

# Análise Exploratória de Dados

Igor Rocha, Isaac Lima,  
João Rupp, ValcÍrio Francisco

2021.2



# Adaptando slide de

Murillo Almeida dos Santos Torres  
2021.1



Associação

# Associação

4

Associação na estatística é o estudo de uma relação entre duas variáveis.

Este tipo de operação é útil para aplicações como:

- Possibilidade de estudar uma através da outra;
- Tentar prever os valores de uma através da outra;

As técnicas de associação são:

- Covariância;
- Correlação linear simples.

Covariância



## Introdução

### COVARIÂNCIA

### CORRELAÇÃO L. SIMPLES

### REGRESSÃO L. SIMPLES

6

O termo **Covariância** no ramo da probabilidade e estatística está relacionado à medida da variabilidade conjunta entre duas variáveis aleatórias. Por exemplo, tomemos duas variáveis  $X$  e  $Y$ , ao analisarmos, se a maioria dos maiores valores de  $X$  corresponde à maioria dos maiores valores de  $Y$ , e o mesmo comportamento se aplica à maioria dos menores valores das mesmas, temos que a covariância entre elas é positiva. Caso o comportamento seja contrário, ou seja, os maiores valores de  $X$  corresponderem aos menores valores de  $Y$ , e vice-versa, temos que a covariância é negativa.

## Definição

### COVARIÂNCIA

### CORRELAÇÃO L. SIMPLES

### REGRESSÃO L. SIMPLES

7

A fórmula da covariância entre duas variáveis reais aleatórias, X e Y, distribuídas em conjunto e com variância finita é definida como a esperança do produto dos seus desvios a partir dos seus valores individuais. A Esperança Matemática (E) representa o valor esperado de um conjunto de resultados, que equivale à soma dos produtos individuais de valor multiplicada pela probabilidade de ocasião, também conhecida como média.

$$\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Executando a distributiva e aplicando a propriedade de linearidade de expectativas, podemos simplificar a fórmula para que a mesma seja o valor esperado do produto de X e Y menos o produto do valor esperado de X e de Y, ficando assim:

$$\text{cov}(X, Y) = E[XY] - E[X] E[Y]$$

## Definição

### COVARIÂNCIA

### CORRELAÇÃO L. SIMPLES

### REGRESSÃO L. SIMPLES

8

Variáveis aleatórias as quais a covariância é **zero** são chamadas de **variáveis não correlacionadas**, ou seja, elas não possuem características de linearidade entre si. As unidades de medida de uma covariância **cov(X, Y)** são as de **X** vezes as de **Y**. Em contrapartida, coeficientes de correlação, os quais dependem da covariância são medidas adimensionais de associação linear.



## Desdobramentos

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

9

Quando trabalhamos com **variáveis discretas**, ou seja variáveis que entre um valor e outro não existe valor intermediário (i.e., pontos), utilizamos a seguinte fórmula:

$$\text{cov}(X, Y) = \sum_X \sum_Y (X - E(X))(Y - E(Y))p(x, y)$$

A covariância padronizada, chama-se **coeficiente de correlação** entre X e Y, o qual denotaremos por **p(x,y)**.

## Desdobramentos

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

10

Quando porém as variáveis trabalhadas forem contínuas, ou seja, quando entre dois valores ( $X_1$  e  $X_2$ ) existirem infinitos valores intermediários (i.e., intervalos), não podemos mais utilizar a fórmula anterior pois agora precisamos de um método capaz de nos dar o resultado de toda uma região, de todo um intervalo, e para isso nós utilizamos integrais definidas como já visto em Cálculo 2:

$$\text{cov}(X, Y) = \int_{-\infty}^{+\infty} \left\langle \int_{-\infty}^{+\infty} (X - E(X))(Y - E(Y)) f(x, y) dx dy \right\rangle$$

## Observações

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

11

Se  $X$  e  $Y$  são variáveis aleatórias **independentes**, a  $\text{Cov}(X,Y)$  será igual a 0.

Porém, se a  $\text{Cov}(X,Y)$  for **igual a 0**, isso **NÃO** significa que as variáveis serão independentes.

## Covariância Positiva

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

12

Obter uma covariância positiva significa, na prática, que as duas variáveis têm o **mesmo comportamento**, ou seja, quando uma delas aumenta, a segunda também aumentará e quando uma delas diminui a outra também se comporta da mesma forma.

Isto fará com que a maior parte das observações recaiam no 1º e 3º quadrante, demonstrando portanto um relacionamento positivo entre as variáveis.



## Propriedades

### COVARIÂNCIA

### CORRELAÇÃO L. SIMPLES

### REGRESSÃO L. SIMPLES

13

Para quaisquer variáveis aleatórias **X**, **Y**, **Z** e uma constante **c**, temos:

$$\text{Cov}(X, X) = \text{Var}(X)$$

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$

$$\text{Cov}(cX, Y) = c\text{Cov}(X, Y)$$

$$\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$$

## Covariância Amostral

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

14

Na **ausência** da distribuição de probabilidade, podemos trabalhar com uma **amostra** da população, assim:

$$cov(X, Y) = \sum_i \frac{(Xi - \bar{X})(Yi - \bar{Y})}{n - 1}$$

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

15

Consideremos duas variáveis aleatórias:

- **M**: rendimento acadêmico em matemática;
- **L**: rendimento acadêmico em línguas

Rendimento acadêmico:

$$\Sigma M = 480$$

$$M = 60$$

$$\Sigma L = 400$$

$$L = 50$$

Obs:	01	02	03	04	05	06	07	08
M:	36	80	50	58	72	60	56	68
L:	35	65	60	39	48	44	48	61

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

16

Obs	M	L	$m = [M - E(M)]$	$l = [L - E(L)]$	$m \cdot l$
01	36	35	-24	-15	360
02	80	65	20	15	300
03	50	60	-10	10	-100
04	58	39	-2	-11	22
05	72	48	12	-2	-24
06	60	44	0	-6	0
07	56	48	-4	-2	8
08	68	61	8	11	88



## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

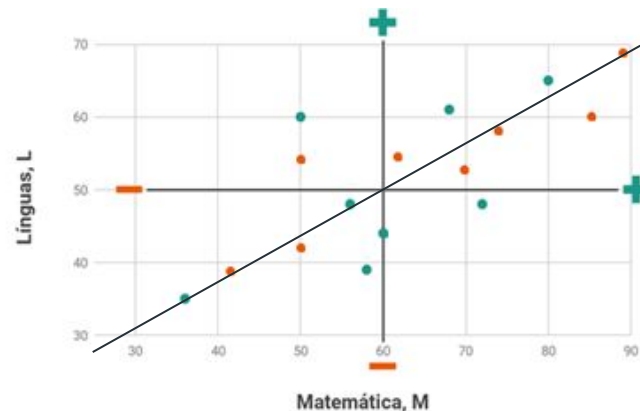
REGRESSÃO L. SIMPLES

17

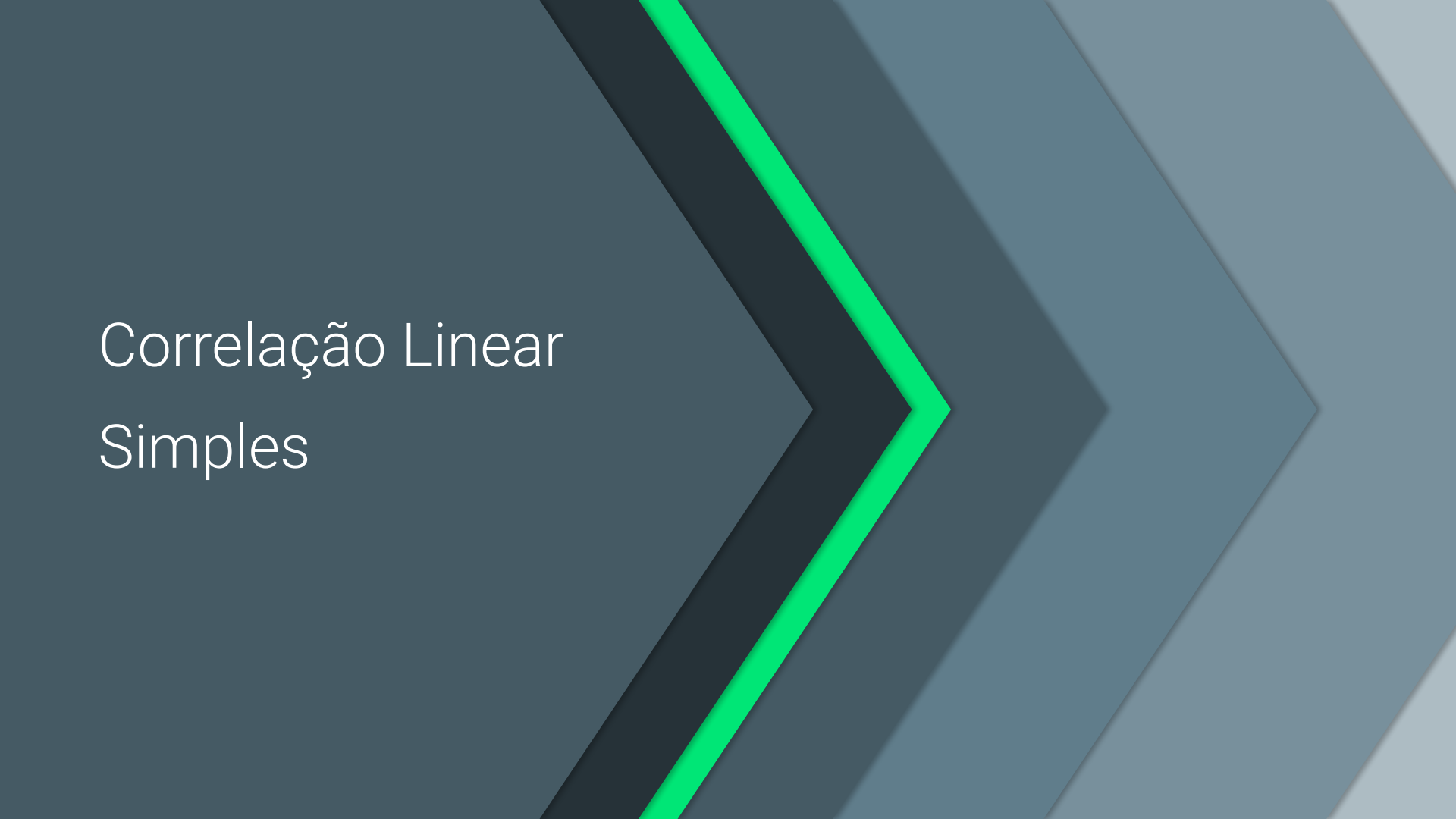
$$\text{Cov}(M, L) = \frac{\sum i(M_i - \bar{M})(L_i - \bar{L})}{n-1} = \frac{654}{7} = 93,43$$

Como possuem comportamentos semelhantes, ou seja, quando uma variável aumenta a outra também aumenta e o mesmo acontece para quando uma diminui, a maior parte das observações recairão nos 1º e 3º quadrantes.

Consequentemente, a maior parte dos produtos (m.l) serão positivos, bem como sua soma ( $\sum ml$ ), demonstrando um relacionamento positivo entre M e L.



# Correlação Linear Simples

The background of the slide features a series of overlapping chevron shapes pointing to the right. The chevrons are in various shades of blue, ranging from a dark navy blue to a light sky blue. A single, thick, bright red chevron is superimposed over the others, starting from the top left and pointing towards the right side of the frame.

## Introdução

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

19

O termo correlação significa **relação** em dois sentidos (co + relação), na **estatística**, a verificação da existência e do grau de relação entre as variáveis é o objeto de estudo da correlação.

## Relação entre duas variáveis

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

20

Para expressar numericamente o quanto as duas variáveis tendem a mudar juntas, utilizamos o coeficiente de correlação. O coeficiente descreve a força e a direção da relação. Para calcular o coeficiente de uma correlação entre duas variáveis, podemos recorrer a dois métodos já solidificados, sendo eles:

- **Pearson**
- **Spearman**

## Pearson ou Spearman?

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

21

A correlação de **Pearson** avalia a relação linear entre duas variáveis quantitativas.

**Exemplo:** Avaliar a quantidade de fertilizante e número de sementes Germinadas.

Já a correlação de **Spearman** avalia a relação monotônica entre duas variáveis contínuas ou ordinais. Em uma relação monotônica, as variáveis tendem a mudar juntas mas não necessariamente a uma taxa constante. A correlação de **Spearman** é muito usada para avaliar relações envolvendo variáveis ordinais.

**Exemplo:** Avaliar indivíduos com enxaqueca de respectivas idades correlacionando com a intensidade da sua dor (leve, moderado, forte, muito forte).

