

---

**EST 105**

**INICIAÇÃO À ESTATÍSTICA**

**CORRELAÇÃO E REGRESSÃO**  
**Resumo**

Departamento de Estatística – UFV

Av. Peter Henry Rolfs, s/n

Campus Universitário

36570.977 – Viçosa, MG

<http://www.det.ufv.br/>



## Motivação:

- Geralmente existe o interesse em se investigar a relação entre duas ou mais variáveis que foram medidas em uma pesquisa.
- Por exemplo, a quantidade vendida de um produto pode estar relacionada ao preço deste produto. Ou, a quantidade de grãos produzida por uma variedade de arroz, pode estar associada à quantidade de adubo utilizada, etc.
- Seja uma amostra de valores de duas variáveis aleatórias (X e Y), por exemplo:

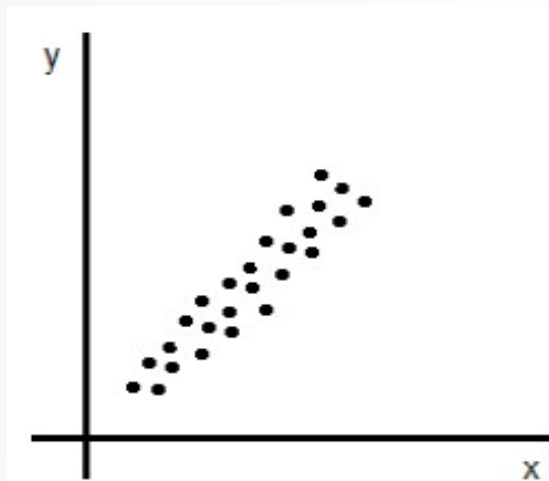
X	$X_1$	$X_2$	$X_3$	$X_4$	...	$X_n$
Y	$Y_1$	$Y_2$	$Y_3$	$Y_4$	...	$Y_n$

# Diagrama de dispersão

- Se representarmos os pares de valores  $(X_i, Y_i)$  num sistema cartesiano, temos um **diagrama de dispersão**. A construção de um gráfico deste tipo pode nos auxiliar a identificar o tipo da associação entre as variáveis aleatórias  $X$  e  $Y$ .

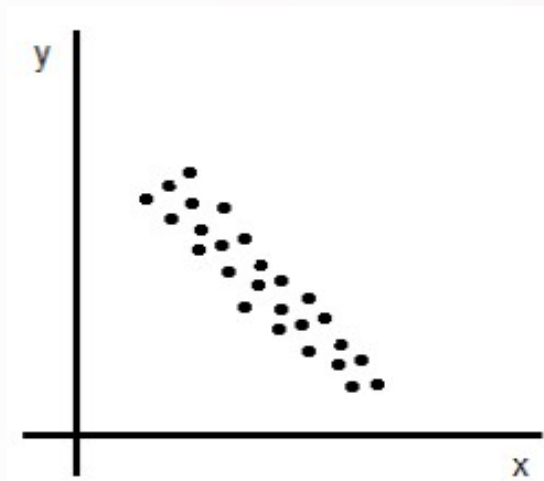
Vejamos algumas possíveis configurações:

(a) Correlação positiva



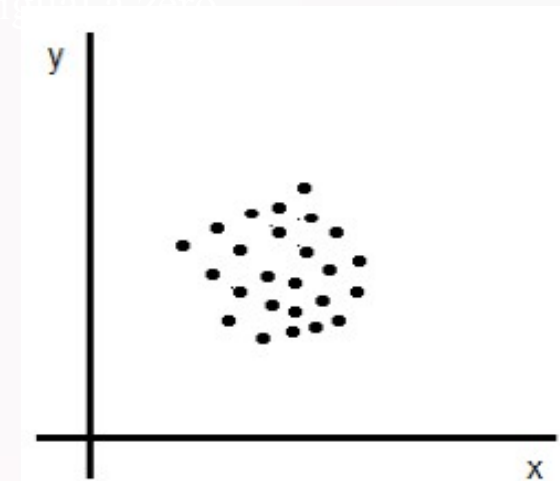
$X$  = altura e  $Y$  = peso

(b) Correlação negativa



$X$  = preço e  $Y$  = número de  
itens vendidos

(c) Correlação aproximadamente  
nula



$X$  = altura e  $Y$  = renda

# Coeficiente de correlação amostral

- Medida usada para avaliar o grau de **associação linear** entre duas variáveis aleatórias  $X$  e  $Y$  é chamada **coeficiente de correlação** ( $\rho$ ).

O **coeficiente de correlação amostral** entre as variáveis  $X$  e  $Y$  pode ser obtido por:

X	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	...	X <sub>n</sub>
Y	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>	Y <sub>4</sub>	...	Y <sub>n</sub>

$$r_{XY} = \frac{SPD_{XY}}{\sqrt{SQD_X \times SQD_Y}}, -1 \leq r_{XY} \leq 1$$

Em que:

$$SQD_X = \sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}; SQD_Y = \sum_{i=1}^n Y_i^2 - \frac{(\sum_{i=1}^n Y_i)^2}{n}; SPD_{XY} = \sum_{i=1}^n X_i Y_i - \frac{(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{n}$$

