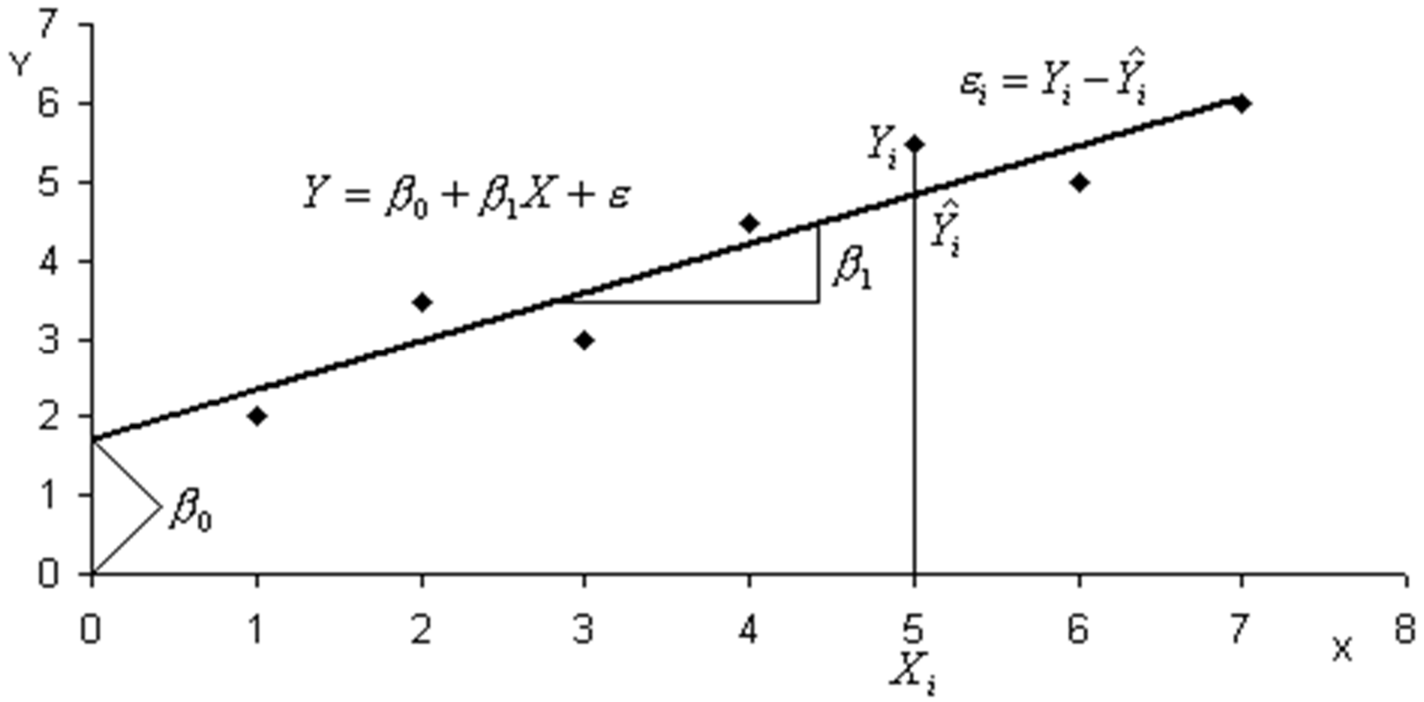
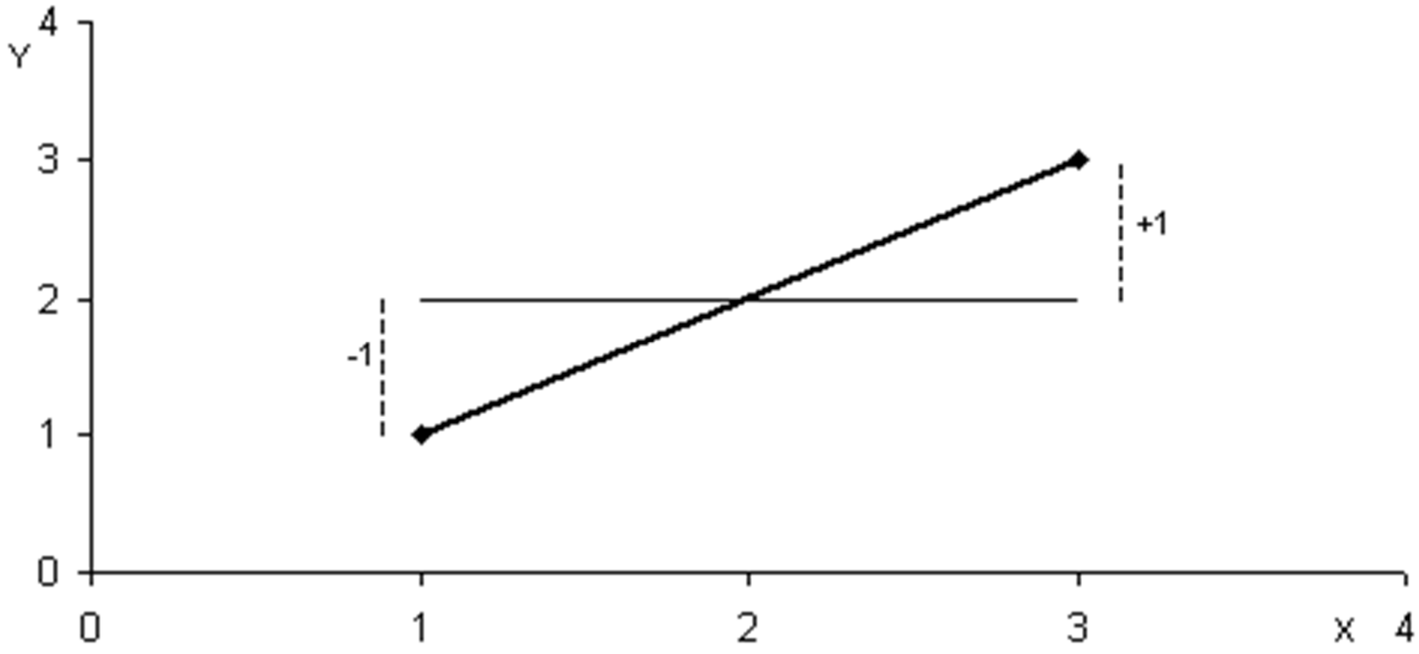
REGRESSÃO E CORRELAÇÃO