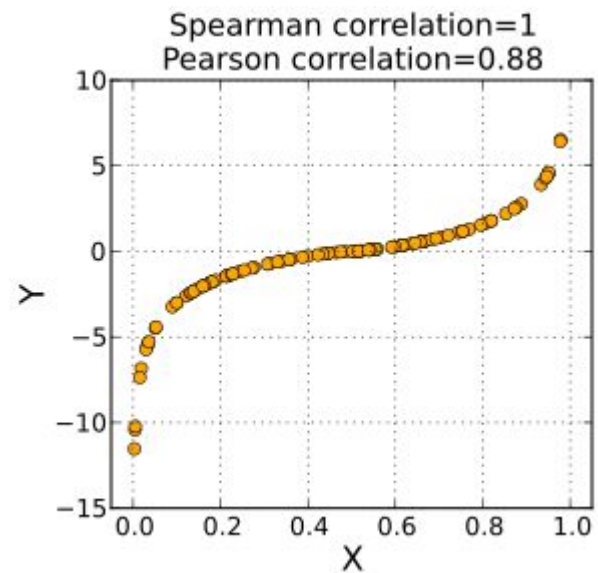
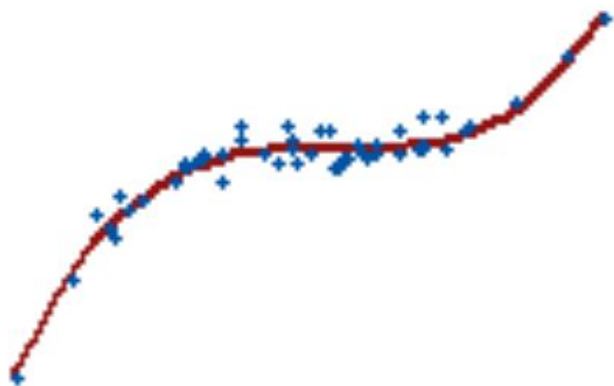
## Relação monotônica

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

22



## Fórmula Coeficiente de Pearson para População

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

23

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}, \sigma_x \sigma_y > 0$$

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \rightarrow \begin{aligned} & \text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y) \\ & \sigma_x \sigma_y = \sqrt{Var(x) Var(y)} \\ & \quad \downarrow \quad \quad \downarrow \\ & Var(y) = \sum (y - E(Y))^2 P(y) = [E(Y^2) - (E(Y))^2] \\ & \quad \quad \downarrow \\ & Var(x) = \sum (x - E(X))^2 P(x) = [E(X^2) - (E(X))^2] \end{aligned}$$

## Fórmula Coeficiente de Pearson para Amostra

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

24

$$r = \frac{\sum (x - \bar{x}) \cdot (y - \bar{y})}{\sqrt{\left[ \sum (x - \bar{x})^2 \right] \cdot \left[ \sum (y - \bar{y})^2 \right]}}$$

$$r = \frac{n \cdot \sum x \cdot y - (\sum x) \cdot (\sum y)}{\sqrt{n \cdot \sum x^2 - (\sum x)^2} \cdot \sqrt{n \cdot \sum y^2 - (\sum y)^2}}$$

$$r = \frac{\sum x \cdot y - \frac{(\sum x) \cdot (\sum y)}{n}}{\sqrt{\left[ \sum x^2 - \frac{(\sum x)^2}{n} \right] \cdot \left[ \sum y^2 - \frac{(\sum y)^2}{n} \right]}}$$



## Interpretando os resultados

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

25

O coeficiente de correlação  $r$  linear é um número puro que varia de  $-1$  a  $+1$  e sua interpretação dependerá do valor numérico e do sinal, como segue:

Coeficiente de Correlação	Correlação
$r = 1$	Perfeito Positivo
$0.8 \leq r < 1$	Forte Positiva*
$0.5 \leq r < 0.8$	Moderado Positiva*
$0.1 \leq r < 0.5$	Fraca Positiva*
$0 \leq r < 0.1$	Íntima Positiva*
$r = 0$	Nula
$-0.1 \leq r < 0$	Íntima Negativa*
$-0.5 \leq r < -0.1$	Fraca Negativa*
$-0.8 \leq r < -0.5$	Moderado Negativa*
$-1 \leq r < -0.8$	Forte Negativa*
$r = -1$	Perfeito Negativo*

## Diagrama de Dispersão

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

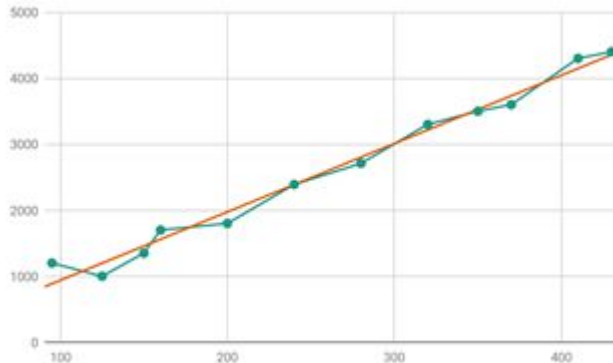
REGRESSÃO L. SIMPLES

26

Os pares de valores das duas variáveis na correlação poderão ser colocados num diagrama cartesiano chamado “**diagrama de dispersão**”. A vantagem de construir um diagrama de dispersão está em que, muitas vezes sua simples observação já nos dá uma idéia bastante boa de como as duas variáveis se relacionam.

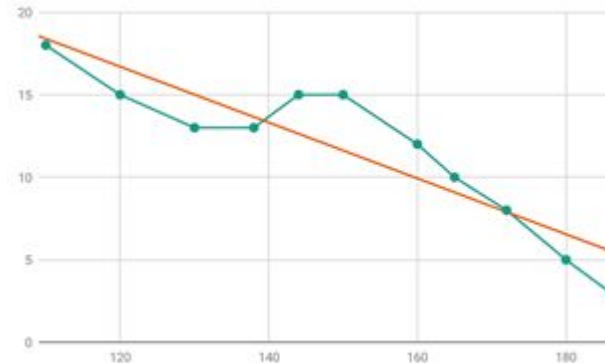
Correlação **positiva e forte**

$$r = 0,984$$



Correlação **negativa e forte**

$$r = -0,819$$



## Diagrama de Dispersão

COVARIÂNCIA

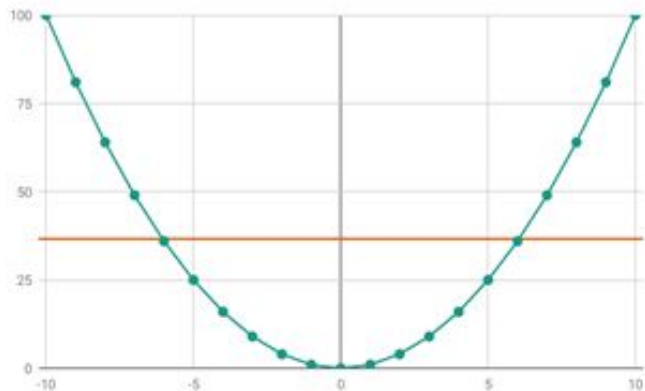
CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

