Regressão

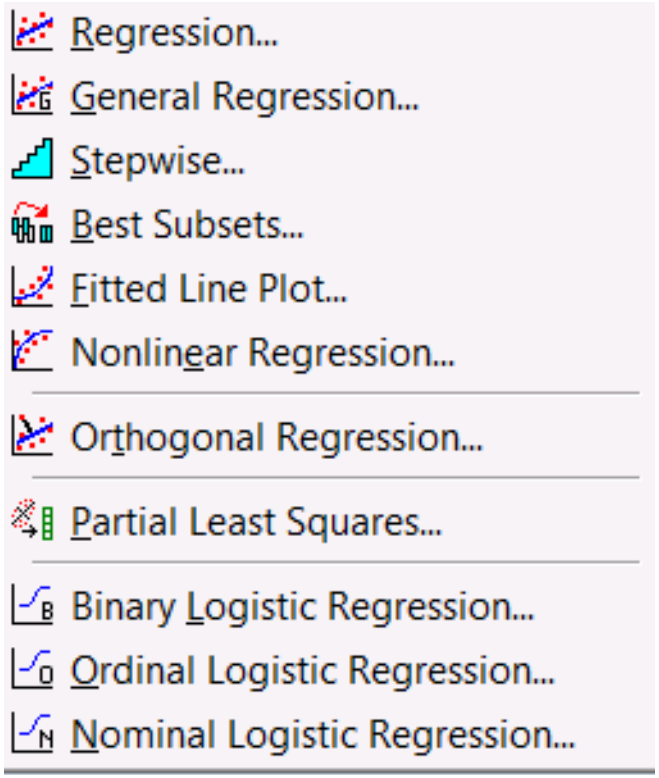
Pedro Paulo Balestrassi

www.pedro.unifei.edu.br

ppbalestrassi@gmail.com

35-36291161 / 999012304 (cel)

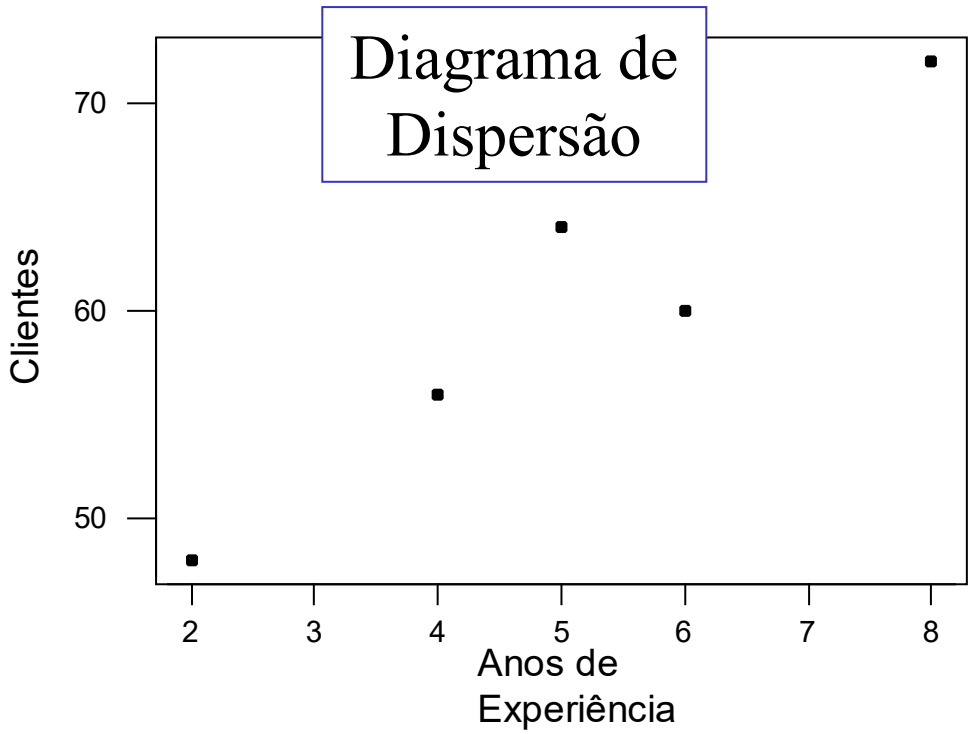
Pratique:
https://pedro.unifei.edu.br/quiz/Regressao



Coeficiente de Correlação

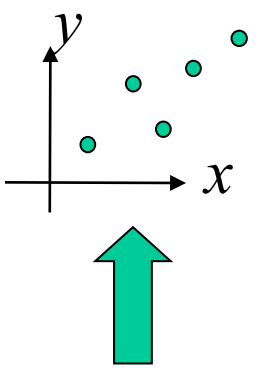
Ex.: Suponha que o nosso desejo seja o de quantificar a associabilidade entre duas variáveis relacionadas a cinco agentes de uma seguradora.
Assim, temos:
$X \equiv$ Anos de experiência do agente.
$Y \equiv$ Número de clientes do agente.

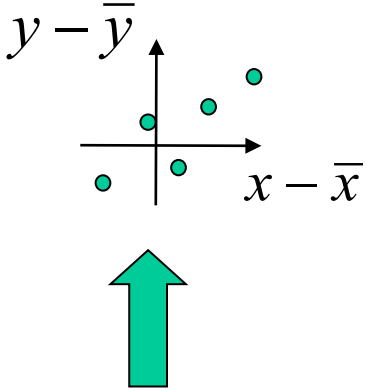
Agente	x	y
A	2	48
B	4	56
C	5	64
D	6	60
E	8	72

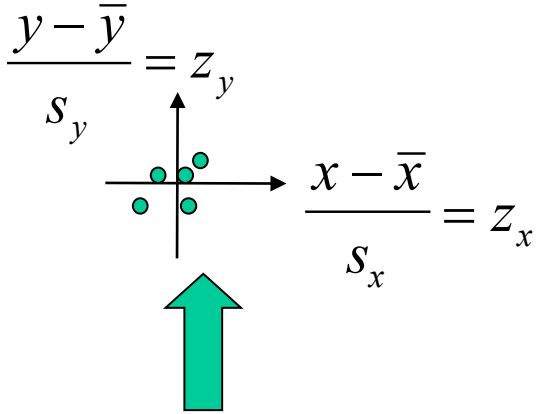


(x, y) é um par aleatório
– Dados emparelhados

r=Correlação de Pearson







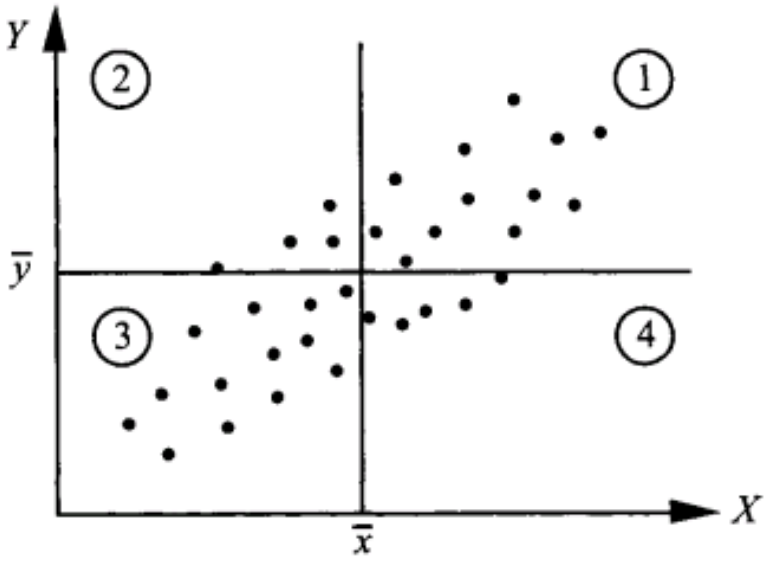
Série de dados originais (x e y) são valores quantitativos.

O conjunto de pontos é deslocado, tendo agora como centro, os valores médios.

A escala de x e y é agora padronizada. Isso torna os valores independente da sua unidade.

$$r = \text{Corr}\,(X,Y) = \frac{1}{n}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$

Quadrantes e Correlação



$$r = \text{Corr}\,(X, Y) = \frac{1}{n}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$

Quadrant	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
1	+	+	+
2	+	−	−
3	−	−	+
4	−	+	−

Coeficiente de Correlação

Agente	x	y	$x - \bar{x}$	$y - \bar{y}$	z_x	z_y	$z_x \cdot z_y$
A	2	48	-3	-12	-1.5	-1.5	2,25
B	4	56	-1	-4	-0.5	-0.5	0,25
C	5	64	0	4	0	0.5	0
D	6	60	1	0	0.5	0	0
E	8	72	3	12	1.5	1.5	2,25
Total	25	300	0	0	0	0	4,75

$\bar{x} = 5 \quad \bar{y} = 60$

$S_x = 2 \quad S_y = 8$

$$r = \text{Correlação}\,(X, Y) = \frac{4,75}{5} = 0,95 = 95\%$$

P_value p/ Correlação

$$r = \text{Corr}\,(X,Y) = \frac{1}{n}\sum_{i=1}^{n} z_{x_i} z_{y_i} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$r = \frac{1}{n}\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} = \frac{\text{Covariância}\,(X,Y)}{s_x \cdot s_y} \quad -1 \le r \le 1$$

A correlação apresentada aqui é linear. Existem outros tipos de correlação!

