11 Regressão Linear Simples e Correlação

ESQUEMA DO CAPÍTULO

Objetivos de Aprendizagem

Após estudo cuidadoso deste capítulo você deverá ser capaz de:

1. Utilizar a regressão linear para construir modelos empíricos para dados em engenharia e ciência;
2. Entender como o métodos dos mínimos quadrados é utilizado para estimar os parâmetros em um modelo de regressão linear;
3. Analisar os resíduos para determinar se o modelo de regressão é uma ajuste adequado aos dados ou para ver se alguma hipótese básica foi violada;
4. Testar hipóteses estatísticas e construir intervalos de confiança para os parâmetros dos modelos de regressão;
5. Utilizar o modelo de regressão para fazer predições de observações futuras e construir um intervalo de predição apropriado para a observação futura;
6. Aplicar o modelo de correlação;
7. Utilizar transformações simples para obter um modelo de regressão linear.

11.1 Modelos Empíricos

- Muitos problemas em engenharia e ciência envolvem a exploração da natureza da relação entre duas ou mais variáveis.
- ***A análise de regressão*** é uma técnica estatística que é bastante útil para tais tipos de problemas.
- Por exemplo, em um processo químico, suponha que o rendimento do produto esteja relacionado à temperatura de operação do processo.
- A análise de regressão pode ser utilizada para construir um modelo para ***prever*** o rendimento em um dado nível de temperatura ou também pode ser utilizada para ***otimização*** de processos, tal como encontrar o nível de temperatura que maximiza o rendimento.

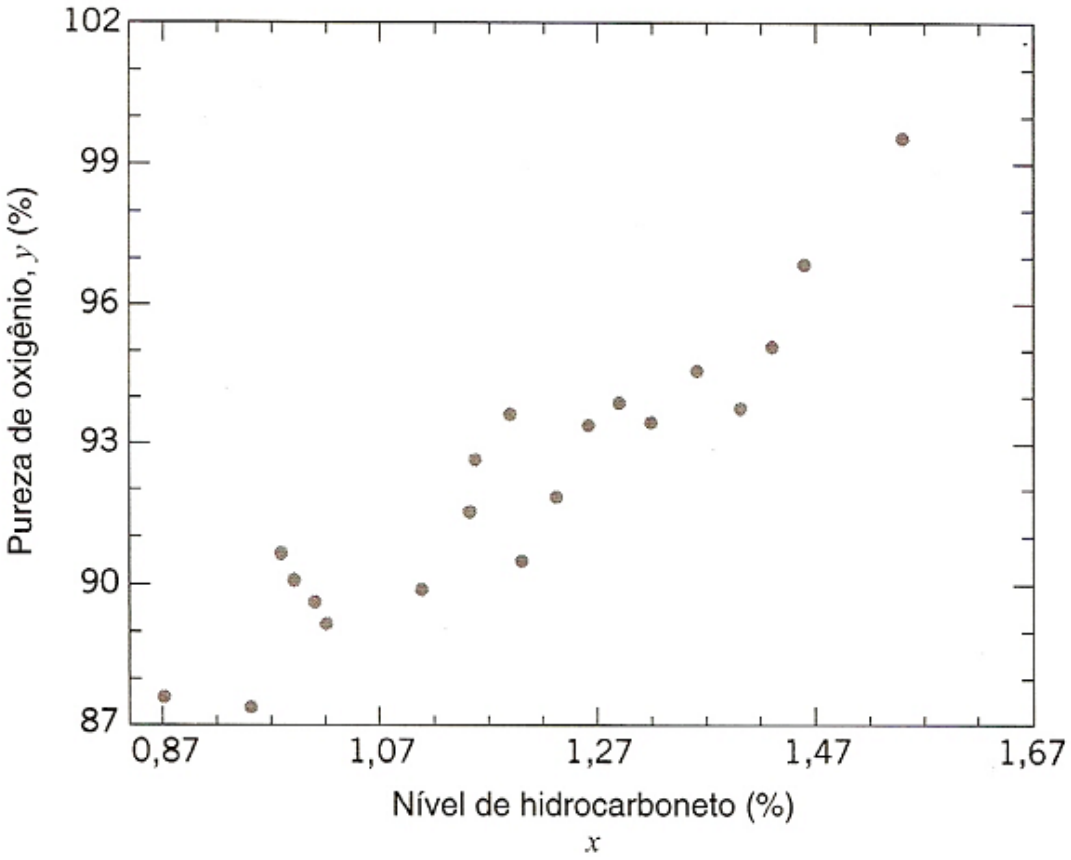
11.1 Modelos Empíricos

Tabela. 11.1 Níveis de Oxigênio e de Hidrocarbonetos.

Número da Observação	Nível de Hidrocarboneto x (%)	Pureza y (%)
1	0,99	90,01
2	1,02	89,05
3	1,15	91,43
4	1,29	93,74
5	1,46	96,73
6	1,36	94,45
7	0,87	87,59
8	1,23	91,77
9	1,55	99,42
10	1,40	93,65
11	1,19	93,54
12	1,15	92,52
13	0,98	90,56
14	1,01	89,54
15	1,11	89,85
16	1,20	90,39
17	1,26	93,25
18	1,32	93,41
19	1,43	94,98
20	0,95	87,33

11.1 Modelos Empíricos

Fig. 11.1 Diagrama de dispersão da pureza de oxigênio *versus* nível de hidrocarbonetos da Tab. 10.1.



11.1 Modelos Empíricos

- Baseado no diagrama de dispersão, é provavelmente razoável assumir que a média da variável aleatória Y esteja relacionada a x pela seguinte relação linear

$$E(Y|\, x) = \mu_{Y|x} = \beta_0 + \beta_1 x,$$

em que a inclinação e a interseção da linha são chamados ***coeficientes de regressão***.

- O ***modelo de regressão linear simples*** (porque possui apenas uma variável independente ou ***regressor***) é dado por

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

sendo ε o termo de erro aleatório.

11.1 Modelos Empíricos

- Pensaremos no modelo de regressão como um ***modelo empírico***.
- Suponha que a média e a variância de ε sejam 0 e σ^2, respectivamente. Então

$$\begin{aligned} E(Y|\, x) &= E(\beta_0 + \beta_1 x + \varepsilon) \\ &= \beta_0 + \beta_1 x + E(\varepsilon) \\ &= \beta_0 + \beta_1 x. \end{aligned}$$

A variância de Y, dado x, é

$$\begin{aligned} V(Y|\, x) &= V(\beta_0 + \beta_1 x + \varepsilon) \\ &= V(\beta_0 + \beta_1 x) + V(\varepsilon) = 0 + \sigma^2. \end{aligned}$$

11.1 Modelos Empíricos

- O modelo verdadeiro de regressão é uma linha de valores médios

  $$\mu_{Y|x} = \beta_0 + \beta_1 x,$$

  em que a inclinação β_1 pode ser interpretada como a mudança na média de Y, para um mudança unitária em *x*.
- Também, a variabilidade de Y, em um valor particular de *x*, é determinada pela variância do erro, σ^2.
- Isso implica que há uma distribuição de valores de Y em cada *x* e que a variância dessa distribuição é a mesma em cada *x*.

