

Regressao linear simples

Análise estatística bivariada

- **Associação entre duas variáveis qualitativas**
Tabelas de contingência
- **Associação entre duas variáveis quantitativas**
Correlação e Regressão

Associação entre variáveis quantitativas



Correlação e Regressão

Suponha que X e Y são duas variáveis quantitativas que podemos assumir serem variáveis aleatórias com distribuições normais.

Estamos interessados em analisar uma possível Associação Linear entre elas.

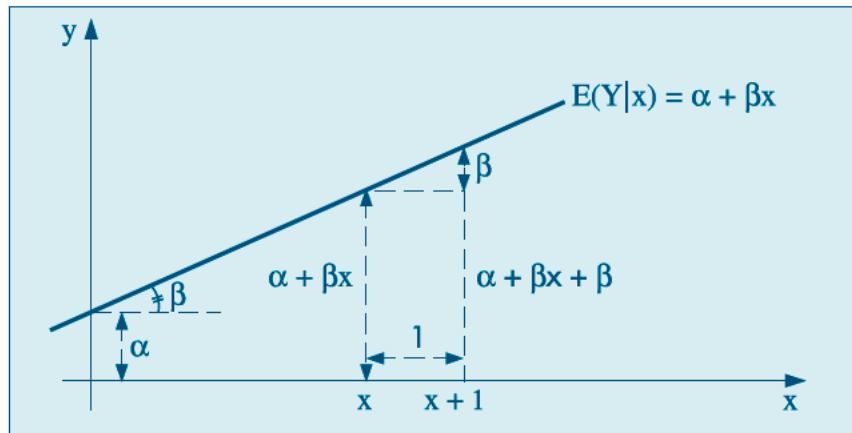
Um modelo simples seria:

$$E(Y|X=x) = \mu(x) = \alpha + \beta x$$

para algum valor de α e β . Ou seja, estamos considerando a possibilidade de que " Y é uma função linear de X ".

$$Y = \alpha + \beta X + \text{"erro aleatório"}$$

Figura 16.2: Representação do modelo $E(Y|x) = \alpha + \beta x$.



(morettin/Bussab)

X

X	Y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

$$y_i = \alpha + \beta x_i + e_i$$

$\underbrace{\hspace{1cm}}$
 "erro"

Exemplos:

Idade e altura das crianças;

Tempo de prática de esportes e ritmo cardíaco;

Tempo de estudo e nota na prova;

Taxa de desemprego e taxa de criminalidade;

Expectativa de vida e taxa de analfabetismo.

Estamos supondo que, para cada valor observado de X , Y é uma variável

aleatoria

X	Y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

$$y_i = \alpha + \beta x_i + e_i$$

valor esperado de y_i para esse valor observado de x_i

"erro"

A aleatoriedade, o "erro", aparece apenas na dependência do Y correspondente a cada valor observado de X .

Note que não estamos levando em conta a aleatoriedade das observações de X : são supostas "dadas" ou "determinísticas".

Em geral supomos:

$$\begin{aligned} E(e_i | X_i = x) &= 0 \\ \text{var}(e_i | X_i = x) &= \sigma^2 \end{aligned} \quad \left. \vphantom{\begin{aligned} E(e_i | X_i = x) &= 0 \\ \text{var}(e_i | X_i = x) &= \sigma^2 \end{aligned}} \right\} \begin{array}{l} \text{não dependem} \\ \text{do} \\ \text{valor } x \end{array}$$

$e_1, e_2, \dots, e_n$ são i.i.d.
erros na amostra.

Por simplicidade, suponha:

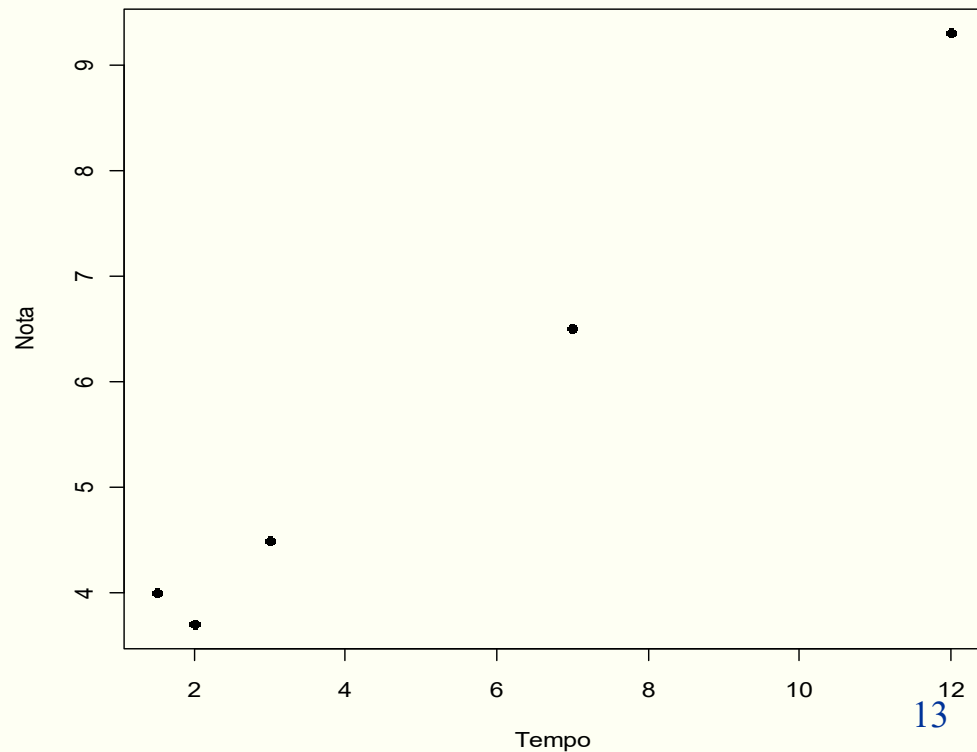
$$e_i \sim N(0; \sigma^2)$$

Diagrama de dispersão

Exemplo 2: Nota na prova e tempo de estudo

X : tempo de estudo (em horas) e Y : nota na prova

Tempo(X)	Nota(Y)
3,0	4,5
7,0	6,5
2,0	3,7
1,5	4,0
12,0	9,3



A partir dos dados observados:

X	Y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

temos duas questoes basicas, que estao intimamente relacionadas:

1) Ha' evidencias que "Y depende linearmente de X" como o modelo supoe?

2) Quais sao as estimativas para os parametros correspondentes?

α, β (e, talvez, σ)

"Reality check": por que estou interessado nestes dois parametros α e β ?

Como estimar α e β ?

Temos dois parâmetros desconhecidos: α e β

Para cada par de valores $(\alpha, \beta) \in \mathbb{R}^2$

teríamos o valor esperado correspondente a cada valor observado X_i :

$$Y_i(\alpha, \beta) = \alpha + \beta X_i$$

e poderíamos calcular os erros correspondentes:

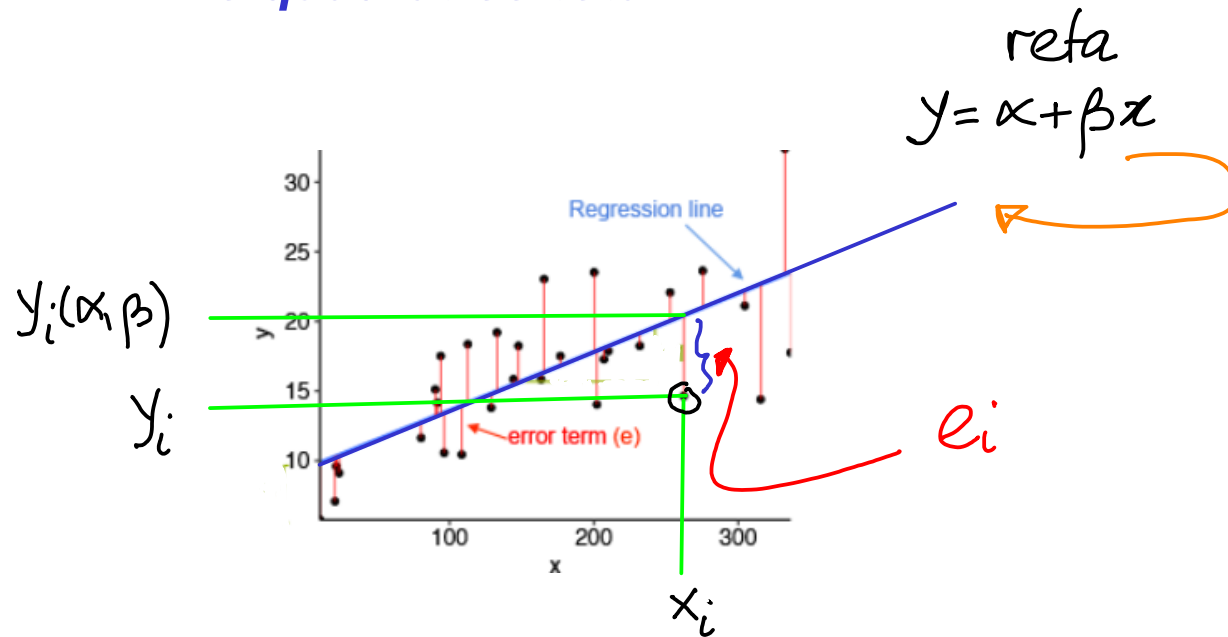
$$e_i = y_i - Y_i(\alpha, \beta) = y_i - (\alpha + \beta X_i), \quad i = 1, 2, \dots, n$$

valor
observado
de y

valor
esperado
("chute" se fossem
esses os parâmetros)

