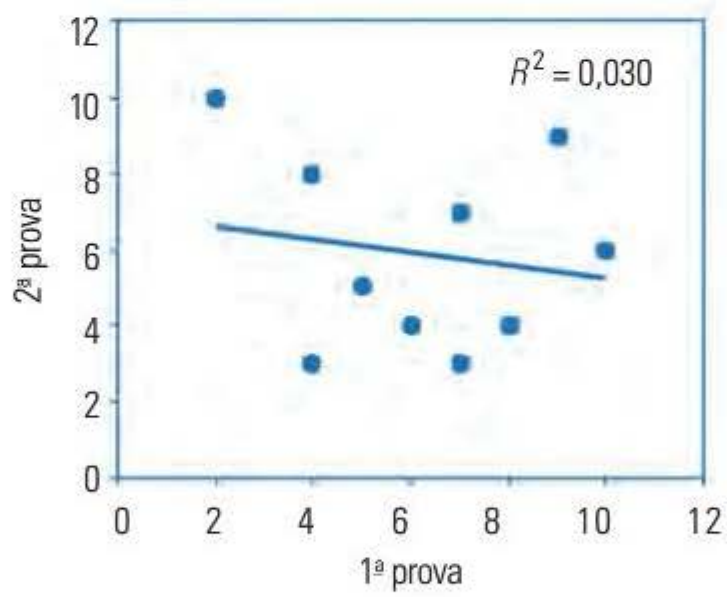


**FIGURA 7.8** Nota na segunda prova em função da nota na primeira prova.

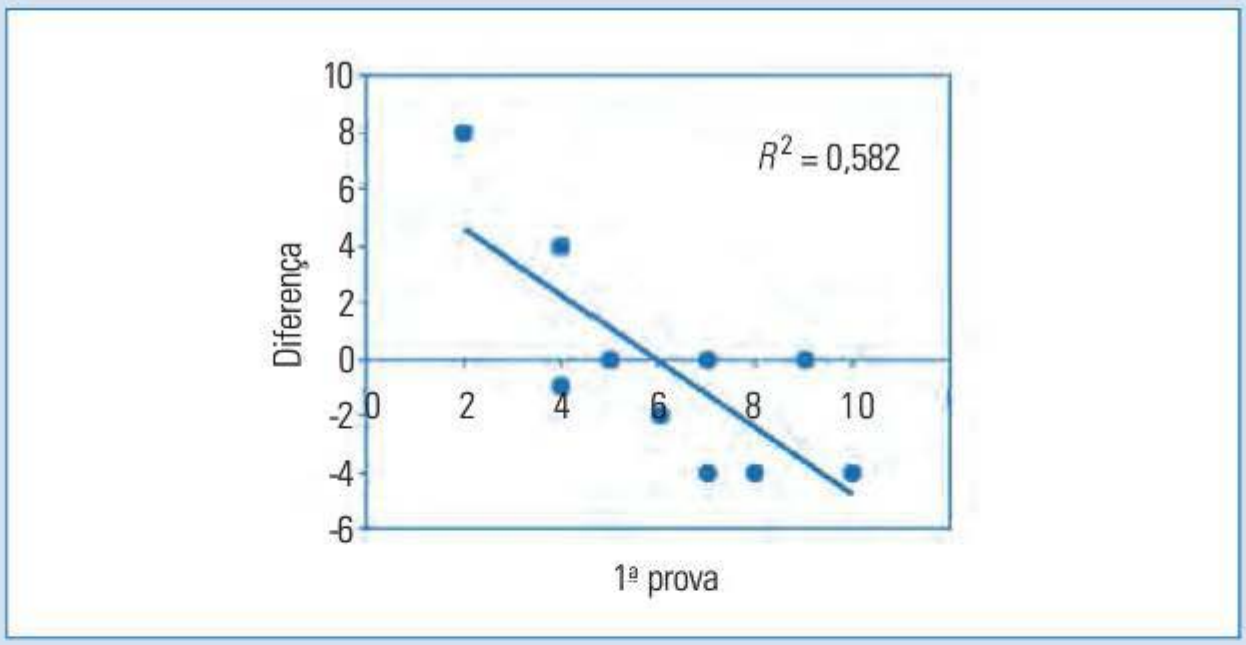


**FIGURA 7.9** Diferença das notas de 10 alunos em duas provas em função da 1ª nota.

## 7.6 – Outros tipos de regressão

Existem situações em que os pares de valores das variáveis *X* e *Y*, apresentados em diagrama de dispersão, não se distribuem em torno de uma reta[5]. Veja o Exemplo 7.11.

### Exemplo 7.11: Uma regressão não-linear.

Observe os dados da Tabela 7.7, apresentados em diagrama de dispersão na Figura 7.10: os pontos estão dispersos em torno de uma curva.

**TABELA 7.7**
**Valores de duas variáveis *X* e *Y*.**

| *X* | *Y* |
|---|---|
| 0 | 4,0 |
| 0,6 | 8,0 |
| 1,2 | 15,0 |
| 1,5 | 22,6 |
| 1,8 | 36,4 |
| 2,1 | 45,3 |
| 2,4 | 60,0 |

[5]No programa EXCEL, você encontra as seguintes opções para ajuste de regressão: linear (que vimos até aqui), logarítmica, polinomial (que não será visto neste livro) potência, exponencial, média móvel (que não será visto neste livro).

