INSTITUTO FEDERAL
Rio Grande do Sul

# Probabilidade e Estatística

## Correlação e regressão

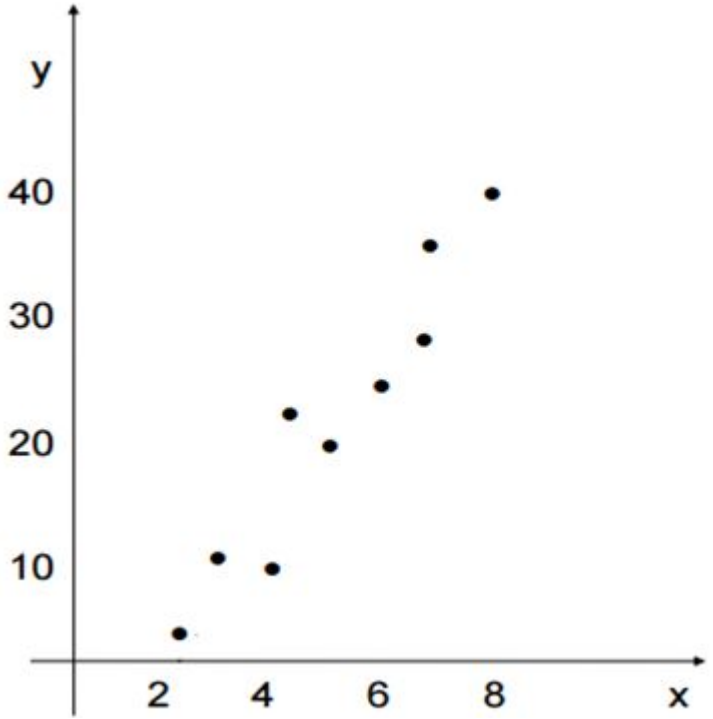
Prof. Ruana Maíra Schneider
Ruana.Schneider@farroupilha.ifrs.edu.br

# Correlação

- Estuda o comportamento de duas variáveis quantitativas distintas.
- “ Mede” o grau de associação entre duas variáveis observadas.
- Representamos cada variável em um eixo

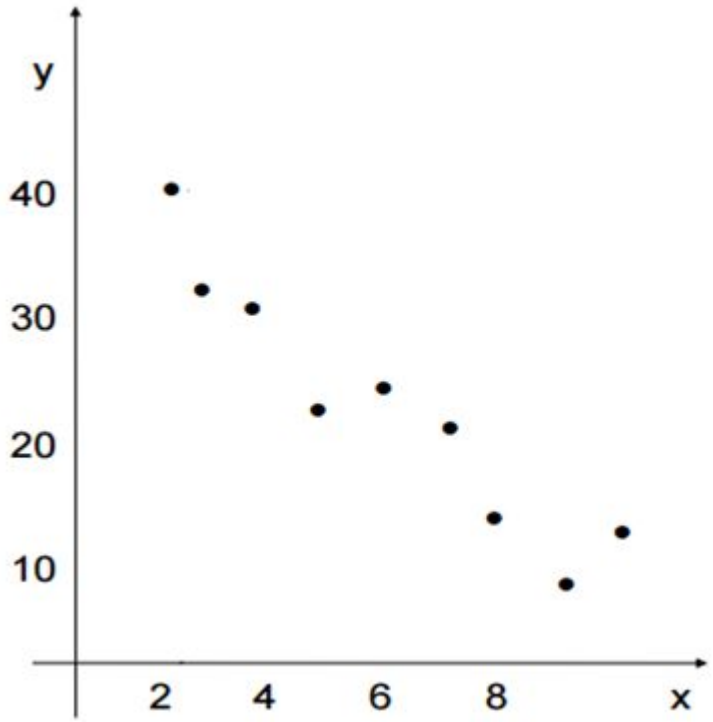
# Três situações possíveis

1) Uma variável cresce e a outra também cresce: correlação positiva

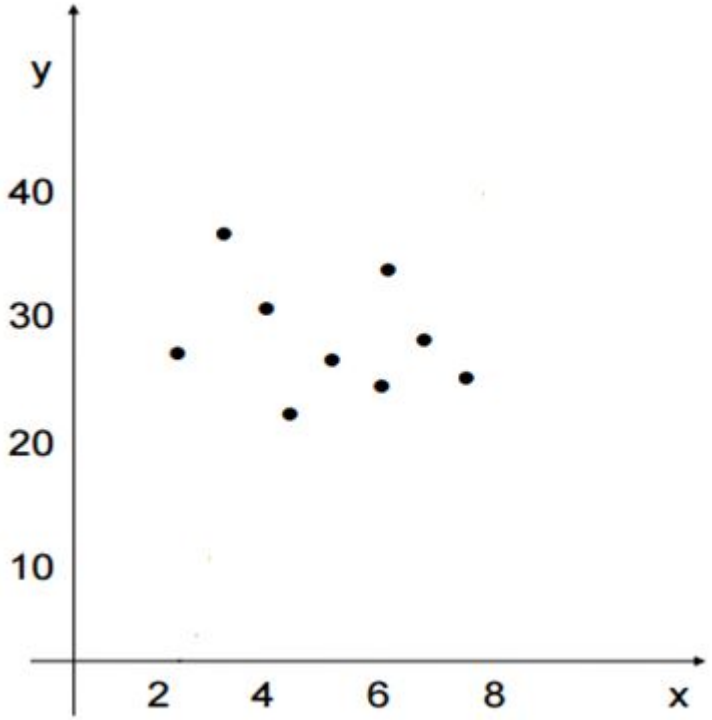



INSTITUTO FEDERAL
Rio Grande do Sul

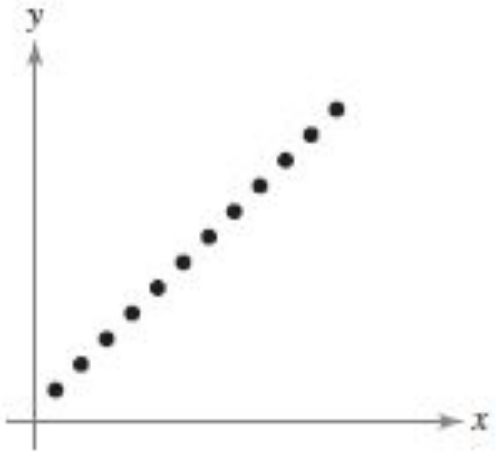
# 2) Uma variável cresce e a outra decresce: correlação negativa