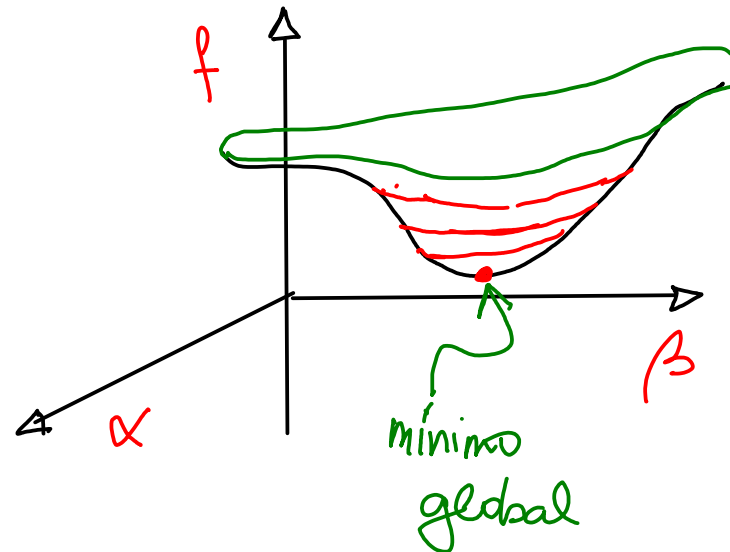
valor
observado
de X

Erro quadrático total:



$$SQ(\alpha, \beta) = \sum_{i=1}^n e_i^2$$

"erro quadrático total"



Encontrar os valores de α e β que minimizam $SQ(\alpha, \beta)$

$\hat{\alpha}$ $\hat{\beta}$ = "estimadores para α e β "

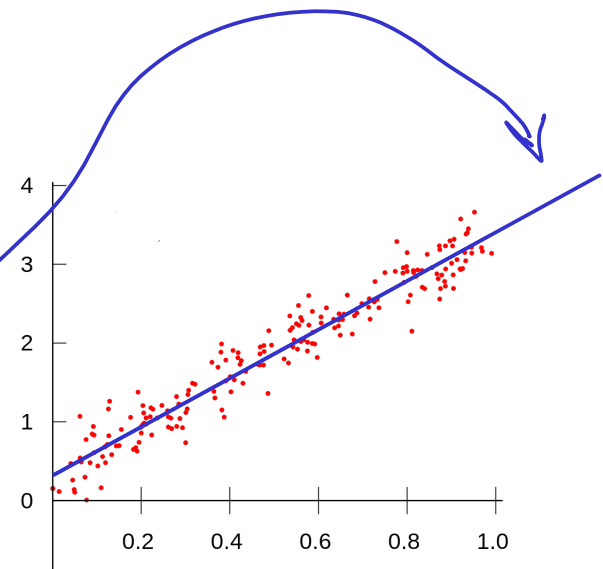
e' um problema simples de calculo. Encontramos:

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

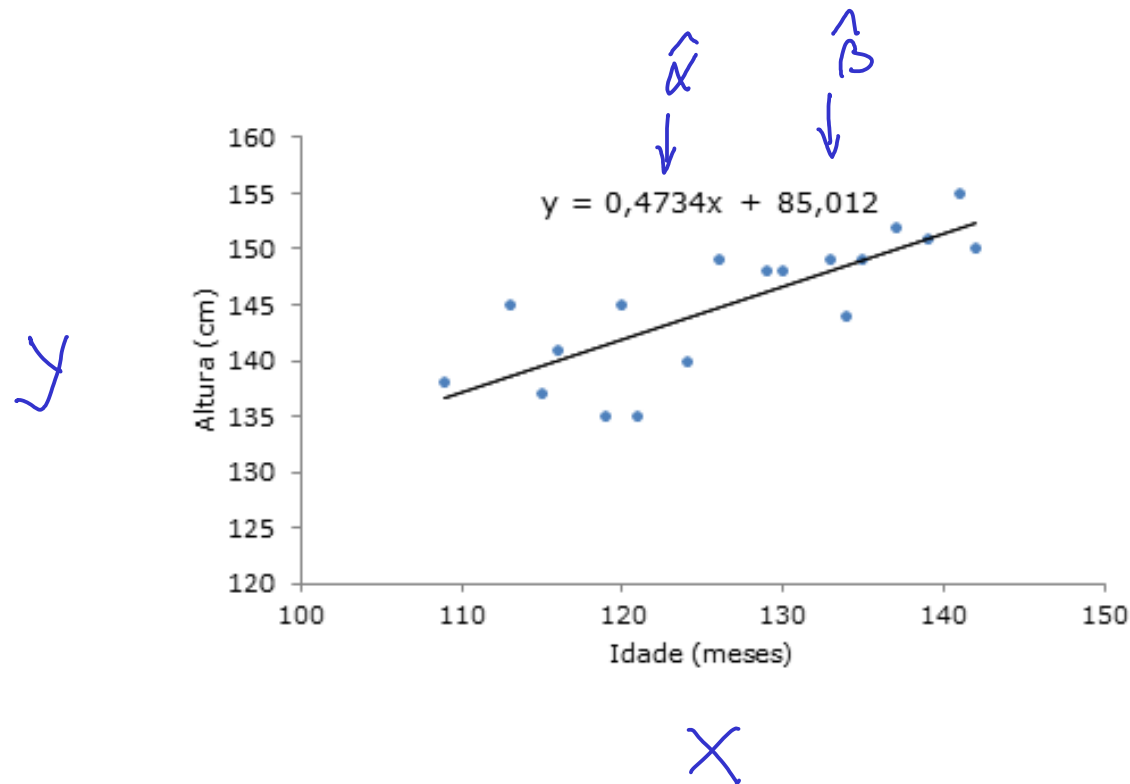
$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

"reta de regressao"



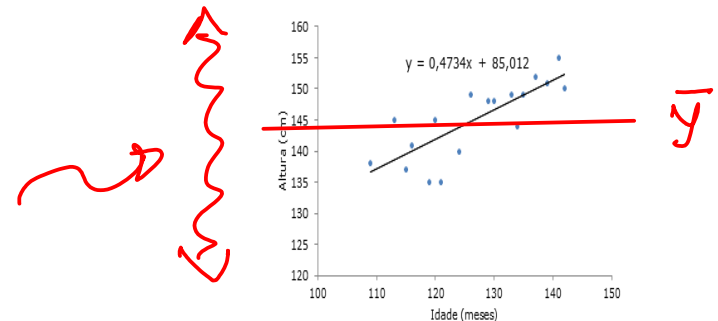
Avaliacao do modelo:

"Parece razoavel que Y depende linearmente de X ?"



"Variabilidade observada em Y":

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



"Erro do modelo"

erro quadrático entre o estimador
e o observado

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

↑
"resíduo"

= "parte da variabilidade de y que o modelo NÃO explicou"

"Erro explicado pelo modelo":

"soma devida à regressão,
ao modelo"

"Variabilidade observada em Y " se o modelo fosse "verdadeiro"

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

↑
Seriam os valores
observados para $X_1, X_2, \dots, X_n$

Temos 3 somas de quadrados:

$$SQT = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n a_i^2$$

$$a_i = y_i - \bar{y}$$

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

$$b_i = y_i - \hat{y}_i$$

$$SQ_{Reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n c_i^2$$

$$c_i = \hat{y}_i - \bar{y}$$

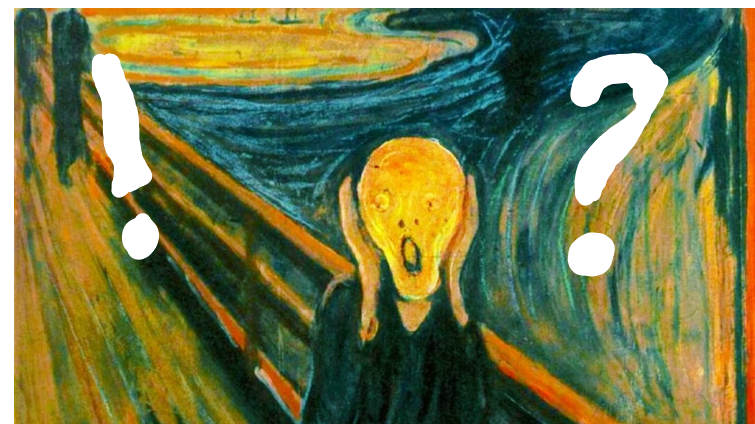


$$a_i = b_i + c_i, \quad i=1, \dots, n$$

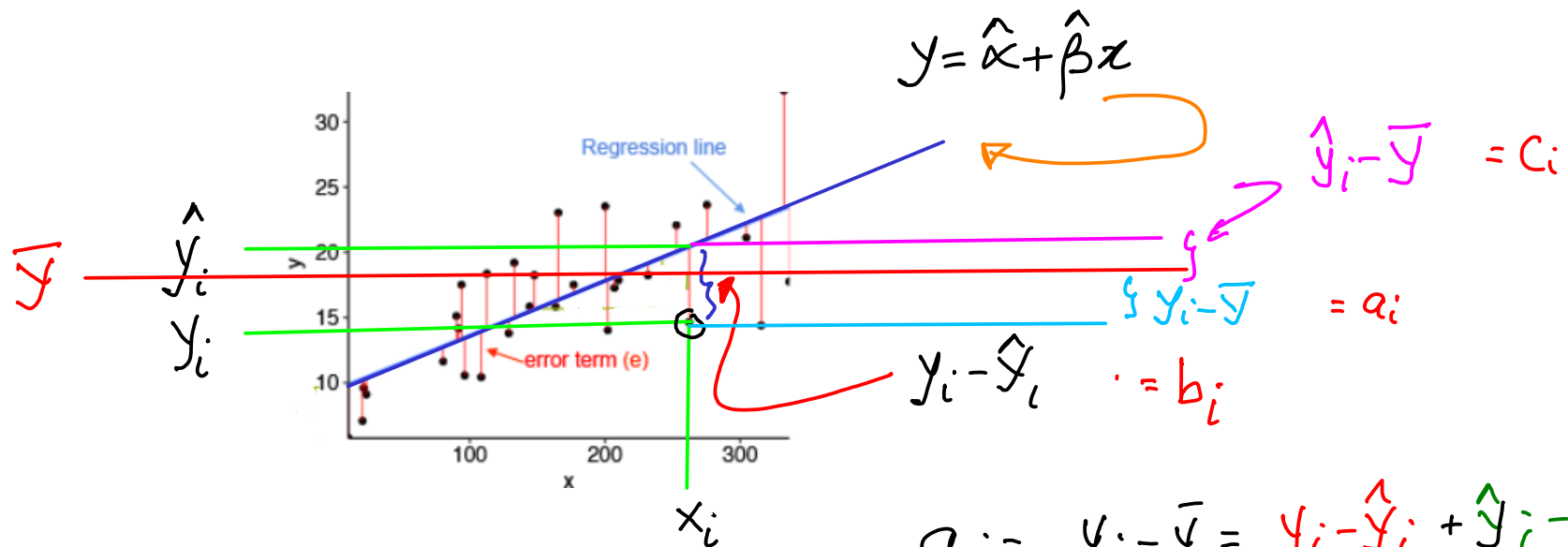
$$\sum_{i=1}^n a_i = \sum_{i=1}^n b_i + \sum_{i=1}^n c_i$$



$$\sum a_i^2 = \sum b_i^2 + \sum c_i^2$$



De novo:



$$a_i = y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y} = b_i + c_i, \quad i=1, \dots, n$$

$$\sum_{i=1}^n a_i = \sum_{i=1}^n b_i + \sum_{i=1}^n c_i$$



$$\sum a_i^2 = \sum b_i^2 + \sum c_i^2$$





Ops, aqui isso vale ...
**pelo teorema
de Pitagoras!**

valores observados
de y

$$y = (y_1, y_2, \dots, y_n)$$

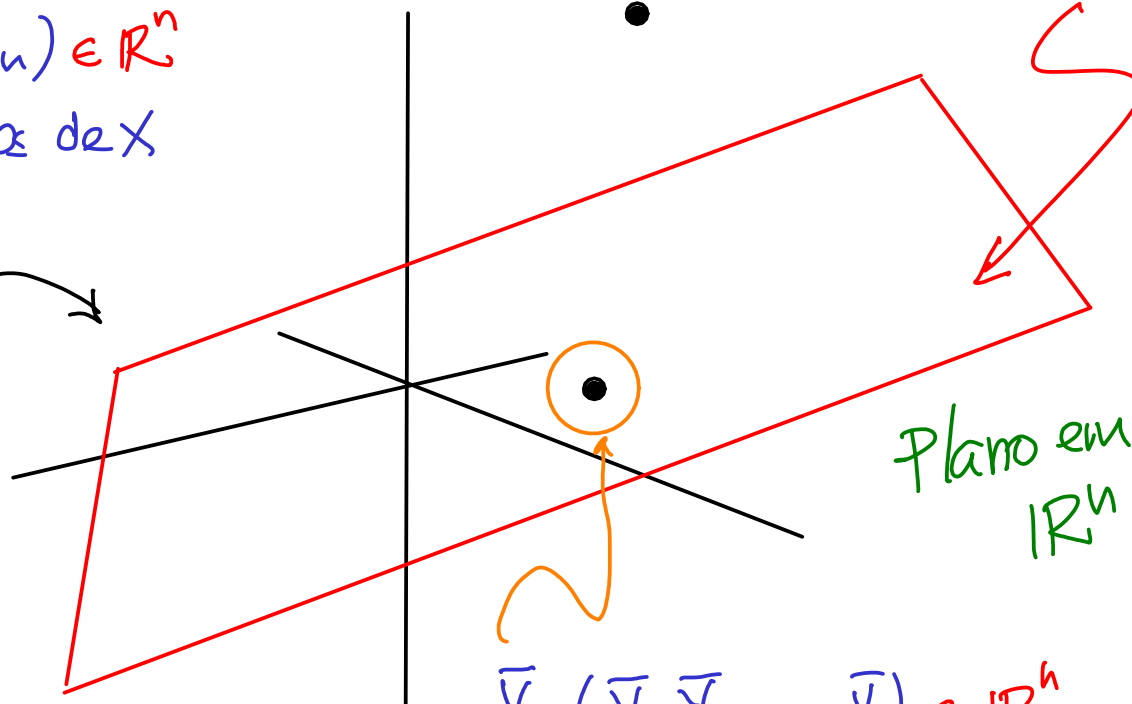
Seja

$$X = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$$

valores observados de x

fixo!

define um
plano em $\mathbb{R}^n$



Coleção de pontos
 $y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$

com

$$y_i = \alpha + \beta x_i$$

para algum par

$$(\alpha, \beta) \in \mathbb{R}^2$$

Plano em
 $\mathbb{R}^n$

$$\bar{y} = (\bar{y}, \bar{y}, \dots, \bar{y}) \in \mathbb{R}^n$$

n termos iguais

ok!: $\bar{y} \in \text{Plano}$

Abuso de
notação...

(tome
 $\alpha = \bar{y} \in \mathbb{R}$
 $\beta = 0$)

Para todo par de valores (α, β) temos $y(\alpha, \beta)$ em algum lugar no plano

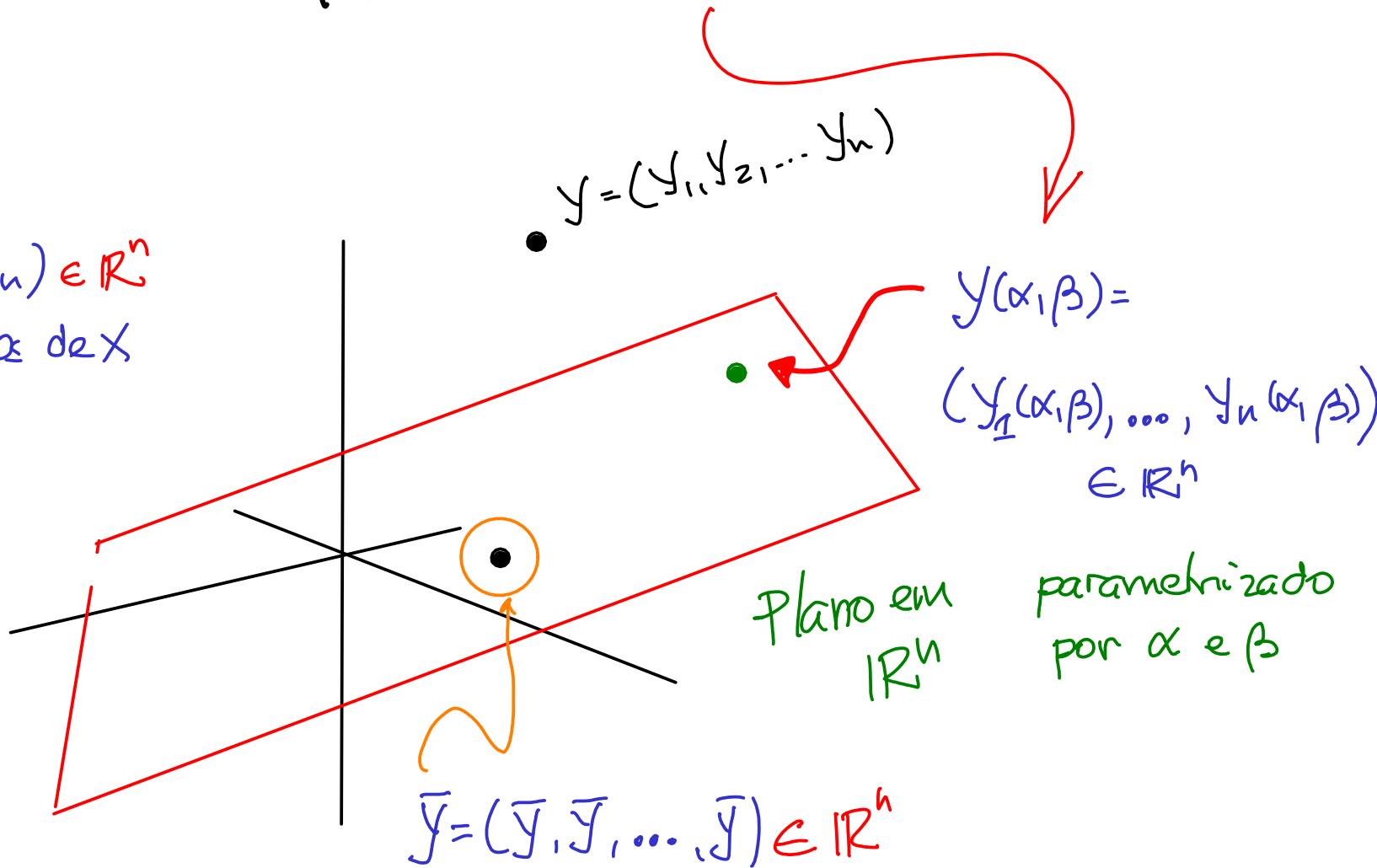
Seja

$$X = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$$

valores observados de X

fixo!