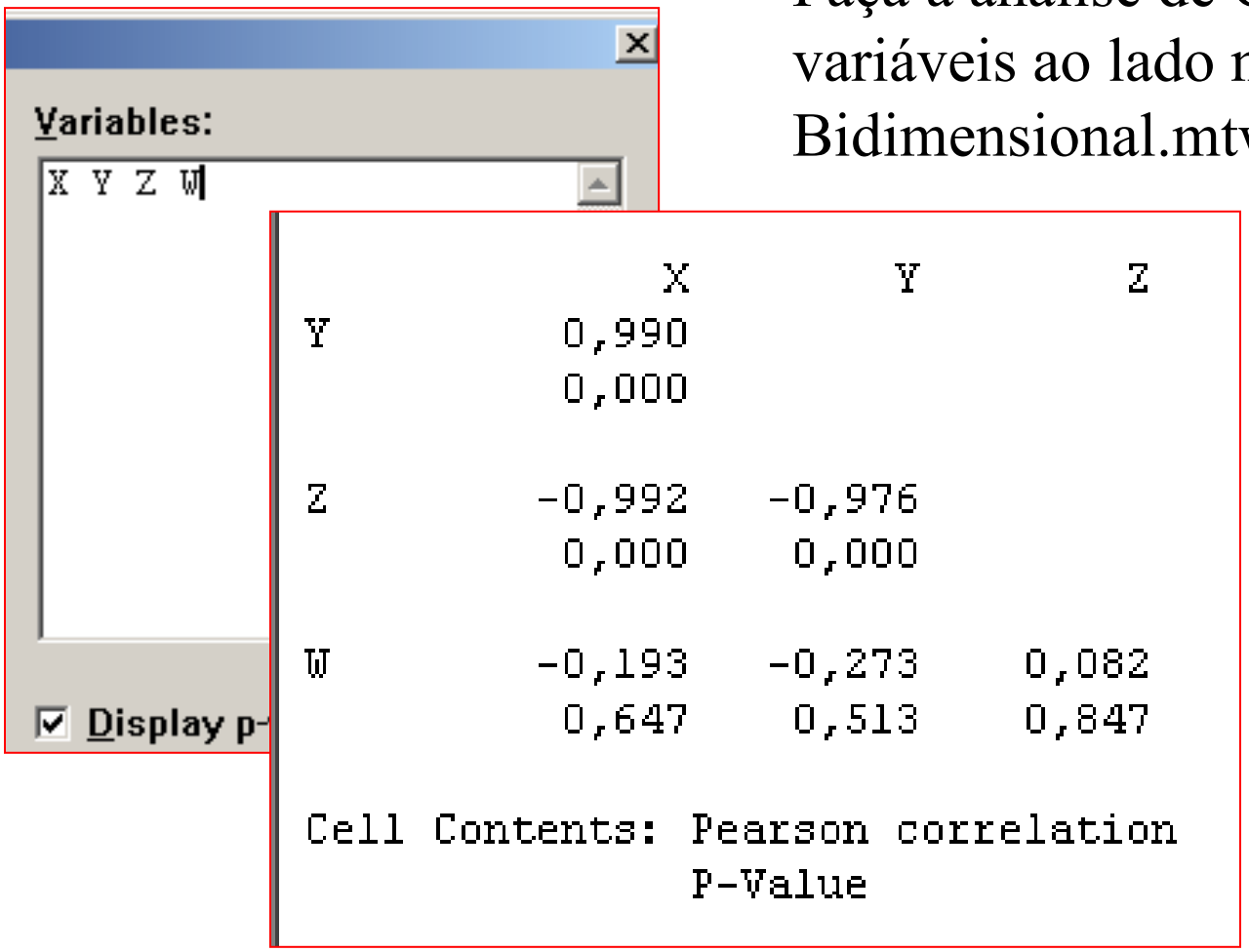
Agente	x	y
A	2	48
B	4	56
C	5	64
D	6	60
E	8	72

Ex.: Cálculo da correlação da tabela ao lado

Pearson correlation of Anos Exp and Clientes = 0,950

P-Value = 0,013 → Forte Correlação pois P-Value <0,05

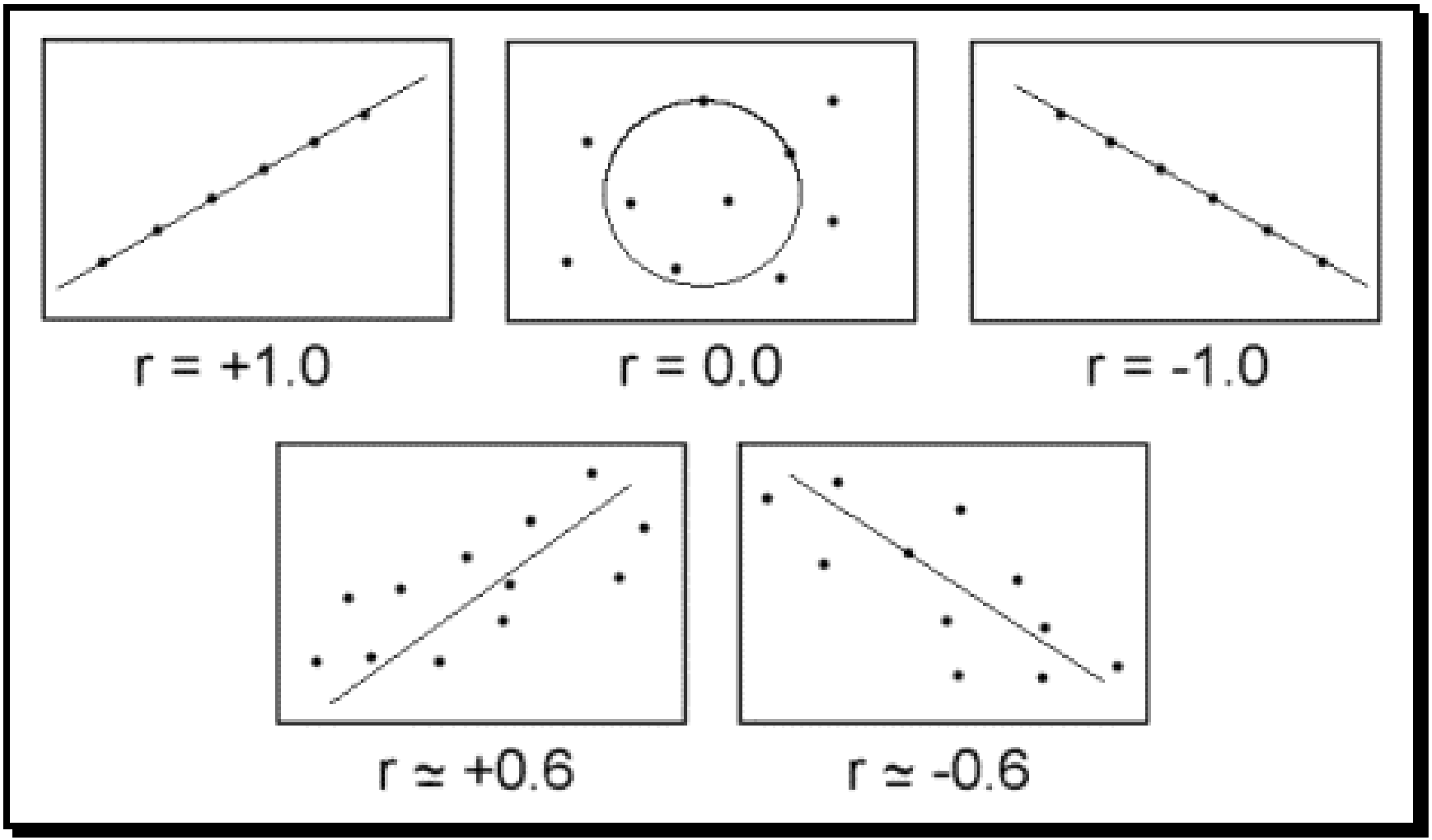
Correlação no Minitab



Faça a análise de Correlação das variáveis ao lado na planilha Bidimensional.mtw

O Coeficiente de Correlação é também chamado de Coeficiente de Pearson.

