Capítulo 11 | Regressão linear simples e correlação

11.1 Introdução à regressão linear

Uma forma razoável de relação entre a resposta *Y* e o regressor *x* é a relação linear

$$Y = \alpha + \beta x,$$

onde, é claro, *α é o intercepto e β* é a *inclinação.* A relação está ilustrada na Figura 11.1.

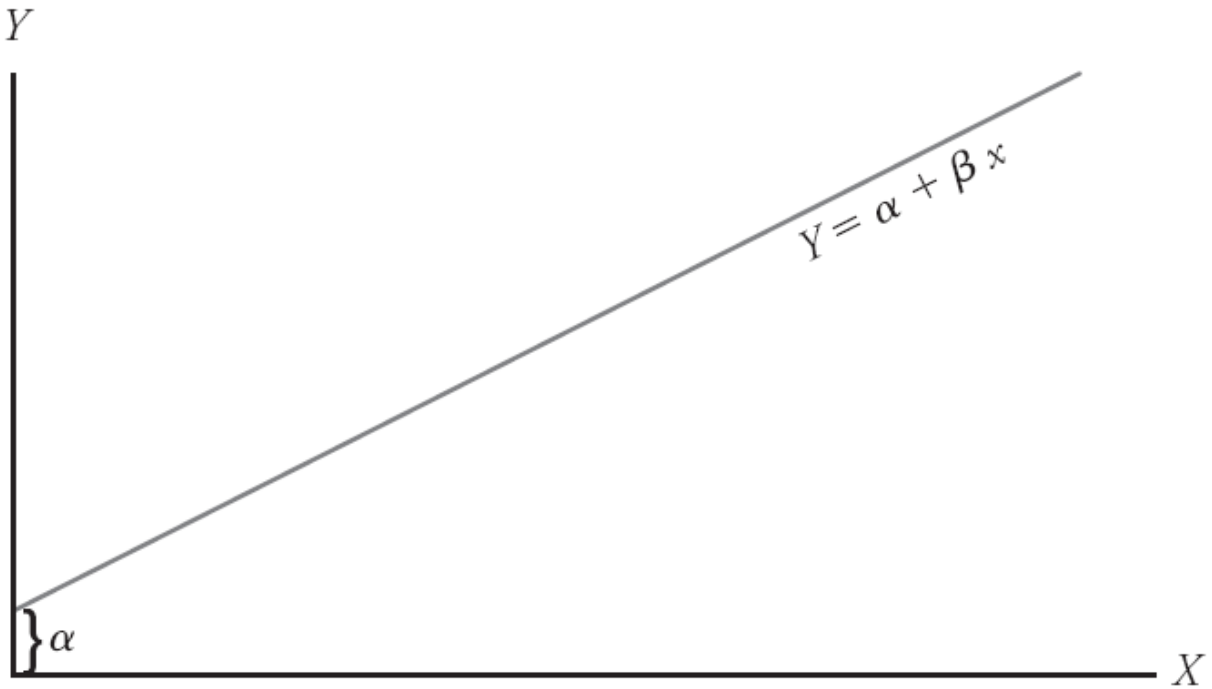


Figura 11.1 Uma relação linear.

11.2 O modelo de regressão linear simples

Modelo de regressão linear simples

A resposta Y está relacionada com a variável independente x por meio da equação

$$Y = \alpha + \beta x + \epsilon.$$

Nesse caso, α e β são os parâmetros desconhecidos de inclinação e de intercepto, respectivamente, e ϵ é uma variável aleatória assumida como sendo distribuída com $E(\epsilon) = 0$ e $Var(\epsilon) = \sigma^2$. A quantidade σ^2 é freqüentemente chamada de variância do erro ou variância residual.

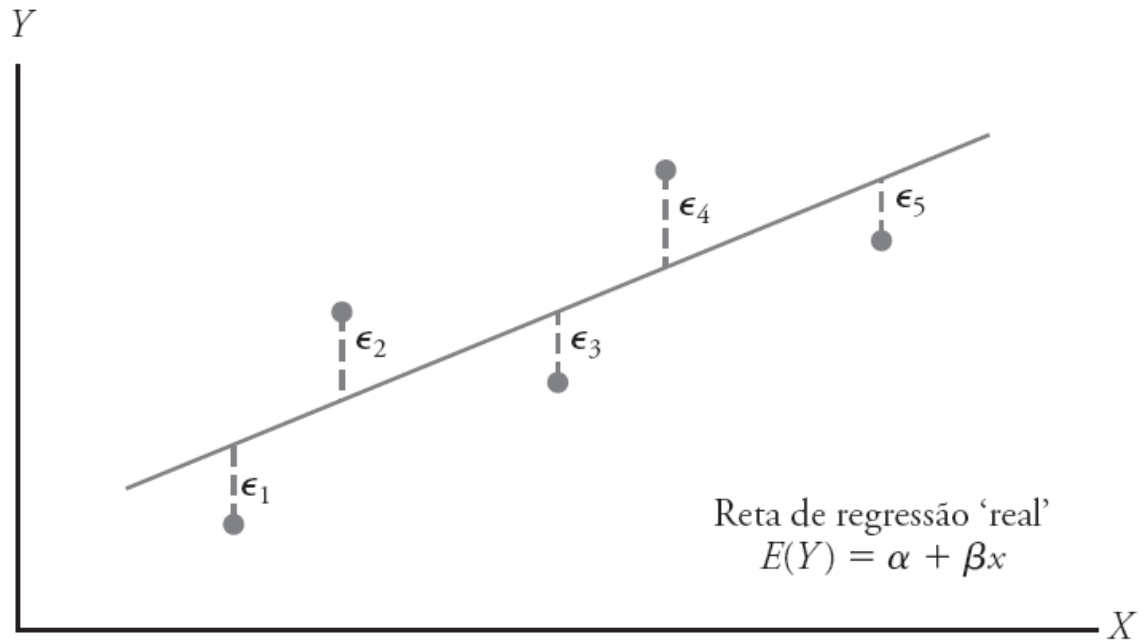


Figura 11.2 Dados hipotéticos (x, y) dispersos ao redor da reta de regressão real, para $n = 5$.

Tabela 11.1 Medidas dos sólidos e demanda de oxigênio químico

Redução de sólidos, x (%)	Demanda de oxigênio químico, y (%)	Redução de sólidos, x (%)	Demanda de oxigênio químico, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

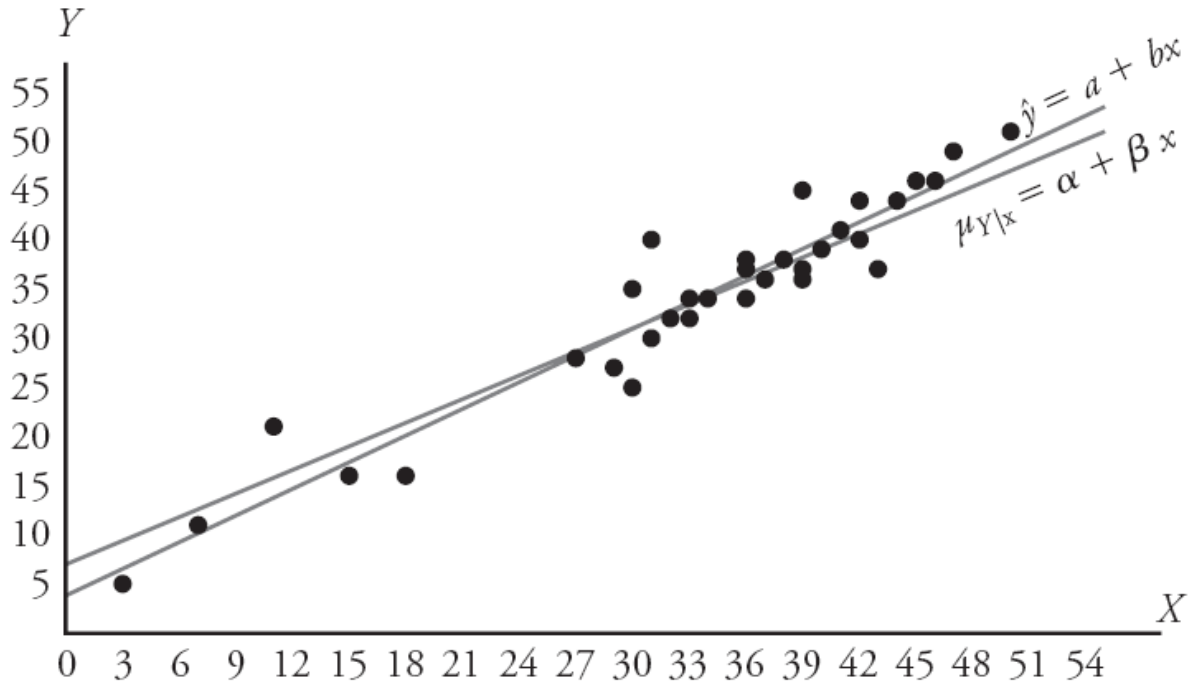


Figura 11.3 Diagrama de dispersão com retas de regressão.

