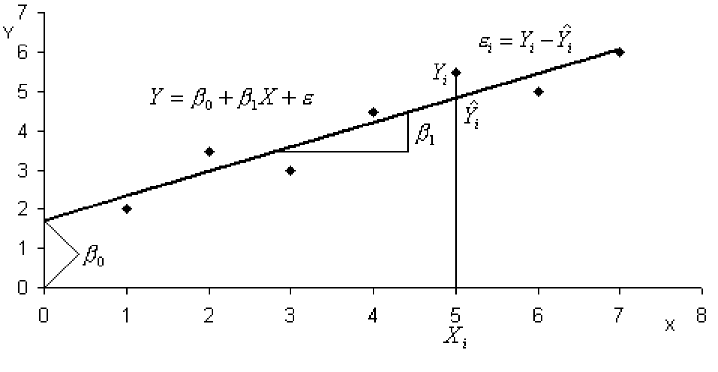
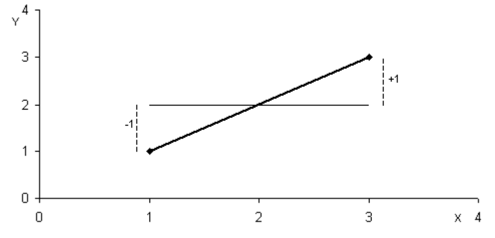
REGRESSÃO E CORRELAÇÃO

AJUSTE DE UMA RETA

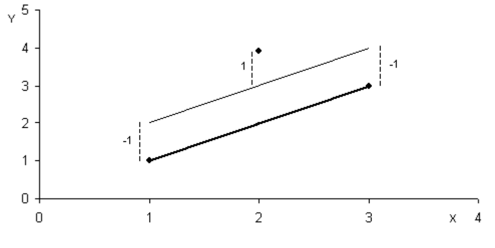


MINIMIZAÇÃO DOS DESVIOS



$$\sum\left(Y_i - \hat{Y}_i\right)$$

MINIMIZAÇÃO DOS DESVIOS ABSOLUTOS



$$\sum\left|Y_i - \hat{Y}_i\right|$$

EXEMPLO 1

- Considere o seguinte conjunto de pontos

X	Y
1	1
2	1
3	2
4	2
5	4

RETAS DE AJUSTE

R1	$Y=-0.1+0.7X$
R2	$Y= 0.5+0.5X$
R3	$Y=-0.7+0.9X$

RETAS

R1	R2	R3
0.6	1	0.2
1.3	1.5	1.1
2	2	2
2.7	2.5	2.9
3.4	3	3.8

DESVIOS

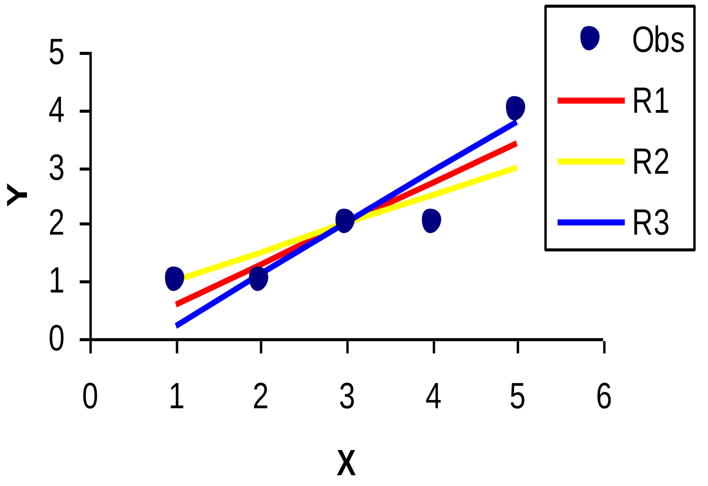
Desv1	Desv2	Desv3
0.4	0	0.8
-0.3	-0.5	-0.1
0	0	0
-0.7	-0.5	-0.9
0.6	1	0.2
0	0	0

DESVIOS ABSOLUTOS

|Desv1|	|Desv2|	|Desv3|
0.4	0	0.8
0.3	0.5	0.1
0	0	0
0.7	0.5	0.9
0.6	1	0.2
2	2	2

QUADRADO DOS DESVIOS

$(Desv1)^2$	$(Desv2)^2$	$(Desv3)^2$
0.16	0	0.64
0.09	0.25	0.01
0	0	0
0.49	0.25	0.81
0.36	1	0.04
1.10	1.50	1.50

