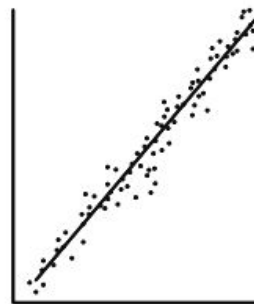




# Medidas de associação com introdução à regressão linear

Ana Luiza Oliveira e João Vitor Guimarães

2024.2



# Adaptação dos Slides de

**2021.1**

Murilo Torres

**2021.2**

Igor Rocha, Isaac Lima, João  
Rupp, ValcÍrio Francisco

**2022.2**

Gustavo Aragão, Henrique Daniel,  
Lucas Pereira Céu

.....

# Conteúdo

## 1 Associação

- 1.1 Introdução e Definição
- 1.2 Variáveis Aleatórias
- 1.3 Valor Esperado

## 2 Covariância

- 2.1 Introdução
- 2.2 Definição
- 2.3 Desdobramento
- 2.4 Observações
- 2.5 Covariância positiva/negativa
- 2.6 Covariância amostral
- 2.7 Exemplo

## 3 Correlação Linear Simples

- 3.1 Introdução
- 3.2 Pearson e Spearman
- 3.3 Relação Monotônica
- 3.4 Interpretação da fórmula
- 3.5 Diagrama de dispersão
- 3.6 Exemplos

## 4 Regressão Linear Simples

- 2.1 Introdução
- 2.2 Regressão x Correlação
- 2.3 Gráficos de análise
- 2.4 Quando usar análise de regressão
- 2.5 Análise de regressão
- 2.6 Análise de variância e regressão



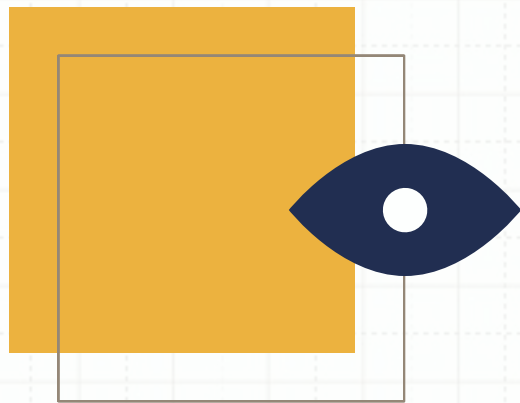
# 01

# Associação

Introdução e Definição

Variáveis Aleatórias

Valor Esperado





# Associação

Associação é o estudo da **variabilidade conjunta** entre duas ou mais **variáveis aleatórias**.

Este estudo é útil nas seguintes aplicações:

- Estudar uma variável através da outra;
- Prever os valores de uma através da outra;

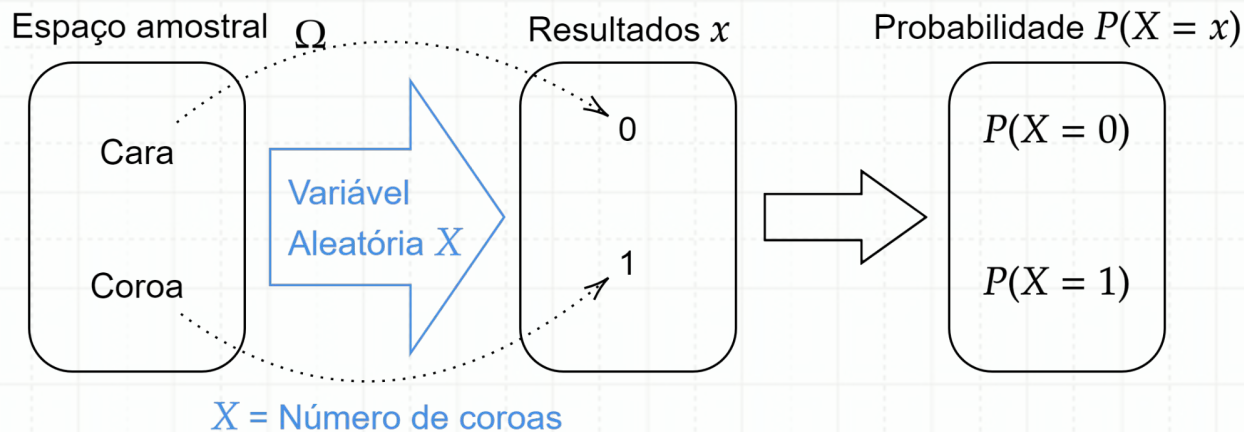
As medidas de associação mais comuns são:

- Covariância;
- Correlação linear simples.



# Variáveis Aleatórias

Uma variável aleatória é uma função que mapeia os resultados de um processo aleatório em valores numéricos.





# Variáveis Aleatórias

- Variáveis aleatórias discretas são aquelas que os resultados são frutos de contagem.

**Exemplo:** Número de filhos, número de sucessos em  $n$  tentativas,...

- Variáveis aleatórias contínuas são aquelas que os resultados são frutos de mensuração.

**Exemplo:** massa (kg) do objeto , altura (m) do indivíduo, temperatura...





# Valor Esperado

O **Valor Esperado de uma Variável aleatória**, é o que se espera de uma variável aleatória em média a longo prazo. O valor esperado também pode ser chamado muitas vezes de média ou expectância ou ainda de esperança matemática de uma variável aleatória.

## Variáveis aleatórias discretas

$$E[X] = \mu = \sum_x x \cdot p_X(x)$$

## Variáveis aleatórias contínuas

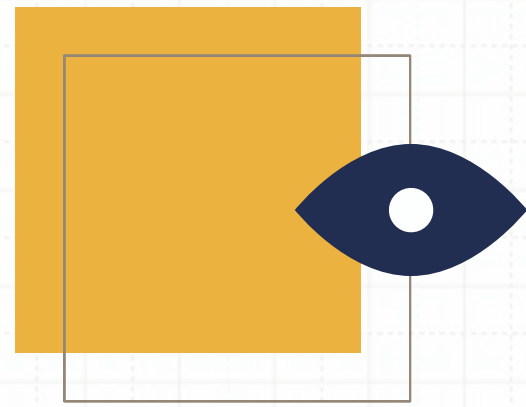
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$



# 02

## Covariância

- Introdução
- Definição
- Desdobramentos
- Observações
- Covariância positiva/negativa
- Covariância amostral
- Exemplo





# Introdução de Covariância

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  $X$  correspondem aos menores valores de  $Y$ , e vice-versa, temos que a covariância é negativa.