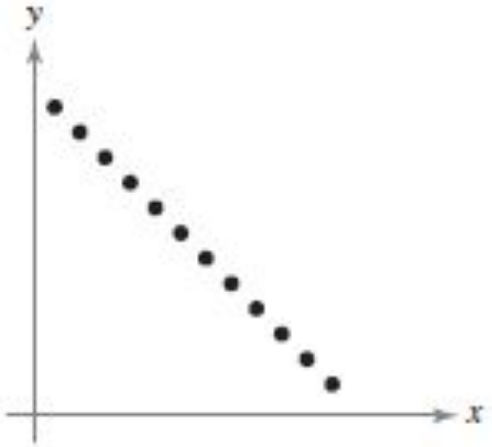


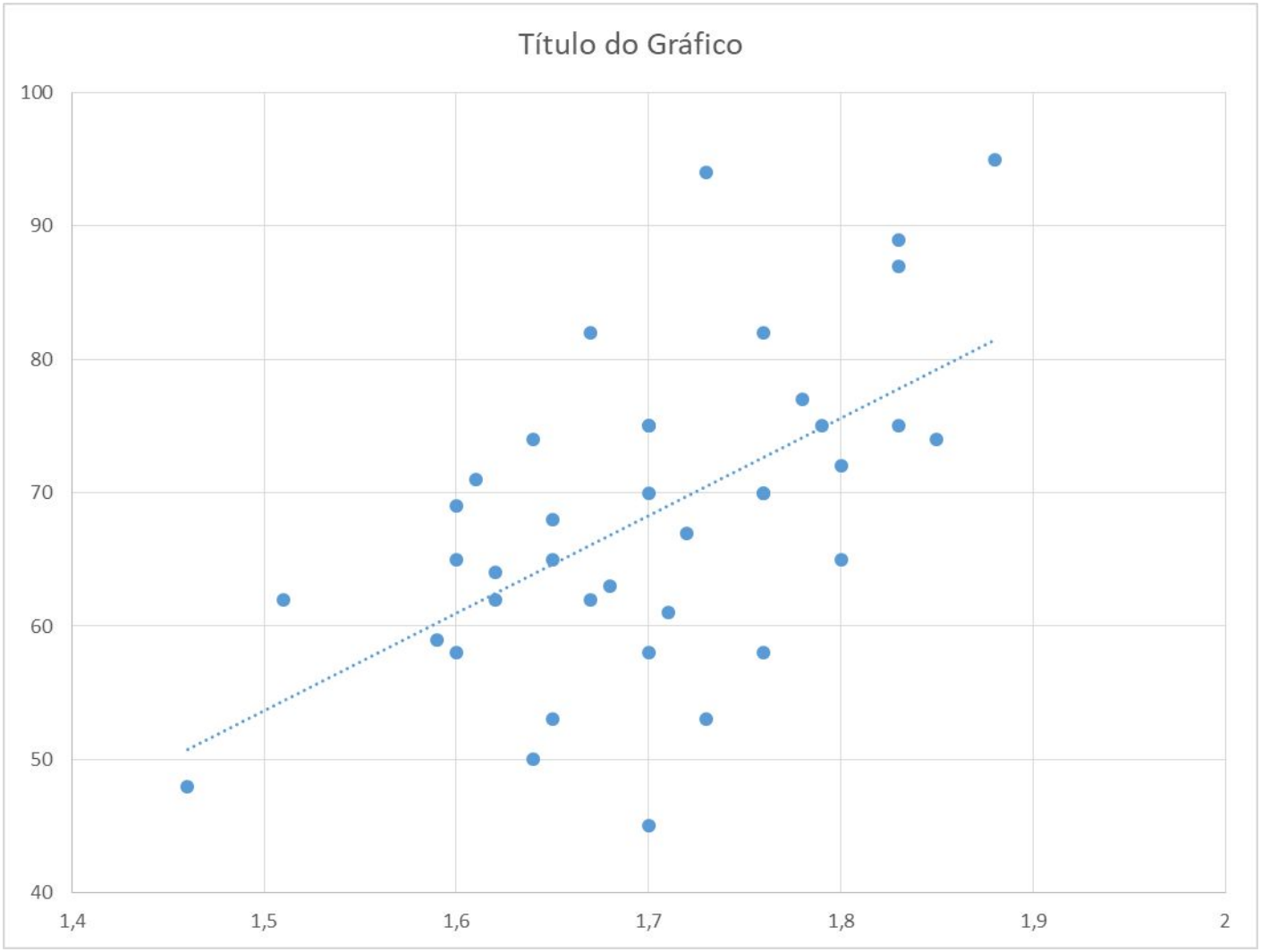
3) Os postos estão dispersos e não é possível estabelecer uma relação direta entre as variáveis: a correlação será muito baixa. As variáveis são ditas não correlacionadas.



Correlação perfeita: 1 ou -1
Pontos perfeitamente alinhados






Correlação fraca

INSTITUTO FEDERAL
Rio Grande do Sul

# Coeficiente de correlação

- É o valor numérico, uma medida, para este grau de associação entre as variáveis.
- Utilizaremos o tipo de relação mais simples: linear.
- Ou seja, vamos julgar que os valores obtidos serão aproximados por uma reta.
- Valores possíveis: de -1 a 1.