11.1 Modelos Empíricos

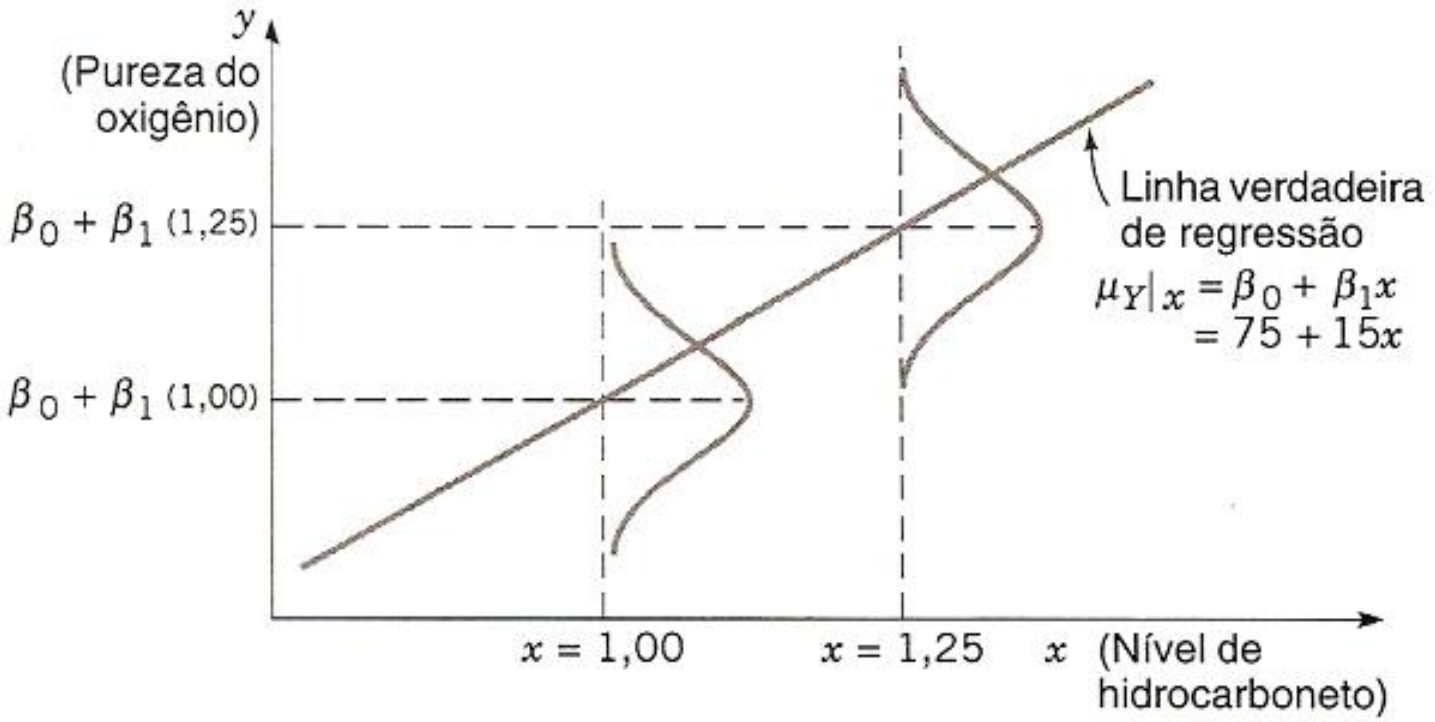


Fig. 11.2 A distribuição de Y para um certo valor de *x*, para os dados da pureza do oxigênio-hidrocarboneto.

11.1 Modelos Empíricos

- **Abusos da Regressão**

  Lembre-se que:

  – uma forte associação observada entre as variáveis ***não*** implica necessariamente que existe relação causal entre aquelas variáveis;

  – o planejamento de experimentos é a ***única*** forma de determinar relações causais;

  – relações de regressão são válidas somente para valores do regressor ***dentro*** da faixa dos dados originais;

  – em outras palavras, modelos de regressão ***não*** são necessariamente válidos para finalidade de extrapolação.

11.2 Regressão Linear Simples

- O caso de ***regressão linear simples*** considera um único ***regressor*** (ou ***preditor***) x e uma variável ***dependente*** (ou variável ***resposta***) Y.
- O valor esperado de Y para cada valor x é

$$E(Y|\, x) = \beta_0 + \beta_1 x.$$

- Consideramos que cada observação, Y, possa ser descrita pelo modelo

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

em que ε é o erro aleatório com média zero e variância σ^2.

11.2 Regressão Linear Simples

- Suponha que tenhamos n pares de observações (x_1, y_1), ..., (x_n, y_n), como visto na Fig. 11.3.

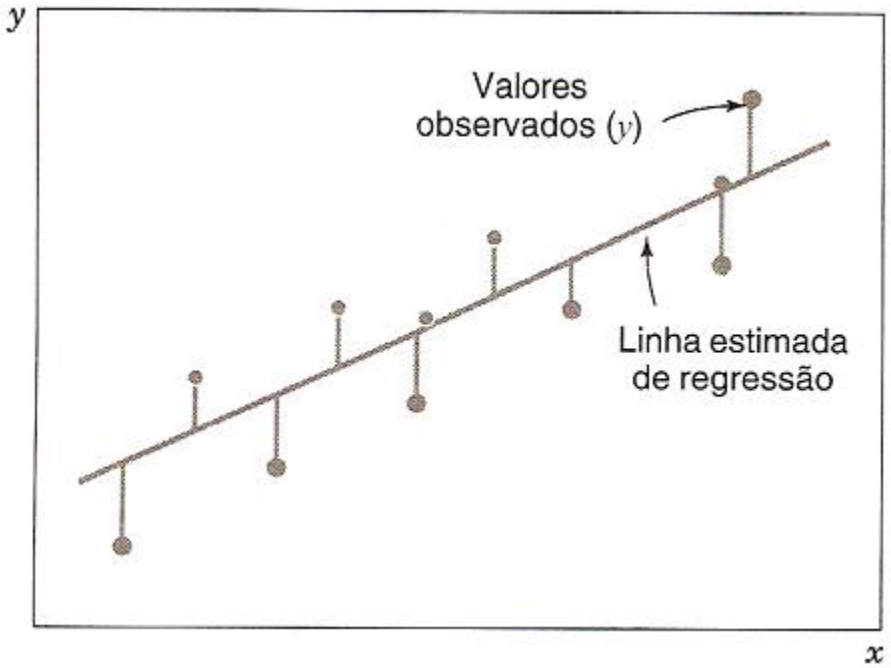


Fig. 11.3 Desvios dos dados em relação ao modelo estimado de regressão.

11.2 Regressão Linear Simples

- O método dos mínimos quadrados é utilizado para estimar os parâmetros β_0 e β_1, pela minimização da soma dos quadrados dos desvios verticais na Fig. 11.3.

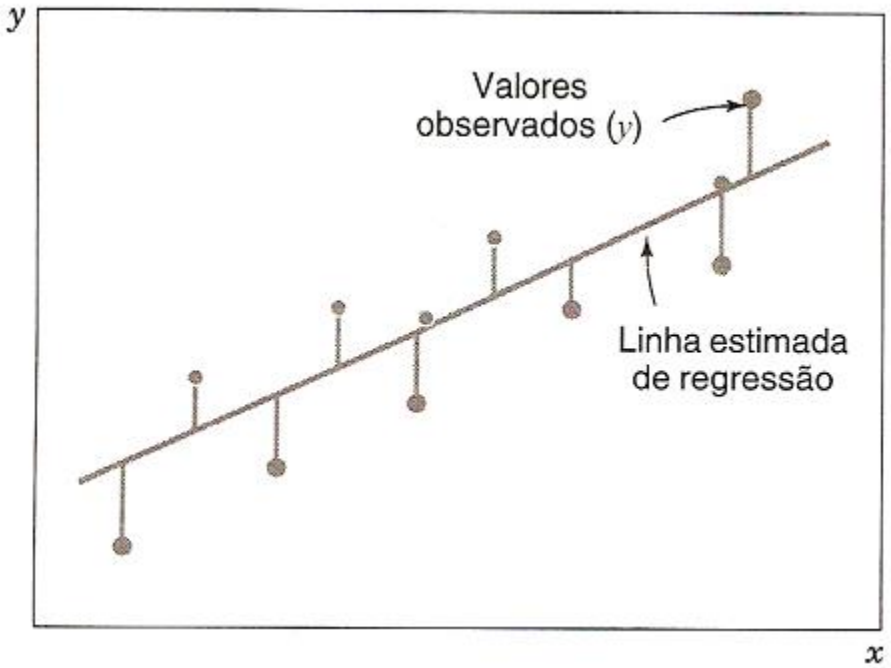


Fig. 11.3 Desvios dos dados em relação ao modelo estimado de regressão.

11.2 Regressão Linear Simples

- Utilizando a equação

  $$Y = \beta_0 + \beta_1 x + \varepsilon,$$

  as n observações na amostra podem ser expressas como

  $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, ..., n.$$

- A soma dos quadrados dos desvios das observações em relação à linha de regressão é dada por

  $$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

11.2 Regressão Linear Simples

- Os estimadores de mínimos quadrados de β_0 e β_1, quais sejam $\hat{\beta}_0$ e $\hat{\beta}_1$,devem satisfazer a

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

11.2 Regressão Linear Simples

- Simplificando essas duas equações, resulta em

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Essas equações são denominadas de equações dos mínimos quadrados. As soluções para as equações normais resulta nos estimadores de mínimos quadrados $\hat{\beta}_0$ e $\hat{\beta}_1$.