# Definição de Covariância

A covariância dá uma ideia da dispersão dos valores da variável bidimensional  $(X,Y)$  em relação ao ponto  $(E(X),E(Y))$ .

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

**Reescrevendo**  $= E[XY - E(Y)X - E(X)Y + E(X)E(Y)]$   
 $= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y)$

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

$$E(E(x)) = E(x)$$

# Variância e Covariância

A variância de uma variável  $X$  mede o grau de dispersão de  $X$  em torno da sua média  $E[X]$ .

Podemos ver a variância de  $X$  como a covariância de  $X$  consigo mesma. Ou seja:

$$\text{Var}(X) = \text{Cov}(X, X) = E[(X - E[X])^2]$$

# Definição de Covariância

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.



Nearly Zero  
Covariance

# Desdobramentos da Covariância

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  **$\rho(x,y)$** .

# Desdobramentos da Covariância

Quando porém as variáveis trabalhadas forem **contínuas**, ou seja, quando entre dois valores ( $x_1$  e  $x_2$  |  $y_1$  e  $y_2$ ) 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:

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



# Observações sobre Covariância

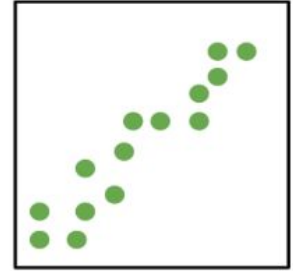
Se  $X$  e  $Y$  são variáveis aleatórias independentes, a  $\text{Cov}(X,Y)$  será igual a 0.

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

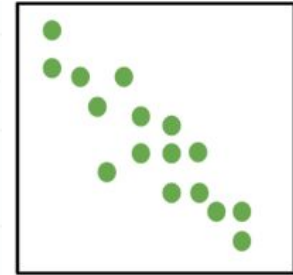
# Covariância Positiva e Negativa

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

Já a covariância negativa ocorre quando uma variável aumenta e a outra diminui.



Positive  
Covariance



Negative  
Covariance

# Covariância Populacional

A covariância populacional entre duas variáveis aleatórias  $X$  e  $Y$  mede a relação linear entre essas duas variáveis na população inteira.

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y)$$

# Covariância Amostral

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

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

\*A distribuição de probabilidade é o processo que descreve o comportamento aleatório de fenômenos.

# Exemplo de Covariância

Consideremos duas variáveis aleatórias:

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

## Rendimento acadêmico:

$$\begin{aligned}\sum M &= 480 \\ m(M) &= 60\end{aligned}$$

$$\begin{aligned}\sum L &= 400 \\ m(L) &= 50\end{aligned}$$

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 de Covariância

Obs	M	L	$m = (M_i - m(M))$	$l = [L_i - m(L)]$	$m * l$
1	36	35	-24	-15	360
2	80	65	20	15	300
3	50	60	-10	10	-100
4	58	39	-2	-11	22
5	72	48	12	-2	-24
6	60	44	0	-6	0
7	56	48	-4	-2	8
8	68	61	8	11	88
$\Sigma$	480	400	40-40=0	36-36=0	654

# Exemplo de Covariância

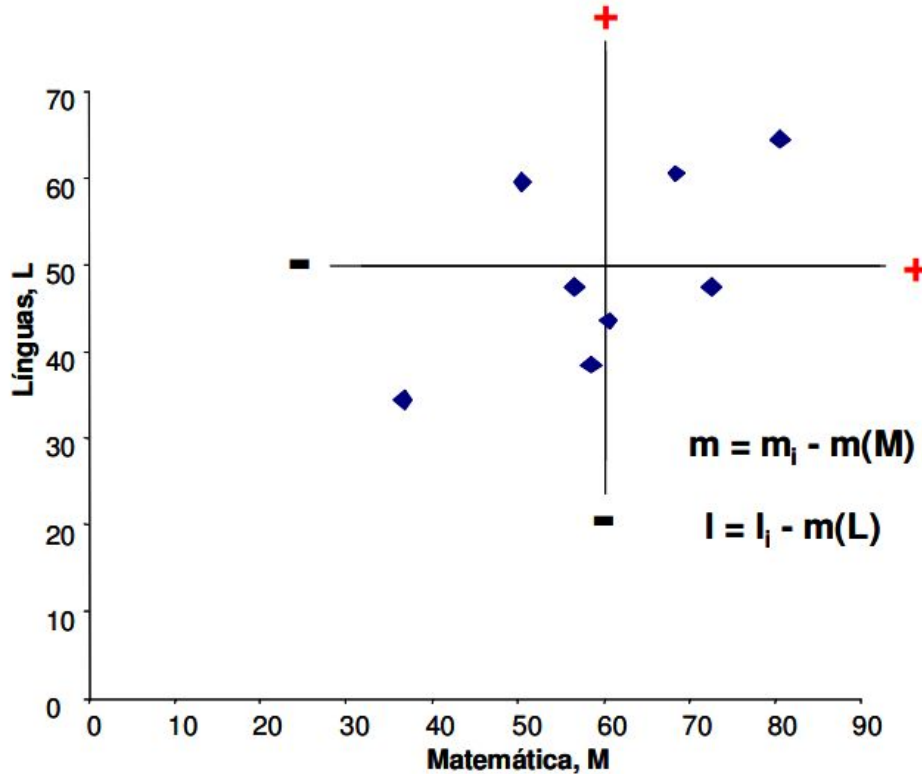
$$\text{cov}(M, L) = \frac{\sum (M_i - m(M)) \times (L_i - m(L))}{n - 1}$$

$$\text{COV}_{M,L} = \frac{654}{7}$$

$$\text{COV}_{M,L} = 93.4285$$



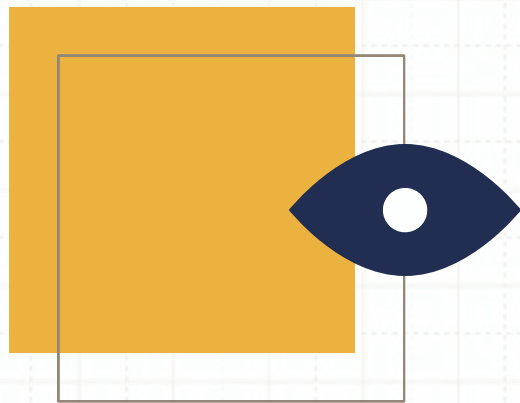
# Exemplo de Covariância



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 ( $\sum m.l$ ), demonstrando um **relacionamento positivo entre M e L.**

# 03



## Correlação Linear Simples

- Introdução
- Pearson e Spearman
- Relação monotônica
- Interpretação da fórmula
- Diagrama de dispersão
- Exemplos práticos





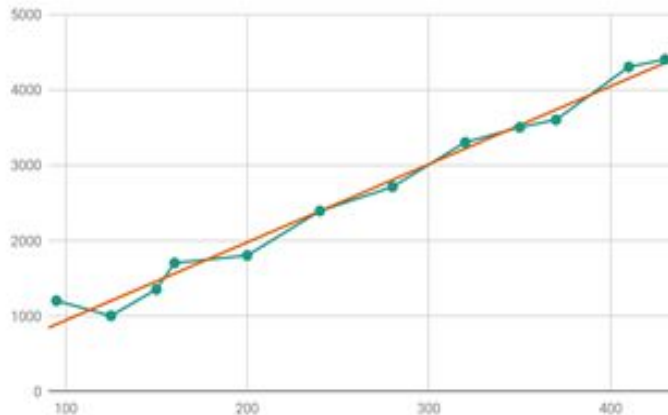
# Introdução

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.