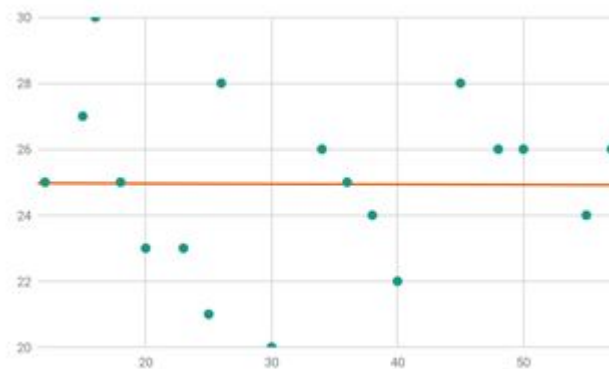
27

Correlação **nula**

$$r = 0$$

Correlação **fraca, quase nula**

$$r = 0,0068$$



## Diagrama de Dispersão

COVARIÂNCIA

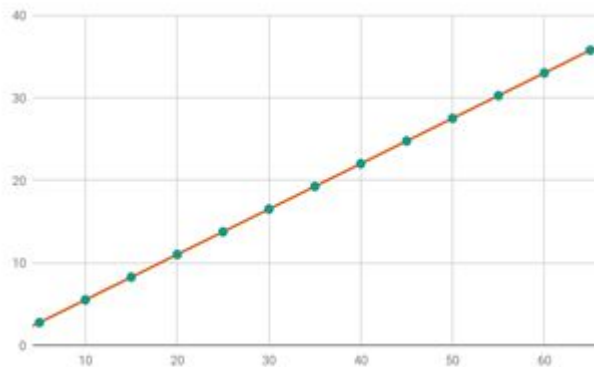
CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

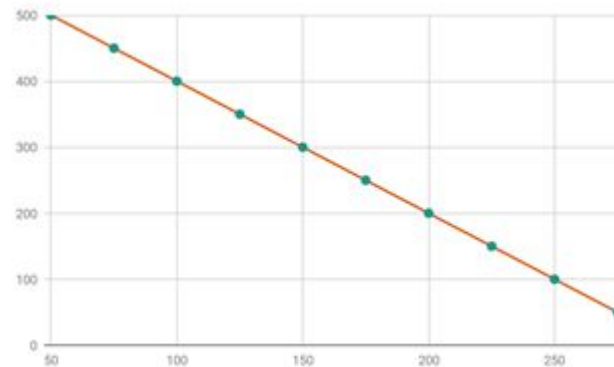
28

Correlação **positiva** e **perfeita**

$$r = 1$$

Correlação **negativa** e **perfeita**

$$r = -1$$



## Vamos por em prática!

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

29

Região	Taxa de Mortalidade Infantil (X)	Taxa de Analfabetismo (Y)	X*Y	X <sup>2</sup>	Y <sup>2</sup>
Norte	18,1	9,1	164,71	327,61	82,81
Nordeste	17,5	16,2	283,5	306,25	262,44
Centro-Oeste	14,8	5,7	84,36	219,04	32,49
Sudeste	10,7	4,3	46,01	114,49	18,49
Sul	9,7	4,1	39,77	94,09	16,81
Somatório	<b>70,8</b>	<b>39,4</b>	<b>618,35</b>	<b>1061,48</b>	<b>413,04</b>
(Somatório) <sup>2</sup>				<b>1.126.740</b>	<b>170.602</b>

$$r = \frac{\sum x,y - \frac{(\sum x) \cdot (\sum y)}{n}}{\sqrt{\left[ \sum x^2 - \frac{(\sum x)^2}{n} \right] \cdot \left[ \sum y^2 - \frac{(\sum y)^2}{n} \right]}}$$

$$r = \frac{618,35 - \frac{70,8 \cdot 39,4}{5}}{\sqrt{\left[ 1061,48 - \frac{1126740}{5} \right] \cdot \left[ 413,04 - \frac{170602}{5} \right]}}$$

$$r = \frac{60.446}{\sqrt{[-224.286,52] \cdot [-33.707,36]}}$$

$$r = 0.7773424$$

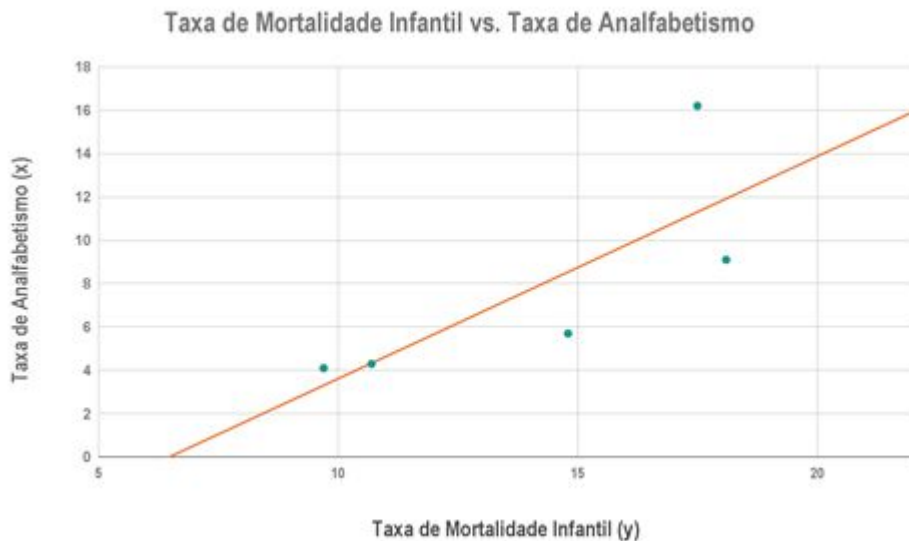
## Gráfico de Dispersão

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES


30



Coeficiente de Correlação	Correlação
$0,5 \leq r < 0,8$	Moderado Positiva

$$r = 0.7773424$$

# Regressão Linear Simples



## Introdução

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

32

A **Regressão Linear Simples**, é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativa de forma que uma variável pode ser predita a partir de outra ou outras.

O modelo de regressão é um dos métodos estatísticos mais usados para investigar a relação entre variáveis.



