BIOESTATÍSTICA

Análise de correlação e regressão

O que é analisado?

- Ao optar por este tipo de análise (correlação ou regressão), o pesquisador tem em mente alguns objetivos.
- A análise de correlação é utilizada para realizar análises **exploratórias** e/ou **descritivas**, enquanto que com a análise de regressão são realizadas as análises **explicativas** e **preditivas**.
- É possível fazer as análises exploratórias e descritivas também através da análise de regressão, mas é um processo demorado e muitas vezes o recomendado é a análise de correlação. Quando se tem uma grande quantidade de variáveis deve-se iniciar pela análise de correlação.

Análise de correlação

- Existe uma associação estatística entre duas variáveis? As duas variáveis são independentes (ou seja, qual o grau da variação das duas juntas)?

Coeficiente de Correlação

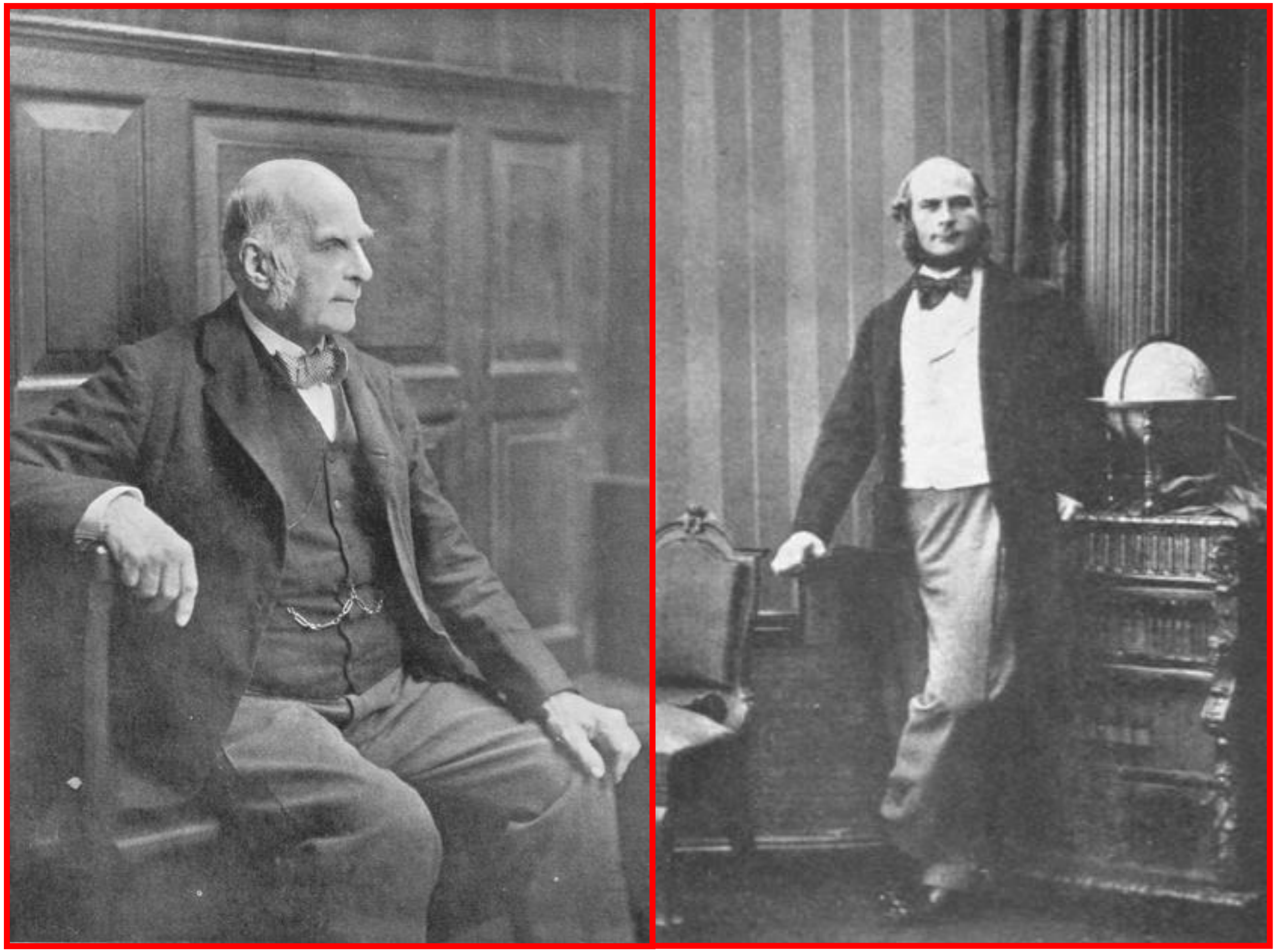
- A relação entre duas ou mais variáveis pode ser explorada através da análise de correlação.
- A análise de correlação fornece o coeficiente de correlação de **Pearson** (r) ou o coeficiente de correlação de postos ou ordens de **Spearman** (dados ordenados) e a sua significância pelo teste t de Student com n-2 graus de liberdade.
- H_0: r = 0; H_A: r ≠ 0.
- Se o coeficiente de correlação for muito baixo (inferior a ±0,20) e se o teste de hipótese não for significativo (teste t de Student) dificilmente se deve ir avante com as demais análises descritivas, explicativas e preditivas.

Francis Galton

- Foi Galton quem inventou a análise de correlação e de regressão (conceito).
- Geógrafo, meteorologista, inventor da identificação pela impressão digital, eugenista.
- Sobrinho de Charles Darwin, ajudou o tio na análise e interpretação da teoria evolucionista.

Biografia de Galton (1822-1911)

Inglês de uma família muito rica, após a morte do pai herdou uma fortuna com a qual realizou diversos experimentos e publicou diversos trabalhos.



Francis Galton

Karl Pearson (1857-1936)

Fez uma grande contribuição para o desenvolvimento da Estatística como uma disciplina científica séria e independente. Ele foi o fundador do Departamento de Estatística Aplicada (Department of Applied Statistics) na University College Londres em 1911. Foi Pearson quem formalizou o método de Galton, em 1896.

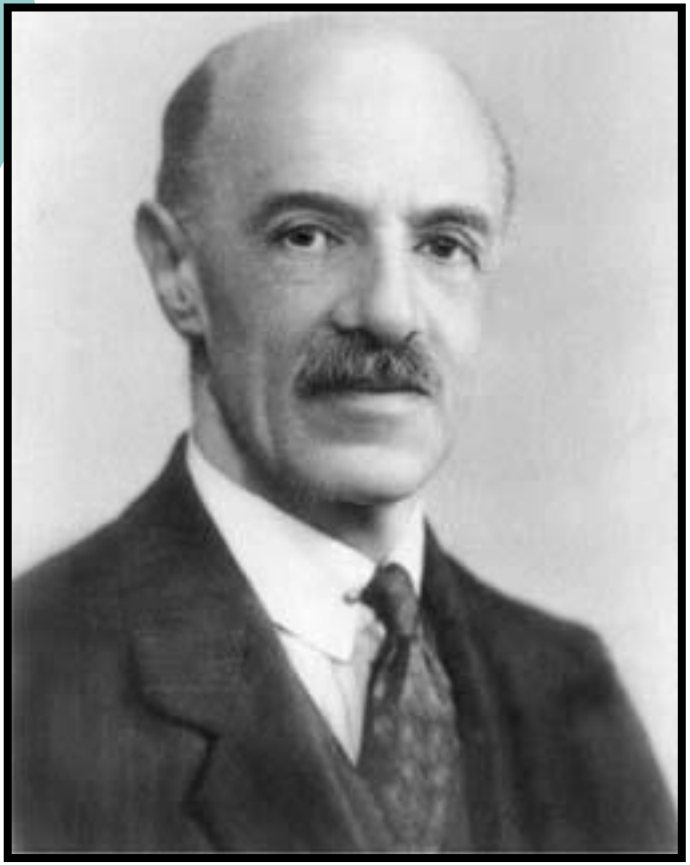
O coeficiente de correlação produto-momento de Pearson foi a primeira medida de força de associação a ser introduzido em estatística. (Wikipedia)

Pearson com Galton, em 1900

Charles Edward Spearman (1863-1945)

- Desenvolveu em 1904 o coeficiente de correlação de Spearman (baseado em postos ou ordens).
- Foi um psicólogo inglês conhecido pelo seu trabalho na área da estatística, como um pioneiro da análise fatorial e pelo coeficiente de correlação de postos de Spearman. Também foi influenciado pelas ideias de Galton.
- Ao encontrar dificuldades em calcular o Coeficiente de Correlação de Pearson, propôs transformar os dados em ordens ou postos e depois calcular o valor do coeficiente.

