PRO2271 ESTATÍSTICA

8. Correlação e Regressão Linear

Correlação Linear

Objetivo:

- medir o grau de associação entre variáveis quantitativas

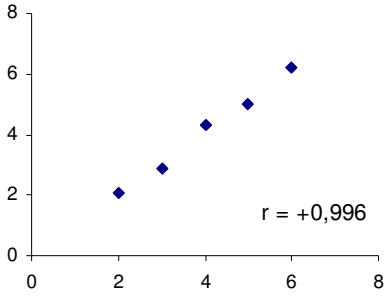
Índice de Correlação Linear de Pearson

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} \qquad (-1 \le r \le +1)$$

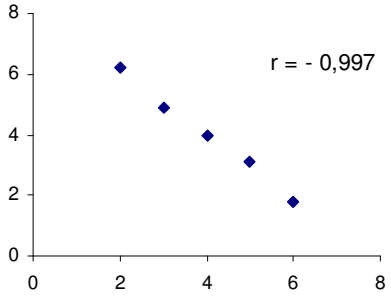
onde: $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

$$S_{xx} = \sum (x_i - \bar{x})^2 \qquad S_{yy} = \sum (y_i - \bar{y})^2$$

Graus de Correlação Linear

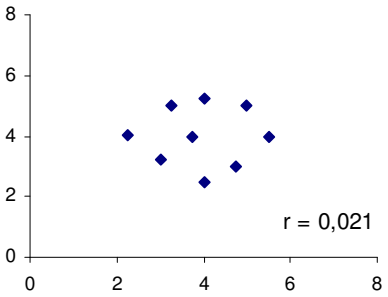


Correlação positiva forte

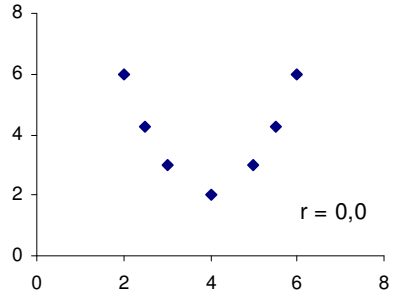


Correlação negativa forte

Graus de Correlação Linear (cont.)



Não há correlação



Correlação não linear

Correlação Linear

Grau de Correlação Linear:

- Forte, se $|r| \geq 0{,}9$
- Moderada, se $0{,}7 \leq |r| < 0{,}9$
- Fraca, se $|r| < 0{,}7$

Significância estatística (tamanho da amostra)

Correlação: associativa ou causa-efeito?

Soma de Quadrados

Índice de Correlação Linear: $r = \dfrac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$

onde:

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

Exemplo

Determine o índice de correlação linear entre x e y.

i	x_i	y_i	$x_i\ y_i$
1	2	12,6	25,2
2	4	13,0	52,0
3	6	15,9	95,4
4	8	18,6	148,8
5	10	19,5	195,0
T	30	79,6	516,4
Q	220	1306,78	

Sxy = 38,8

Sxx = 40,0

Syy = 39,548

r = 0,9755

Excel – Matriz de Correlação

X	*Y*	*Z*
2	15	10
8	35	24
11	40	30
10	150	35
8	90	25
4	60	17
2	100	20
2	15	9
9	30	24
8	90	28

	X	*Y*	*Z*
X	1		
Y	0,3074	1	
Z	0,8931	0,6685	1

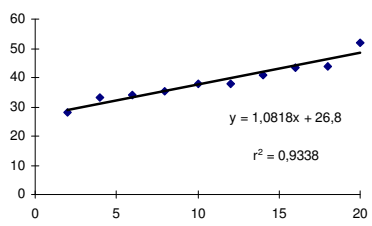
Análise de Regressão

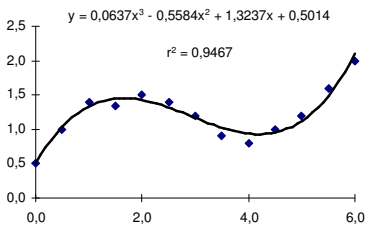
O objetivo da Análise de Regressão é identificar relações de causa e efeito entre variáveis independentes Xi e uma variável de resposta Y.