define
o plano



Para o par de valores (α, β) o vetor $y(\alpha, \beta)$ "mais perto" de Y

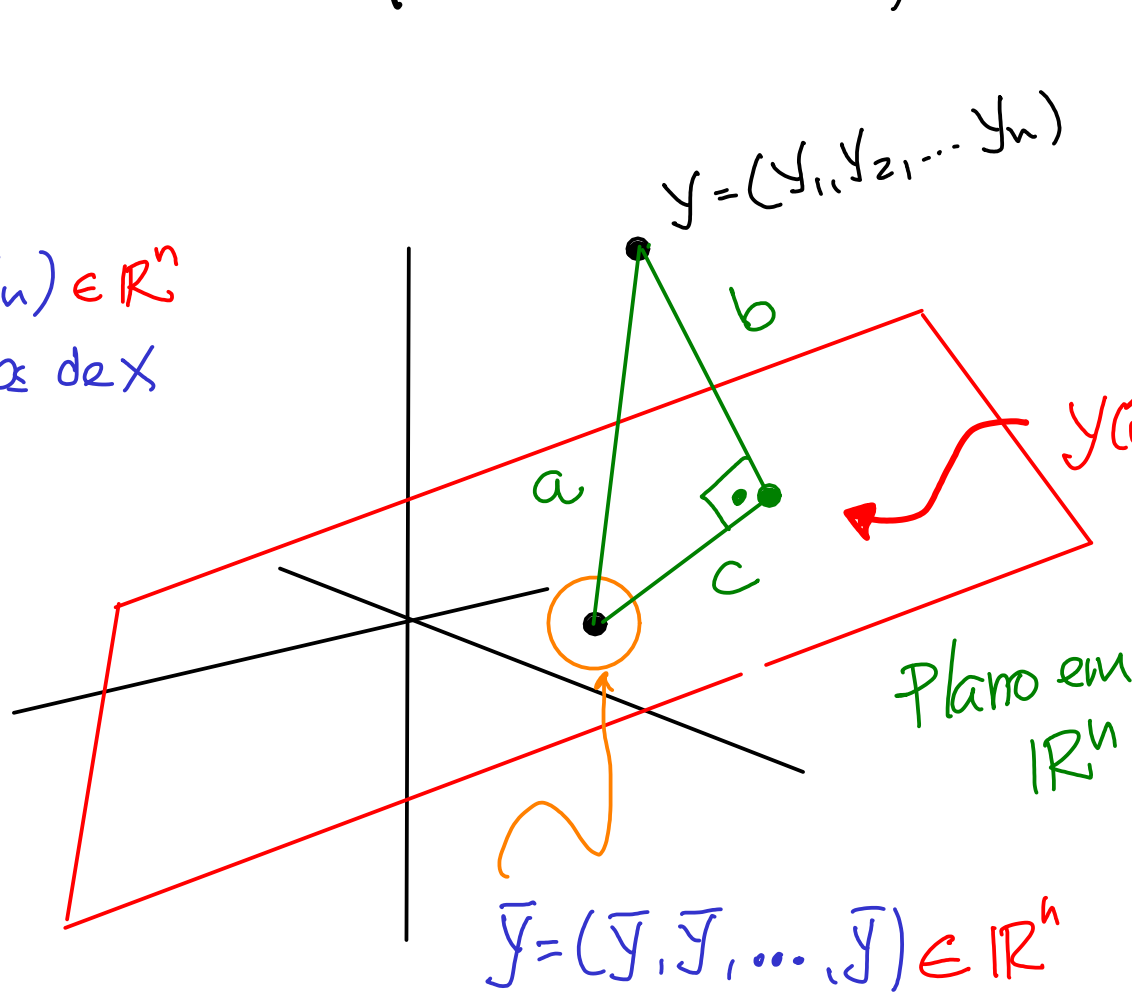
Seja

$$X = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$$

valores observados de X

fixo!

define
o plano



$$y(\hat{\alpha}, \hat{\beta}) = \hat{y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$$

parametrizado
por α e β

$$|a|^2 = \sum (y_i - \bar{y})^2$$

$$|b|^2 = \sum (y_i - \hat{y}_i)^2$$

$$|c|^2 = \sum (\hat{y}_i - \bar{y})^2$$

Entao temos

$$SQ_{Tot} = SQ_{Reg} + SQ_{Res}$$

variabilidade de y

variabilidade explicada pelo modelo

"Resíduo"

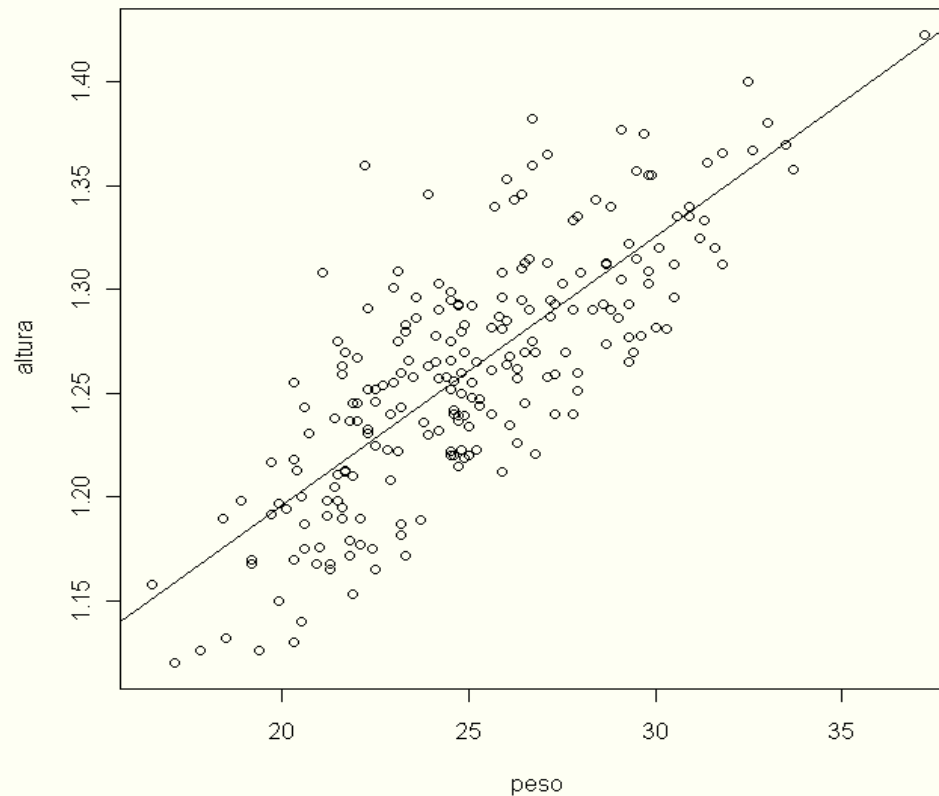
Parte não explicada da variabilidade

O modelo sera' bom se explicou a variabilidade observada em Y!

$$r^2 = \frac{SQ_{Reg}}{SQ_{Tot}} \in [0,1] \quad \text{deve ser grande!}$$

Coeficiente de correlação linear

É uma medida que avalia o quanto a “nuvem de pontos” no diagrama de dispersão aproxima-se de uma reta.



O coeficiente de correlação linear de Pearson é dado por

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y},$$

$\bar{X}$ e $\bar{Y}$ são as médias de X e Y, respectivamente;

S_X e S_Y são os desvios padrão de X e Y, respectivamente.