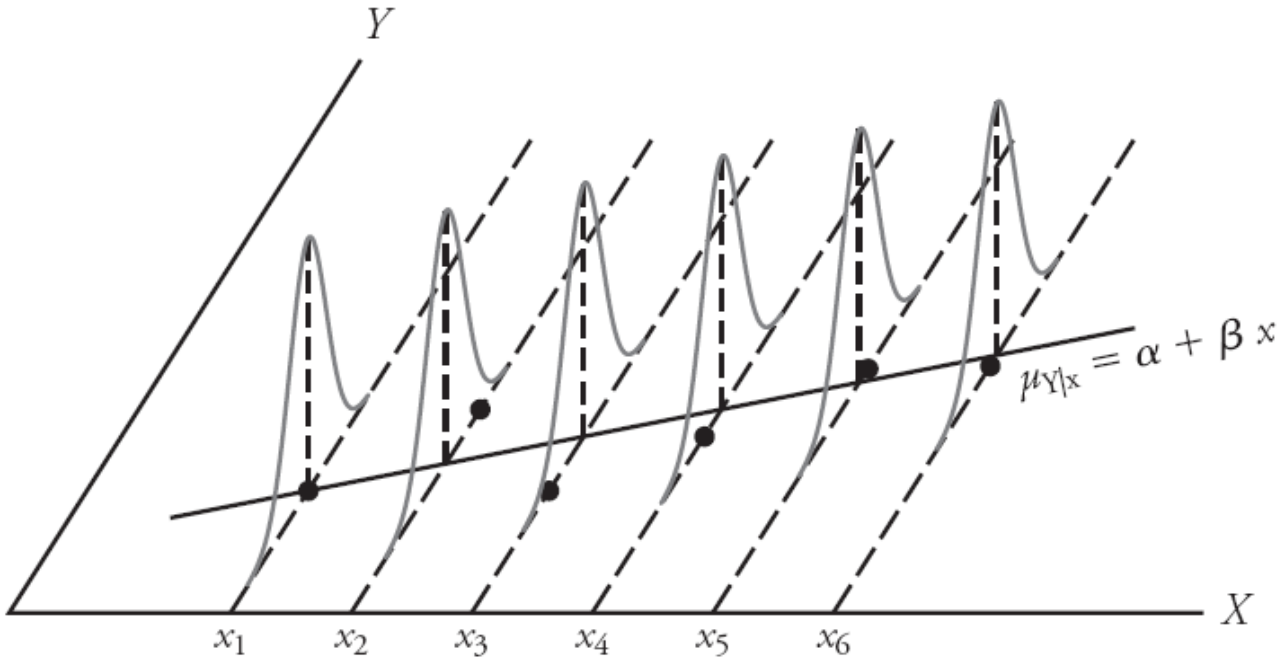


Figura 11.4 Observações individuais em torno da reta de regressão real.

11.3 Mínimos quadrados e o modelo ajustado

Resíduo: Erro no ajuste

Dado um conjunto de dados de regressão $[(x_i, y_i); i = 1, 2, \ldots, n]$ e um modelo ajustado, $\hat{y}_i = a + bx_i$, o i-ésimo resíduo e_i é dado por

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

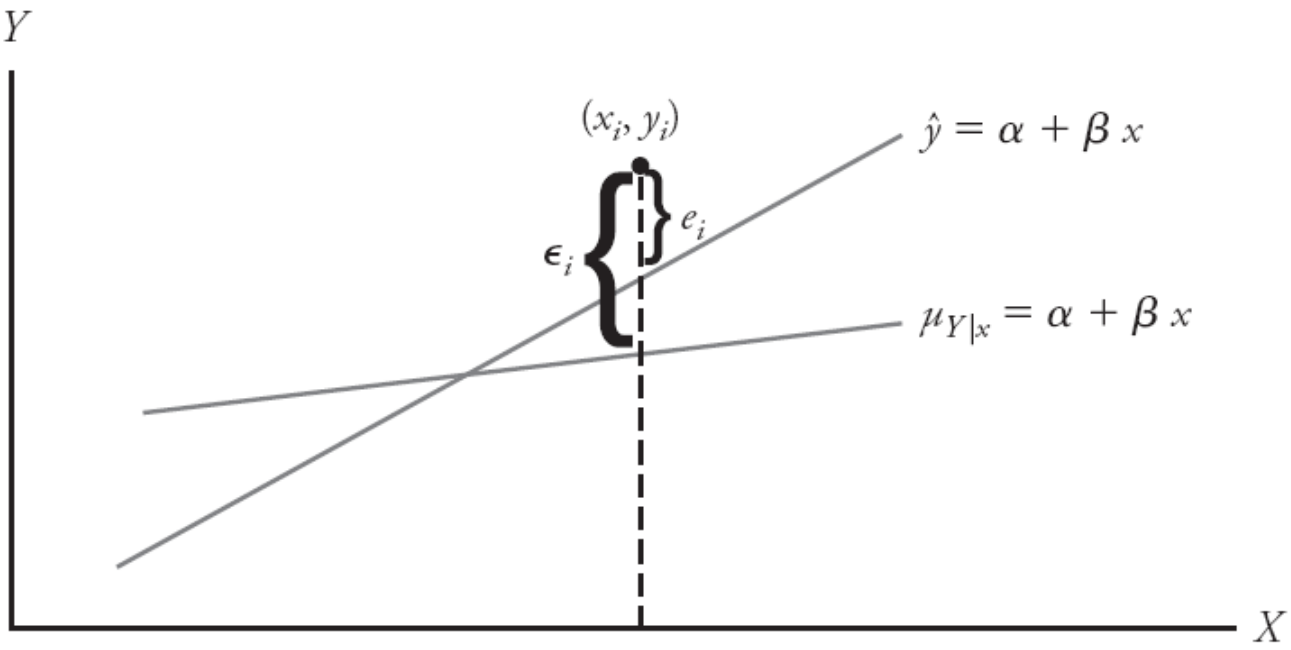


Figura 11.5 Comparação de ϵ_i com o resíduo, ϵ_i.

Estimando os coeficientes de regressão

Dada a amostra $\{(x_i, y_i); i = 1, 2, ..., n,\}$, as estimativas de mínimos quadrados, a e b, dos coeficientes de regressão α e β são calculadas das fórmulas

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

e

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}.$$

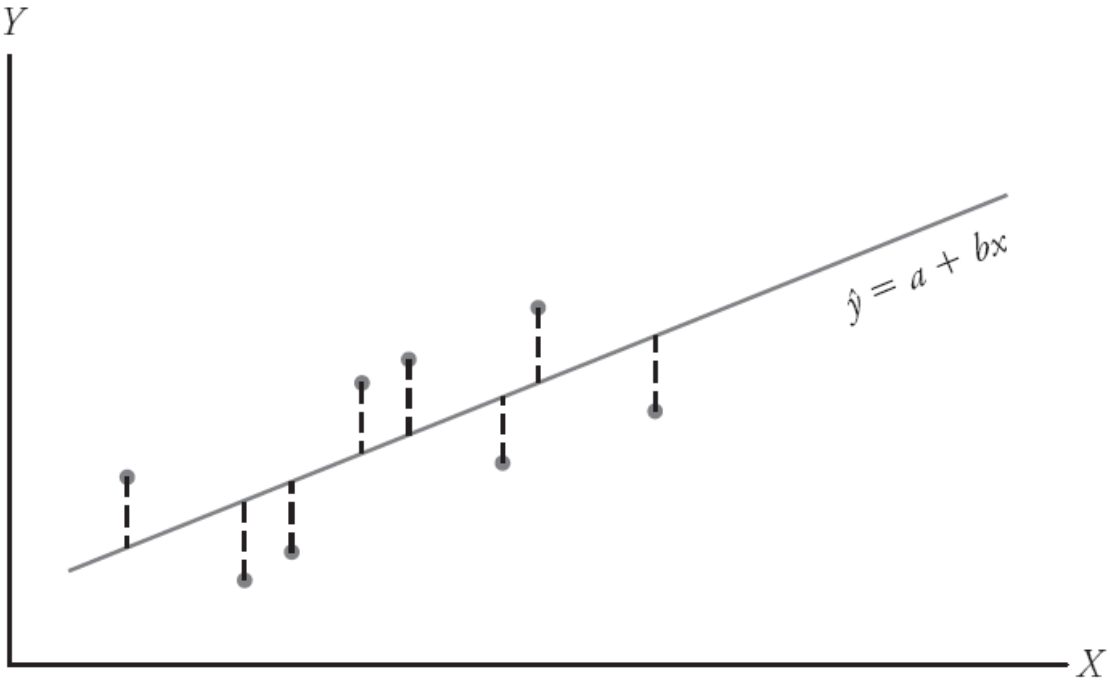


Figura 11.6 Resíduos como desvios verticais.

11.4 Propriedades dos estimadores de mínimos quadrados

Teorema 11.1

Uma estimativa não-viciada de σ^2 é

$$s^2 = \frac{SQE}{n-2} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}.$$

11.5 Inferências sobre os coeficientes de regressão

Intervalo de confiança para β

Um intervalo de confiança de $100(1-\alpha)\%$ para o parâmetro β da reta de regressão $\mu_{Y|x_0} = \alpha + \beta x$ é

$$b - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta < b + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}},$$

onde $t_{\alpha/2}$ é um valor da distribuição t com $n-2$ graus de liberdade.

```
Regression Analysis: COD versus Per_Red

The regression equation is COD = 3,83 + 0,904
  Per_Red

Predictor     Coef  SE Coef      T      P
Constant     3,830    1,768   2,17  0,038
Per_Red    0,90364  0,05012  18,03  0,000

S = 3,22954   R-Sq = 91,3%   R-Sq(adj) = 91,0%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       1  3390,6  3390,6  325,08  0,000
Residual Error  31   323,3    10,4
Total           32  3713,9
```

Figura 11.7 Impressão *Minitab* do teste t para os dados do Exemplo 11.1.