# Coeficiente de correlação linear ou de Pearson: r

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2] \ . \ [n\sum y_i^2 - (\sum y_i)^2]}}$$

No Excel: (não importa a ordem de x e y)
=PEARSON(valores de x; valores de y)
exemplo
=PEARSON(A1:A5;B1:B5)

# Coeficiente de correlação linear ou de Pearson: r

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2] \;.\; [n\sum y_i^2 - (\sum y_i)^2]}}$$

- 0 a 0,3: correlação desprezível
- 0,3 a 0,5 correlação fraca
- 0,5 a 0,7 correlação moderada
- A cima de 0,7: correlação forte

# Covariância

A covariância ( ou variância conjunta) é “ à variância, porém, considera as duas variáveis:

$$Cov(x,y) = \sum (x_i - \bar{x}).(y_i - \bar{y})$$

No Excel: (não importa a ordem de x e y)
=COVAR.P(valores de x; valores de y)
(populacional)

=COVAR.A(x;y)
(amostral)

Outra forma de calcular o coeficiente r:

$$r = \frac{Cov(x,y)}{\sqrt{var(x).var(y)}}$$

Em que a covariância é calculada por

$$Cov(x,y) = \sum (x_i - \bar{x}).(y_i - \bar{y})$$

# Exemplo

Calcular a correlação linear e a covariância entre as variáveis (use o excel):

| | | | | | | |
|---|---|---|---|---|---|---|
| Amostra A | 4 | 8 | 3 | 9 | 7 | 5 |
| Amostra B | 1 | 5 | 2 | 14 | 3 | 11 |

$$r = \frac{Cov(x,y)}{\sqrt{var(x).var(y)}}$$

$$Cov(x,y) = \sum (x_i - \bar{x}).(y_i - \bar{y})$$

R.: 0,575

# Coeficiente de determinação – $r^2$

- É uma medida de ajustamento dos valores observados.

- É sempre um valor entre 0 e 1 e representa a **porcentagem** em qualidade do modelo em relação à realidade.

- Quanto mais próximo de 1, melhor é o  modelo.

# Coeficiente de determinação – $r^2$

No Excel:

=RQUAD(valores de x ; valores de y)

Ou basta calcular o r de Pearson, e elevar ao quadrado.

Exemplo

| | | | | | | |
|---|---|---|---|---|---|---|
| Amostra A | 4 | 8 | 3 | 9 | 7 | 5 |
| Amostra B | 1 | 5 | 2 | 14 | 3 | 11 |

R.: 0,575

R2= 0,3306

R2= 33,06%

# Regressão linear

Em um conjunto (finito) de dados **x e y,** em que a correlação linear é significativa é possível "prever" uma aproximação para valor esperado para medidas não utilizadas ou dados não coletados.

Para isso, os pontos obtidos são "aproximados" por uma função que melhor se encaixa sobre eles. No caso linear, utilizamos uma **reta de regressão.**

**Exemplo:** Suponha que em um experimento para verificar a variação de uma amostra obteve-se os seguintes dados:

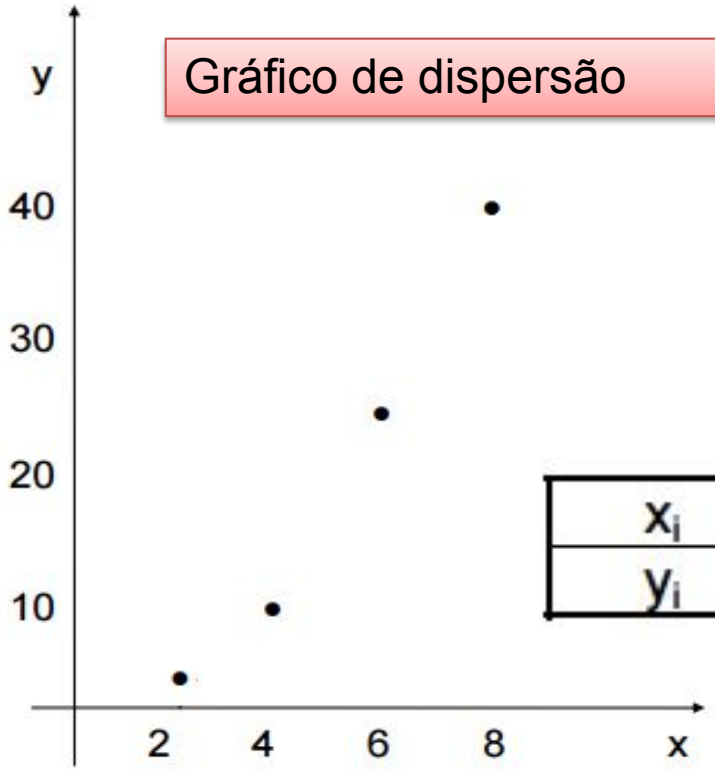
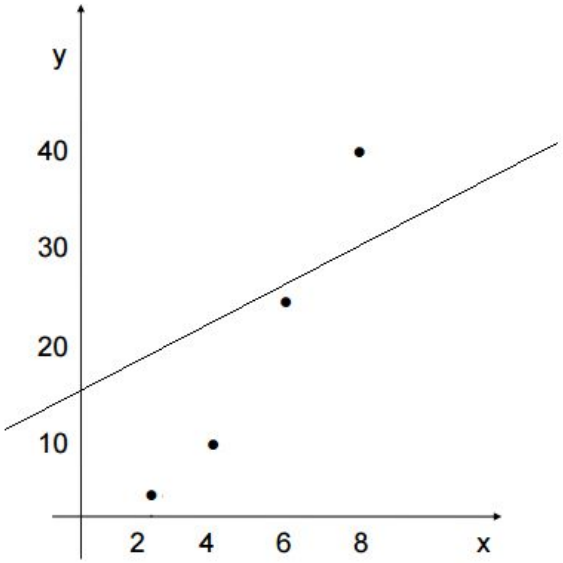