Fórmula alternativa

$$r = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{(n-1) S_X S_Y}$$

$$S^2 = \frac{\sum_{i=1}^n X_i^2 - n \bar{X}^2}{(n-1)}$$

Voltando ao Exemplo 2

Tempo (X)	Nota (Y)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
3,0	4,5	-2,1	-1,1	2,31
7,0	6,5	1,9	0,9	1,71
2,0	3,7	-3,1	-1,9	5,89
1,5	4,0	-3,6	-1,6	5,76
12,0	9,3	6,9	3,7	25,53
25,5	28,0	0	0	41,2
$\bar{X} = 5,1$	$\bar{Y} = 5,6$			

$$S_x^2 = \frac{(-2,1)^2 + \dots + (6,9)^2}{4} = \frac{78,2}{4} = 19,55 \Rightarrow S_x = 4,42$$

$$S_y^2 = \frac{(-1,1)^2 + \dots + (3,7)^2}{4} = \frac{21,9}{4} = 5,47 \Rightarrow S_y = 2,34$$

Então,

$$r = \frac{41,2}{4 \cdot 4,42 \cdot 2,34} = 0,9959$$

Propriedade: $-1 \leq r \leq 1$

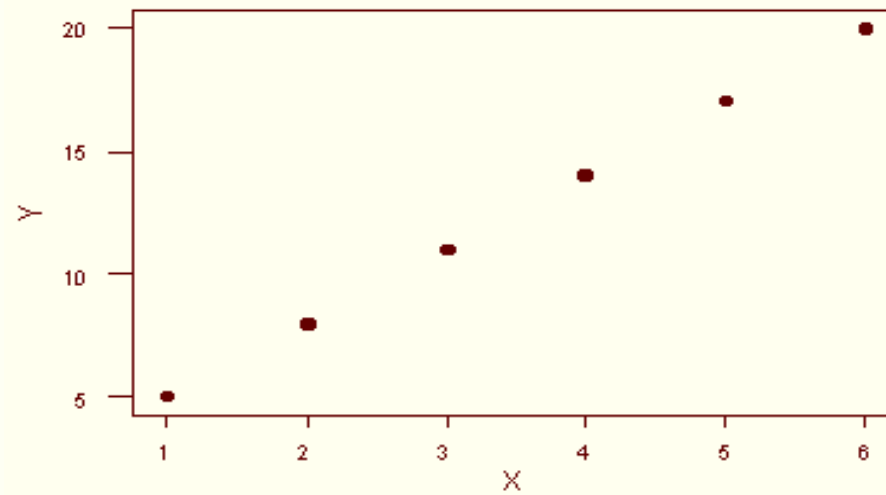
Casos particulares

$r = 1 \Rightarrow$ correlação linear positiva e perfeita;

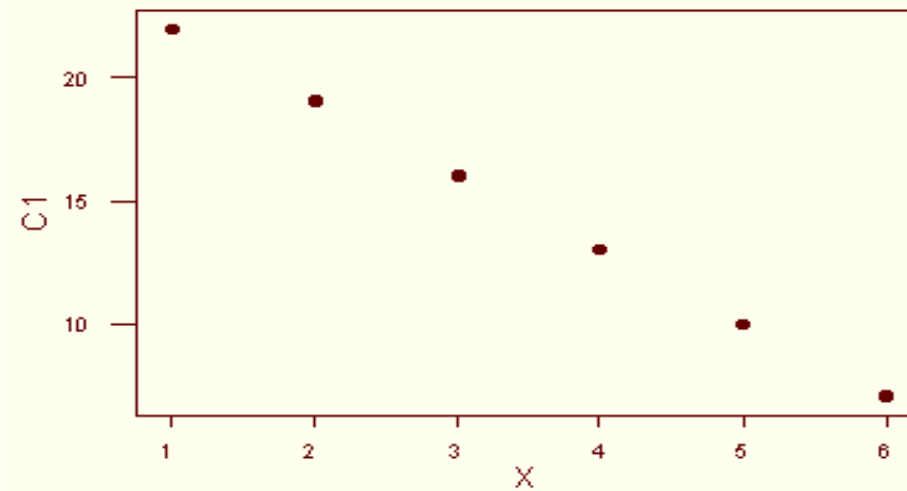
$r = -1 \Rightarrow$ correlação linear negativa e perfeita;

$r = 0 \Rightarrow$ inexistência de correlação linear.

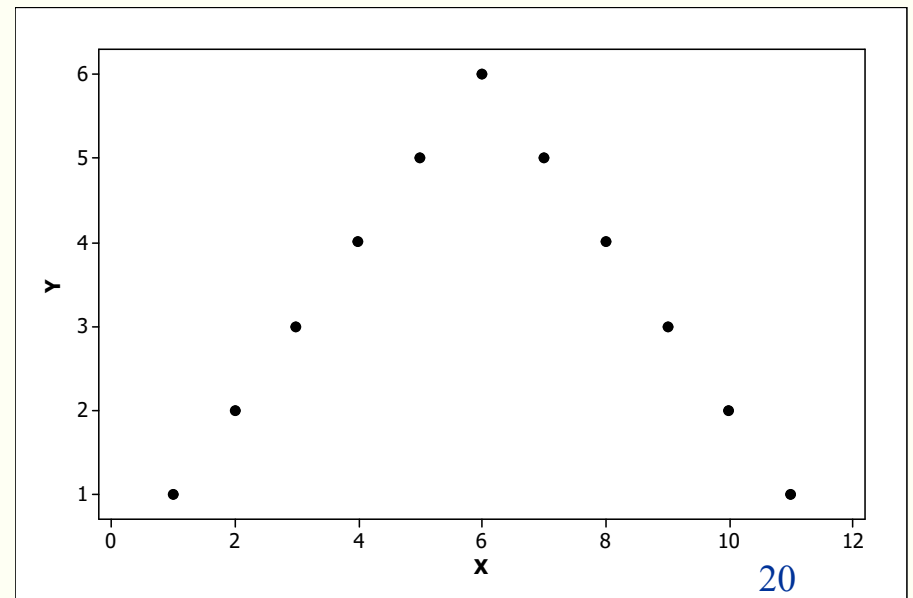
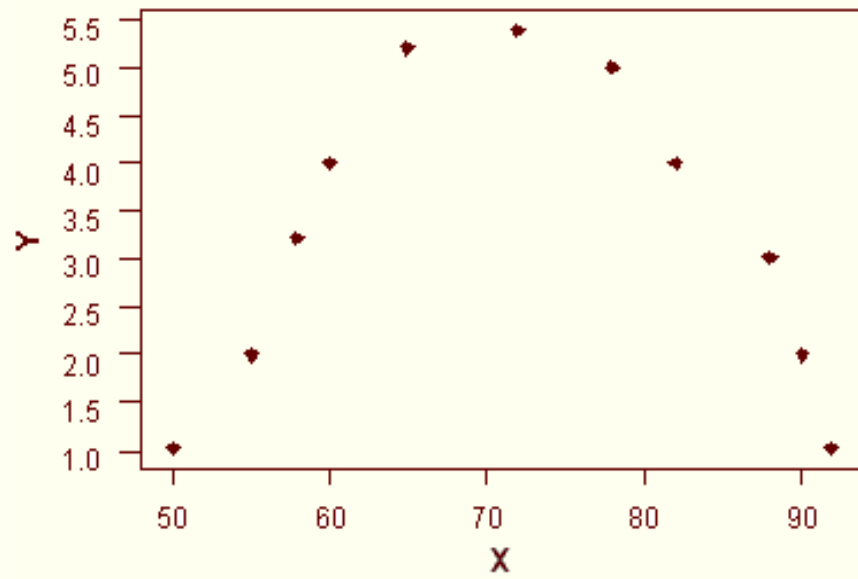
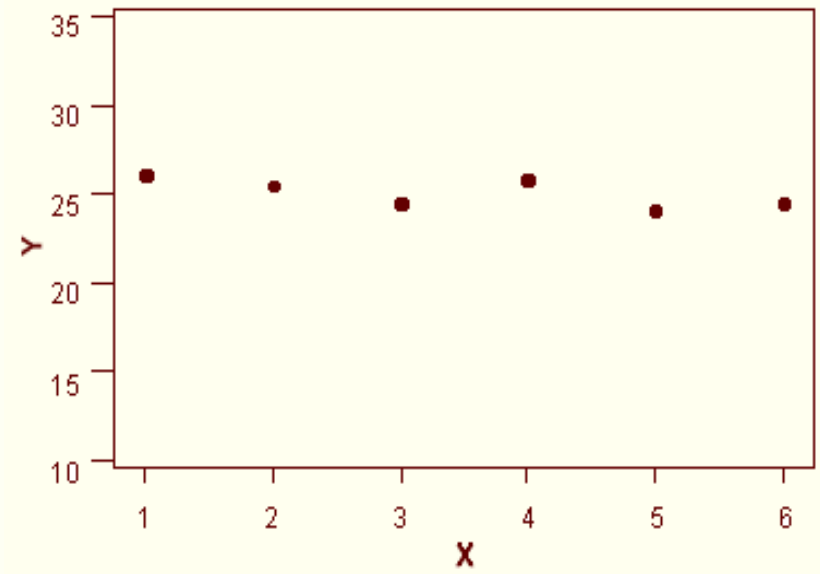
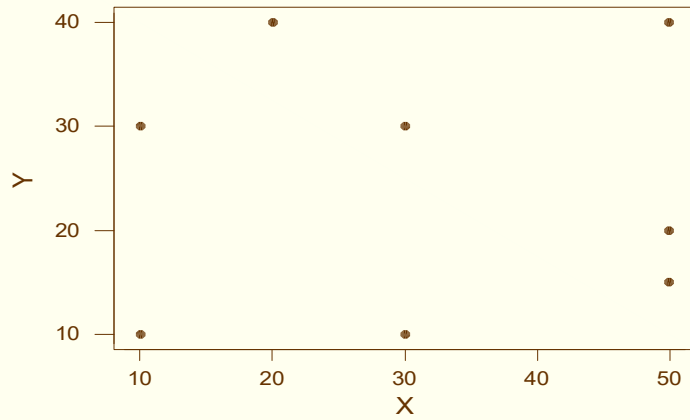
$r = 1$, correlação linear positiva e perfeita