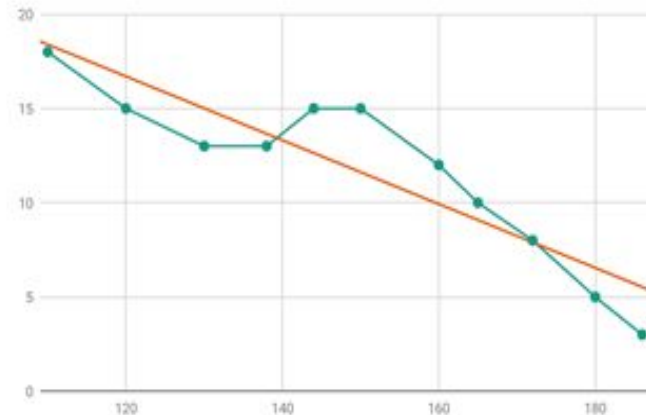
# Diagrama de dispersão

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  
 $r = 0,984$

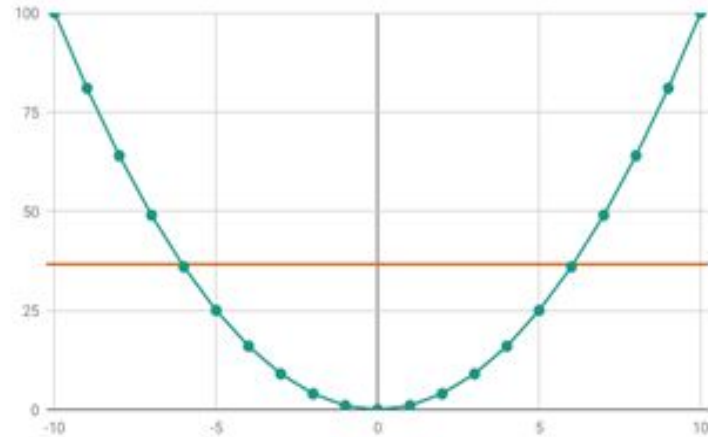


Correlação **negativa** e forte  
 $r = -0,819$

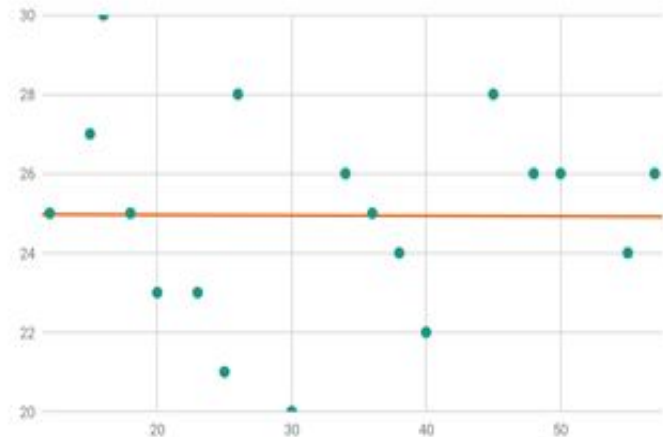




**Correlação nula**  
 $r = 0$



**Correlação fraca, quase nula**  
 $r = 0,0068$





# Coeficiente de correlação

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

# Definição

**$\rho$**  : Correlação populacional

$$\rho = \frac{COV_{Pop}(Y_1, Y_2)}{\sigma(Y_1) \cdot \sigma(Y_2)}$$

**$r$**  : Estimativa da correlação ou correlação amostral

$$r = \frac{cov_{Amo}(Y_1, Y_2)}{s(Y_1) \cdot s(Y_2)}$$





# Pearson e Spearman

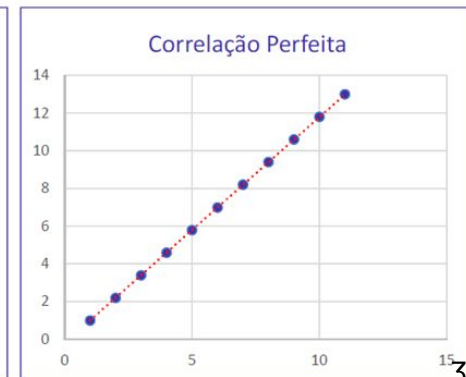
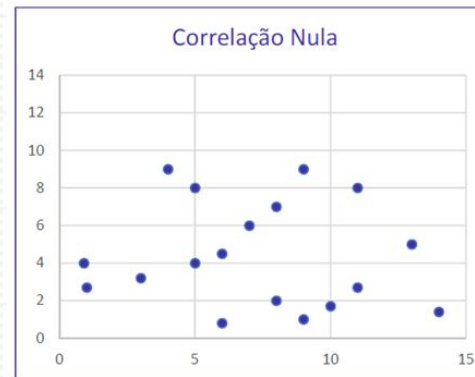
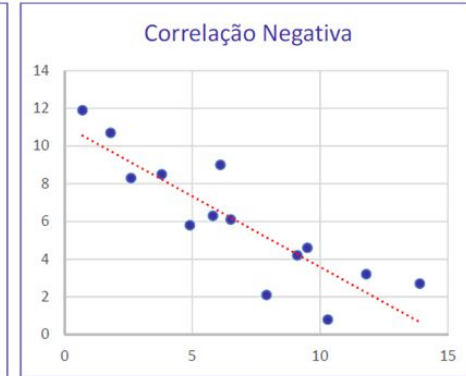
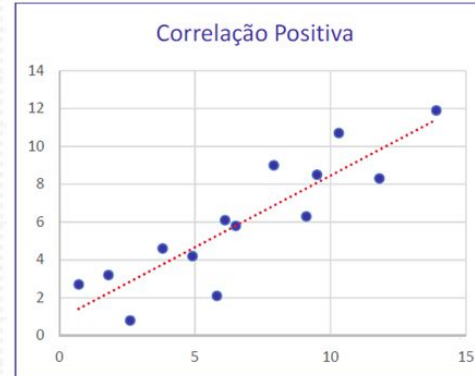
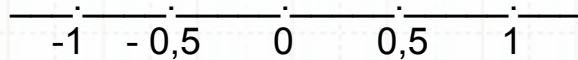
A correlação de Pearson avalia a **relação linear entre duas variáveis quantitativas**.