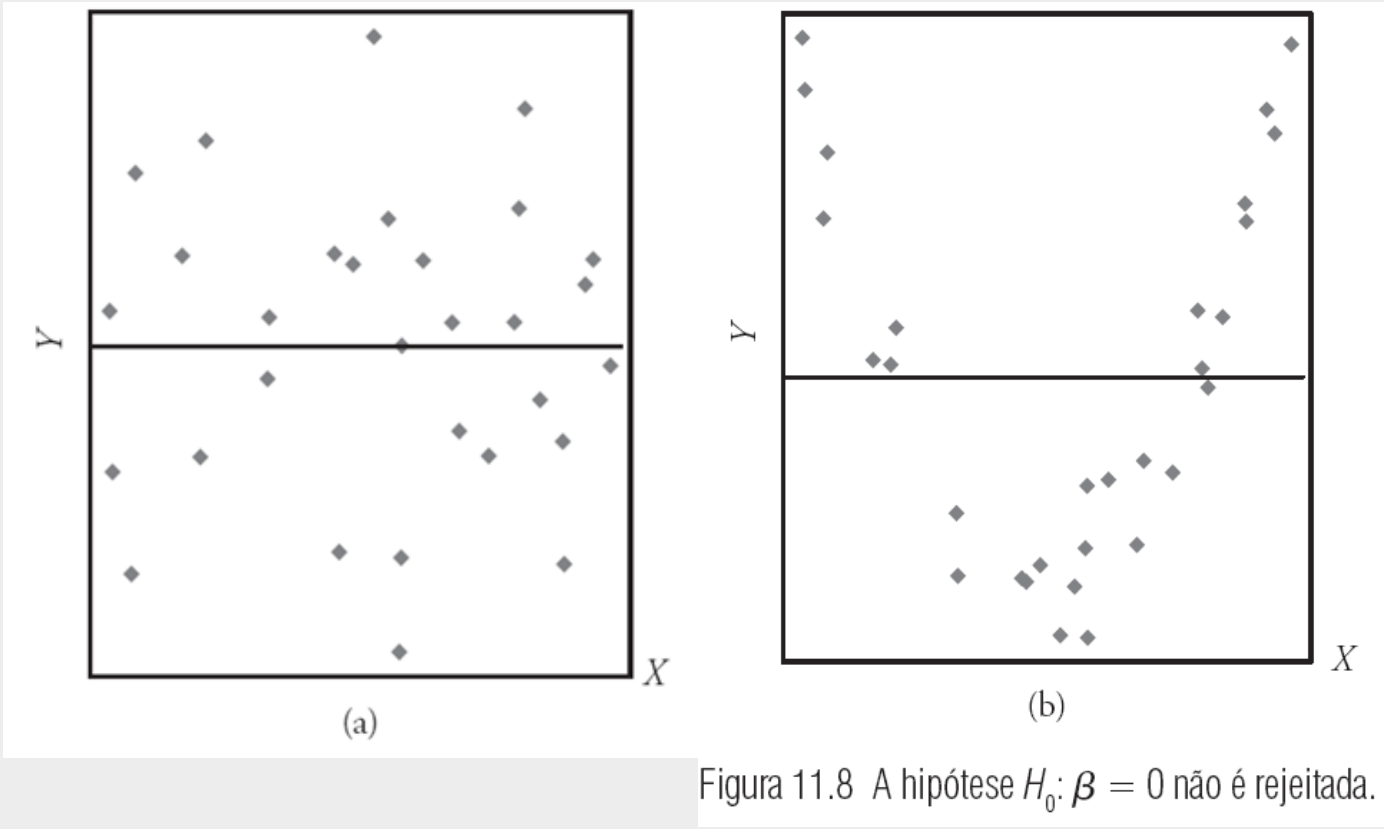


Figura 11.8 A hipótese H_0: $\beta = 0$ não é rejeitada.

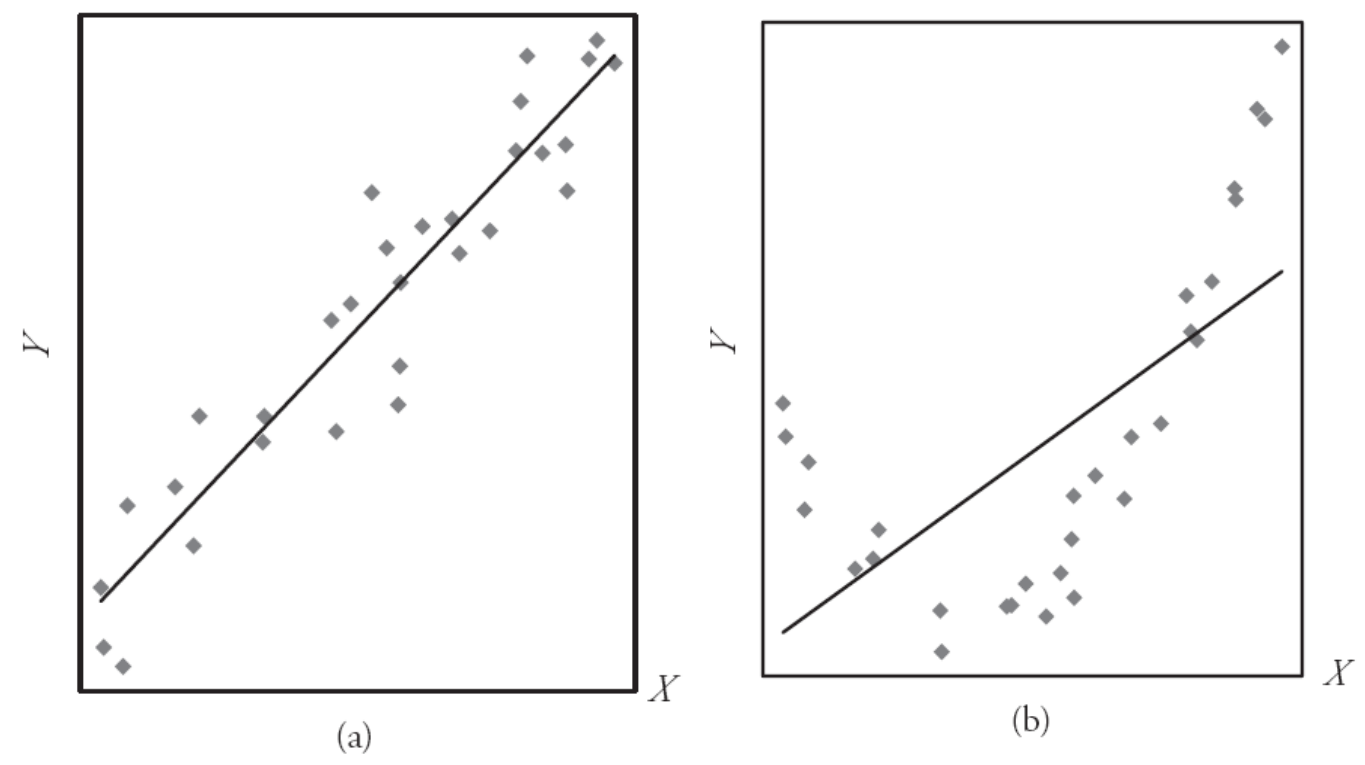


Figura 11.9 A hipótese H_0: $\beta = 0$ é rejeitada.

Intervalo de confiança para α

Um intervalo de confiança de $100(1-\alpha)\%$ para o parâmetro α da reta de regressão $\mu_{Y|x_0} = \alpha + \beta x$ é

$$a - t_{\alpha/2}\frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}} < \alpha < a + t_{\alpha/2}\frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}},$$

onde $t_{\alpha/2}$ é um valor da distribuição t com $n-2$ graus de liberdade.

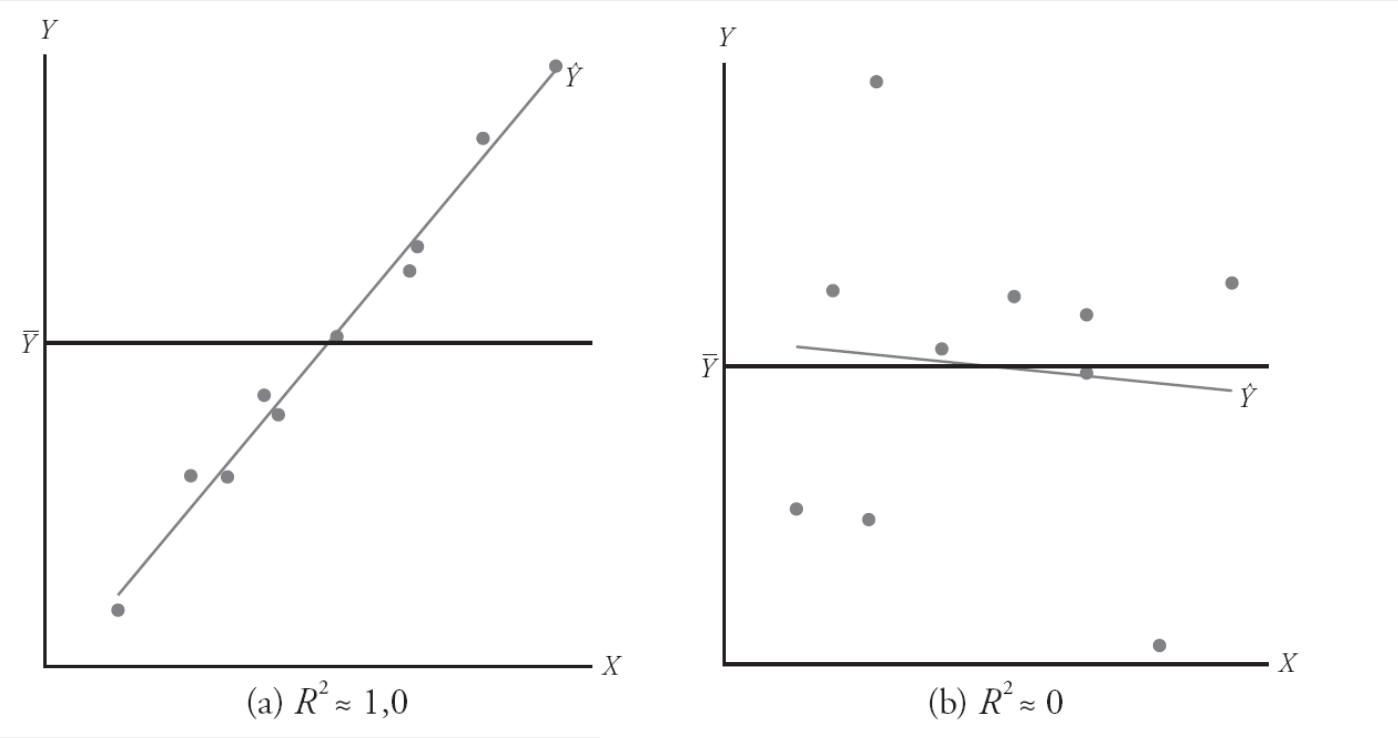


Figura 11.10 Gráficos mostrando um ajuste muito bom e um ajuste pobre.