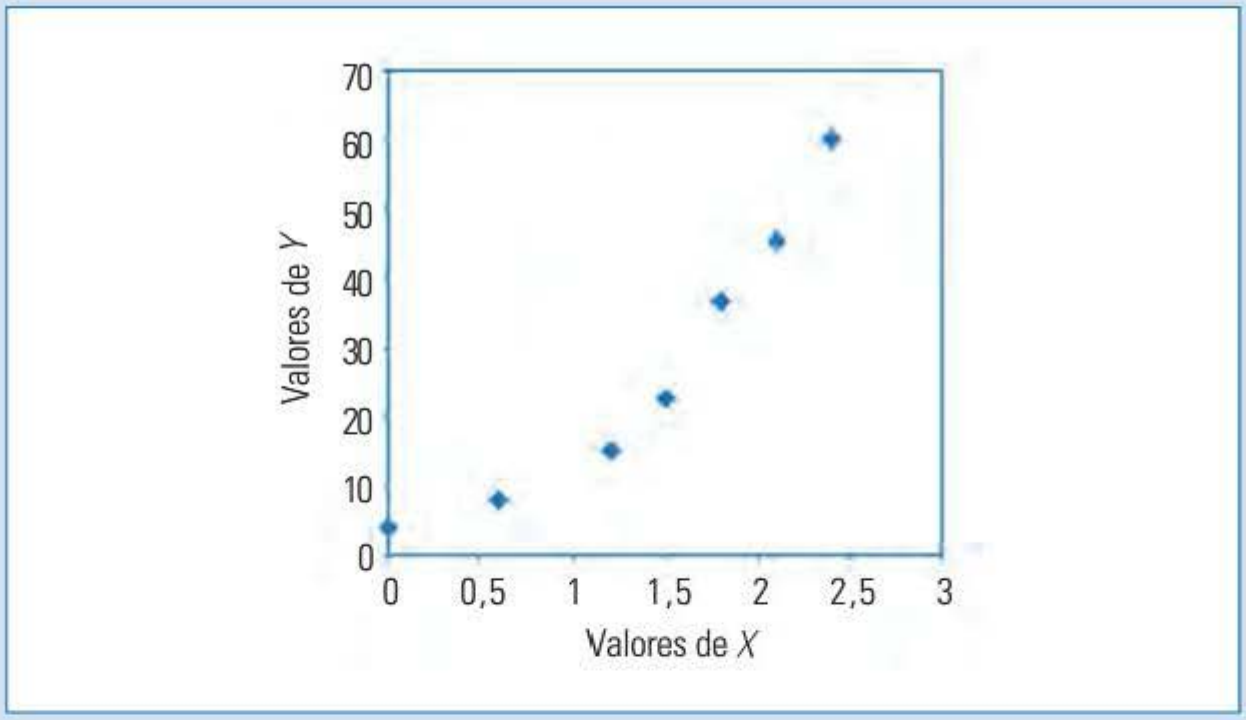


**FIGURA 7.10** Diagrama de dispersão para os valores *X* e *Y* apresentados na Tabela 7.7.

Quando os pontos apresentados em diagrama de dispersão *não* estão em torno de uma reta, devemos experimentar *transformar* a variável *Y*. Por exemplo, podemos experimentar fazer um diagrama de dispersão colocando, em lugar de valores de *Y*, os valores do logaritmo neperiano[6] de *Y*.

Para os dados apresentados no Exemplo 7.11, os valores de *X* e dos logaritmos neperianos de *Y* estão apresentados na Tabela 7.8 e na Figura 7.11.

**TABELA 7.8**
**Valores de *X* e valores dos logaritmos neperianos de *Y*.**

| *X* | ln*Y* |
|---|---|
| 0 | 1,3863 |
| 0,6 | 2,0794 |
| 1,2 | 2,7081 |
| 1,5 | 3,1179 |
| 1,8 | 3,5946 |
| 2,1 | 3,8133 |
| 2,4 | 4,0943 |

[6]No Excel, procure a opção *exponencial*.