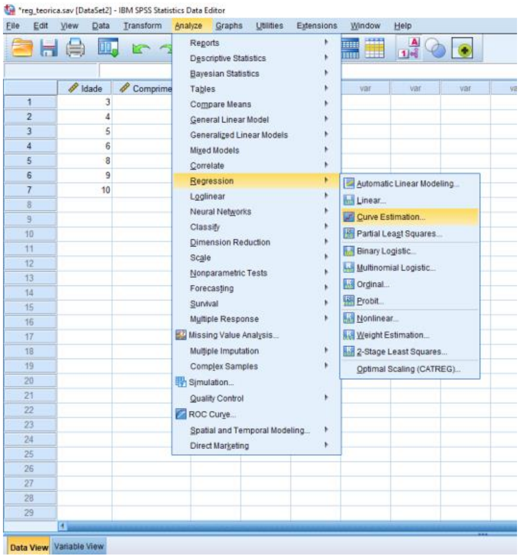


EXEMPLO 2

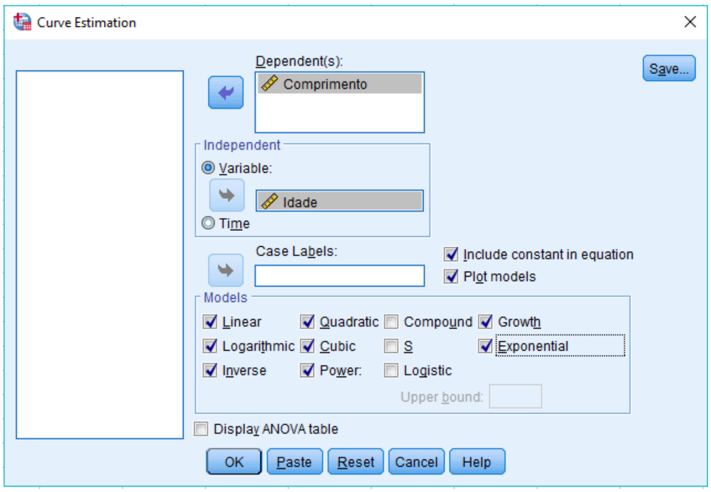
Comprimento alar (cm) em função da idade (dias) para andorinhas

Dias	Comp.
3	1,4
4	1,5
5	2,1
6	2,4
8	3,1
9	3,2
10	3,3

EXEMPLO 2 (SPSS)



EXEMPLO 2



Models (curve estimation algorithms)

Previous Next

CURVEFIT allows the user to specify a model with or without a constant term designated by β_0. If this constant term is excluded, simply set it zero or one depending upon whether it appears in an additive or multiplicative manner in the models listed below.

(1) Linear	$E(Y_t) = \beta_0 + \beta_1 t$
(2) Logarithmic	$E(Y_t) = \beta_0 + \beta_1 \ln(t)$
(3) Inverse	$E(Y_t) = \beta_0 + \beta_1 / t$
(4) Quadratic	$E(Y_t) = \beta_0 + \beta_1 t + \beta_2 t^2$
(5) Cubic	$E(Y_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$
(6) Compound	$E(Y_t) = \beta_0 \beta_1^t$
(7) Power	$E(Y_t) = \beta_0 t^{\beta_1}$
(8) S	$E(Y_t) = \exp(\beta_0 + \beta_1 / t)$
(9) Growth	$E(Y_t) = \exp(\beta_0 + \beta_1 t)$
(10) Exponential	$E(Y_t) = \beta_0 e^{\beta_1 t}$
(11) Logistic	$E(Y_t) = \left(\frac{1}{u} + \beta_0 \beta_1^t\right)^{-1}$

EXEMPLO 2

*Output1 [Document1] - IBM SPSS Statistics Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help

Output
- Log
- Curve Fit
  - Title
  - Notes
  - Active Dataset
  - Model Description
  - Case Processing
  - Variable Processi
  - Model Summary a
  - Curvefit for Compr

Curve Fit

[DataSet2] C:\Users\ACB\OneDrive\Aulas2017_18\reg_teorica.sav

Model Description

Model Name		MOD_1
Dependent Variable	1	Comprimento
Equation	1	Linear
	2	Logarithmic
	3	Inverse
	4	Quadratic
	5	Cubic
	6	Power[a]
	7	Growth[a]
	8	Exponential[a]
Independent Variable		Idade
Constant		Included
Variable Whose Values Label Observations in Plots		Unspecified
Tolerance for Entering Terms in Equations		0,0001

a. The model requires all non-missing values to be positive.