# Diagrama de dispersão

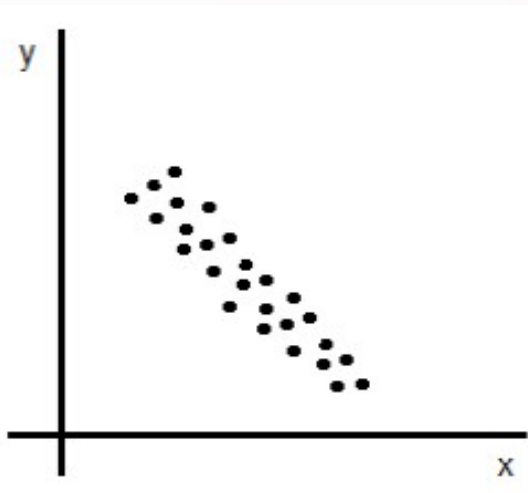
- Vejamos algumas possíveis configurações:

(a) Correlação positiva



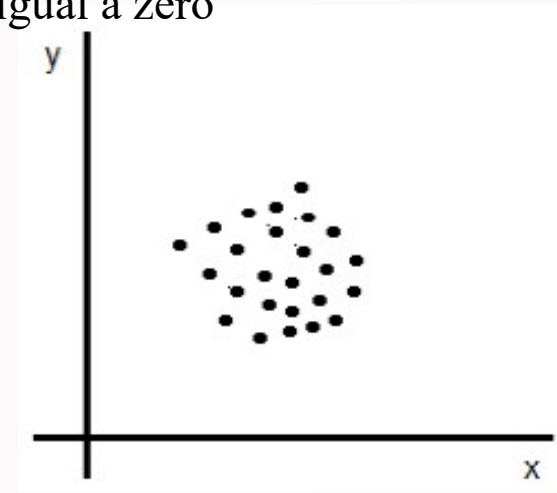
X = altura e Y = peso

(b) Correlação negativa



X = preço e Y = número de  
itens vendidos

(c) Correlação aproximadamente  
igual a zero



X = altura e Y = renda

## Exemplo

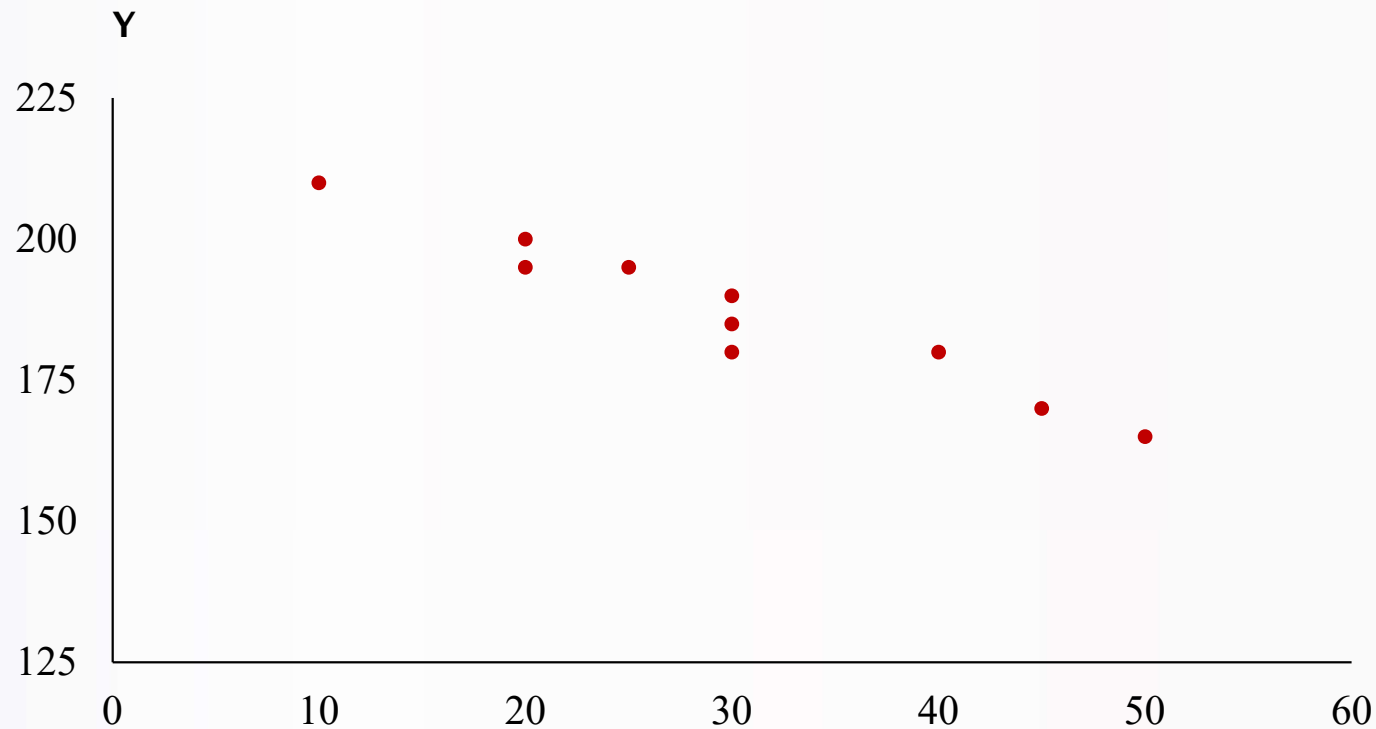
A tabela a seguir apresenta informações sobre a idade ( $X$ , em anos) e o número máximo de batimentos cardíacos ( $Y$ , em minutos) de 10 pacientes amostrados em um estudo médico. Calcule o **coeficiente de correlação amostral** entre as variáveis  $X$  e  $Y$ .