11.2 Regressão Linear Simples

- **Definição:**

As **estimativas de mínimos quadrados** da interseção e da inclinação no modelo de regressão linear simples são

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (10.7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad (10.8)$$

em que $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ e $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

11.2 Regressão Linear Simples

- A ***linha estimada de regressão*** ou ***linha ajustada de regressão*** é consequentemente

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Note que cada par de observação satisfaz a relação

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, \ldots, n,$$

sendo que

$$e_i = y_i - \hat{y}_i$$

é chamado de ***resíduo***. O resíduo descreve o erro no ajuste do modelo para a i-ésima observação y_i. Mais adiante, usaremos os resíduos para fornecer informação sobre a ***adequação*** do modelo ajustado.

11.2 Regressão Linear Simples

- Em termos de notação, é conveniente dar símbolos especiais aos numerador e ao denominador da Eq. (10.8). Tendo os dados (x_1, y_1), ..., (x_n, y_n), faça

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}$$

e

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i \right)\left(\sum_{i=1}^{n} y_i \right)}{n}.$$

11.2 Regressão Linear Simples

- **Exemplo 11.1:**

Ajustaremos um modelo de regressão linear simples aos dados de pureza de oxigênio na Tabela 11.1. As seguintes quantidades podem ser computadas:

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23{,}92 \quad \sum_{i=1}^{20} y_i = 1.843{,}21$$

$$\bar{x} = 1{,}20 \quad \bar{y} = 92{,}16$$

$$\sum_{i=1}^{20} y_i^2 = 170.044{,}53 \quad \sum_{i=1}^{20} x_i^2 = 29{,}29 \quad \sum_{i=1}^{20} x_i y_i = 2.214{,}66$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29{,}29 - \frac{(23{,}92)^2}{20} = 0{,}68$$

11.2 Regressão Linear Simples

- **Exemplo 11.1 (cont.):**

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20}$$

$$= 2.214{,}66 - \frac{(23{,}92)(1.843{,}21)}{20} = 10{,}18$$

Logo, as estimativas de mínimos quadrados da inclinação e da interseção são

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10{,}18}{0{,}68} = 14{,}97$$

e

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 92{,}16 - (14{,}97)1{,}20 = 74{,}20$$

11.2 Regressão Linear Simples

- **Exemplo 11.1 (cont.):**

O modelo ajustado da regressão linear simples é

$$\hat{y} = 74{,}20 + 14{,}97x$$

Esse modelo é plotado na Fig. 11.4, juntamente com os dados da amostra. Temos considerado duas casas decimais no cálculo desses coeficientes de regressão.

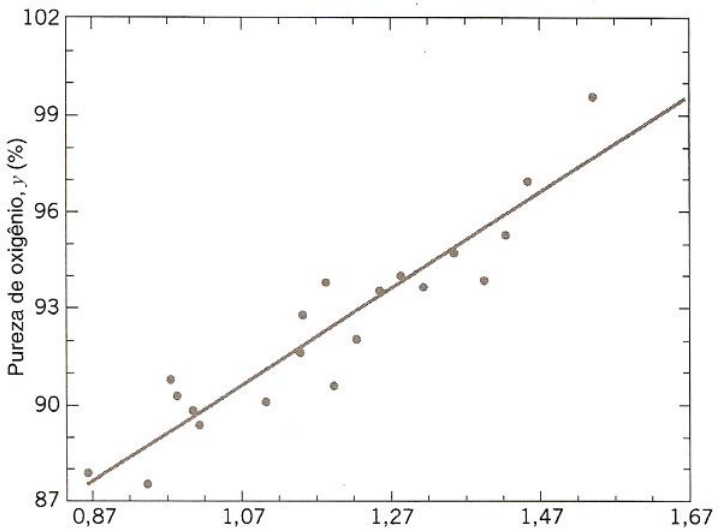


Fig. 11.4 Diagrama de dispersão da pureza do oxigênio, *y*, *versus* o nível de hidrocarbonetos, *x*, e o modelo de regressão $\hat{y} = 74{,}20 + 14{,}97x$.

11.2 Regressão Linear Simples

- **Exemplo 11.1 (final):**

  Programas computacionais são largamente empregados nos modelos de regressão.

  A Tabela 11.2 mostra parte de uma saída do Minitab® para esse problema.

  Nas próximas seções, daremos explicações para as outras informações disponibilizadas nessa saída do computador.

Tabela 11.2 Saída do Minitab® para os dados de pureza do oxigênio.

Análise de Regressão

A equação de regressão é

$y = 74{,}3 + 14{,}9x$

Preditor	Coeficiente	Desvio-padrão	T	P
Constante	74,283	1,593	46,62	0,000
x	14,947	1,317	11,35	0,000

S = 1,087 R^2 = 87,7% R^2 (ajustado) = 87,1%

Análise de Variância

Fonte	DF	SQ	MQ	F	P
Regressão	1	152,13	152,13	128,86	0,000
Erro	18	21,25	1,18		
Total	19	173,38			

11.3 Propriedades dos Estimadores de Mínimos Quadrados

- **Propriedades da inclinação:**

$$E(\hat{\beta}_1) = \beta_1$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

- **Propriedades do intercepto:**

$$E(\hat{\beta}_0) = \beta_0,$$

$$V(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right].$$

11.3 Propriedades dos Estimadores de Mínimos Quadrados

- A soma dos erros quadráticos é

$$SQ_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Pode ser mostrado que o valor esperado da soma dos erros quadráticos é

$$E(SQ_E) = (n - 2)\sigma^2.$$

- Por conseguinte, um estimador não-tendencioso de σ^2 é

$$\hat{\sigma}^2 = \frac{SQ_E}{n-2}, \qquad (10.17)$$

em que o SQ_E pode ser facilmente calculado como

$$SQ_E = SQ_T - \hat{\beta}_1 S_{xy}, \text{ com } SQ_T = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

11.3 Propriedades dos Estimadores de Mínimos Quadrados

- **Definição:**

Em uma regressão linear simples, o **erro-padrão estimado da inclinação** é

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

e o **erro-padrão estimado da interseção** é

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

em que σ^2 é calculada a partir da Eq. 10.17.

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.1. Uso de Testes t

- Suponha que desejamos testar a hipótese de a inclinação ser igual a uma constante, como $\beta_{1,0}$.

  As hipóteses apropriadas são

  $H_0: \beta_1 = \beta_{1,0}$

  $H_1: \beta_1 \neq \beta_{1,0}$

- Uma estatística apropriada seria

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}.$$

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.1. Uso de Testes t

- A estatística de teste poderia também ser escrita como

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}.$$

- Rejeitaríamos a hipótese nula se

$$|t_0| > t_{\alpha/2,n-2}.$$

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.1. Uso de Testes t

- Similarmente, podemos testar a hipótese sobre a interseção ser igual a uma constante, digamos, $\beta_{0,0}$. Para testar

  $H_0: \beta_0 = \beta_{0,0}$

  $H_1: \beta_0 \neq \beta_{0,0}$

  usaremos a estatística

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}.$$

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.1. Uso de Testes t

- Rejeitaríamos a hipótese nula se

$$|t_0| > t_{\alpha/2,n-2}.$$

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.1. Uso de Testes t

- Um caso especial importante das hipóteses da inclinação é

  $H_0: \beta_1 = 0$

  $H_1: \beta_1 \neq 0$

  Estas hipóteses se relacionam à ***significância da regressão***.
- Falhar em rejeitar H_0 é equivalente a concluir que não há relação linear entre *x* e Y, situação ilustrada na Fig. 11.5.