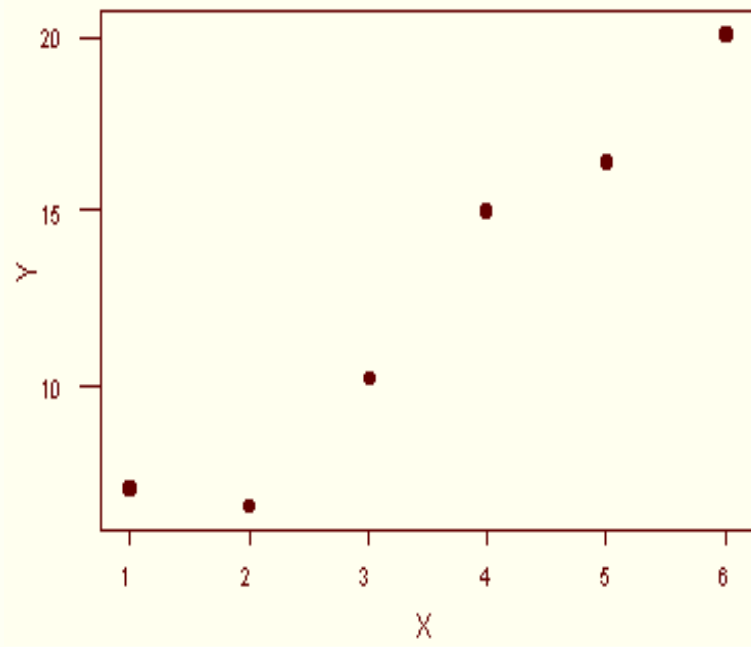
$r = -1$, correlação linear negativa e perfeita



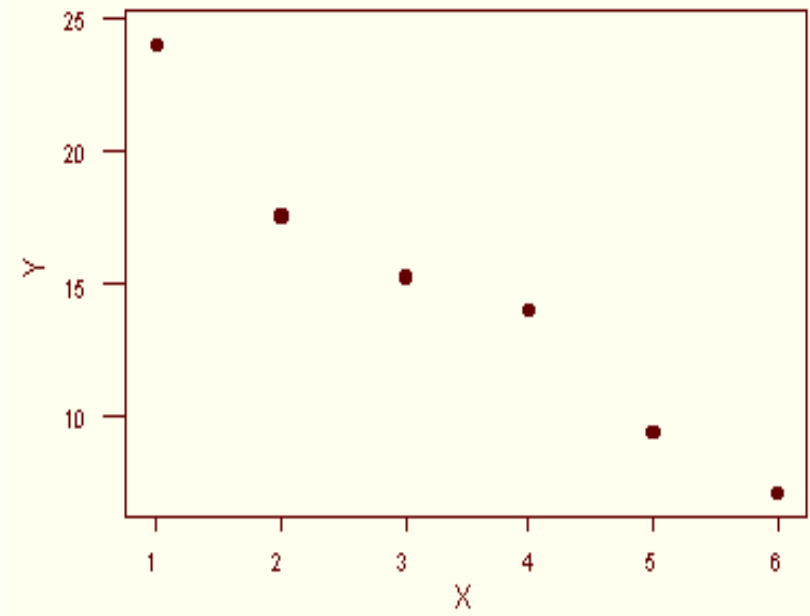
$$r \cong 0$$



$$r \cong 1$$



$$r \cong -1$$



Exemplo 3: criminalidade e analfabetismo

Considere duas variáveis observadas em 50 estados norte-americanos:

Y: taxa de criminalidade

X: taxa de analfabetismo

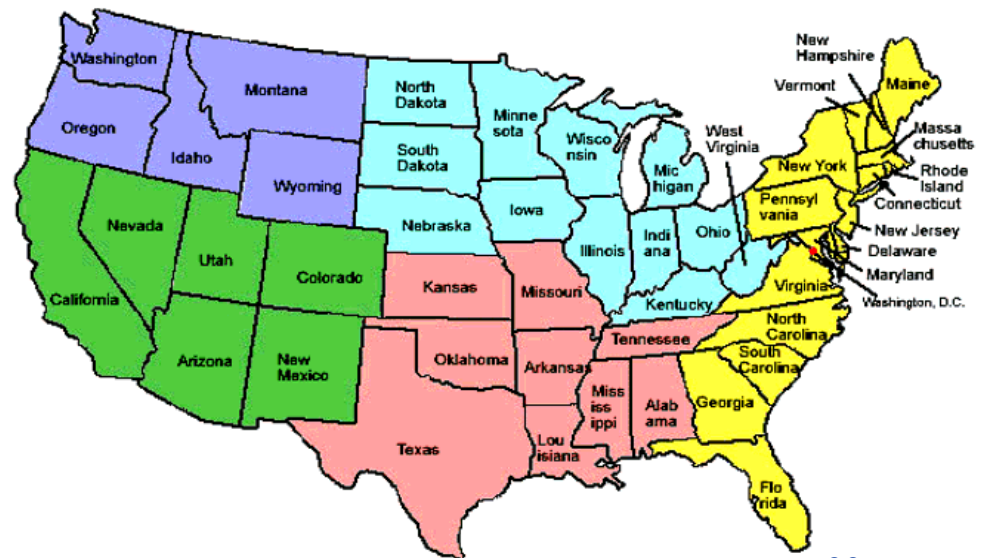
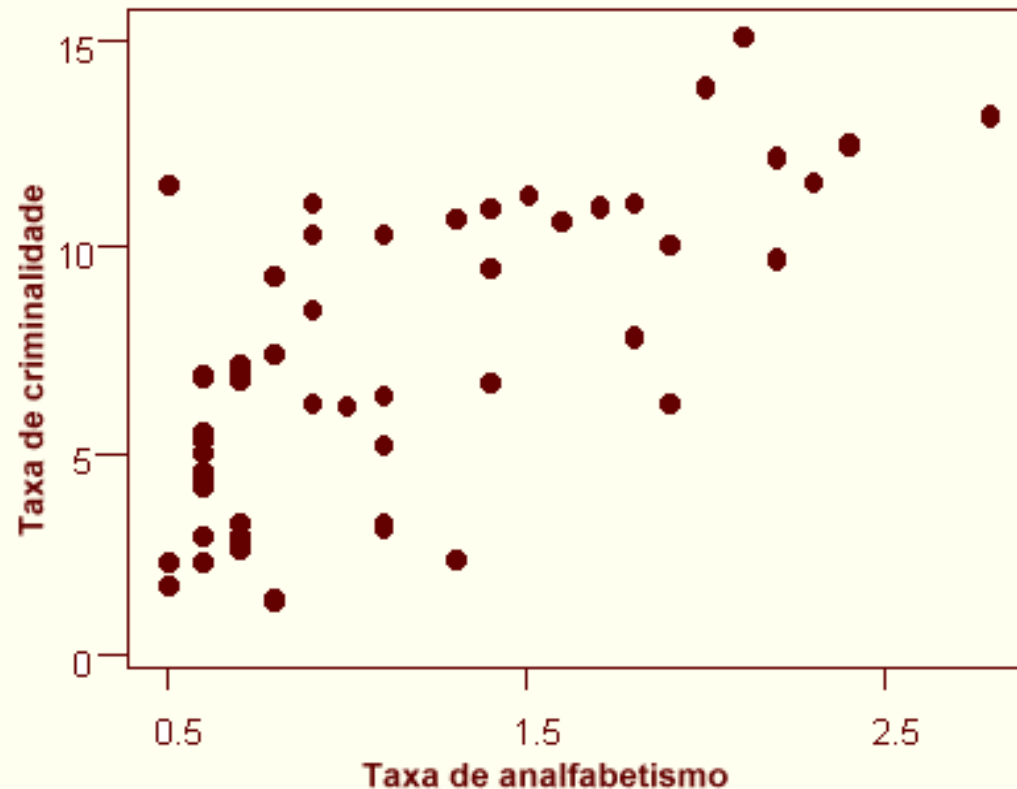


Diagrama de dispersão



Podemos notar que, conforme aumenta a taxa de analfabetismo (X), a taxa de criminalidade (Y) tende a aumentar. Nota-se também uma tendência linear.

Cálculo da correlação

$\bar{Y} = 7,38$ (média de Y) e $S_Y = 3,692$ (desvio padrão de Y)

$\bar{X} = 1,17$ (média de X) e $S_X = 0,609$ (desvio padrão de X)

$$\sum X_i Y_i = 509,12$$

Correlação entre X e Y:

$$r = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{(n-1) S_X S_Y}$$

$$r = \frac{509,12 - 50 \cdot 7,38 \cdot 1,17}{49 \cdot 3,692 \cdot 0,609} = \frac{77,39}{110,17} = 0,702$$