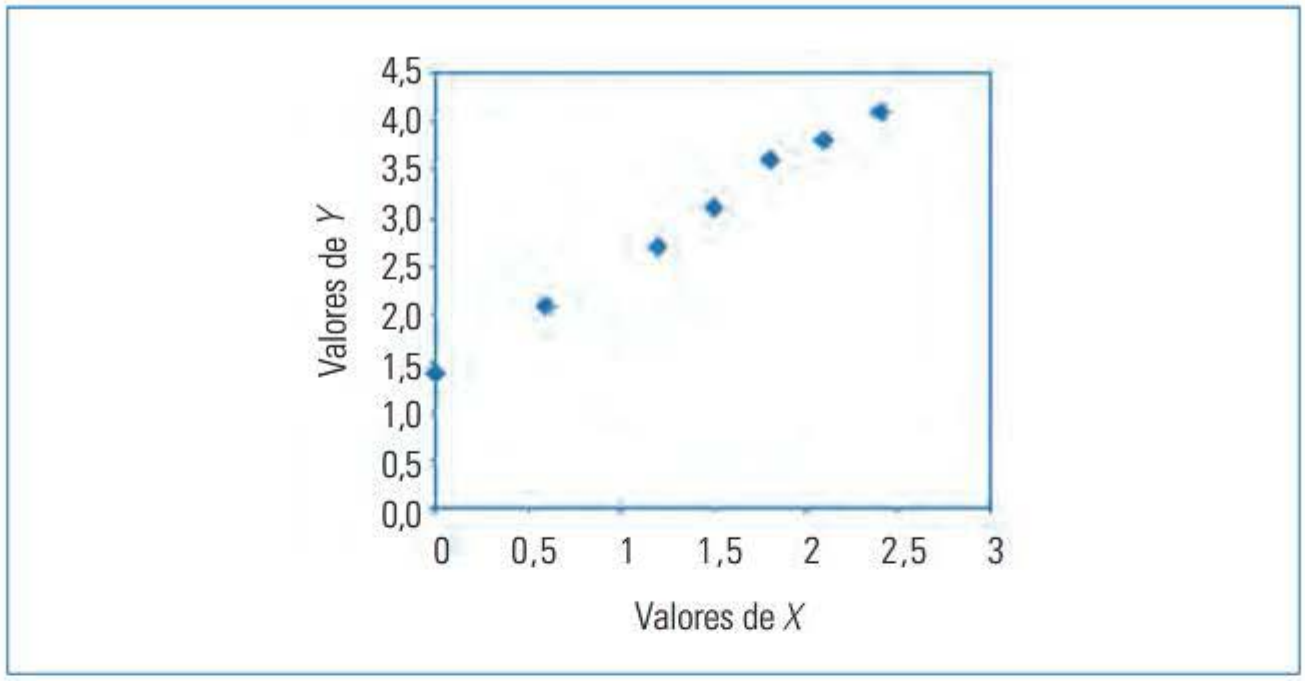


**FIGURA 7.11** Diagrama de dispersão.

O diagrama de dispersão apresentado na Figura 7.11 mostra pontos praticamente sobre uma reta. Então é possível ajustar uma regressão linear de ln$Y$ contra $X$. Para calcular $a$ e $b$, são necessários os cálculos intermediários apresentados na Tabela 7.9.

**TABELA 7.9**
**Cálculos intermediários para a obtenção de *a* e *b*.**

| $X$ | ln$Y$ | $X$ ln$Y$ | $X^2$ |
|---|---|---|---|
| 0 | 1,3863 | 0,0000 | 0 |
| 0,6 | 2,0794 | 1,2477 | 0,36 |
| 1,2 | 2,7081 | 3,2497 | 1,44 |
| 1,5 | 3,1179 | 4,6769 | 2,25 |
| 1,8 | 3,5946 | 6,4702 | 3,24 |
| 2,1 | 3,8133 | 8,0079 | 4,41 |
| 2,4 | 4,0943 | 9,8264 | 5,76 |
| 9,6 | 20,7940 | 33,4788 | 17,46 |

Com base nos cálculos apresentados na Tabela 7.9, é possível obter:

$$b = \frac{33,4788 - \dfrac{9,6 \times 20,7940}{7}}{17,46 - \dfrac{9,6^2}{7}} = 1,1554$$

$$a = \frac{20,7940}{7} - 1,1554 \times \frac{9,6}{7} = 1,3861$$

A equação de reta de regressão de $\ln\hat{Y}$ contra $X$ é:

$$\ln\hat{Y} = 1{,}3861 + 1{,}1554X$$

Se você quiser voltar ao valor da variável Y, é preciso calcular o antilogaritmo da equação. Então, você obtém:

$$\hat{Y} = \text{antiln}(1{,}3861)\, e^{1{,}1554X}$$

ou:

$$\hat{Y} = 3{,}999\, e^{1{,}1554X}$$

Esta equação é chamada de *exponencial* porque traz a variável explanatória no expoente.

Para que uma regressão linear possa ser ajustada aos dados, muitas vezes basta transformar uma das variáveis[7]. Outras vezes, é preciso transformar ambas as variáveis[8]. Também podem ser utilizadas outras transformações, além da *transformação logarítmica*, mostrada aqui. Assim, são também usadas a *extração de raiz quadrada* e a *inversão*, além de outras, mais complicadas.

As transformações são, em geral, *empíricas*, isto é, dados $n$ pares de valores $X$ e $Y$, é preciso fazer várias tentativas até achar a transformação que permita ajustar uma regressão linear aos pares de dados. Algumas vezes, porém, o modelo é *especificado* teoricamente. Por exemplo, a equação de Arrenhius dá a velocidade de uma reação química em função da temperatura em que a reação se processa. Se $T$ é a temperatura em graus Kelvin em que ocorre a reação química, a equação de Arrenhius estabelece que a velocidade $V$ é dada por:

$$\ln V = C - \frac{A}{R} \times \frac{1}{T}$$

em que $\ln V$ é o logaritmo neperiano da velocidade da reação química à temperatura $T$ e $R$ é uma constante (1,987 cal/grau/mol). Para ajustar a equação de Arrenhius aos dados de temperatura e de velocidade de uma reação química, é preciso calcular os valores das variáveis transformadas, isto é, o *logaritmo neperiano da velocidade* e o *inverso da temperatura*. Depois se ajusta uma regressão linear do logaritmo neperiano de $V$ contra o inverso de $T$, isto é:

$$\ln V = a + b\frac{1}{T}$$

Então, $C = a$ e $A = -Rb$.