11.6 Predição

Intervalo de confiança para $\mu_{Y|x_0}$

Um intervalo de confiança de $100(1-\alpha)\%$ para a resposta média $\mu_{Y|x_0}$ é

$$\hat{y}_0 - t_{\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} <$$

$$\hat{y}_0 + t_{\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

onde $t_{\alpha/2}$ é um valor da distribuição t com $n-2$ graus de liberdade.

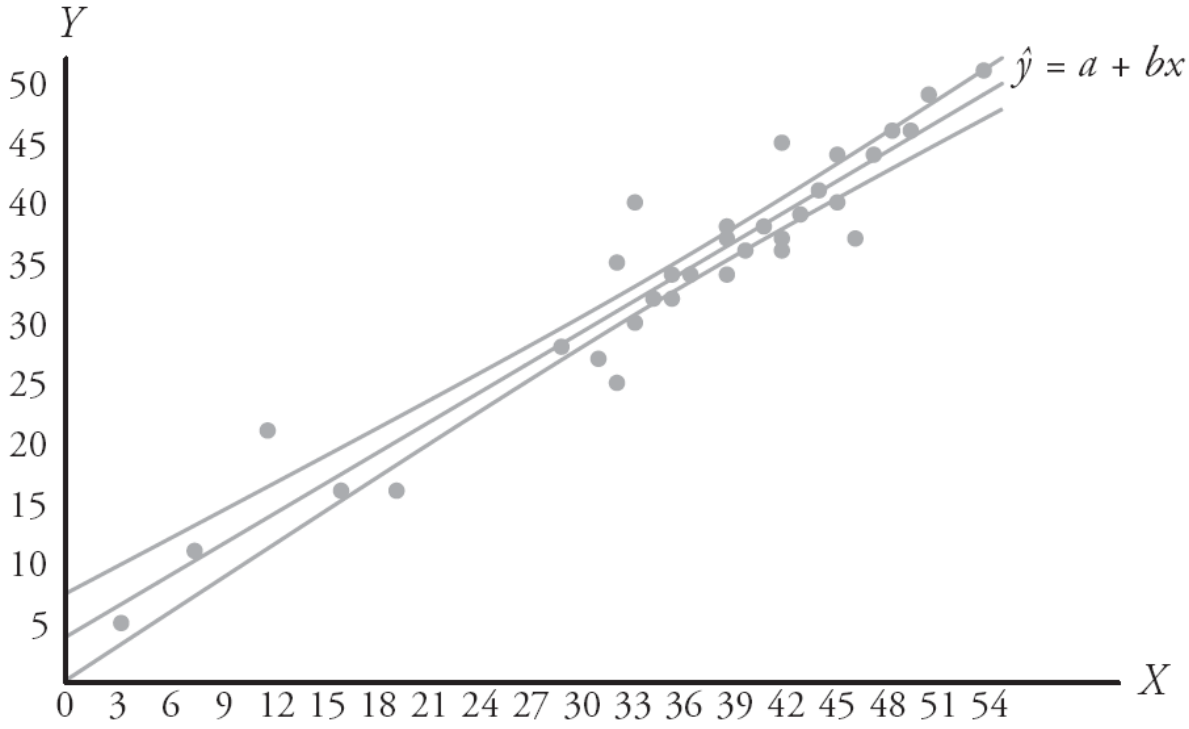


Figura 11.11 Limites de confiança para o valor médio de $Y|x$.

Intervalo de predição para y_0

Um intervalo de predição de $100(1 - \alpha)\%$ para uma única resposta y_0 é dado por

$$\hat{y}_0 - t_{\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 <$$

$$\hat{y}_0 + t_{\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

onde $t_{\alpha/2}$ é um valor da distribuição t com $n - 2$ graus de liberdade.

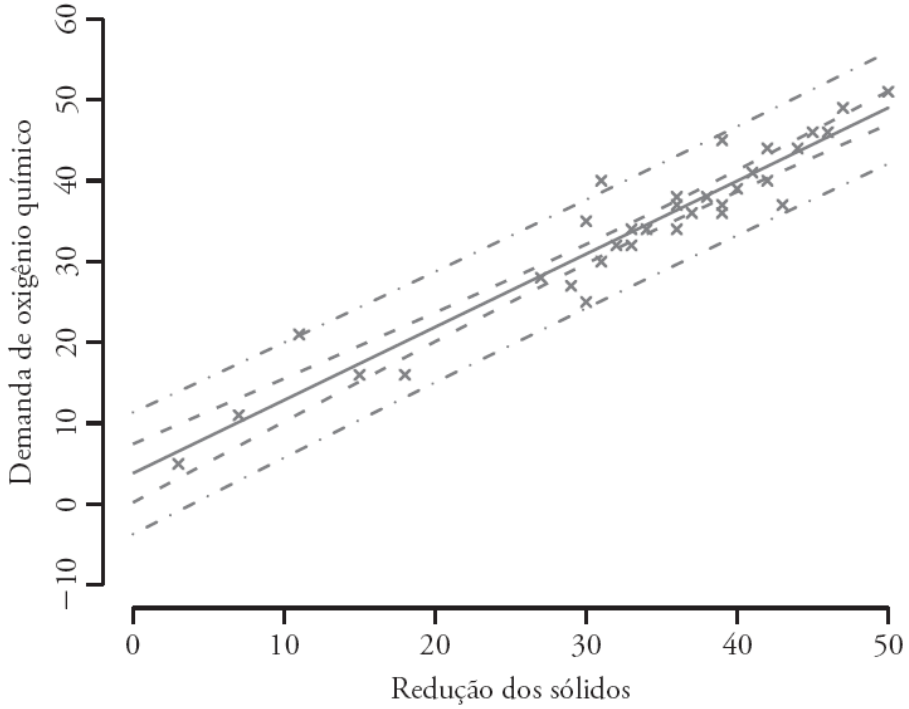


Figura 11.12 Intervalos de confiança e de predição para os dados da demanda de oxigênio químico; as faixas internas indicam os limites de confiança para as respostas médias e as faixas externas indicam os limites de predição para as respostas futuras.

Root MSE			1,48794	R-Square	0,9509			
Dependent Mean			21,50000	Adj R-Sq	0,9447			
Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|			
Intercept	1	44,78018	1,92919	23,21	< 0,0001			
WT	1	-0,00686	0,00055133	-12,44	< 0,0001			
MODEL	WT	MPG	Predict	LMean	UMean	Lpred	Upred	Residual
GMC	4520	15	13,7720	11,9752	15,5688	9,8988	17,6451	1,22804
Geo	2065	29	30,6138	28,6063	32,6213	26,6385	34,5891	-1,61381
Honda	2440	31	28,0412	26,4143	29,6681	24,2439	31,8386	2,95877
Hyundai	2290	28	29,0703	27,2967	30,8438	25,2078	32,9327	-1,07026
Infinit	3195	23	22,8618	21,7478	23,9758	19,2543	26,4693	0,13825
Isuzu	3480	21	20,9066	19,8160	21,9972	17,3062	24,5069	0,09341
Jeep	4090	15	16,7219	15,3213	18,1224	13,0158	20,4279	-1,72185
Land	4535	13	13,6691	11,8570	15,4811	9,7888	17,5493	-0,66905
Lexus	3390	22	21,5240	20,4390	22,6091	17,9253	25,1227	0,47599
Lincoln	3930	18	17,8195	16,5379	19,1011	14,1568	21,4822	0,18051

Figura 11.13 Impressão *SAS* para o Exercício 11.29.

11.8 Abordagem da análise de variância

Tabela 11.2 Análise de variância para testar $\beta = 0$

Fonte da variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	f calculado
Regressão	SQR	1	SQR	$\frac{SQR}{s^2}$
Erro	SQE	$n-2$	$s^2 = \frac{SQE}{n-2}$	
Total	SQT	$n-1$		

11.9 Teste da linearidade da regressão: dados com observações repetidas

The regression equation is COD = 3,83 + 0,904 Per_Red

Predictor	Coef	SE Coef	T	P
Constant	3,830	1,768	2,17	0,038
Per_Red	0,90364	0,05012	18,03	0,000

S = 3,22954 R-Sq = 91,3% R-Sq(adj) = 91,0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3390,6	3390,6	325,08	0,000
Residual Error	31	323,3	10,4		
Total	32	3713,9			

Obs	Per_Red	COD	Fit	SE Fit	Residual	St Resid
1	3,0	5,000	6,541	1,627	-1,541	-0,55
2	36,0	34,000	36,361	0,576	-2,361	-0,74
3	7,0	11,000	10,155	1,440	0,845	0,29
4	37,0	36,000	37,264	0,590	-1,264	-0,40
5	11,0	21,000	13,770	1,258	7,230	2,43
6	38,0	38,000	38,168	0,607	-0,168	-0,05
7	15,0	16,000	17,384	1,082	-1,384	-0,45
8	39,0	37,000	39,072	0,627	-2,072	-0,65
9	18,0	16,000	20,095	0,957	-4,095	-1,33
10	39,0	36,000	39,072	0,627	-3,072	-0,97
11	27,0	28,000	28,228	0,649	-0,228	-0,07
12	39,0	45,000	39,072	0,627	5,928	1,87
13	29,0	27,000	30,035	0,605	-3,035	-0,96
14	40,0	39,000	39,975	0,651	-0,975	-0,31
15	30,0	25,000	30,939	0,588	-5,939	-1,87
16	41,0	41,000	40,879	0,678	0,121	0,04
17	30,0	35,000	30,939	0,588	4,061	1,28
18	42,0	40,000	41,783	0,707	-1,783	-0,57
19	31,0	30,000	31,843	0,575	-1,843	-0,58
20	42,0	44,000	41,783	0,707	2,217	0,70
21	31,0	40,000	31,843	0,575	8,157	2,57
22	43,0	37,000	42,686	0,738	-5,686	-1,81
23	32,0	32,000	32,746	0,567	-0,746	-0,23
24	44,0	44,000	43,590	0,772	0,410	0,13
25	33,0	34,000	33,650	0,563	0,350	0,11
26	45,0	46,000	44,494	0,807	1,506	0,48
27	33,0	32,000	33,650	0,563	-1,650	-0,52
28	46,0	46,000	45,397	0,843	0,603	0,19
29	34,0	34,000	34,554	0,563	-0,554	-0,17
30	47,0	49,000	46,301	0,881	2,699	0,87
31	36,0	37,000	36,361	0,576	0,639	0,20
32	50,0	51,000	49,012	1,002	1,988	0,65
33	36,0	38,000	36,361	0,576	1,639	0,52

Figura 11.14 Impressão *Minitab* da regressão linear simples para os dados sobre a demanda de oxigênio químico: parte I.