## Aplicação

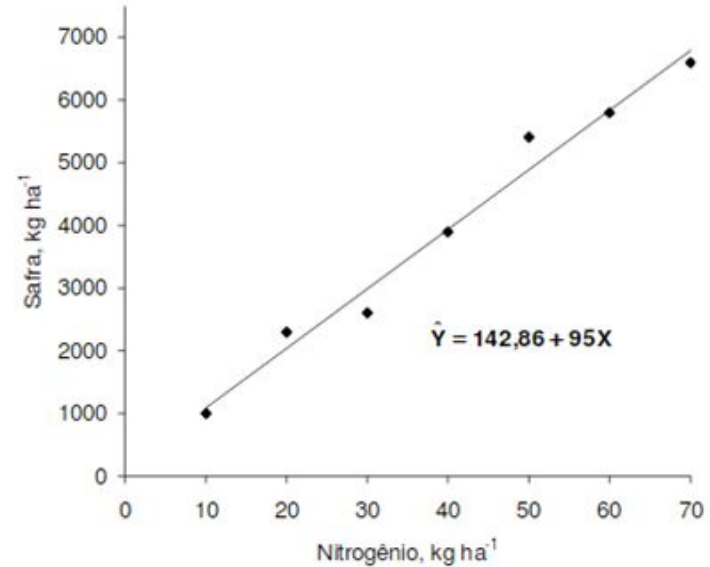
COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

33

Exemplo ilustrativo de regressão linear simples. A safra do milho em função de doses crescentes de adubo nitrogenado aplicado em cobertura.



## Aplicação

COVARIÂNCIA

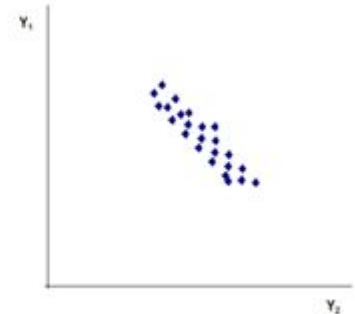
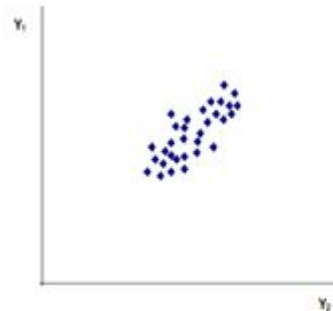
CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

34

A análise de correlação é indicada para estudar o grau de associação linear entre variáveis aleatórias. Ou seja, essa técnica é empregada, especificamente, para se avaliar o grau de covariação entre duas variáveis aleatórias: se uma variável aleatória **Y1** aumenta, o que acontece com uma outra variável aleatória **Y2**?

Aumenta, diminui ou não altera?



## Informações

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

35

Na análise de regressão uma **resposta unilateral** é esperada: alterações em X (fator quantitativo) podem implicar em alterações em Y, mas alterações em Y não resultam em alterações em X.

Quando se deseja verificar a existência de alguma relação estatística entre uma ou mais variáveis fixas, independentes, sobre uma variável aleatória, denominada dependente, utiliza-se a análise de regressão.

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

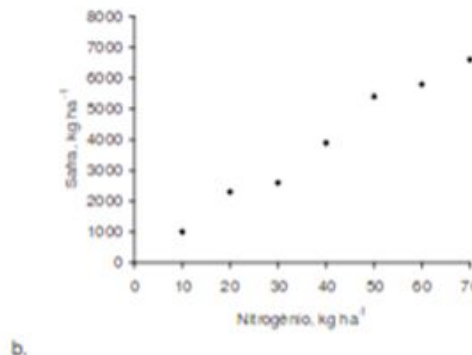
REGRESSÃO L. SIMPLES

36

Para exemplificar, vamos considerar que conduzimos um experimento submetendo plantas de milho a doses crescentes de nitrogênio.

Naturalmente, a produção será dependente da quantidade aplicada desse fertilizante, X.  
(unilateralidade)

Assim, o fertilizante nitrogenado aplicado é a variável independente, e cada uma das quantidades aplicadas são seus níveis,  $x_i$  ( $10 \rightarrow 70 \text{ kg ha}^{-1}$ ).  
Cada variável mensurada na cultura do milho, sujeita a influência dos níveis  $x_i$  da variável independente, é chamada “variável dependente” ou “fator resposta”.



## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

37

Poderia-se medir, por exemplo, o número de espigas por planta ( $Y_1$ ), a altura média das plantas ( $Y_2$ ), o peso de 1.000 grãos ( $Y_3$ ), o teor de proteínas dos grãos ( $Y_4$ ), o teor de gordura dos grãos ( $Y_5$ ), etc.

Como a aplicação do fertilizante não depende da safra, designamos-lá “variável independente” ou “regressor”.

Podemos estudar via análise de regressão o efeito da variável, neste caso, fixa, independente,  $X$  (dose de nitrogênio), sobre as variáveis aleatórias, ou dependentes,  $Y_i$  (produção de matéria seca, teor de proteínas dos grãos, teor de gordura dos grãos, etc.). Diz-se regressão de  $Y$  sobre  $X$ .

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

38

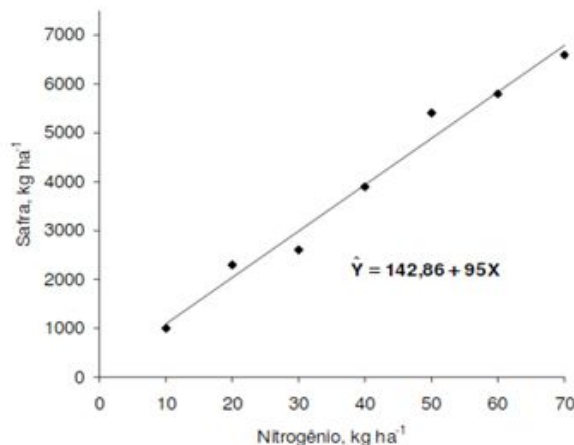
Se grafarmos a safra, Y, decorrente das diversas aplicações, X, de nitrogênio, poderemos observar uma dispersão análoga a figura ao lado.

A aplicação de nitrogênio afeta a safra.

Podemos, por meio de uma equação, relacionando X e Y, descrever como afeta.

Estimar uma equação é geometricamente equivalente a ajustar uma curva àqueles dados dispersos, isto é, a “regressão de Y sobre X”.

Esta equação será útil como descrição breve e precisa de prever a safra Y para qualquer quantidade X de nitrogênio.



## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

39

Vamos considerar um estudo sobre a influência do nitrogênio aplicado em cobertura sobre a safra do milho.

Suponhamos que só dispomos de recursos para fazer sete observações experimentais.

X Nitrogênio kg ha <sup>-1</sup>	Y Safra kg ha <sup>-1</sup>
10	1.000
20	2.300
30	2.600
40	3.900
50	5.400
60	5.800
70	6.600

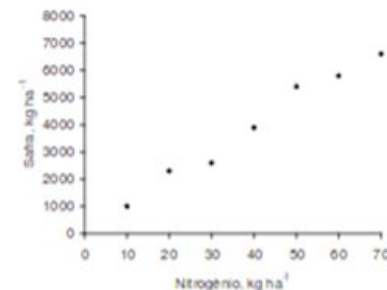


Figura 13 - Dados e reta ajustada a olho aos dados apresentados da Safra em função do Nitrogênio

O pesquisador fixa então sete valores de X (sete níveis do regressor), fazendo apenas uma observação Y (fator resposta), em cada caso, tal como se vê na figura acima.