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.1. Uso de Testes t

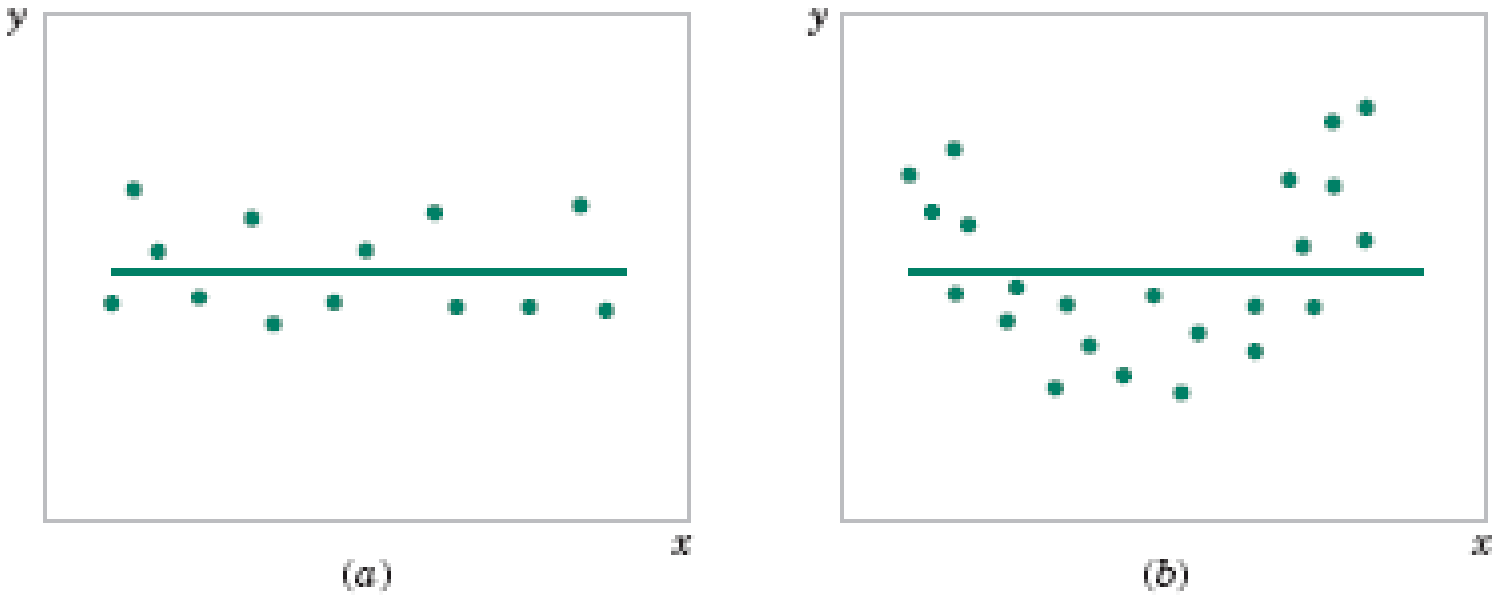


Fig. 11.5 A hipótese H_0: $\beta_1 = 0$ não é rejeitada.

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.1. Uso de Testes t

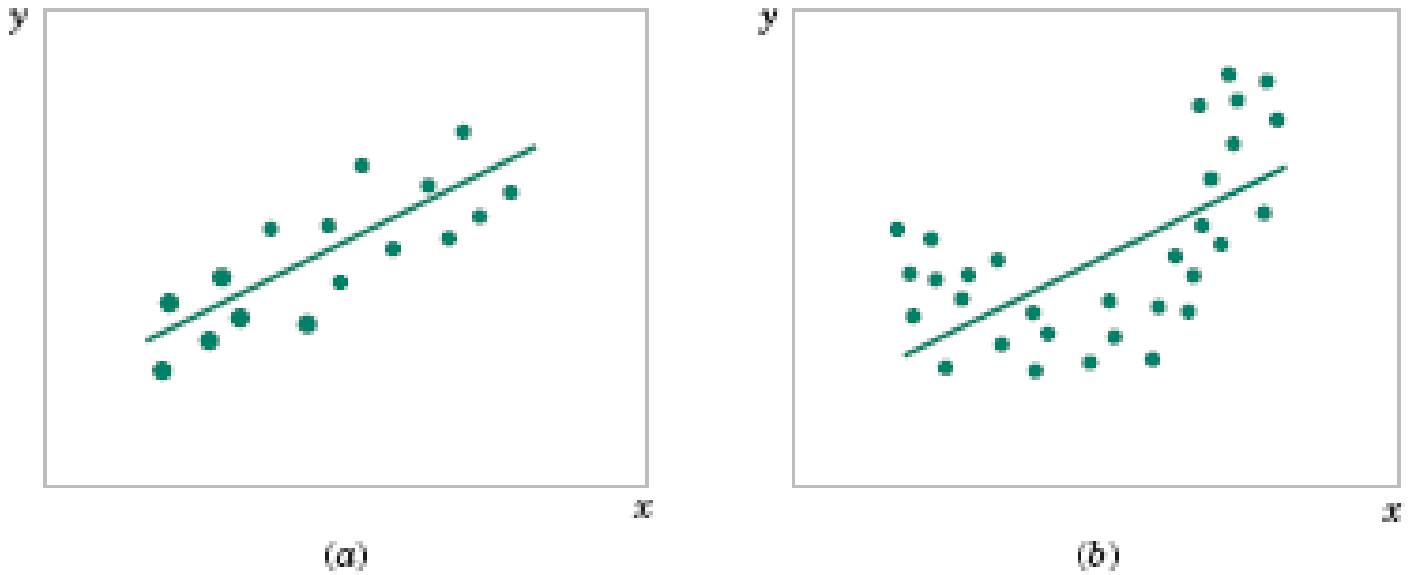


Fig. 11.6 A hipótese H_0: $\beta_1 = 0$ é rejeitada.

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.1. Uso de Testes t

- **Exemplo 11.2:**

Testaremos a significância da regressão usando o modelo para os dados de pureza do oxigênio do Exemplo 11.1. As hipóteses são

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

e usaremos $\alpha = 0{,}01$. Do Exemplo 11.1, temos

$$\hat{\beta}_1 = 14{,}97 \quad n = 20, \quad S_{xx} = 0{,}68, \quad \hat{\sigma}^2 = 1{,}17$$

logo, a estatística t na Eq. 10.20 se torna

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{14{,}97}{\sqrt{1{,}17/0{,}68}} = 11{,}41$$

Já que o valor de referência de t é $t_{0{,}005;18} = 2{,}88$, o valor da estatística de teste está bem inserido na região crítica, implicando que $H_0: \beta_1 = 0$ deve ser rejeitada. O valor P para esse teste é $P \simeq 1{,}13 \times 10^{-9}$. Ele foi obtido manualmente com uma calculadora.

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.1. Uso de Testes t

- **Exemplo 11.2 (final):**

  O resultado anterior estava disponível na saída do Minitab®, reproduzida na Tabela 11.2

Tabela 11.2 Saída do Minitab® para os dados de pureza do oxigênio.

Análise de Regressão

A equação de regressão é

$y = 74{,}3 + 14{,}9x$

Preditor	Coeficiente	Desvio-padrão	T	P
Constante	74,283	1,593	46,62	0,000
x	14,947	1,317	11,35	0,000

S = 1,087 $R^2 = 87{,}7\%$ R^2 (ajustado) = 87,1%

Análise de Variância

Fonte	DF	SQ	MQ	F	P
Regressão	1	152,13	152,13	128,86	0,000
Erro	18	21,25	1,18		
Total	19	173,38			

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.2. Abordagem de Análise de Variância para Testar Significância de Regressão

- A ***equação básica da análise de variância*** é dada a seguir

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (10.25)$$

- Simbolicamente

$$SQ_T = SQ_R + SQ_E \quad (10.26)$$

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.2. Abordagem de Análise de Variância para Testar Significância de Regressão

- Se a hipótese nula, H_0: $\beta_1 = 0$, for verdadeira, a estatística

$$F_0 = \frac{SQ_R/1}{SQ_E/(n-2)} = \frac{MQ_R}{MQ_E} \qquad (10.28)$$

segue uma distribuição $F_{1,n\text{-}2}$ e rejeitamos H_0 se acontece que $f_0 > f_{\alpha;\,1;\,n\text{-}2}$.

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.2. Abordagem de Análise de Variância para Testar Significância de Regressão

- As quantidades $MQ_R = SQ_R/1$ e $MQ_E = SQ_E/(n-2)$ são denominadas ***médias quadráticas***.
- O procedimento de teste é geralmente arranjado em uma ***tabela de análise de variância***

Tabela 11.3 Análise de Variância para Testar a Significância da Regressão.