- Output
  - Log
  - Curve Fit
    - Title
    - Notes
    - Active Dataset
    - Model Description
    - Case Processing
    - Variable Processi
    - Model Summary a
    - Curvefit for Compr

Variable Processing Summary

		Variables	
		Dependent Comprimento	Independent Idade
Number of Positive Values		7	7
Number of Zeros		0	0
Number of Negative Values		0	0
Number of Missing Values	User-Missing	0	0
	System-Missing	0	0

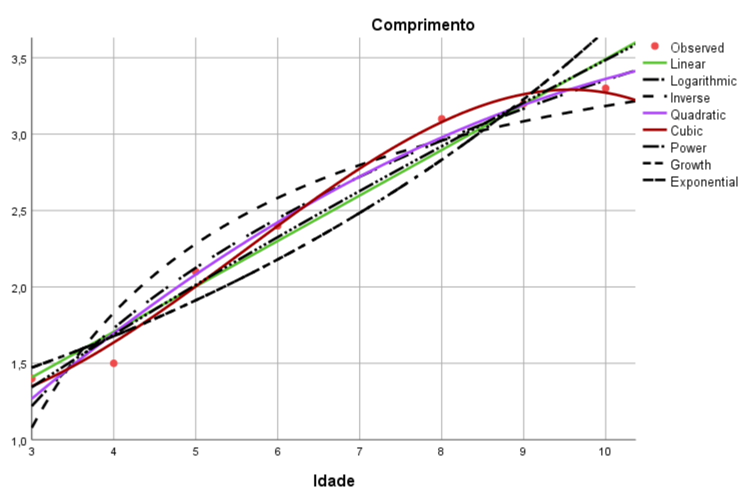
Model Summary and Parameter Estimates

Dependent Variable: Comprimento

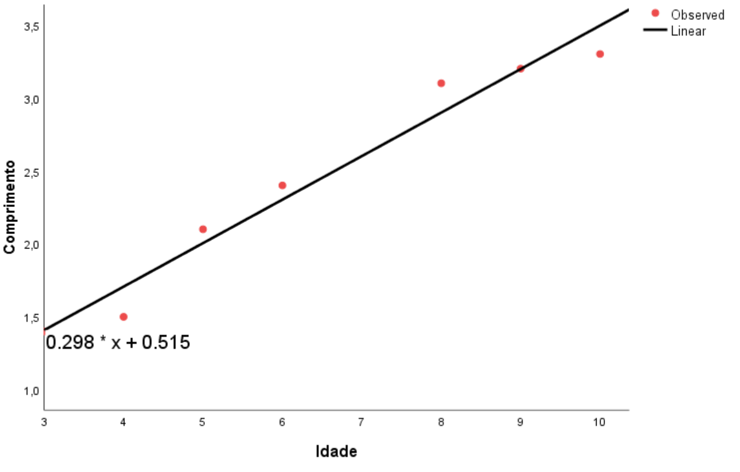
Equation	Model Summary: R Square	F	df1	df2	Sig.	Parameter Estimates: Constant	b1	b2	b3
Linear	$\boxed{0,964}$	132,174	1	5	0,000	0,515	0,298		
Logarithmic	0,971	165,753	1	5	0,000	-0,727	1,772		
Inverse	0,915	53,833	1	5	0,001	4,087	-9,026		
Quadratic	$\boxed{0,980}$	99,685	2	4	0,000	-0,274	0,579	-0,021	
Cubic	$\boxed{0,991}$	106,896	3	3	0,002	1,471	-0,387	0,141	-0,008
Power	0,968	149,638	1	5	0,000	0,563	0,792		
Growth	0,931	67,190	1	5	0,000	-0,006	0,131		
Exponential	0,931	67,190	1	5	0,000	0,994	0,131		

The independent variable is Idade.