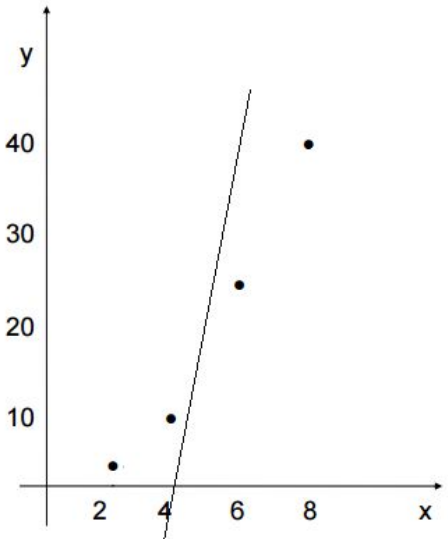

Gráfico de dispersão

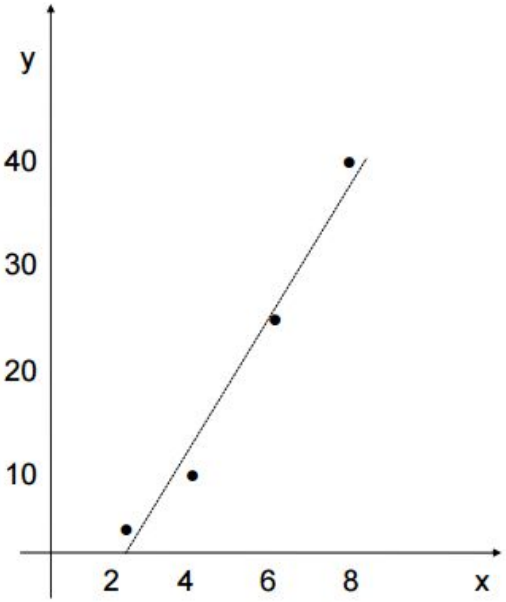
**Variável independente: x**
**Variável dependente: y**

| $x_i$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $y_i$ | 2 | 11 | 28 | 40 |

# Qual das retas melhor se aproxima dos dados?

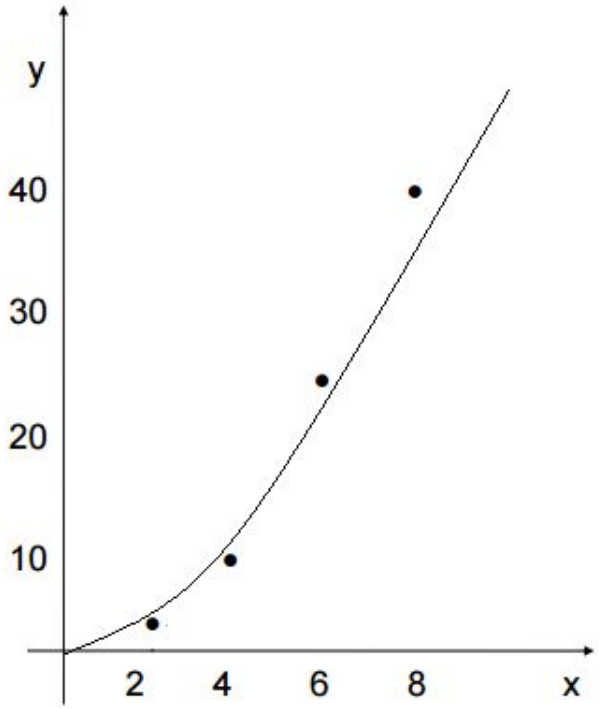


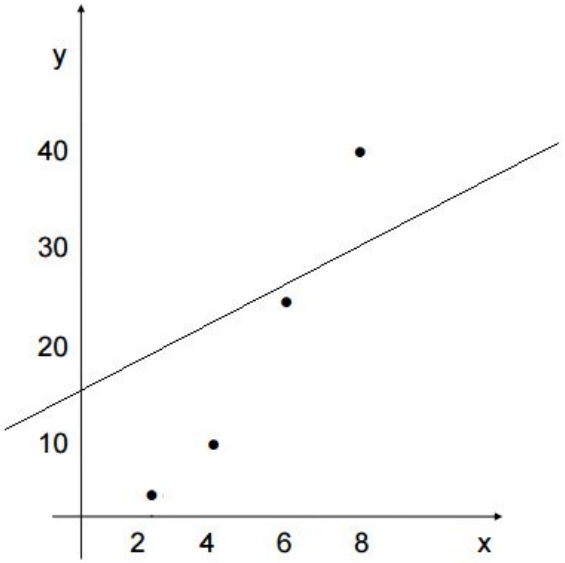
Observação: utilizaremos  apenas função de primeiro grau **(reta)** que melhor aproxima os dados obtidos.

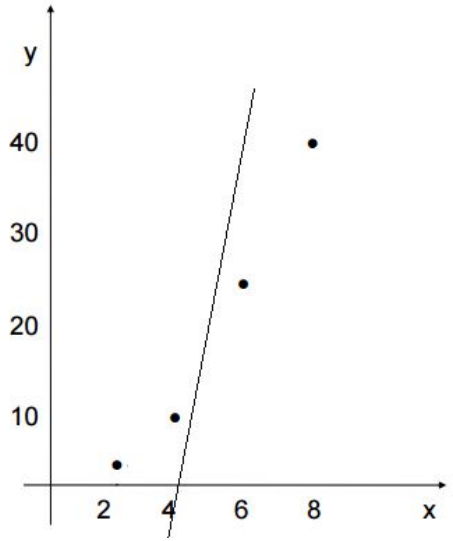
Dependendo da variável estudada, pode ser necessário utilizar outras funções para interpolar os dados. (quadrática, polinômio, logaritmo...)