AJUSTE DE UMA RECTA

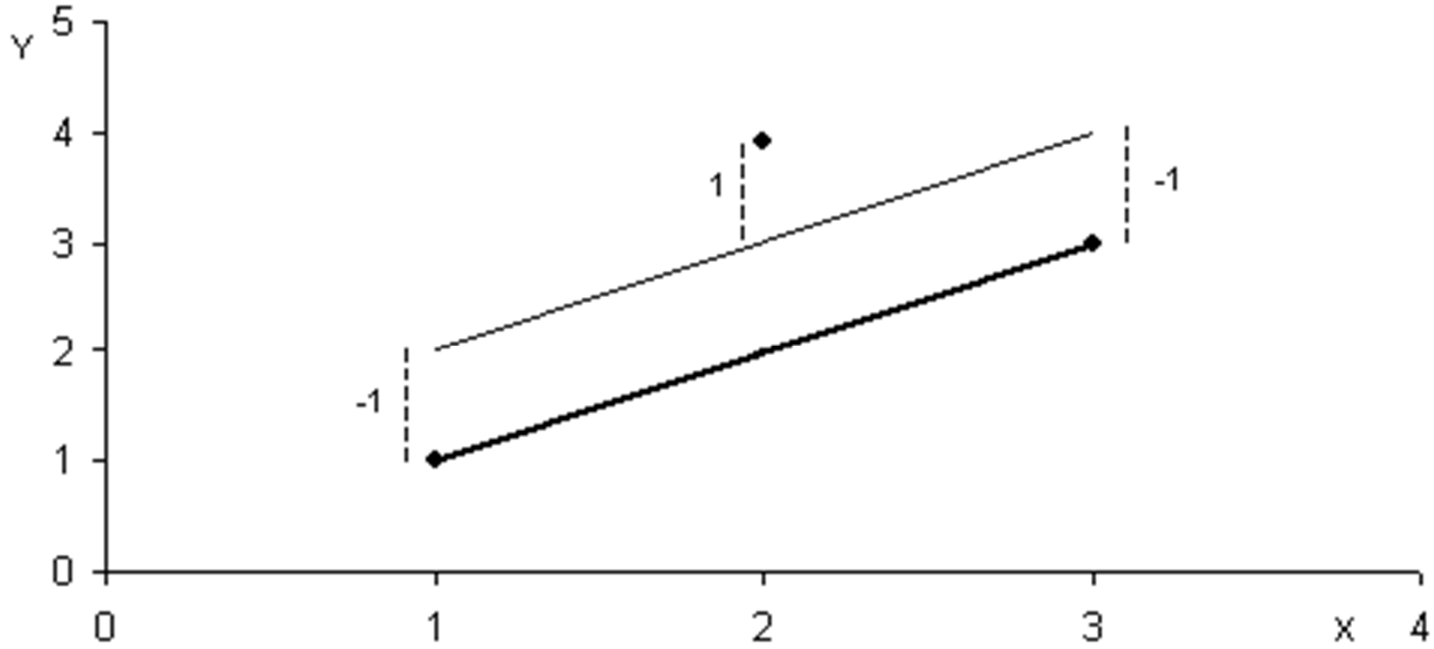


MINIMIZAÇÃO DOS DESVIOS



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

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

RECTAS DE AJUSTE

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

RECTAS

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

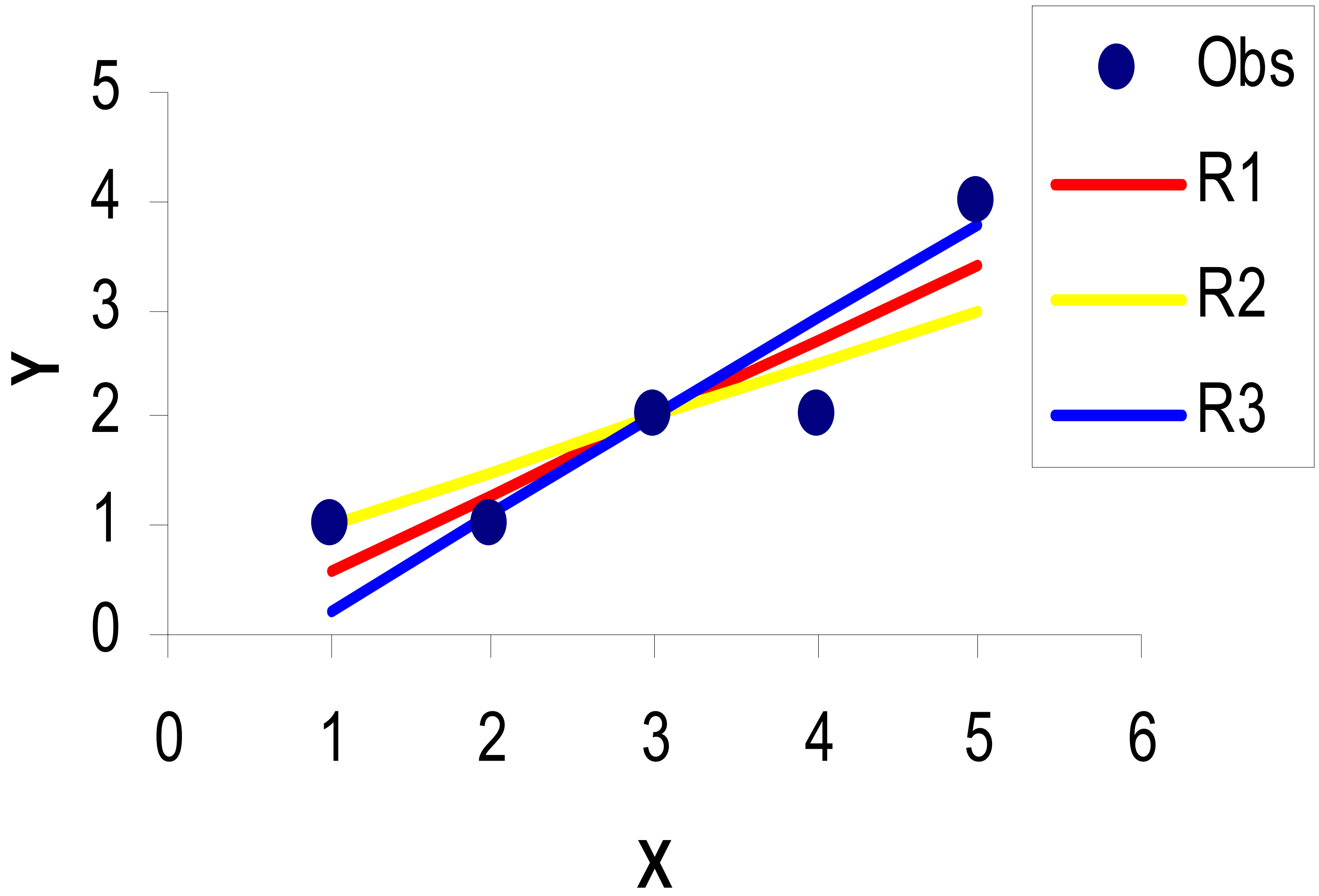
Desv1	Desv2	Desv3
0.4	0	0.8
-0.3	-0.5	-0.1
0	0	0
-0.7	-0.5	-0.9
0.6	1	0.2
0	0	0

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



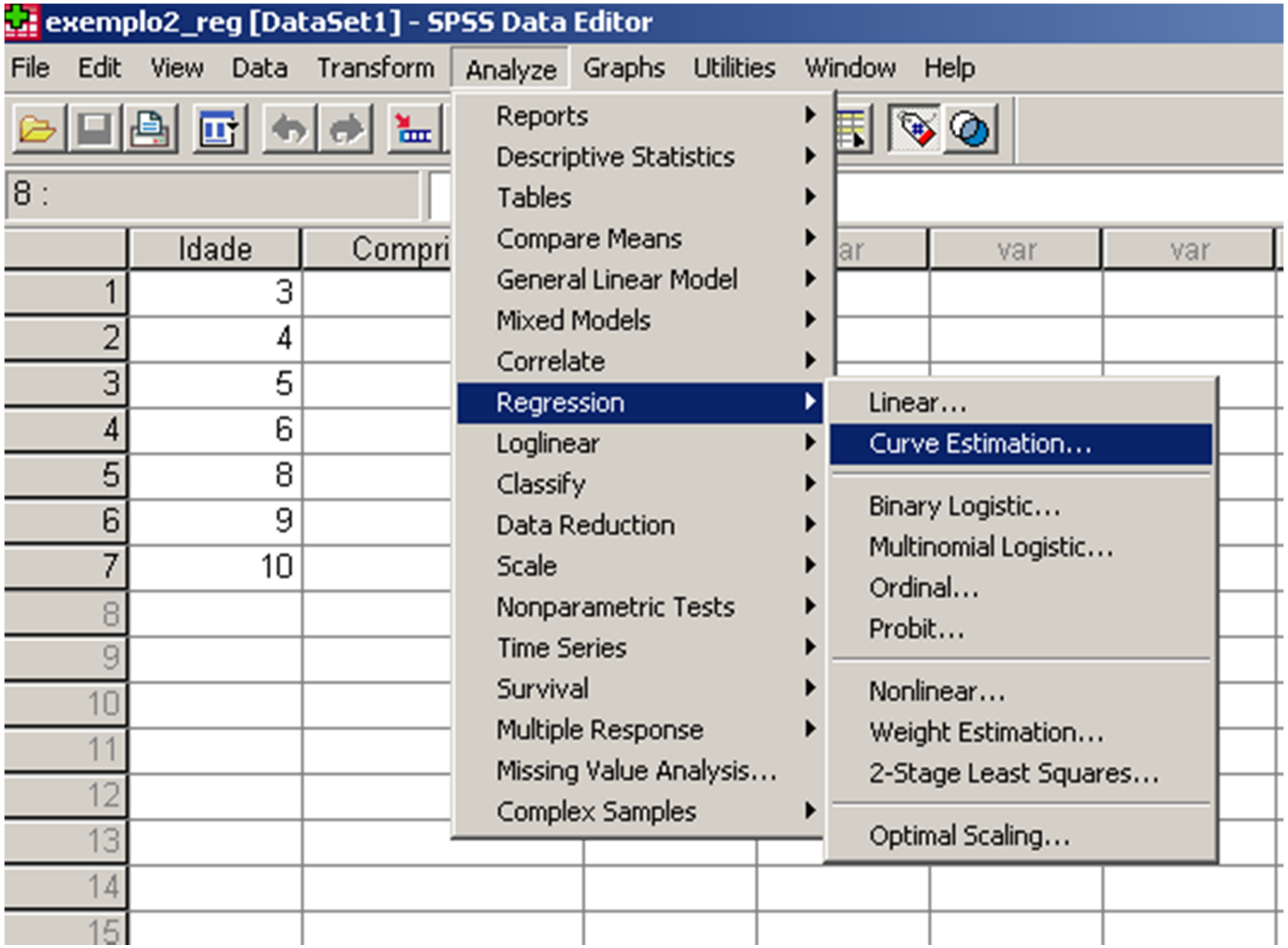
5
4
3
2
1
0
Y
0
1
2
3
4
5
6
X
Obs
R1
R2
R3

EXEMPLO 2

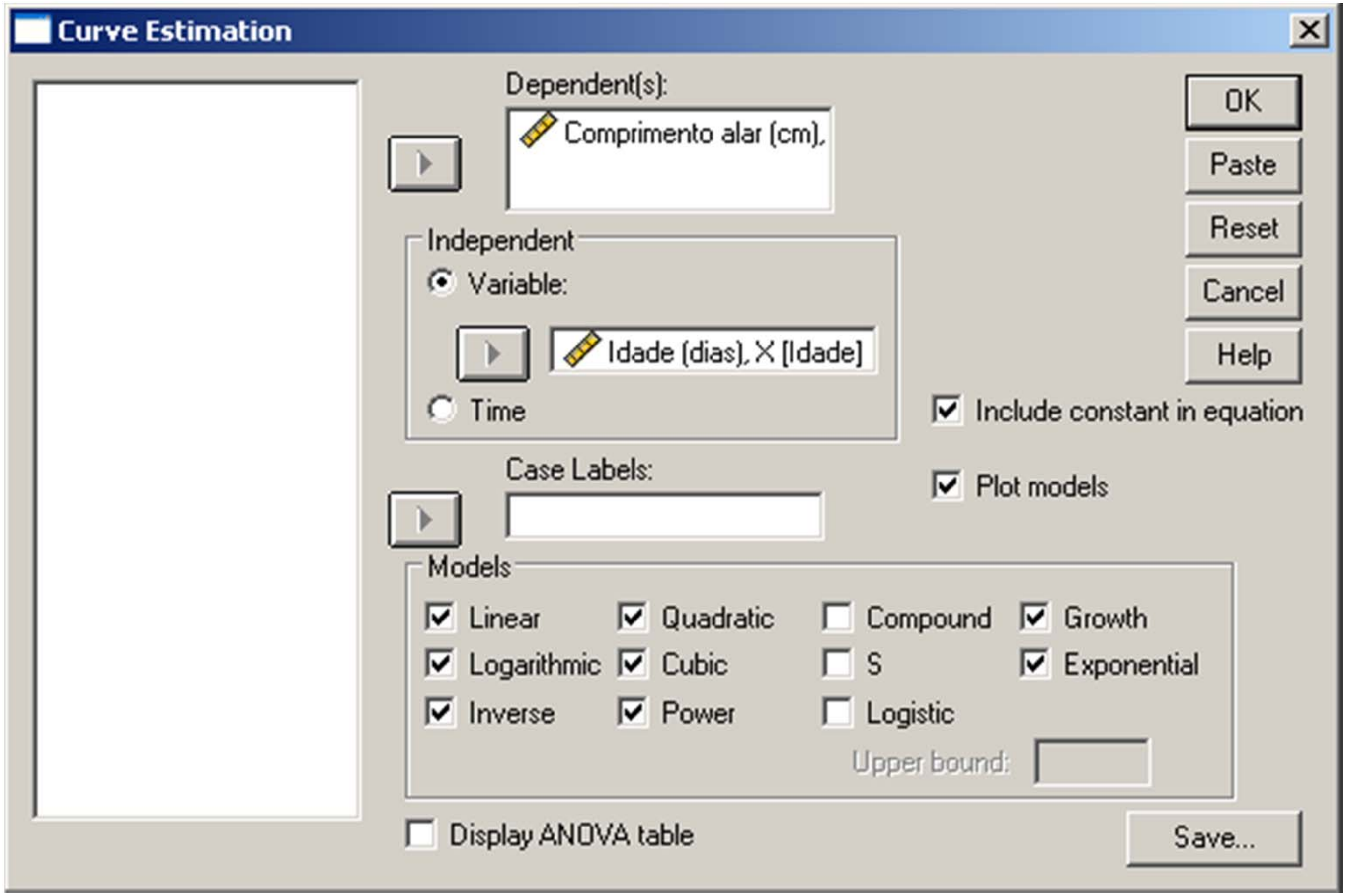
- Comprimento alar (cm) em função da idade (dias) para andorinhas