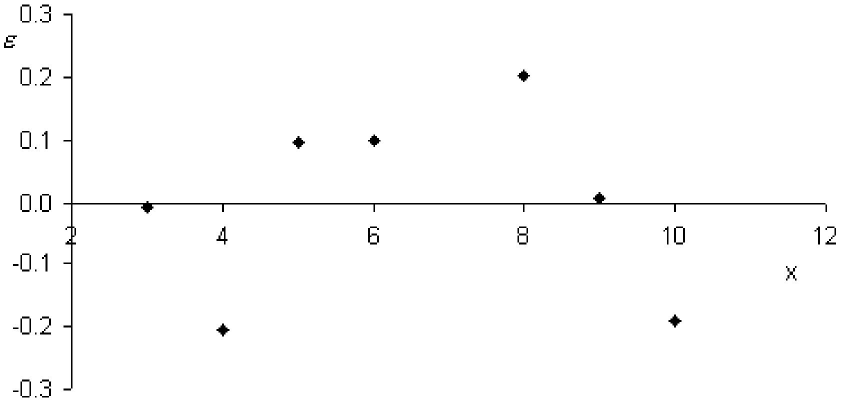
EXEMPLO 2



RECTA DE MÍNIMOS QUADRADOS



RESÍDUOS



Estimadores

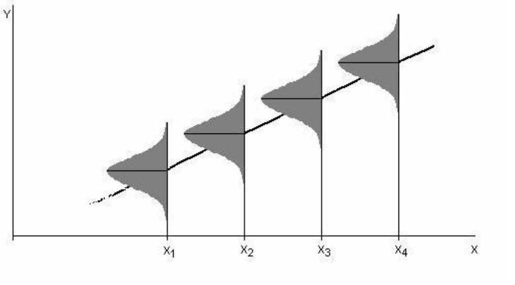
$$Y_i = \beta_0 + \beta_1.\left(X_i - \overline{X}\right) + \varepsilon_i \qquad \text{i}=1,\ldots,\text{n}$$

β_0 $\quad \boxed{\hat{\beta}_0 = \frac{1}{n}\sum_i Y_i = \overline{Y}}$

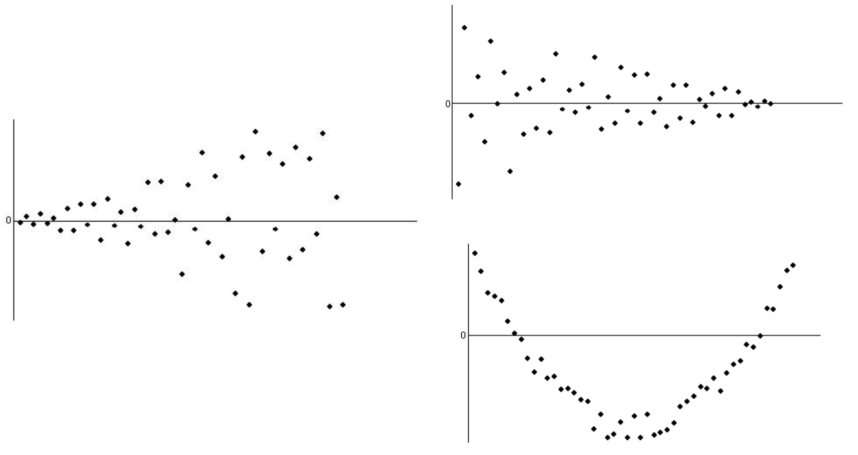
β_1 $\quad \boxed{\hat{\beta}_1 = \frac{\sum_i \left(X_i - \overline{X}\right).\left(Y_i - \overline{Y}\right)}{\sum_i \left(X_i - \overline{X}\right)^2} = \frac{s_{XY}}{s_{xx}}}$

σ^2 $\quad \boxed{s^2 = \frac{1}{n-2}\sum_i \hat{e}_i^2 = \frac{1}{n-2}\sum_i \left\{Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1.\left(X_i - \overline{X}\right)\right]\right\}^2}$