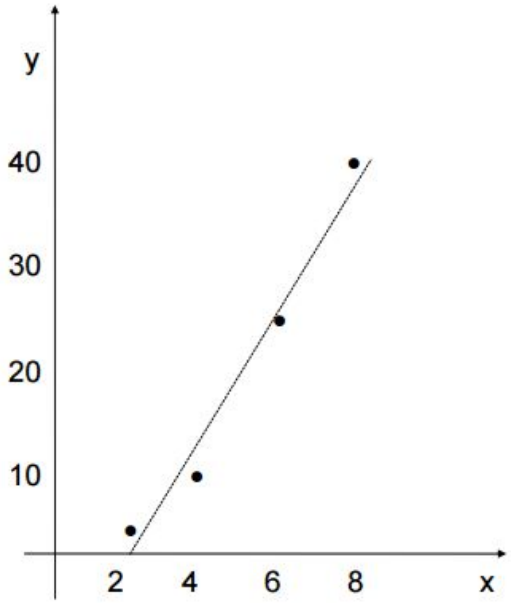



INSTITUTO FEDERAL
Rio Grande do Sul

# Como construir a equação da reta que melhor se encaixa entre os pontos?




Equação da reta: $f(x) = ax + b$

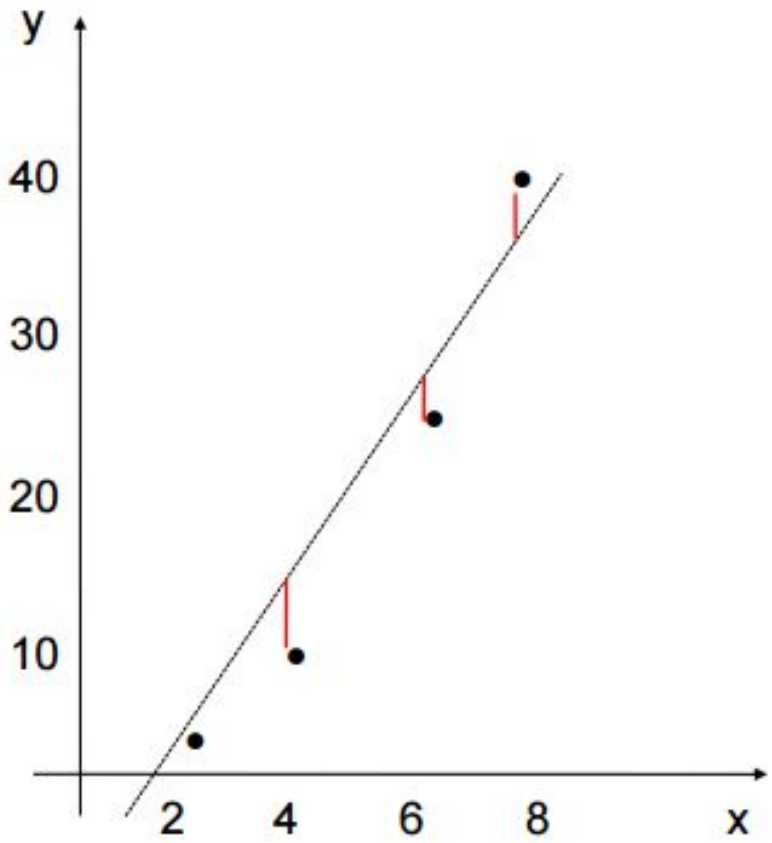
INSTITUTO FEDERAL
Rio Grande do Sul

# Equação da reta (função de primeiro grau):

$$f(x) = ax + b$$

É preciso determinar o coeficiente angular “a” (inclinação) e linear “b”

Procuramos uma reta em que a distância entre ela e os pontos seja a menor possível:

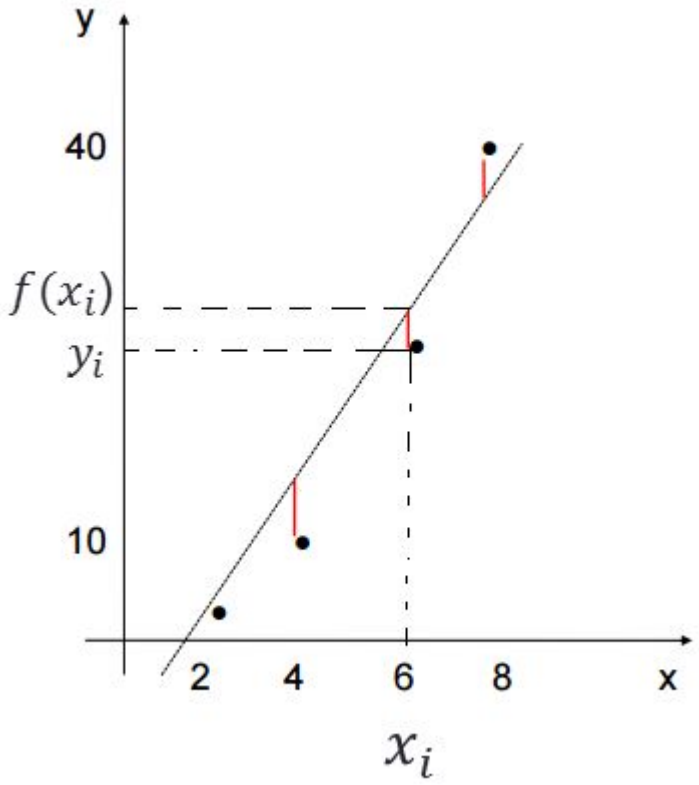


Para isso, calculamos cada uma das distâncias que serão obtidas e elevamos ao quadrado.

