

# Análise de Regressão

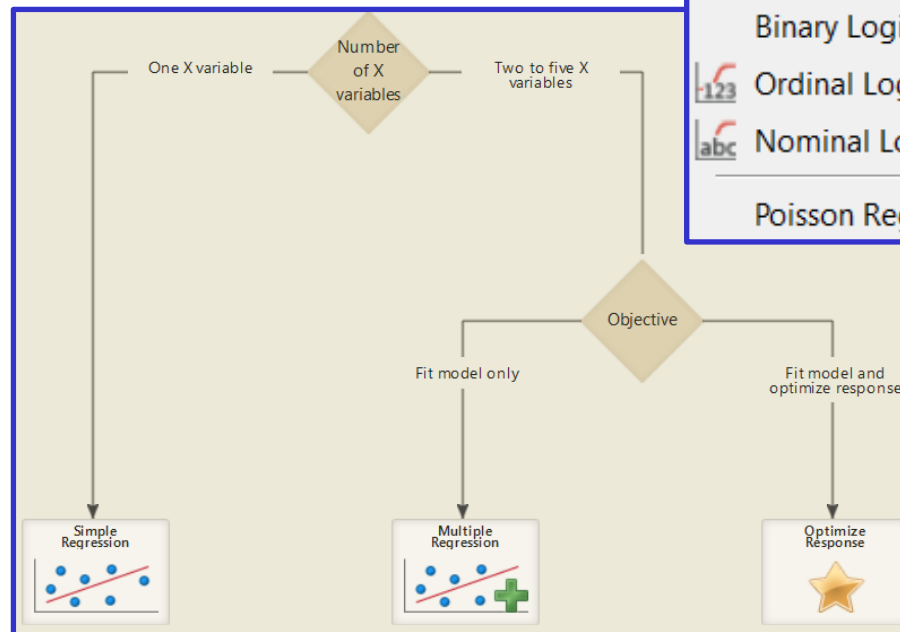
**Pedro Paulo Balestrassi**

[www.pedro.unifei.edu.br](http://www.pedro.unifei.edu.br)

[ppbalestrassi@gmail.com](mailto:ppbalestrassi@gmail.com)

35-36291161 / 999012304 (cel)

- Fitted Line Plot...  
Regression
- Nonlinear Regression...
- Stability Study
- Orthogonal Regression...
- Partial Least Squares...
- Binary Fitted Line Plot...  
Binary Logistic Regression
- Ordinal Logistic Regression...
- Nominal Logistic Regression...
- Poisson Regression



# Coeficiente de Correlação

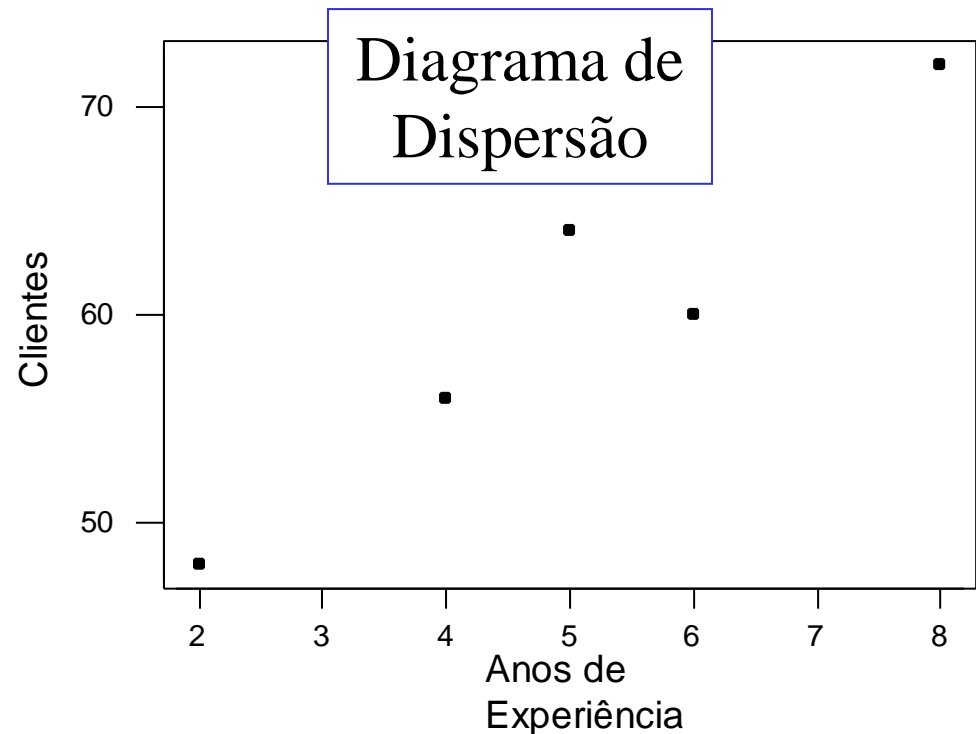
**Ex.:** Suponha que o nosso desejo seja o de quantificar a associabilidade entre duas variáveis relacionadas a cinco agentes de uma seguradora.

Assim, temos:

$X \equiv$  Anos de experiência do agente.

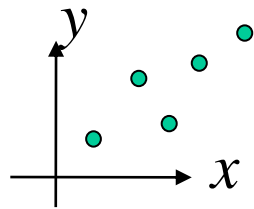
$Y \equiv$  Número de clientes do agente.

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

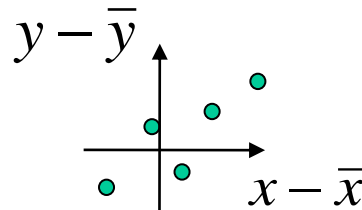


$(x, y)$  é um par aleatório  
– Dados emparelhados

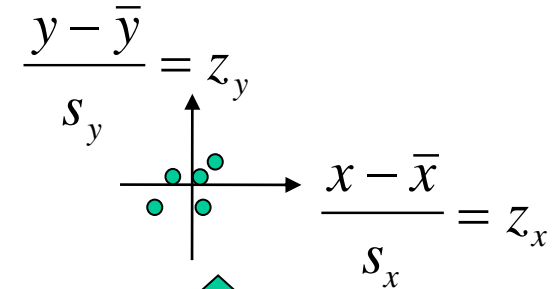
# r=Correlação de Pearson



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



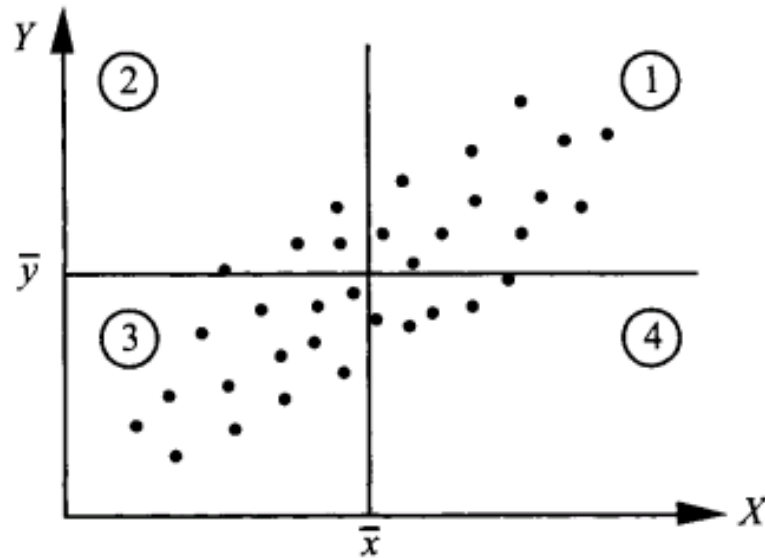
O conjunto de pontos é deslocado, tendo agora como centro, os valores médios.