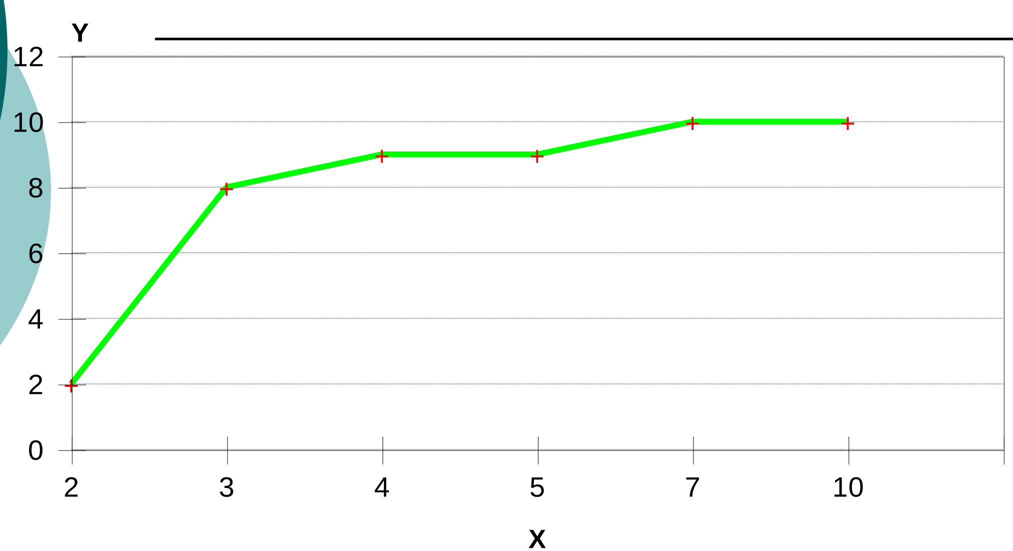
Coeficiente de Correlação de Spearman



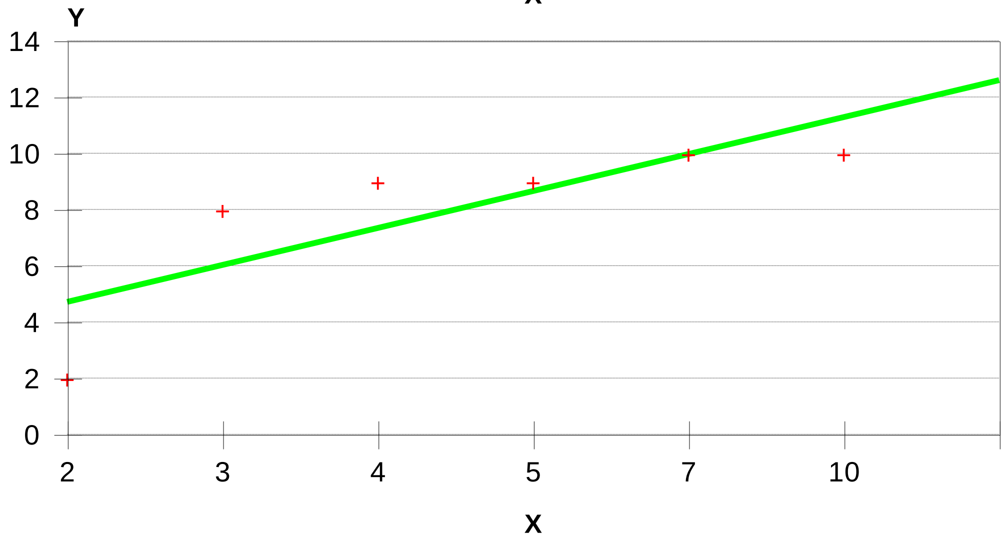
O coeficiente de correlação de Pearson detecta se existe uma correlação linear entre duas variáveis. As variáveis devem ser contínuas.

O coeficiente de correlação de Spearman detecta se existe correlação entre duas variáveis (esta relação pode ser não linear). As variáveis podem ser discretas ou contínuas.

Exemplo



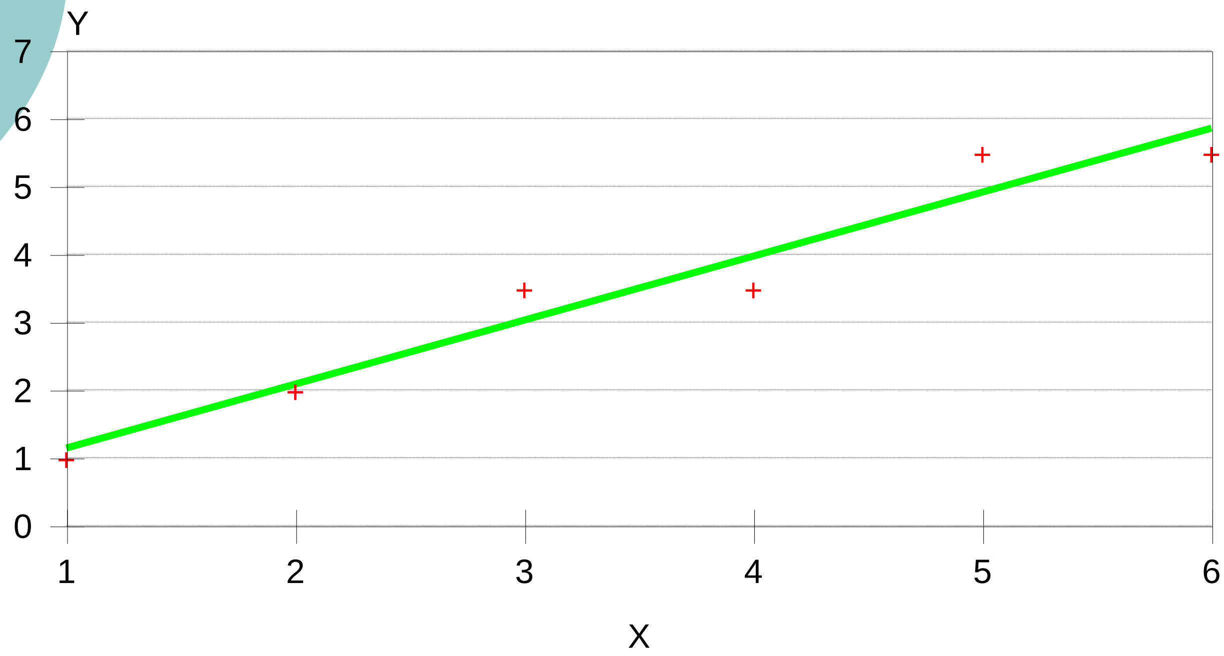
Não é uma reta (não linear)



Pearson: r = 0,69838

P > 0,1227 n.s.