---

[7]Para ajustar uma regressão *logarítmica*, transforme $X$, isto é, ajuste a regressão dos logaritmos de $X$ contra $Y$. Para ajustar uma regressão *potência*, transforme $X$ e $Y$, isto é, ajuste a regressão dos logaritmos de $X$ contra os logaritmos de $Y$.

[8]Veja mais sobre o assunto em VIEIRA, S. **Bioestatística: tópicos avançados. 2 ed.** Rio de Janeiro: Campus, 2004.

Uma regra, porém, é básica: antes de ajustar uma reta de regressão aos dados, devem-se colocar os pontos (*X; Y*) em um diagrama de dispersão e estudar o conhecimento disponível na literatura sobre o fenômeno. A inspeção dos dados numéricos é obrigatória. Às vezes, é possível ajustar mais de um modelo aos dados e depois escolher, com base nas estatísticas obtidas (coeficientes de determinação etc.), o modelo que melhor se ajusta aos dados.

Neste Capítulo vimos como se ajusta uma *regressão linear simples* aos dados: *linear*, porque é uma reta, e *simples*, porque está no plano, isto é, existe uma só variável dependente e uma só variável explanatória. Mas a variação da variável dependente pode ser posta em função de diversas variáveis, isto é, podem existir diversas variáveis explanatórias. É o caso, por exemplo, da pressão arterial que depende não apenas de peso como mostrado no exemplo, mas da idade, de fatores hereditários, da alimentação etc. Nesses casos, ajusta-se aos dados uma *regressão múltipla*, isto é, uma função com diversas variáveis explanatórias. Mas este tema não será tratado aqui.

## 7.7 – Exercícios resolvidos

### *7.7.1 – Faça um gráfico de linhas para os dados apresentados na Tabela 7.10. Discuta.*

**TABELA 7.10**
**Número de nascidos vivos no Brasil, no período de 1996 a 2005.**

| *Ano* | *Número de nascidos vivos* |
|---|---|
| 1996 | 2.945.425 |
| 1997 | 3.026.658 |
| 1998 | 3.148.037 |
| 1999 | 3.256.433 |
| 2000 | 3.206.761 |
| 2001 | 3.115.474 |
| 2002 | 3.059.402 |
| 2003 | 3.038.251 |
| 2004 | 3.026.548 |
| 2005 | 3.035.096 |

Fonte: DATASUS (2008)[9]

[9]Disponível em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?idb2006/a02.def em 10 de abril de 2008.