X	10	20	20	25	30	30	30	40	45	50
Y	210	200	195	195	190	180	185	180	170	165

# Diagrama de dispersão

Idade ( $X$ , em anos) e o número máximo de batimentos cardíacos ( $Y$ , em minutos) de 10 pacientes amostrados em um estudo médico.

X	10	20	20	25	30	30	30	40	45	50
Y	210	200	195	195	190	180	185	180	170	165



# Regressão Linear Simples (RLS)

- Tem por objetivo **estabelecer uma relação funcional** entre uma variável aleatória e dependente ( $Y$ ) e uma variável fixa e independente ( $X$ ).

## 1. O modelo de RLS

- **Modelo estatístico:**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

Em que

$X_i$  é i-ésimo valor da variável explicativa ou independente ( $X$ );

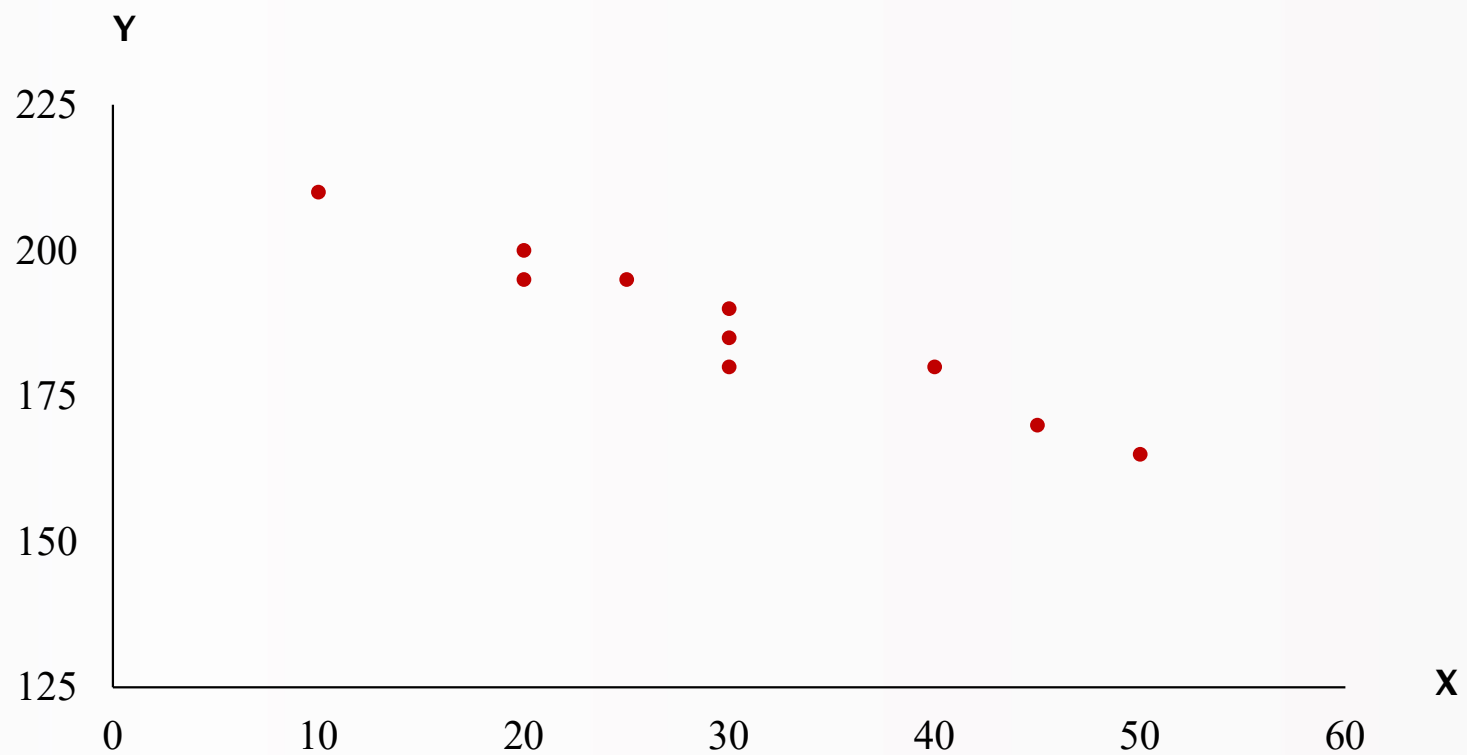
$Y_i$  é i-ésimo valor da variável dependente ou resposta ( $Y$ );

$\beta_0$  é a constante da regressão ou intercepto (parâmetro);