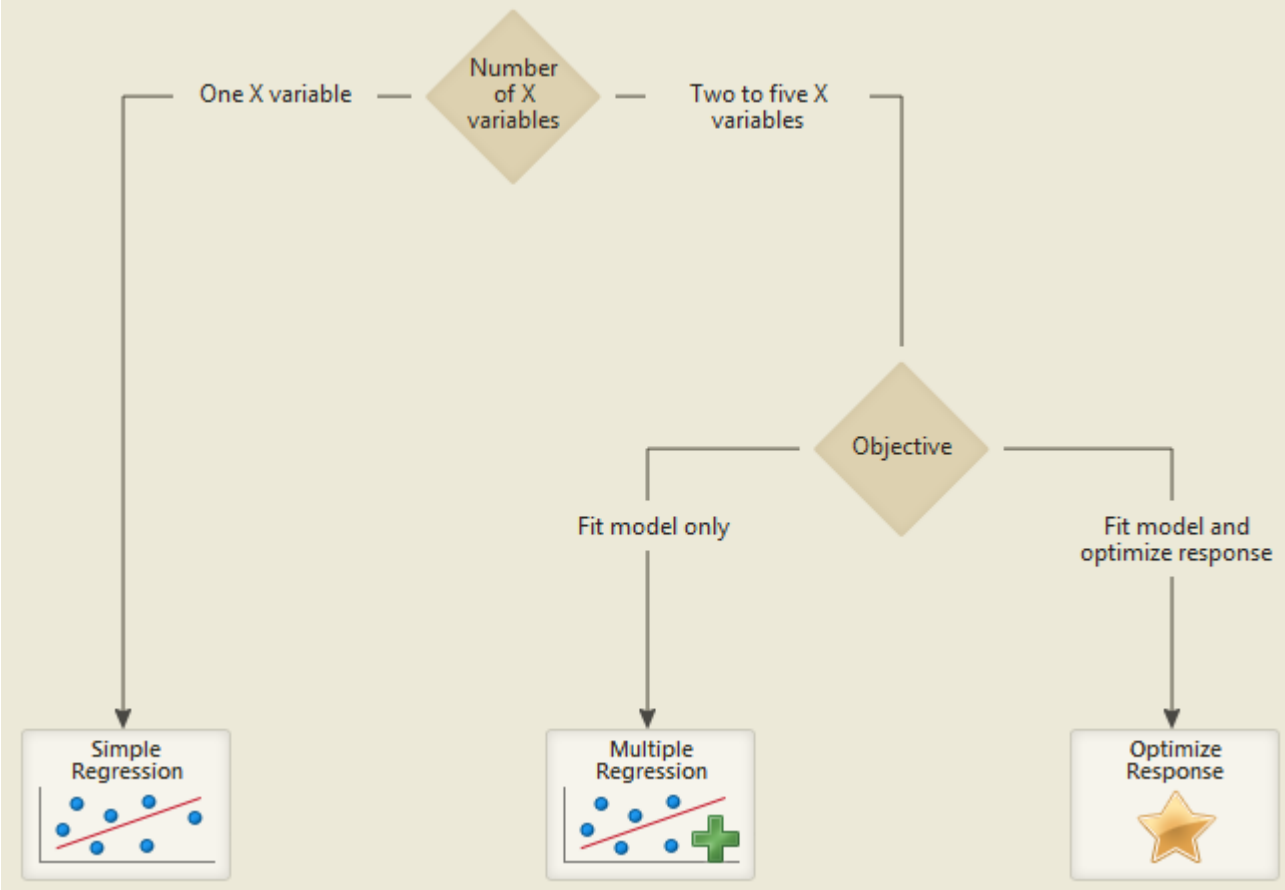
Correlação significa Causa/Efeito?



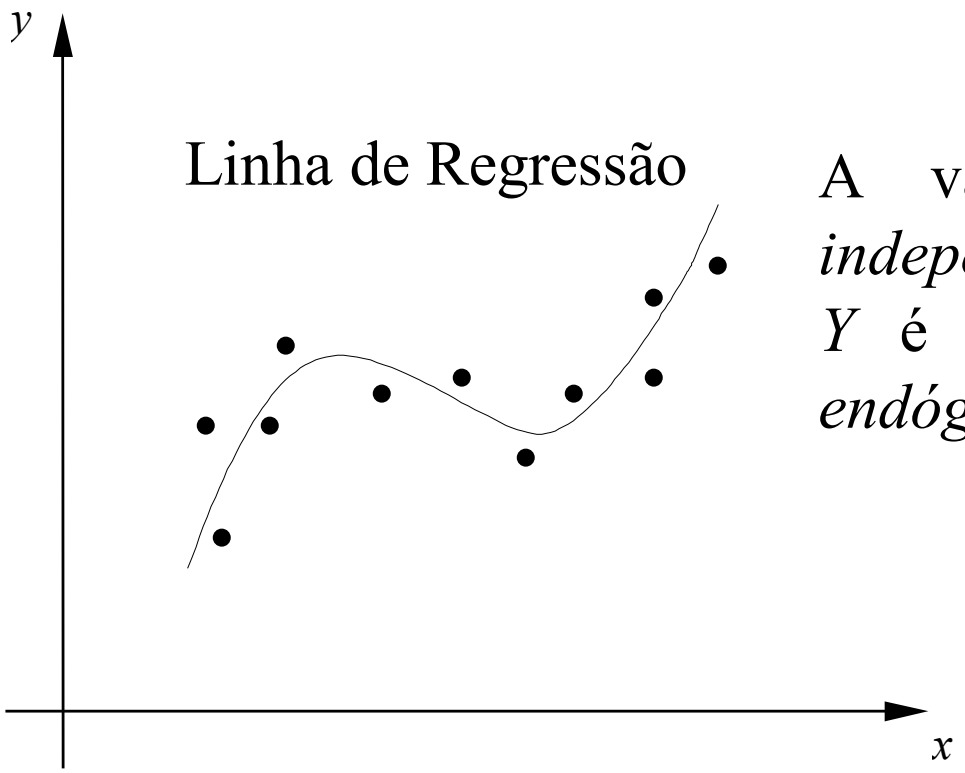
Análise de Regressão

- *A análise de regressão é uma técnica estatística usada para modelar e investigar a relação entre duas ou mais variáveis. O modelo é freqüentemente usado para **previsões**.*
- *Regressão é um **teste de hipótese***

H_a: O modelo permite significativamente prever a resposta.



Y=f(x)

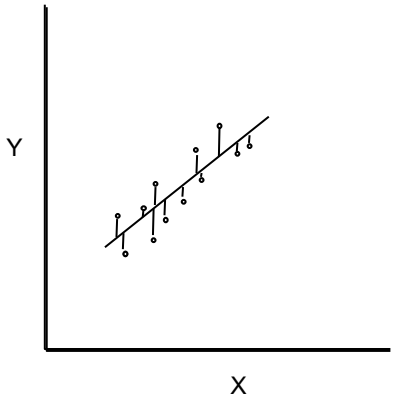


A variável *X* é dita *variável independente (ou exógena)*, enquanto *Y* é dita *variável dependente (ou endógena)*.

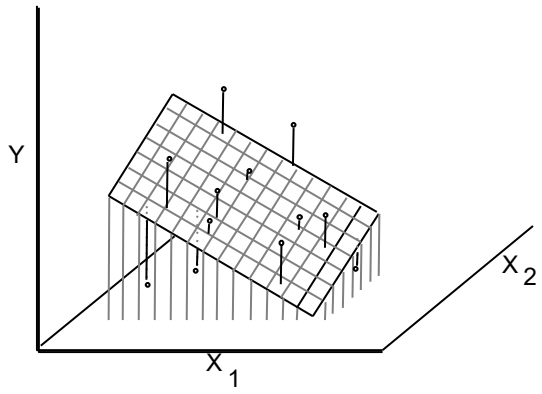
- **Y=f(x) Simples**
- **Y=f(x,y,z...) Múltipla**

Regressão

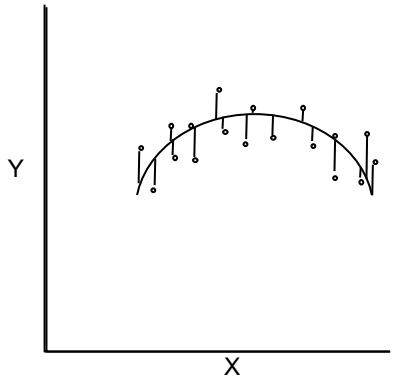
Linear Simples (Um X)



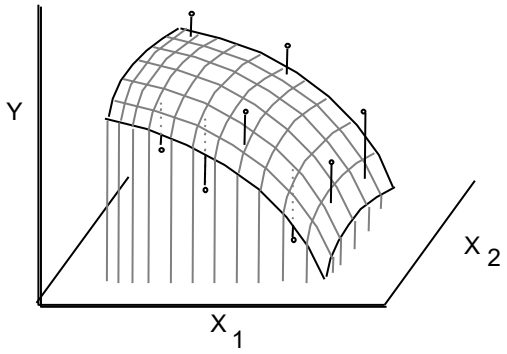
Múltipla (Dois ou mais Xs)



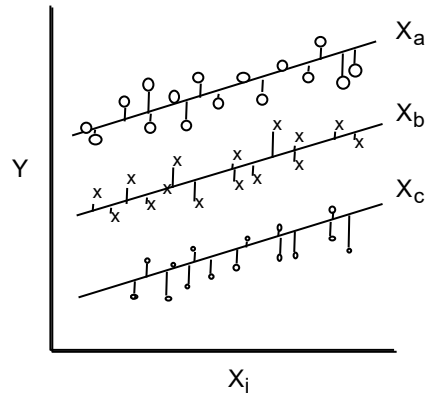
Curvilínea (Um X)



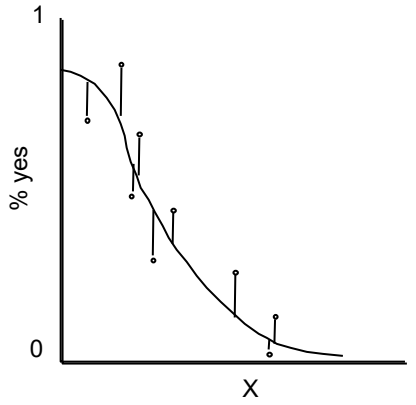
Curvilínear (Dois ou mais Xs)