Solução

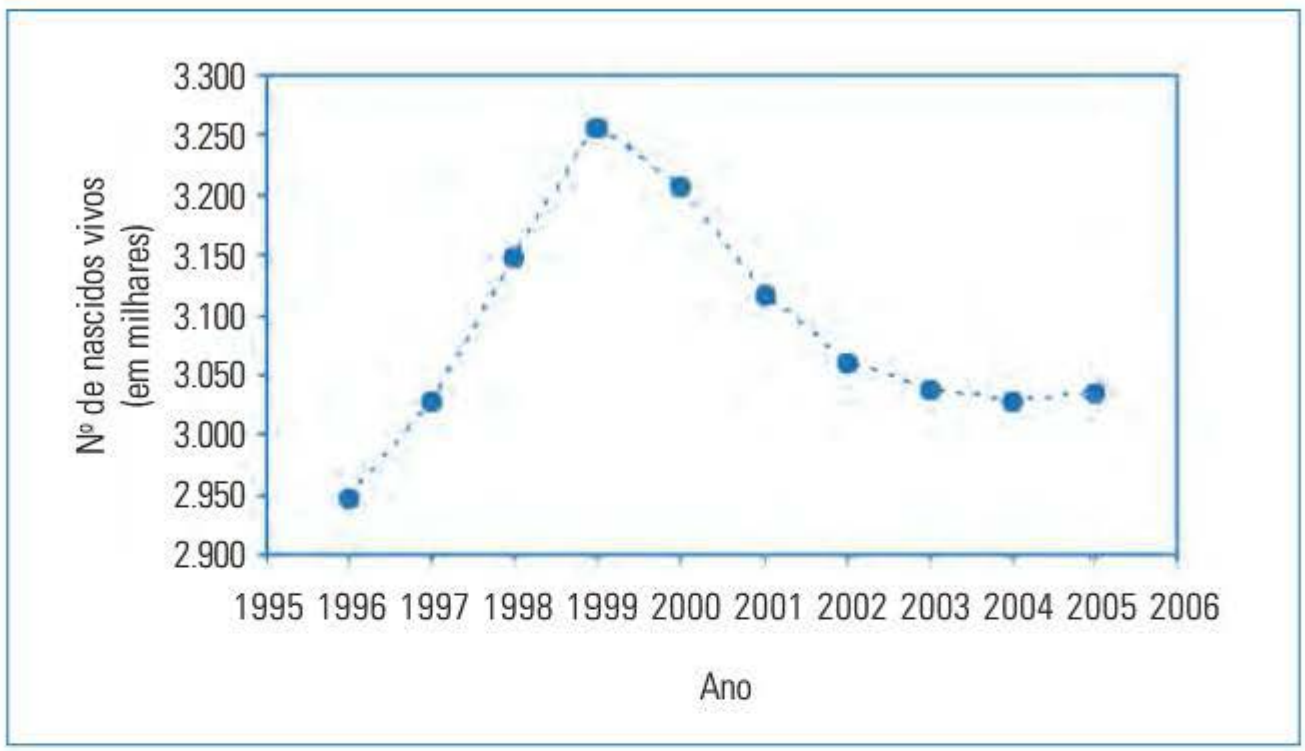


**FIGURA 7.12** Número de nascidos vivos no Brasil, no período de 1996 a 2005.

O número de nascidos vivos no Brasil aumentou até 1999. De lá para 2006, observa-se decréscimo.

***7.7.2 – Faça um gráfico de linhas para os dados apresentados no Exercício 6.5.2 do Capítulo 6, para mostrar como o índice clínico varia em função do peso seco das placas. Discuta.***

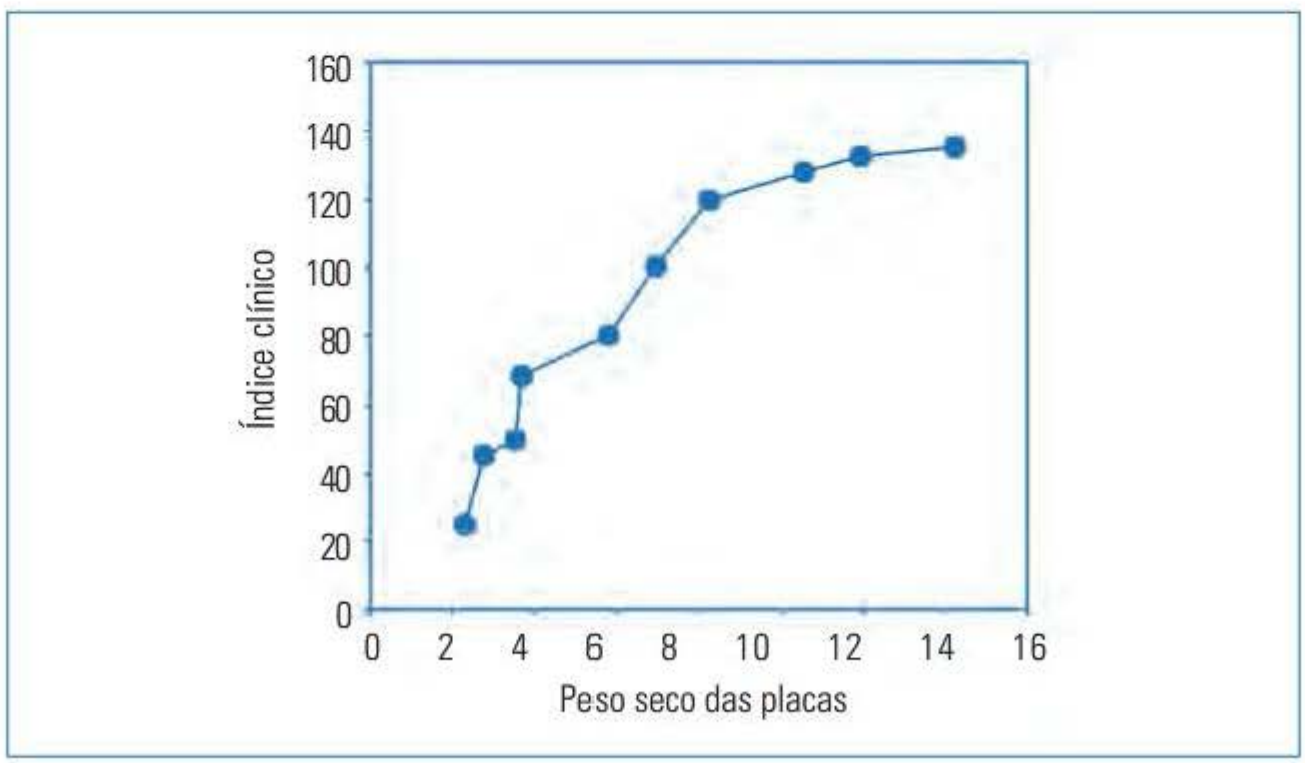


**FIGURA 7.13** Índice clínico em função do peso seco das placas bacterianas.

A Figura 7.13 mostra que o índice clínico usado para medir a quantidade de placa aumenta linearmente (e aceleradamente) com o peso seco das placas, em miligramas, até cerca de 8 mg. Depois, tende a estabilizar. Isto talvez se explique pelo fato de o índice clínico medir a área dos dentes com placas bacterianas, mas não o volume. Ora, o peso leva em conta o volume, que aumenta quando o acúmulo de placas é grande.