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	-	+	-

# Cálculo simplificado de r

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

$$\bar{x} = 5 \quad \bar{y} = 60$$
$$S_x = 2 \quad S_y = 8$$

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

# Pvalue p/ Correlação

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

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

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

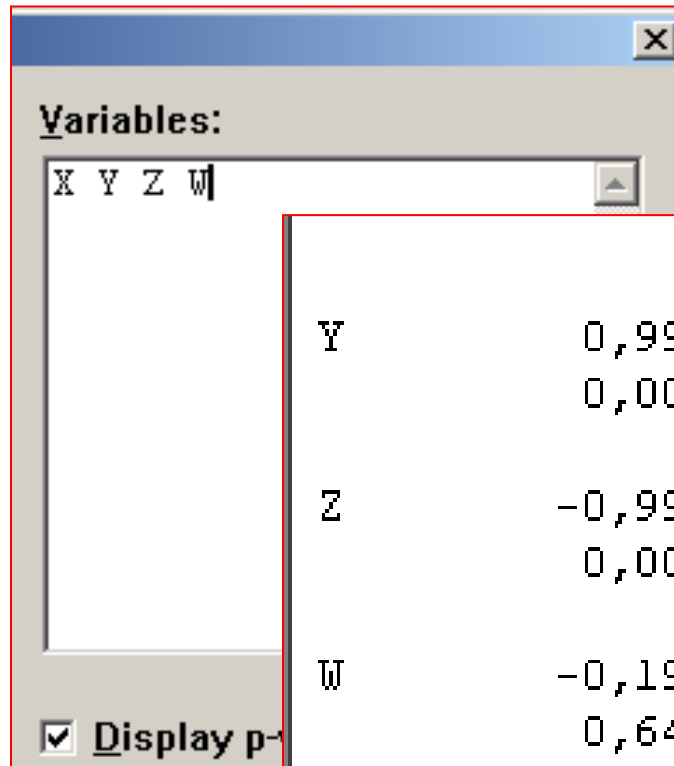
Pearson correlation of Anos Exp and Clientes = 0,950

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

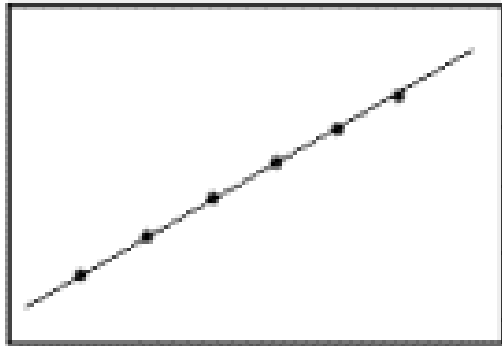


	X	Y	Z
Y	0,990 0,000		
Z	-0,992 0,000	-0,976 0,000	
W	-0,193 0,647	-0,273 0,513	0,082 0,847

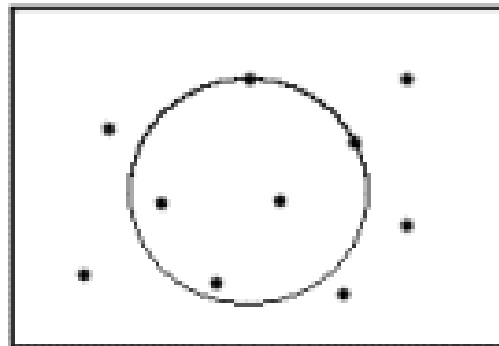
Cell Contents: Pearson correlation  
P-Value

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

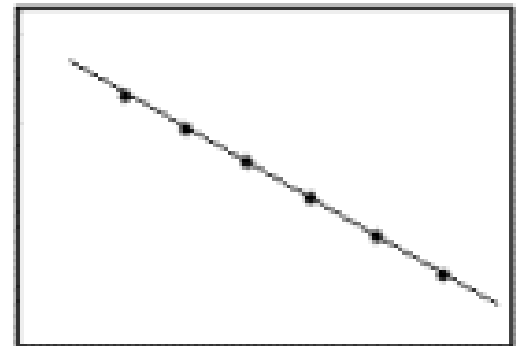
# Correlação significa Causa/Efeito?