Fonte de Variação	Soma Quadrática*	Graus de Liberdade	Média Quadrática*	F_0
Regressão	$SQ_R = \hat{\beta}_1 S_{xy}$	1	MQ_R	MQ_R/MQ_E
Erro	$SQ_E = SQ_T - \hat{\beta}_1 S_{xy}$	$n-2$	MQ_E	
Total	SQ_T	$n-1$		

Note que $MQ_E = \hat{\sigma}^2$.
*Em inglês, soma quadrática e média quadrática são abreviadas por *SS* e *MS*, respectivamente. (*N.T.*)

11.4 Testes de Hipóteses na Regressão Linear Simples

11.4.2. Abordagem de Análise de Variância para Testar Significância de Regressão

- **Exemplo 11.3:** Usaremos a abordagem de análise de variância para testar a significância da regressão usando os dados de pureza do oxigênio do Exemplo 11.1. Lembre-se de que $SQ_T = 173{,}37$, $\hat{\beta}_1 = 14{,}97$, $S_{xy} = 10{,}18$ e $n = 20$. A soma quadrática devido à regressão é

$$SQ_R = \hat{\beta}_1 S_{xy} = (14{,}97)10{,}18 = 152{,}39$$

e a soma quadrática devido ao erro é

$$\begin{aligned} SQ_E &= SQ_T - SQ_R \\ &= 173{,}37 - 152{,}39 \\ &= 20{,}98 \end{aligned}$$

A análise de variância para testar H_0: $\beta_1 = 0$ está resumida na Tabela 11.2. A estatística de teste $f_0 = MQ_R/MQ_E = 152{,}39/1{,}17 = 130{,}25$, para a qual encontramos o valor P como sendo $P \simeq 1{,}13 \times 10^{-9}$; logo, concluímos que β_1 não é zero.

11.5 Intervalos de Confiança

11.5.1. Intervalos de Confiança para a Inclinação e Interseção

- **Definição:**

Sob a suposição de que as observações sejam normal e independentemente distribuídas, um **intervalo de confiança** de $100(1-\alpha)\%$ **para a inclinação** β_1 na regressão linear simples é

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad (10.31)$$

Similarmente, um **intervalo de confiança** de $100(1-\alpha)\%$ **para a interseção** β_0 na regressão linear simples é

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \quad (10.32)$$

11.5 Intervalos de Confiança

11.5.1. Intervalos de Confiança para a Inclinação e Interseção

- **Exemplo 11.4:**

Encontraremos um intervalo de confiança de 95% para a inclinação da linha de regressão, usando os dados no Exemplo 11.1. Lembre-se de que $\hat{\beta}_1 = 14,97$, $S_{xx} = 0,68$ e $\hat{\sigma}^2 = 1,17$ (ver Tabela 10.2). Então, da Eq. 10.31, encontramos

$$\hat{\beta}_1 - t_{0,025;18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{0,025;18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

ou

$$14,97 - 2,101\sqrt{\frac{1,17}{0,68}} \leq \beta_1 \leq 14,97 + 2,101\sqrt{\frac{1,17}{0,68}}$$

Isso simplifica para

$$12,21 \leq \beta_1 \leq 17,73$$

11.5 Intervalos de Confiança

11.5.2. Intervalos de Confiança para a Resposta Média

- **Definição:**

Um **intervalo de confiança** de $100(1-\alpha)\%$ **em torno da resposta média** no valor de $x = x_0$, como $\mu_{Y|x_0}$ é dado por

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

$$\leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} \quad (10.33)$$

sendo $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ calculado a partir do modelo ajustado de regressão.

11.5 Intervalos de Confiança

11.5.2. Intervalos de Confiança para a Resposta Média

- **Exemplo 11.5:** Para os dados do Exemplo 11.1, construiremos um intervalo de confiança de 95% em torno da resposta média. O modelo ajustado é $\hat{\mu}_{Y|x_0} = 74{,}20 + 14{,}97x_0$ e o intervalo de confiança de 95% para $\mu_{Y|x_0}$ é encontrado da Eq. 10.33 como

$$\hat{\mu}_{Y|x_0} \pm 2{,}101 \sqrt{1{,}17\left[\frac{1}{20} + \frac{(x_0 - 1{,}20)^2}{0{,}68}\right]}$$

Suponha que estejamos interessados em prever a pureza média do oxigênio quando $x_0 = 1{,}00\%$. Então

$$\hat{\mu}_{Y|x_0} = 74{,}20 + 14{,}97(1{,}00) = 89{,}17$$

11.5 Intervalos de Confiança

11.5.2. Intervalos de Confiança para a Resposta Média

- **Exemplo 11.5 (cont.):**

e o intervalo de confiança de 95% é

$$\left[89{,}17 \pm 2{,}101 \sqrt{1{,}17\left[\frac{1}{20} + \frac{(1{,}00 - 1{,}20)^2}{0{,}68}\right]}\right]$$

ou

$$89{,}17 \pm 0{,}75$$

Por conseguinte, o intervalo de confiança de 95% para $\mu_{Y|1{,}00}$ é

$$88{,}42 \leq \mu_{Y|1{,}00} \leq 89{,}92$$

11.5 Intervalos de Confiança

11.5.2. Intervalos de Confiança para a Resposta Média

- **Exemplo 11.5 (cont.):**

Repetindo esses cálculos para vários valores de x_0, podemos obter limites de confiança para cada valor correspondente de $\mu_{Y|x_0}$. A Fig. 11.7 apresenta o diagrama de dispersão com o modelo ajustado e os correspondentes limites de confiança de 95%, plotados como linhas inferior e superior. O nível de confiança de 95% se aplica apenas ao intervalo obtido de um valor de x e não ao conjunto inteiro de valores de x. Note que a largura do intervalo de confiança para $\mu_{Y|x_0}$ aumenta à medida que $|x_0 - \bar{x}|$ aumenta.

11.5 Intervalos de Confiança

11.5.2. Intervalos de Confiança para a Resposta Média

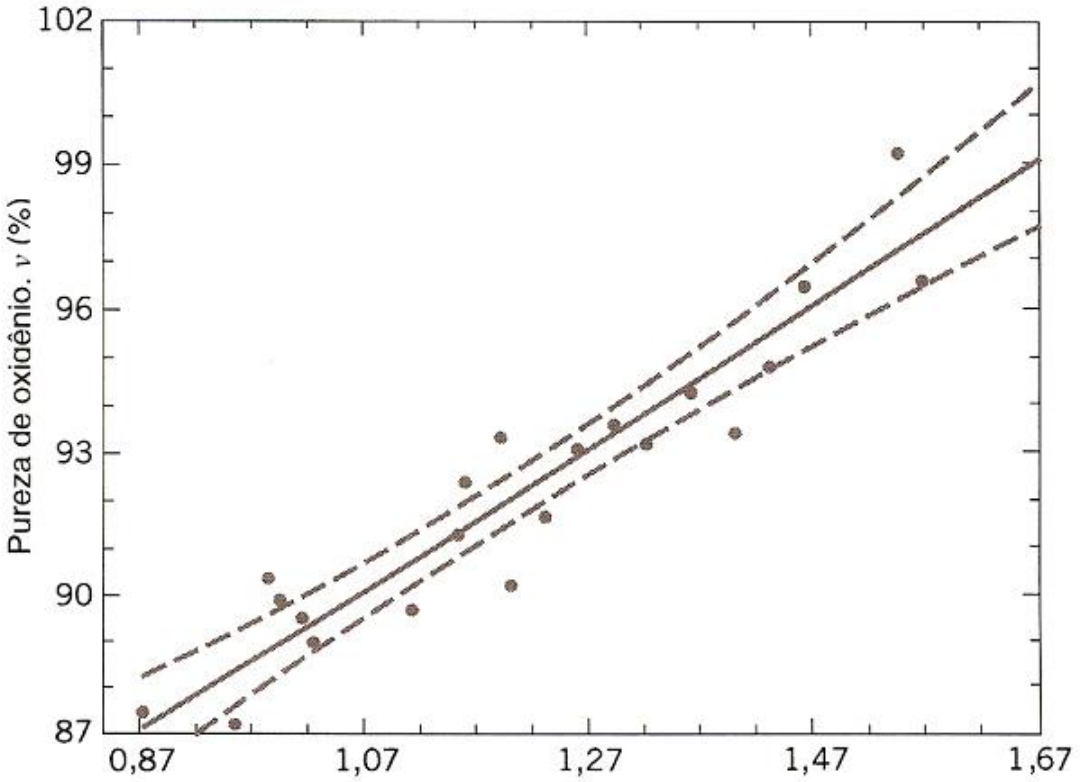


Fig. 11.7 Diagrama de dispersão dos dados de pureza de oxigênio do Ex. 10.1, com a linha ajustada de regressão e os limites de confiança de 95% para $\mu_{Y|x_0}$.