***7.7.3 – Ajuste uma reta de regressão aos dados apresentados no Exercício 6.5.3 do Capítulo 6, para estudar peso em função do comprimento dos recém-nascidos. Calcule o coeficiente de determinação.***

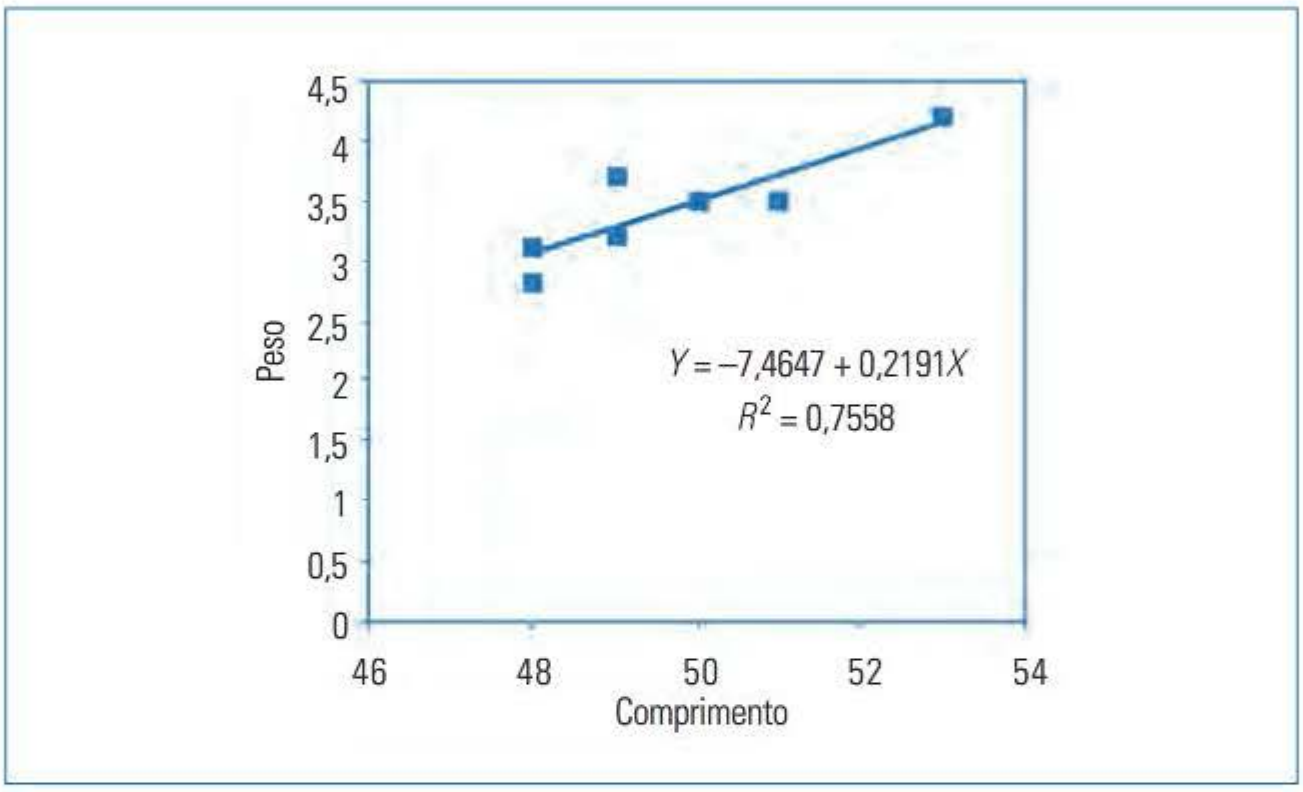


**FIGURA 7.14** Reta de regressão para peso de recém-nascidos em função do comprimento.

***7.7.4 – Ajuste uma reta de regressão aos dados apresentados no Exercício 6.5.4 do Capítulo 6, para estudar peso em função da altura. Calcule o coeficiente de determinação.***