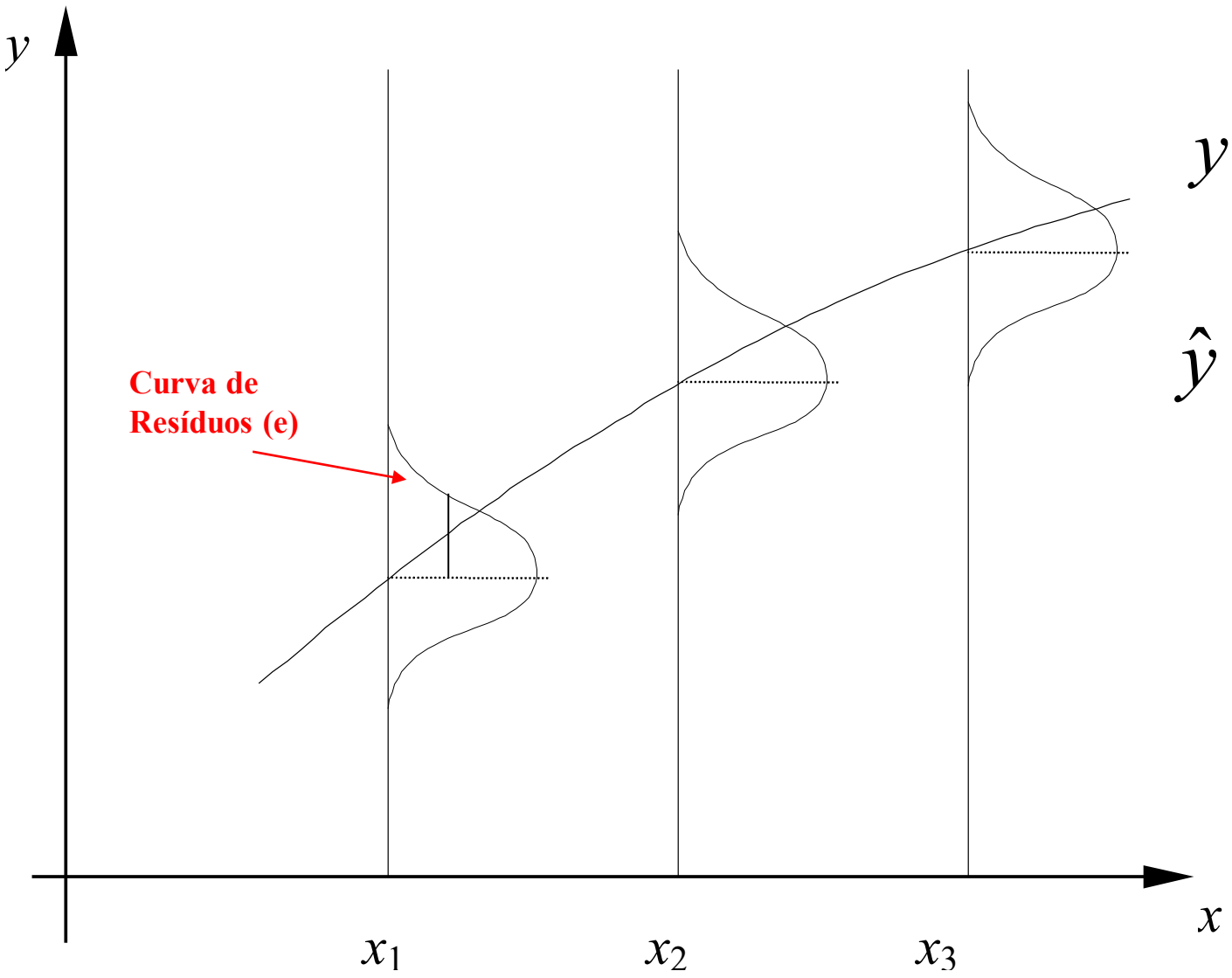
Variáveis Indicativas (para Xs Discretos)



Logística (Ys Discretos)



Resíduos

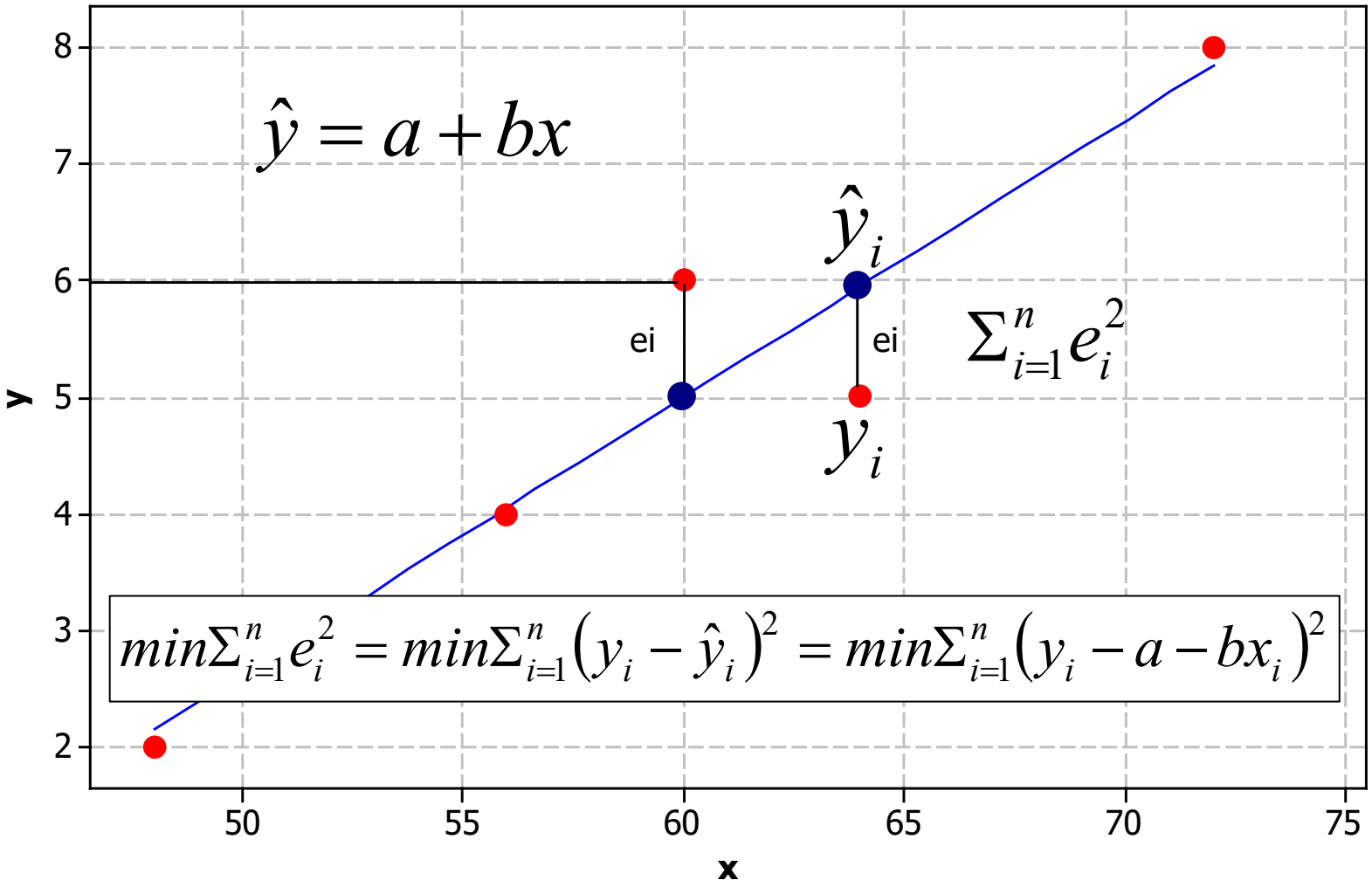


$$y = \alpha + \beta x$$

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

Uma importante condição para o uso de regressão simples é que os resíduos (**e**) sejam independentes de x. **Porque?**

Regressão Linear Simples



A matemática da Regressão Linear

$$\Sigma_{i=1}^{n} e_i^2 \qquad \hat{y} = a + bx$$

$$min\Sigma_{i=1}^{n} e_i^2 = min\Sigma_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = min\Sigma_{i=1}^{n} \left(y_i - a - bx_i\right)^2$$

$$\frac{\partial}{\partial a}\sum_{i=1}^{n} d_i^2 = 0 \text{ e } \frac{\partial}{\partial b}\sum_{i=1}^{n} d_i^2 = 0.$$

$$-2\sum_{i=1}^{n} (y_i - a - bx_i) = 0,$$
$$-2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0,$$

Ufa!

$$\begin{cases} \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i, \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \end{cases}$$

$$\begin{cases} b = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}}, \\ a = \bar{y} - b\bar{x}, \end{cases}$$

Exemplo

Obter a equação da reta (chamada de reta dos mínimos quadrados) para os seguintes pontos experimentais:

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

Traçar a reta no diagrama de dispersão. Calcular o coeficiente de correlação linear.

Qual o valor previsto para x=9?

Regressão: By Hand

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	0,5	0,5	1	0,25
2	0,6	1,2	4	0,36
3	0,9	2,7	9	0,81
4	0,8	3,2	16	0,64
5	1,2	6,0	25	1,44
6	1,5	9,0	36	2,25
7	1,7	11,9	49	2,89
8	2,0	16,0	64	4,00
36	9,2	50,5	204	12,64