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

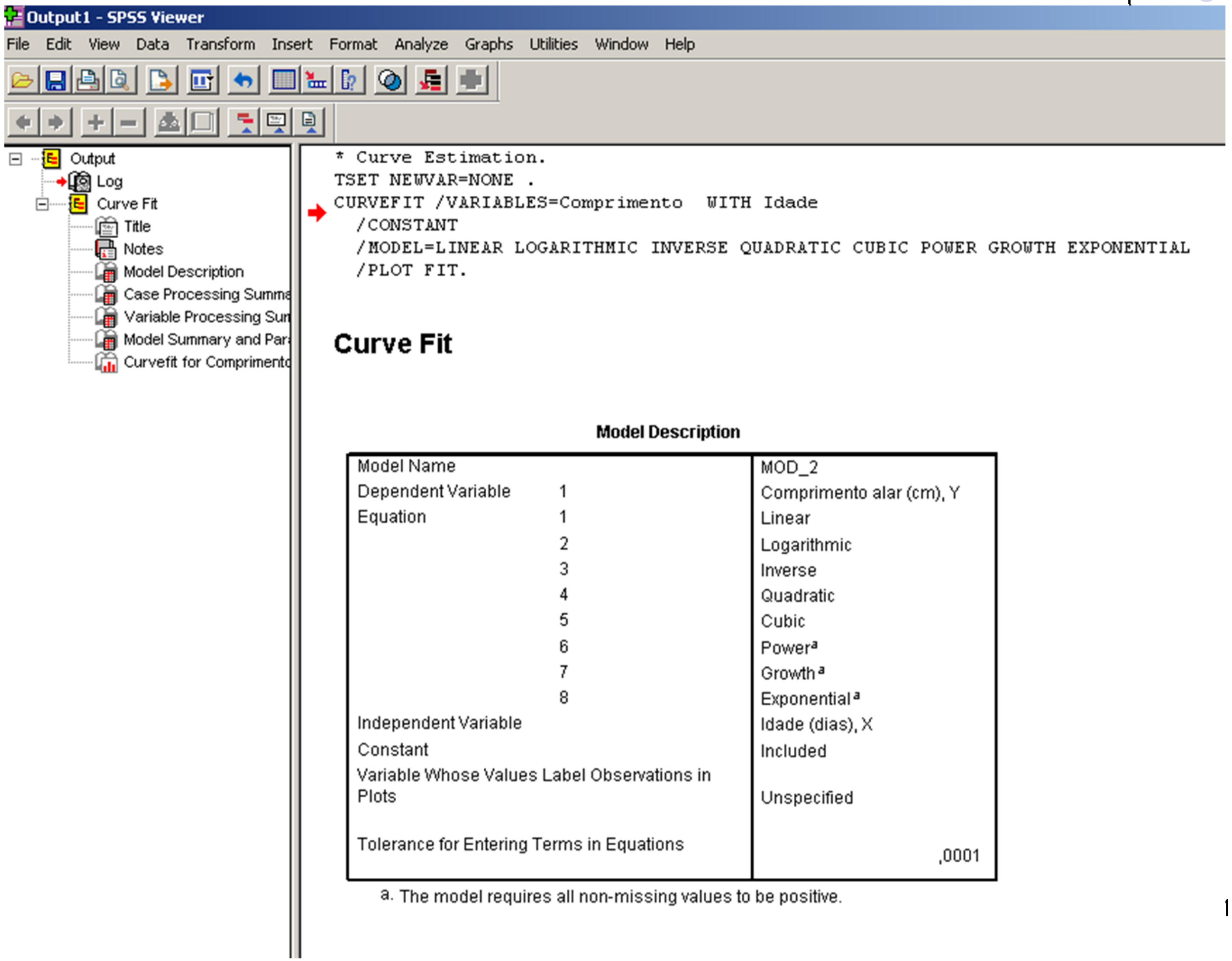
EXEMPLO 2


exemplo2_reg [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
8 :
Idade
Compri
var
var
1 3
2 4
3 5
4 6
5 8
6 9
7 10
8
9
10
11
12
13
14
15
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Linear...
Curve Estimation...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling...

EXEMPLO 2


Curve Estimation
Dependent(s):
Comprimento alar (cm),
OK
Paste
Reset
Cancel
Help
Independent
Variable:
Idade (dias), X [Idade]
Time
Include constant in equation
Case Labels:
Plot models
Models
Linear
Quadratic
Compound
Growth
Logarithmic
Cubic
S
Exponential
Inverse
Power
Logistic
Upper bound:
Display ANOVA table
Save...

EXEMPLO 2




```
* Curve Estimation.
TSET NEWVAR=NONE .
CURVEFIT /VARIABLES=Comprimento  WITH Idade
  /CONSTANT
  /MODEL=LINEAR LOGARITHMIC INVERSE QUADRATIC CUBIC POWER GROWTH EXPONENTIAL
  /PLOT FIT.
```

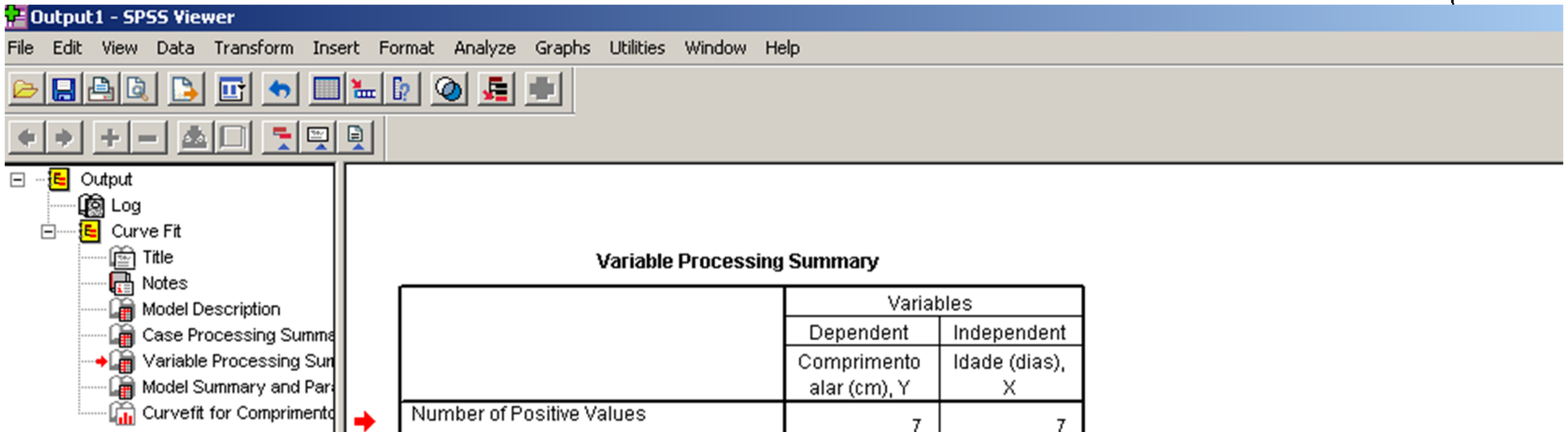
Curve Fit

Model Description

Model Name		MOD_2
Dependent Variable	1	Comprimento alar (cm), Y
Equation	1	Linear
	2	Logarithmic
	3	Inverse
	4	Quadratic
	5	Cubic
	6	Power[a]
	7	Growth[a]
	8	Exponential[a]
Independent Variable		Idade (dias), X
Constant		Included
Variable Whose Values Label Observations in Plots		Unspecified
Tolerance for Entering Terms in Equations		,0001

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

EXEMPLO 2



Variable Processing Summary

		Variables	
		Dependent	Independent
		Comprimento alar (cm), Y	Idade (dias), X
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

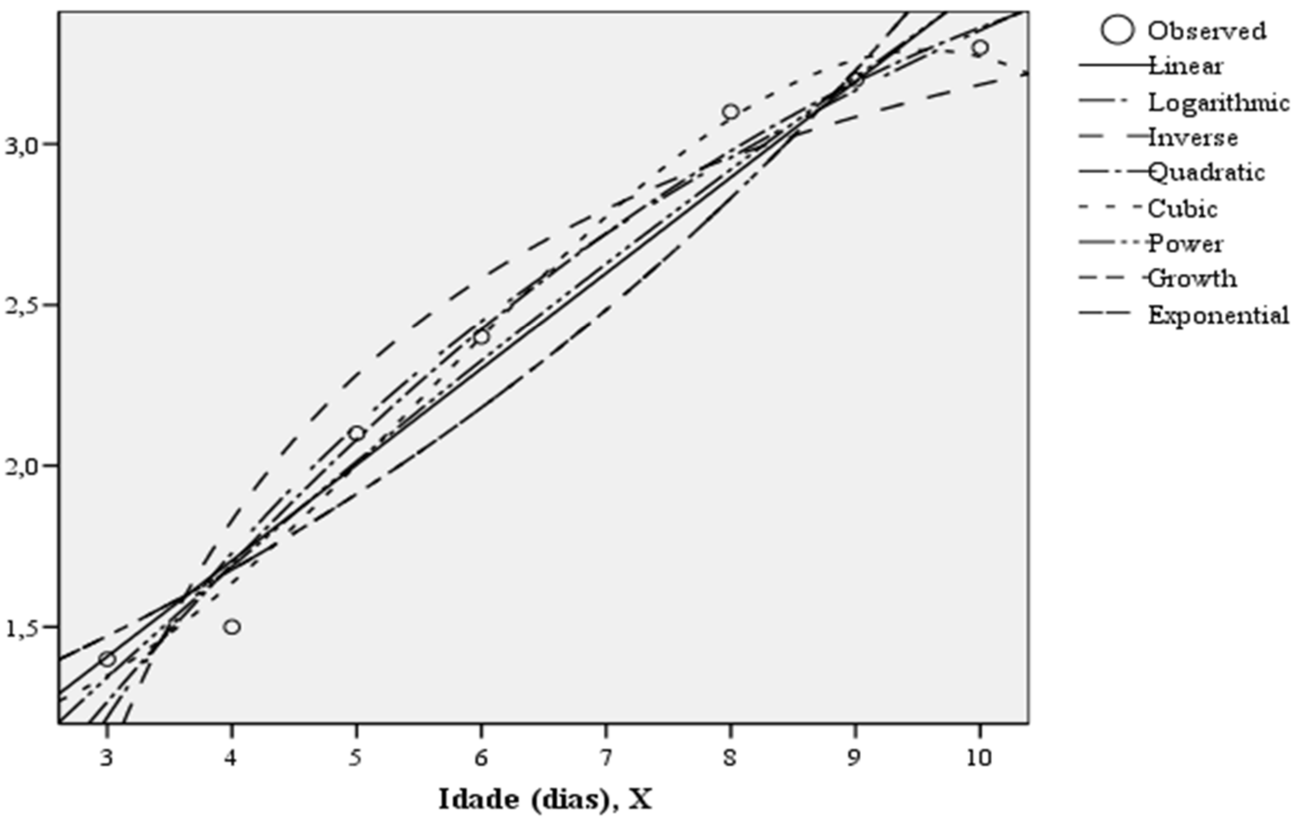
Model Summary and Parameter Estimates

Dependent Variable: Comprimento alar (cm), Y

	Model Summary					Parameter Estimates			
Equation	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	$\boxed{,964}$	132,174	1	5	,000	,515	,298		
Logarithmic	,971	165,753	1	5	,000	-,727	1,772		
Inverse	,915	53,833	1	5	,001	4,087	-9,026		
Quadratic	$\boxed{,980}$	99,685	2	4	,000	-,274	,579	-,021	
Cubic	$\boxed{,991}$	106,896	3	3	,002	1,471	-,387	,141	-,00
Power	,968	149,638	1	5	,000	,563	,792		
Growth	,931	67,190	1	5	,000	-,006	,131		
Exponential	,931	67,190	1	5	,000	,994	,131		