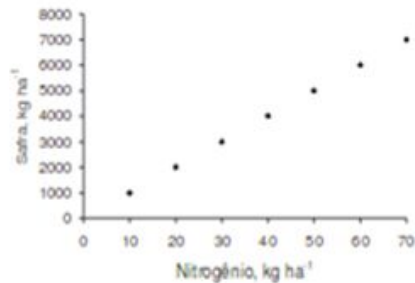
## Exemplo

COVARIÂNCIA

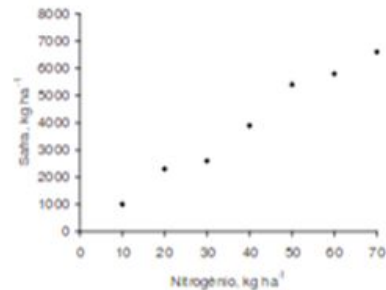
CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

40



a.



b.

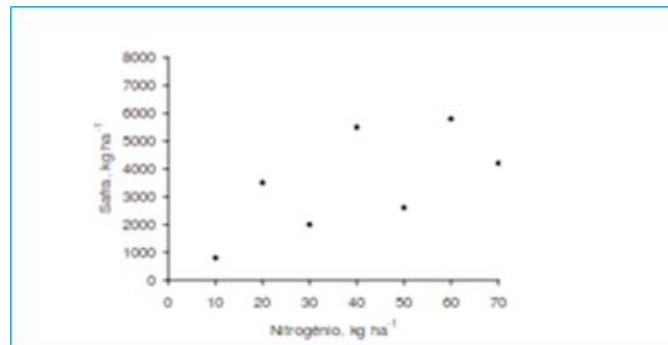


Ilustração de diversos graus de dispersão.



## Exemplo

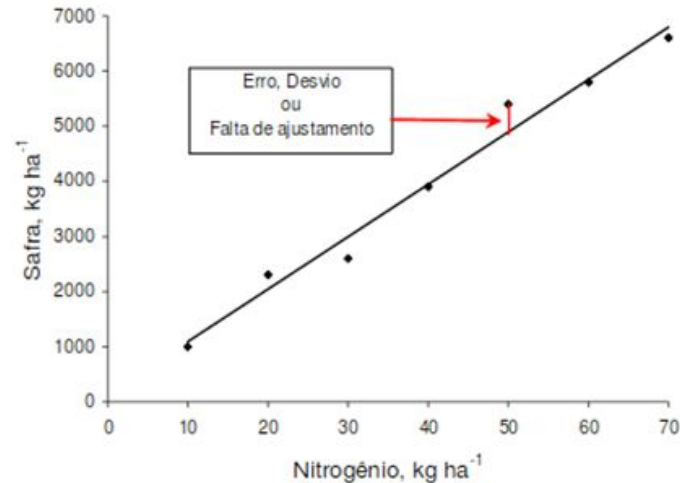
COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

41

Erro ou a falta de ajustamento é definido como a distância vertical entre o valor observado (real)  $Y_i$  e o valor ajustado (predito)  $\hat{Y}_i$  na reta, isto é,  $(Y_i - \hat{Y}_i)$ :



Erro típico no ajustamento de uma reta.

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

42

O método mais comumente utilizado para se ajustar uma reta aos pontos dispersos é o que minimiza a soma de quadrados dos erros:

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

Conhecido como critério dos “mínimos quadrados” ou “mínimos quadrados dos erros”. Sua justificativa inclui as seguintes observações:

- O quadrado elimina o problema do sinal, pois torna positivos todos os erros.
- A álgebra dos mínimos quadrados é de manejo relativamente fácil.

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

43

Passo extra que facilita o cálculo futuro:

- Ajustando uma reta:
  - Estágio 1: Expressar  $X$  em termos de desvios a contar de sua média, isto é, definir uma nova variável  $x$  (minúsculo), tal que:

$$x = X - \bar{X}$$

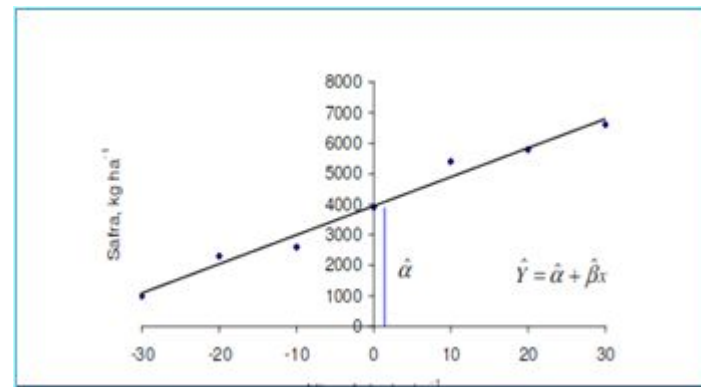
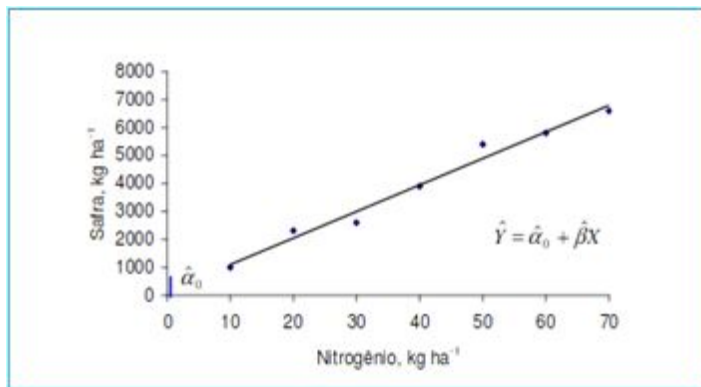
## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

44



Observa-se que o eixo Y foi deslocado para a direita, de 0 a  $\bar{X}$ .

O novo valor x torna-se positivo, ou negativo, conforme X esteja a direita ou à esquerda de  $\bar{X}$ . Não há modificação nos valores de Y. O intercepto  $\alpha$  difere do intercepto original,  $\alpha_0$ , mas o coeficiente angular, permanece o mesmo.