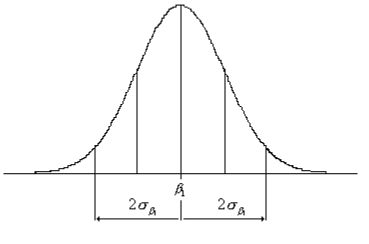
DISTRIBUIÇÃO DOS ERROS



RESÍDUOS



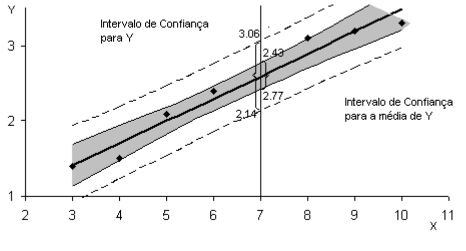
DISTRIBUIÇÃO DO DECLIVE



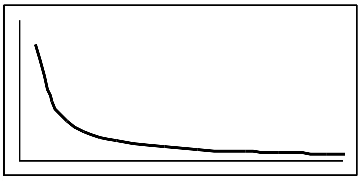
IC e Testes de hipóteses

	IC	TH
β_0	$\hat{\beta}_0 \pm t_{n-2,\left(\alpha/2\right)} \cdot \frac{s}{\sqrt{n}}$	$H_0 : \beta_0 = b_0$ $H_1 : \beta_0 \neq b_0, \beta_0 > b_0$ ou $\beta_0 < b_0$ $ET = \frac{\hat{\beta}_0 - b_0}{s/\sqrt{n}}$ H_0 verdadeira $\Rightarrow ET \sim t_{n-2}$
β_0'	$\left(\hat{\beta}_0 - \bar{X}.\hat{\beta}_1\right) \pm t_{n-2,\left(\alpha/2\right)}.s.\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{s_{XX}}}$	$H_0 : \beta_0' = b_0'$ $H_1 : \beta_0' \neq b_0', \beta_0' > b_0'$ ou $\beta_0' < b_0'$ $ET = \frac{\left(\hat{\beta}_0 - \bar{X}.\hat{\beta}_1\right) - b_0'}{s.\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{s_{XX}}}}$ H_0 verdadeira $\Rightarrow ET \sim t_{n-2}$
β_1	$\hat{\beta}_1 \pm t_{n-2,\left(\alpha/2\right)} \cdot \frac{s}{\sqrt{s_{XX}}}$	$H_0 : \beta_1 = b_{10}$ $H_1 : \beta_1 \neq b_{10}, \beta_1 > b_{10}$ ou $\beta_1 < b_{10}$ $ET = \frac{\hat{\beta}_1 - b_{10}}{\frac{s}{\sum_i \left(x_i - \bar{x}\right)^2}}$ H_0 verdadeira $\Rightarrow ET \sim t_{n-2}$

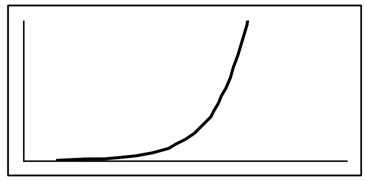
INTERVALO DE CONFIANÇA



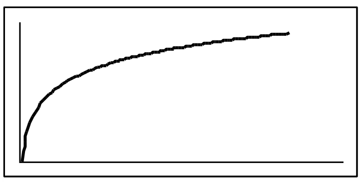
REGRESSÃO NÃO LINEAR



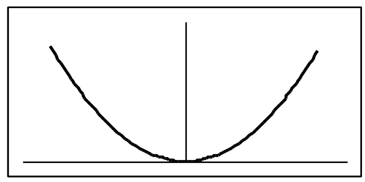
$\hat{Y} = \beta_0 + \beta_1 \frac{1}{X}$



$\hat{Y} = \beta_0 + \beta_1 e^X$



$\hat{Y} = \beta_0 + \beta_1 \ln X$



$\hat{Y} = \beta_0 + \beta_1 X^2$

REGRESSÃO NÃO LINEAR

Modelo	**Transformação**
• $Y_i = \alpha' + \frac{\beta}{X_i} + e_i$	$U_i = \frac{1}{X_i}$ $Y_i = \alpha' + \beta . U_i + e_i$
• $Y_i = e^{\alpha' + \beta . X_i + e_i}$	$Z_i = \ln Y_i$ $Z_i = \alpha' + \beta . X_i + e_i$
• $Y_i = e^{\alpha' + \frac{\beta}{X_i} + e_i}$ com $\alpha' > 0, \beta < 0$	$U_i = \frac{1}{X_i}$ $Z_i = \ln Y_i$ $Z_i = \alpha' + \beta . U_i + e_i$

REGRESSÃO LINEAR E MÚLTIPLA

Um modelo de regressão linear múltipla descreve uma relação entre várias variáveis quantitativas **independentes**, X_1, X_2, ..., X_j, e uma variável quantitativa **dependente**, Y, nos termos seguintes:

$$Y_i = \beta_0 + \beta_1.\left(X_{1i} - \overline{X}_1\right) + \beta_2.\left(X_{2i} - \overline{X}_2\right) + \ldots + \beta_j.\left(X_{ji} - \overline{X}_j\right) + \varepsilon_i \qquad i = 1,\ldots,n \quad j = 1,\ldots,J$$

onde:

- $(X_{1i}, X_{2i}, \ldots, X_{Ji}, Y_i)$ i-ésima observação das variáveis $X_{1i}, X_{2i}, \ldots, X_{Ji}$ e Y.
- $\overline{X}_j$ média aritmética das observações X_{ji}
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_J$ parâmetros fixos da relação linear entre $X_{1i}, X_{2i}, \ldots, X_{Ji}$ e Y
- ε_i erro aleatório associado ao valor observado Y_i

RESÍDUOS

Pressupostos para ε_i

- Têm valor esperado nulo e variância constante, σ^2;
- São mutuamente independentes;
- São normalmente distribuídos.