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.

# Interpretando resultados

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

forte    fraca    fraca    forte

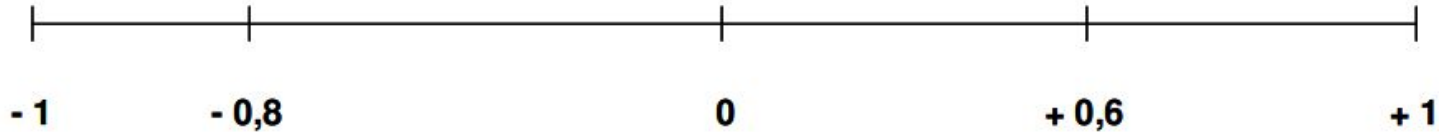
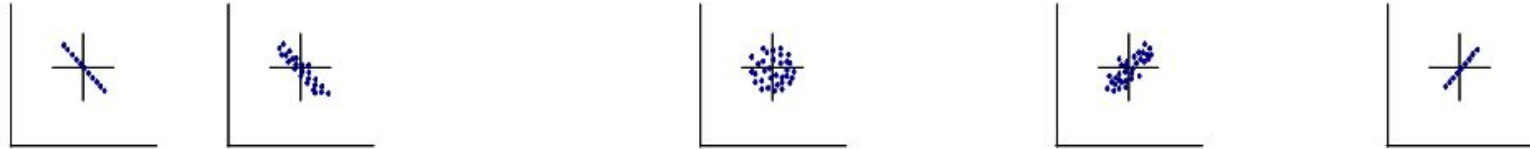


# Interpretando resultados

**Perfeita negativa**

**Não correlacionadas**

**Perfeita positiva**



**Aumenta grau de correlação  
negativa**

**Aumenta grau de correlação  
positiva**

# Utilizando a Fórmula

	x	y	x <sup>2</sup>	y <sup>2</sup>	xy
	3	7	9	49	21
	2	5	4	25	10
	-1	-1	1	1	1
	4	9	16	81	36
<b>Soma</b>	8	20	30	158	68

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

Substituindo os valores na fórmula, temos:

$$r = 4.68 - 8.20 / \sqrt{4.30 - 8^2} \cdot \sqrt{4.156 - 20^2}$$
$$r = 1$$

## Exemplo prático: Mortalidade infantil x Taxa de alfabetismo

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	70,8	39,4	618,35	1061,48	413,04
(Somatório) <sup>2</sup>				1.126.740	170.602

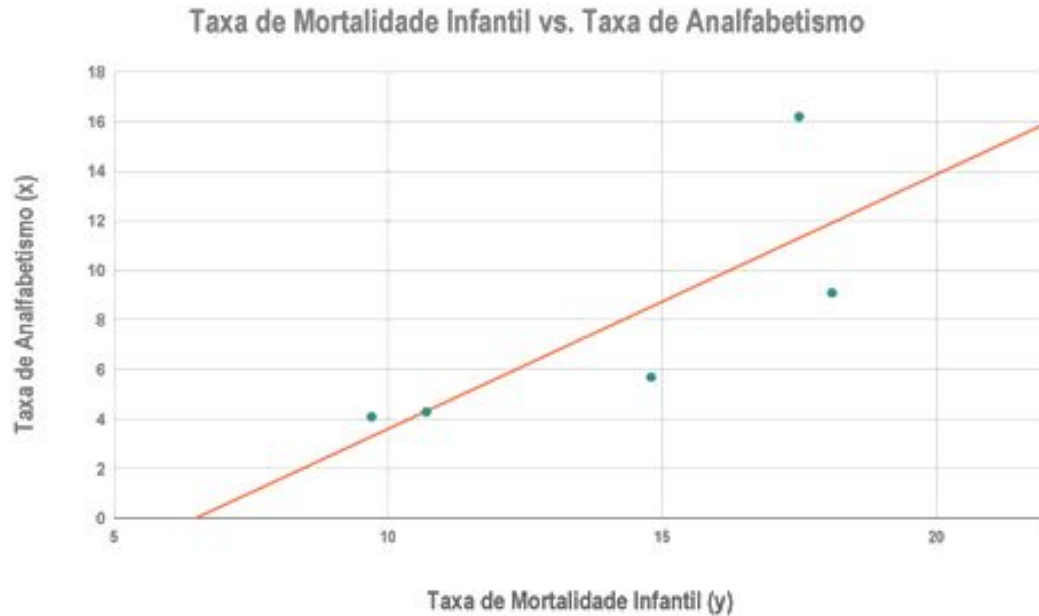
$$r = \frac{\sum x,y - \frac{(\sum x) \cdot (\sum y)}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] \cdot \left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}$$

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

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

$$r = 0.78$$

# Gráfico de dispersão



Coeficiente de  
Correlação

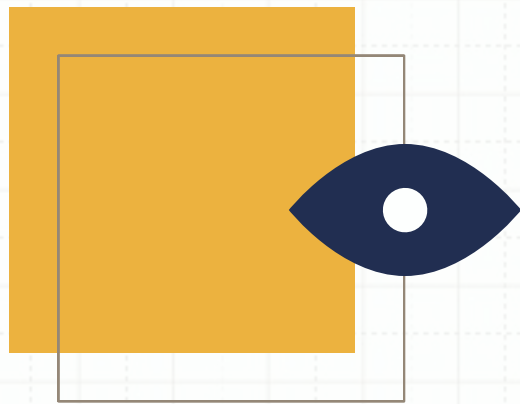
$0,5 \leq r < 0,8$

Correlação

Moderado  
Positiva

**$r = 0.78$**

# 04



## Regressão Linear Simples

- Introdução
- Regressão x Correlação
- Gráficos de análise
- Quando usar análise de regressão
- Análise de Regressão
- Análise de variância e regressão







# Introdução

A **regressão linear simples**, é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativas 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 quantitativas.