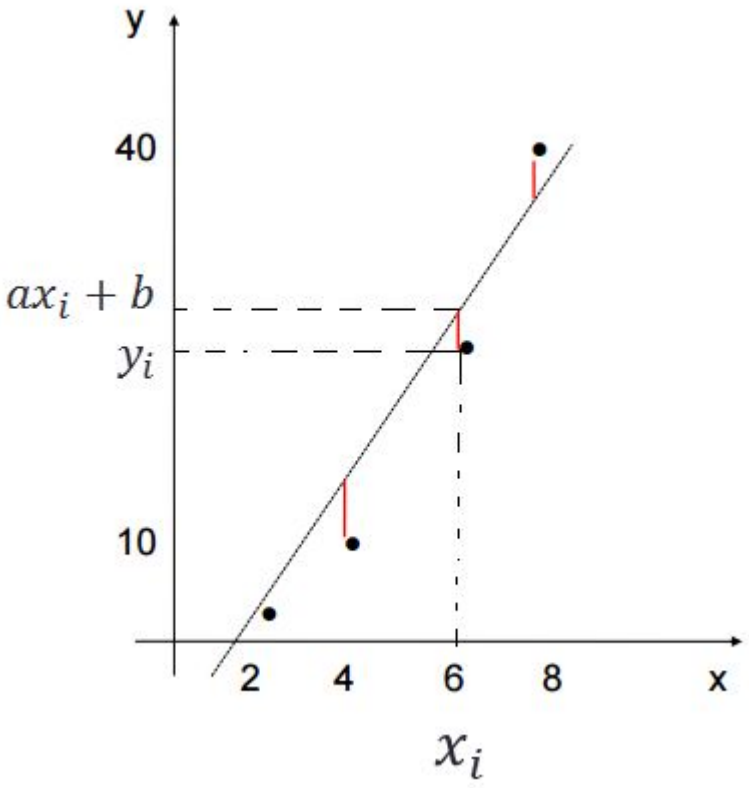
(onde isso já apareceu?)

Minimizamos a soma dos **quadrados** das diferenças

Para cada medida $x_i$, temos o valor obtido no experimento, e o valor que está na reta aproximada:



Para cada medida $x_i$, temos o valor obtido no experimento, e o valor que está na reta aproximada:




Diferença entre a medida e a reta:

$$(ax_i+b) - y_i$$

Quadrado da diferença:

$$((ax_i+b) - y_i)^2$$

Portanto, queremos minimizar as somas destes quadrados:

$$\sum_{i=1}^{n} \left( (ax_i + b) - y_i \right)^2$$

Para minimizar esta soma utiliza-se Derivada de uma função (Cálculo diferencial)

Utilizando derivadas é possível mostrar que os coeficientes a e b serão obtidos utilizando as equações (chamadas de equações normais):

$$a = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right).\left(\sum_{i=1}^{n} y_i\right)}{n.\left(\sum_{i=1}^{n} x_i^{\,2}\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad e$$

$$b = \frac{\left(\sum_{i=1}^{n} x_i^{\,2}\right).\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i y_i\right).\left(\sum_{i=1}^{n} x_i\right)}{n.\left(\sum_{i=1}^{n} x_i^{\,2}\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Para simplificar a notação, podemos utilizar apenas:

$$a = \frac{n \sum x_i y_i - \sum x_i \cdot \sum y_i}{n. \sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{\sum x_i^2 \cdot \sum y_i - \sum x_i y_i \cdot \sum x_i}{n. \sum x_i^2 - (\sum x_i)^2}$$

Portanto, precisamos das seguintes informações para fazer o cálculo:

| $x_i$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $y_i$ | 2 | 11 | 28 | 40 |

$$\sum x_i y_i$$

$$\sum x_i$$

$$\sum y_i$$

$$\sum x_i^2$$

| $x_i$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $y_i$ | 2 | 11 | 28 | 40 |

| xi | yi | xi.yi | xi^2 |
|---|---|---|---|
| 2 | 2 | | |
| 4 | 11 | | |
| 6 | 28 | | |
| 8 | 40 | | |

| Soma(xi) | Soma(yi) | Soma(xi.yi) | Soma(xi^2) |
|---|---|---|---|
| | | | |

| $x_i$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $y_i$ | 2 | 11 | 28 | 40 |

| xi | yi | xi.yi | xi^2 |
|---|---|---|---|
| 2 | 2 | 4 | 4 |
| 4 | 11 | 44 | 16 |
| 6 | 28 | 168 | 36 |
| 8 | 40 | 320 | 64 |

| Soma(xi) | Soma(yi) | Soma(xi.yi) | Soma(xi^2) |
|---|---|---|---|
| 20 | 81 | 536 | 120 |

# substituindo nas “ fórmulas”:

| Soma(xi) | Soma(yi) | Soma(xi.yi) | Soma(xi^2) |
|---|---|---|---|
| 20 | 81 | 536 | 120 |

$$a = \frac{n \sum x_i y_i - \sum x_i \, . \, \sum y_i}{n. \sum x_i^2 - (\sum x_i)^2}$$