Identificadas as variáveis que apresentam correlação forte com Y, determina-se a função matemática que melhor expressa esta relação.

Esta função pode ser linear ou não linear, simples ou com múltiplas variáveis. O primeiro modelo a ser estudado é denominado regressão linear simples.

Exemplos





No primeiro exemplo, tem-se uma função linear e, no segundo, um polinômio de grau 3. Outras funções também podem ser utilizadas como a exponencial, potência, logarítmica etc.

Regressão Linear Simples

Modelo: $Y = \alpha + \beta x + \varepsilon \qquad \varepsilon \sim N(0, \sigma_R)$

Estatísticas: $\hat{y}_i = a + b x_i$

$$b = \frac{S_{xy}}{S_{xx}} \qquad a = \bar{y} - b\bar{x}$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum (x_i - \bar{x})^2 \qquad S_{yy} = \sum (y_i - \bar{y})^2$$

As expressão acima são obtidas pelo método dos mínimos quadrados.

Exemplo RL

A partir dos dados abaixo, determine a equação da ***reta dos mínimos quadrados***.

i	x_i	y_i	$x_i y_i$
1	2	12,6	25,2
2	4	13,0	52,0
3	6	15,9	95,4
4	8	18,6	148,8
5	10	19,5	195,0
T	30	79,6	516,4
Q	220	1306,78	

Sxy = 38,8

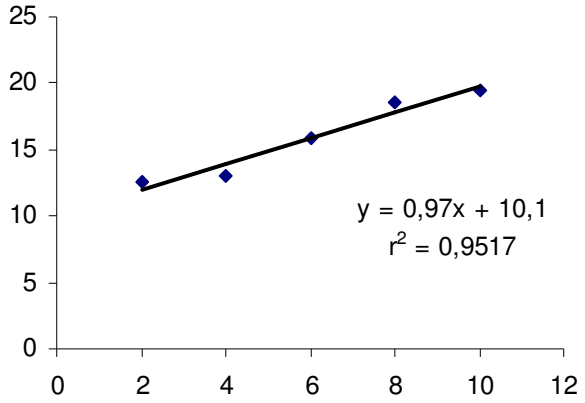
Sxx = 40,0

Syy = 39,548

b = 0,97

a = 10,1

Exemplo RL (cont.)



Variância Residual

Variância Residual:

$$SRQ = \sum (y_i - \hat{y}_i)^2 = S_{yy} - b \cdot S_{xy} \qquad s_R^2 = \frac{SRQ}{n-2} = \frac{S_{yy} - b \cdot S_{xy}}{n-2}$$

Coeficiente de Determinação:

$$STQ = S_{yy} = \sum (y_i - \bar{y})^2 \qquad r^2 = 1 - \frac{SRQ}{STQ} = \frac{S_{xy}^2}{S_{xx} \cdot S_{xy}}$$

O coeficiente $\mathbf{r^2}$ permite avaliar o grau de ajuste da reta. Valores maiores que 0,8 indicam alto grau de determinação

Exemplo RL (cont.)

A partir dos dados abaixo, determine a variância residual e o coeficiente de determinação da ***reta dos mínimos quadrados***.

i	x_i	y_i	y_i^	$(y_i - y_i\hat{})^2$
1	2	12,6	12,04	0,3136
2	4	13,0	13,98	0,9604
3	6	15,9	15,92	0,0004
4	8	18,6	17,86	0,5476
5	10	19,5	19,80	0,0900
				1,912

SRQ = 1,912
STQ = 39,548

s_R^2 = 0,6373
s_R = 0,798

r^2 = 0,9517
r = 0,976

Exemplo RL (cont.)