Obs	Fit	SE Fit	95% CI	95% PI
1	6,541	1,627	(3,223, 9,858)	(-0,834, 13,916)
2	36,361	0,576	(35,185, 37,537)	(29,670, 43,052)
3	10,155	1,440	(7,218, 13,092)	(2,943, 17,367)
4	37,264	0,590	(36,062, 38,467)	(30,569, 43,960)
5	13,770	1,258	(11,204, 16,335)	(6,701, 20,838)
6	38,168	0,607	(36,931, 39,405)	(31,466, 44,870)
7	17,384	1,082	(15,177, 19,592)	(10,438, 24,331)
8	39,072	0,627	(37,793, 40,351)	(32,362, 45,781)
9	20,095	0,957	(18,143, 22,047)	(13,225, 26,965)
10	39,072	0,627	(37,793, 40,351)	(32,362, 45,781)
11	28,228	0,649	(26,905, 29,551)	(21,510, 34,946)
12	39,072	0,627	(37,793, 40,351)	(32,362, 45,781)
13	30,035	0,605	(28,802, 31,269)	(23,334, 36,737)
14	39,975	0,651	(38,648, 41,303)	(33,256, 46,694)
15	30,939	0,588	(29,739, 32,139)	(24,244, 37,634)
16	40,879	0,678	(39,497, 42,261)	(34,149, 47,609)
17	30,939	0,588	(29,739, 32,139)	(24,244, 37,634)
18	41,783	0,707	(40,341, 43,224)	(35,040, 48,525)
19	31,843	0,575	(30,669, 33,016)	(25,152, 38,533)
20	41,783	0,707	(40,341, 43,224)	(35,040, 48,525)
21	31,843	0,575	(30,669, 33,016)	(25,152, 38,533)
22	42,686	0,738	(41,181, 44,192)	(35,930, 49,443)
23	32,746	0,567	(31,590, 33,902)	(26,059, 39,434)
24	43,590	0,772	(42,016, 45,164)	(36,818, 50,362)
25	33,650	0,563	(32,502, 34,797)	(26,964, 40,336)
26	44,494	0,807	(42,848, 46,139)	(37,704, 51,283)
27	33,650	0,563	(32,502, 34,797)	(26,964, 40,336)
28	45,397	0,843	(43,677, 47,117)	(38,590, 52,205)
29	34,554	0,563	(33,406, 35,701)	(27,868, 41,239)
30	46,301	0,881	(44,503, 48,099)	(39,473, 53,128)
31	36,361	0,576	(35,185, 37,537)	(29,670, 43,052)
32	49,012	1,002	(46,969, 51,055)	(42,115, 55,908)
33	36,361	0,576	(35,185, 37,537)	(29,670, 43,052)

Figura 11.15 Impressão *Minitab* para a regressão linear simples para os dados da demanda de oxigênio químico; parte II.

Cálculo da falta de ajuste na soma dos quadrados

1. Calcule a soma dos quadrados do erro puro

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2.$$

Essa soma dos quadrados tem $n - k$ graus de liberdade associados a ela e a média quadrada resultante é o nosso estimador imparcial s^2 de σ^2.

2. Subtraia a soma do erro puro dos quadrados da soma dos erros dos quadrados SQE e, com isso, obtenha a soma dos quadrados devida à falta de ajuste. Os graus de liberdade para a falta de ajuste também são obtidos simplesmente subtraindo-se $(n - 2) - (n - k) = k - 2$.

Tabela 11.3 Análise de variância para testar a linearidade da regressão

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	f calculado
Erro de regressão	SQR	1	SQR	$\frac{SQR}{s^2}$
	SQE	$n-2$		
Falta de ajuste	$SQE - SQE$ (puro)	$k-2$	$\frac{SQE - SQE\ (\text{puro})}{k-2}$	$\frac{SQE - SQE\ (\text{puro})}{s^2(k-2)}$
Erro puro	SQE (puro)	$n-k$	$s^2 = \frac{SQE\ (\text{puro})}{n-k}$	
Total	SQT	$n-1$		

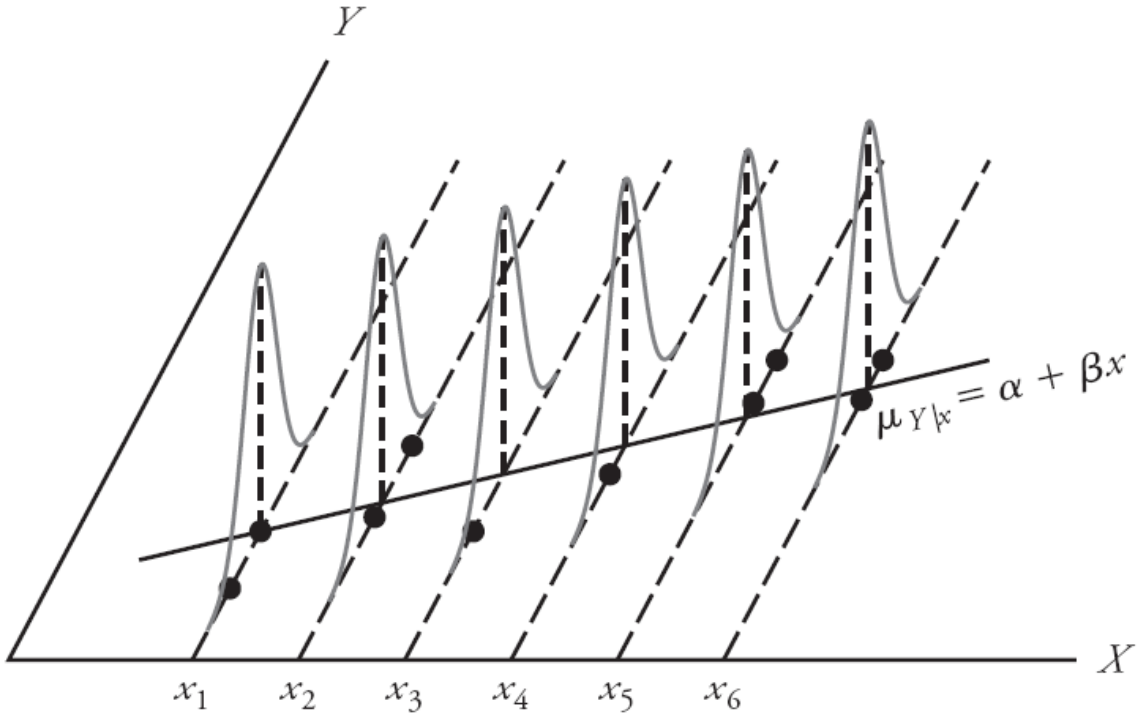


Figura 11.16 Modelo linear correto sem o componente da falta de ajuste.

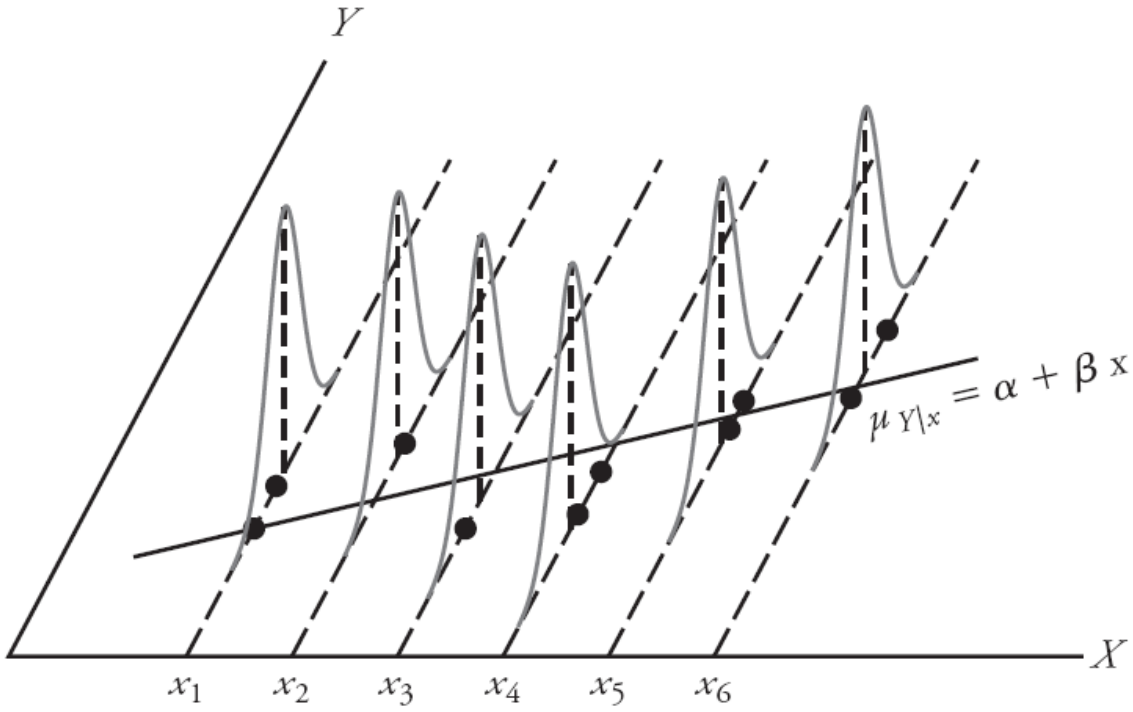


Figura 11.17 Modelo linear incorreto com o componente da falta de ajuste.