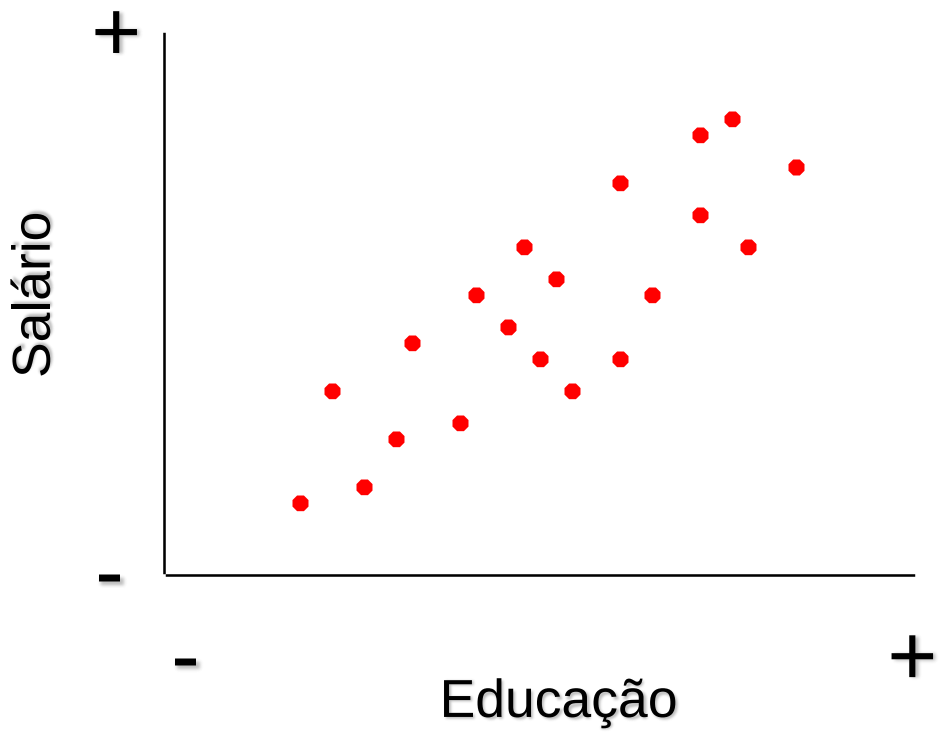
Spearman



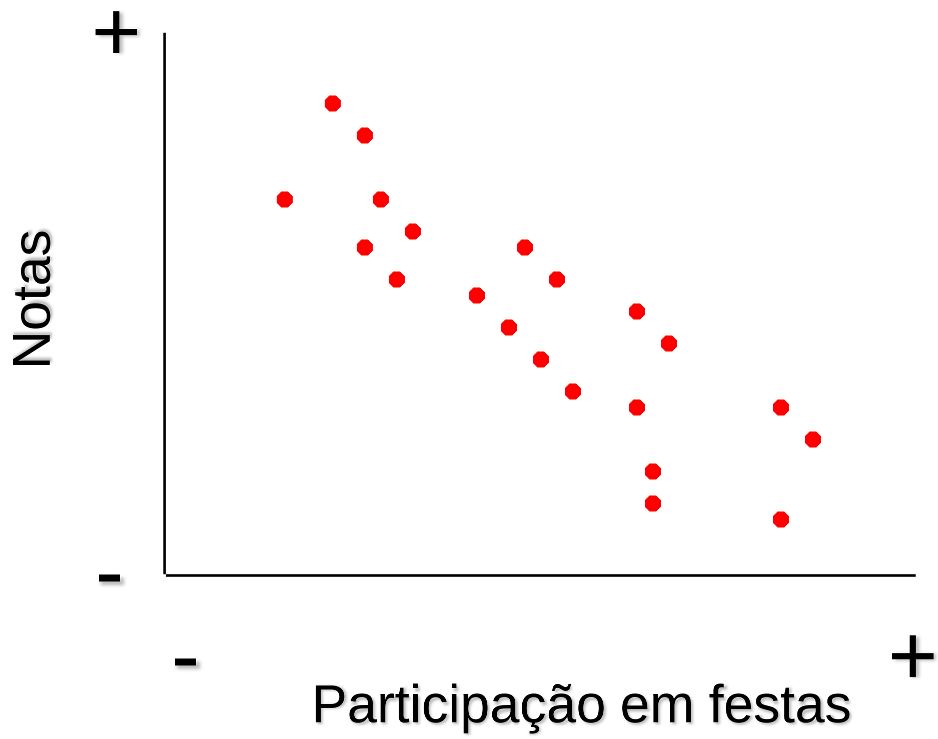
r = 0,97101

P > 0,0012 **

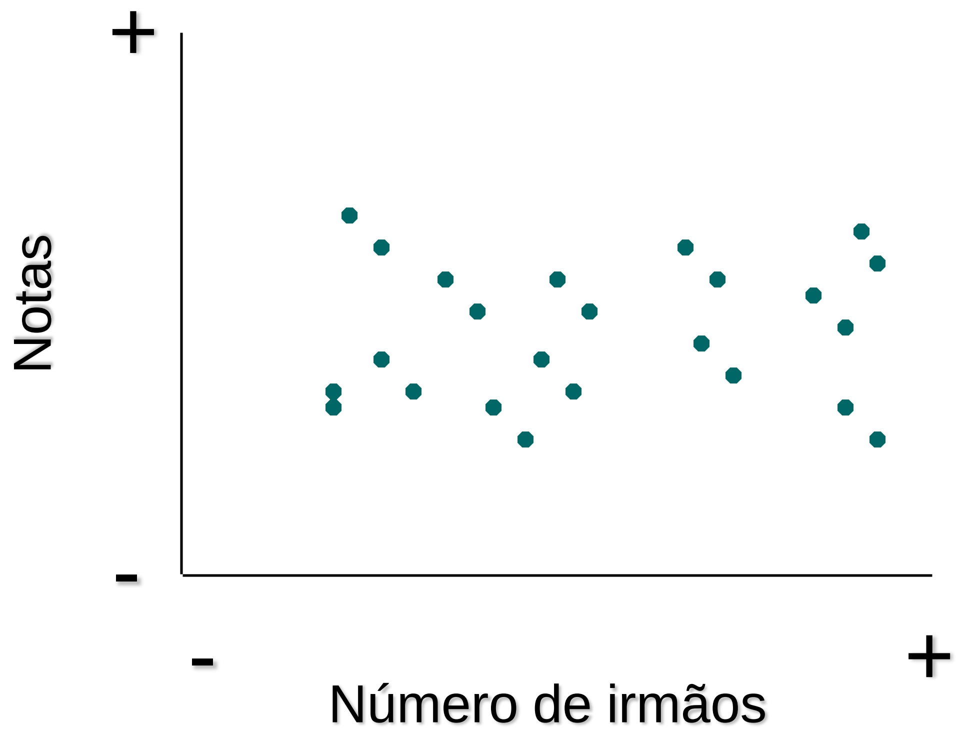
Correlação positiva: Educação e salário.



Correlação negativa: festa com notas dos alunos.



Sem correlação: notas dos alunos e número de irmãos.



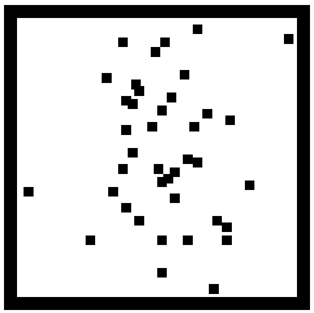
Teste de correlação.

- O coeficiente de correlação de Pearson, r, é calculado pela fórmula:

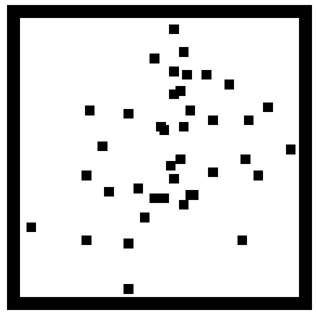
$$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right)}}$$

- Os valores do coeficiente de correlação sempre variarão de -1 a +1.Os maiores valores (se negativo ou positivo) implicam em maior grau de correlação.
- Os teste de correlção são realizados pelo SAS através do Proc Corr.