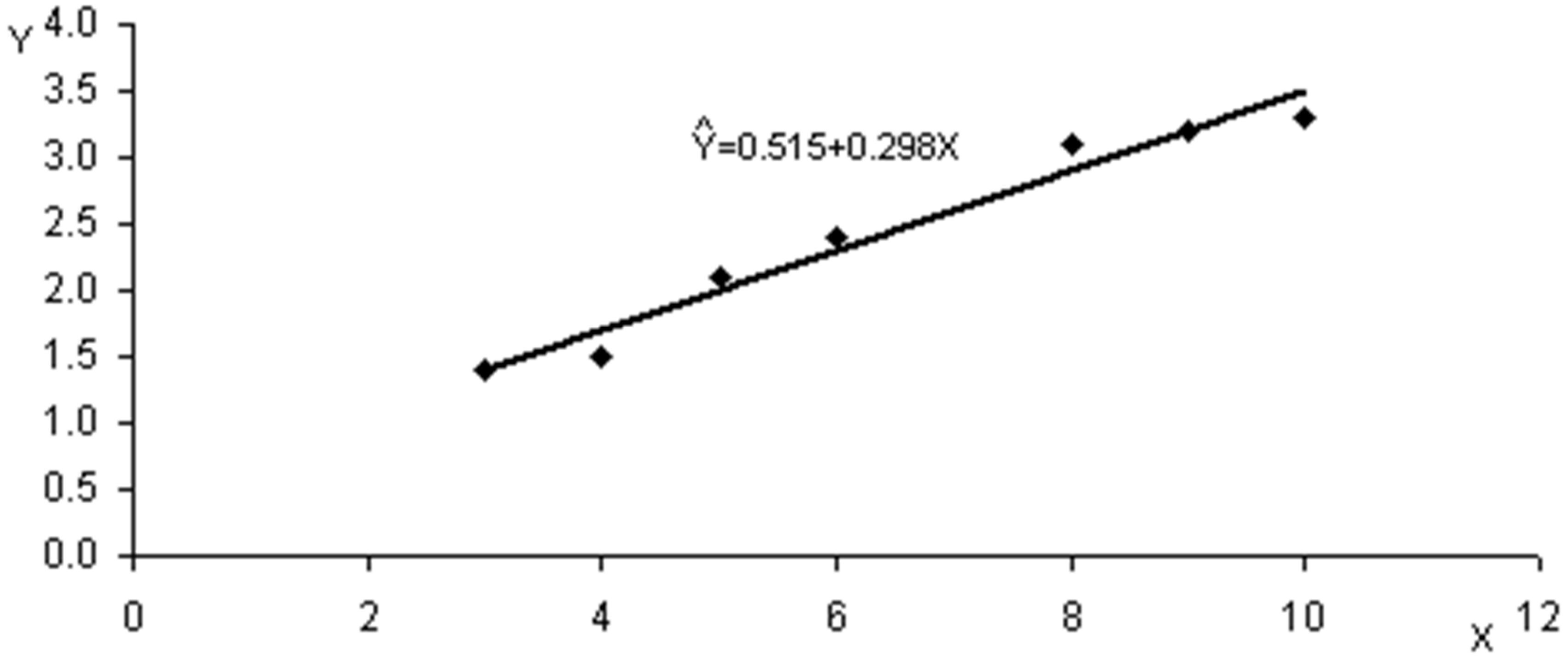
The independent variable is Idade (dias), X.

EXEMPLO 2

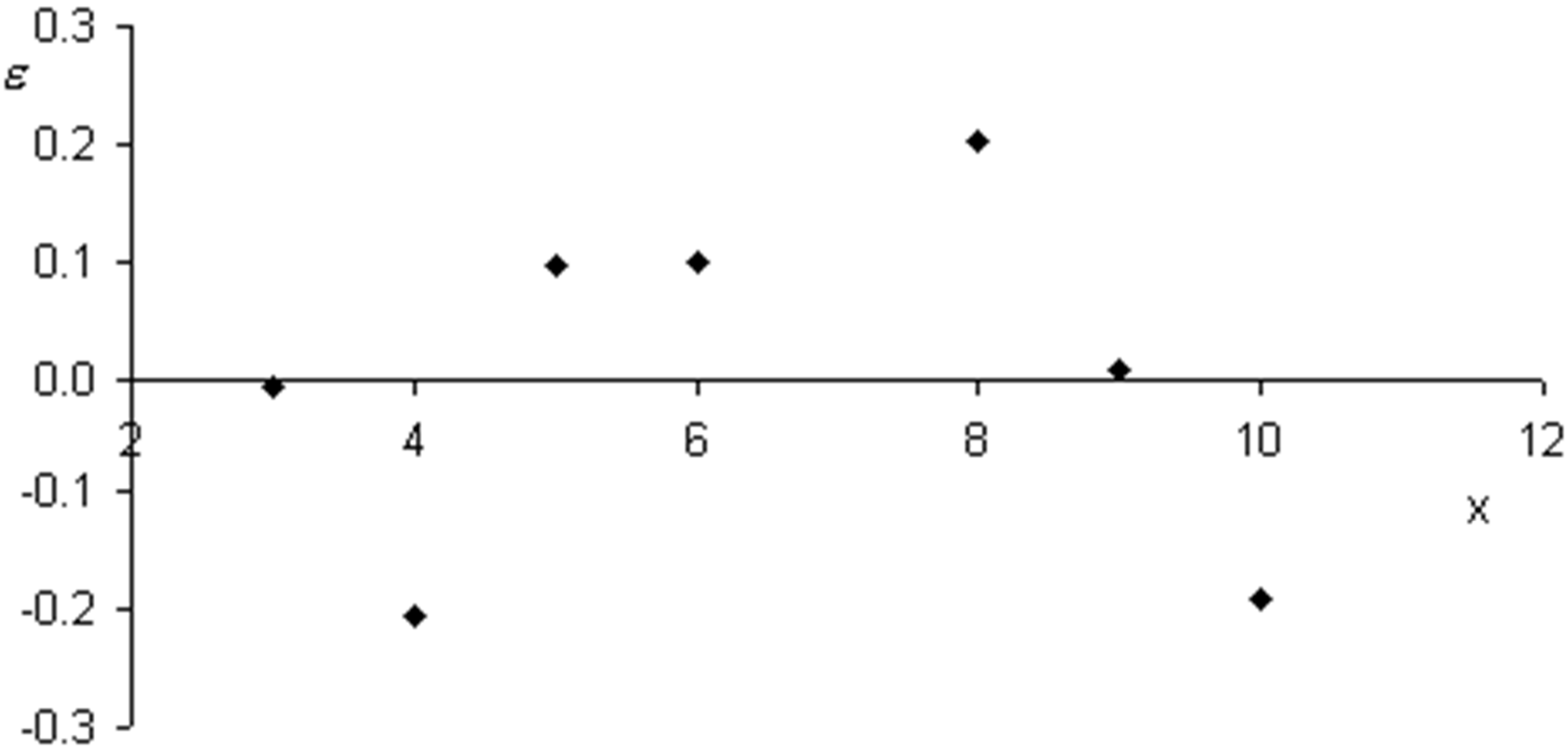


3,0
2,5
2,0
1,5
3
4
5
6
7
8
9
10
Idade (dias), X
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Power
Growth
Exponential

RECTA DE MÍNIMOS QUADRADOS



RESÍDUOS



Estimadores

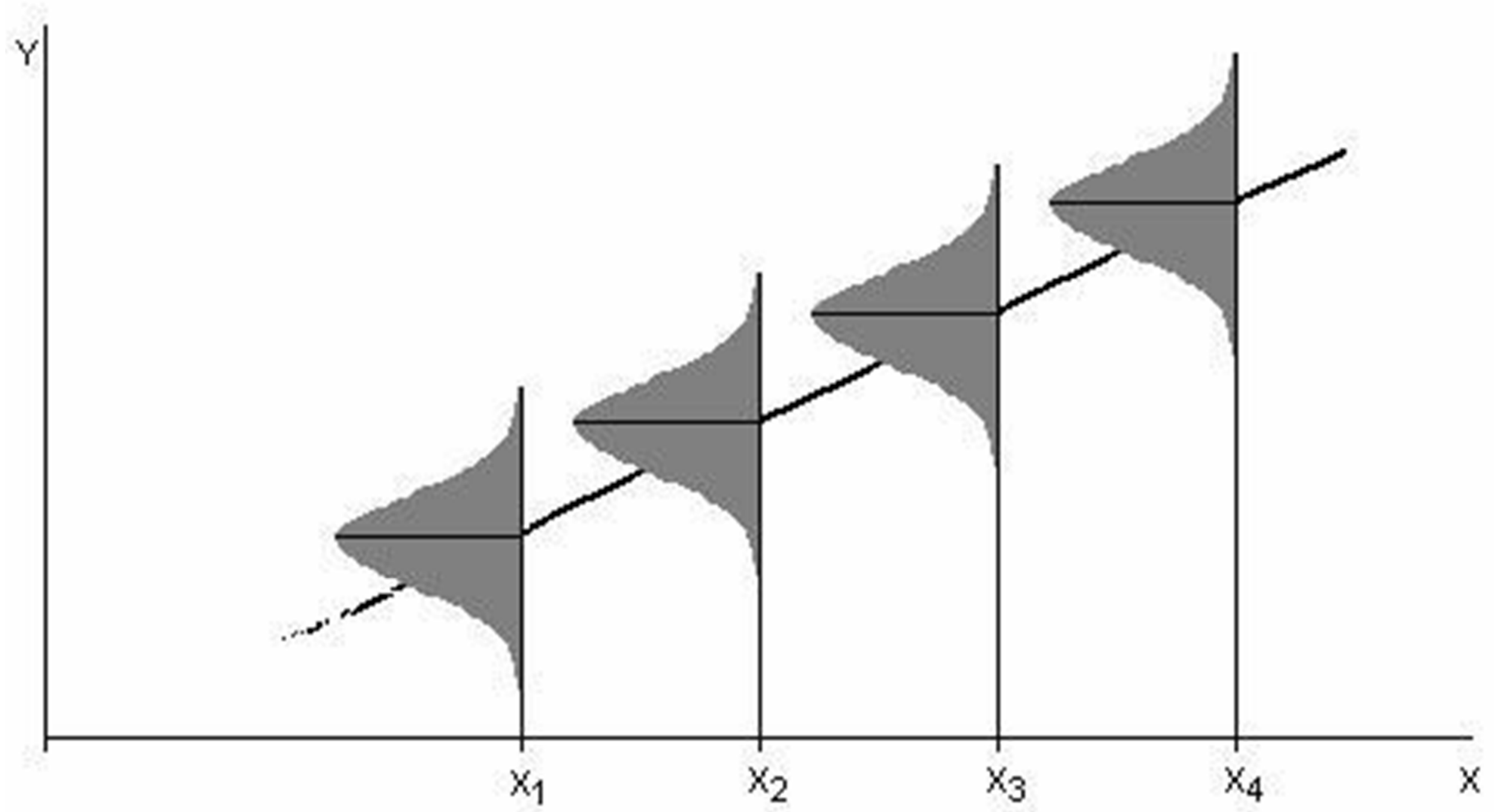
$$Y_i = \beta_0 + \beta_1 . \left(X_i - \overline{X} \right) + \varepsilon_i \qquad \text{i} = 1,\ldots,\text{n}$$

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

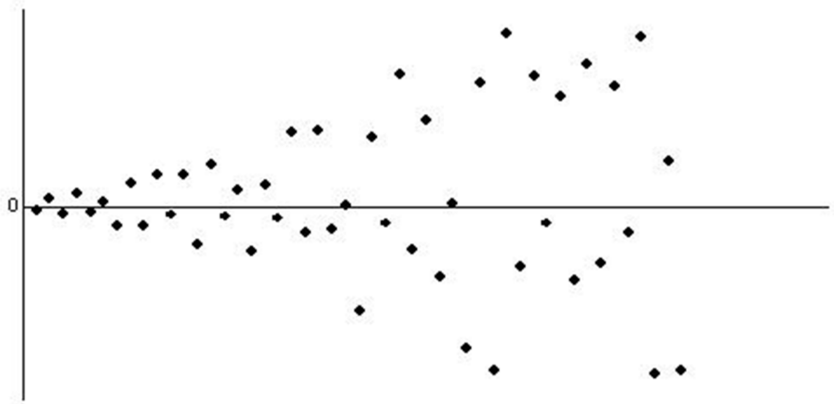
β_1 $$\boxed{\hat{\beta}_1 = \frac{\sum_i \left(X_i - \overline{X} \right) . \left(Y_i - \overline{Y} \right)}{\sum_i \left(X_i - \overline{X} \right)^2} = \frac{s_{XY}}{s_{xx}}}$$