RESUMO DOS RESULTADOS

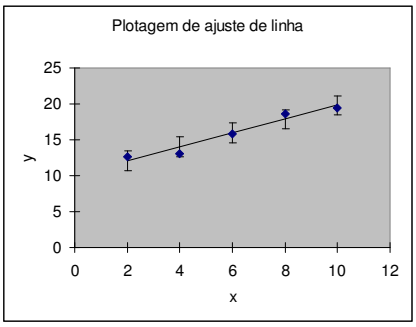
Estatística de regressão	
R múltiplo	0,976
R-Quadrado	0,9517
R-quadrado ajustado	0,9355
Erro padrão	0,798
Observações	5

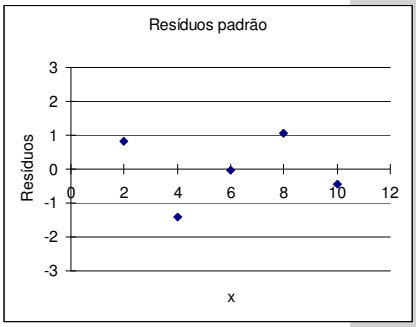
ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	1	37,636	37,636	59,05	0,00458
Resíduo	3	1,912	0,6373		
Total	4	39,548			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*Inferior 95%*	*Superior 95%*
Interseção	10,10	0,84	12,063	0,00123	7,44	12,76
Variável X 1	0,97	0,13	7,685	0,00458	0,57	1,37

Gráficos





Avaliação da Adequação do Modelo

Análise do diagrama de dispersão e dos resíduos
- relação não linear
- pontos extremos
- dispersão variável
- variáveis omitidas etc

Teste estatísticos
- testes sobre os coeficientes
- análise de variância para validação do modelo
- testes de normalidade dos resíduos

Transformação de Variáveis

É possível, a partir da transformação de variáveis, utilizar a regressão linear para correlacionar duas ou mais variáveis. Abaixo, algumas transformações usuais utilizadas para correlacionar duas variáveis x e y.

Função	Transformação	Forma Linear
$Y=\alpha e^{\beta x}$	y'=ln(y)	$Y'=\ln(\alpha)+\beta x$
$Y=\alpha x^{\beta}$	y'=ln(y) e x'=ln(x)	$Y'=\ln(\alpha)+\beta x'$
$Y=\alpha+\beta\log(x)$	x'=ln(x)	$Y'=\alpha+\beta x'$
$Y=\alpha+\beta(1/x)$	x'=(1/x)	$Y'=\alpha+\beta x'$

Inferências em RL

Intervalos de confiança para o coeficiente angular

$$b \pm t_{n-2,\alpha/2} \cdot \frac{s_R}{\sqrt{S_{xx}}}$$ ——— s_b

Intervalos de confiança para a média de Y em x

$$(a+bx) \pm t_{n-2,\alpha/2} \cdot s_R \cdot \sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}}$$

Intervalos de previsão

$$(a+bx) \pm t_{n-2,\alpha/2} \cdot s_R \cdot \sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}}$$

A.V. na Regressão

Regressão Linear Simples

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrados Médios	F
Modelo	$SQRegr = \sum(\hat{y}_i - \bar{y})^2 = b_1 S_{xy}$	1	$SQRegr$	$\frac{SQRegr}{SRQ/(n-2)}$
Residual	$SQR = \sum(y_i - \hat{y}_i)^2 = S_{yy} - b_1 S_{xy}$	$n-1$	$s_R^2 = \frac{SRQ}{(n-2)}$	
Total	$SQR = \sum(y_i - \bar{y})^2 = S_{yy}$	$n-2$		

Regressão Múltipla

Modelos mais complexos, com maior número de variáveis, podem ser construídos.

Estes modelos envolvem dificuldades tanto de coleta de dados quanto de conhecimento técnico dos analistas.