Exemplo 4: expectativa de vida e analfabetismo

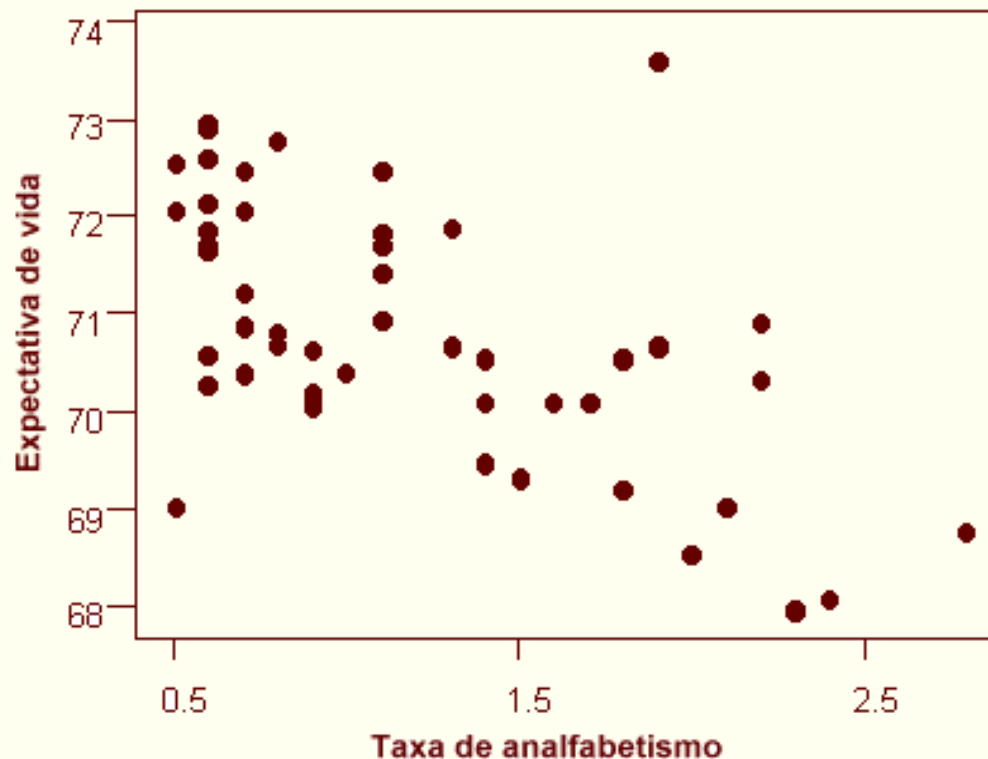
Considere duas variáveis observadas em 50 estados norte-americanos:

Y : expectativa de vida

X : taxa de analfabetismo



Diagrama de Dispersão



Podemos notar que, conforme aumenta a taxa de analfabetismo (X), a expectativa de vida (Y) tende a diminuir. Nota-se também uma tendência linear.

Cálculo da correlação

$\bar{Y} = 70,88$ (média de Y) e $S_Y = 1,342$ (desvio padrão de Y)

$\bar{X} = 1,17$ (média de X) e $S_X = 0,609$ (desvio padrão de X)

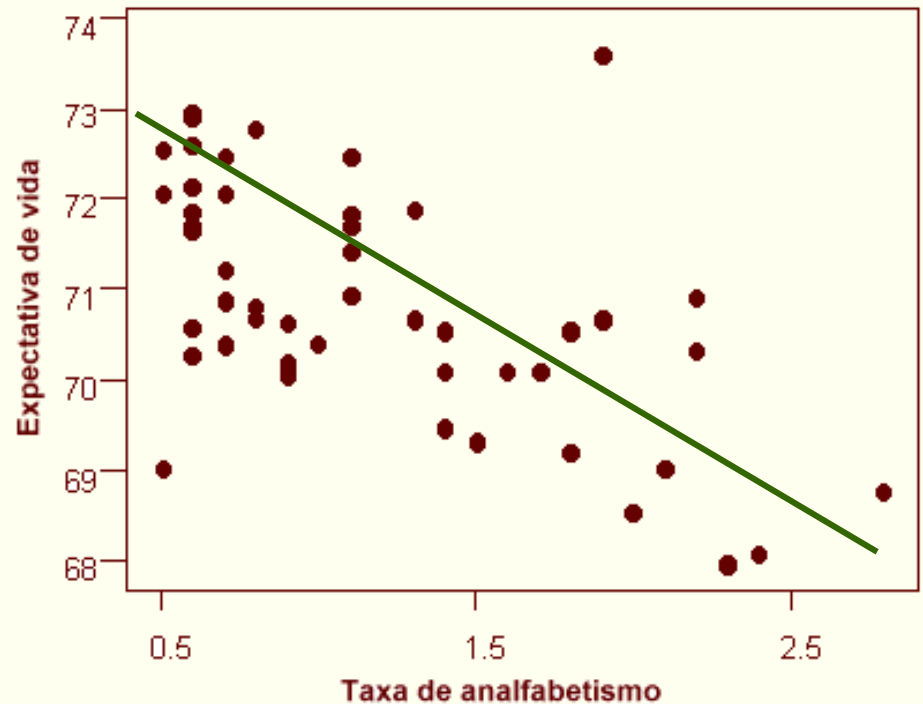
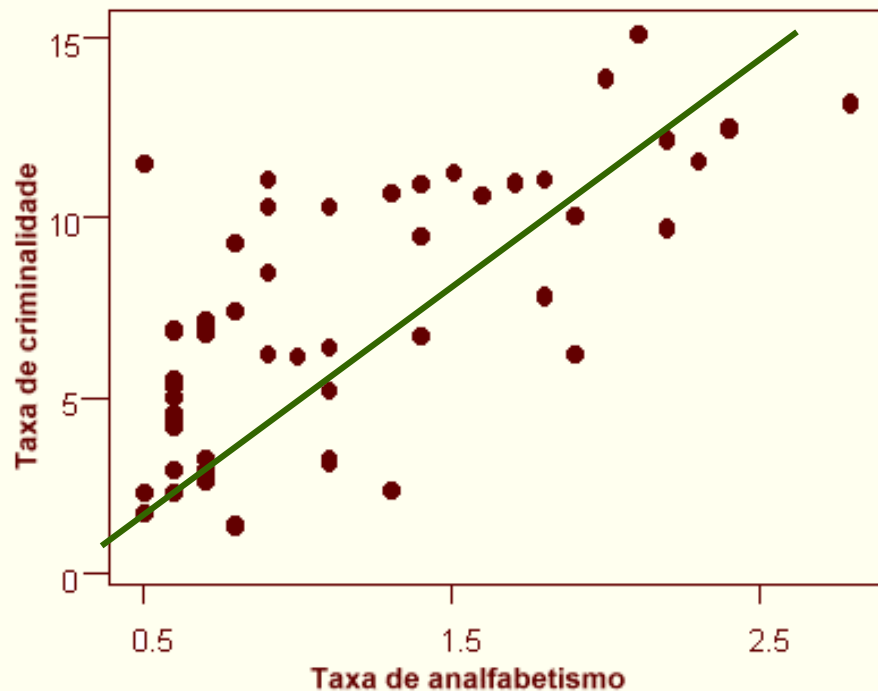
$$\sum X_i Y_i = 4122,8$$

Correlação entre X e Y:

$$r = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{(n-1) S_X S_Y}$$
$$r = \frac{4122,8 - 50 \cdot 70,88 \cdot 1,17}{49 \cdot 1,342 \cdot 0,609} = \frac{-23,68}{40,047} = -0,59$$

Análise de Regressão

Diagramas de Dispersão



⇒ Explicar a forma da relação por meio de uma função matemática: $Y = a + bX$

Análise de Regressão

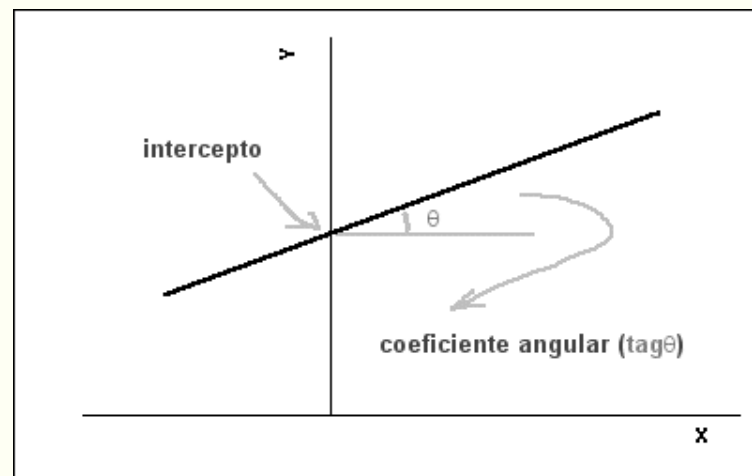
Reta ajustada:

$$\hat{Y} = a + bX$$

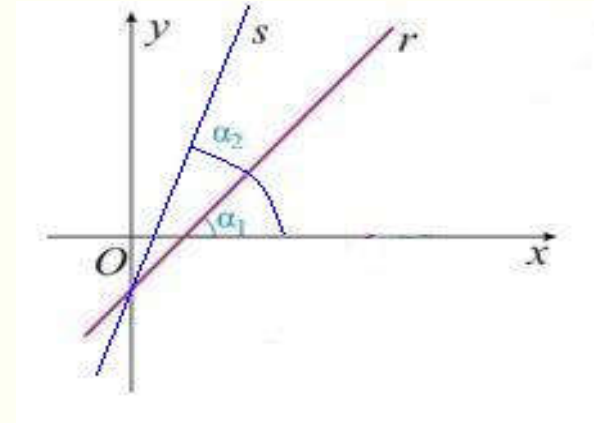
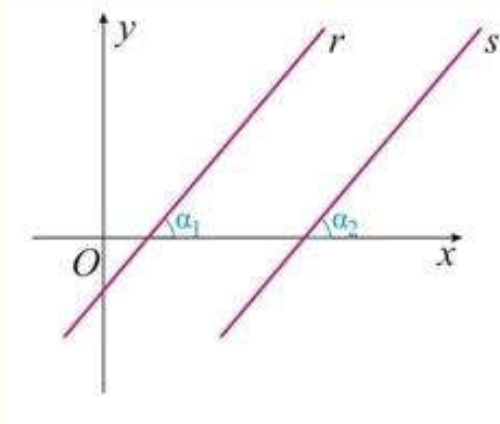
O que são ***a*** e ***b***?

