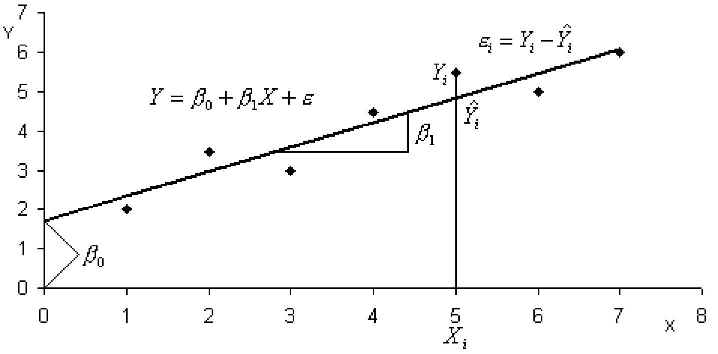
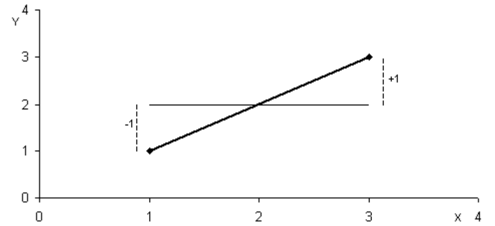
REGRESSÃO E CORRELAÇÃO

AJUSTE DE UMA RETA

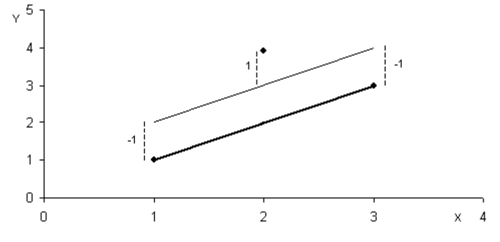


MINIMIZAÇÃO DOS DESVIOS



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

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

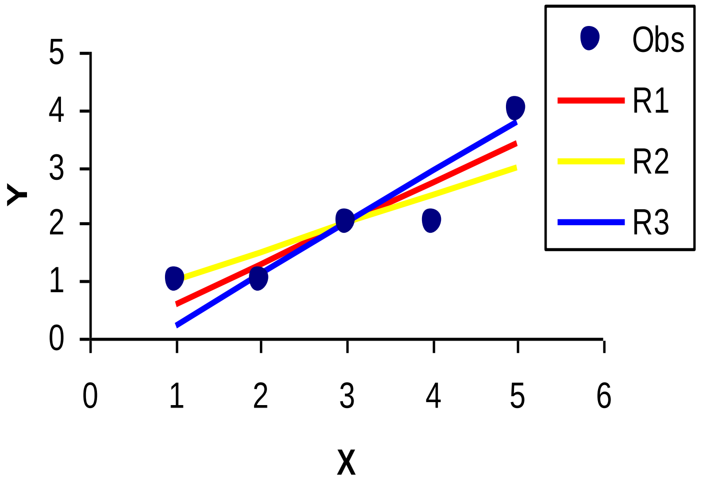
Desv1	Desv2	Desv3
0.4	0	0.8
-0.3	-0.5	-0.1
0	0	0
-0.7	-0.5	-0.9
0.6	1	0.2
0	0	0