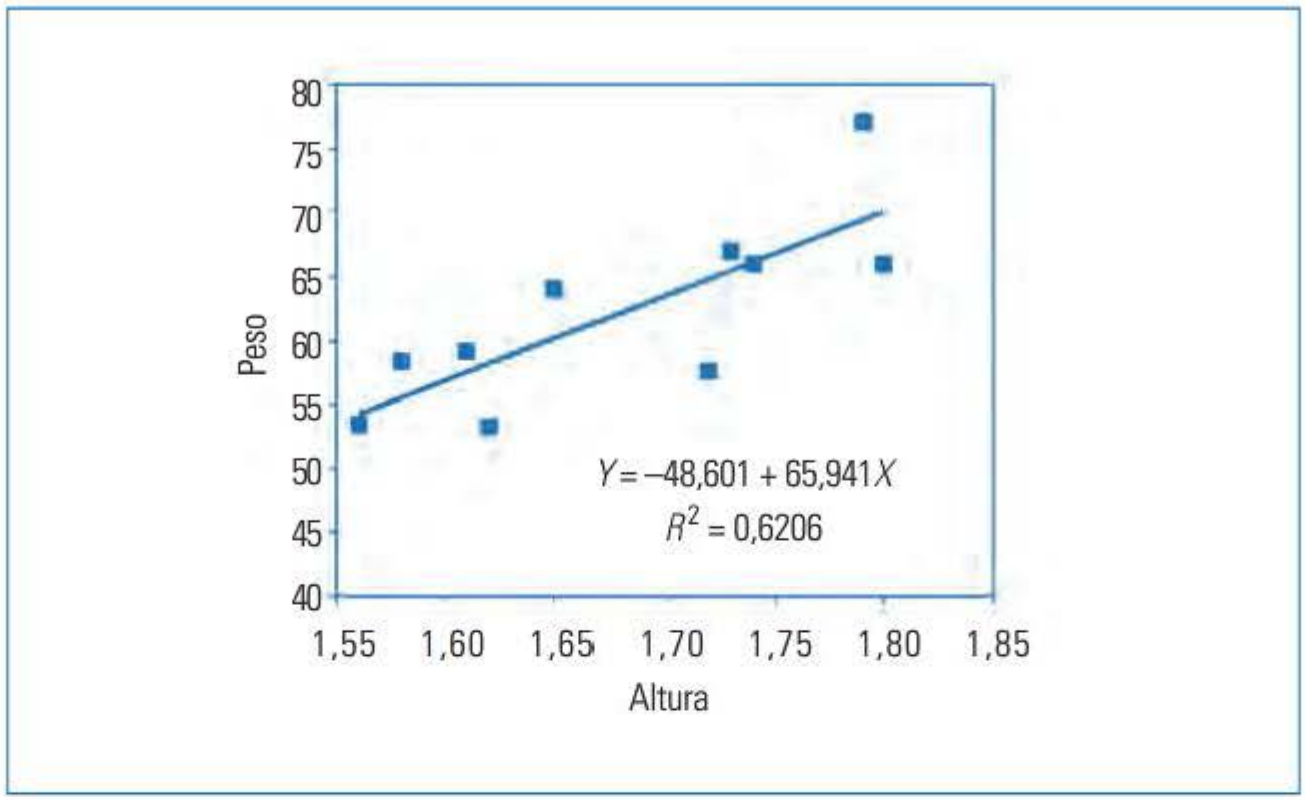


**FIGURA 7.15** Reta de regressão para peso em função da altura.

## 7.8 – Exercícios propostos

***7.8.1 – Faça um gráfico de linhas para os dados apresentados na Tabela 7.11. Discuta.***

**TABELA 7.11**
**Razão de sexos[10] no Brasil, em 2005.**

| *Faixa etária* | *Razão de sexos* |
|---|---|
| Menos de 1 ano | 104,36 |
| De 1 a 4 anos | 103,59 |
| De 5 a 9 anos | 103,49 |
| De 10 a 14 anos | 103,16 |
| De 15 a 19 anos | 102,29 |
| De 20 a 24 anos | 100,05 |
| De 25 a 29 anos | 97,57 |
| De 30 a 34 anos | 95,13 |
| De 35 a 39 anos | 94,41 |
| De 40 a 44 anos | 92,84 |
| De 45 a 49 anos | 92,61 |
| De 50 a 54 anos | 93,63 |
| De 55 a 59 anos | 90,40 |
| De 60 a 64 anos | 87,09 |
| De 65 a 69 anos | 81,49 |
| De 70 a 74 anos | 80,08 |
| De 75 a 79 anos | 77,81 |
| 80 e mais anos | 64,49 |

Fonte: DATASUS[11] (2008)

[10]Razão de sexos: número de homens por 100 mulheres.

[11]Disponível em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?idb2006/a02.def em 10 de abril de 2008.

***7.8.2 – Faça um gráfico de linhas para os dados apresentados na Tabela 7.12. Discuta.***

**TABELA 7.12**
**Coeficiente de mortalidade infantil[12] no Brasil, de 1889 a 1998**

| *Ano* | *Coeficiente de mortalidade infantil* |
|---|---|
| 1989 | 52,02 |
| 1990 | 49,40 |
| 1991 | 46,99 |
| 1992 | 44,79 |
| 1993 | 42,80 |
| 1994 | 41,01 |
| 1995 | 39,40 |
| 1996 | 37,97 |
| 1997 | 36,70 |
| 1998 | 36,10 |

Fonte DATASUS (2008)[13]

***7.8.3 – Ajuste uma reta de regressão aos dados apresentados na Tabela 7.13.***

**TABELA 7.13**
**Teor de vitamina C (mg de ácido ascórbico/100 ml de suco de maçã) em função do período de armazenamento em dias.**

| *Período de armazenamento* | *Teor de vitamina C* |
|---|---|
| 1 | 4,09 |
| 45 | 3,27 |
| 90 | 2,45 |
| 135 | 3,27 |
| 180 | 1,64 |

[12]Taxa ou coeficiente de mortalidade infantil é a razão entre o total de óbitos de menores de 1 ano de idade (excluídos os nascidos mortos) e o total de nascidos vivos, em determinado período de tempo (normalmente 1 ano). Essa razão é multiplicada por 1.000. A taxa de mortalidade infantil estima o risco que um nascido vivo tem de morrer antes de completar 1 ano de idade. A Organização Mundial de Saúde considera *altas* as taxas de 50 por 1.000 ou mais, *médias* as que ficam entre 20 e 49 e *baixas as menores do que* 20.