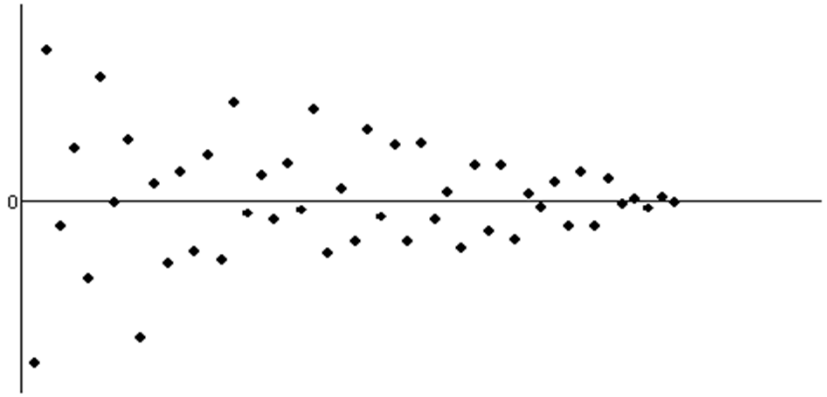
σ^2 $$\boxed{s^2 = \frac{1}{n-2} \sum_i \hat{e}_i^2 = \frac{1}{n-2} \sum_i \left\{ Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 . \left(X_i - \overline{X} \right) \right] \right\}^2}$$