Para ter sucesso nesta empreitada, deve-se dispor de um bom software profissional de estatística.

Os conceitos apresentados aqui são um ponto de partida.

Exercício 4 - Minitab

Data Display

Row	x	y
1	20	15,0
2	40	18,6
3	60	27,1
4	90	26,4
5	120	38,0
6	150	34,2
7	180	44,6

Regression Analysis: y versus x

$t_{5;5\%} = 2{,}015$

The regression equation is
y = 13,1 + 0,170 x

IC(β; 90%)
0,12≤β≤0,22

Predictor	Coef	SE Coef	T	P
Constant	13,064	2,678	4,88	0,005
x	0,17039	0,02459	6,93	0,001

S = 3,54392 R-Sq = 90,6% R-Sq(adj) = 88,7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	603,02	603,02	48,01	0,001
Residual Error	5	62,80	12,56		
Total	6	665,81			

Predicted Values for New Observations

x = 200

New Obs	Fit	SE Fit	90% CI	90% PI
1	47,14	2,92	(41,25; 53,03)	(37,88; 56,40)

Exercício 5 - Minitab

Regression Analysis: y versus x

The regression equation is
y = - 11,4 + 0,250 x

Predictor	Coef	SE Coef	T	P
Constant	-11,360	3,925	-2,89	0,034
x	0,25021	0,04813	5,20	0,003

S = 0,560738 R-Sq = 84,4% R-Sq(adj) = 81,3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	8,4964	8,4964	27,02	0,003
Residual Error	5	1,5721	0,3144		
Total	6	10,0686			

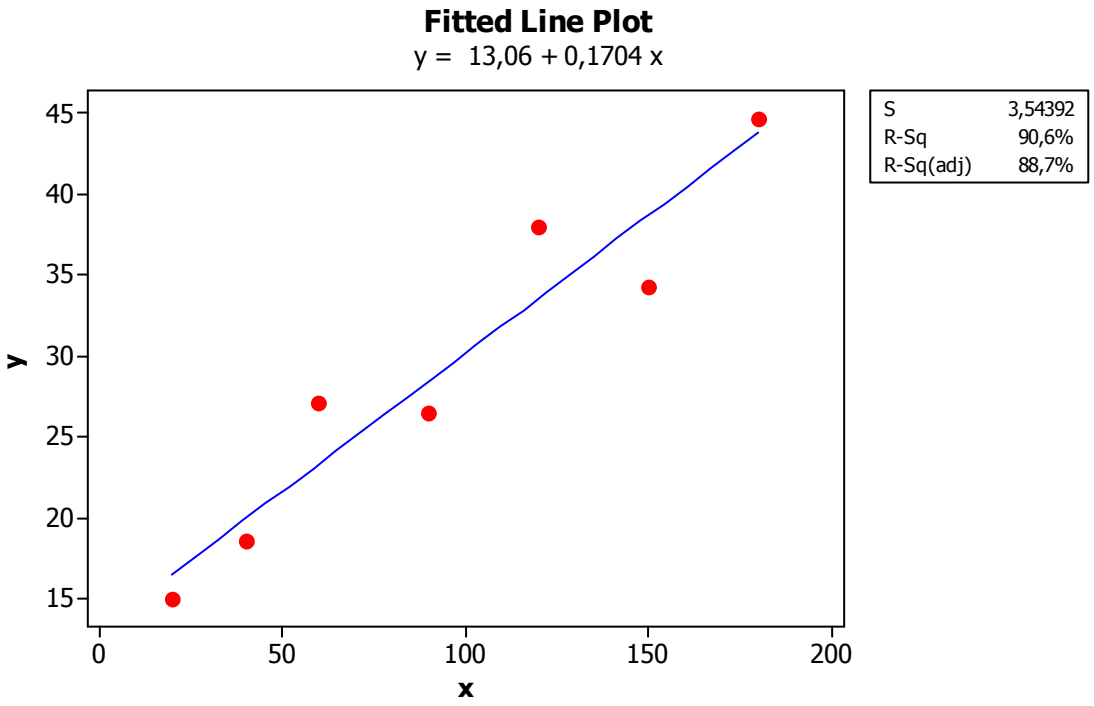
EXERCÍCIOS

1) Os dados abaixo representam as notas finais de dez alunos nas disciplinas de Matemática (x) e Física (y). Verifique se há correlação linear entre as variáveis:

x	y
51	74
68	70
72	88
97	93
55	67
95	99
20	33
91	91
74	80
80	86

2) Sete lotes, de diferentes quantidades, foram produzidos em um centro de produção, tendo sido anotados o tempo total de produção de cada lote, conforme dados da tabela abaixo. Determine a equação da reta dos mínimos quadrados e uma estimativa do tempo de produção para um lote com 200 unidades.

x (qte) :	20	40	60	90	120	150	180
y (min) :	15,0	18,6	27,1	26,4	38,0	34,2	44,6



i	x	y	x^2	y^2	x y
1	20	15,0			
2	40	18,6			
3	60	27,1			
4	90	26,4			
5	120	38,0			
6	150	34,2			
7	180	44,6			

$$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right) \cdot \left(\sum y_i\right)}{n} =$$

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} =$$

$$S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} =$$

$$b = \frac{S_{xy}}{S_{xx}} = \qquad a = \bar{y} - b\bar{x} =$$

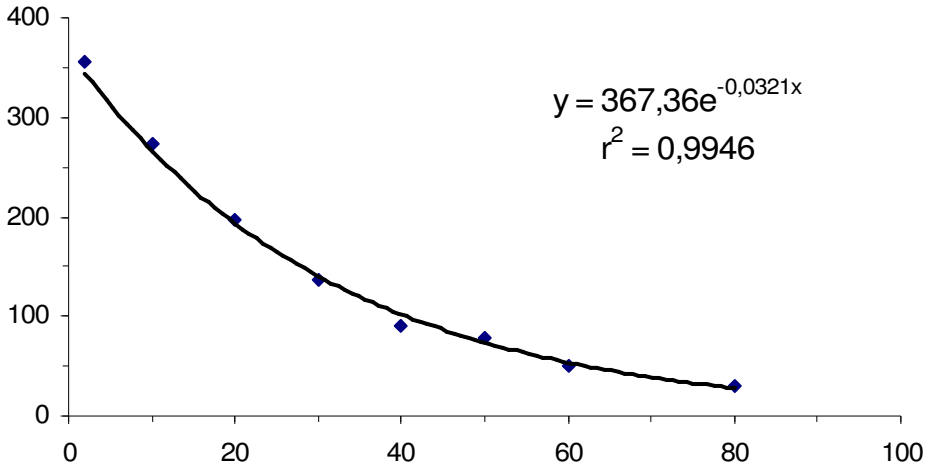
$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} =$$

$$s = \sqrt{\frac{S_{yy} - bS_{xy}}{n-2}} =$$

3) A relação entre duas variáveis pode ser representada pela seguinte equação: $Y = \alpha\, e^{\beta x}$. Para verificar a validade do modelo, foram realizadas experiências em laboratório, que produziram os seguintes resultados :