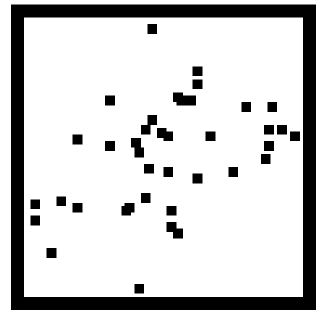
Valores de r e os gráficos respectivos



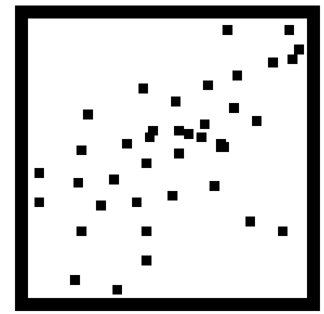
r=0



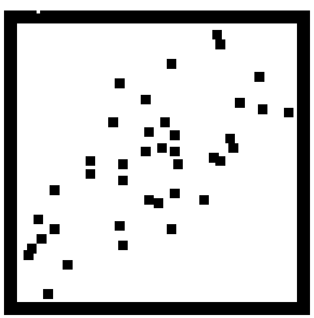
r=.28



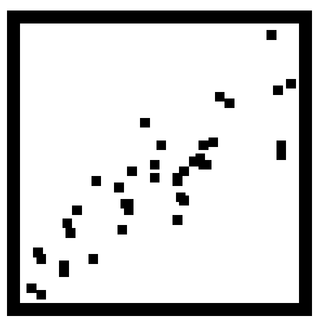
r=.42



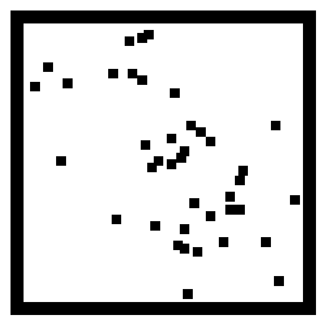
r=.55



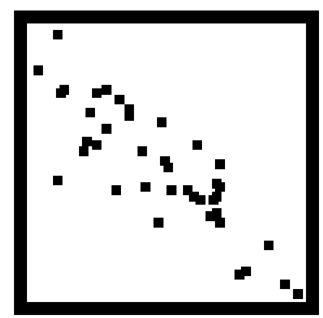
r=.67



r=.86



r=-.55



r=-.85

O PROC CORR DO SAS

```
DATA SOLO;
INPUT PONTO PH MO P CA;
  1 5.0 39  9 27
  2 5.1 40  7 23
  3 4.7 37  8 24
  4 5.9 45 10 31
  5 4.9 38  7 22
  6 4.5 35 11 33
  7 6.0 46  9 28
  8 6.2 48 10 29
  9 5.2 40  6 18
 10 4.0 31  7 20
;;;
ODS PDF FILE='C:\Arquivos2015\Bioestatistica2015\SOLO.PDF';
TITLE2'*** Análise de correlação entre variáveis do solo ***';
TITLE4'*** Experimento na Fazenda Cerradinho - Catanduva - SP ***';
PROC CORR DATA=SOLO ;
VAR PH MO P CA;
RUN;
ODS PDF CLOSE;
```

Pearson é o default. Para Spearman especificar

RESULTADO DA ANÁLISE DE CORRELAÇÃO

The SAS System

****** Análise de correlação das propriedades do solo ******

****** Experimento na Fazenda Cerradinho - Catanduva - SP ******

The CORR Procedure

4 Variables:	PH MO P CA

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
PH	10	5.14347	0.71111	51.43474	3.98277	6.20259
MO	10	39.84571	5.29896	398.45707	31.06792	47.71510
P	10	8.28539	1.55779	82.85389	5.85324	10.55488
CA	10	25.39057	4.73955	253.90574	18.25328	32.50041

Pearson Correlation Coefficients, N = 10 Prob > |r| under H0: Rho=0				
	PH	**MO**	**P**	**CA**
PH	1.00000	0.99918 <.0001	0.43796 0.2055	0.44121 0.2018
MO	0.99918 <.0001	1.00000	0.42892 0.2161	0.43180 0.2127
P	0.43796 0.2055	0.42892 0.2161	1.00000	0.99835 <.0001
CA	0.44121 0.2018	0.43180 0.2127	0.99835 <.0001	1.00000

Pearson ou Spearman?

- Os dois coeficientes fornecem informações importantes, principalmente durante a fase das análises exploratórias quando não conhecemos o comportamento dos dados.
- Os valores do coeficiente já são suficientes para tomar decisão sobre o prosseguimento das análises (dedutivas, explicativas e preditivas)? 0,50, por exemplo?

<u>Análise de regressão</u>: estudo da relação linear entre duas ou mais variáveis.

A mais usada das técnicas estatísticas para análise de dados.

Também chamada de Análise de Regressão Linear

Teste de hipótese e predição

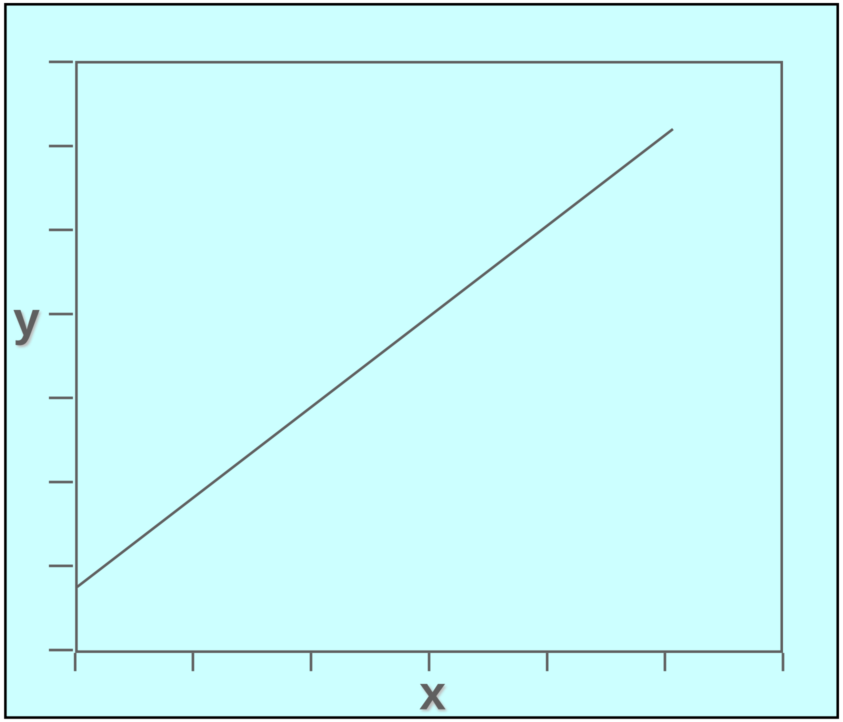
Por quê usar regressão?

- Uma variável ou conjunto de variáveis independentes ou preditoras possuem um efeito causal sobre a variável dependente ou resposta (exemplo: será que a temperatura influencia na germinação das sementes)?
- As suposições sobre a normalidade de Y, independência das observações e normalidade dos erros são cruciais.