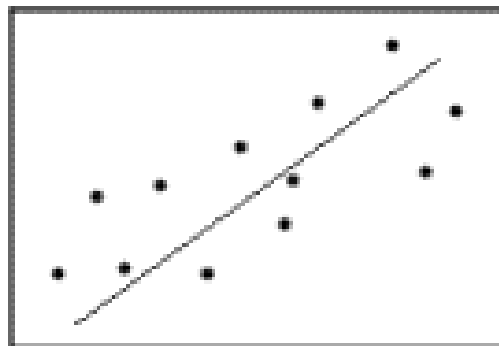
$$r = +1.0$$



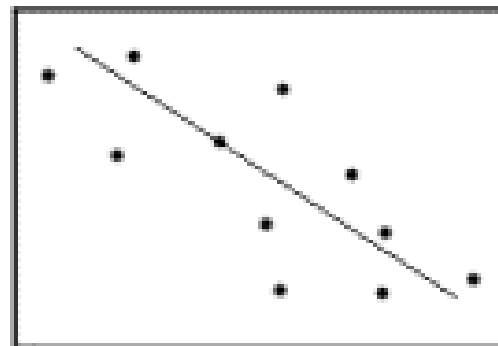
$$r = 0.0$$



$$r = -1.0$$



$$r \approx +0.6$$



$$r \approx -0.6$$

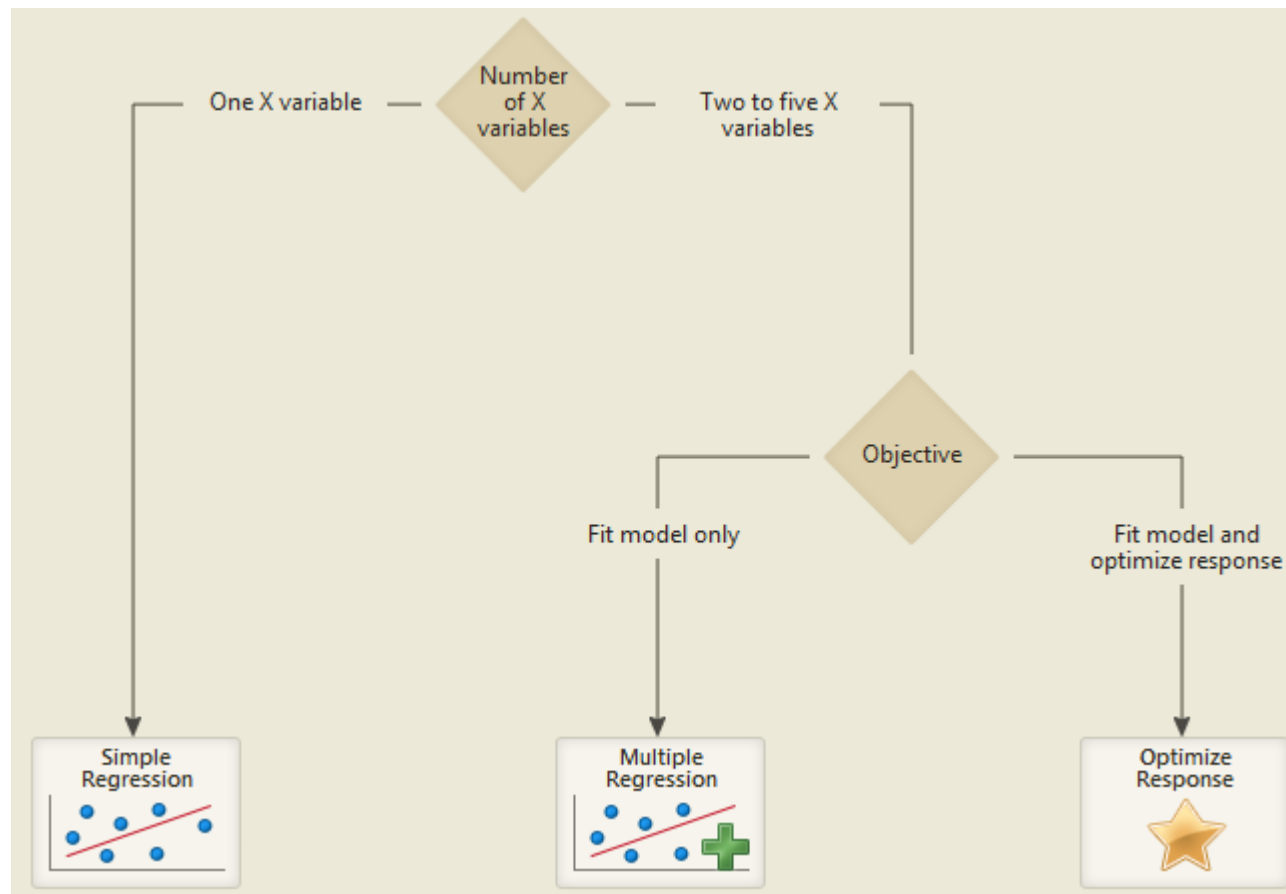


# Análise de Regressão

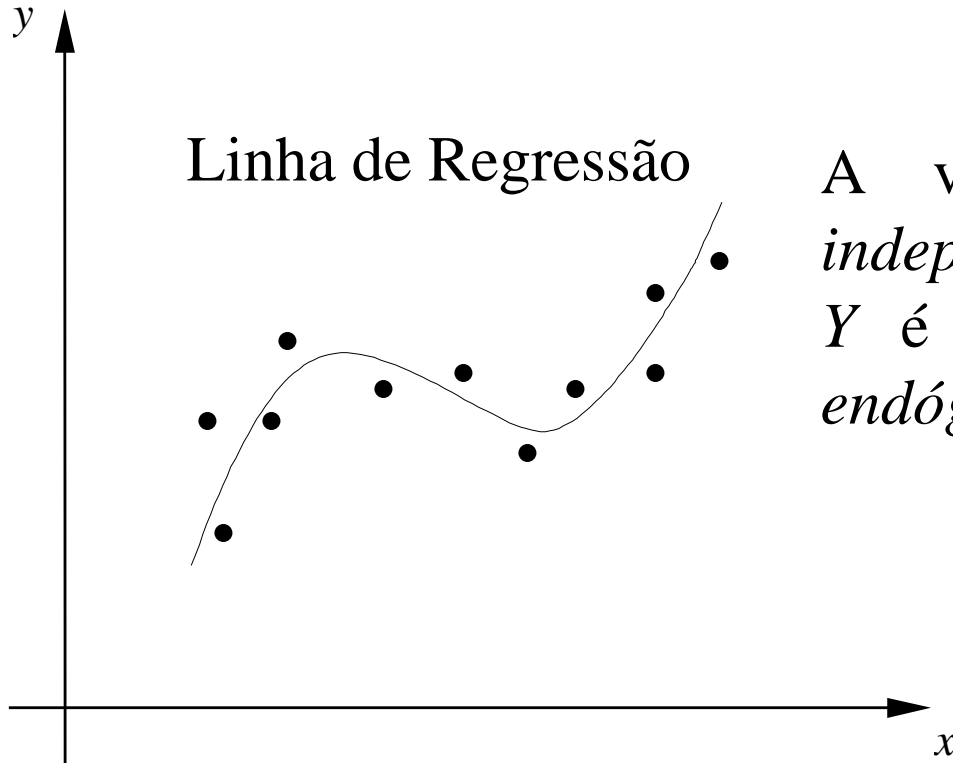
□ A análise de regressão é uma técnica estatística usada para modelar e investigar a relação entre duas ou mais variáveis. O modelo é freqüentemente usado para **previsões**.

□ Regressão é um **teste de hipótese**

$H_a$ : O modelo permite significativamente prever a resposta.



# $Y=f(x)$



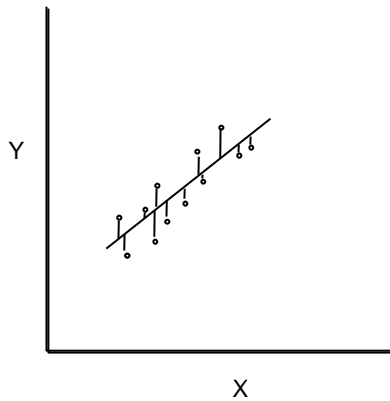
A variável  $X$  é dita *variável independente (ou exógena)*, enquanto  $Y$  é dita *variável dependente (ou endógena)*.