x :	2	10	20	30	40	50	60	80
y :	356	274	196	137	90	78	51	30

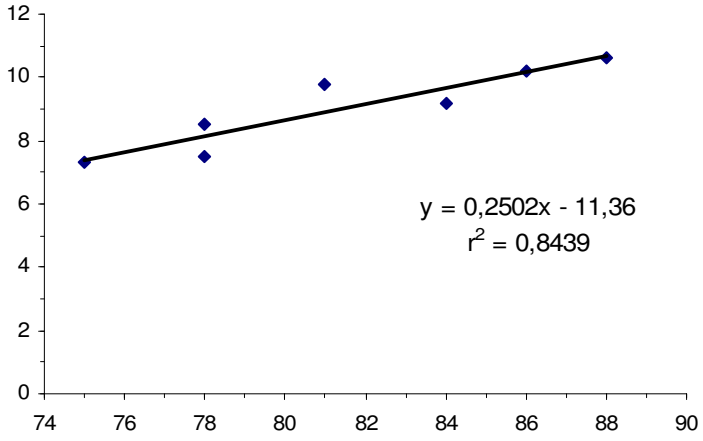
Aplicando a transformação de variáveis e o método dos mínimos quadrados, estime os parâmetros do modelo em questão.



4) (Cont. 2) Em relação ao exercício anterior, determine um intervalo com 90% de confiança para o coeficiente angular da reta e estime o erro de previsão de y para x = 200.

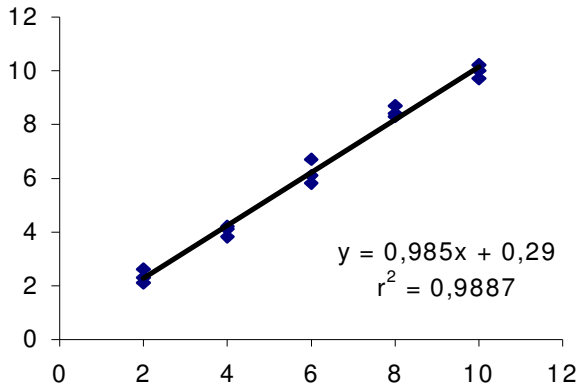
5) Estime os parâmetros de um modelo de regressão linear simples para os dados abaixo e teste a hipótese de que o coeficiente angular seja maior que 0,20.

x	y
75	7,3
78	8,5
78	7,5
81	9,8
84	9,2
86	10,2
88	10,6



6) Os dados foram obtidos em um ensaio de calibração, onde x representa cinco valores de referência e y, três medições para cada padrão. Estime os coeficientes da reta dos mínimos quadrados com 95% de confiança:

x	y
2	2,1
	2,6
	2,3
4	3,8
	4,2
	4,1
6	5,8
	6,1
	6,7
8	8,3
	8,4
	8,7
10	10,0
	10,2
	9,7



RESUMO DOS RESULTADOS

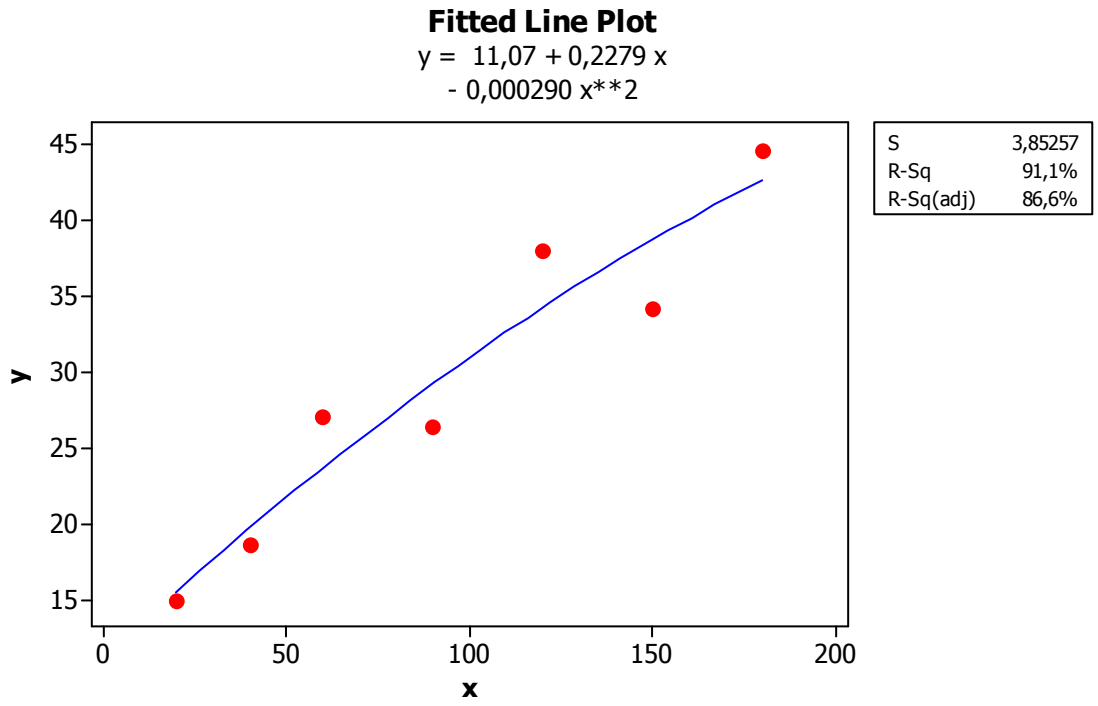
Estatística de regressão	
R múltiplo	0,994
R-Quadrado	0,989
R-Quadrado ajustado	0,988
Erro padrão	0,320
Observações	15