$$S_{xy} = 50{,}5 - \frac{36 \cdot 9{,}2}{8} = 50{,}5 - 41{,}4 = 9{,}1,$$

$$S_{xx} = 204 - \frac{(36)^2}{8} = 204 - 162 = 42.$$

Regressão: Cálculos

$$S_{xy} = 50{,}5 - \frac{36 \cdot 9{,}2}{8} = 50{,}5 - 41{,}4 = 9{,}1,$$

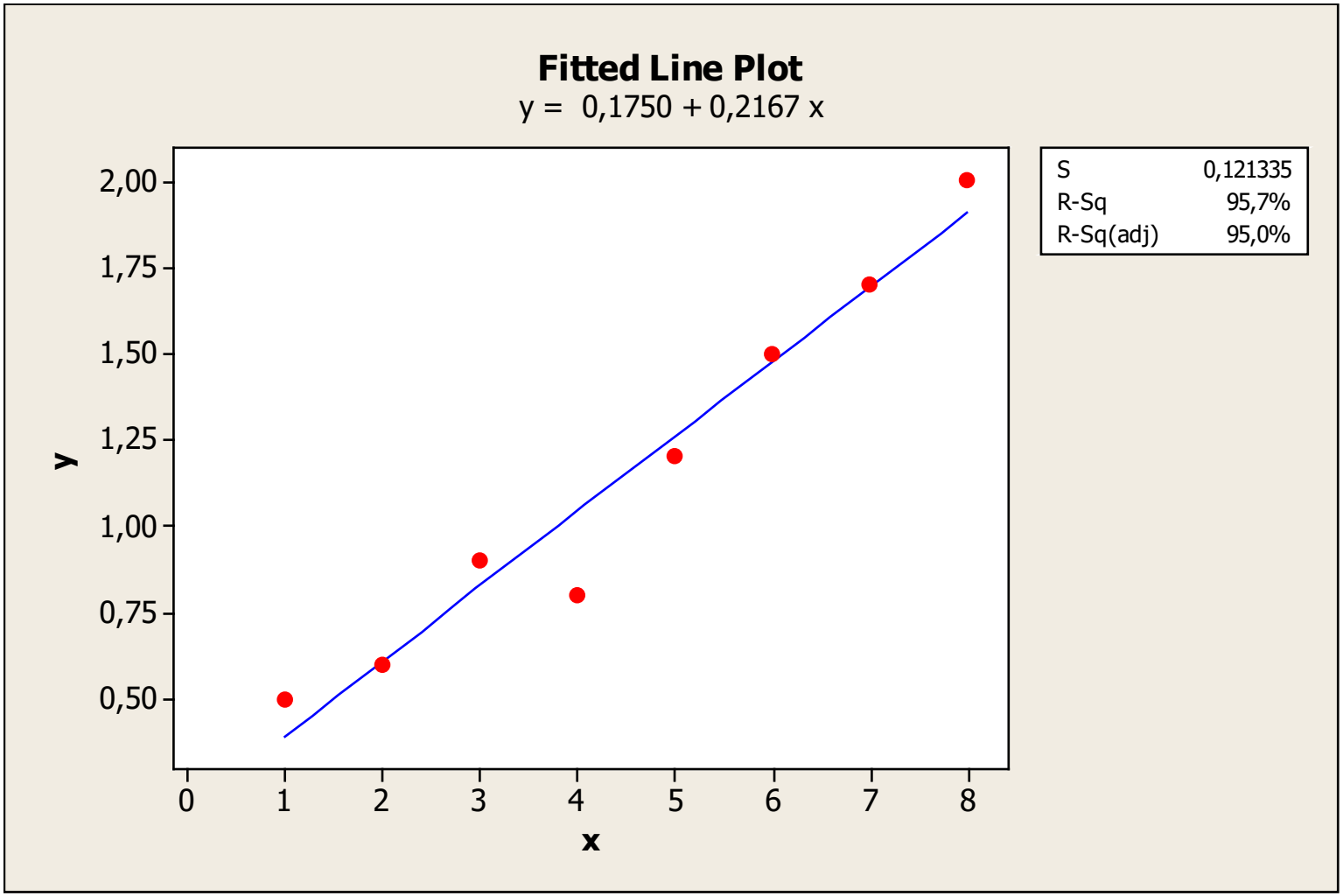
$$S_{xx} = 204 - \frac{(36)^2}{8} = 204 - 162 = 42.$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{9{,}1}{42} \cong 0{,}217,$$

$$a = \bar{y} - b\bar{x} \cong \frac{9{,}2}{8} - 0{,}217 \cdot \frac{36}{8} = 1{,}150 - 0{,}976 = 0{,}174.$$

$$\hat{y} = 0{,}174 + 0{,}217x$$

Regressão: Gráfico



Regressão: Correlação

$$S_{yy} = 12{,}64 - \frac{(9{,}2)^2}{8} = 12{,}64 - 10{,}58 = 2{,}06, \therefore$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{9{,}1}{\sqrt{42 \cdot 2{,}06}} \cong 0{,}98$$ Relembre Correlação!

Regressão:Teste Hipóteses

Para Teste de Hipóteses, considera-se:

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

Ho:a=0

T=a / SE Coef(a)

SE Coef (a)= $\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$

Ho:b=0

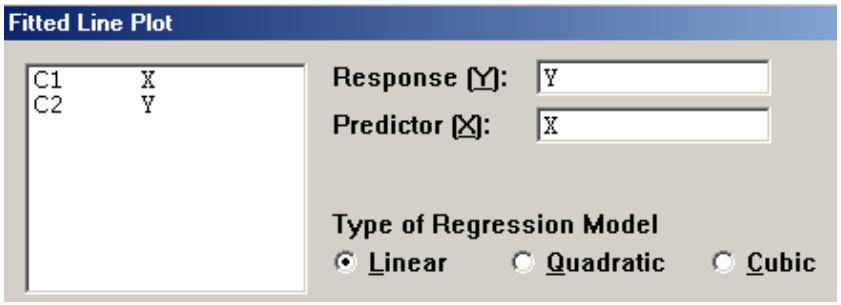
T=b / SE Coef(b)

SE Coef (b)= $\frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

Regressão linear simples no Minitab

```
The regression equation is
Y = 0,175 + 0,217 X

Predictor        Coef     SE Coef          T        P
Constant      0,17500     0,09454       1,85    0,114
X             0,21667     0,01872      11,57    0,000

S = 0,1213      R-Sq = 95,7%     R-Sq(adj) = 95,0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      1,9717      1,9717    133,92    0,000
Residual Error     6      0,0883      0,0147
Total              7      2,0600

Unusual Observations
Obs          X          Y         Fit      SE Fit    Residual    St Resid
  4       4,00     0,8000      1,0417      0,0439     -0,2417       -2,14R

R denotes an observation with a large standardized residual

Predicted Values for New Observations

New Obs     Fit     SE Fit         95,0% CI             95,0% PI
1        2,1250     0,0945   (  1,8936;  2,3564)  (  1,7485;  2,5015)

Values of Predictors for New Observations

New Obs         X
1            9,00
```