Regressão Linear Simples

Fórmula da linha reta

$$y = b_0 + b_1x$$

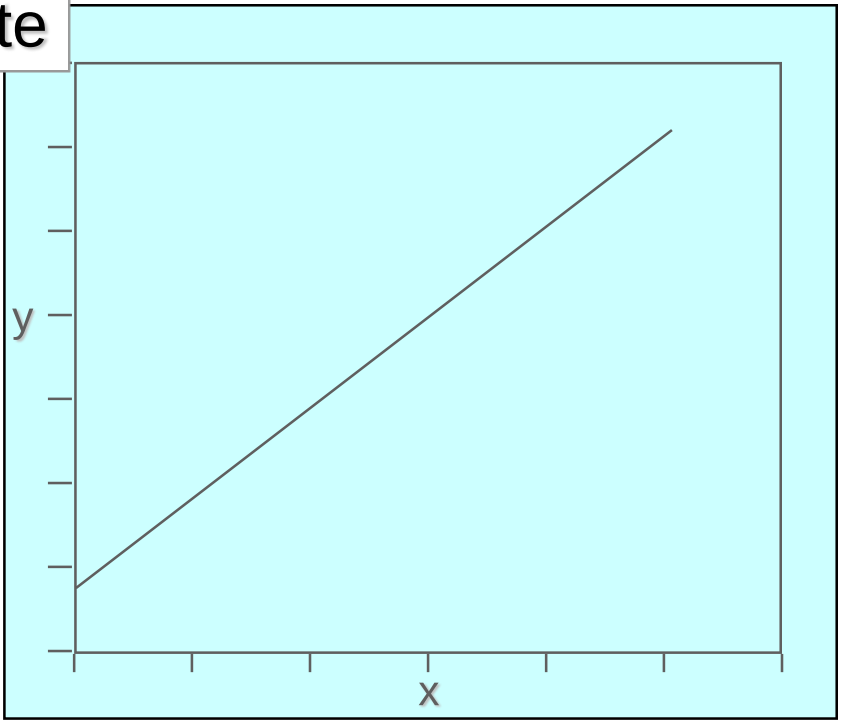


Modelo de Regressão Linear

Fórmula da linha reta

Dependente | Independente

$$y = b_0 + b_1 x$$

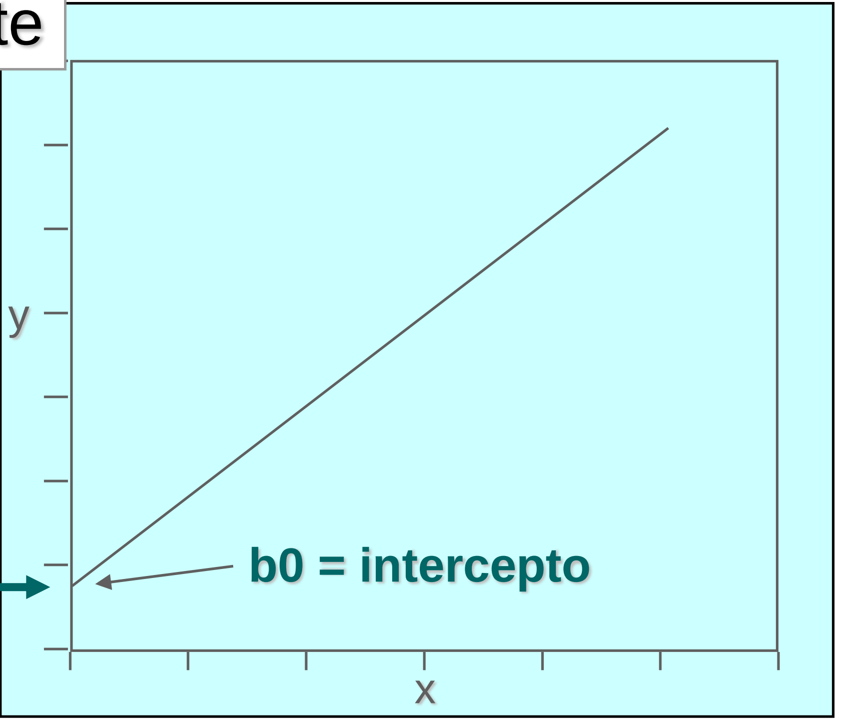


Modelo de Regressão Linear

Fórmula da linha reta

Dependente | Independente

$$y = b_0 + b_1 x$$

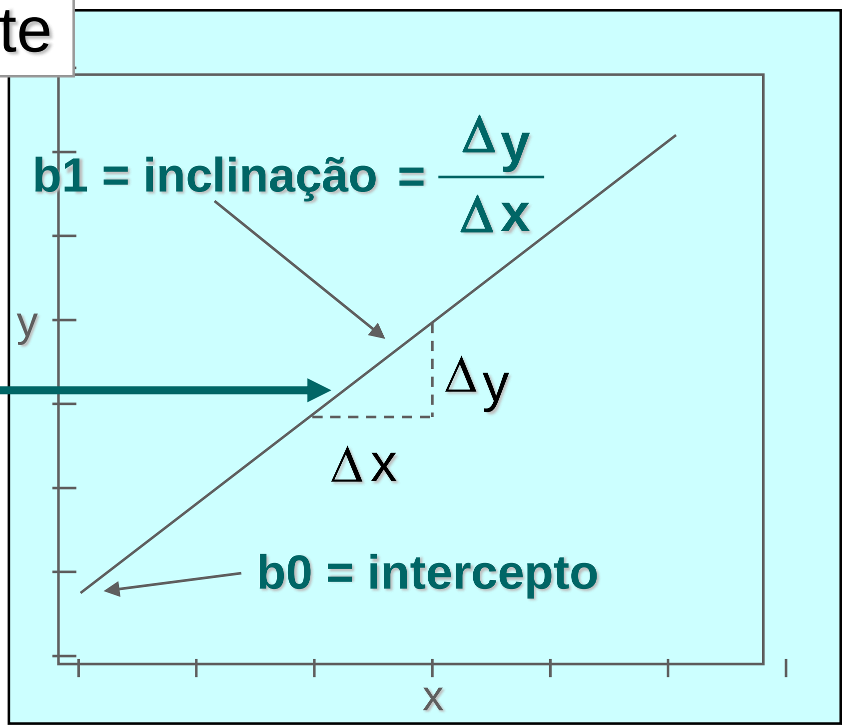


Modelo de Regressão Linear

Fórmula da linha reta

Dependente | Independente

$$y = b_0 + b_1 x$$



Modelo de Regressão Linear

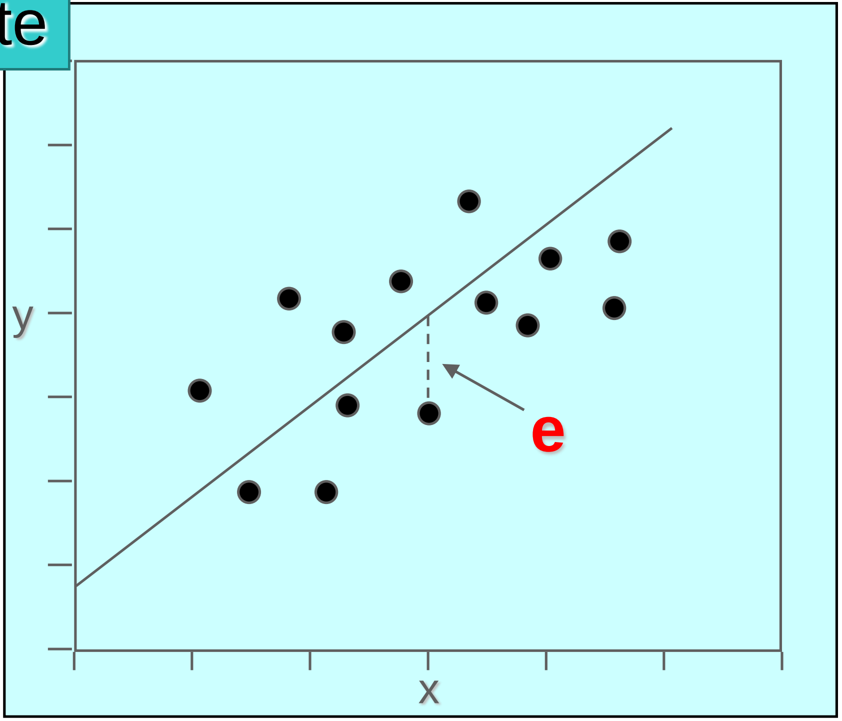
Fórmula da linha reta

Dependente

Independente

$$y = b_0 + b_1 x + e$$

Erro na medição

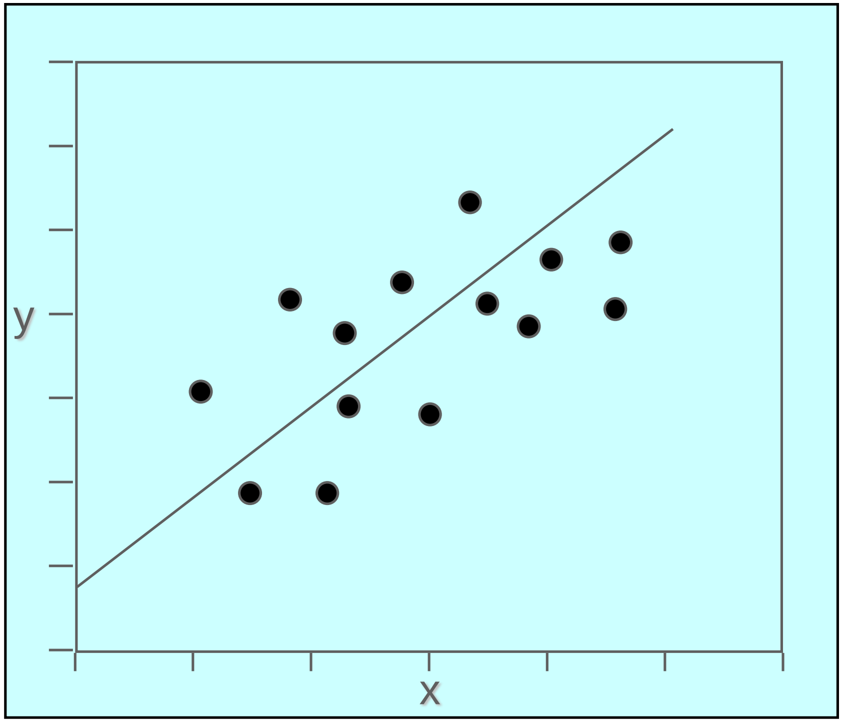


O modelo de regressão linear

Fórmula para a linha reta

$$y = b_0 + b_1 x + e$$

Procuramos estimar esses valores



O método dos quadrados mínimos.