ANOVA

	gl	*SQ*	*MQ*	*F*	*valor-P*
Regressão	1	116,427	116,4270	1135,45	0,000
Resíduo	13	1,333	0,1025		
Total	14	117,760			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*Inferior 95%*	*Superior 95%*
Interseção	0,290	0,194	1,496	0,1586	-0,1289	0,7089
x	0,985	0,029	33,696	0,0000	0,9218	1,0482

7) (Cont. 2) A partir dos dados do exercício 2, determine a curva do segundo grau pelo método dos mínimos quadrados. Há melhoria significativa com a adoção do modelo quadrático em relação ao modelo linear?



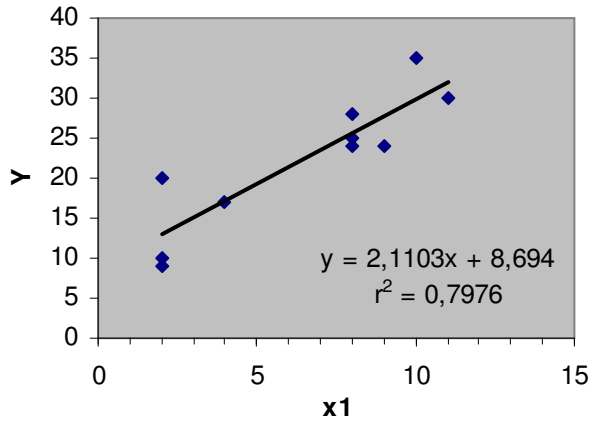
Polynomial Regression Analysis: y versus x

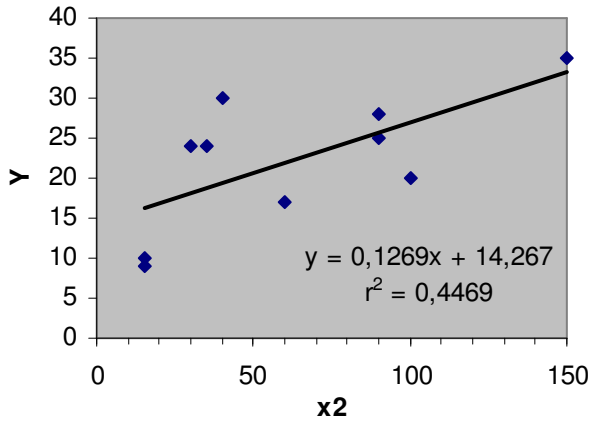
```
The regression equation is
y = 11,07 + 0,2279 x - 0,000290 x**2


S = 3,85257   R-Sq = 91,1%   R-Sq(adj) = 86,6%


Analysis of Variance

Source       DF       SS       MS      F      P
Regression    2  606,445  303,223  20,43  0,008
Error         4   59,369   14,842
Total         6  665,814


Sequential Analysis of Variance

Source     DF       SS      F      P
Linear      1  603,017  48,01  0,001
Quadratic   1    3,428   0,23  0,656
```