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

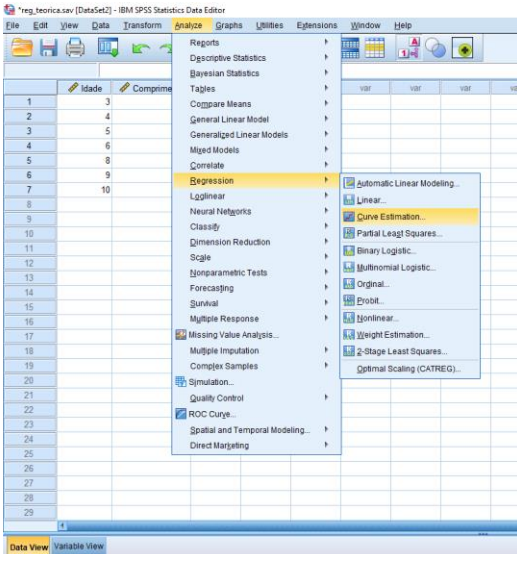


EXEMPLO 2

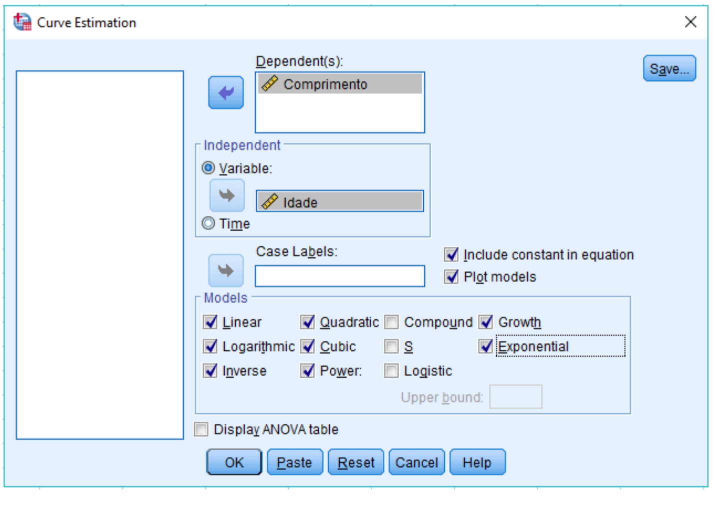
Comprimento alar (cm) em função da idade (dias) para andorinhas

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	Independent
		Comprimento	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

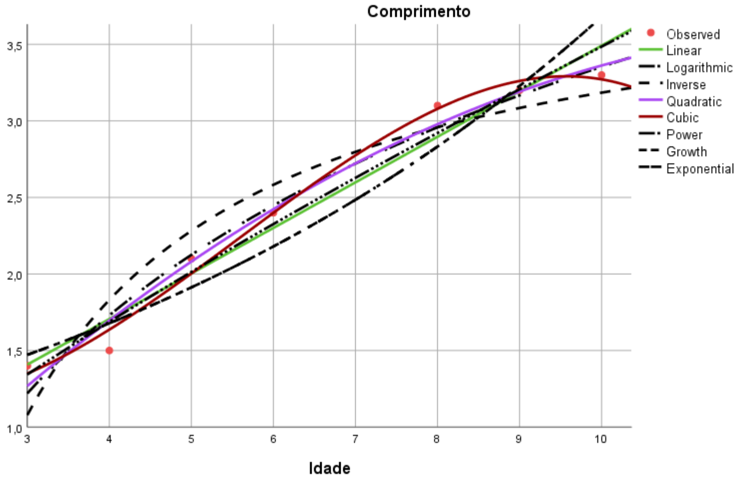
Model Summary and Parameter Estimates

Dependent Variable: Comprimento

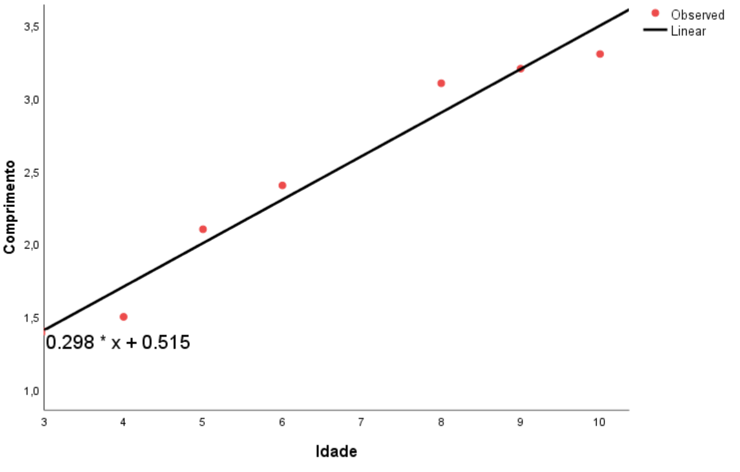
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	$\boxed{0,964}$	132,174	1	5	0,000	0,515	0,298		
Logarithmic	0,971	165,753	1	5	0,000	-0,727	1,772		
Inverse	0,915	53,833	1	5	0,001	4,087	-9,026		
Quadratic	$\boxed{0,980}$	99,685	2	4	0,000	-0,274	0,579	-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.