•  $Y=f(x)$  **Simple**s

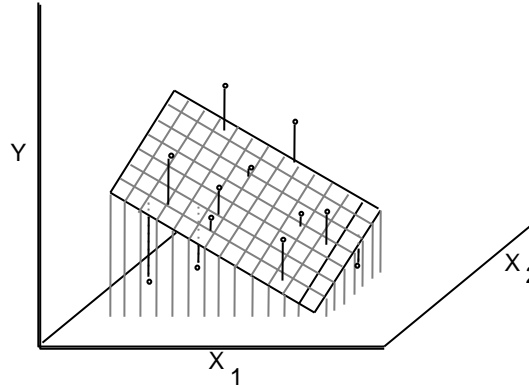
•  $Y=f(x,y,z\dots)$  **Múltipla**

# Alguns Tipos de Regressão

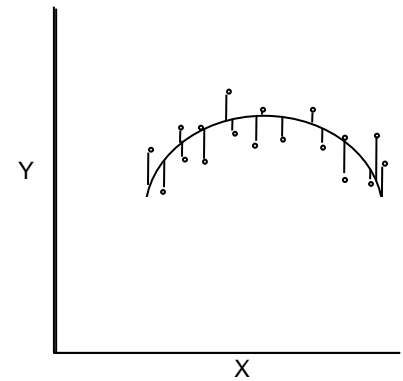
**Linear Simples (Um X)**



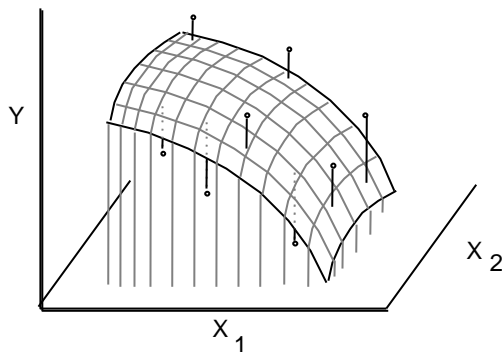
**Múltipla (Dois ou mais Xs)**



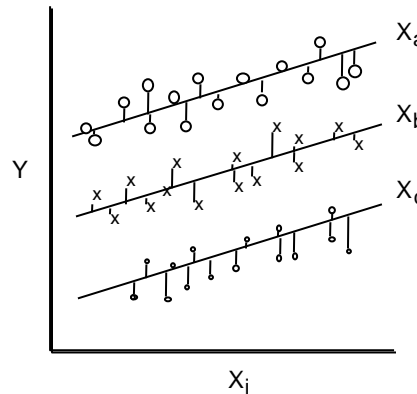
**Curvilínea (Um X)**



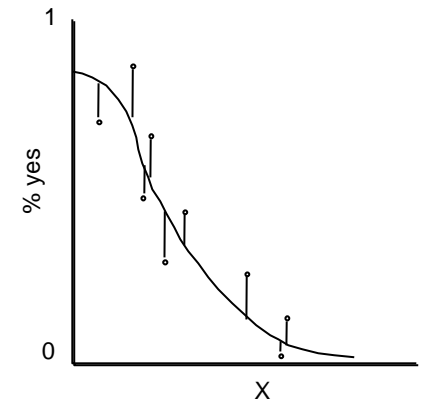
**Curvilínea (Dois ou mais Xs)**



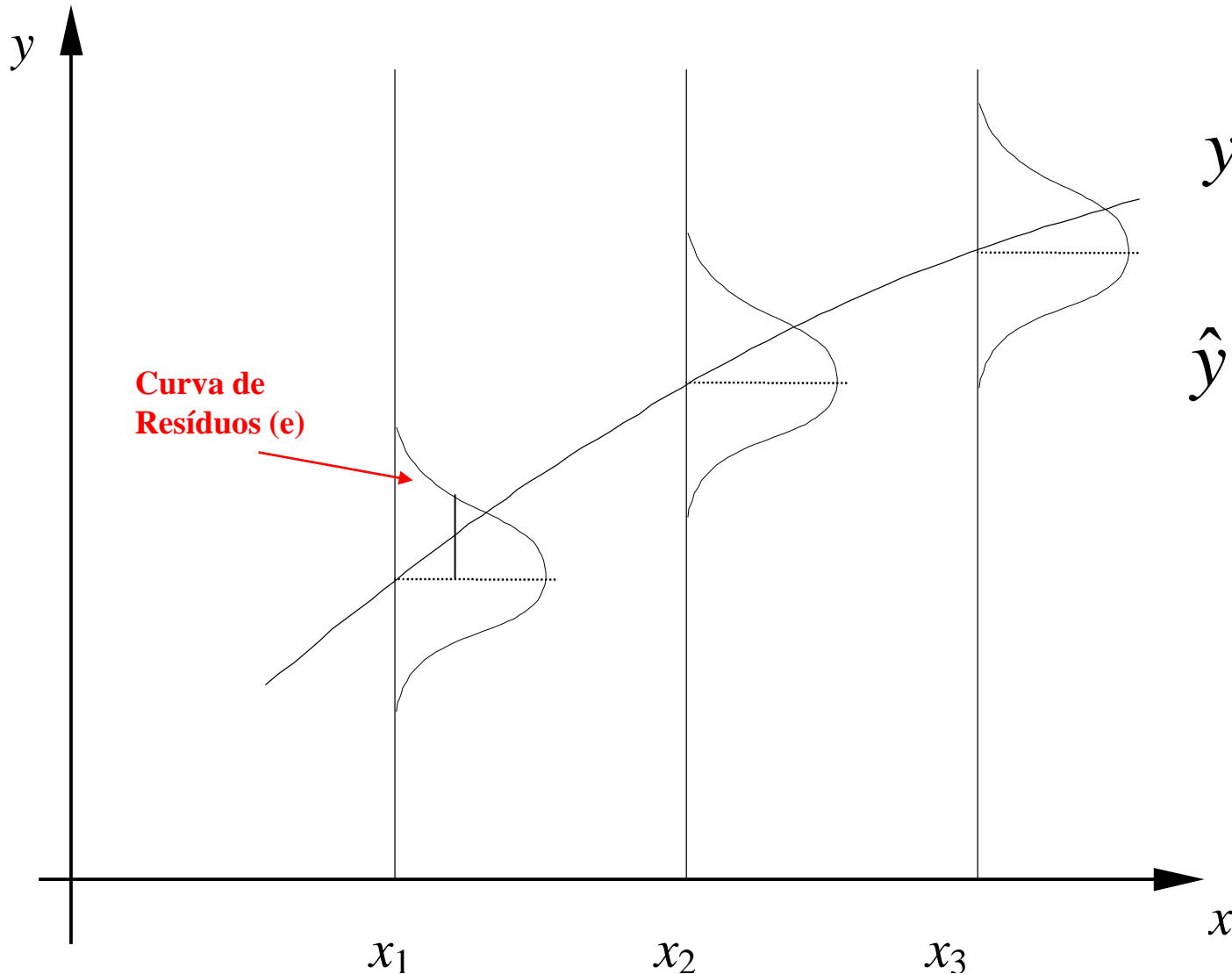
**Variáveis Indicativas  
(para Xs Discretos)**