8) A tabela abaixo apresenta o tempo de suprimento de uma máquina automática de refrigerantes em relação ao número de caixas supridas e a distância da máquina ao local de estacionamento do veículo. Estime os parâmetros do modelo de regressão linear múltipla que permita prever a variável de respostas y (tempo) em função das variáveis independentes x_1 (número de caixas) e x_2 (distância).

Tabela 8.1 Tempos de suprimento.

i	x_1 (un.)	x_2 (m)	y (min)
1	2	15	10
2	8	35	24
3	11	40	30
4	10	150	35
5	8	90	25
6	4	60	17
7	2	100	20
8	2	15	9
9	9	30	24
10	8	90	28





Regression Analysis: y versus x1; x2

```
The regression equation is
y = 5,55 + 1,79 x1 + 0,0826 x2


Predictor      Coef  SE Coef      T      P
Constant      5,553    1,242   4,47  0,003
x1           1,7943   0,1651  10,87  0,000
x2          0,08261  0,01327   6,23  0,000


S = 1,66590   R-Sq = 96,9%   R-Sq(adj) = 96,0%


Analysis of Variance

Source           DF      SS      MS       F      P
Regression        2  608,17  304,09  109,57  0,000
Residual Error    7   19,43    2,78
Total             9  627,60


Source  DF  Seq SS
x1       1  500,57
x2       1  107,61


Unusual Observations

Obs   x1       y     Fit  SE Fit  Residual  St Resid
  7  2,0  20,000  17,403   1,129     2,597      2,12R

R denotes an observation with a large standardized residual.
```

Stepwise Regression: y versus x1; x2

```
  Alpha-to-Enter: 0,15  Alpha-to-Remove: 0,15


Response is y on 2 predictors, with N = 10


Step             1      2
Constant     8,694  5,553

x1            2,11   1,79
T-Value       5,61  10,87
P-Value      0,001  0,000

x2                  0,083
T-Value              6,23
P-Value             0,000

S             3,98   1,67
R-Sq         79,76  96,90
R-Sq(adj)    77,23  96,02
Mallows C-p   39,8    3,0
```

EXERCÍCIOS PROPOSTOS

1) Refaça os cálculos do exercício 2, adicionando-se o par de valores x = 200 e y = 10. Interprete os resultados obtidos.

2) Construa um diagrama de dispersão e estime os coeficientes do modelo linear.

x :	**10**	**10**	**12**	**12**	**14**	**14**	**16**	**16**	**18**	**18**	**20**	**20**
y :	**17,2**	**16,0**	**21,1**	**23,0**	**27,0**	**24,4**	**26,9**	**28,5**	**29,2**	**28,4**	**29,4**	**28,0**

3) (Cont. 2) Estime os coeficientes de um polinômio de segundo grau para os dados do exercício anterior. Compare os resultados obtidos.

4) Considere-se os dados abaixo. A partir de transformações lineares, verifique qual o modelo de regressão mais adequado para representar a relação entre as variáveis.

x :	10	20	30	40	50
y :	4,9	5,8	7,6	8,1	8,4

5) Verifique, a partir dos dados na tabela abaixo, qual variável dependente apresentar maior grau de correlação com a variável de resposta Y.

x_1	x_2	x_3	y
48,4	0,192	1,82	9,8
56,3	0,212	1,93	10,6
43,6	0,188	1,95	8,9
51,2	0,191	1,90	9,4
55,9	0,186	2,09	9,1

6) Quais são os objetivos e premissas da Análise de Regressão. Explique a lógica e cite três possíveis exemplos de aplicação da técnica em sua área de formação.