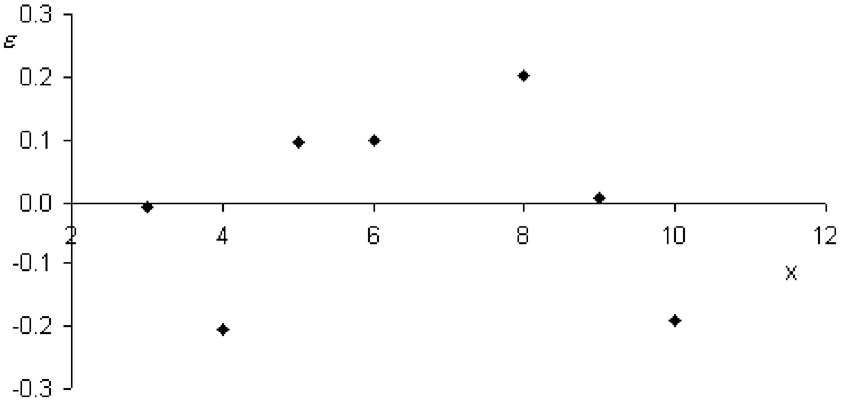
EXEMPLO 2



RECTA DE MÍNIMOS QUADRADOS



RESÍDUOS



Estimadores

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

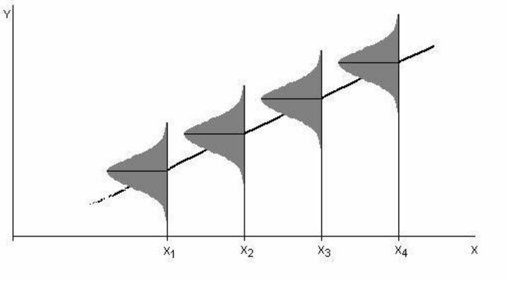
β_0 $$\boxed{\hat{\beta}_0 = \frac{1}{n}\sum_i Y_i = \overline{Y}}$$

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

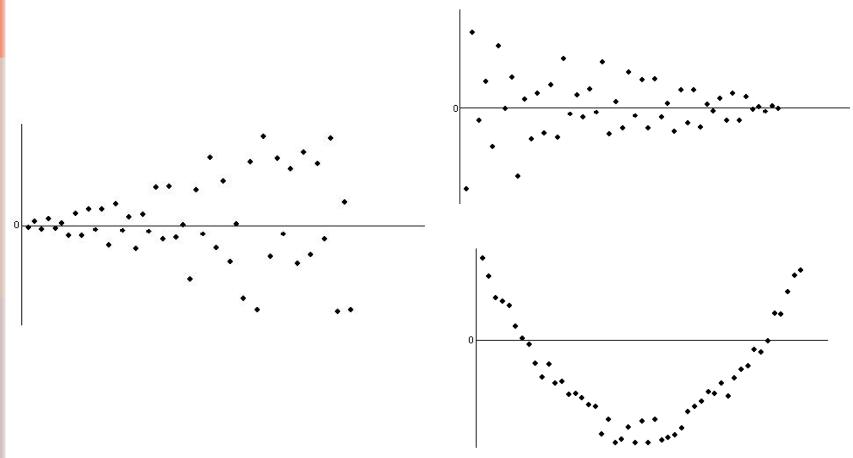
21

DISTRIBUIÇÃO DOS ERROS

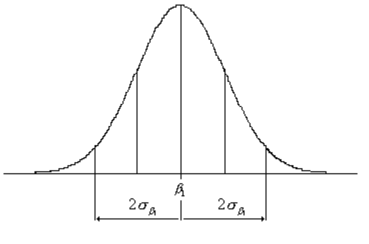


22

RESÍDUOS



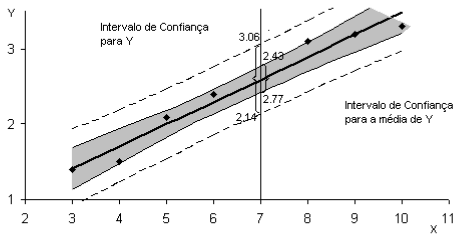
DISTRIBUIÇÃO DO DECLIVE



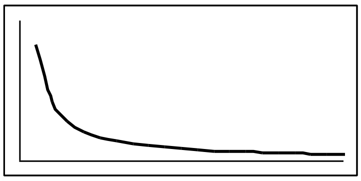
IC e Testes de hipóteses

	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 $\beta_1 < b_{10}$ $ET = \frac{\hat{\beta}_1 - b_{10}}{\frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}}$ H_0 verdadeira $\Rightarrow ET \sim t_{n-2}$