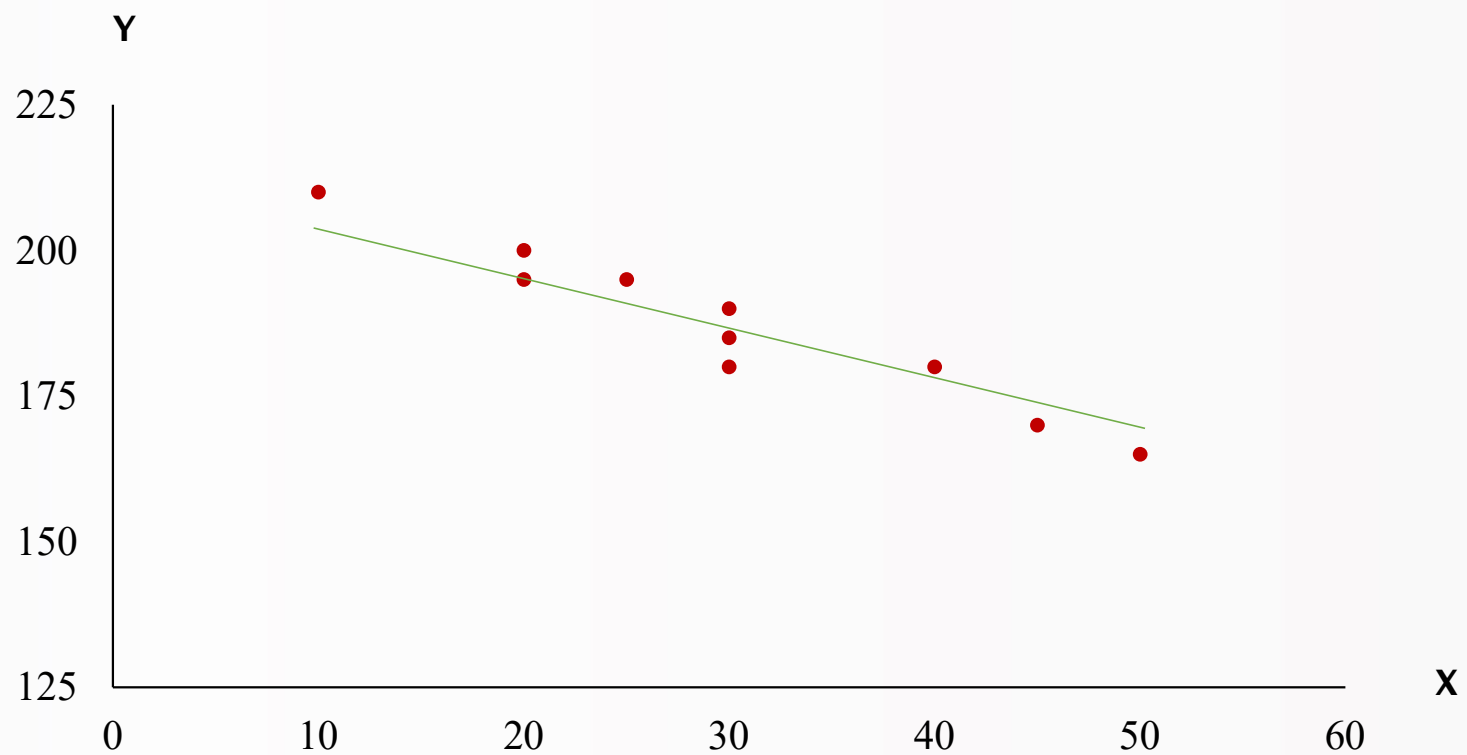
$\beta_1$  é o coeficiente de regressão ou coeficiente angular (parâmetro);

$\varepsilon_i$  é o i-ésimo erro aleatório (não observável).

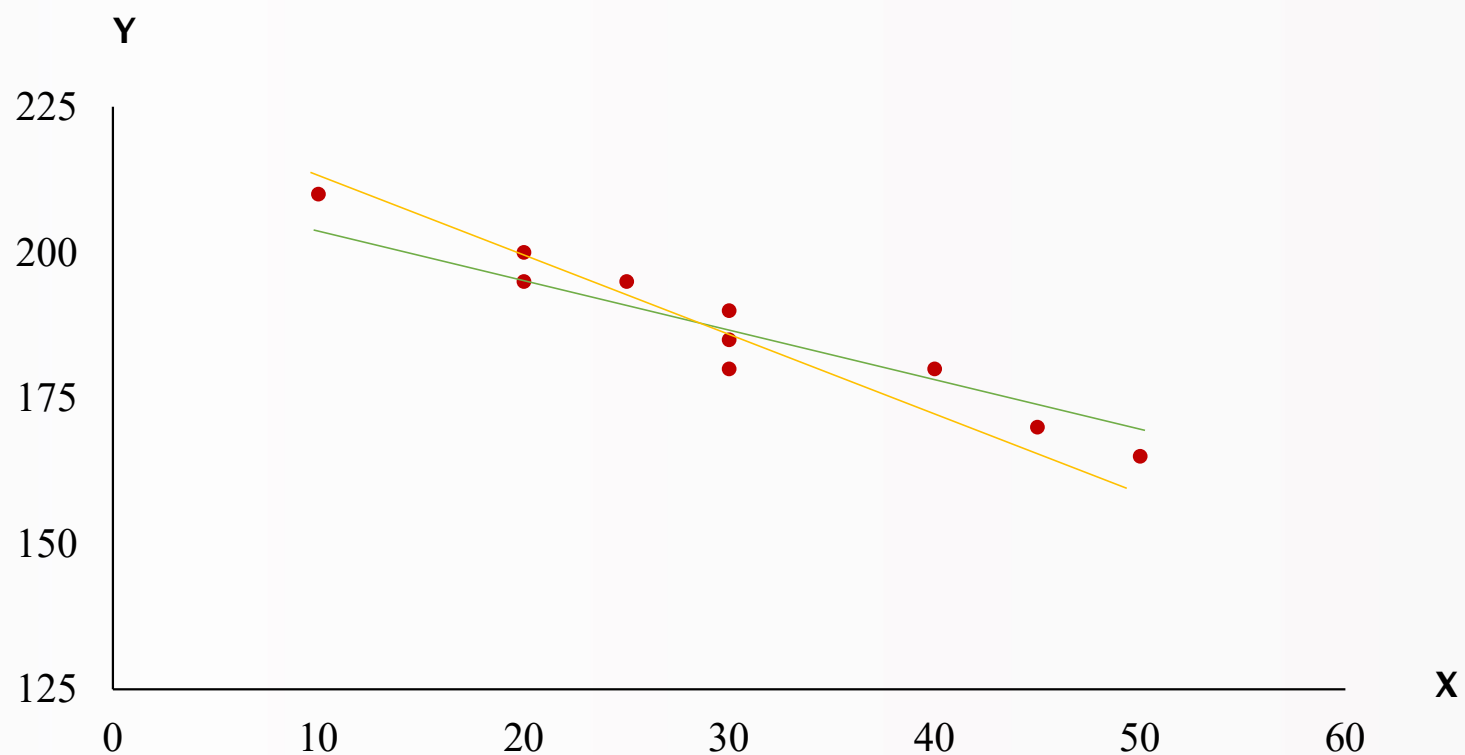
# Regressão Linear Simples (RLS)