**Logística (Ys Discretos)**



# Resíduos

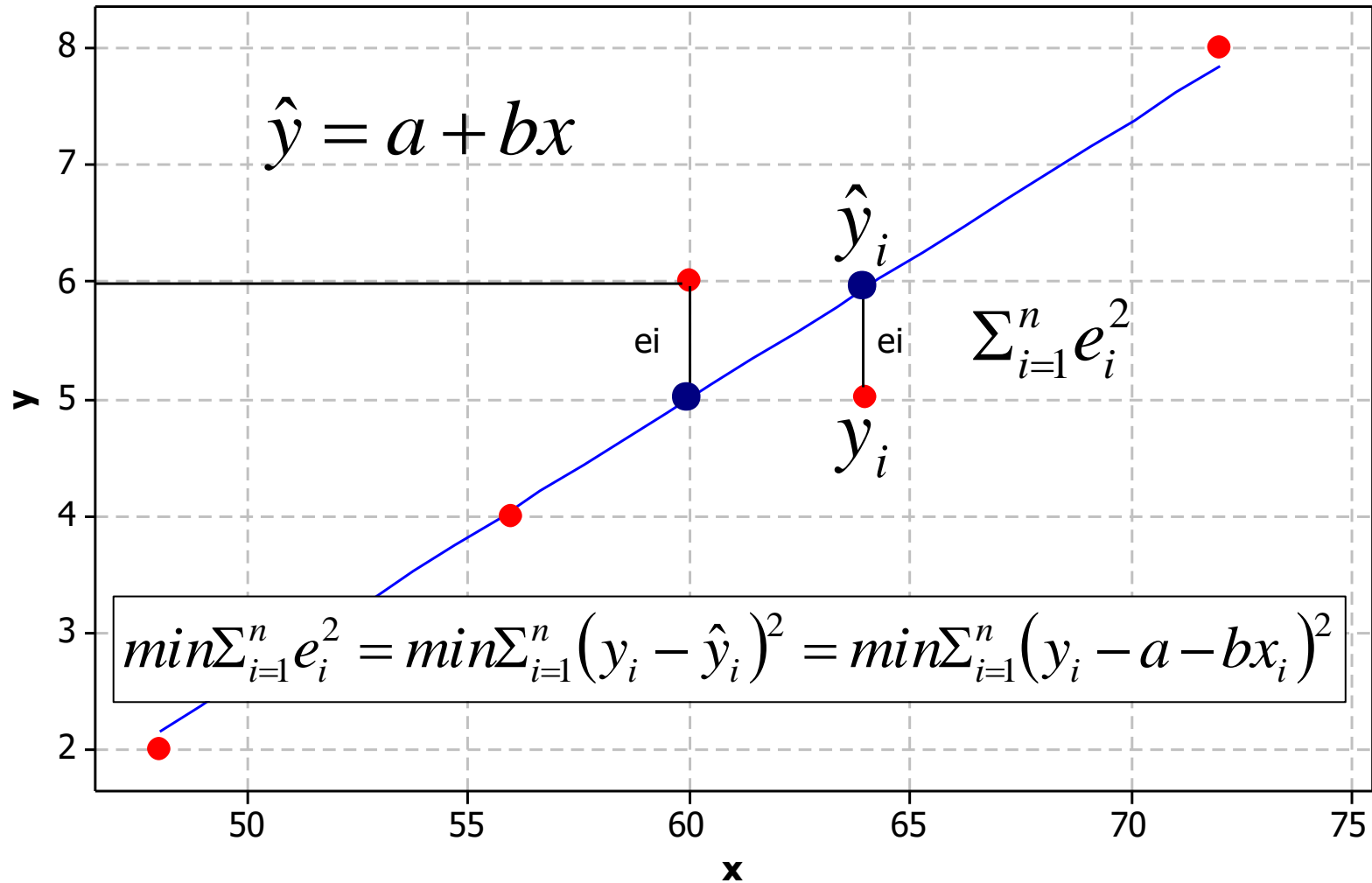


$$y = \alpha + \beta x$$

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

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

$$\sum_{i=1}^n e_i^2$$

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

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

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

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

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

# Ufa!

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

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

# 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

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

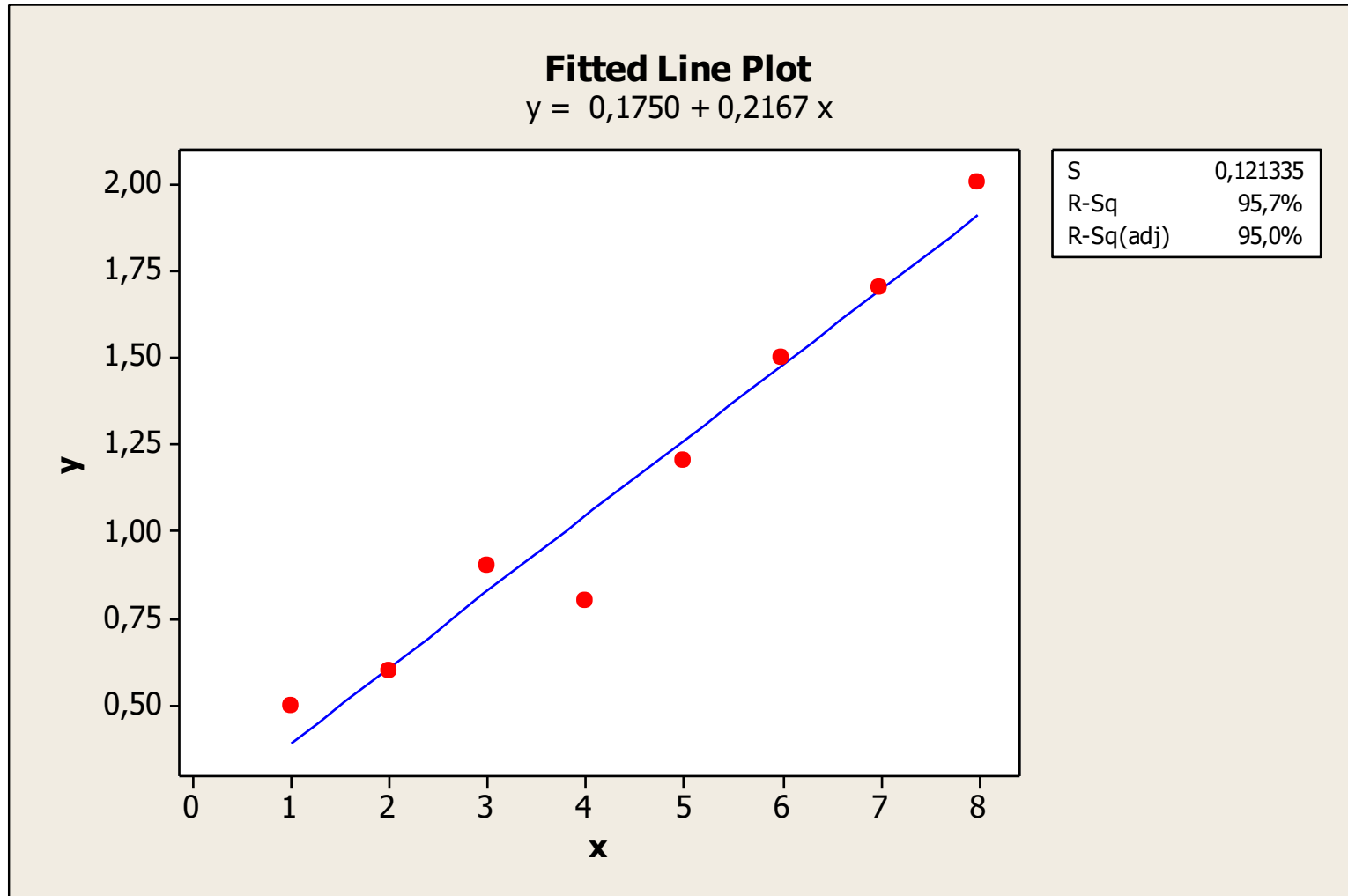
$$S_{xx} = 204 - \frac{(36)^2}{8} = 204 - 162 = 42.$$