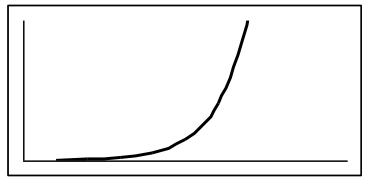
INTERVALO DE CONFIANÇA



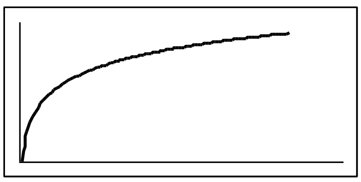
REGRESSÃO NÃO LINEAR



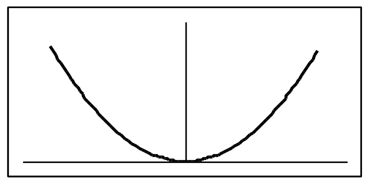
$\hat{Y} = \beta_0 + \beta_1 \frac{1}{X}$



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



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



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

REGRESSÃO NÃO LINEAR

Modelo	Transformação
• $Y_i = \alpha' + \frac{\beta}{X_i} + e_i$	$U_i = \frac{1}{X_i}$ $Y_i = \alpha' + \beta . U_i + e_i$
• $Y_i = e^{\alpha' + \beta . X_i + e_i}$	$Z_i = \ln Y_i$ $Z_i = \alpha' + \beta . X_i + e_i$
• $Y_i = e^{\alpha' + \frac{\beta}{X_i} + e_i}$ 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:

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

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

29

RESÍDUOS

Pressupostos para ε_i

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

$$\left.\right\} \varepsilon_i \sim IN(0, \sigma^2)$$

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

30

ESTIMADORES MÍNIMOS QUADRADOS

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

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

$$S_{X_{j_1}X_{j_2}} = \sum_n \left(X_{j_1 i} - \bar{X}_{j_1}\right)\left(X_{j_2 i} - \bar{X}_{j_2}\right)$$

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

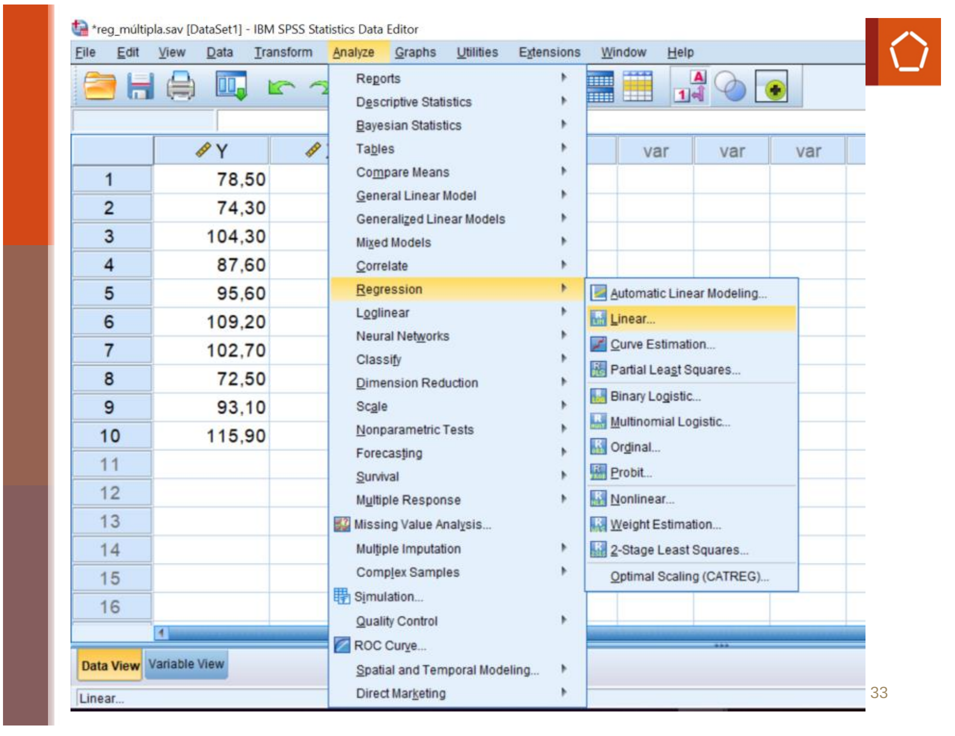
$$s^2 = \frac{1}{n-J-1}\sum_i \hat{\varepsilon}_i^2 =$$