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

45

Medir  $X$  como desvio a contar de  $\bar{X}$  simplifica os cálculos porque a soma dos novos valores  $x$  é igual a zero, isto é:

$$\sum x_i = 0 \quad \therefore \quad \sum x_i = \sum (X_i - \bar{X}) = \sum X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0$$

## Exemplo

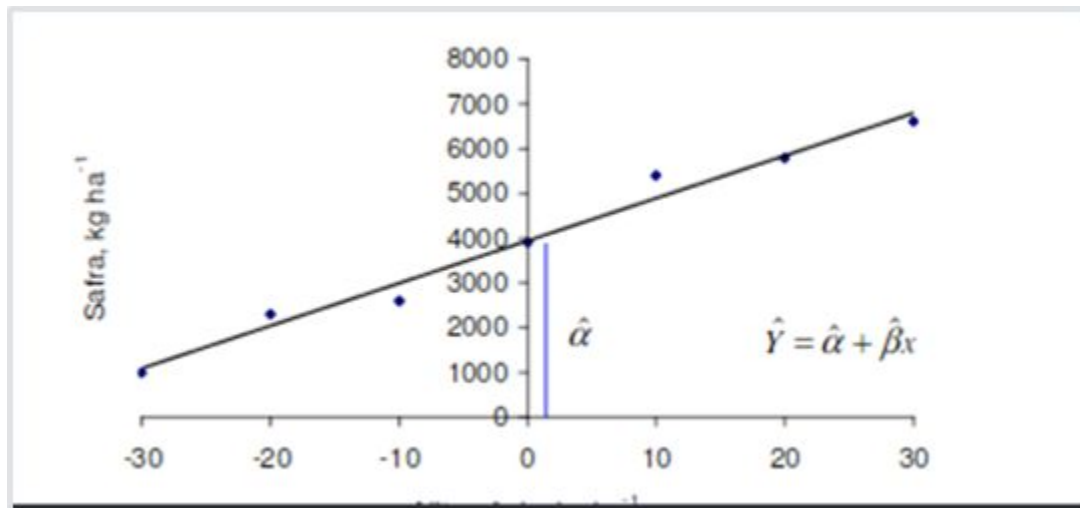
COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

46

## Estágio 2: Ajustar a reta:



## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

47

Devemos ajustar a reta aos dados, escolhendo valores para  $\alpha$  e  $\beta$ , que satisfaçam o critério dos mínimos quadrados. Ou seja, escolher valores de  $\alpha$  e  $\beta$  que minimizem

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

Equação 01

Cada valor ajustado  $\hat{y}_i$  estará sobre a reta estimada:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

Equação 02

Assim, estamos diante da seguinte situação: devemos encontrar os valores  $\alpha$  e  $\beta$  de modo a minimizar a soma de quadrados dos erros.

## Exemplo

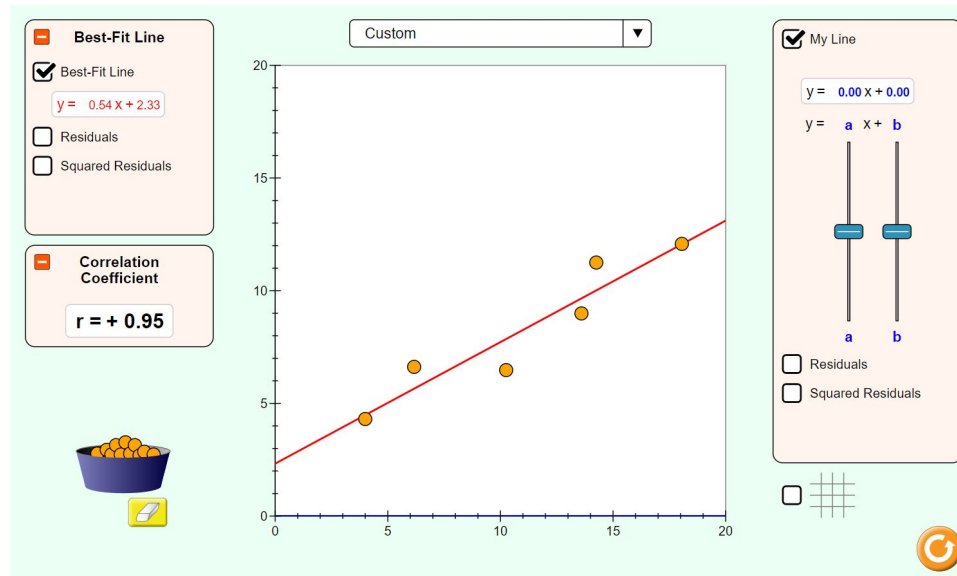
COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

48

Visualização gráfica:



[https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression\\_en.html](https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html)



## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

49

Considerando as Equações 01 e 02, isto pode ser expresso algebricamente como:

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

Equação 01

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

Equação 02

$$s(\hat{\alpha}, \hat{\beta}) = \sum (y_i - (\hat{\alpha} + \hat{\beta}x_i))^2 = \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

50

Uma possível técnica é fornecida pelo cálculo. A minimização de  $S(\alpha, \beta)$  exige a anulação simultânea de suas derivadas parciais:

Igualando a zero a derivada parcial em relação a  $\alpha$ :

$$\frac{\partial}{\partial \hat{\alpha}} \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \sum 2(-1)(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$$

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

51

Dividindo ambos os termos por  $(-2)$  e reagrupando:

$$\sum Y_i - n\hat{\alpha} - \hat{\beta}\sum x_i = 0 \quad \therefore \quad \sum x_i = 0$$

$$\sum Y_i - n\hat{\alpha} - 0 = 0$$

$$\sum Y_i - n\hat{\alpha} = 0$$

$$n\hat{\alpha} = \sum Y_i$$

$$\hat{\alpha} = \frac{\sum Y_i}{n} = \bar{Y}$$

Assim, a estimativa de mínimos quadrados para  $\hat{\alpha}$  é simplesmente o valor médio de  $Y$ .

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

52

É preciso também anular a derivada parcial em relação a  $\beta$ :

$$\frac{\partial}{\partial \beta} \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \sum 2(-x_i)(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$$

Dividindo ambos os termos por (-2):

$$\sum x_i (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$$

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

53

Reagrupando:

$$\sum x_i Y_i - \hat{\alpha} \sum x_i - \hat{\beta} \sum x_i^2 = 0 \quad \therefore \quad \sum x_i = 0$$

$$\sum x_i Y_i - 0 - \hat{\beta} \sum x_i^2 = 0$$

$$\sum x_i Y_i - \hat{\beta} \sum x_i^2 = 0$$

$$\hat{\beta} \sum x_i^2 = \sum x_i Y_i$$

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}$$

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

54

Para os dados da Figura 13 (Dados e reta ajustada a olho aos dados apresentados da Safra em função do Nitrogênio),  $\alpha$  e  $\beta$  acham-se calculados no Quadro 14.1.