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

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

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

# Regressão: Gráfico



# Regressão: Correlação

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

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

# Regressão: Teste Hipóteses

Para Teste de Hipóteses, considera-se:

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

$$H_0: a=0$$

$$T = a / \text{SE Coef}(a)$$

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

$$H_0: b=0$$

$$T = b / \text{SE Coef}(b)$$

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

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

# 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

Fitted Line Plot

C1	X
C2	Y

Response (Y):

Predictor (X):

Type of Regression Model

Linear     Quadratic     Cubic

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**

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(1)

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

(2)

$$L = \sum_{i=1}^n \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

(3)

$$\left. \frac{\partial L}{\partial \boldsymbol{\beta}} \right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{0}$$

(4)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

(5)

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

(6)

$$SS_E = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

(7)

$$\hat{\sigma}^2 = \frac{SS_E}{n - p}$$

(8)

$$\hat{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

(9)

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

(10)

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

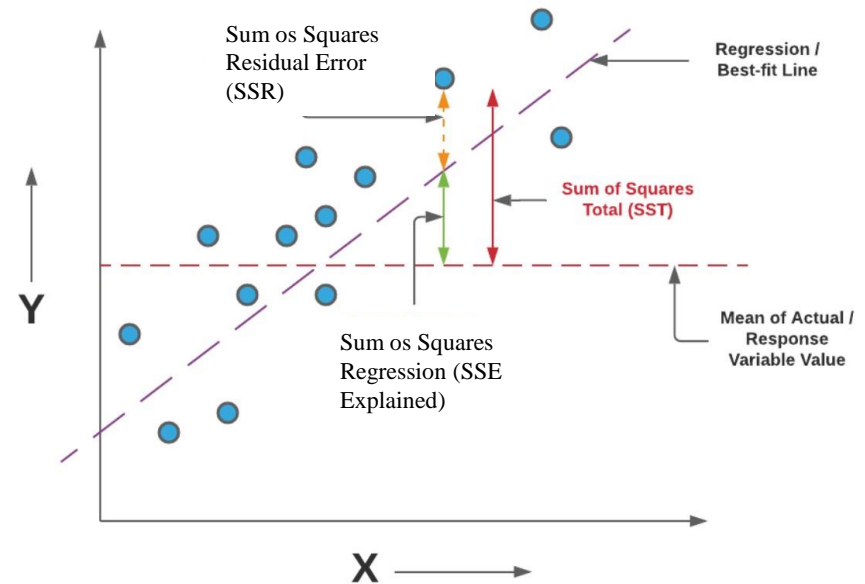
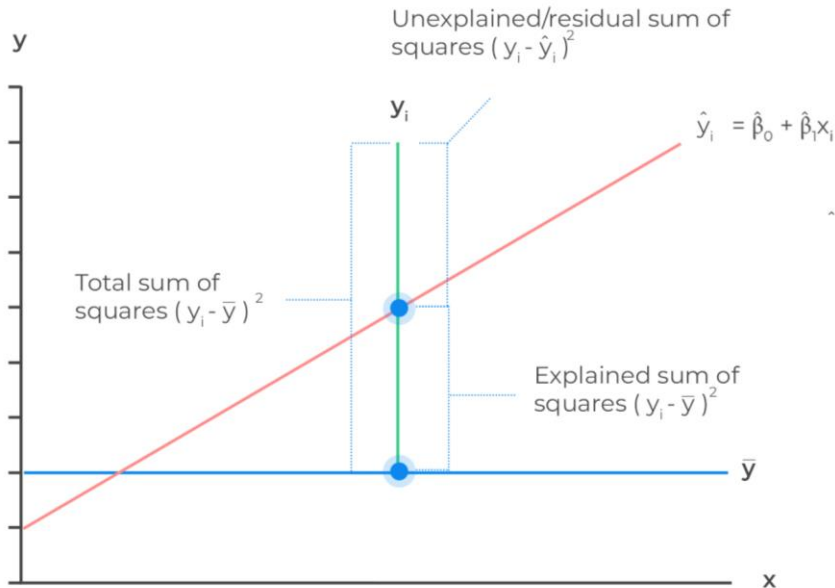
(11)

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

(12)

1. Modelo de Regressão Linear;
2. Termos da Equação de Regressão Linear;
3. Equação de Erros Quadráticos;
4. Minimização de Erros;
5. Coeficientes Estimados do Modelo;
6. Equação Normal de Erros;
7. Soma dos Quadrados dos Resíduos;
8. Variância dos Resíduos;
9. Equação Estimada de Regressão;
10. Vetor de erros;
11. BLUE (Best Linear Unbiased Estimator);
12. Variância dos Coeficientes Estimados.

# Coefficient of Determination and Regression Ratio



$$SST_{\text{total}} = SSE_{\text{explained}} + SSR_{\text{residual}}$$

or, equivalently,

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SSR}{SST}$$

**$R^2$** : Coefficient of Determination (the greater the better)

**SSR**: Sum of Squares of Residuals

**TSS**: Total Sum of Squares

$\frac{SSR}{SST}$ : Regression Ratio (the lower the better)