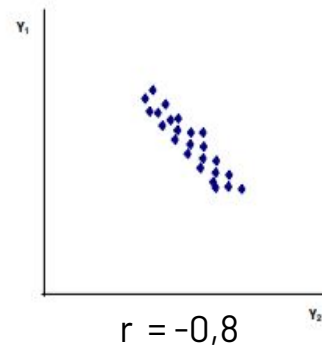
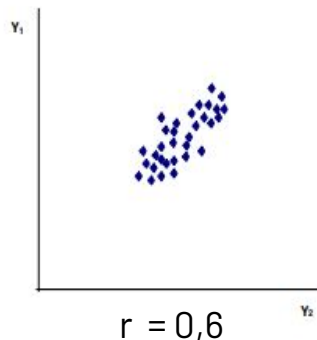
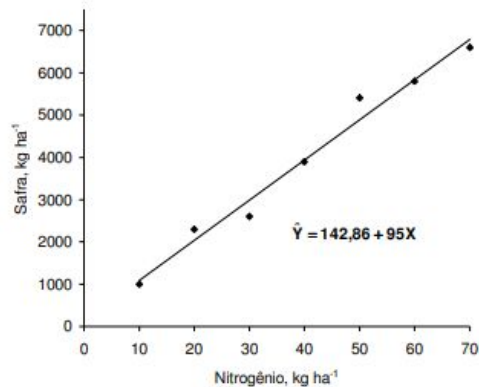
# Regressão x Correlação

Quando os tratamentos são níveis crescentes de pelo menos um fator quantitativo, os ensaios devem ser analisados por intermédio da análise quantitativa de experimentos, isto é, regressão, e ou, correlação.

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  $Y_1$  aumenta, o que acontece com uma outra variável aleatória  $Y_2$ : aumenta, diminui ou não altera?

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.

# Gráficos das Análises



Na análise de regressão estimamos toda uma função  $Y=f(X)$ , a equação de regressão.

A análise de correlação nos fornece apenas um número, um índice, que quantifica o grau da associação linear entre duas variáveis aleatórias.

Enquanto a análise de regressão linear nos mostra como as variáveis se relacionam linearmente, a análise de correlação vai nos mostrar apenas o grau desse mesmo relacionamento.



# Quando usar análise de Regressão?

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 (obs: também pode ser utilizada para estabelecer a relação funcional entre duas ou mais variáveis aleatórias).

**Exemplo:** Vamos considerar que conduzimos um experimento submetendo plantas de milho a doses crescentes de nitrogênio (fertilizante).



# Quando usar análise de Regressão?

Naturalmente, a produção será dependente da quantidade aplicada desse fertilizante,  $X$ .

Cada variável aleatória mensurada na cultura do milho, sujeita a influência dos níveis  $x_i$ , ou seja, das doses de nitrogênio, é chamada de **"variável dependente" ou "fator resposta"**. (o número de espigas por planta ( $Y_1$ ), a altura média das plantas ( $Y_2$ ), o peso de 1.000 grãos ( $Y_3$ ) etc.).

Como a aplicação do fertilizante não depende da safra, chamamos a mesma de **"variável independente" ou "regressor"**.



# Quando usar análise de Regressão?

Seria **incorreto** estudar via análise de correlação o efeito do nitrogênio (variável fixa) sobre a produção de matéria seca dos grãos de milho (variável aleatória), ou sobre outras variáveis aleatórias.

Neste caso, podemos estudar, via análise de regressão, o efeito da variável fixa, independente,  $X$  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  $X$ .**

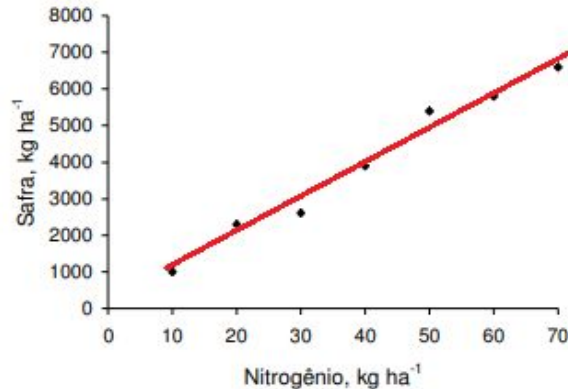


# Análise de Regressão

Vamos continuar considerando um estudo sobre a influência do N (nitrogênio) aplicado em cobertura sobre a safra do milho.

Temos uma amostra de sete valores de X (sete níveis do regressor), fazendo apenas uma observação Y (fator resposta), em cada caso.

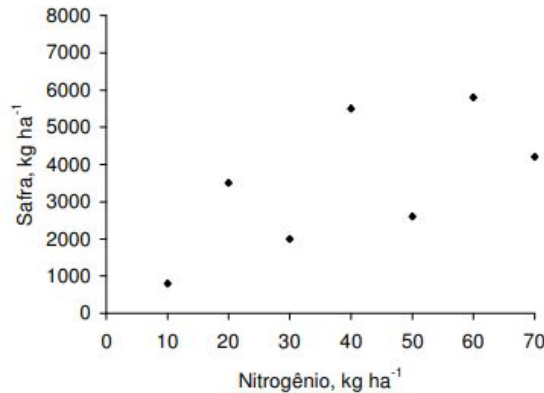
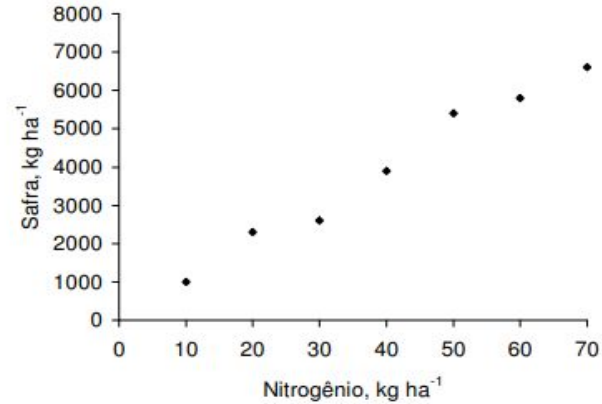
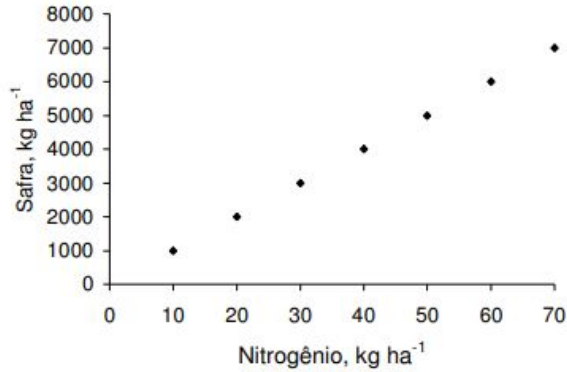
X Nitrogênio kg ha <sup>-1</sup>	Y Safra kg ha <sup>-1</sup>
10	1.000
20	2.300
30	2.600
40	3.900
50	5.400
60	5.800
70	6.600