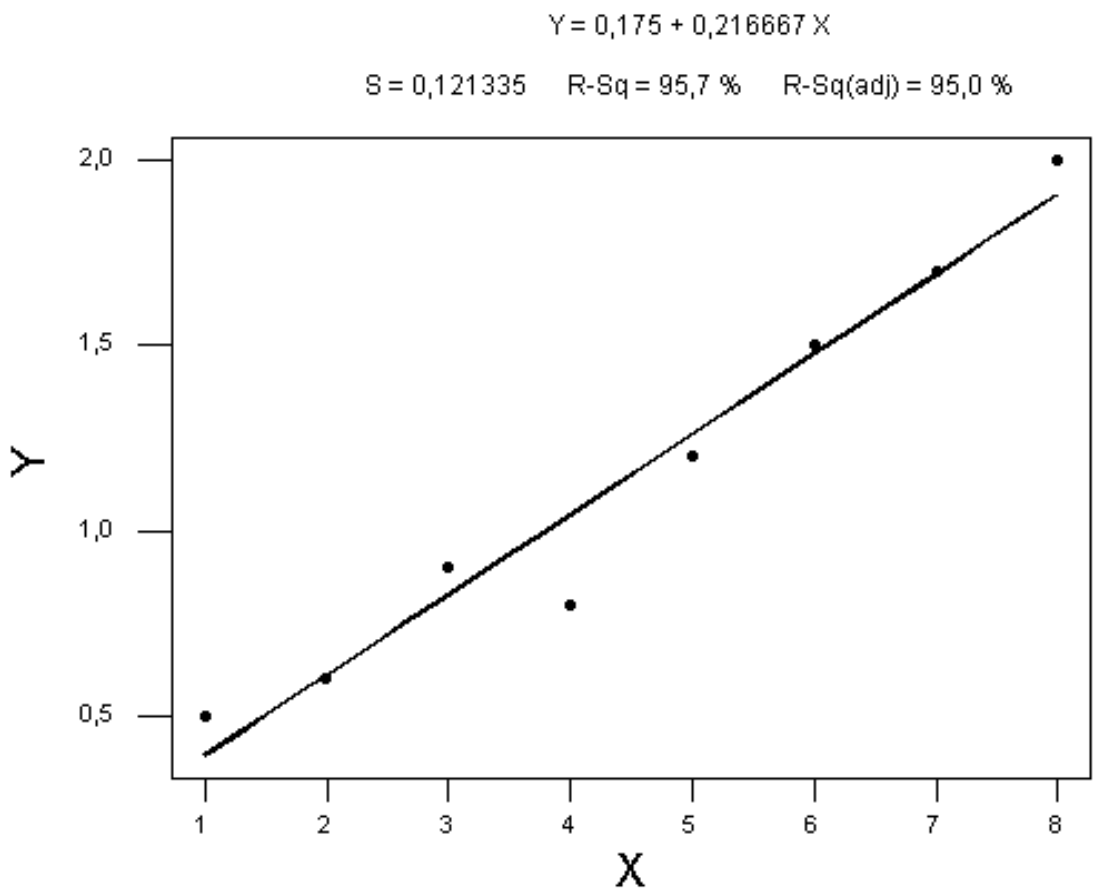


X	Y
1	0,5
2	0,6
3	0,9
4	0,8
5	1,2
6	1,5
7	1,7
8	2,0

Previsão

Ajuste da Regressão

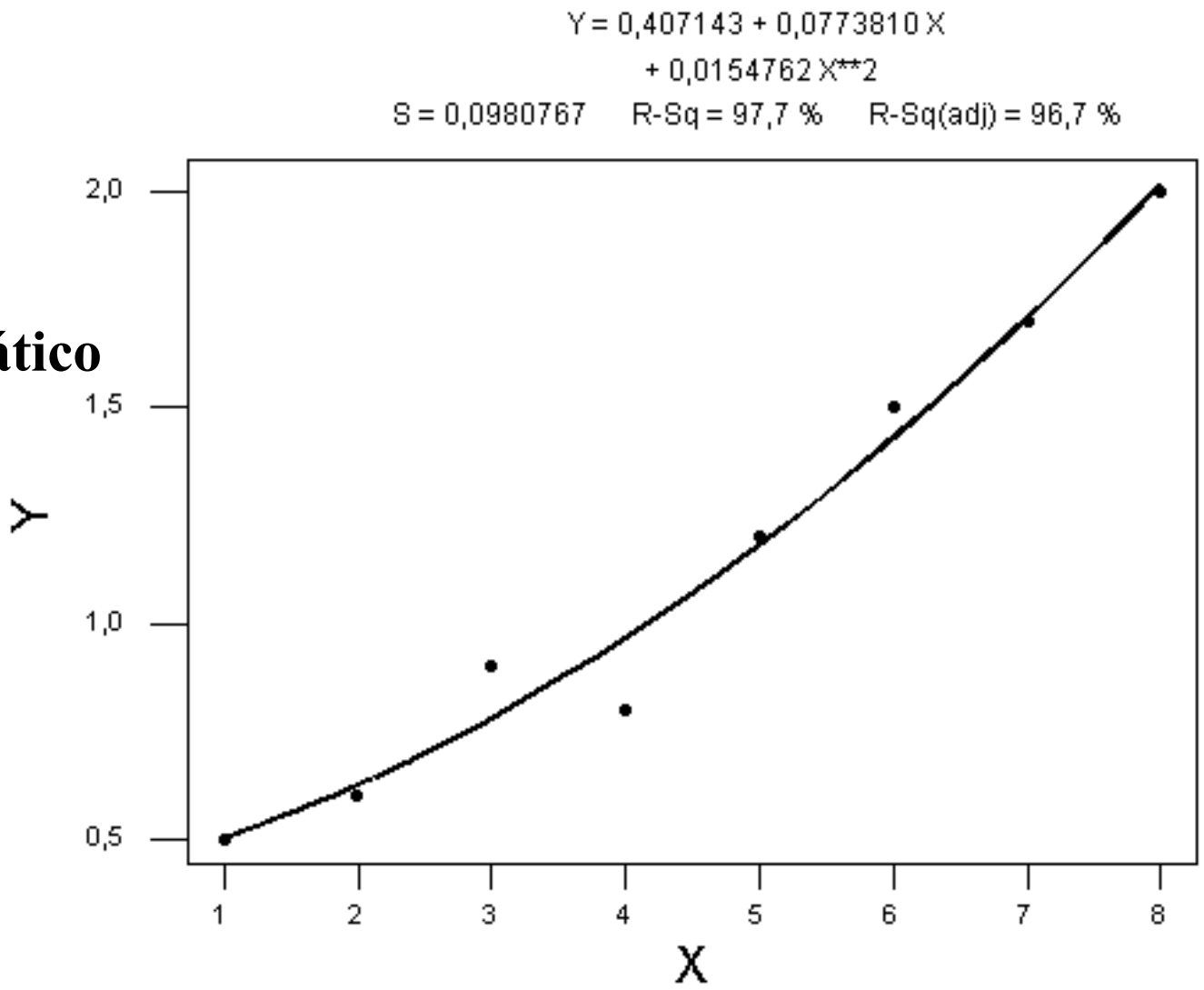
Linear



- *R-quadrado é a porcentagem da variação explicada pelo seu modelo.*
- *R-quadrado (ajustado) é a porcentagem da variação explicada pelo seu modelo, ajustada para o número de termos em seu modelo e o número de pontos de dados.*
- *O "valor-p" para a regressão é para ver se o modelo de regressão inteiro é significativo.*
  - *H_a: O modelo permite significativamente prever a resposta.*

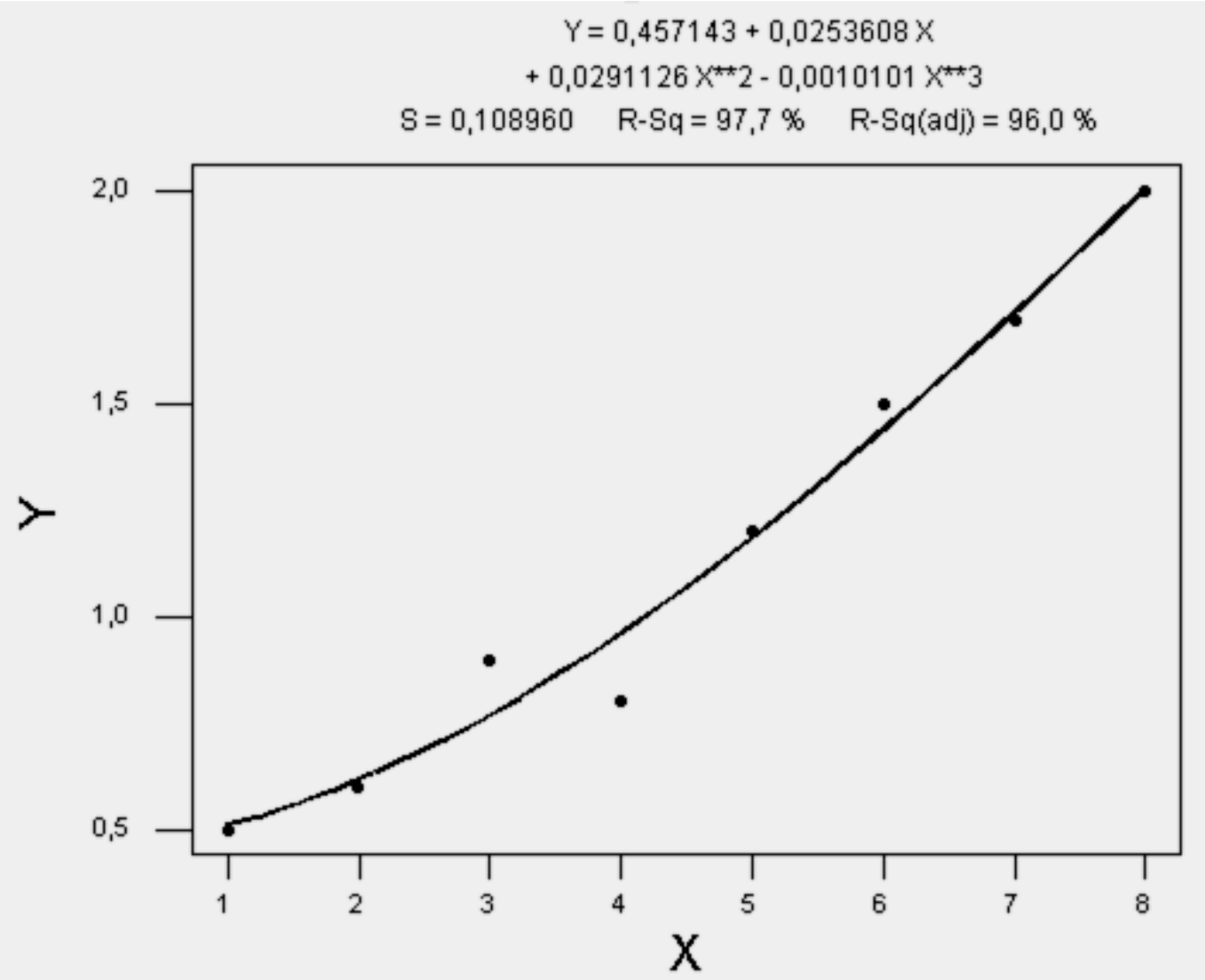
Ajuste Quadrático

Quadrático

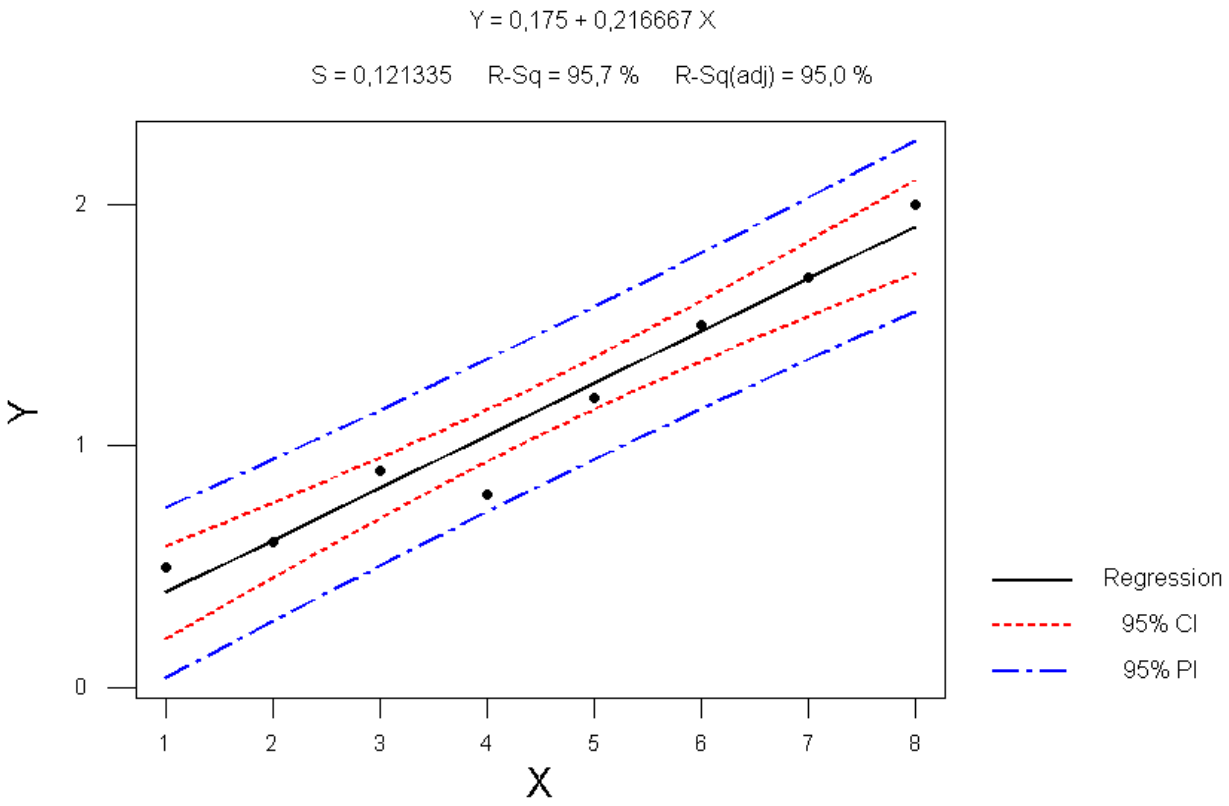


Ajuste Cúbico

Cúbico



Intervalos de confiança e de previsão



- *Uma faixa (ou intervalo) de confiança é uma medida da certeza da forma da linha de regressão ajustada. Em geral, uma faixa de 95% implica em uma chance de 95% de que as linha verdadeira fique dentro da faixa. [Linhas vermelhas]*