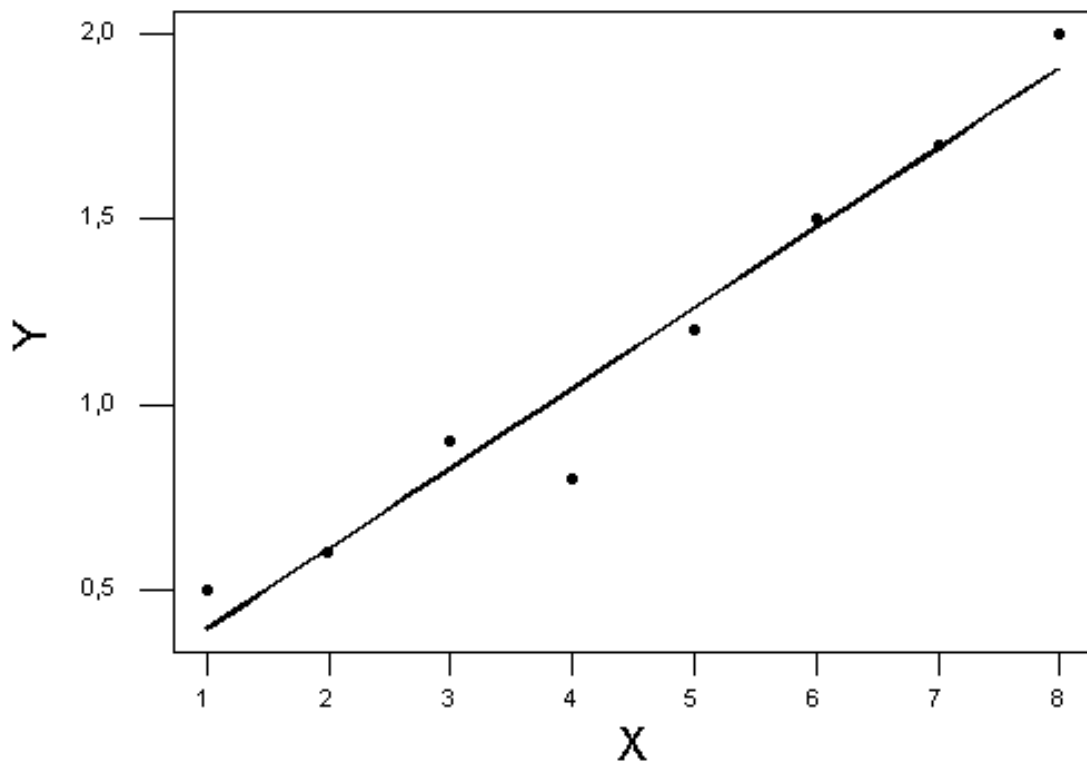


# Ajuste da Regressão

## Linear

$$Y = 0,175 + 0,216667 X$$

$$S = 0,121335 \quad R\text{-Sq} = 95,7 \% \quad R\text{-Sq}(\text{adj}) = 95,0 \%$$



□ *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.

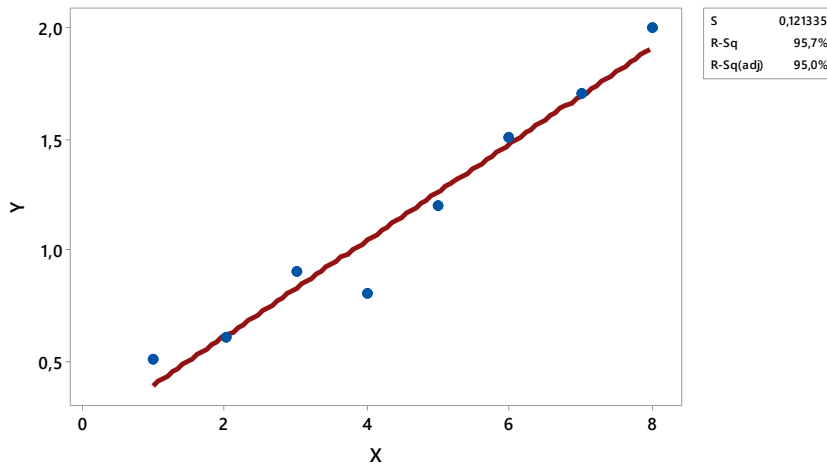
□ O “valor-p” para a regressão é para ver se o modelo de regressão inteiro é significativo.

—  $H_a$ : O modelo permite significativamente prever a resposta.