**Até onde é bom um ajustamento da reta feito a olho?**



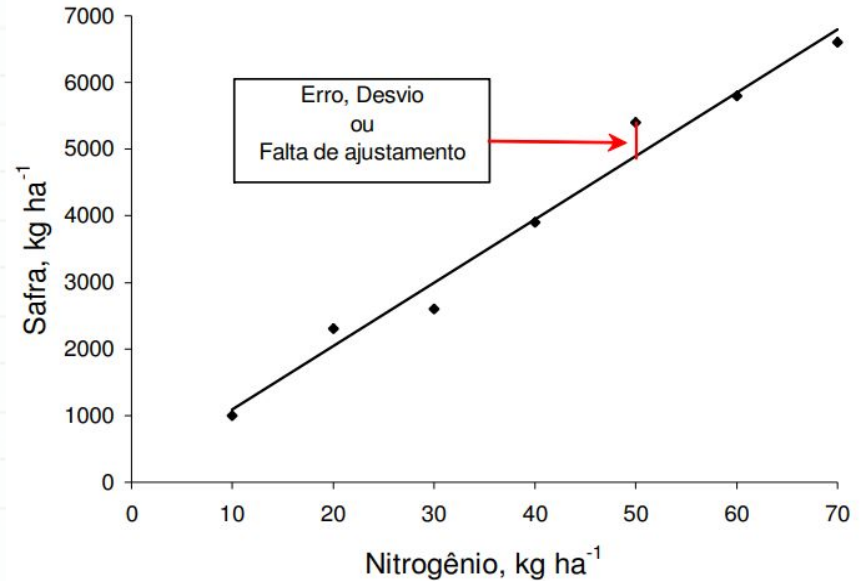
# Até onde é bom um ajustamento da reta feito a olho?



# Análise de Regressão

Precisamos então de um método objetivo, que faça o melhor ajustamento da reta, ou seja, minimizar o erro total.

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,  $(Y_i - \hat{Y}_i)$ :





# Análise de Regressão

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.



# Análise de Regressão

Passo extra que facilita o cálculo futuro:

- Ajustando uma reta:

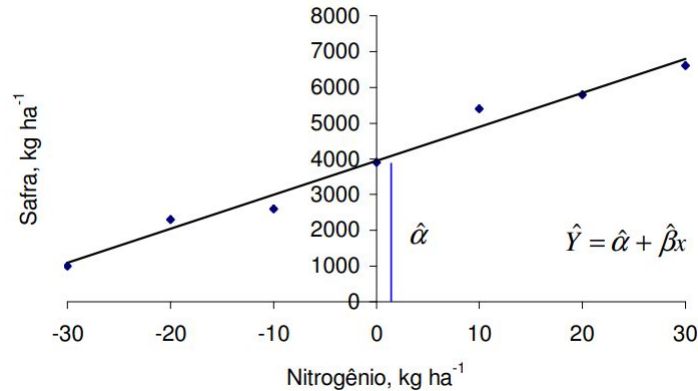
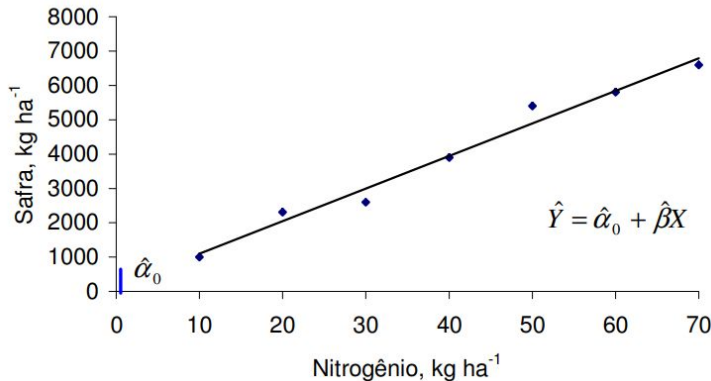
**Estágio 1:** Expressar  $X$  em termos de desvios a contar de sua média, isto é, definir uma nova variável  $x$  (minúsculo), tal que:

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

# Análise de Regressão

## Estágio 1

Analisando os gráficos abaixo, 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 à direita ou à esquerda do  $\bar{X}$  médio. Não há modificação nos valores de  $Y$ . O intercepto  $\alpha$  difere do intercepto original,  $\alpha_0$ , mas o coeficiente angular permanece o mesmo.



# Análise de Regressão

## Estágio 1

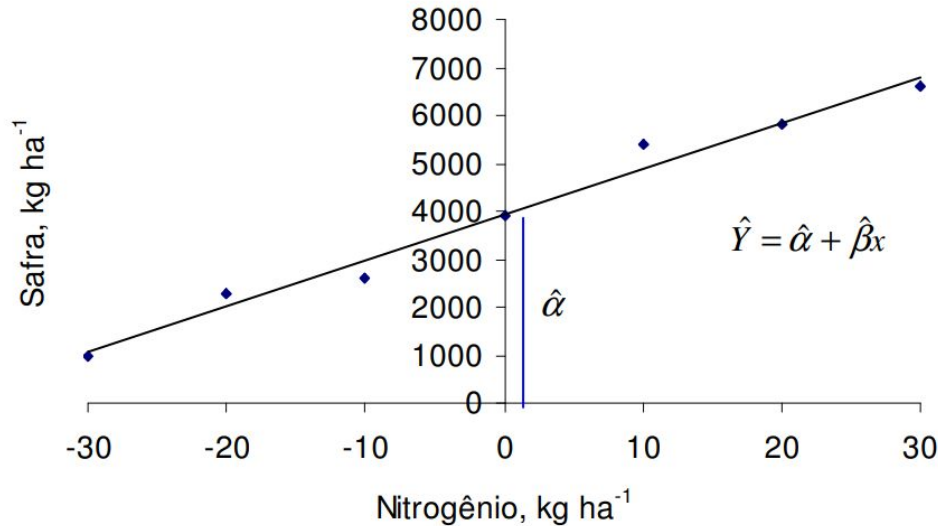
Medir  $x$  como desvio a contar de  $\bar{X}$  médio simplifica os cálculos porque a soma dos novos valores  $x$  é igual a zero, isto é:

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

# Análise de Regressão

## Estágio 2

Ajustar a reta:  $\hat{Y} = \hat{\alpha} + \hat{\beta}x$





# Análise de Regressão

## Estágio 2

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.