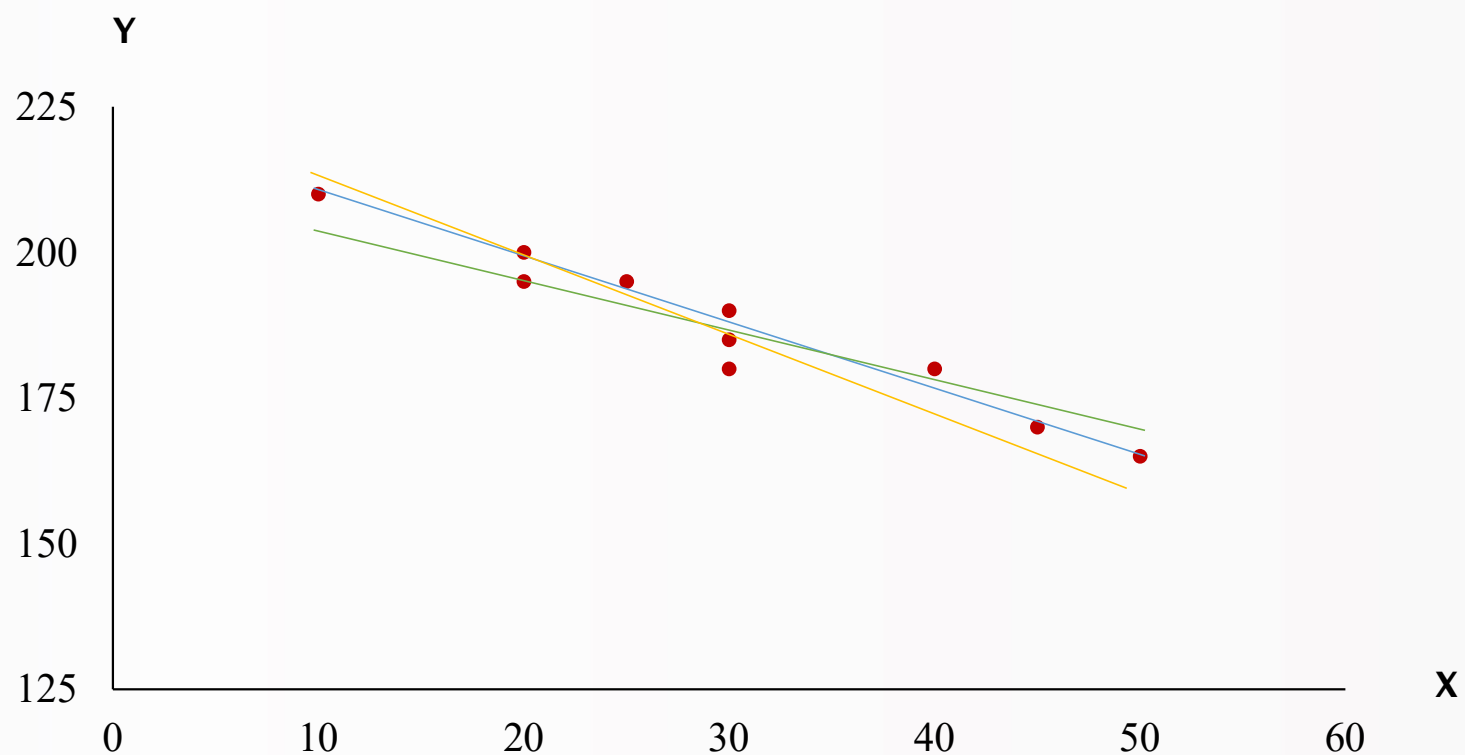
# Regressão Linear Simples (RLS)