Tabela 11.4 Dados para o Exemplo 11.8

y(%)	x(°C)	y(%)	x(°C)
77,4	150	88,9	250
76,7	150	89,2	250
78,2	150	89,7	250
84,1	200	94,8	300
84,5	200	94,7	300
83,7	200	95,9	300

Tabela 11.5 Análise de variância dos dados de temperatura e rendimento

Fonte da variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	*f* calculado	Valores *P*
Regressão	590,2507	1	590,2507	1531,58	< 0,0001
Erro	3,8660	10			
Falta de ajuste	1,2060	2	0,6030	1,81	0,2241
Erro puro	2,6600	8	0,3325		
Total	531,1167	11			

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	510,4566667	170,1522222	511,74	<,0001
Error	8	2,6600000	0,3325000		
Corrected Total	11	513,1166667			

R-Square	Coeff Var	Root MSE	yield Mean
0,994816	0,666751	0,576628	86,48333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temperature	1	509,2506667	509,2506667	1531,58	<,0001
LOF	2	1,2060000	0,6030000	1,81	0,2241

Figura 11.18 Impressão *SAS*, mostrando a análise dos dados do Exemplo 11.8.

11.10 Gráficos dos dados e transformações

Tabela 11.6 Algumas transformações úteis para linearização.

Forma funcional relacionando y a x	**Transformação apropriada**	**Forma da regressão linear simples**
Exponencial: $y = \alpha e^{\beta x}$	$y^* = \ln y$	Regresse y^* contra x
Potência: $y = \alpha x^{\beta}$	$y^* = \log y; \quad x^* = \log x$	Regresse y^* contra x^*
Recíproca: $y = \alpha + \beta\left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Regresse y contra x^*
Função hiperbólica: $y = \frac{x}{\alpha + \beta x}$	$y^* = \frac{1}{y}; \quad x^* = \frac{1}{x}$	Regresse y^* contra x^*

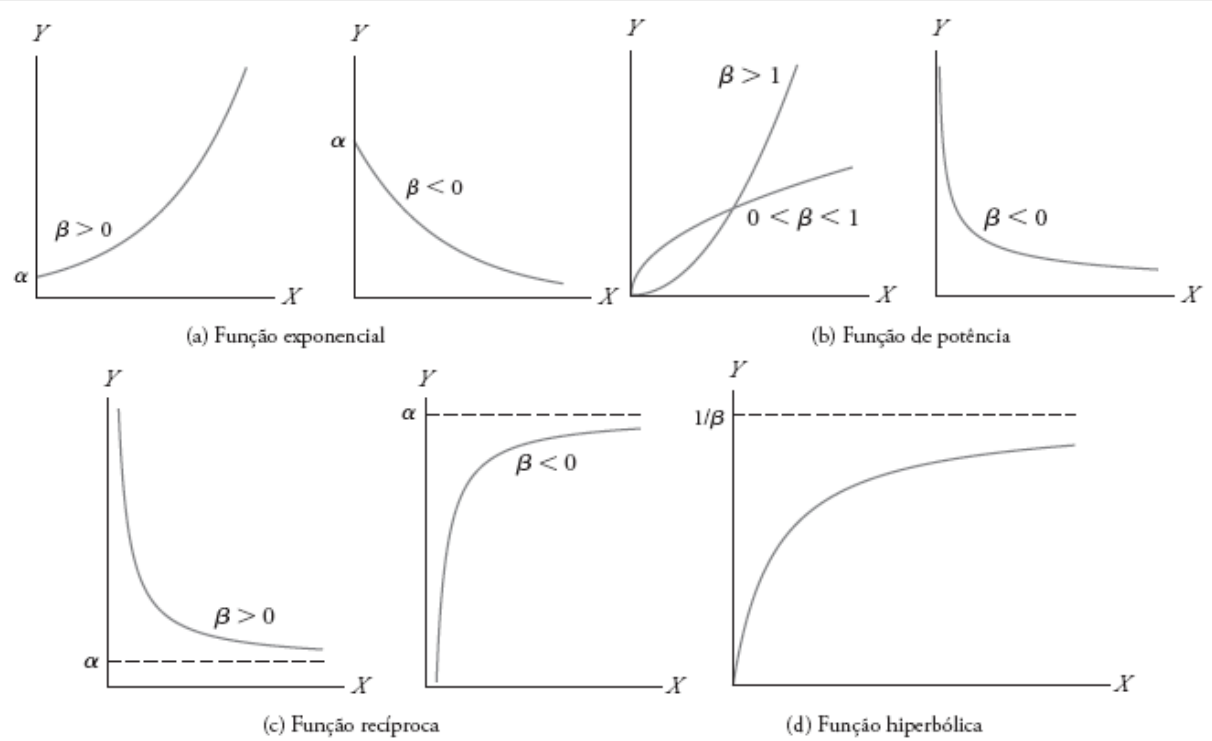


Figura 11.19 Diagramas que representam as funções listadas na Tabela 11.6.

Tabela 11.7 Dados para o Exemplo 11.9

V (cm^3)	50	60	70	90	100
P (kg/cm^2)	64,7	51,3	40,5	25,9	7,8

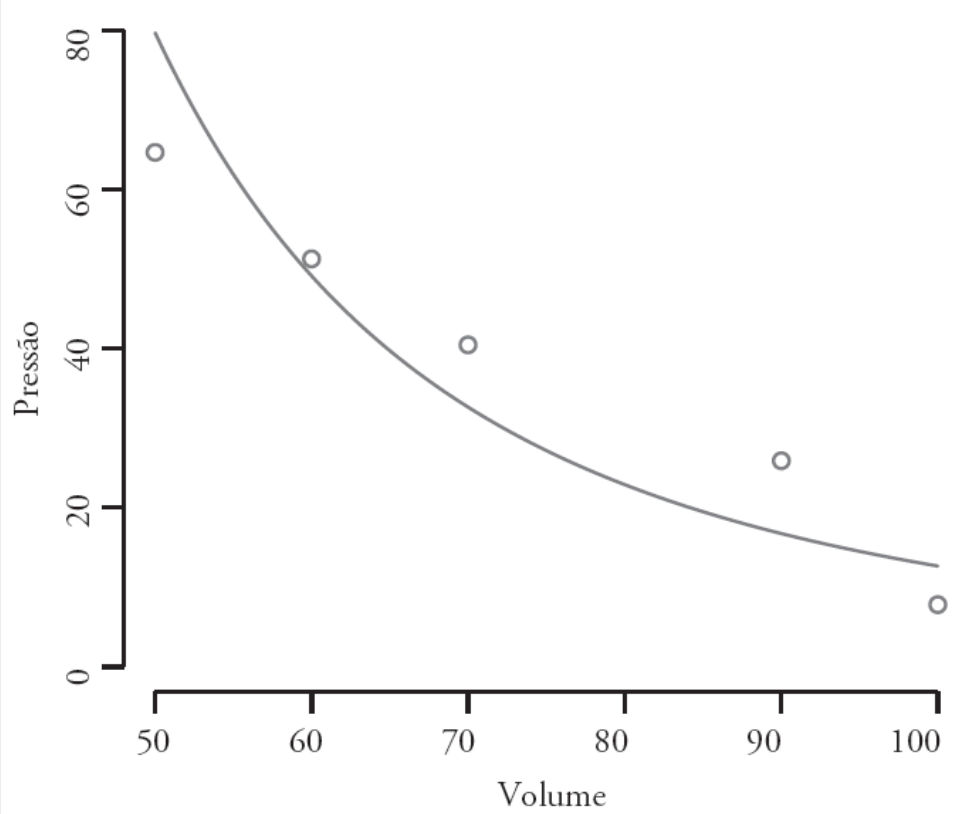


Figura 11.20 Dados de pressão e volume e a regressão ajustada.