11.6 Previsão de Novas Observações

- Uma aplicação importante de um modelo de regressão é prever novas (ou futuras) observações Y, correspondentes a um valor especificado do regressor *x*.
- Se x_0 for o valor de interesse do regressor, então

  $$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

  será a estimativa de um novo (ou futuro) valor da resposta Y_0.

11.6 Previsão de Novas Observações

- **Definição:**

Um **intervalo de previsão** de $100(1-\alpha)\%$ **para uma observação futura** y_0, em um certo valor x_0, é dado por

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$
$$\leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \quad (10.35)$$

O valor $\hat{y}_0$ é calculado a partir do modelo de regressão $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

11.6 Previsão de Novas Observações

- **Exemplo 11.6:**

Para ilustrar a construção de um intervalo de previsão, suponha que usemos os dados no Exemplo 11.1 para encontrar um intervalo de previsão de 95% para a próxima observação da pureza de oxigênio em $x_0 = 1{,}00\%$. Usando a Eq. 10.35 e lembrando, do Exemplo 10.6, que $\hat{y}_0 = 89{,}17$, encontramos que o intervalo de previsão é

$$89{,}17 - 2{,}101\sqrt{1{,}17\left[1 + \frac{1}{20} + \frac{(1{,}00 - 1{,}20)^2}{0{,}68}\right]}$$

$$\leq y_0 \leq 89{,}17 + 2{,}101\sqrt{1{,}17\left[1 + \frac{1}{20} + \frac{(1{,}00 - 1{,}20)^2}{0{,}68}\right]}$$

11.6 Previsão de Novas Observações

- **Exemplo 11.6 (cont):**

que simplifica para

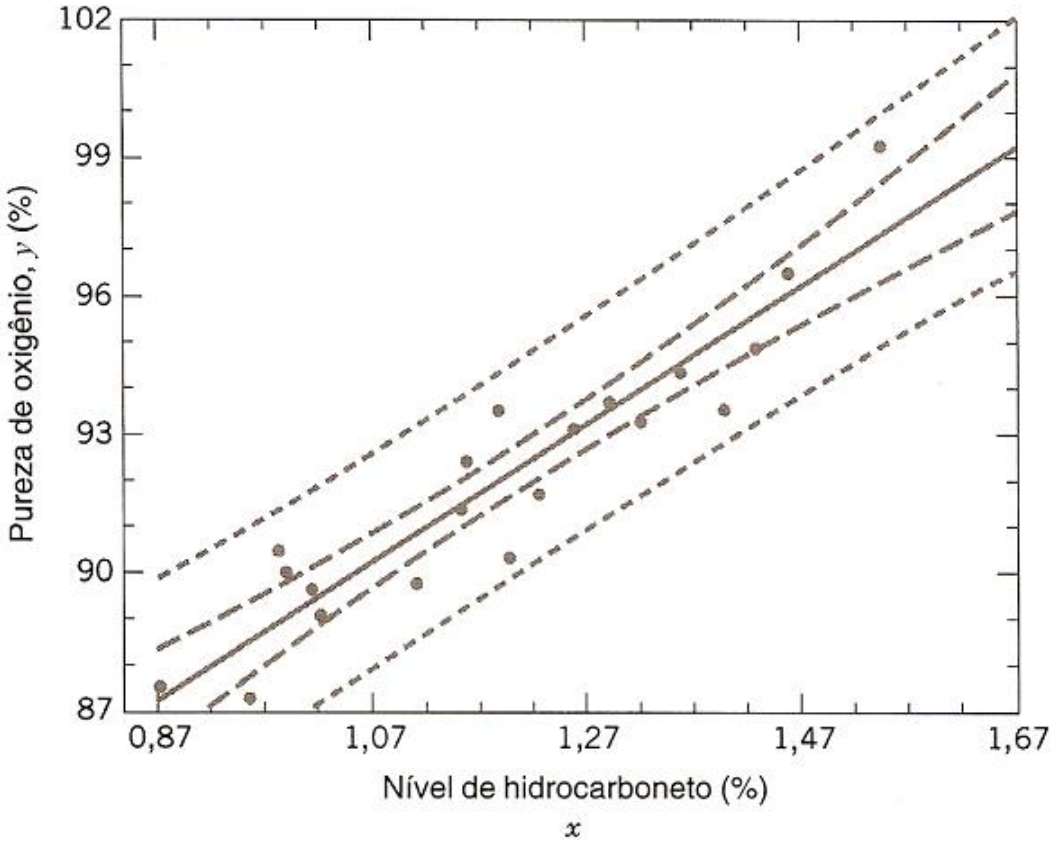
$$86{,}78 \leq y_0 \leq 91{,}56$$

Repetindo os cálculos anteriores para diferentes valores de x_0, podemos obter os intervalos de previsão de 95%, mostrados graficamente na Fig. 11.8, através das linhas superior e inferior em torno do modelo ajustado de regressão. Observe que esse gráfico mostra também os limites de confiança de 95% para $\mu_{Y|x_0}$ calculado no Exemplo 11.5. Ele ilustra que os limites de previsão são sempre mais largos que os limites de confiança.

11.6 Previsão de Novas Observações

- **Exemplo (cont):**

Fig. 11.8 Diagrama de dispersão dos dados de pureza de oxigênio do Ex. 11.1, com a linha ajustada de regressão e os limites de previsão (linhas mais externas) de 95% e os limites de confiança de 95% para $\mu_{Y|x_0}$.



11.7 Cálculo da Adequação do Modelo de Regressão

- Ajustar um modelo de regressão requer várias ***suposições***:
  1. Os erros são variáveis aleatórias não correlacionadas com média zero;
  2. Os erros têm variância constante; e
  3. Os erros são normalmente distribuídos.
- Os analistas devem sempre duvidar da validade dessas suposições e conduzir análises para examinar a adequação do modelo que se está testando.

11.7 Cálculo da Adequação do Modelo de Regressão

11.7.1 Análise Residual

- Os resíduos de um modelo de regressão são

  $e_i = y_i - \hat{y}_i$, $i = 1, 2, ..., n$,

  em que y_i é uma observação real e $\hat{y}_i$ é o valor ajustado corresponde, proveniente do modelo de regressão;
- A análise dos resíduos é frequentemente útil na verificação da suposição de que os erros sejam distribuídos de forma aproximadamente normal, com variância constante, assim como na determinação da utilidade dos termos adicionais no modelo.