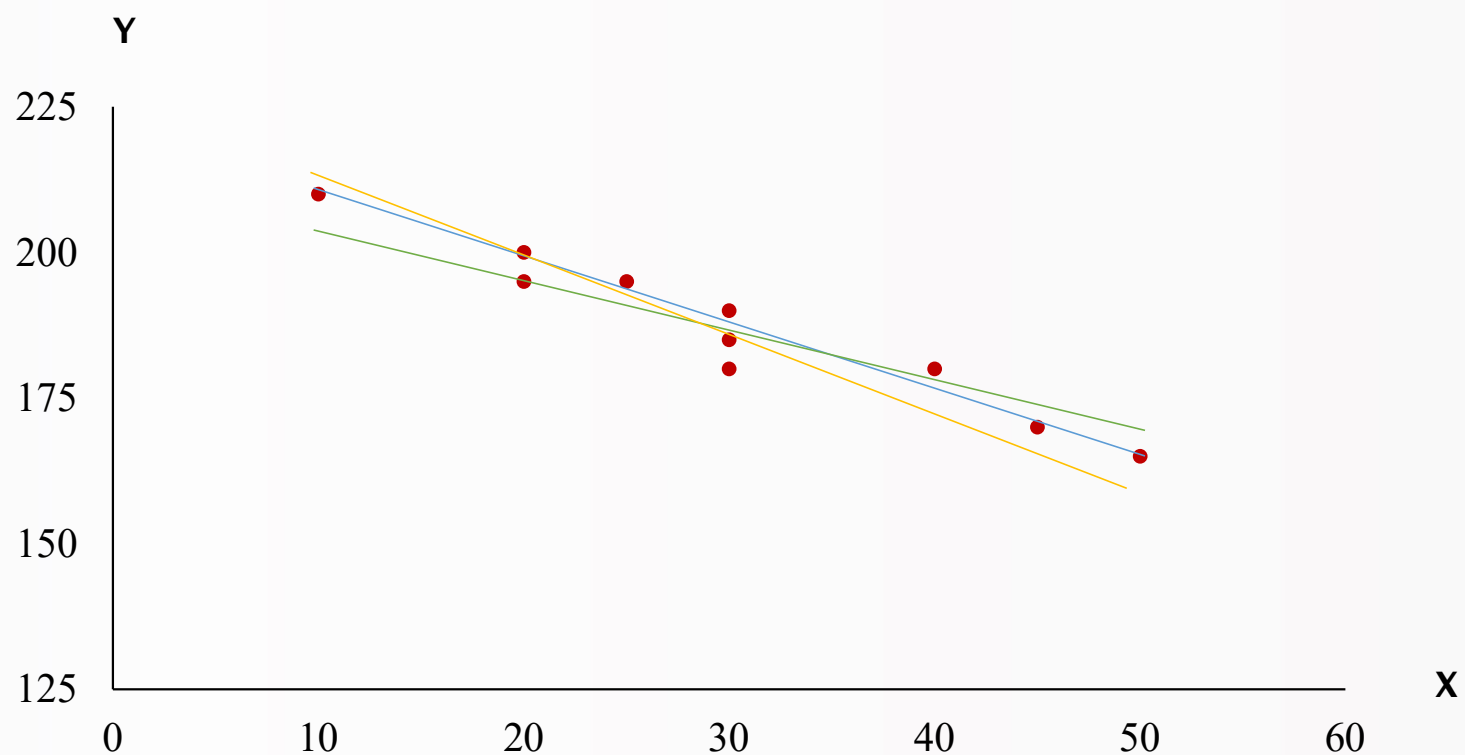
# Regressão Linear Simples (RLS)



# Regressão Linear Simples (RLS)



# Regressão Linear Simples (RLS)



# Regressão Linear Simples

## 2. Método de Estimação

- Apenas uma **amostra** de pares  $(x, y)$  é observada, logo, a verdadeira relação linear entre  $X$  e  $Y$  não será conhecida e sim estimada pela análise de regressão linear simples.
- **Método dos Mínimos Quadrados** (MMQ): o objetivo deste método é obter as estimativas dos parâmetros que minimizam o valor da soma de quadrados dos erros aleatórios.
- Definido o modelo  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ , então,  $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$ . O MMQ define  $\min Z = \min \sum_{i=1}^n \varepsilon_i^2 = \min \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$ .
- Os estimadores (fórmulas) que produzem estimativas dos parâmetros (valores) que minimizam  $Z$ , são obtidos pela derivação parcial de  $Z$  em relação aos parâmetros ( $\beta_0$  e  $\beta_1$ ) do modelo. Isto é,  $\frac{\partial Z}{\partial \beta_0}$  e  $\frac{\partial Z}{\partial \beta_1}$ .

# Regressão Linear Simples

- $\hat{\beta}_1$  ou  $b_1$  é o estimador (fórmula) do parâmetro  $\beta_1$ .

$$\hat{\beta}_1 = b_1 = \frac{SPD_{XY}}{SQD_X} = \frac{\left[ \sum_{i=1}^n X_i Y_i - \frac{(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{n} \right]}{\left[ \sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n} \right]}$$

- $\hat{\beta}_0$  ou  $b_0$  é o estimador (fórmula) do parâmetro  $\beta_0$ .