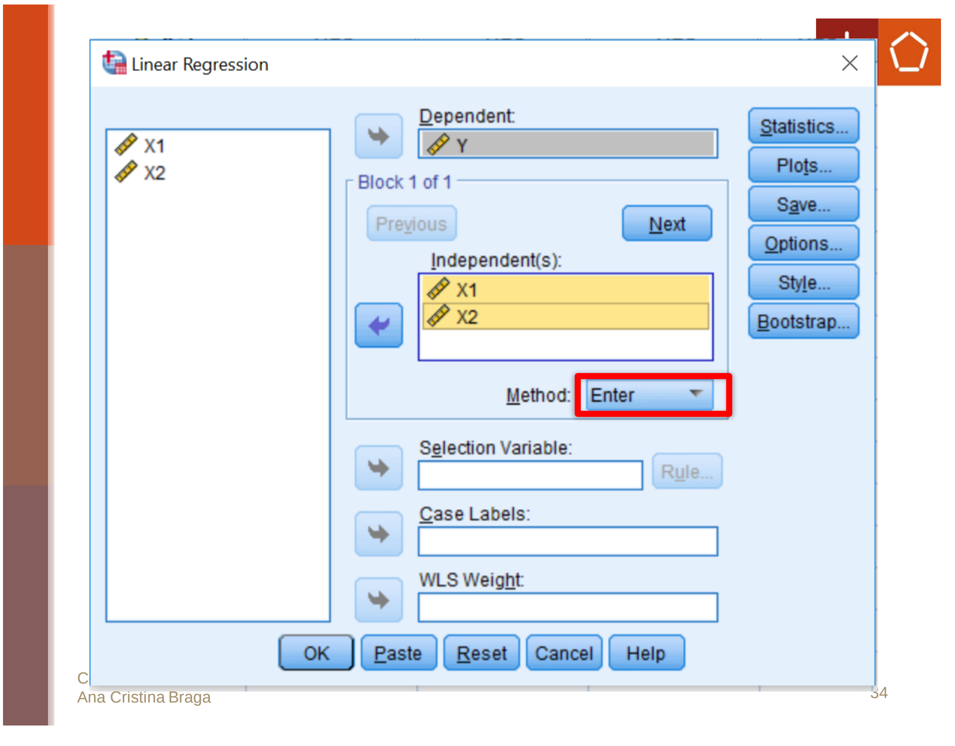
$$= \frac{1}{n-J-1}\sum_i \left\{Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 . \left(X_{1i} - \bar{X}_1\right) + \hat{\beta}_2 . \left(X_{2i} - \bar{X}_2\right) + \ldots + \hat{\beta}_j . \left(X_{Ji} - \bar{X}_J\right)\right]\right\}^2$$

EXEMPLO 3

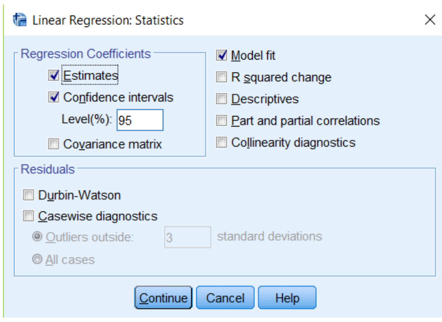
Determine a relação existente entre o calor envolvido no endurecimento, representado pela variável Y e os pesos de duas substâncias X_1 e X_2, tendo em consideração os seguintes valores obtidos numa experiência:

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


*reg_múltipla.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Y
78,50
74,30
104,30
87,60
95,60
109,20
102,70
72,50
93,10
115,90
Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Spatial and Temporal Modeling...
Direct Marketing
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling (CATREG)...
var var var
Data View Variable View
Linear...
33


Linear Regression
X1
X2
Dependent:
Y
Block 1 of 1
Previous
Next
Independent(s):
X1
X2
Method: Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...
OK Paste Reset Cancel Help
Ana Cristina Braga
34

Regression>Linear>Statistics



Regression>Linear>Save

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	X2, X1[b]	.	Enter

a. Dependent Variable: Y

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,988[a]	,977	,970	2,57617

a. Predictors: (Constant), X2, X1

b. Dependent Variable: Y

ANOVA (Modelo)

H_0: O modelo de regressão considerado não serve

Decisão: Como valor p <0,05, rejeita-se a H_0, pelo que o modelo de regressão considerado é estatisticamente significativo