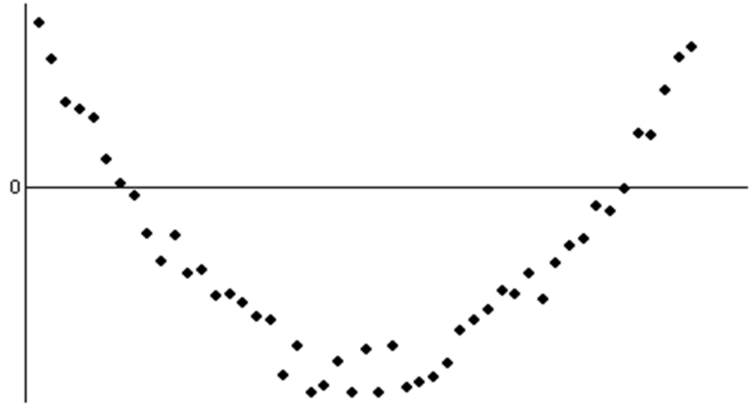
DISTRIBUIÇÃO DOS ERROS



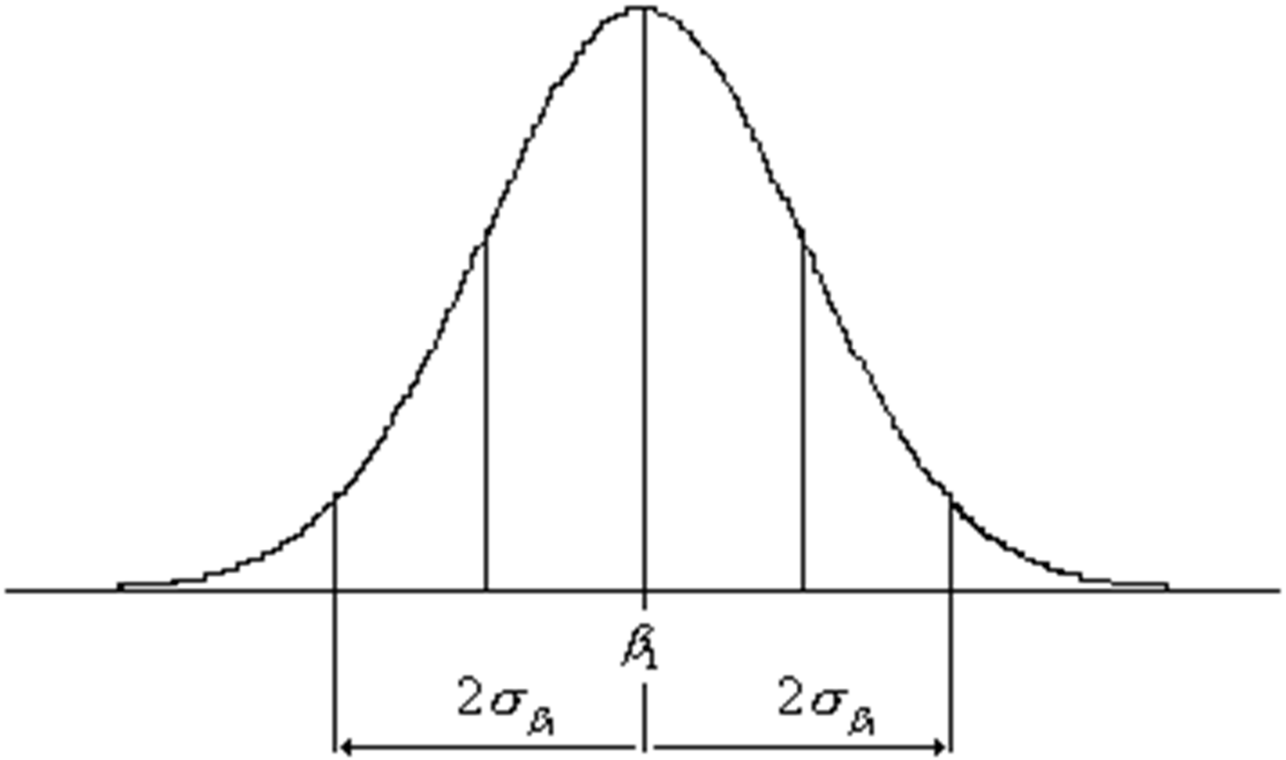
RESÍDUOS







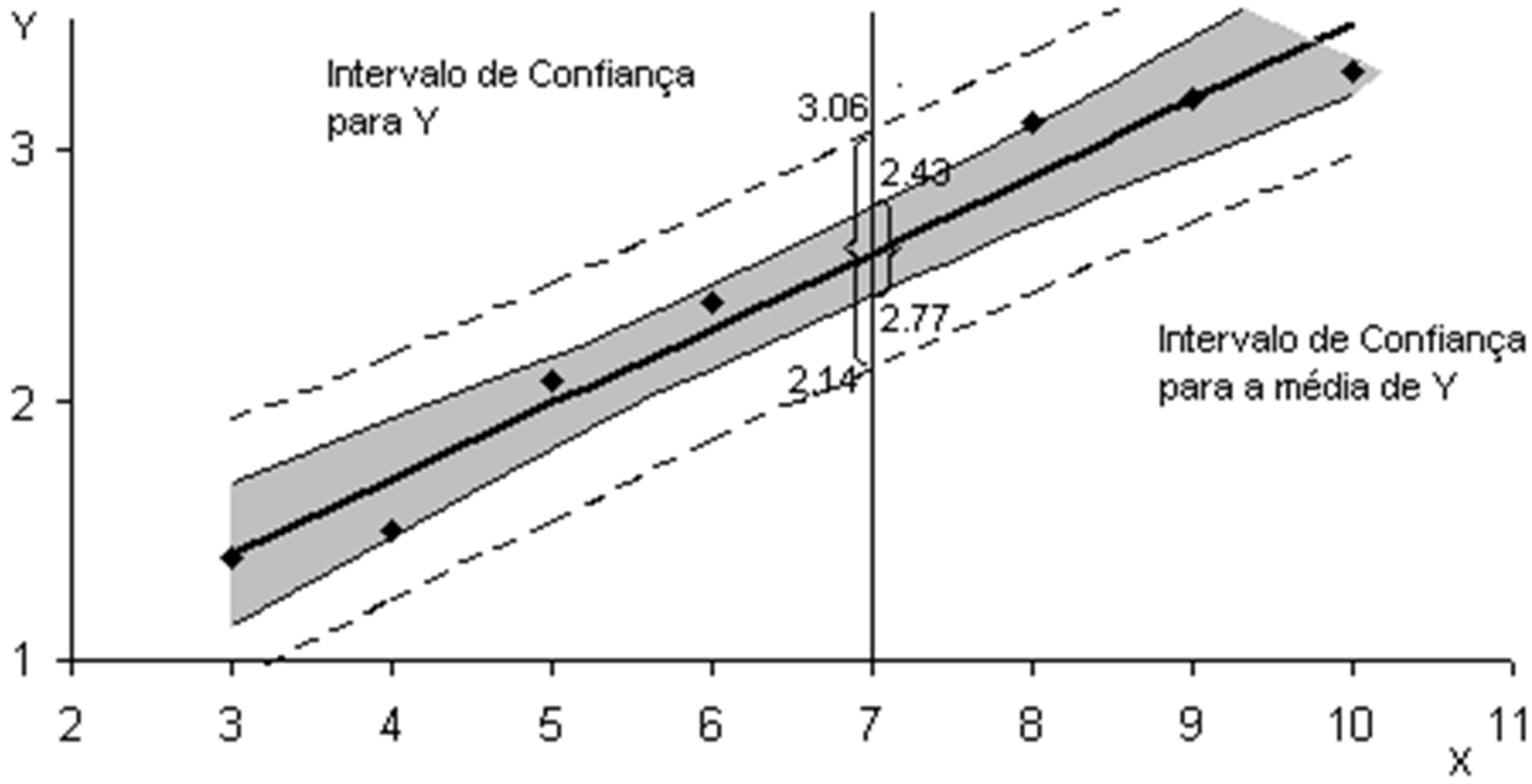
DISTRIBUIÇÃO DO DECLIVE



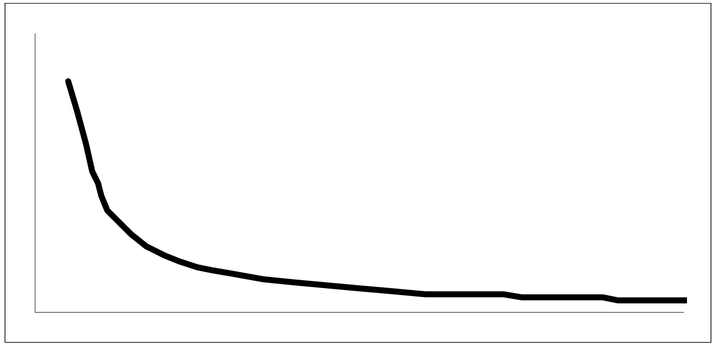
IC e Testes de hipóteses

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

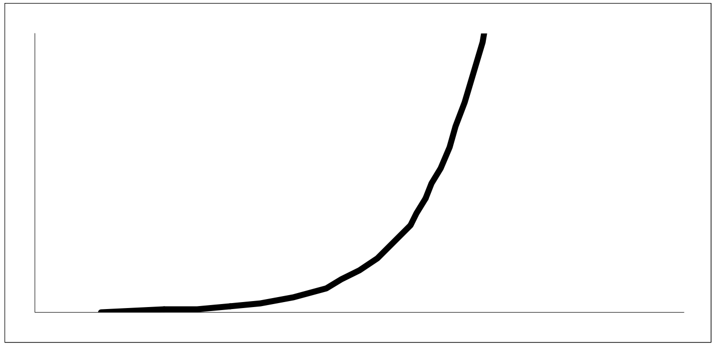
INTERVALO DE CONFIANÇA



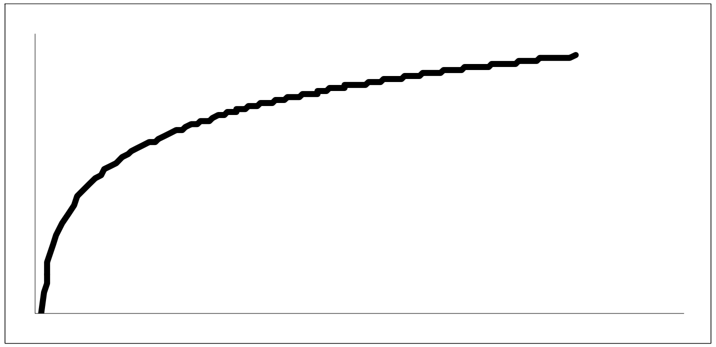
REGRESSÃO NÃO LINEAR



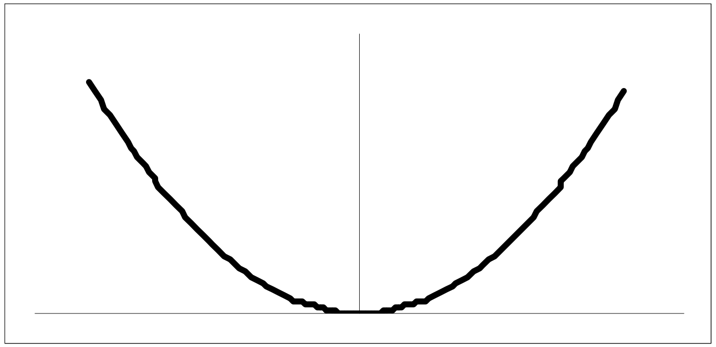
$\hat{Y} = \beta_0 + \beta_1 \frac{1}{X}$



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



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



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

REGRESSÃO NÃO LINEAR

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

COEFICIENTE DE CORRELAÇÃO

Coeficiente de correlação de Pearson

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

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

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

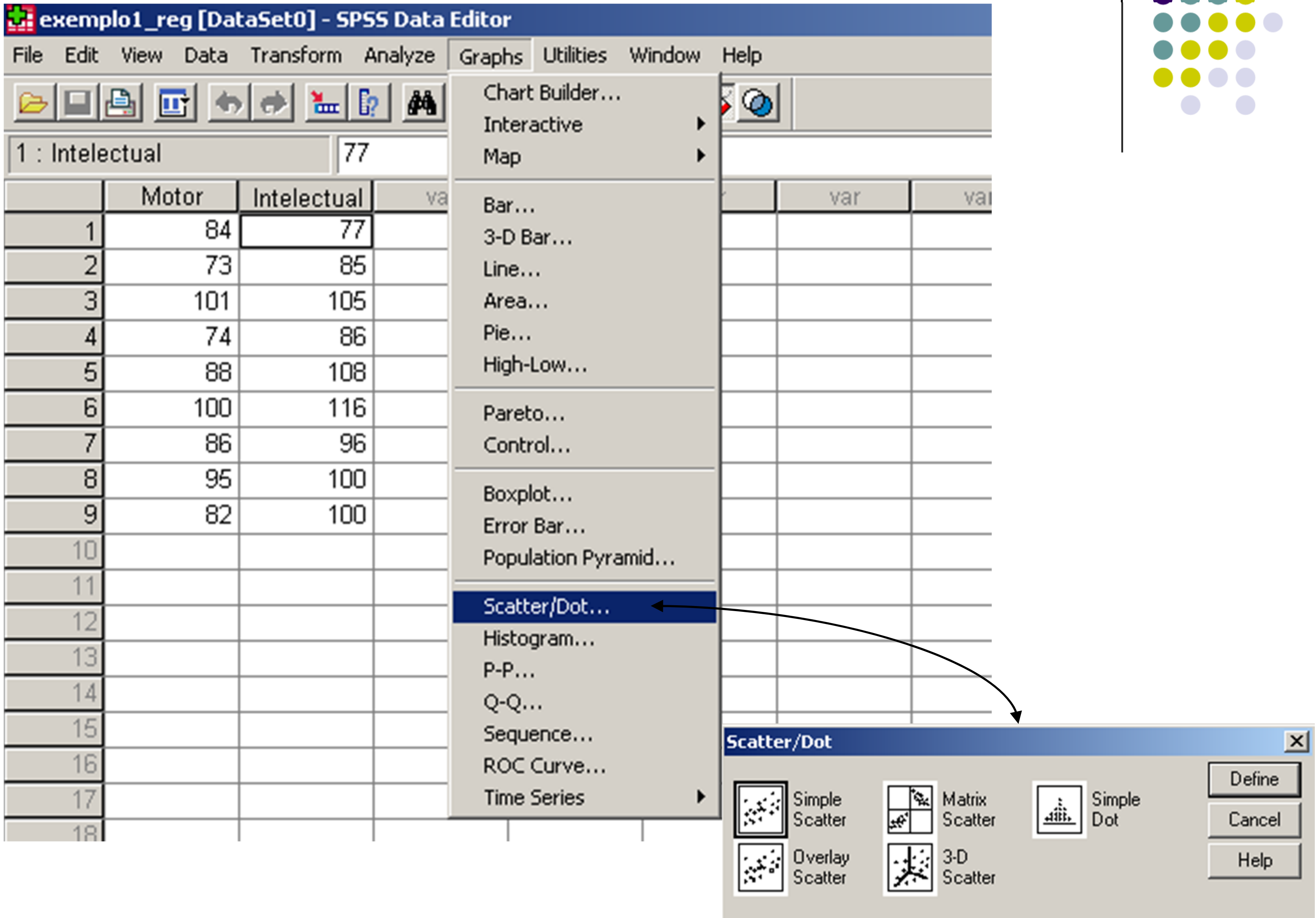
Região de Rejeição:

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

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


exemplo1_reg [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : Intelectual 77
Motor Intelectual var var var
1 84 77
2 73 85
3 101 105
4 74 86
5 88 108
6 100 116
7 86 96
8 95 100
9 82 100
Chart Builder...
Interactive
Map
Bar...
3-D Bar...
Line...
Area...
Pie...
High-Low...
Pareto...
Control...
Boxplot...
Error Bar...
Population Pyramid...
Scatter/Dot...
Histogram...
P-P...
Q-Q...
Sequence...
ROC Curve...
Time Series
Scatter/Dot
Simple Scatter
Matrix Scatter
Simple Dot
Overlay Scatter
3-D Scatter
Define
Cancel
Help