# Análise de Regressão

## Estágio 2

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

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



# Análise de Regressão

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  $\alpha$ :

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



# Análise de Regressão

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.



# Análise de Regressão

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):

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



# Análise de Regressão

Reagrupando:

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

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

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

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

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

# Análise de Regressão

## Estágio 2: Fórmulas

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

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

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

$\hat{\alpha}$  é nosso intercepto (onde a reta toca o eixo y);  
 $\hat{\beta}$  é nosso coeficiente angular.



# Análise de Regressão

## Estágio 2

Substituindo valores calculados na fórmula,  $\alpha$  e  $\beta$  acham-se calculados no Quadro 14.1.

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

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

$$\begin{aligned}\sum X &= 280 \\ \bar{X} &= \frac{1}{N} \sum X \\ \bar{X} &= \frac{280}{7} = 40\end{aligned}$$

$$\sum x = 0$$

$$\sum Y = 27.600$$

$$\bar{Y} = \frac{1}{N} \sum Y$$

$$\bar{Y} = \frac{27.600}{7}$$

$$\bar{Y} = 3.942,86$$

$$\sum xY = 266.000$$

$$\sum x^2 = 2.800$$

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

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

# Análise de Regressão

## Estágio 3

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

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.

$$\hat{Y} = 3.942,86 + 95x \quad \therefore \quad x = (X - \bar{X})$$

$$\hat{Y} = 3.942,86 + 95(X - \bar{X})$$

$$\hat{Y} = 3.942,86 + 95(X - 40)$$

$$\hat{Y} = 3.942,86 + 95X - 3.800$$

$$\text{Equação 04} \quad \hat{Y} = 142,86 + 95X$$

$$\text{Equação 03} \quad \hat{Y} = 3.942,86 + 95x$$

# Análise de Regressão

## ● Estágio 3

Essa equação é útil para estimar a safra em relação a quantidade de nitrogênio.

Se nenhum nitrogênio for aplicado à cultura, a safra estimada será de 142,86 kg. Esta safra se deve à absorção pela cultura do N disponível no solo, possivelmente associado ao ciclo Orgânico.

No intervalo das doses aplicadas (10 a 70 kg), considerando-se um hectare, para cada kg de nitrogênio aplicado, a cultura responde com 95 kg de grãos.

# Análise de Regressão

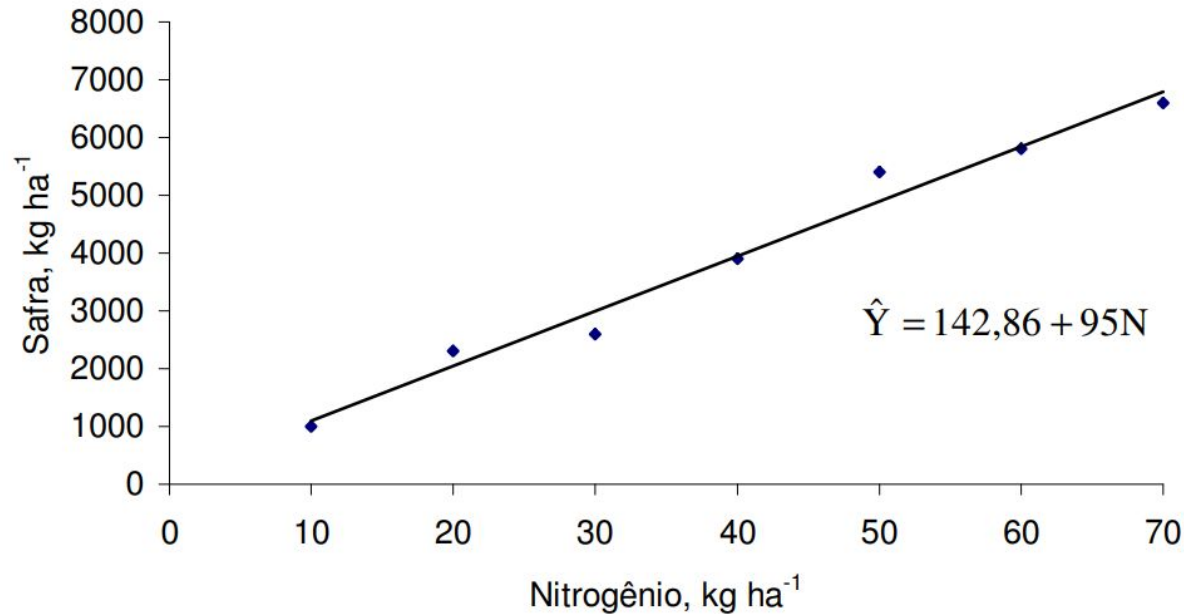


Figura X – Produção de milho em função da aplicação de doses de nitrogênio – 2023

# Confiabilidade da Regressão

Para se decidir quão bem o modelo ajustado é adequado à natureza dos dados experimentais, pode-se lançar mão da **análise de variância da regressão**.

# Análise de Variância da Regressão

Vamos particionar a variação total (**SQDtot**) da variável dependente em função das variações nos níveis da variável independente em duas partes:

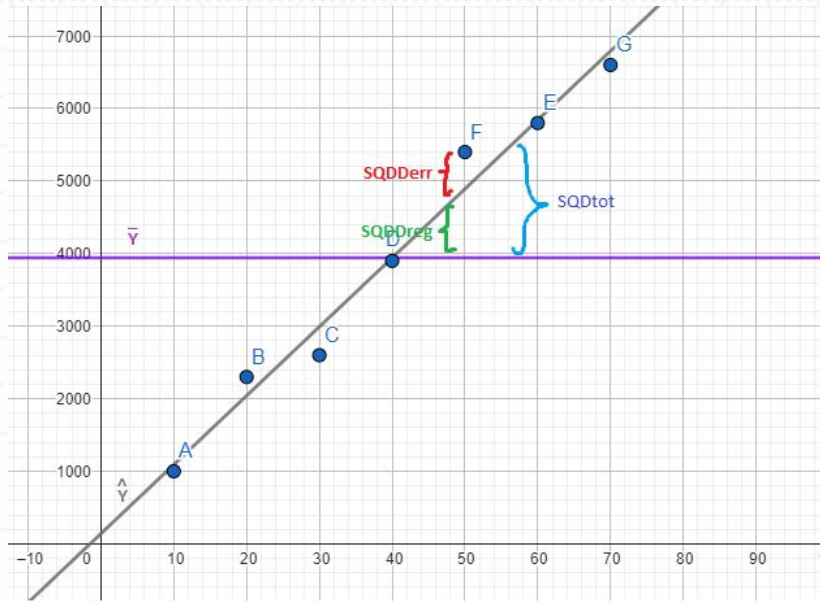
- Uma parte associada ao modelo ajustado (**SQDreg**): soma de quadrados dos desvios devido à regressão, que quantifica o quanto da variação total da safra é explicada pelo modelo ajustado
- Uma outra parte associada à falta de ajuste (**SQDerr**): soma de quadrados dos desvios devido ao erro, que quantifica o montante da variação total da safra que não é explicada pelo modelo ajustado.