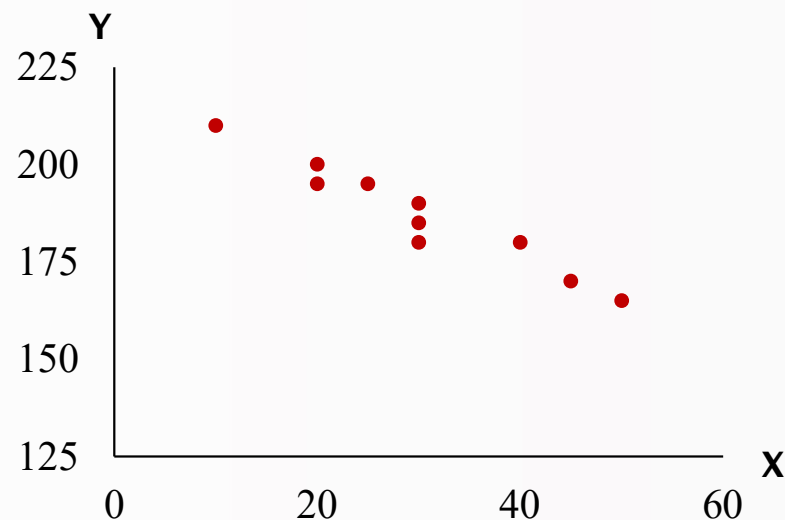
$$\hat{\beta}_0 = b_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{\sum_{i=1}^n Y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^n X_i}{n}$$

**Equação estimada (ou modelo ajustado):**

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = b_0 + b_1 X_i$$

# Exemplo

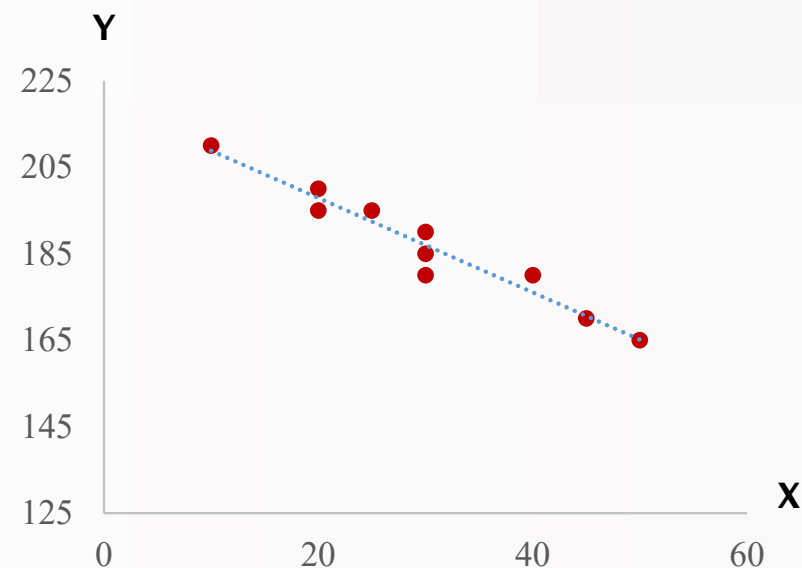
Considere novamente, os dados do exemplo inicial referentes à idade ( $X$ , em anos) e o número máximo de batimentos cardíacos por minuto ( $Y$ ) de  $n = 10$  pacientes amostrados.



**a) Apresente a equação ajustada.**

- A equação ajustada ou o modelo ajustado:

$$\hat{Y}_i = 219,78 - 1,093X_i$$



### 3. Interpretação

- $\hat{\beta}_0$  representa o valor estimado de  $Y$  ( $\hat{Y}$ ) quando  $X$  é igual a zero. Algumas vezes essa estimativa não possuirá uma interpretação prática.
- $\hat{\beta}_1$  representa o aumento ( $\hat{\beta}_1 > 0$ ) ou a redução ( $\hat{\beta}_1 < 0$ ) média(o) estimada em  $Y$  para cada aumento unitário em  $X$ .

# Exemplo

**b) Interprete o coeficiente de regressão.**

# Regressão Linear Simples

## 4. Desvios da regressão (ou resíduos)

São estimativas para os erros aleatórios. Em um modelo bem ajustado, isto é, aquele no qual a variável  $X$  é útil para explicar as variações na variável resposta  $Y$ , espera-se que os desvios sejam pequenos.

Os resíduos/desvios podem ser calculados como:

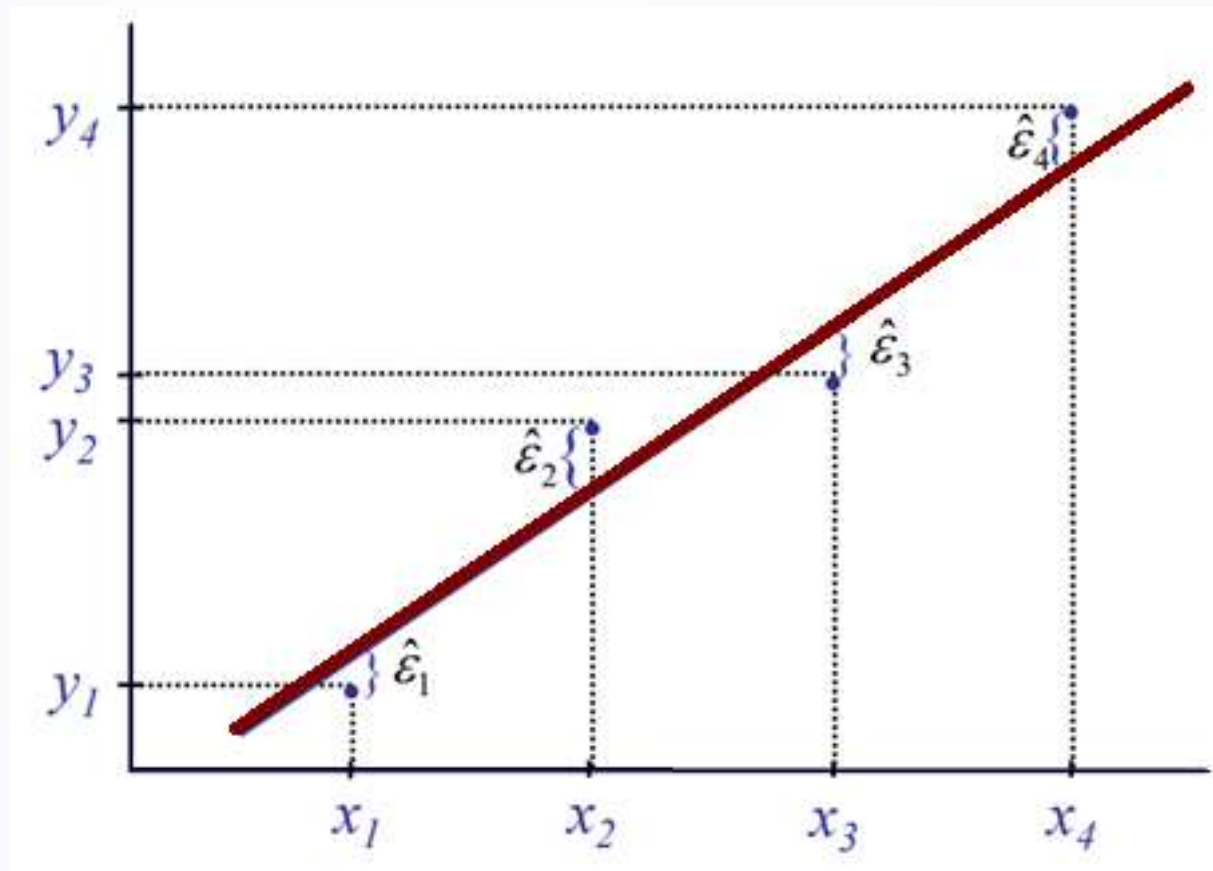
$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

Diagram illustrating the calculation of the residual  $\hat{\varepsilon}_i$ :

- $Y_i$  is labeled as **Valor observado** (Observed value).
- $\hat{Y}_i$  is labeled as **Valor estimado** (Estimated value).

# Regressão Linear Simples

$$(b_0, b_1) = (\hat{\beta}_0, \hat{\beta}_1) = \arg \min (\sum_{i=1}^n \varepsilon_i^2)$$



$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

## Exemplo

X	10	20	20	25	30	30	30	40	45	<b>50</b>
Y	210	200	195	195	190	180	185	180	170	<b>165</b>

d) Qual é a estimativa do número máximo de batimentos cardíacos para um indivíduo de 50 anos?

e) Calcule o desvio da regressão para a observação  $X = 50$ .

# Regressão Linear Simples

## 5. Extrapolação

- É possível obter estimativas para  $Y$  usando valores de  $X$  que não foram estudados. **Entretanto, estes devem estar dentro do intervalo coberto pela amostra.**
- Utilizar o modelo ajustado fora da amplitude estudada significa fazer uma **extrapolação**. A equação ajustada é razoável para interpolar dentro do intervalo coberto pela amostra, mas pode ser inapropriada para fazer uma extrapolação.
- **ATENÇÃO:** Por este motivo, no nosso exemplo, como o intervalo observado de idade  $X$  não continha  $X = 0$ , então, interpretar  $\hat{\beta}_0$  seria uma EXTRAPOLAÇÃO DO MODELO.

# Exemplo

**f) Estime o número máximo de batimentos cardíacos para um indivíduo de 60 anos.  
Comente a respeito desta estimativa.**

# Regressão Linear Simples

## 6. Coeficiente de Determinação ( $r^2$ )

- O coeficiente de determinação é uma medida da qualidade do ajuste do modelo.
- Indica a proporção da variação na variável dependente  $Y$  que está sendo explicada pela variável independente  $X$  ou pela regressão nos valores de  $X$ .
- O  $r^2$  é expresso em porcentagem e calculado a partir da seguinte expressão:

$$r^2(\%) = \frac{SQ_{\text{Regressão}}}{SQ_{\text{Total}}} 100\%, \quad 0 \leq r^2 \leq 100\%$$

em que

$$SQ_{\text{Total}} = SQD_Y = \sum_{i=1}^n Y_i^2 - \frac{(\sum_{i=1}^n Y_i)^2}{n} \quad \text{e} \quad SQ_{\text{Regressão}} = \hat{\beta}_1 SPD_{XY}.$$

- Quanto maior for o  $r^2$ , melhor é a qualidade do ajuste.

**Obs.:** No caso da RLS, o coeficiente de determinação é igual ao quadrado coeficiente de correlação amostral entre  $X$  e  $Y$ , isto é,  $r^2(\%) = (r_{XY})^2 \times 100(\%)$ .

# Exemplo

**g) Calcule e interprete o coeficiente de determinação.**

# Atividade Proposta

Resolver os exercícios do Roteiro de Aulas abaixo relacionados:

- Exercício 4 – pág. 164
- Exercício 6 – pág. 165
- Exercício 8 – pág. 166
- Exercício 9 – pág. 167
- Exercício 10– pág. 167