- *Uma faixa (ou intervalo) de previsão é uma medida da certeza da dispersão dos pontos individuais em torno da linha de regressão. Em geral, 95% dos pontos individuais (da população em que a linha de regressão se baseia) estarão contidos dentro da faixa. [Linhas azuis]*

Pratique Regressão Linear Simples

Determine a função de transferência entre o Número de Setups e o Tempo de Ciclo para diversas operações em uma certa empresa. Use a planilha **cycletime.mtw.**

Faça a análise de Resíduos.

Qual a previsão do Tempo de Ciclo para uma operação que consiste em 10 Setups de equipamento?

A equação final é adequada? Se não for, como melhorá-la?

Regressão Múltipla

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Parâmetros ou coeficientes

Erro aleatório

Ex.: Uma reação Química foi realizada sob seis pares de diferentes condições de pressão e temperatura. Em cada caso foi medido o tempo necessário para que a reação se completasse. Obter a equação de regressão do tempo em relação a pressão e temperatura.

Temperatura	Pressão	Tempo
20	1,5	9,4
30	1,5	8,2
30	1,2	9,7
40	1,0	9,5
60	1,0	6,9
80	0,8	6,5

Regressão.mtw

Regressão Múltipla: Resultados

```
The regression equation is
Tempo = 18,4 - 0,106 Temperatura - 4,68 Pressão

Predictor         Coef      SE Coef          T          P
Constant        18,428        1,827      10,08      0,002
Temperat     -0,10617      0,01391      -7,63      0,005
Pressão        -4,681        1,089      -4,30      0,023

S = 0,3383      R-Sq = 96,5%      R-Sq(adj) = 94,2%

Analysis of Variance

Source              DF          SS          MS          F          P
Regression           2      9,4500      4,7250      41,29      0,007
Residual Error       3      0,3433      0,1144
Total                5      9,7933

Source        DF      Seq SS
Temperat       1      7,3369
Pressão        1      2,1131
```

Menores que 0,05

Maior melhor

Best Subsets

92 estudantes americanos participam de um simples experimento. Cada estudante registra o seu peso, altura, gênero, pulso e se é fumante ou não. Todos eles jogam uma moeda e sorteiam se vão dar uma corrida (cara) ou não por um minuto. Após a corrida, todos os alunos registram o seu pulso novamente. Um aluno sugere que seja inserida a seguinte "importante" consideração: Se a pessoa pinta o cabelo ou não.

Deseja-se fazer uma regressão do segundo pulso em relação a todas as outras variáveis.

Pulso1	Pulso2	Corrida	Fumante	Sexo	Altura	Peso	Cabelo
64	88	1	2	1	66,00	140	1
58	70	1	2	1	72,00	145	1
62	76	1	1	1	73,50	160	1
66	78	1	1	1	73,00	190	1
64	80	1	2	1	69,00	155	1
74	84	1	2	1	73,00	165	2
84	84	1	2	1	72,00	150	1
68	72	1	2	1	74,00	190	1
62	75	1	2	1	72,00	195	1
76	118	1	2	1	71,00	138	1
90	94	1	1	1	74,00	160	1
80	96	1	2	1	72,00	155	1

Regressão.mtw

Best Subsets: Resultados

Equação de regressão inicial. Muito complexa

Regression Analysis: Pulso2 versus Pulso1; Corrida; ...

The regression equation is
Pulso2 = 41,5 + 0,819 Pulsol - 20,3 Corrida + 0,82 Fumante + 6,21 Sexo
+ 0,135 Altura - 0,0545 Peso + 0,30 Cabelo

Predictor	Coef	SE Coef	T	P
Constant	41,54	32,16	1,29	0,200
Pulsol	0,81870	0,09555	8,57	0,000
Corrida	-20,326	2,089	-9,73	0,000
Fumante	0,819	2,247	0,36	0,716
Sexo	6,206	5,768	1,08	0,285
Altura	0,1348	0,4754	0,28	0,777
Peso	-0,05451	0,07316	-0,75	0,458
Cabelo	0,303	4,930	0,06	0,951

S = 9,342 R-Sq = 72,4% R-Sq(adj) = 70,1%

Correlations: Pulso1; Corrida; Fumante; Sexo; Altura; Peso; Cabelo

	Pulsol	Corrida	Fumante	Sexo	Altura	Peso
Corrida	-0,052					
	0,621					
Fumante	-0,129	0,066				
	0,221	0,535				
Sexo	0,285	0,107	0,129			
	0,006	0,311	0,220			
Altura	-0,212	-0,228	-0,056	-0,714		
	0,043	0,028	0,594	0,000		
Peso	-0,202	-0,218	-0,200	-0,709	0,785	
	0,053	0,037	0,056	0,000	0,000	
Cabelo	0,193	0,107	0,178	0,908	-0,613	-0,615
	0,065	0,311	0,090	0,000	0,000	0,000

Cell Contents: Pearson correlation
P-Value

Correlação muito alta. Quem pinta cabelo é "geralmente" mulher

Best Subsets: Resultados

Best Subsets Regression: Pulso2 versus Pulso1; Corrida; ...

Response is Pulso2

Melhor ajuste

Vars	R-Sq	R-Sq(adj)	C-p	S	Pulsol	Course	Fumar	Sexo	Altura	Peso	Cabelo
1	38,0	37,3	101,0	13,538	X						
1	33,3	32,5	115,3	14,041		X					
2	67,7	67,0	12,4	9,8219	X	X					
2	47,2	46,0	74,9	12,560		X		X			
3	72,1	71,2	0,9	9,1751	X	X		X			
3	71,4	70,4	3,1	9,2969	X	X					X
4	72,3	71,1	2,3	9,1948	X	X		X		X	
4	72,2	71,0	2,6	9,2115	X	X	X	X			
5	72,4	70,8	4,1	9,2379	X	X	X	X		X	
5	72,4	70,8	4,1	9,2410	X	X		X	X	X	
6	72,4	70,5	6,0	9,2875	X	X	X	X	X	X	
6	72,4	70,5	6,1	9,2917	X	X	X	X		X	X
7	72,4	70,1	8,0	9,3424	X	X	X	X	X	X	X

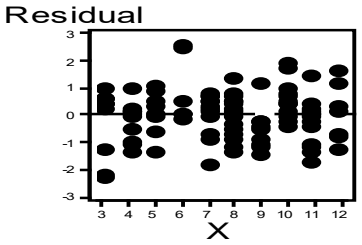
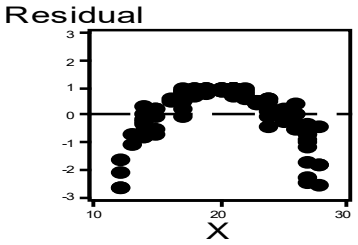
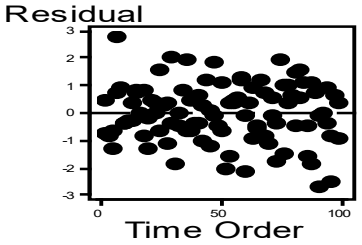
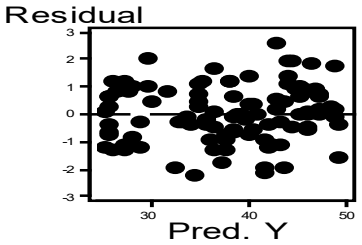
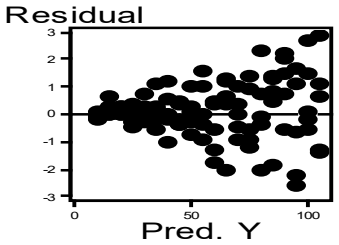
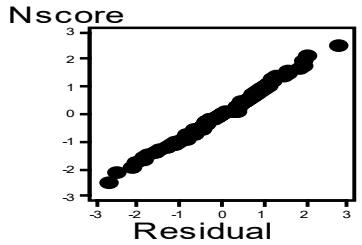
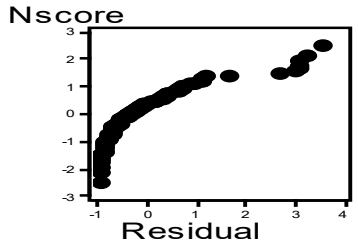
Regression Analysis: Pulso2 versus Pulso1; Corrida; Sexo