- O Método dos Quadrados Mínimos Ordinários (OLS) encontra o <u>modelo linear que minimiza a soma do quadrado dos erros.</u>
- Este modelo apresenta a melhor explicação/predição dos dados.

$$\textbf{SQNC} = \sum (\widehat{Y_i} - Y_i)^2$$

$$= \sum e_i^{\,2}$$

Fórmulas para calcular os valores dos parâmetros pelo MQMO.

$$\widehat{b}_1 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^{\,2} - (\sum X_i)^2}$$

$$= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

$$\widehat{a} = \overline{Y} - \widehat{b}_1 \overline{X}$$

Testes de inferência

- Teste t para os coeficientes.
- Teste F para o modelo.

Medidas de ajustamento do modelo

- O Coeficiente de Correlação.
- O R^2 (Coeficiente de determinação).

$$\boxed{R^2 = \left(1 - \frac{SQResíduo}{SQTotal}\right) x\, 100}$$

Programa SAS para análise de regressão simples

```
DATA A;
INPUT DAP BIOMASSA;
DATALINES;
12 34
14 45
23 89
56 138
87 379
;
PROC REG DATA = A;
MODEL BIOMASSA = DAP;
RUN;
```

Colocar os comandos ODS e TITLE

Outros modelos

```
DATA A;
INPUT DAP BIOMASSA;
LBIOMA=LOG(BIOMASSA);
LDAP=LOG(DAP);
DATALINES;
12 34
14 45
23 89
56 138
87 379
;
PROC REG DATA = A PLOTS(ONLY)=PREDICTIONS(X=DAP);
MODEL LBIOMA = DAP;
MODEL LBIOMA = LDAP;
RUN;
```

Colocar os comandos ODS e TITLE

The SAS System

****** ANÁLISE DE REGRESSÃO - BIOMASSA E DAP ******

The REG Procedure

Model: MODEL1

Dependent Variable: LBIOMA

Number of Observations Read	5
Number of Observations Used	5

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	3.43791	3.43791	44.26	0.0069
Error	3	0.23302	0.07767		
Corrected Total	4	3.67092			

Root MSE	0.27870	**R-Square**	0.9365
Dependent Mean	4.53729	**Adj R-Sq**	0.9154
Coeff Var	6.14236		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	3.43881	0.20687	16.62	0.0005
DAP	1	0.02861	0.00430	6.65	0.0069

The SAS System

**** ANÁLISE DE REGRESSÃO - BIOMASSA E DAP ****

The REG Procedure

Model: MODEL2

Dependent Variable: LBIOMA

Number of Observations Read	5
Number of Observations Used	5

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	3.47696	3.47696	53.78	0.0052
Error	3	0.19396	0.06465		
Corrected Total	4	3.67092			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	0.93137	0.50469	1.85	0.1622
LDAP	1	1.07635	0.14677	7.33	0.0052

Root MSE	0.25427	**R-Square**	0.9472
Dependent Mean	4.53729	**Adj R-Sq**	0.9295
Coeff Var	5.60405		

Gráfico do modelo 1

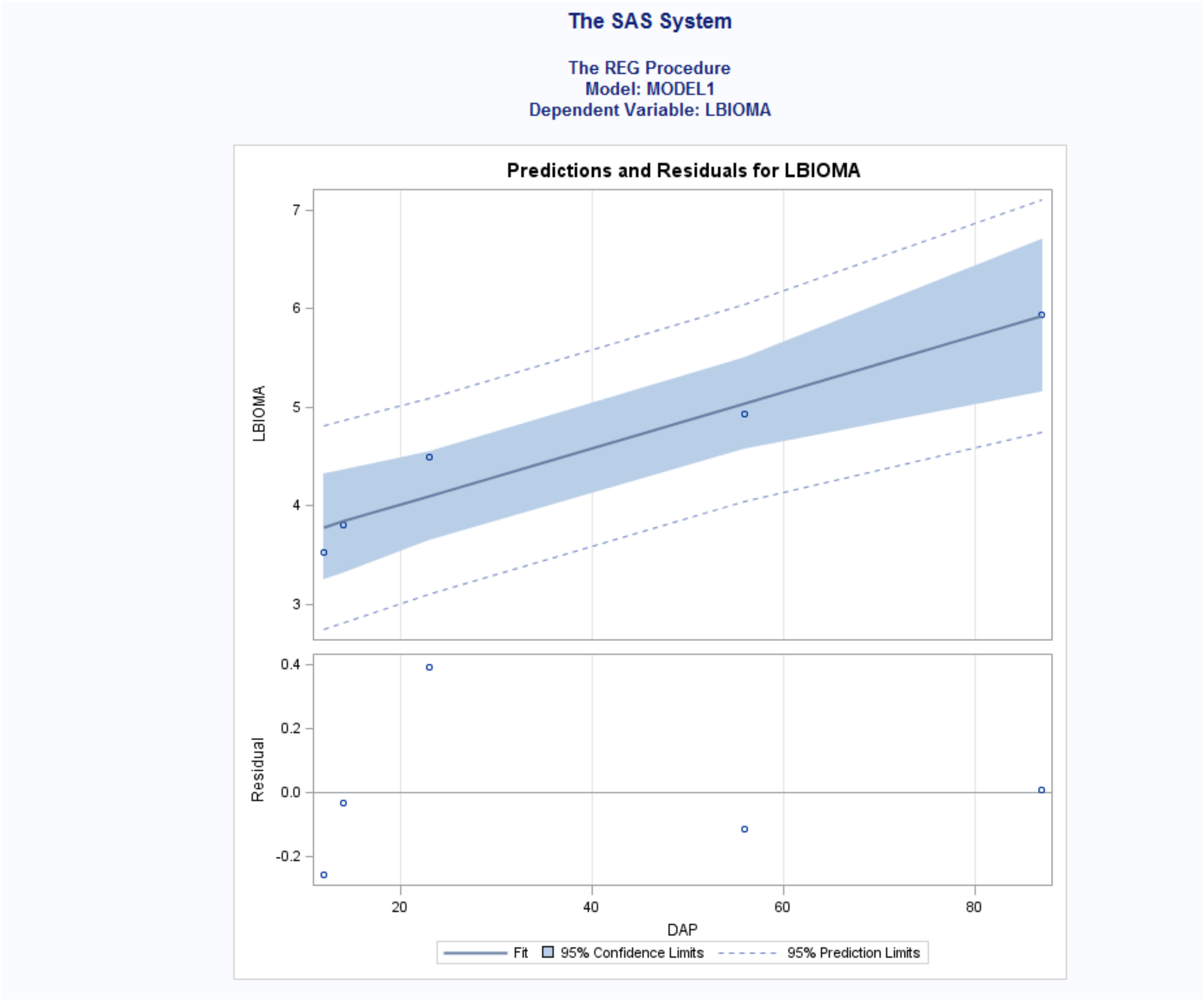
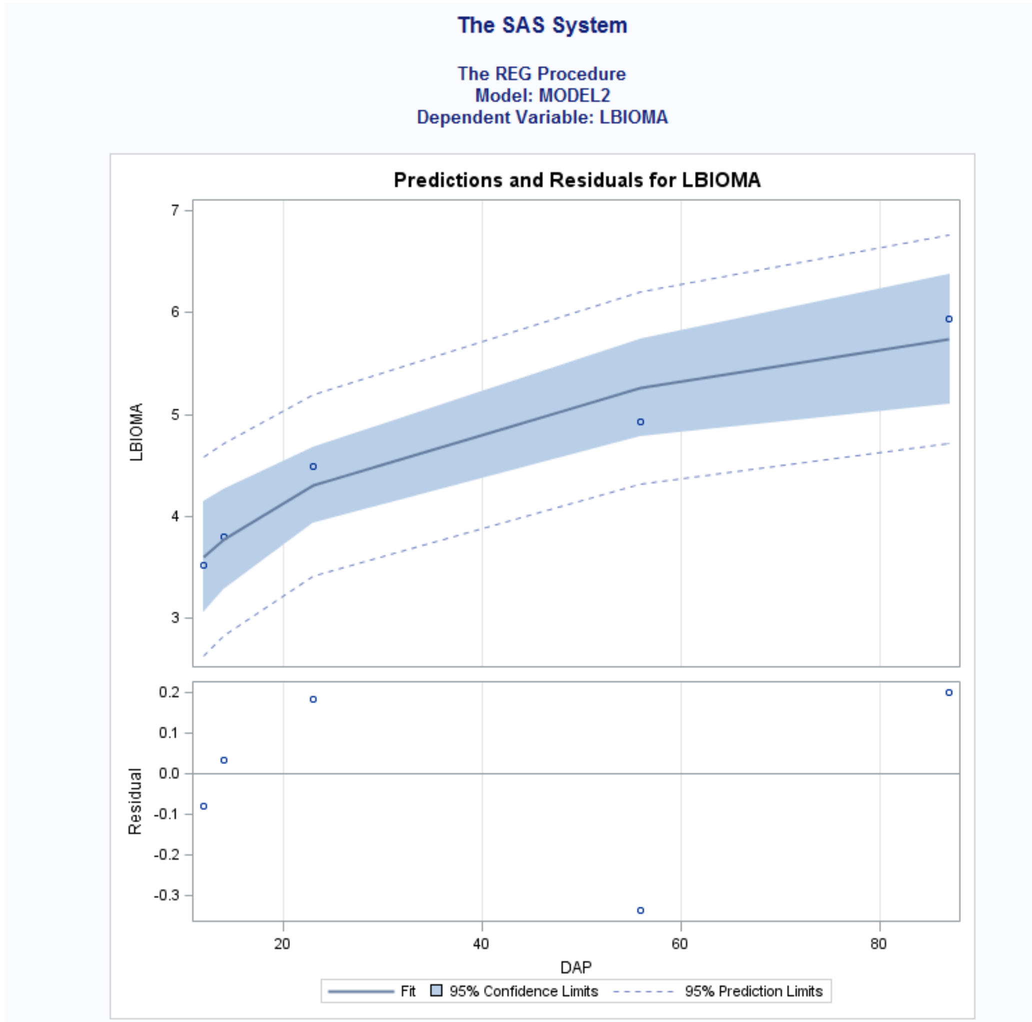