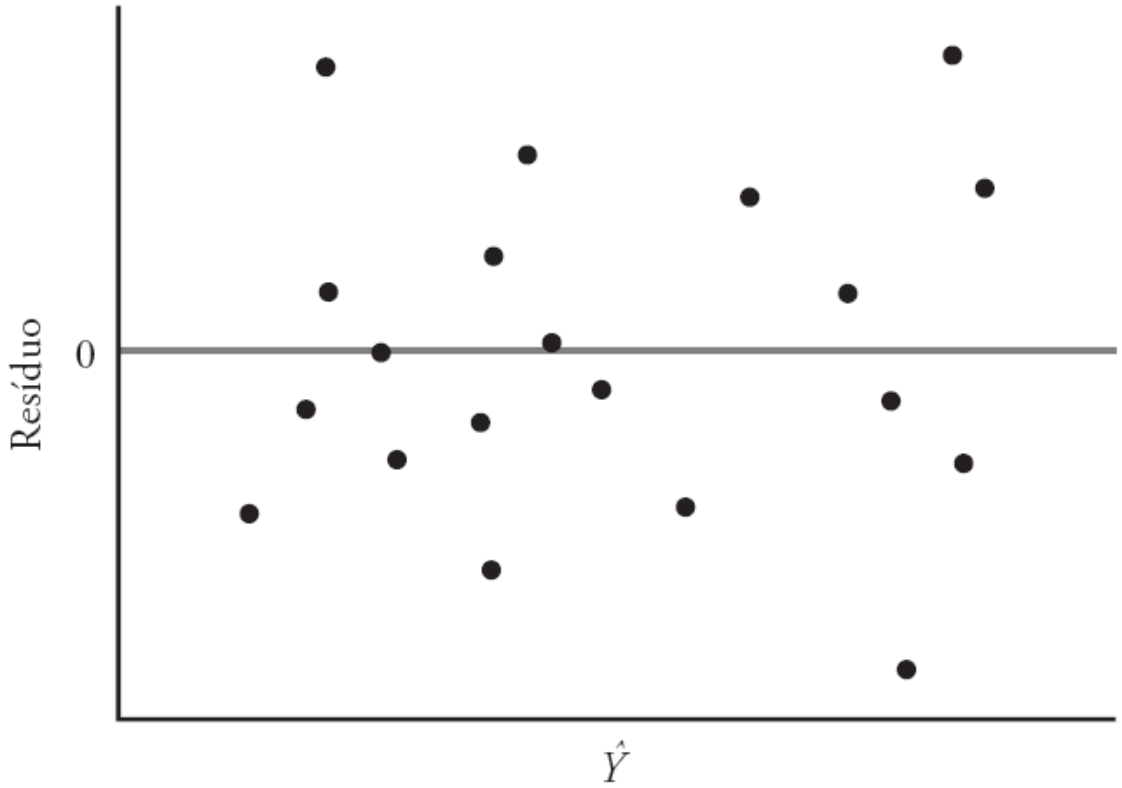


Figura 11.21 Representação ideal dos resíduos.

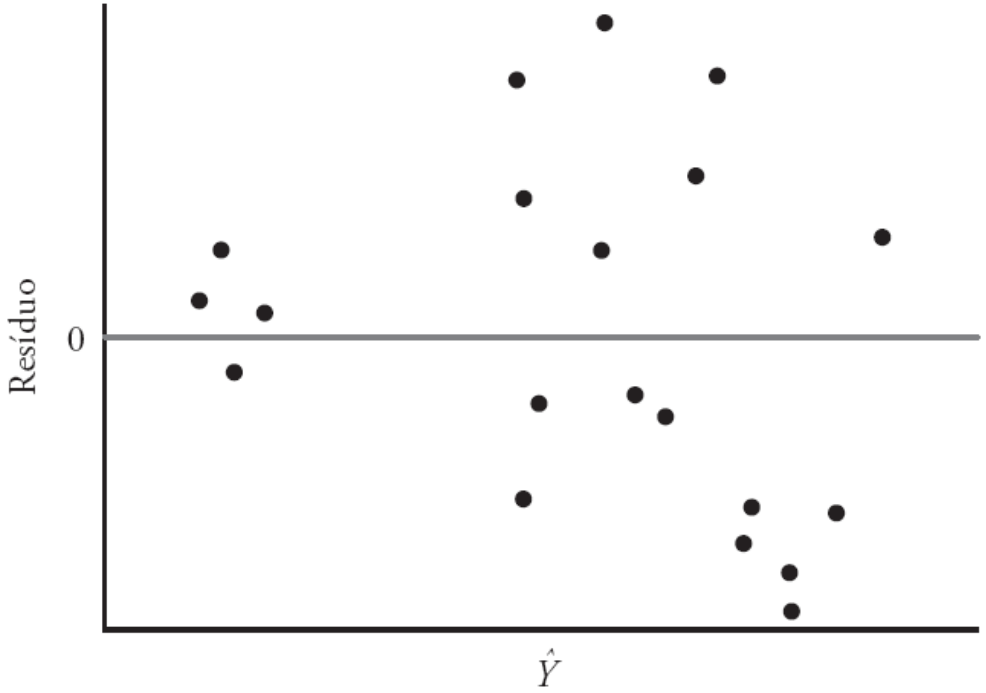


Figura 11.22 Representação dos resíduos que mostram variância do erro heterogênea.

11.11 Estudo de caso de regressão linear simples

Tabela 11.8 Densidade e rigidez para os 30 compensados

Densidade, x	Rigidez, y	Densidade, x	Rigidez, y
9,50	14.814,00	8,40	17.502,00
9,80	14.007,00	11,00	19.443,00
8,30	7.573,00	9,90	14.191,00
8,60	9.714,00	6,40	8.076,00
7,00	5.304,00	8,20	10.728,00
17,40	43.243,00	15,00	25.319,00
15,20	28.028,00	16,40	41.792,00
16,70	49.499,00	15,40	25.312,00
15,00	26.222,00	14,50	22.148,00
14,80	26.751,00	13,60	18.036,00
25,60	96.305,00	23,40	104.170,00
24,40	72.594,00	23,30	49.512,00
19,50	32.207,00	21,20	48.218,00
22,80	70.453,00	21,70	47.661,00
19,80	38.138,00	21,30	53.045,00

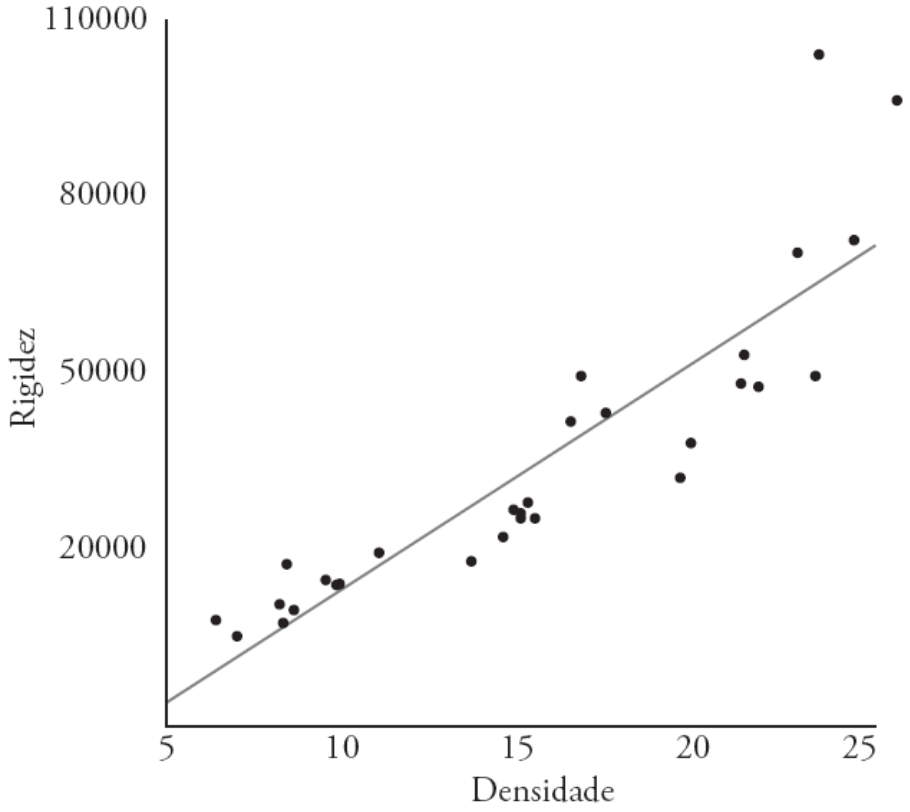


Figura 11.23 Gráfico de dispersão dos dados da densidade da madeira.

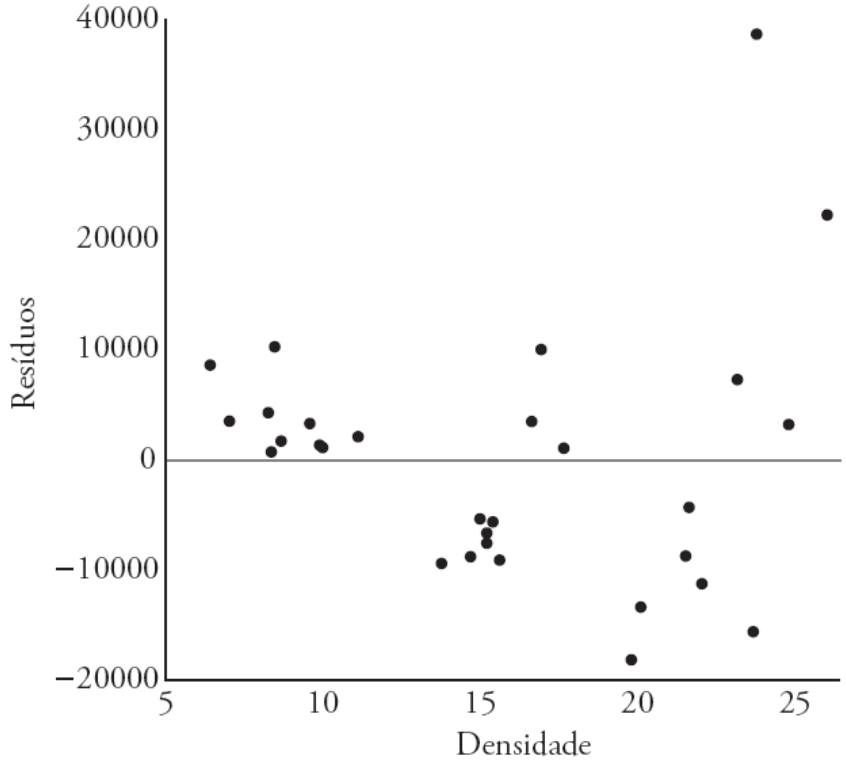


Figura 11.24 Representação dos resíduos para os dados da densidade da madeira.