The regression equation is
Pulso2 = 42,6 + 0,812 Pulsol - 20,1 Corrida + 7,75 Sexo

Predictor	Coef	SE Coef	T	P
Constant	42,618	7,358	5,79	0,000
Pulsol	0,81217	0,09151	8,88	0,000
Corrida	-20,069	1,989	-10,09	0,000
Sexo	7,753	2,073	3,74	0,000

S = 9,175 R-Sq = 72,1% R-Sq(adj) = 71,2%

Análise de Resíduos

	Bom	Ruim
Residuals vs Each X		
Time Plot of Residuals		
Residuals vs Predicted Y (Fits)		
Normal Probability Plot of Residuals		

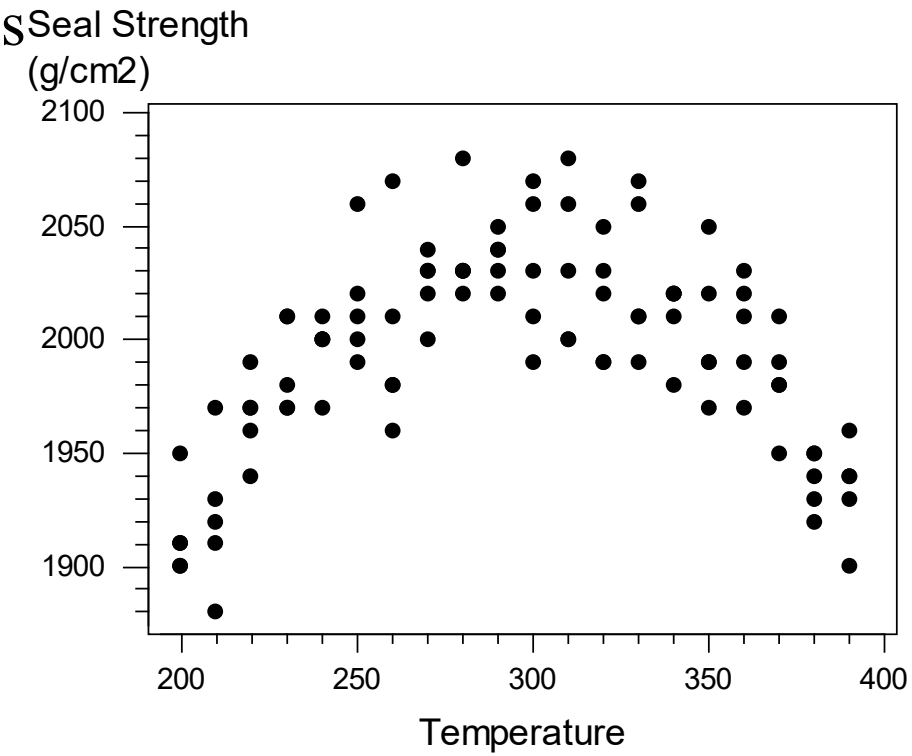
Nos casos ruins tente uma transformação em X,em Y ou ambos. Use *Box-Cox Transformation*

Considere a possibilidade da existência de variáveis ocultas que não foram consideradas no modelo (Lurking)

Entenda que X e Y não precisam ser normalmente distribuídos. Os resíduos, contudo, deveriam ser.

Regressão Curvilínea

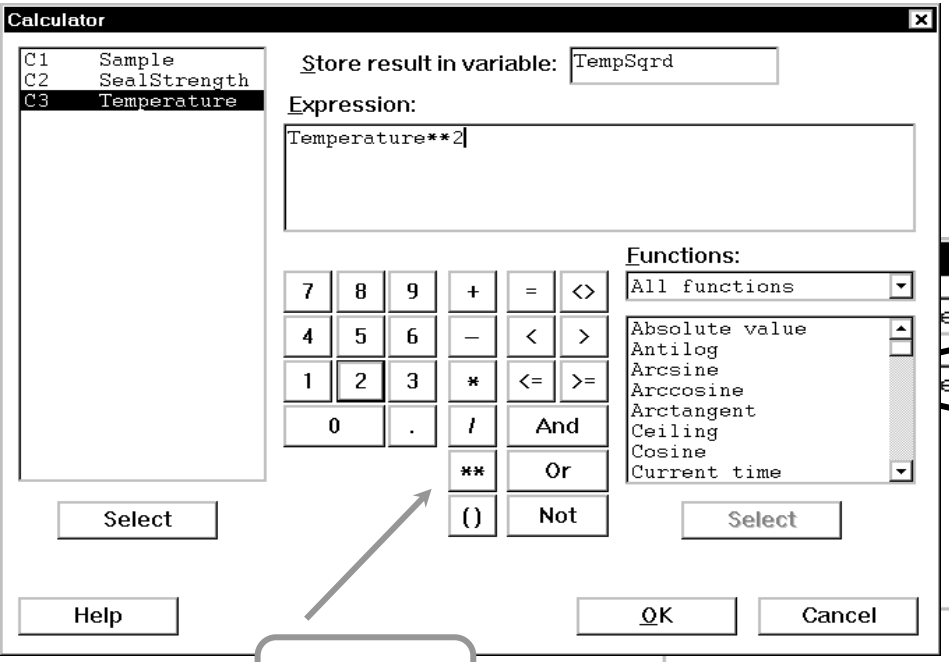
Um laboratório está fazendo testes em adesivos em função da temperatura. Quando a temperatura aumenta a força do contato entre duas superfícies aumenta Em um determinado ponto, contudo a força desse contato começa a diminuir em função de propriedades térmicas do adesivo. Qual o modelo empírico da força (Seal Strength) em função da temperatura?

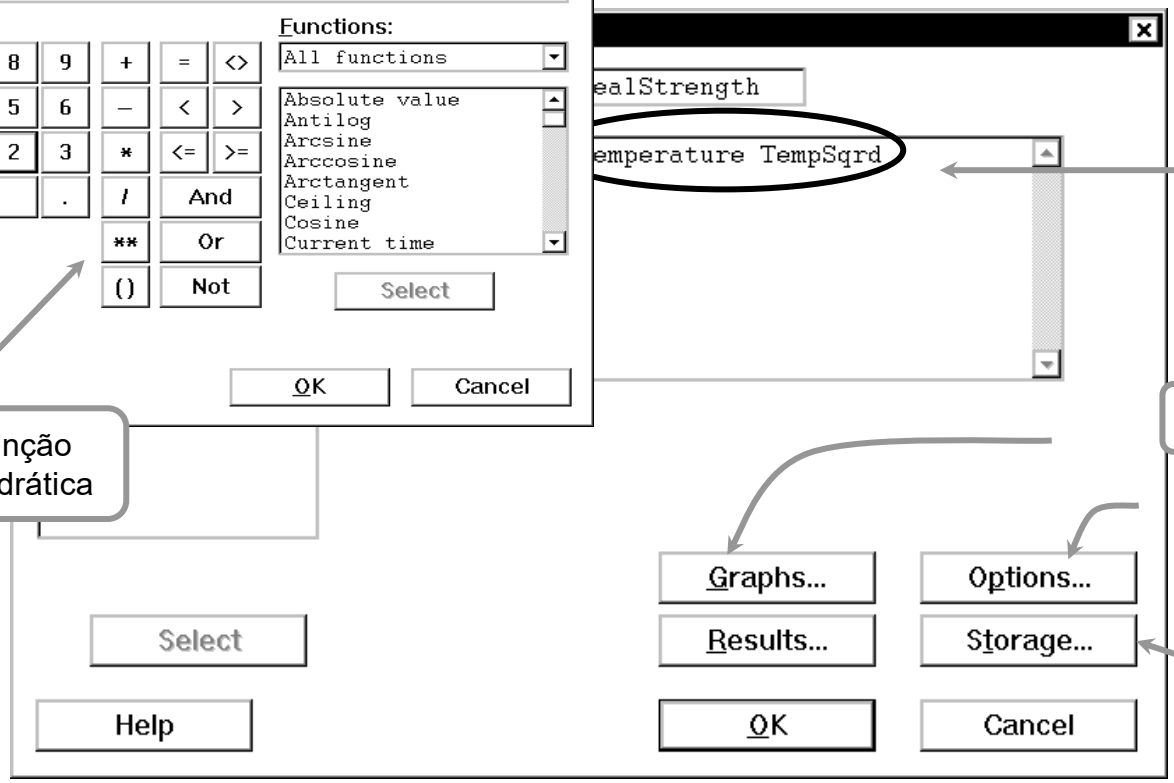


Curve.mtw

Termo quadrático da regressão

Deve-se criar a variável quadrática e em seguida rodar o modelo em **Regression**





Termo quadrático

Função quadrática

Observe resíduos

VIF

Armazena resíduos

Regressão Curvilínea

```
The regression equation is
SealStrength = 923 + 7.45 Temperature - 0.0125 TempSqrd

Predictor         Coef       StDev          T        P       VIF
Constant        922.98       72.33      12.76    0.000
Temperat        7.4469      0.5033      14.80    0.000     132.9
TempSqrd   -0.0124596   0.0008499     -14.66    0.000     132.9

S = 25.18       R-Sq = 69.4%     R-Sq(adj) = 68.7%

Analysis of Variance

Source             DF          SS
Regression          2      139321       69661    109.87    0.000
Residual Error     97       61498         634
Total              99      200819

Source        DF      Seq SS
Temperat       1        3051
TempSqrd       1      136270
```

X →
X^2 →

X e X^2 são fortemente correlacionados. Nenhuma surpresa

Conclusão: Existe uma curvatura significativa

Pratique!

- Livro Texto: Montgomery/Runger 5e
  - **Chapters 11 and 12**(Resolver todos os exercícios com análise de dados pelo Minitab).