11.7 Cálculo da Adequação do Modelo de Regressão

11.7.1 Análise Residual

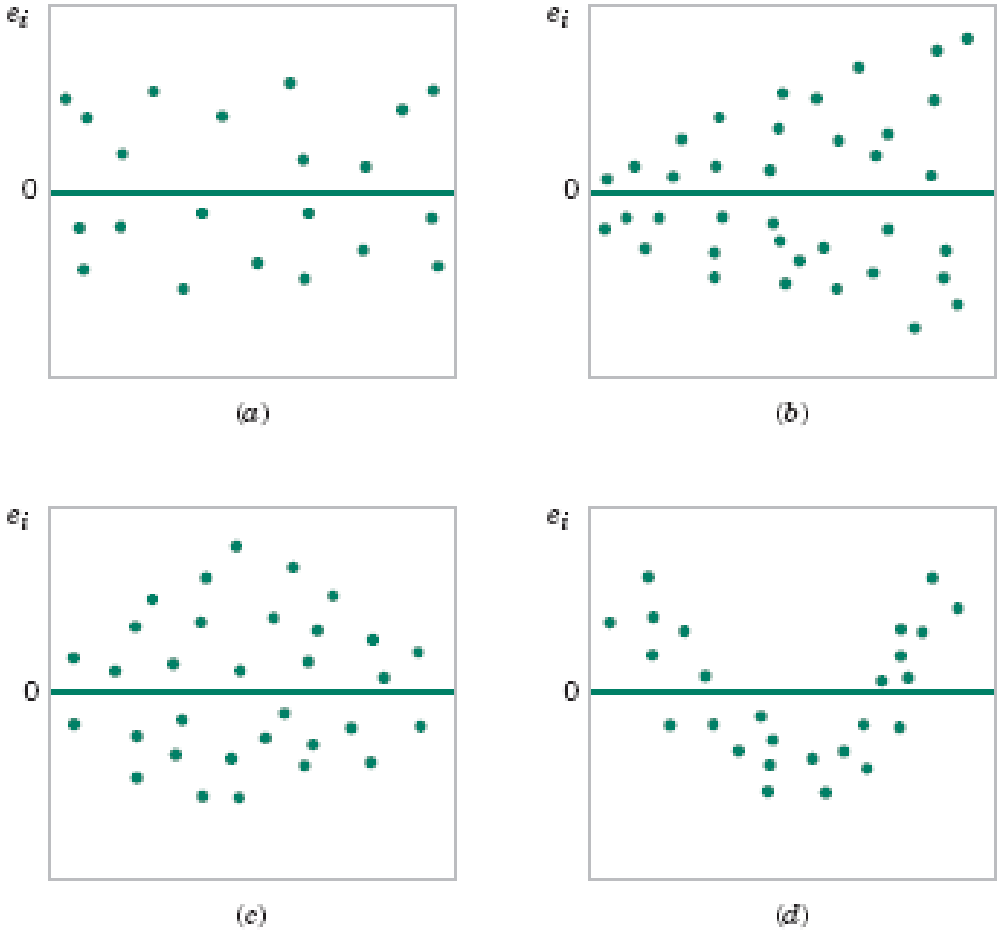


Fig. 11.9 Padrões de comportamento para gráficos de resíduos, (a) satisfatório, (b) funil, (c) arco duplo, (d) não linear.

11.7 Cálculo da Adequação do Modelo de Regressão

11.7.1 Análise Residual

- **Exemplo 11.7:** O modelo de regressão para os dados de pureza de oxigênio no Exemplo 11.1 é $\hat{y} = 74{,}20 + 14{,}97x$. A Tabela 11.5 apresenta os valores observados e previstos de y, para cada valor de x proveniente desse conjunto de dados, juntamente com o resíduo correspondente. Esses valores foram calculados usando o Minitab e mostram o número típico de casas decimais na saída do computador. Um gráfico de probabilidade normal dos resíduos é mostrado na Fig. 11.10. Visto que os resíduos caem aproximadamente ao longo de uma linha reta na figura, concluímos que não há um sério desvio da normalidade. Os resíduos são também plotados contra os valores previstos $\hat{y}_i$ na Fig. 11.11 e contra os níveis de hidrocarbonetos x_i na Fig. 11.12. Esses gráficos não indicam qualquer inadequação séria do modelo.

11.7 Cálculo da Adequação do Modelo de Regressão

11.7.1 Análise Residual

- **Exemplo 11.7 (cont.):**

	Nível de Hidrocarboneto, x	Pureza de Oxigênio, y	Valor Previsto, $\hat{y}$	Resíduo, $e = y - \hat{y}$
1	0,99	90,01	89,069009	0,940991
2	1,02	89,05	89,518136	−0,468136
3	1,15	91,43	91,464353	−0,034353
4	1,29	93,74	93,560279	0,179721
5	1,46	96,73	96,105332	0,624668
6	1,36	94,45	94,608242	−0,158242
7	0,87	87,59	87,272501	−0,317499
8	1,23	91,77	92,662025	−0,892025
9	1,55	99,42	97,452713	1,967287
10	1,40	93,65	95,207078	−1,557078
11	1,19	93,54	92,063189	1,476811
12	1,15	92,52	91,614062	0,905938
13	0,98	90,56	88,919300	1,640700
14	1,01	89,54	89,368427	0,171573
15	1,11	89,85	90,865517	−1,015517
16	1,20	90,39	92,212898	−1,822898
17	1,26	93,25	93,111152	0,138848
18	1,32	93,41	94,009406	−0,599406
19	1,43	94,98	95,656205	−0,676205
20	0,95	87,33	88,470173	−1,140173

Tabela 11.5 Dados de Pureza de Oxigênio do Ex. 11.1, Valores Previsto e Resíduos.

11.7 Cálculo da Adequação do Modelo de Regressão

11.7.1 Análise Residual

- **Exemplo 11.7 (cont.):**

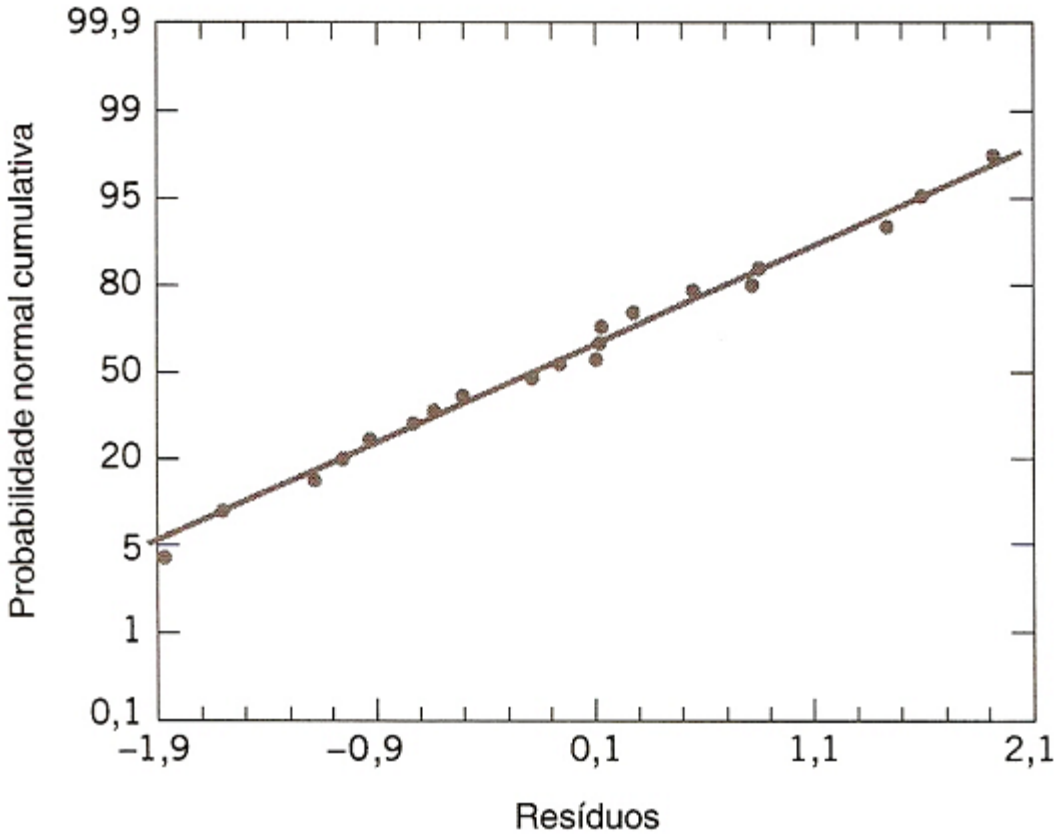


Fig. 11.10 Gráfico de probabilidade normal dos resíduos, Ex. 11.7.