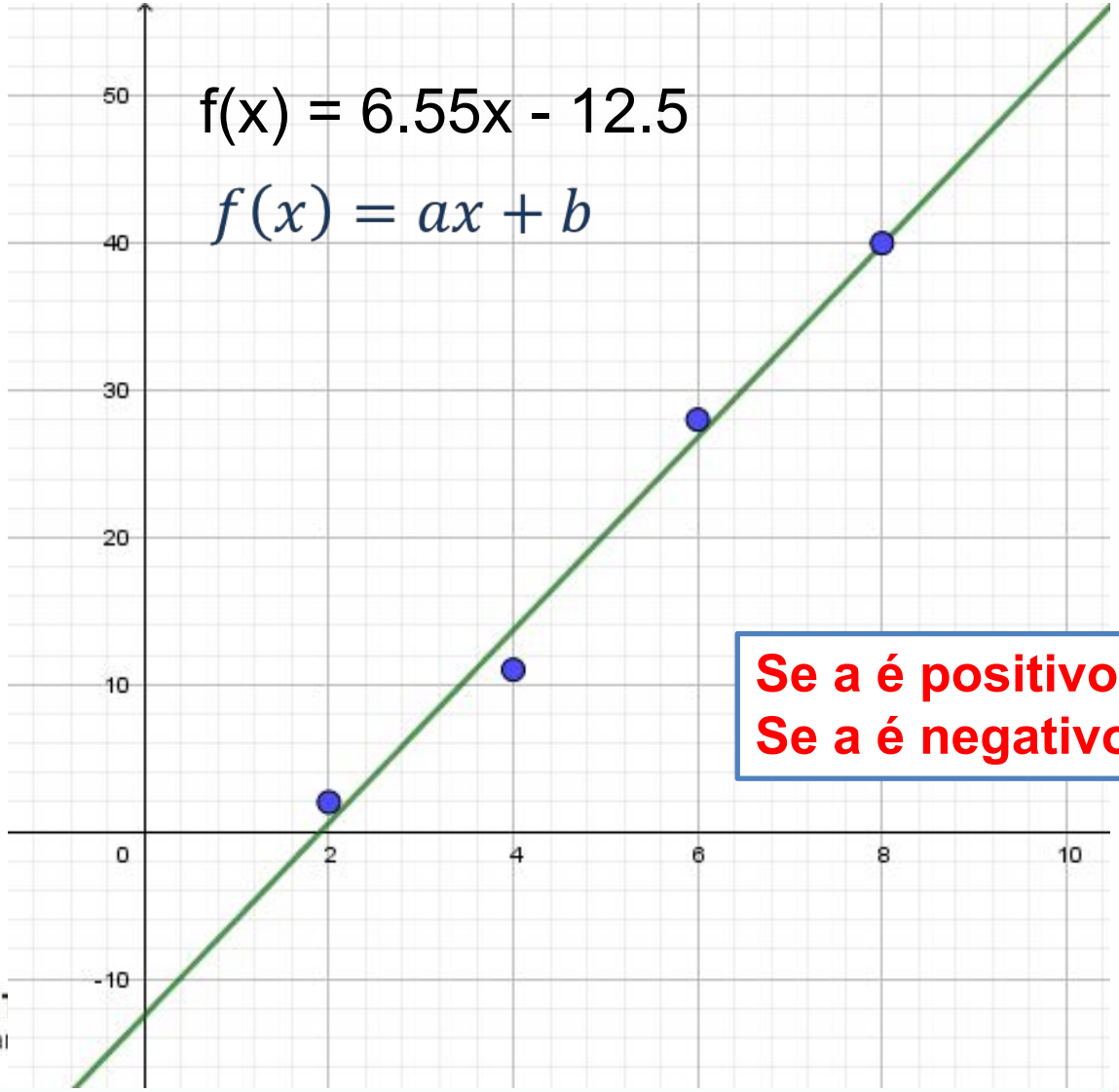
$$b = \frac{\sum x_i^2 \, . \, \sum y_i - \sum x_i y_i \, . \, \sum x_i}{n. \sum x_i^2 - (\sum x_i)^2}$$

Portanto, a reta será:

$$f(x) = ax + b$$

f(x) = 6.55x - 12.5

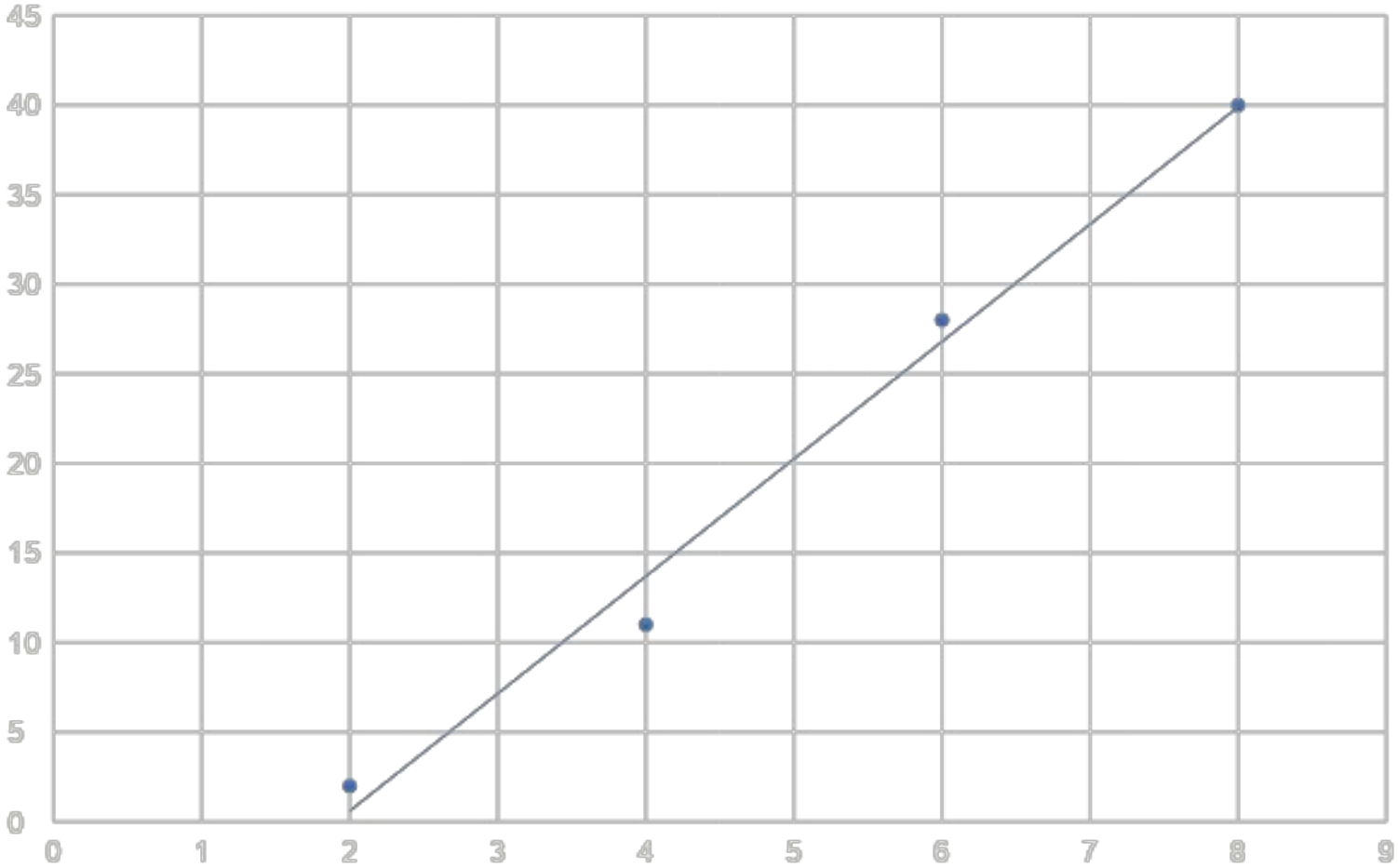
- **a** é a inclinação da reta  (tangente do ângulo)
- **b** é o valor de y em que a reta cruza o eixo vertical.