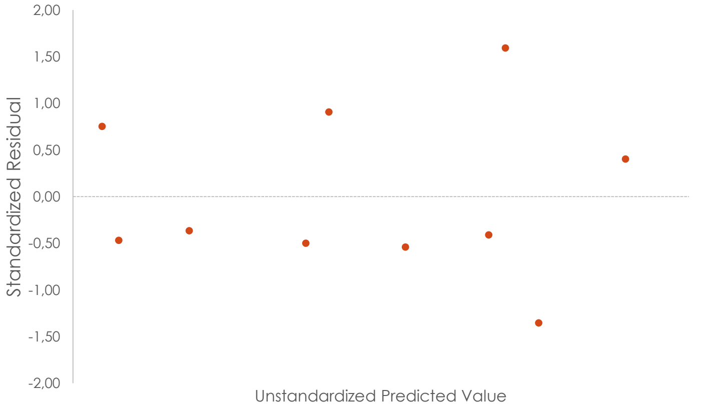
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1976,924	2	988,462	148,940	,000[b]
	Residual	46,457	7	6,637		
	Total	2023,381	9			

a. Dependent Variable: Y

b. Predictors: (Constant), X2, X1

RESÍDUOS (homoscedasticidade)



Resíduos (Normalidade)

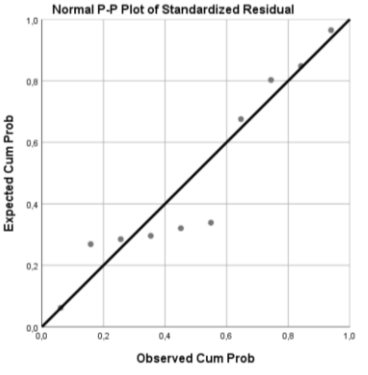
- Teste analítico (KS com correção de Lilliefors)
- Método gráfico (P-P ou Q-Q plot)

NPar Tests

One-Sample Kolmogorov-Smirnov Test

		Standardized Residual
N		10
Normal Parameters[a,b]	Mean	,0000000
	Std. Deviation	,88191710
Most Extreme Differences	Absolute	,261
	Positive	,261
	Negative	-,169
Test Statistic		,261
Asymp. Sig. (2-tailed)		,051[c]

a. Test distribution is Normal.
b. Calculated from data.
c. Lilliefors Significance Correction.



Resíduos (média zero)

T-Test

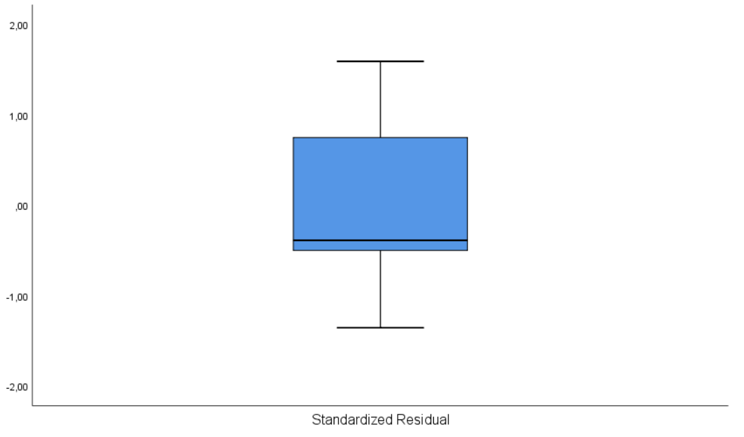
One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Standardized Residual	10	,0000000	,88191710	,27888668

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Standardized Residual	1,19E-014	9	1,000	3,32E-015	-,631	,631

Verificação de outliers



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

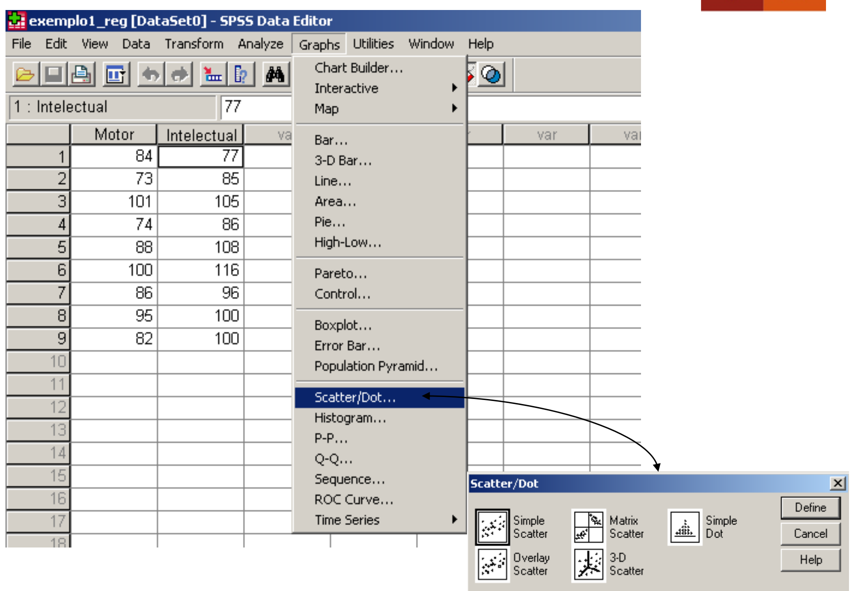
Região de Rejeição:

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

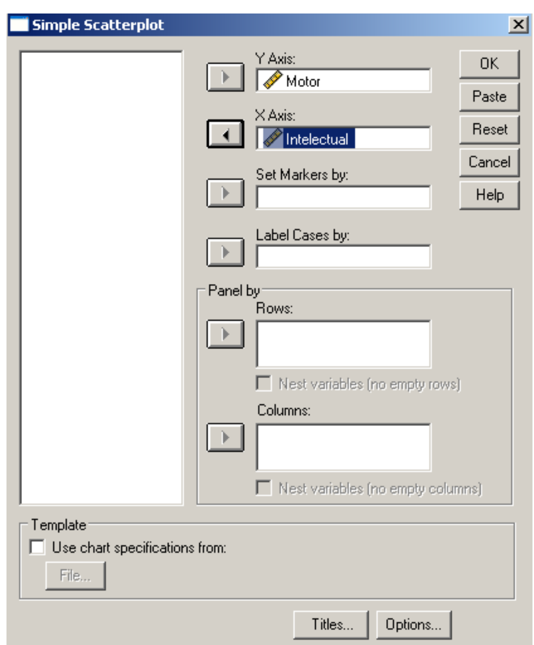
EXEMPLO

- 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

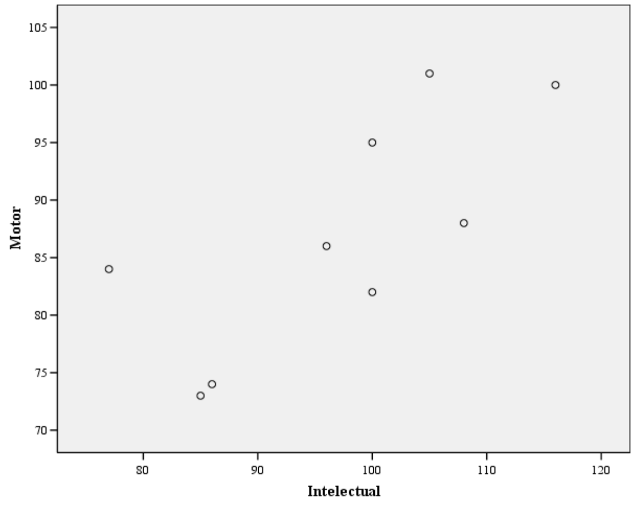

exemplo1_reg [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : Intelectual 77
Motor Intelectual var var
1 84 77
2 73 85
3 101 105
4 74 86
5 88 108
6 100 116
7 86 96
8 95 100
9 82 100
Chart Builder...
Interactive
Map
Bar...
3-D Bar...
Line...
Area...
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