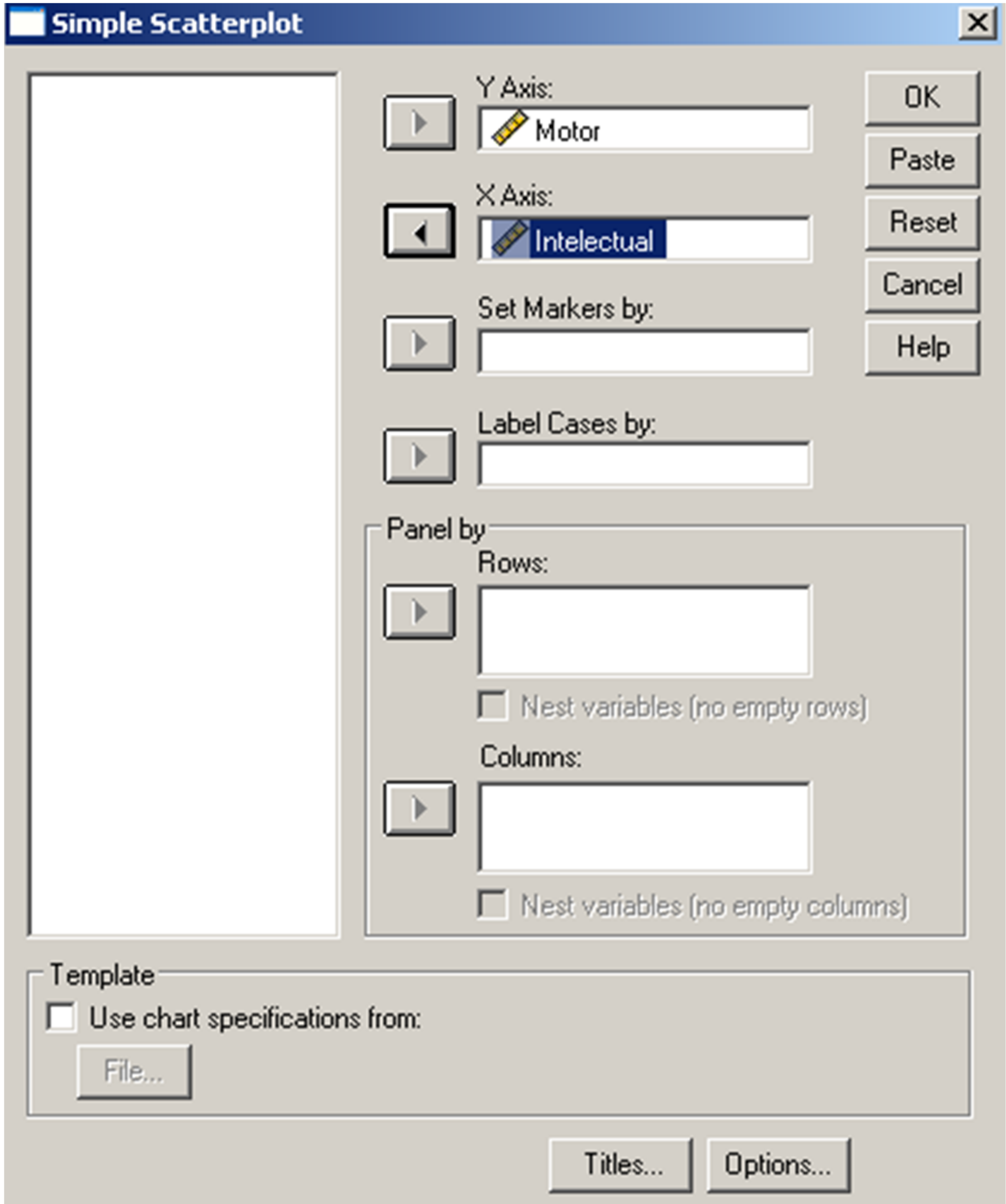
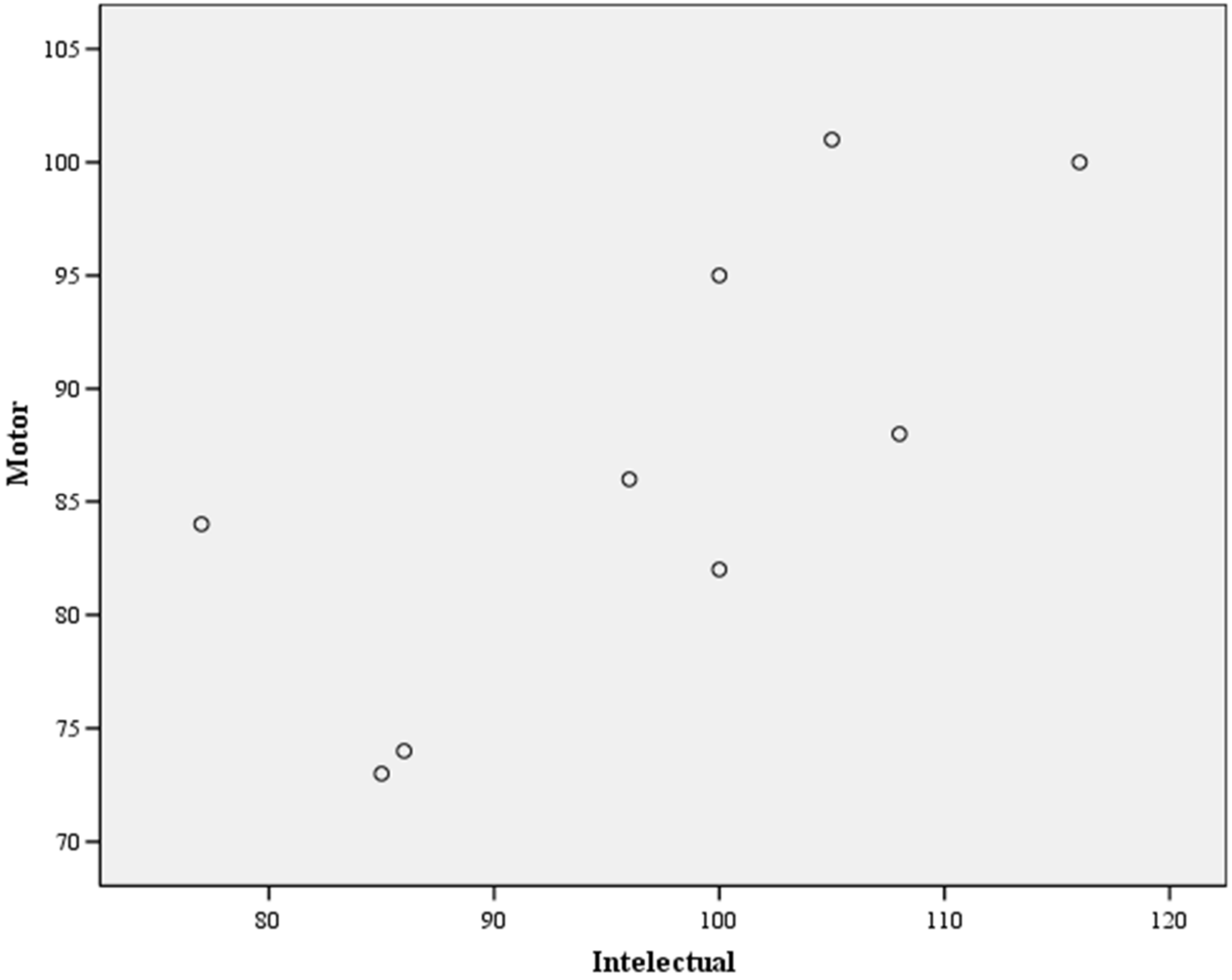
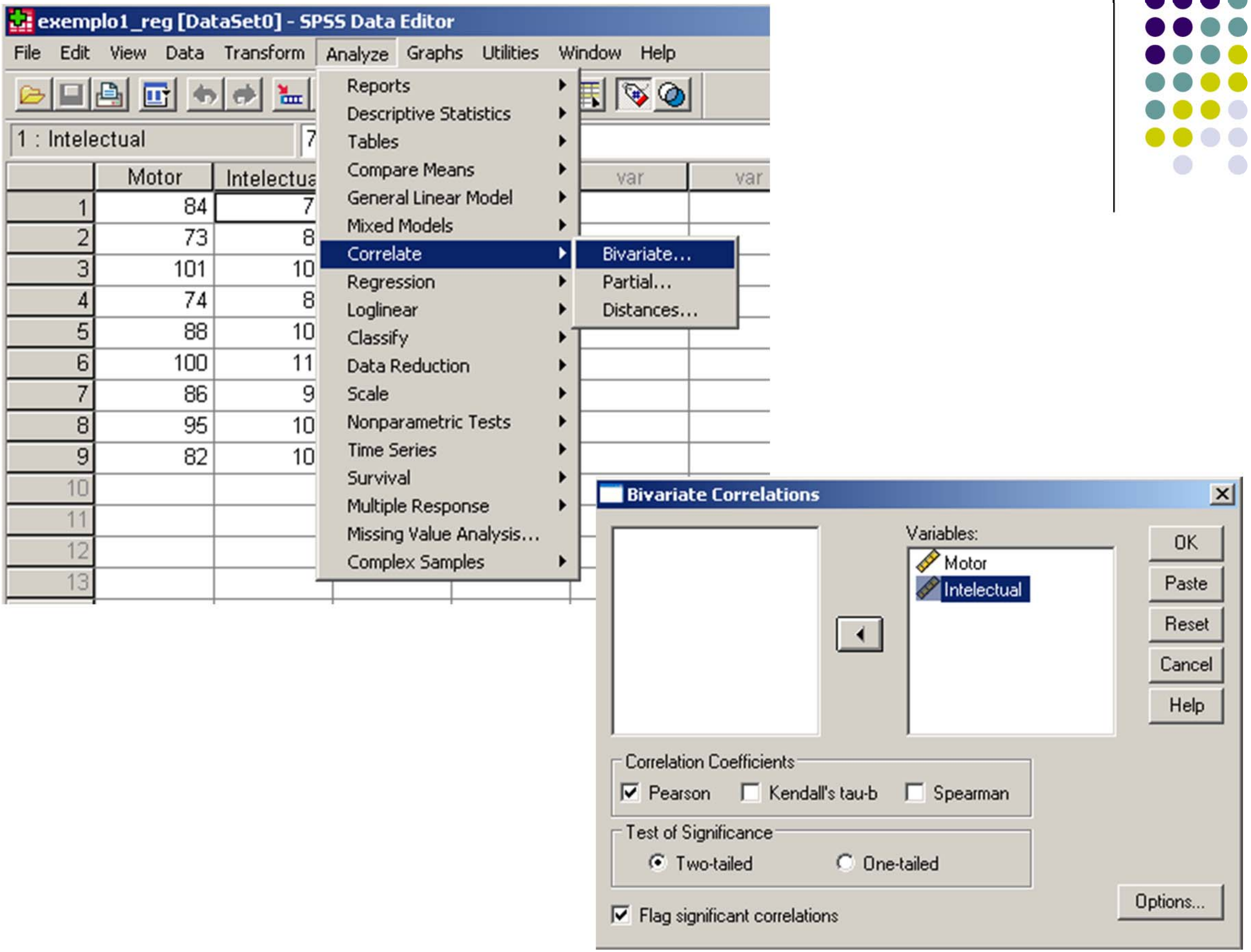
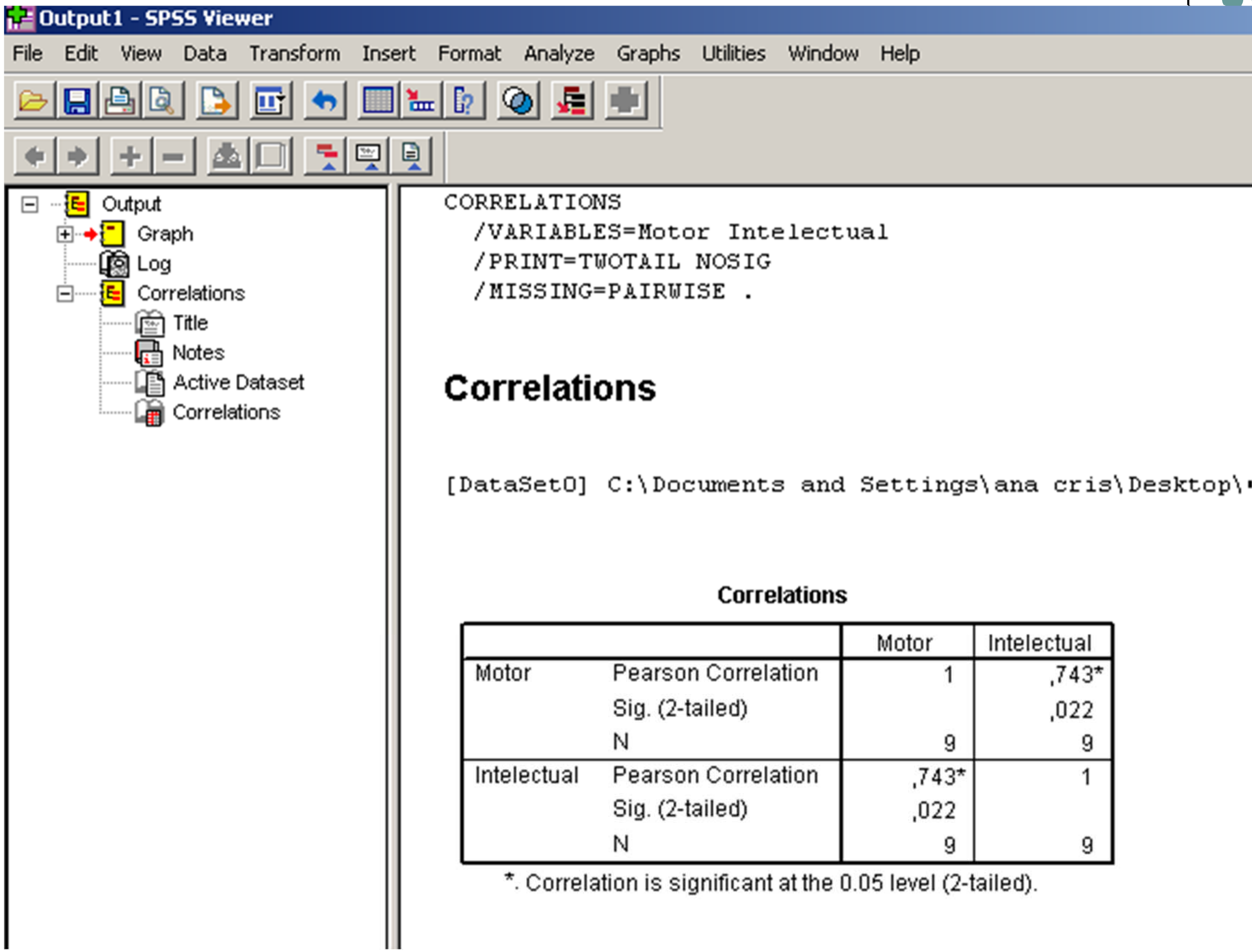

Simple Scatterplot
Y Axis:
Motor
X Axis:
Intelectual
Set Markers by:
Label Cases by:
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
Template
Use chart specifications from:
File...
OK
Paste
Reset
Cancel
Help
Titles...
Options...