11.7 Cálculo da Adequação do Modelo de Regressão

11.7.1 Análise Residual

- **Exemplo 11.7 (cont.):**

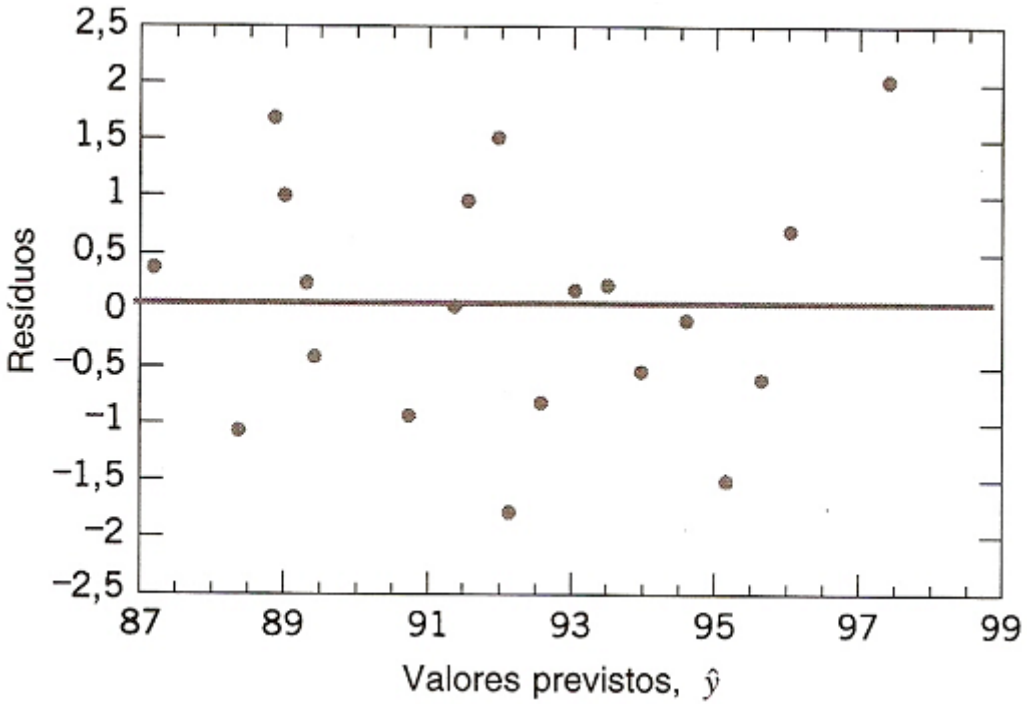


Fig. 11.11 Gráfico dos resíduos *versus* pureza prevista do oxigênio, $\hat{y}$, Ex. 11.7.

11.7 Cálculo da Adequação do Modelo de Regressão

11.7.2 Coeficiente de Determinação (R^2)

- A quantidade

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T} = \frac{S_{XY}^2}{S_{XX} S_{YY}}$$

é chamada de ***coeficiente de determinação***, sendo frequentemente usada para julgar a adequação de um modelo de regressão.

- Da análise de variância

$0 \leq R^2 \leq 1$.

- Frequentemente, referimo-nos a R^2 como a quantidade de variabilidade nos dados explicada (ou considerada) pelo modelo de regressão.

11.7 Cálculo da Adequação do Modelo de Regressão

11.7.2 Coeficiente de Determinação (R^2)

- Para o modelo de regressão da pureza do oxigênio, temos

$$R^2 = SQ_R/SQ_T$$
$$= 152{,}39/173{,}37$$
$$= 0{,}8790$$

- Portanto, o modelo ***explica*** 87,90% da variabilidade dos dados.

11.8 Correlação

- **Definição:**

  O ***coeficiente de correlação*** é definido como

  $\rho = \sigma_{XY}/\sigma_X\sigma_Y$,

  sendo σ_X e σ_Y as variâncias de X e Y, respectivamente, e σ_{XY}, a covariância entre X e Y, definido como

  $\sigma_{XY} = E(XY) - \mu_X\mu_Y$.

- O estimador de ρ é o ***coeficiente de correlação da amostra***

$$R = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}.$$

11.8 Correlação

- **Definição:**

  Note que

$$\hat{\beta}_1 = \left(\frac{S_{YY}}{S_{XX}} \right)^{1/2} R,$$

  Assim, podemos escrever também

$$R^2 = \hat{\beta}_1^2 \frac{S_{XX}}{SQ_T} = \frac{\hat{\beta}_1 S_{XY}}{SQ_T} = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T} = \frac{S_{XY}^2}{S_{XX} S_{YY}}$$

11.8 Correlação

- É frequentemente útil testar as hipóteses

  H_0: $\rho = 0$

  H_1: $\rho \neq 0$

- A estatística apropriada de teste para essas hipóteses é

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}},$$

- Rejeita-se H_0 se

  $|t_0| > t_{\alpha/2;\, n\text{-}2}$.

11.8 Correlação

- O procedimento de teste para as hipóteses

  $H_0: \rho = \rho_0$

  $H_1: \rho \neq \rho_0$

  em que $\rho_0 \neq 0$ é um pouco mais complicado.
- Neste caso, a estatística de teste apropriada é

  $Z_0 = (\text{arctgh } R - \text{arctgh } \rho_0)(n-3)^{1/2}$.

- Rejeita-se H_0 se

  $|z_0| > z_{\alpha/2}$.

11.8 Correlação

- **Exemplo:** No Cap. 1 (Seção 1.4), é descrita uma aplicação de análise de regressão em que um engenheiro, em uma planta de montagem de semicondutores, está implementando a relação entre a resistência ao puxamento de um fio colado e dois fatores: comprimento do fio e altura do molde. Nesse exemplo, consideraremos somente um dos fatores: o comprimento do fio. Uma amostra aleatória de 25 unidades é selecionada e testada, sendo a resistência ao puxamento do fio e o comprimento do fio observados para cada unidade. Os dados são mostrados na Tabela 1.2. Consideramos que a resistência ao puxamento e o comprimento do fio sejam distribuídos normal e conjuntamente.

11.8 Correlação

- **Exemplo (cont.):** Usando os dados na Tabela 1.2, podemos calcular

$$SQ_T = 6105{,}9447 \quad S_{xx} = 698{,}5600 \text{ e } S_{xy} = 2027{,}7132$$

O modelo de regressão é

$$\hat{y} = 5{,}1145 + 2{,}9027x$$

O coeficiente de correlação da amostra entre X e Y da amostra é calculado pela Eq. 10.50 como

$$r = \frac{S_{xy}}{[S_{xx}S_{yy}]^{1/2}} = \frac{2027{,}7132}{[(698{,}560)(6105{,}9447)]^{1/2}} = 0{,}9818$$

Note que $r^2 = (0{,}9818)^2 = 0{,}9640$ ou que, aproximadamente, 96,40% da variabilidade na resistência ao puxamento é explicada pela relação linear com o comprimento do fio.

11.8 Correlação

- **Exemplo (cont.):**

Agora suponha que desejemos testar a hipótese

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

com $\alpha = 0{,}05$. Podemos calcular a estatística t pela Eq. 10.53 como

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}9818\sqrt{23}}{\sqrt{1-0{,}9640}} = 24{,}82$$

e uma vez que $t_{0,\,025;23} = 2{,}069$, rejeitamos H_0 e concluímos que o coeficiente de correlação é $\rho \neq 0$.

11.8 Correlação

- **Exemplo (final):**

Finalmente, podemos construir um intervalo aproximado de confiança de 95% para ρ a partir da Eq. 10.57. Uma vez que arctgh r = arctgh 0,9818 = 2,3452, a Eq. 10.57 se torna

$$\text{tgh}\left(2{,}3452 - \frac{1{,}96}{\sqrt{22}}\right) \leq \rho \leq \text{tgh}\left(2{,}3452 + \frac{1{,}96}{\sqrt{22}}\right)$$

que se reduz a

$$0{,}9585 \leq \rho \leq 0{,}9921$$

11.9 Transformações

- Ocasionalmente, um diagrama de dispersão exibirá uma relação aparentemente não linear entre Y e *x* e, em algumas dessas situações, uma função não linear pode ser expressa como uma linha reta, usando uma transformação adequada.
- Tais modelos são denominados ***intrinsecamente lineares***.
- Por exemplo, a função

  $$Y = \beta_0 \times e^{\beta_1 x} \times \varepsilon$$

  é intrinsecamente linear, uma vez que pode ser transformada em uma linha reta por uma transformação logarítmica

  $$\ln Y = \ln \beta_0 + \beta_1 x + \ln \varepsilon.$$

  Esta transformação requer que os termos transformados do erro $\ln \varepsilon$, sejam normal e independentemente distribuídos com média 0 e variância σ^2.

TERMOS E CONCEITOS IMPORTANTES

- **Teste da análise de variância em regressão**
- **Intervalos de confiança da resposta média**
- **Coeficiente de correlação**
- **Modelo empírico**
- **Intervalos de confiança dos parâmetros do modelo**
- **Modelo intrinsecamente linear**
- **Estimativa dos mínimos quadrado dos parâmetros do modelo de regressão**
- **Determinação da adequação do modelo**
- **Intervalos de previsão de novas observações**
- **Gráfico dos resíduos**
- **Resíduos**
- **Diagramas de dispersão**
- **Significância da regressão**
- **Erros padrão da regressão linear simples**
- **Transformações**