[13]Disponível em http://tabnet.datasus.gov.br/cgi/mortinf/mibr.htm#topo em 10 de abril de 2008.

***7.8.4 – A reta de regressão será a mesma se você trocar X por Y? O coeficiente de correlação muda?***

***7.8.5 – É preciso que X e Y tenham as mesmas unidades para poder se calcular a reta de regressão?***

***7.8.6 – Se os filhos fossem exatamente 5 cm mais altos do que seus pais, como ficaria a reta de regressão que daria a altura dos filhos em função da altura de seus pais?***

***7.8.7 – Como seria a reta de regressão se todos os pontos de X tivessem o mesmo valor?***

***7.8.8 – Os dados da Tabela 7.14 foram apresentados com a finalidade de mostrar que existe relação entre CPO-D médio (a média de um índice de cáries, ou seja, a média da soma do número de dentes afetados pela cárie em uma amostra de crianças: C = cariados; P = perdidos por cárie; O = obturados, ou seja, restaurados devido ao ataque de cárie) e a média do número de anos de estudo do responsável pelas crianças. O que você acha?***

**TABELA 7.14**
**Número médio de anos de estudo do responsável pelas crianças de uma amostra e CPO-D médio.**

| *Anos de estudo do responsável* | *CPO-D médio* |
|---|---|
| 0 | 1,70 |
| 1 - 4 | 1,85 |
| 5 - 8 | 0,75 |
| 9 - 11 | 0,44 |

***7.8.9 – Uma cadeia de padarias queria saber se a quantidade de dinheiro gasto em propaganda faz aumentar as vendas. Durante seis semanas fez, em ordem aleatória, gastos com propaganda de valores variados conforme mostra a Tabela 7.15 e anotou os valores recebidos nas vendas. Calcule a reta de regressão e coloque em gráfico. O que você acha?***

**TABELA 7.15**
**Gastos com propaganda, em reais, na semana e valores recebidos, em reais, nas vendas.**

| *Gastos* | *Valores recebidos* |
|---|---|
| 100,00 | 1.020,00 |
| 150,00 | 1.610,00 |
| 200,00 | 2.030,00 |
| 250,00 | 2.560,00 |
| 300,00 | 2.800,00 |

***7.8.10 – Com os dados[14] apresentados no Exercício 6.6.14 do Capítulo 6, obtidos de pacientes com enfisema, calcule a reta de regressão.***

***7.8.11 – Com os dados[14] apresentados no Exercício 6.6.15 do Capítulo 6 sobre o volume máximo de oxigênio inalado ($VO_2$máx), você diria que a variável diminui linearmente quando a atividade aumenta? Calcule a reta de regressão.***

***7.8.12 – Os dados[15] apresentados na Tabela 7.16 referem-se à pressão sangüínea diastólica, em milímetros de mercúrio, quando a pessoa está em repouso. Os valores de X indicam o tempo em minutos desde o início do repouso e os valores Y são valores de pressão sangüínea. Desenhe um diagrama de dispersão. Por que não se deve ajustar uma reta de regressão aos dados?***

**TABELA 7.16**
**Tempo em minutos desde o início do repouso e pressão sangüínea diastólica, em milímetros de mercúrio.**

| *Tempo em minutos desde o início do repouso* | *Pressão sangüínea diastólica* |
|---|---|
| 0 | 72 |
| 5 | 66 |
| 10 | 70 |
| 15 | 64 |
| 20 | 66 |

[14]OTT, L e MENDENHALL, W. **Understanding Statistics**. Belmont, Wadsworth, 6 ed. 1994. p. 487.

[15]SCHORK, M. A. e REMINGTON, R. D**. Statistics with applications to the biological and health sciences**. New Jersey, Prentice Hall, 3 ed. 2000. p. 297.

*7.8.13 – Faça um diagrama de dispersão para apresentar os dados da Tabela 7.17. Calcule a reta de regressão. Coloque a reta no gráfico. Quanto devem pesar 10 ratos com 32 dias?*

**TABELA 7.17**
**Idade, em dias, e peso médio, em gramas, de 10 ratos machos da raça Wistar.**

| *Idade* | *Peso médio* |
|---|---|
| 30 | 64 |
| 34 | 74 |
| 38 | 82 |
| 42 | 95 |
| 46 | 106 |

*7.8.14 – Ajuste uma equação exponencial aos dados da Tabela 7.18.*

**TABELA 7.18**
**Dados de *X* e *Y*.**

| *X* | *Y* |
|---|---|
| 28 | 1,25 |
| 32 | 1,25 |
| 35 | 1,75 |
| 38 | 2,25 |
| 39 | 3,25 |
| 41 | 3,25 |
| 42 | 4,25 |

(página deixada intencionalmente em branco)