**Se a é positivo:** crescente
**Se a é negativo:** decrescente.

# No EXCEL:

1. Inserir os dados x e y em duas colunas
2. Construir o gráfico de dispersão
3. Selecione “exibir linha de tendência” (Linear)
4. “Exibir equação no gráfico”


45
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9

## Utilizando funções:

Para encontrar **a**:

=INCLINAÇÃO (valores de y; valores de x)

Para encontrar o **b**:

=INTERÇEPÇÃO(valores de y; valores de x)

**ATENÇÃO: inserir primeiro os valores de y**

Com a equação linearizada, ou seja, com a reta aproximada, podemos obter uma “ previsão” para qualquer valor de x, basta substituir x na função e obter f(x) =y.

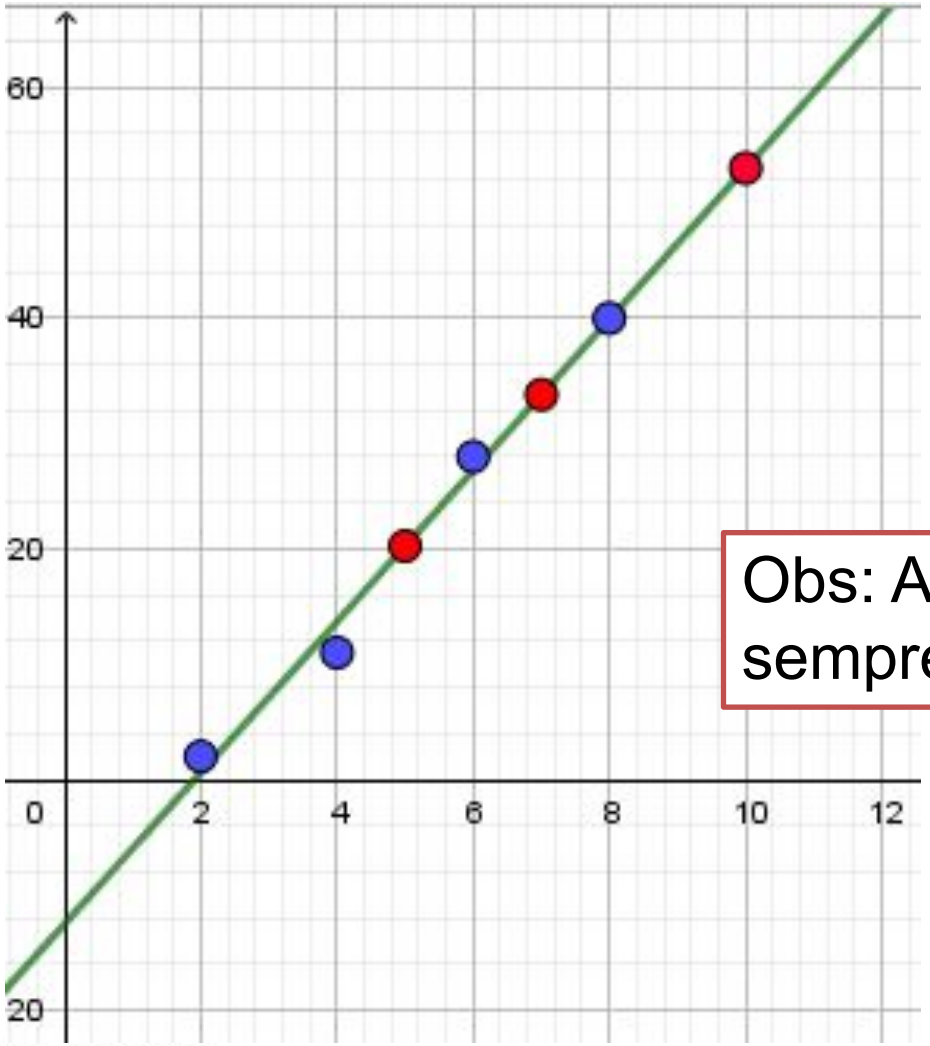
$$y = 6.55x - 12.5$$

Ex.: No exemplo anterior, obtenha o valor da medida esperada para

$$x = 5 \; e \qquad x = 7 \qquad x = 10,$$

y=20,25 y=33,35 y=53

# Observe essas “previsões” no gráfico de dispersão:




$x = 5$ $x = 7$ $x = 10,$

y=20,25 y=33,35 y=53

Obs: As “previsões” sempre pertencem à reta.

## Para praticar no excel:

a) Calcule o coeficiente de correlação $r$ e $r^2$. Os dados são correlacionados? Positivo ou negativo?
b) o gráfico de dispersão
c) Encontre a reta de regressão (crescente ou decrescente?)

**Exercício 1:**

**Tabela 1.** Valores experimentais da posição de um carrinho em função do tempo.

| X - tempo (s) | Y - posição (m) |
|---|---|
| 0,100 | 0,51 |
| 0,200 | 0,59 |
| 0,300 | 0,72 |
| 0,400 | 0,80 |
| 0,500 | 0,92 |

Resposta: 1,08x +0,38

**Exercício 2:** Em uma clínica de Endocrinologia foi feita uma pesquisa com 5 mulheres de 50 anos de idade. Nessa pesquisa foram feitas duas perguntas.

Qual é o nível de HDL – Colesterol em seu sangue?
Quantas horas semanais você pratica exercícios físicos?

Os resultados estão descritos na tabela a seguir.

| HDL – Colesterol mg/dl | Número de horas de prática de exercícios físicos |
|---|---|
| 40 | 0 |
| 50 | 2 |
| 55 | 3 |
| 60 | 4 |
| 65 | 6 |

R.: Y = 0,23x - 9,42