$\varepsilon_i \sim IN(0, \sigma^2)$

Se estas hipóteses se verificarem então: $Y_i \sim IN(\mu_{Y_i}, \sigma^2)$

ESTIMADORES MÍNIMOS QUADRADOS

$$\hat{\beta}_0 = \frac{1}{n}\sum_i Y_i = \overline{Y}$$

$$\begin{cases} \hat{\beta}_1.S_{X_1X_1} + \hat{\beta}_2.S_{X_1X_2} + \ldots + \hat{\beta}_J.S_{X_1X_J} = S_{X_1Y} \\ \hat{\beta}_1.S_{X_2X_1} + \hat{\beta}_2.S_{X_2X_2} + \ldots + \hat{\beta}_J.S_{X_2X_J} = S_{X_2Y} \\ (\ldots) \\ \hat{\beta}_1.S_{X_JX_1} + \hat{\beta}_2.S_{X_JX_2} + \ldots + \hat{\beta}_J.S_{X_JX_J} = S_{X_JY} \end{cases}$$

$$S_{X_{j_1}X_{j_2}} = \sum_n \left(X_{j_1i} - \overline{X}_{j_1}\right)\left(X_{j_2i} - \overline{X}_{j_2}\right)$$

$$S_{X_jY} = \sum_n \left(X_{ji} - \overline{X}_j\right)\left(Y_i - \overline{Y}\right)$$

$$s^2 = \frac{1}{n-J-1}\sum_i \hat{\varepsilon}_i^2 =$$
$$= \frac{1}{n-J-1}\sum_i \left\{Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1.\left(X_{1i} - \overline{X}_1\right) + \hat{\beta}_2.\left(X_{2i} - \overline{X}_2\right) + \ldots + \hat{\beta}_j.\left(X_{Ji} - \overline{X}_J\right)\right]\right\}^2$$

EXEMPLO 3

Determine a relação existente entre o calor envolvido no endurecimento, representado pela variável Y e os pesos de duas substâncias X_1 e X_2, tendo em consideração os seguintes valores obtidos numa experiência:

Y	78.5	74.3	104.3	87.6	95.6	109.2	102.7	72.5	93.1	115.9
X1	7	1	11	11	7	11	3	1	2	21
X2	26	29	59	31	52	55	71	31	54	47

COEFICIENTE DE CORRELAÇÃO

Coeficiente de correlação de Pearson

$$R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} = \frac{s_{XY}}{\sqrt{s_{xx}} \cdot \sqrt{s_{YY}}}$$

33

TESTES DE ASSOCIAÇÃO

Unilateral à direita	Unilateral à esquerda	Bilateral
$H_0 : \rho = 0$	$H_0 : \rho = 0$	$H_0 : \rho = 0$
$H_1 : \rho > 0$	$H_1 : \rho < 0$	$H_1 : \rho \neq 0$

Estatística de teste $t = \frac{r.\sqrt{n-2}}{\sqrt{1-r^2}}$

Região de Rejeição:

$t > t_{n-2,(\alpha)}$	$t < -t_{n-2,(\alpha)}$	$\lvert t \rvert > t_{n-2,(\alpha/2)}$