DIAGRAMA DE DISPERSÃO




exemplo1_reg [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : Intelectual
Motor
Intelectua
var
var
1 84 7
2 73 8
3 101 10
4 74 8
5 88 10
6 100 11
7 86 9
8 95 10
9 82 10
10
11
12
13
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Bivariate...
Partial...
Distances...
Bivariate Correlations
Variables:
Motor
Intelectual
OK
Paste
Reset
Cancel
Help
Correlation Coefficients
Pearson
Kendall's tau-b
Spearman
Test of Significance
Two-tailed
One-tailed
Flag significant correlations
Options...




```
CORRELATIONS
  /VARIABLES=Motor Intelectual
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

Correlations

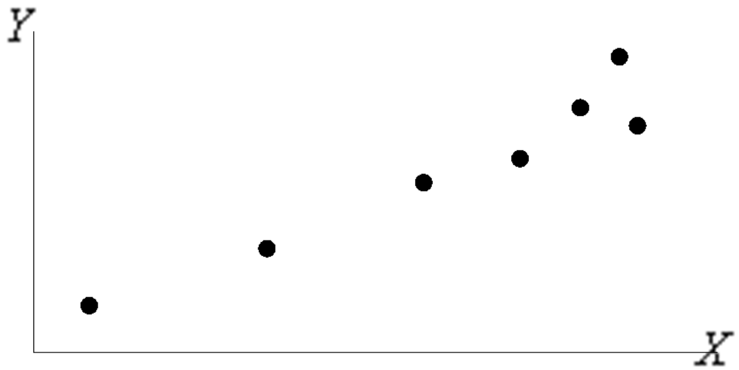
[DataSet0] C:\Documents and Settings\ana cris\Desktop\

Correlations

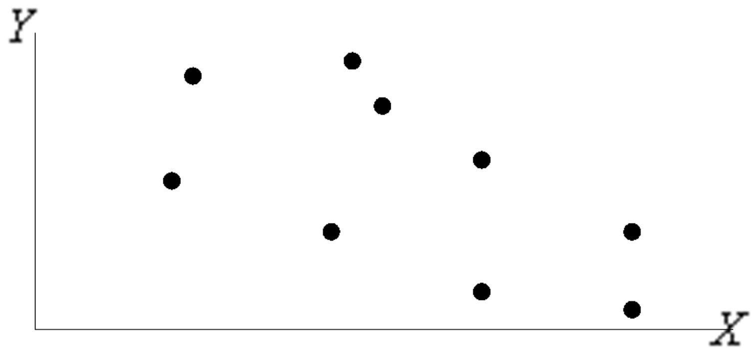
		Motor	Intelectual
Motor	Pearson Correlation	1	,743*
	Sig. (2-tailed)		,022
	N	9	9
Intelectual	Pearson Correlation	,743*	1
	Sig. (2-tailed)	,022	
	N	9	9

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

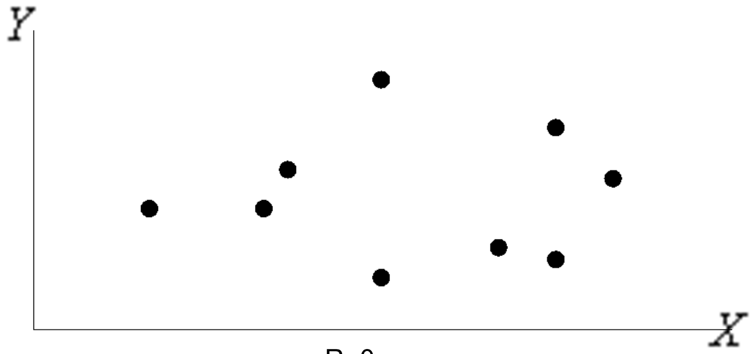
CORRELAÇÃO



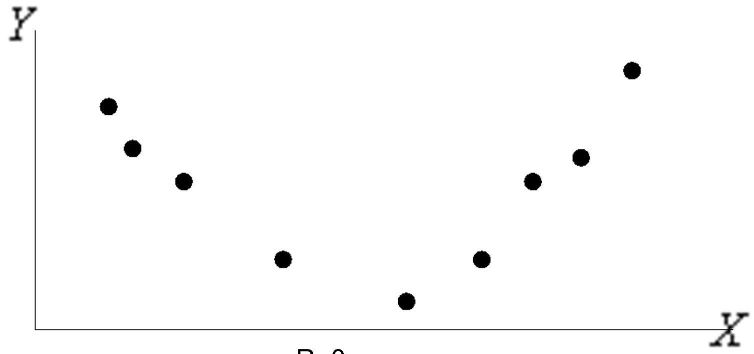
R=0.93



R=-0.62

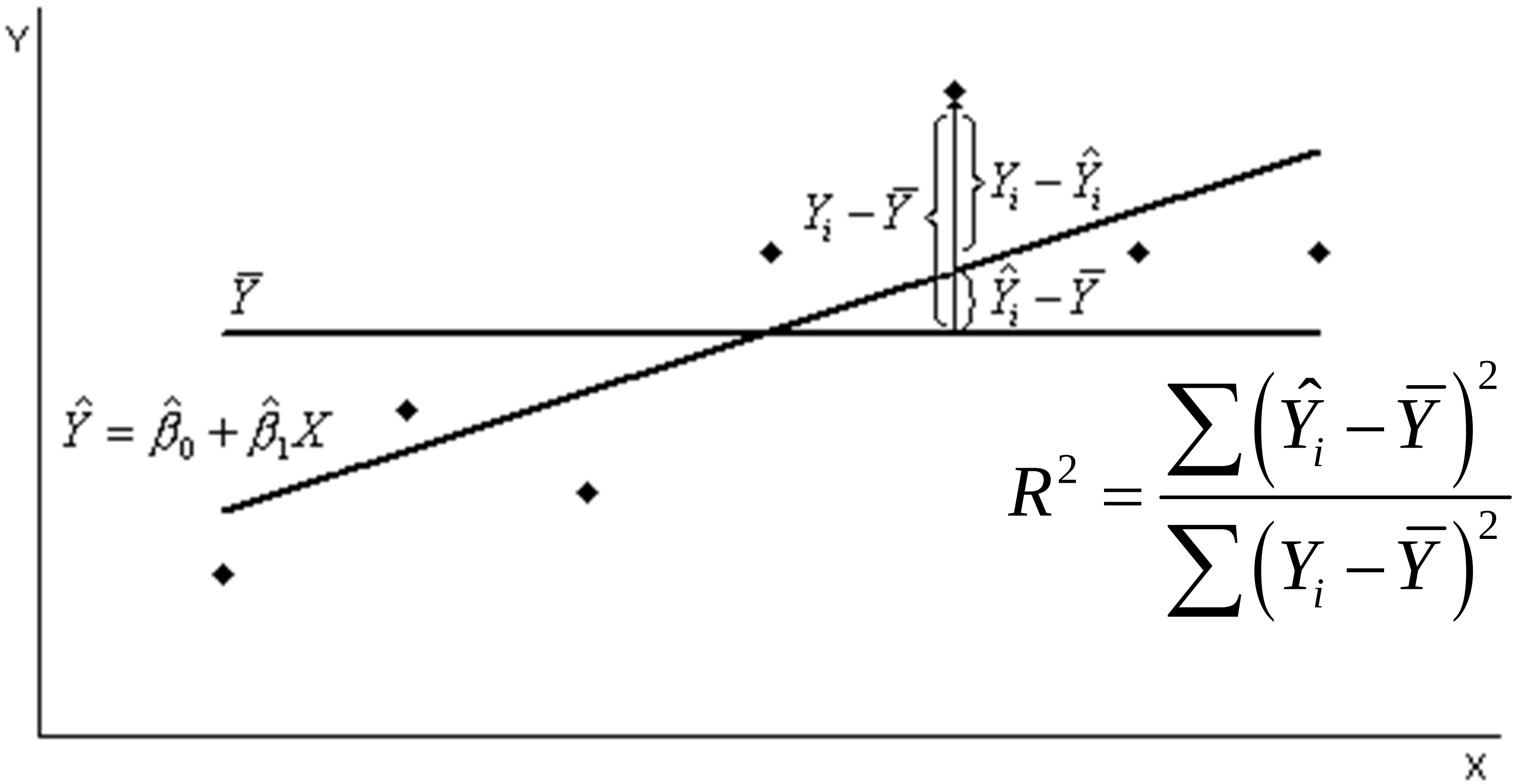


R=0



R=0

COEFICIENTE DE DETERMINAÇÃO



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

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

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

CORRELAÇÃO DE SPEARMAN

- O coeficiente ρ de Spearman mede a intensidade da relação entre variáveis ordinais. Usa, em vez do valor observado, apenas a ordem das observações.
- Aplica-se igualmente em variáveis intervalares/rácio como alternativa ao R de Pearson, quando neste último se viola a normalidade.
- Nos caso em que os dados não formam uma nuvem "bem comportada", com alguns pontos muito afastados dos restantes, ou em que parece existir uma relação crescente ou decrescente em formato de curva, o coeficiente ρ de Spearman é mais apropriado.

CORRELAÇÃO DE SPEARMAN

- Sem observações repetidas

$$R_S = \frac{\sum_{i=1}^{n}\left[R(X_i) - \frac{n+1}{2}\right]\left[R(Y_i) - \frac{n+1}{2}\right]}{\frac{n\left(n^2-1\right)}{12}}$$

ou

$$R_S = 1 - \frac{6T}{n\left(n^2-1\right)} \qquad T = \sum_{i=1}^{n}\left[R(X_i) - R(Y_i)\right]^2$$

- Com observações repetidas

$$R_S = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^{n} R(X_i)^2 - n\left(\frac{n+1}{2}\right)^2} \cdot \sqrt{\sum_{i=1}^{n} R(Y_i)^2 - n\left(\frac{n+1}{2}\right)^2}}$$

TESTES DE CORRELAÇÃO DE SPEARMAN

A. Teste bilateral

H_0: As variáveis X e Y são independentes.

H_1: (a) Existe uma tendência para os maiores valores de X formarem pares com os maiores valores de Y, ou

(b) Existe uma tendência para os menores valores de X formarem pares com os maiores valores de Y.

R.R: $R_S > c_1$ ou $R_S < c_2$, sendo c_1 o ponto crítico que corresponde a $1 - \frac{\alpha}{2}$ e c_2 o ponto crítico que corresponde a $\frac{\alpha}{2}$

B. Teste unilateral para correlação positiva

H_0: As variáveis X e Y são independentes.

H_1: Existe uma tendência para os maiores valores de X e de Y formarem pares.

R.R: $R_S > c$, em que c é o ponto crítico que corresponde a $1 - \alpha$

C. Teste unilateral para correlação negativa

H_0: As variáveis X e Y são independentes.

H_1: Existe uma tendência para os menores valores de X formarem pares com os maiores valores de Y e vice-versa.

R.R: $R_S < c$ sendo c o ponto crítico que corresponde a α.

Exemplo (Ficha 13 exerc 15)

O Sr. José foi seleccionado para júri num concurso de beleza. Concorreram oito jovens, e os resultados foram registados.
As idades são também apresentadas porque se suspeita que o Sr. José tem um favoritismo em relação às mais jovens. Verifique se existe ou não favoritismo.

		$R(X_i)$	lugar	$R(Y_i)$	$R(X_i)*R(Y_i)$	$R(X_i)^2$	$R(Y_i)^2$
Amélia	17 anos	2	1	1	2	4	1
Bela	16 anos	1	2	2,5	2,5	1	6,25
Carolina	18 anos	4	2	2,5	10	16	6,25
Deolinda	20 anos	6,5	4	4	26	42,25	16
Eva	18 anos	4	5	5	20	16	25
Francisca	18 anos	4	6	6	24	16	36
Georgina	20 anos	6,5	7	7,5	48,75	42,25	56,25
Helena	23 anos	8	7	7,5	60	64	56,25
Somas		**36**		**36**	**193,25**	**201,5**	**203**

$$R_S = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^{n} R(X_i)^2 - n\left(\frac{n+1}{2}\right)^2} \cdot \sqrt{\sum_{i=1}^{n} R(Y_i)^2 - n\left(\frac{n+1}{2}\right)^2}} = \frac{193{,}25 - 8 * 4{,}5^2}{\sqrt{201{,}5 - 8 * 4{,}5^2} * \sqrt{203 - 8 * 4{,}5^2}} = 0{,}7765$$