# Análise de Regressão

Gráfico feito a partir dos valores encontrados em nosso exemplo:



$$\text{SQDtot} = \sum (y_i - \bar{y})^2 \quad \text{SQDDerr} = \sum (y_i - \hat{y}_i)^2 \quad \text{SQDDreg} = \sum (\hat{y}_i - \bar{y})^2$$



# Análise de Variância da Regressão

Sabemos que  $SQD_{tot} = SQD_{err} + SQD_{reg}$

Se considerarmos que nossa reta de regressão não teve erros, ou seja,  $SQD_{err} = 0$ , vamos obter  $SQD_{tot} = SQD_{reg}$ .

Fazendo:  $SQD_{reg}/SQD_{tot} = 1$

Significa que quanto mais próximo de 1, menos erros, e quanto mais próximo de 0, mais erros e esse é nosso **coeficiente de determinação da regressão**.

$$r^2 = \frac{SQD_{reg}}{SQD_{tot}} \therefore 0 \leq r^2 \leq 1$$

Obs: Se tirarmos a raiz obtemos o **coeficiente de correlação r**.

# Calculando SQDtot, SQDerr e SQDreg do exemplo:

**SQDtot**

Obs	$m_{(Obs)}$	$Obs - m_{(Obs)}$	$[Obs - m_{(Obs)}]^2$
1.000	3.942,86	-2.942,86	8.660.408,16
2.300	3.942,86	-1.642,86	2.698.979,59
2.600	3.942,86	-1.342,86	1.803.265,31
3.900	3.942,86	-42,86	1.836,73
5.400	3.942,86	1.457,14	2.123.265,31
5.800	3.942,86	1.857,14	3.448.979,59
6.600	3.942,86	2.657,14	7.060.408,16
			25.797.142,86

Note que se fizermos  
SQDreg + SQDerr obtemos **SQDtot**

**SQDreg**

Est	$m_{(Est)}$	$Est - m_{(Est)}$	$[Est - m_{(Est)}]^2$
1.093	3.942,86	-2.850,00	8.122.500,00
2.043	3.942,86	-1.900,00	3.610.000,00
2.993	3.942,86	-950,00	902.500,00
3.943	3.942,86	0,00	0,00
4.893	3.942,86	950,00	902.500,00
5.843	3.942,86	1.900,00	3.610.000,00
6.793	3.942,86	2.850,00	8.122.500,00
			25.270.000,00

**SQDerr**

Obs	Est	$Erro(Obs - Est)$	$m_{(Erro)}$	$Erro - m_{(Erro)}$	$[Erro - m_{(Erro)}]^2$
1.000	1.092,86	-92,86	0,00	-92,86	8.622,45
2.300	2.042,86	257,14	0,00	257,14	66.122,45
2.600	2.992,86	-392,86	0,00	-392,86	154.336,73
3.900	3.942,86	-42,86	0,00	-42,86	1.836,73
5.400	4.892,86	507,14	0,00	507,14	257.193,88
5.800	5.842,86	-42,86	0,00	-42,86	1.836,73
6.600	6.792,86	-192,86	0,00	-192,86	37.193,88
					527.142,86

# Análise de Variância da Regressão

**Calculando o coeficiente de determinação da regressão:**

$$r^2 = \frac{25.270.000,00}{25.797.142,86} = 0,9796 = 97,96\%$$

**Interpretação:**

97,96% da variação total da safra, em decorrência da variação da dose de nitrogênio, é explicada pelo modelo de regressão ajustado.

# Análise de Variância da Regressão

Cálculos alternativos da soma de quadrados dos desvios:

$$SQD_{tot} = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

$$SQD_{reg} = \hat{\alpha}_0 \sum Y_i + \hat{\beta} \sum X_i Y_i - \frac{(\sum Y_i)^2}{n}$$

$$SQD_{err} = SQD_{tot} - SQD_{reg}$$

Embora menos compreensível à primeira vista, é a mais prática e deve ser preferencialmente utilizada.

X	Y	Y <sup>2</sup>	XY
10	1.000	1.000.000	10.000
20	2.300	5.290.000	46.000
30	2.600	6.760.000	78.000
40	3.900	15.210.000	156.000
50	5.400	29.160.000	270.000
60	5.800	33.640.000	348.000
70	6.600	43.560.000	462.000
	27.600	134.620.000	1.370.000

$$SQD_{tot} = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = 134.620.000 - \frac{(27.600)^2}{7} = 25.797.142,86$$

$$SQD_{reg} = \hat{\alpha}_0 \sum Y_i + \hat{\beta} \sum X_i Y_i - \frac{(\sum Y_i)^2}{n}$$

$$SQD_{reg} = 142,85714286 \times 27.600 + 95 \times 1.370.000 - \frac{(27.600)^2}{7}$$

$$SQD_{reg} = 25.270.000$$

$$SQD_{err} = SQD_{tot} - SQD_{reg}$$

$$SQD_{err} = 25.797.142,86 - 25.270.000$$

$$SQD_{err} = 527.142,86$$

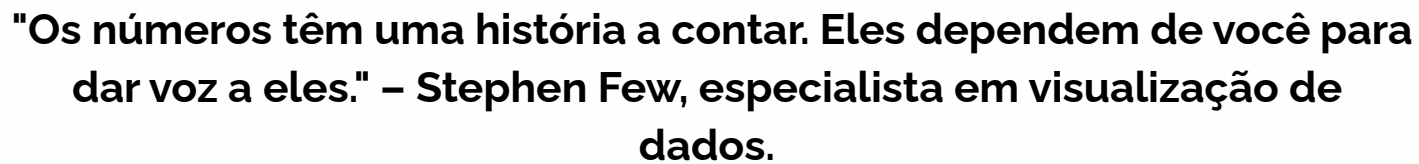


# Referências

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

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







Obrigado!