# Ajustes

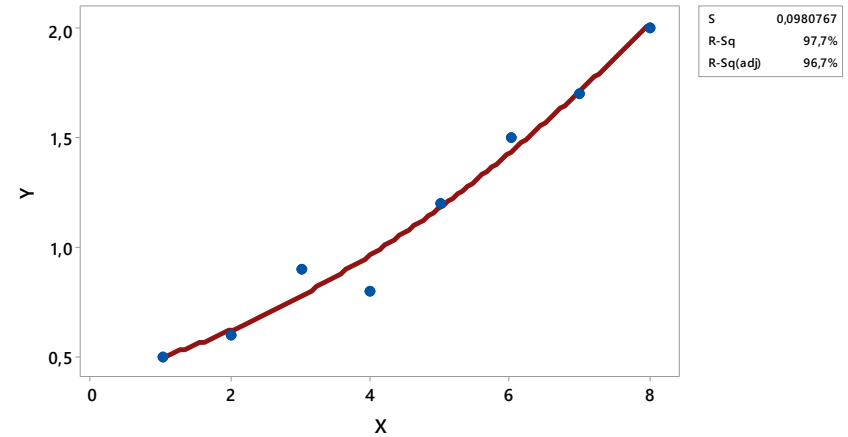
Fitted Line Plot

$$Y = 0,1750 + 0,2167 X$$



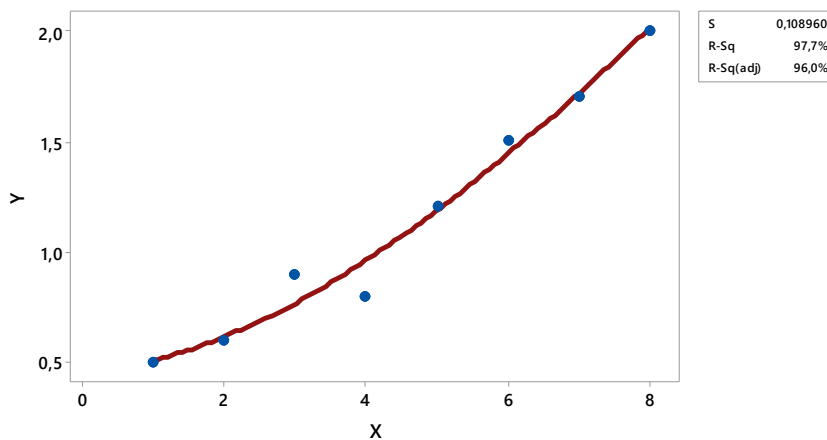
Fitted Line Plot

$$Y = 0,4071 + 0,07738 X + 0,01548 X^2$$



Fitted Line Plot

$$Y = 0,4571 + 0,0254 X + 0,02911 X^2 - 0,001010 X^3$$



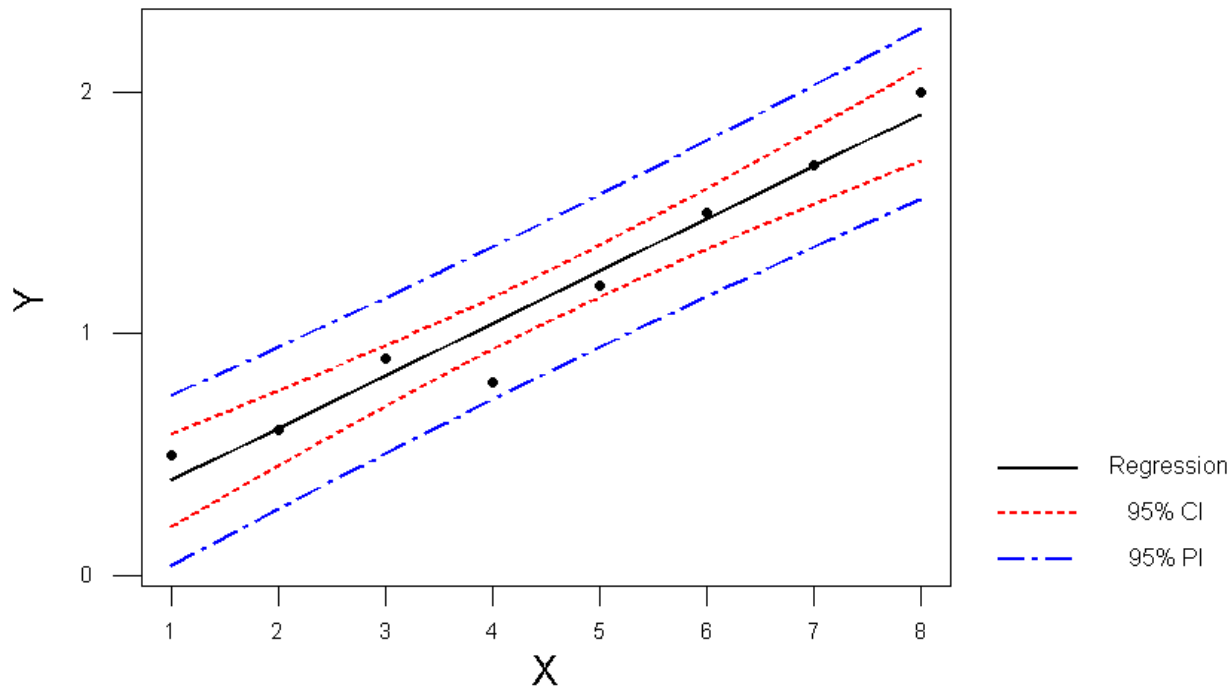
## Sequential Analysis of Variance

Source	DF	SS	F	P
Linear	1	1,97167	133,92	0,000
Quadratic	1	0,04024	4,18	0,096
Cubic	1	0,00061	0,05	0,832

# Intervalos de confiança e de previsão

$$Y = 0,175 + 0,216667 X$$

S = 0,121335   R-Sq = 95,7 %   R-Sq(adj) = 95,0 %



□ 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 a 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**.

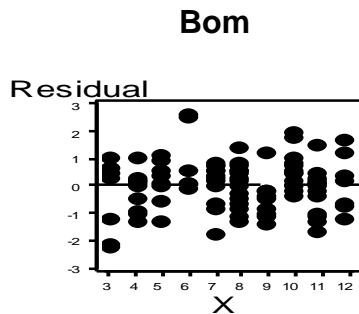
Faça a análise de Resíduos.

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

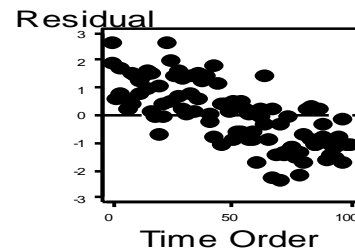
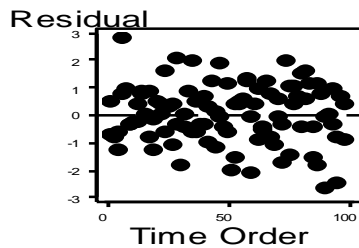
A equação final é adequada? Se não for, como melhorá-la?

# Análise de Resíduos

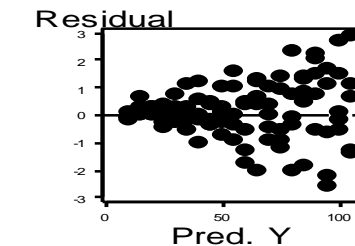
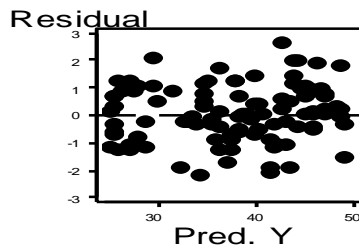
Residuals vs  
Each X



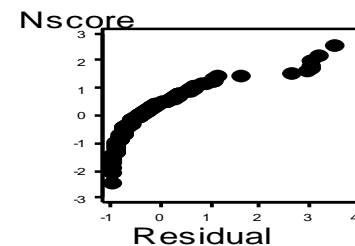
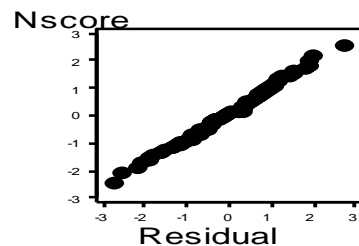
Time Plot of  
Residuals



Residuals vs  
Predicted Y  
(Fits)



Normal  
Probability Plot  
of Residuals



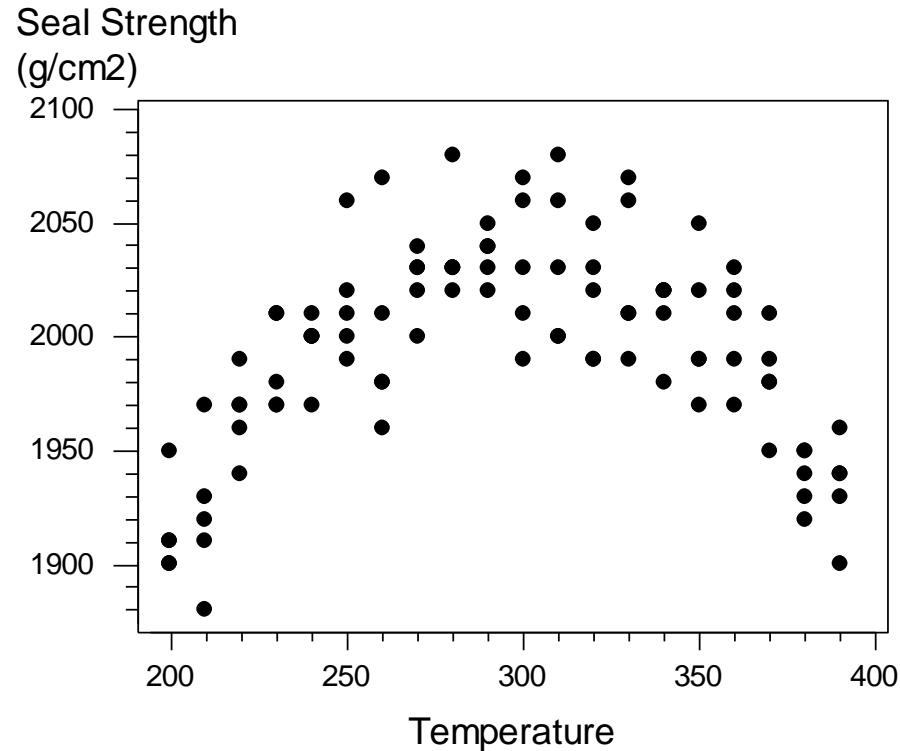
Nos casos ruins tente  
uma transformação  
em X, em Y ou  
ambos. Use *Box-Cox  
Transformation*

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**

# Regressão Curvilínea

The regression equation is

$$\text{SealStrength} = 923 + 7.45 \text{ Temperature} - 0.0125 \text{ TempSqrd}$$

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	F	P
Regression	2	139321	69661	109.87
Residual Error	97	61498	634	0.000
Total	99	200819		

Source	DF	Seq SS
Temperat	1	3051
TempSqrd	1	136270

X e X<sup>2</sup> são fortemente correlacionados. Nenhuma surpresa

**Conclusão:** Existe uma curvatura significativa

# Regressão Múltipla

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

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

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

$$\text{Tempo} = 18,4 - 0,106 \text{ Temperatura} - 4,68 \text{ Pressão}$$

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

Menores  
que 0,05

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

**Maior melhor**

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

# Regressão para muitos fatores

**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**