Gráfico do modelo 2



Critérios para a seleção de modelo.

- Se possível justificar através de algum valor quantificável (R_2, teste F, por exemplo)
- A experiência pode ser importante (conhecimento sobre o assunto)
- Algumas ideias pré-concebidas podem ser importantes.
- Como o modelo pode ser aplicado na prática (modelos complexos podem ser de difícil aceitação ou utilização)
- **Seleção de Modelo**: a tarefa de selecionar um modelo matemático a partir de diversos modelos **potenciais**, com fortes **evidências.**

Regressão Linear Múltipla.

- O modelo de regressão linear múltipla é uma extensão do modelo simples com apenas duas variáveis (independente e dependente). Ao adicionar no modelo mais uma variável independente é criado um espaço de mútipla dimensão. Por exemplo, se existirem duas variáveis independentes estamos ajustando os pontos a um “plano no espaço”.

O modelo linear básico.

$$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki} + e_i$$

As suposições do modelo:

- Os erros possuem a distribuição normal.
- Os resíduos são homoscedásticos.
- Não há correlação serial.
- Não há multicolinearidade.
- As variáveis independentes são fixas. (não-estocásticas)
- Existem mais dados que estimativas de parâmetros.
- O modelo é linear.

PROGRAMA SAS PARA ANÁLISE DE REGRESSÃO MÚLTIPLA

```
DATA A;
INPUT DAP ALT BIOMASSA;
LBIOMA=LOG(BIOMASSA);
LDAP=LOG(DAP);
LALT=LOG(ALT);
DATALINES;
12 10 34
14 11 45
18 9 69
23 16 89
31 14 80
56 18 138
66 19 190
87 23 379
91 22 408
;
PROC REG DATA = A PLOTS(ONLY)=PREDICTIONS(X=DAP);
MODEL BIOMASSA = DAP ALT;
MODEL LBIOMA = DAP ALT;
MODEL LBIOMA = LDAP LALT;
RUN;
```

Colocar os comandos ODS e TITLE

*** ANÁLISE DE REGRESSÃO - BIOMASSA COM DAP E ALT ***

The REG Procedure

Model: MODEL1

Dependent Variable: BIOMASSA

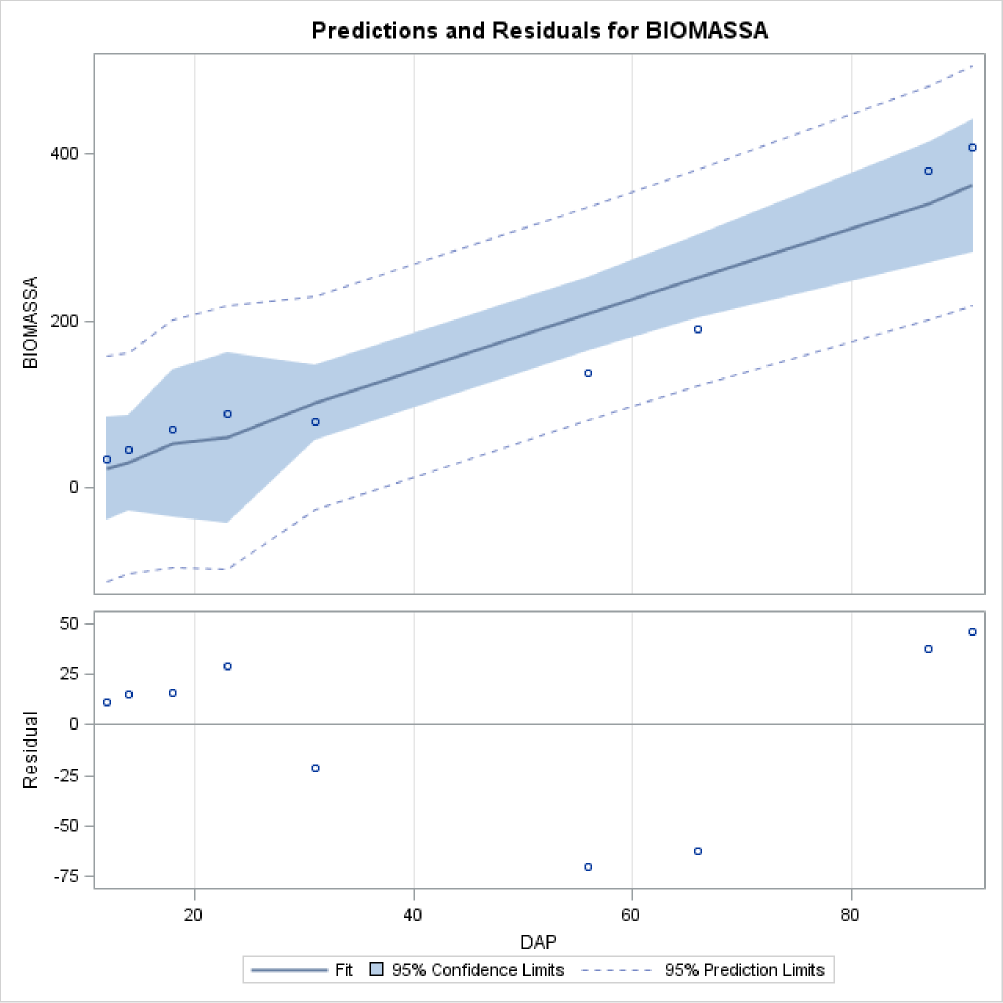
Number of Observations Read	9
Number of Observations Used	9

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	145243	72622	30.21	0.0007
Error	6	14422	2403.61163		
Corrected Total	8	159665			

Root MSE	49.02664	**R-Square**	0.9097
Dependent Mean	159.11111	**Adj R-Sq**	0.8796
Coeff Var	30.81283		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-9.39304	97.84918	-0.10	0.9267
DAP	1	4.65325	1.71180	2.72	0.0347
ALT	1	-2.36237	10.45763	-0.23	0.8288

Gráfico de modelo 1



*** ANÁLISE DE REGRESSÃO - BIOMASSA COM DAP E ALT ***

The REG Procedure

Model: MODEL2

Dependent Variable:

LBIOMA

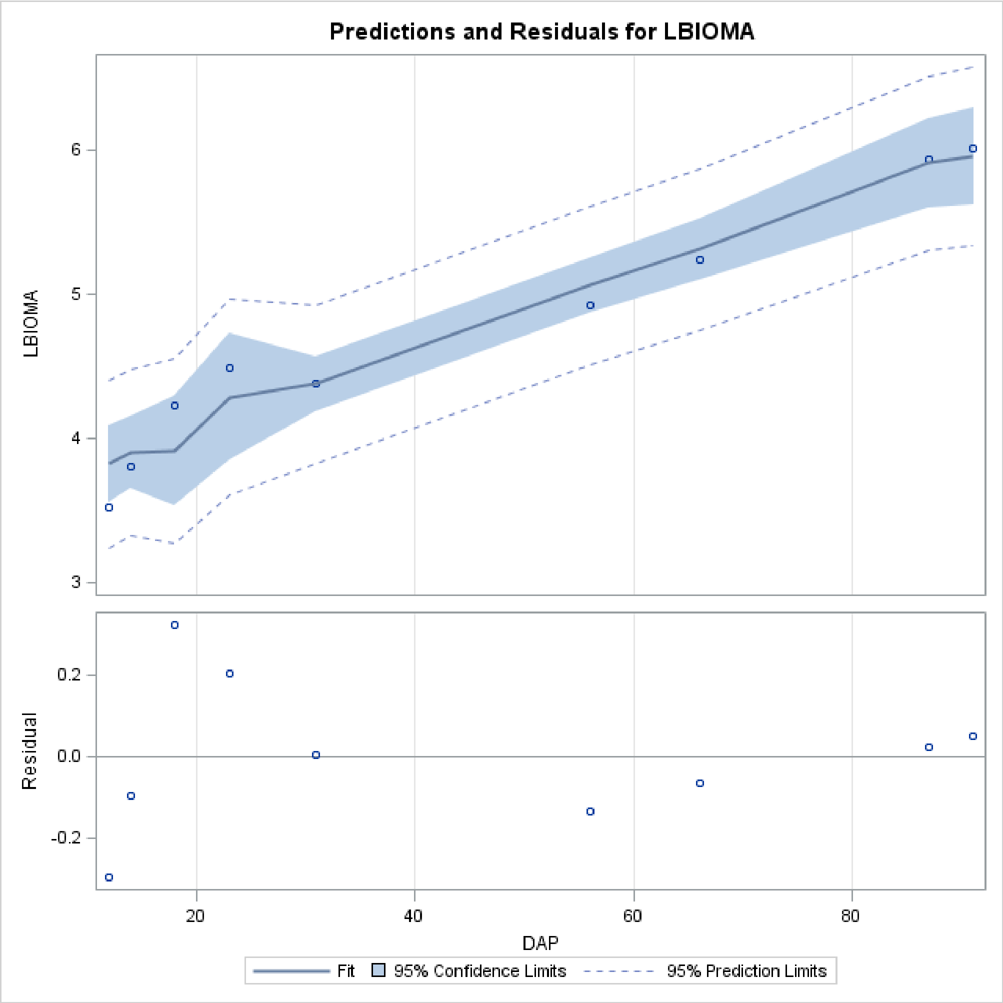
Number of Observations Read	9
Number of Observations Used	9

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	5.86419	2.93209	65.55	<.0001
Error	6	0.26837	0.04473		
Corrected Total	8	6.13256			

Root MSE	0.21149	**R-Square**	0.9562
Dependent Mean	4.72899	**Adj R-Sq**	0.9417
Coeff Var	4.47222		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	3.18668	0.42210	7.55	0.0003
DAP	1	0.02127	0.00738	2.88	0.0280
ALT	1	0.03814	0.04511	0.85	0.4303

Gráfico do modelo 2



Análise do modelo 3

The SAS System

The REG Procedure
Model: MODEL3
Dependent Variable: LBIOMA

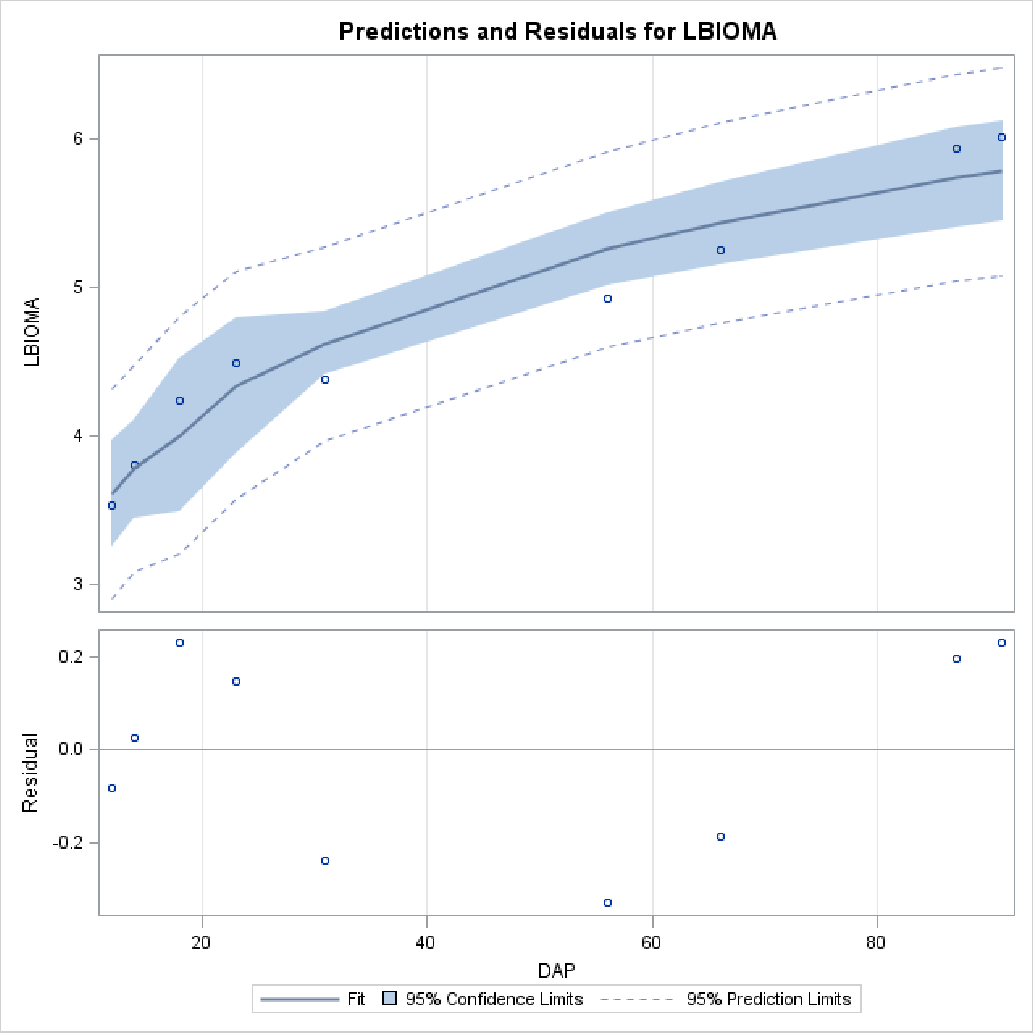
Number of Observations Read	9
Number of Observations Used	9

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.75655	2.87828	45.93	0.0002
Error	6	0.37600	0.06267		
Corrected Total	8	6.13256			

Root MSE	0.25033	R-Square	0.9387
Dependent Mean	4.72899	Adj R-Sq	0.9183
Coeff Var	5.29359		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.73518	0.99748	0.74	0.4889
LDAP	1	1.01128	0.31383	3.22	0.0181
LALT	1	0.15634	0.71880	0.22	0.8350

Gráfico de modelo 3



Exercício em classe

- Com dados de experimento com doses de N em cultivares de cevada plantadas no Cerrado de altitude (arquivo excel Dados_doses_N-Cevada_Bahia.xlsx) testar 4 modelos matemáticos: linear simples, linear quadrático, log-log e log-inverso. Devo fazer um modelo por cultivar ou posso ter um modelo para todos os cultivares?

BOAS ANÁLISES ESTATÍSTICAS!!!