Quadro 14.1 - Cálculos dos valores necessários

X	$x = X - \bar{X}$ $x = X - 40$	Y	xY	$x^2$
10	-30	1.000	-30.000	900
20	-20	2.300	-46.000	400
30	-10	2.600	-26.000	100
40	0	3.900	0	0
50	10	5.400	54.000	100
60	20	5.800	116.000	400
70	30	6.600	198.000	900
$\sum X = 280$		$\sum Y = 27.600$		
$\bar{X} = \frac{1}{N} \sum X$		$\bar{Y} = \frac{1}{N} \sum Y$		
$\bar{X} = \frac{280}{7} = 40$		$\bar{Y} = \frac{27.600}{7}$		
		$\bar{Y} = 3.942,86$		
$\sum x = 0$		$\sum xY = 266.000$		$\sum x^2 = 2.800$

$$\hat{\alpha} = \frac{\sum Y_i}{n} = \bar{Y} \therefore \hat{\alpha} = \frac{27.600}{7} = 3.942,86$$

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} \therefore \hat{\beta} = \frac{266.000}{2.800} = 95,00$$

$$\hat{Y} = 3.942,86 + 95x$$

Equação 03

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

55

**Estágio 3:** A regressão pode agora ser transformada para o sistema original de referência:

$\hat{Y} = 3.942,86 + 95x$	$\therefore x = (X - \bar{X})$
$\hat{Y} = 3.942,86 + 95(X - \bar{X})$	
$\hat{Y} = 3.942,86 + 95(X - 40)$	
$\hat{Y} = 3.942,86 + 95X - 3.800$	
$\hat{Y} = 142,86 + 95X$	Equação 04
$\hat{Y} = 3.942,86 + 95x$	Equação 03

Comparando as Equações 03 e 04, observa-se que:

- O coeficiente angular da reta de regressão ajustada ( $\beta = 95X$ ) permanece inalterado.
- A única diferença é o intercepto,  $\alpha$ , onde a reta tangencia o eixo Y.
- O intercepto original foi facilmente obtido.

## Exemplo

COVARIÂNCIA

CORRELAÇÃO L. SIMPLES

REGRESSÃO L. SIMPLES

56

Outra alternativa ao cálculo da regressão linear simples:

Cost Function:

$$\rightarrow J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

repeat until convergence {  
     $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$   
}



# Apresentação de Scripts

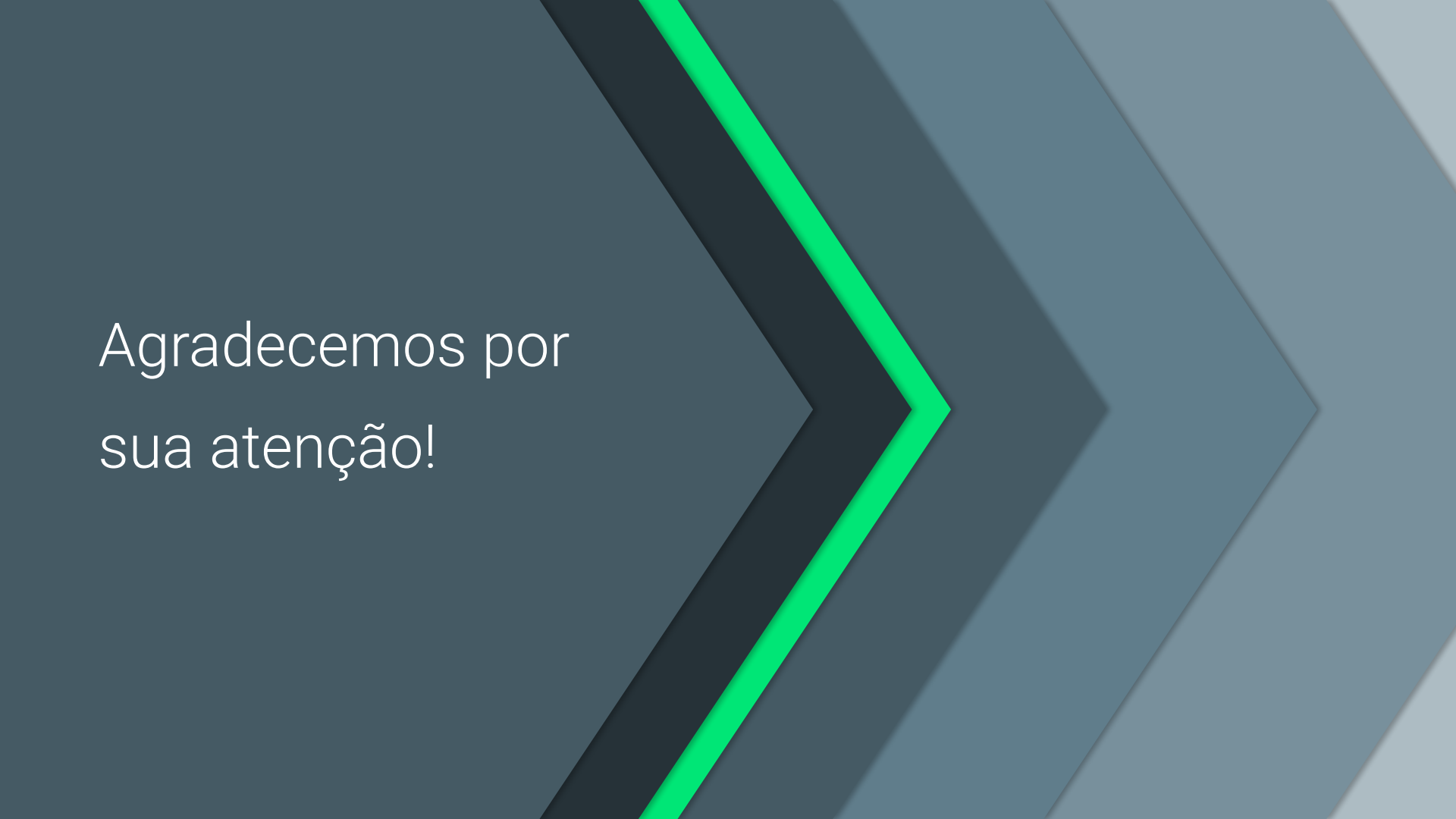
The background of the slide features a series of overlapping chevron shapes pointing to the right. The chevrons are in various shades of blue, from a dark navy blue to a light sky blue. A single, bright cyan chevron is superimposed over the others, creating a strong visual focal point.

## Referências

58

**FARIA, José Cláudio. Notas de aulas expandidas – Ilhéus, UESC/DCET, 10 ed. 2009.**

**LARSON, Ron; FARBER, Betsy. Estatística Aplicada 4ª Edição – São Paulo.**



Agradecemos por  
sua atenção!