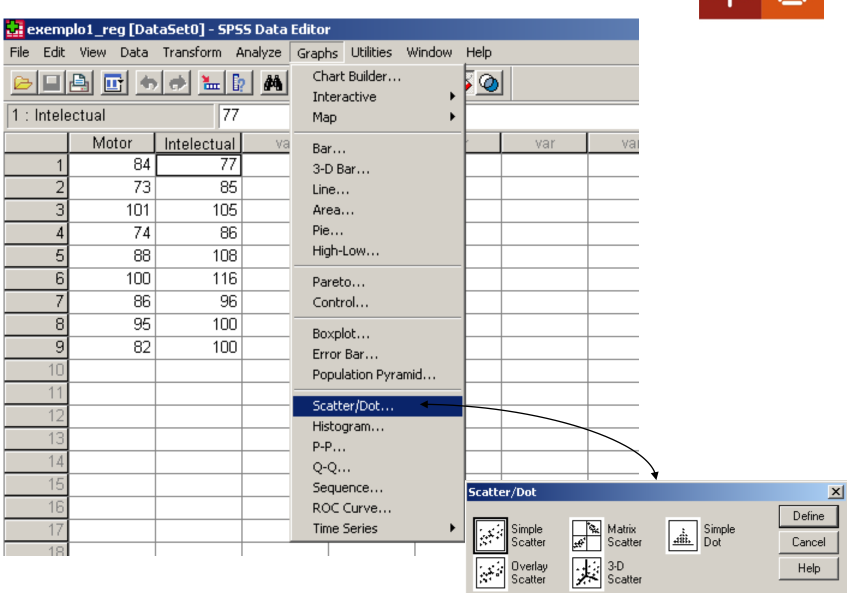
34

EXEMPLO

- Índice de Desenvolvimento de Griffiths
  - avaliações motora e intelectual para 9 crianças com a idade de 4 anos

Motor	Intelectual
84	77
73	85
101	105
74	86
88	108
100	116
86	96
95	100
82	100

35



36

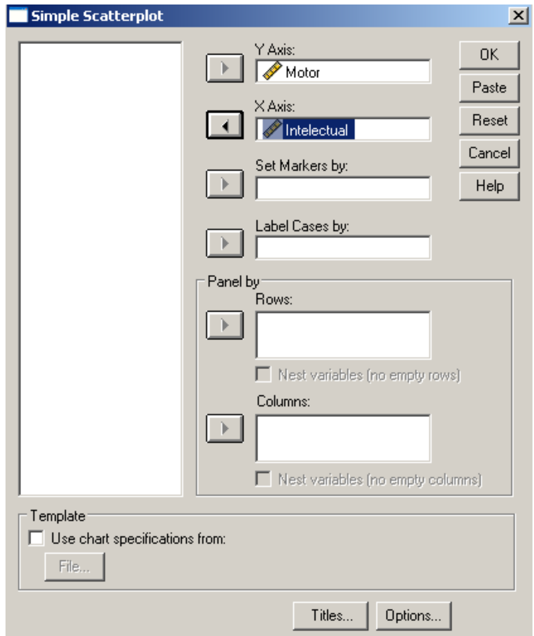
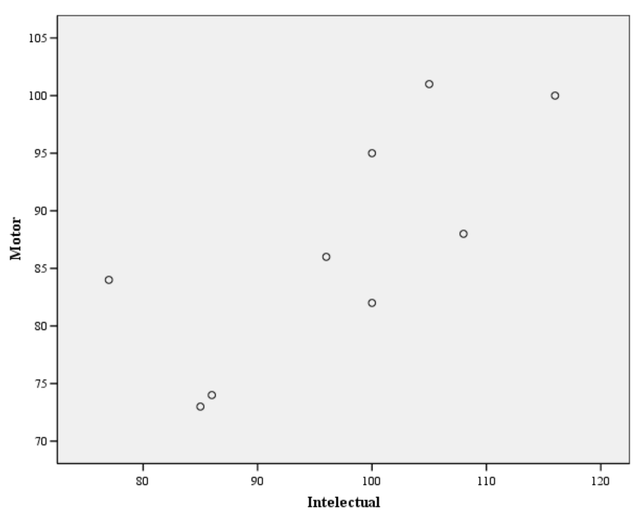
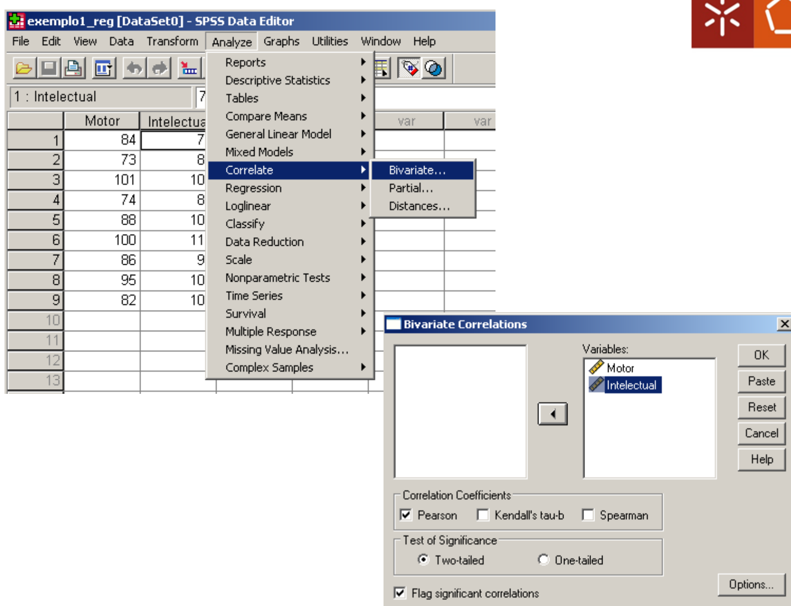
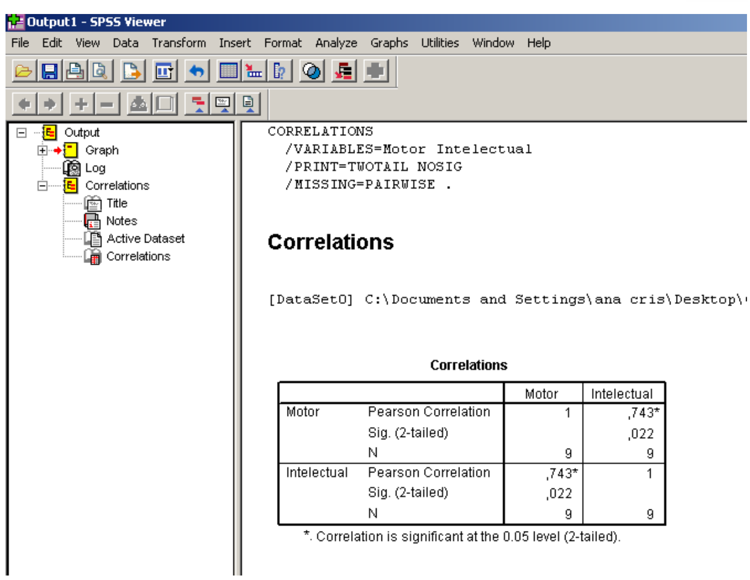