45

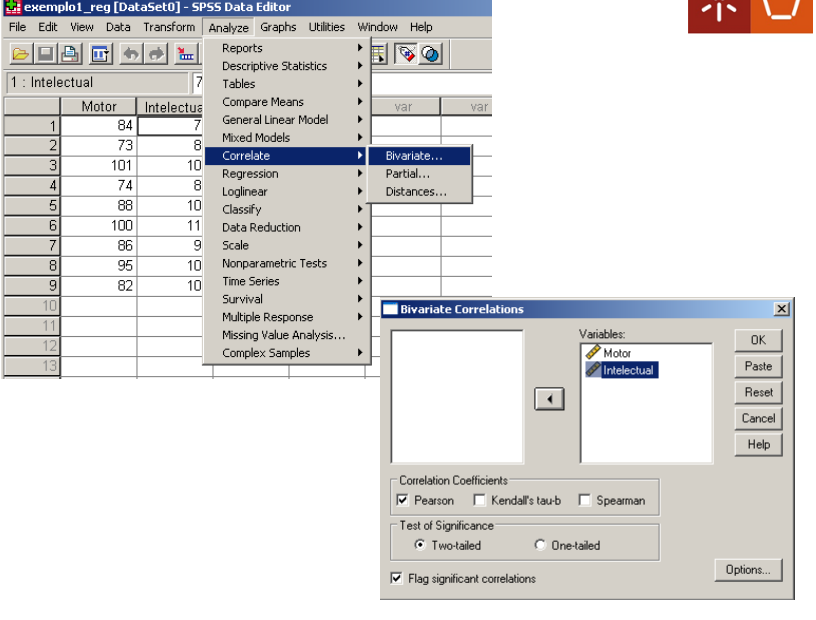

Simple Scatterplot
Y Axis:
Motor
X Axis:
Intelectual
Set Markers by:
Label Cases by:
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
OK
Paste
Reset
Cancel
Help
Template
Use chart specifications from:
File...
Titles...
Options...

46

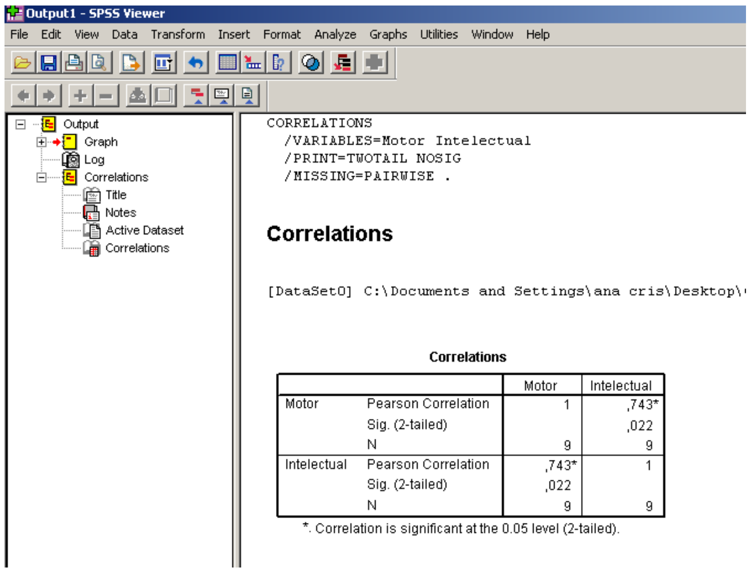
DIAGRAMA DE DISPERSÃO



47



48




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

Correlations

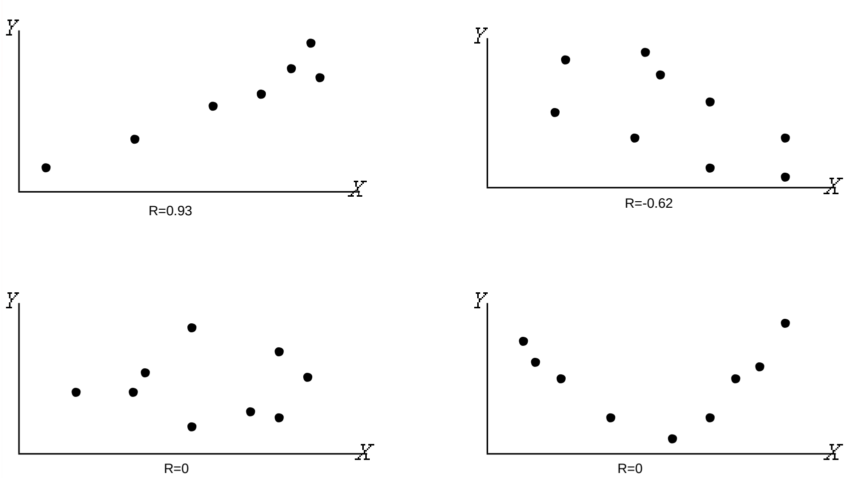
[DataSet0] C:\Documents and Settings\ana cris\Desktop\

Correlations

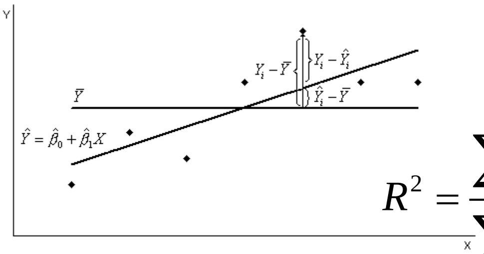
		Motor	Intelectual
Motor	Pearson Correlation	1	,743*
	Sig. (2-tailed)		,022
	N	9	9
Intelectual	Pearson Correlation	,743*	1
	Sig. (2-tailed)	,022	
	N	9	9

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

CORRELAÇÃO



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

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