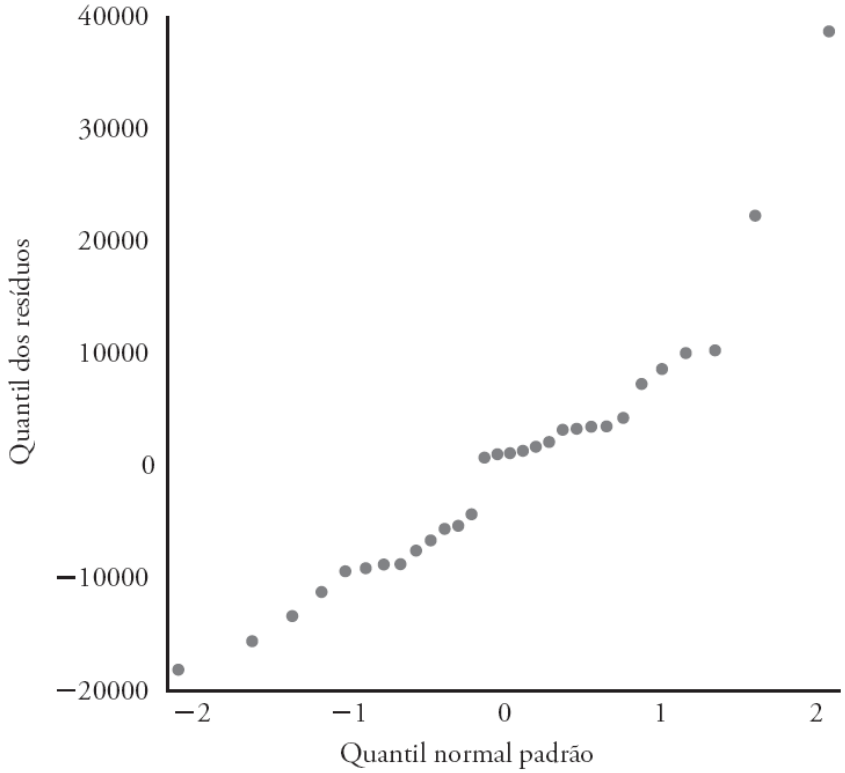


Figura 11.25 Gráfico de probabilidade normal de resíduos para os dados de densidade da madeira.

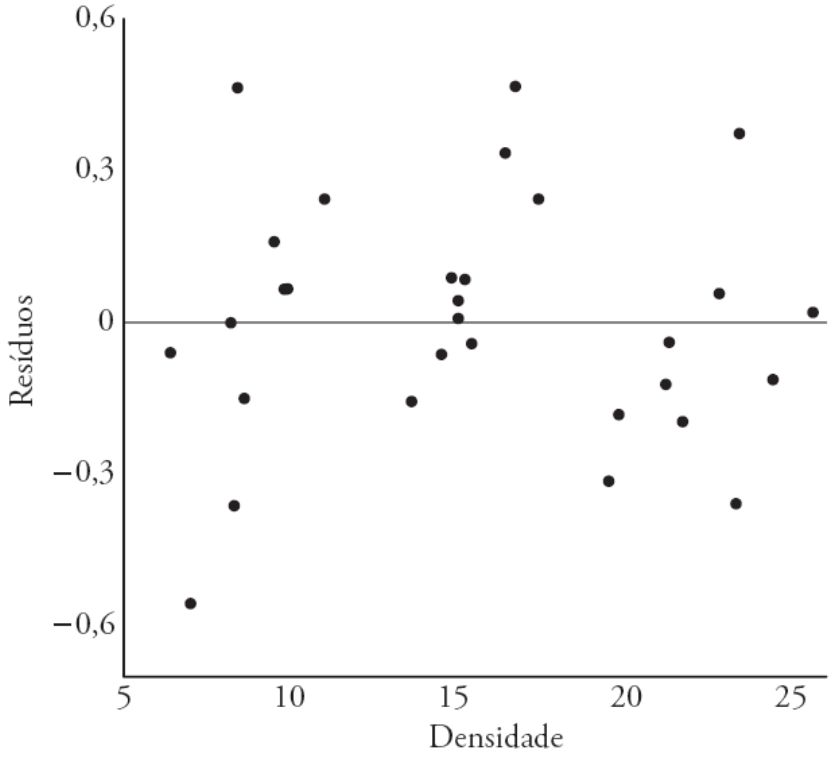


Figura 11.26 Gráfico residual do modelo que usa a transformação log para os dados de densidade da madeira.

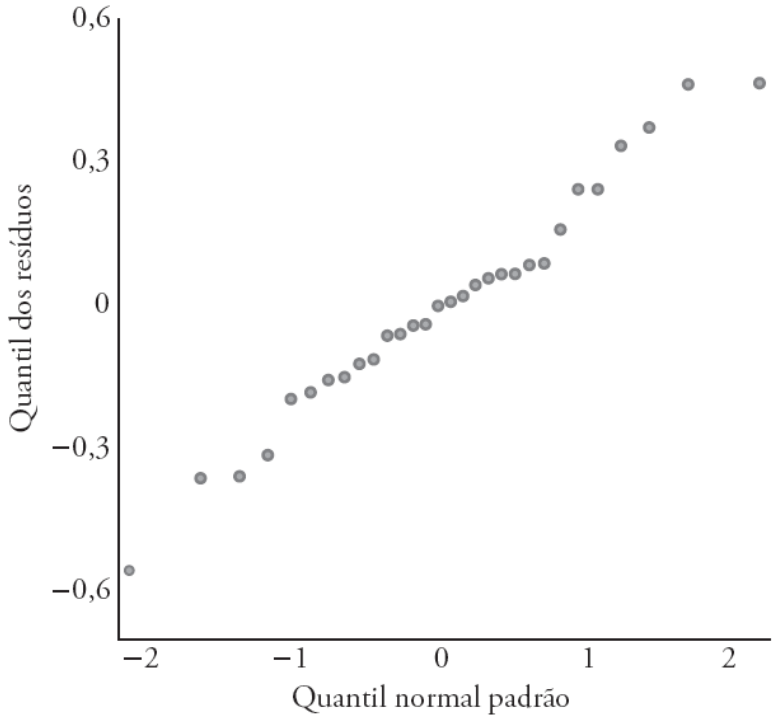


Figura 11.27 Gráfico de probabilidade normal dos resíduos do modelo-que usa a transformação log para os dados da densidade da madeira.

11.12 Correlação

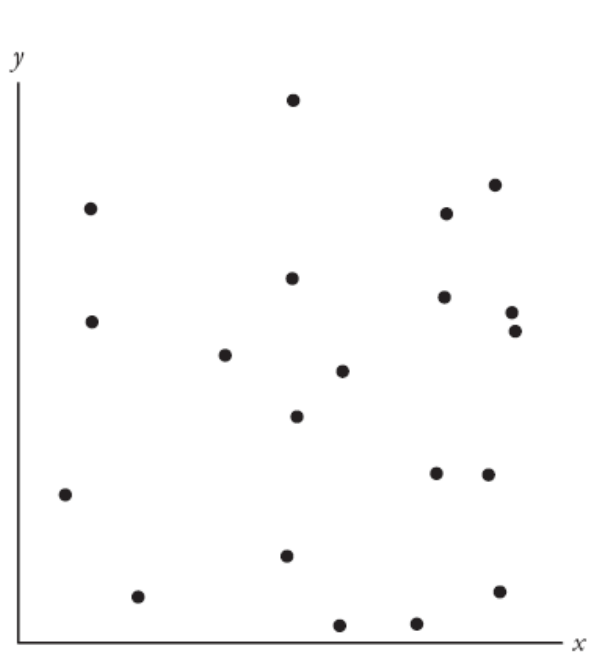
Coeficiente de correlação

A medida ρ da associação linear entre duas variáveis X e Y é estimada pelo *coeficiente de correlação amostral* r, onde

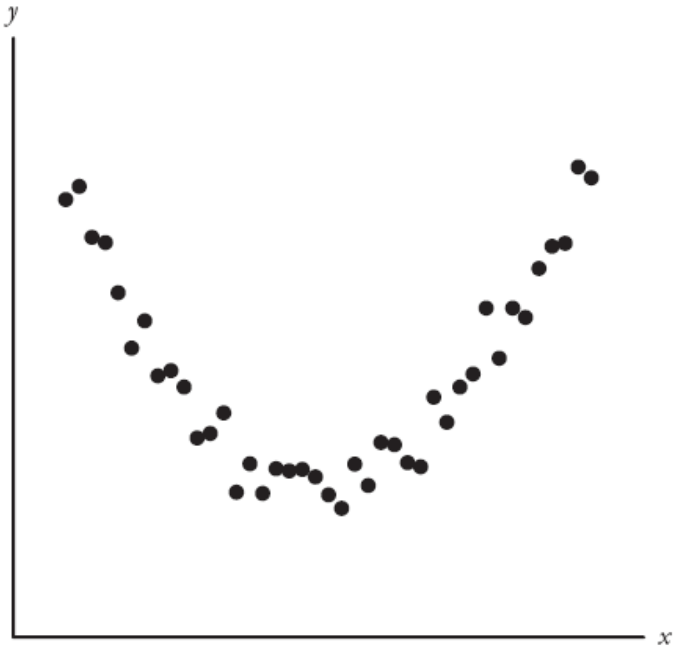
$$r = b\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}}.$$

Tabela 11.9 Dados de 29 pinheiros para o Exemplo 11.10

Gravidade específica, x (g/cm³)	Módulo de ruptura, y (kPa)	Gravidade específica, x (g/cm³)	Módulo de ruptura, y (kPa)
0,414	29.186	0,581	85.156
0,383	29.266	0,557	69.571
0,399	26.215	0,550	84.160
0,402	30.162	0,531	73.466
0,442	38.867	0,550	78.610
0,422	37.831	0,556	67.657
0,466	44.576	0,523	74.017
0,500	46.097	0,602	87.291
0,514	59.698	0,569	86.836
0,530	67.705	0,544	82.540
0,569	66.088	0,557	81.699
0,558	78.486	0,530	82.096
0,577	89.869	0,547	75.657
0,572	77.369	0,585	80.490
0,548	67.095		



(a) Sem associação



(b) Relação causal

Figura 11.28 Diagrama de dispersão que mostra correlação zero.