DIAGRAMA DE DISPERSÃO





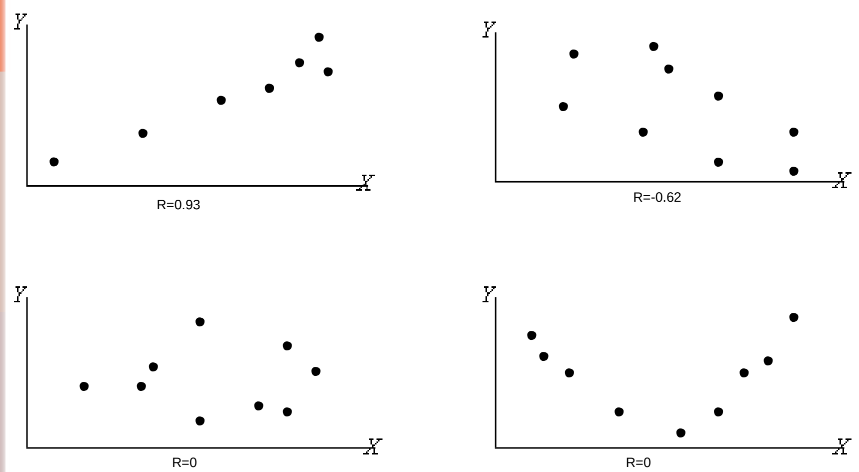


Correlations

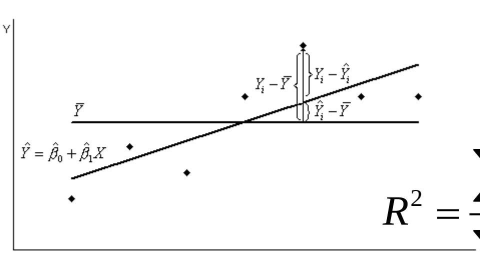
		Motor	Intelectual
Motor	Pearson Correlation	1	,743*
	Sig. (2-tailed)		,022
	N	9	9
Intelectual	Pearson Correlation	,743*	1
	Sig. (2-tailed)	,022	
	N	9	9

*. Correlation is significant at the 0.05 level (2-tailed).

CORRELAÇÃO



COEFICIENTE DE DETERMINAÇÃO



$$R^2 = \frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum\left(Y_i - \bar{Y}\right)^2}$$

Coeficiente de determinação (r^2), representa a proporção da variação de Y que é explicada pela regressão

$$r^2 = \frac{\hat{\beta}_1^2 . s_{XX}}{s_{YY}} = \frac{\hat{\beta}_1^2 . \sum_i \left(X_i - \bar{X}\right)^2}{\sum_i \left(Y_i - \bar{Y}\right)^2} = \frac{\text{variação de } Y \text{ explicada pela regressão}}{\text{variação total de } Y}$$