a : intercepto

b : inclinação ou coeficiente angular



Análise de Regressão



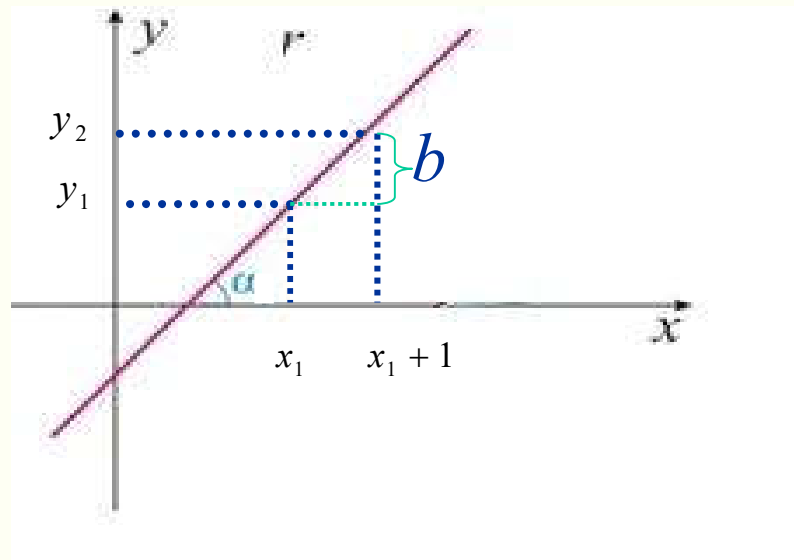
- Iguais coeficientes angulares
- Diferentes interceptos
- Diferentes coeficientes angulares
- Iguais interceptos

Reta ajustada:

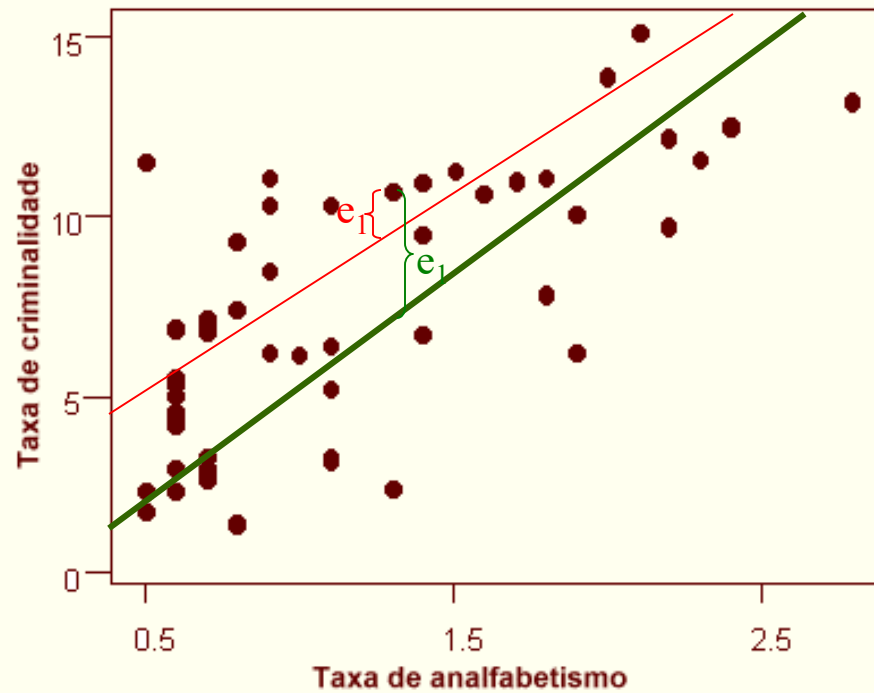
$$\hat{Y} = a + bX$$

Interpretação de b : Para cada aumento de uma unidade em X , temos um aumento médio de b unidades em Y .

$$\begin{aligned} \text{tag}(\alpha) &= \frac{y_2 - y_1}{x_2 - x_1} = \frac{y_2 - y_1}{x_1 + 1 - x_1} \\ &= y_2 - y_1 = b \end{aligned}$$



Reta ajustada (método de mínimos quadrados)



Reta ajustada (método de mínimos quadrados)

Os coeficientes a e b são calculados da seguinte maneira:

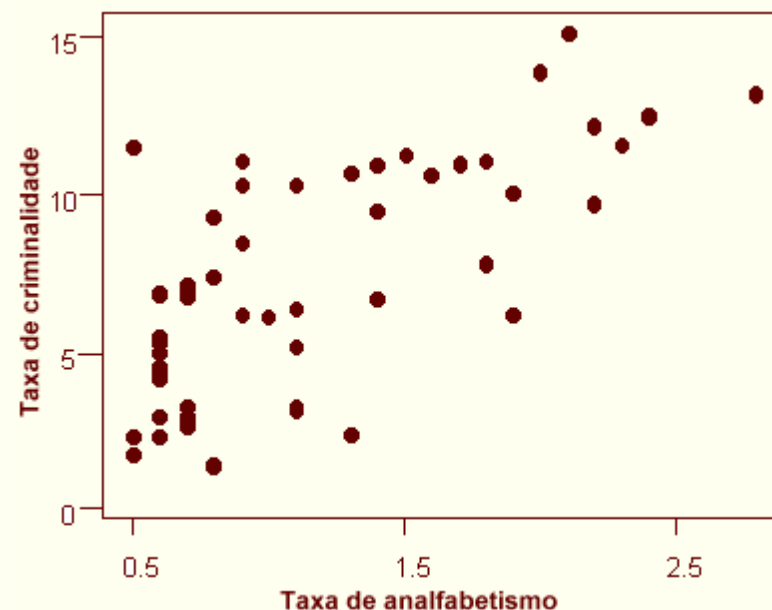
$$b = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{(n-1) S_X^2}$$

$$a = \bar{Y} - b \bar{X}$$

No Exemplo 3,

A reta ajustada fica dada por

$$\hat{Y} = 2,397 + 4,257 X$$



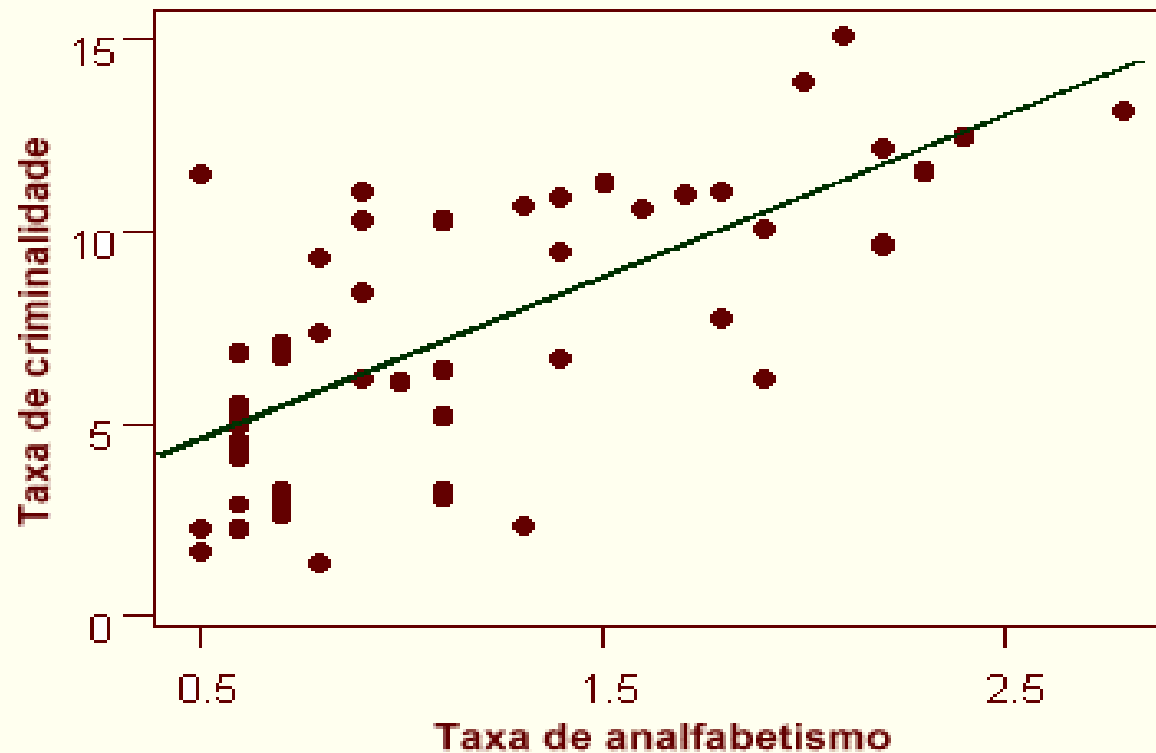
$\hat{Y}$: valor predito para a taxa de criminalidade

X : taxa de analfabetismo

Interpretação de b : Para um aumento de uma unidade na taxa do analfabetismo (X), a taxa de criminalidade (Y) aumenta, em média, 4,257 unidades.

Graficamente, temos

$$\hat{Y} = 2,397 + 4,257 X$$

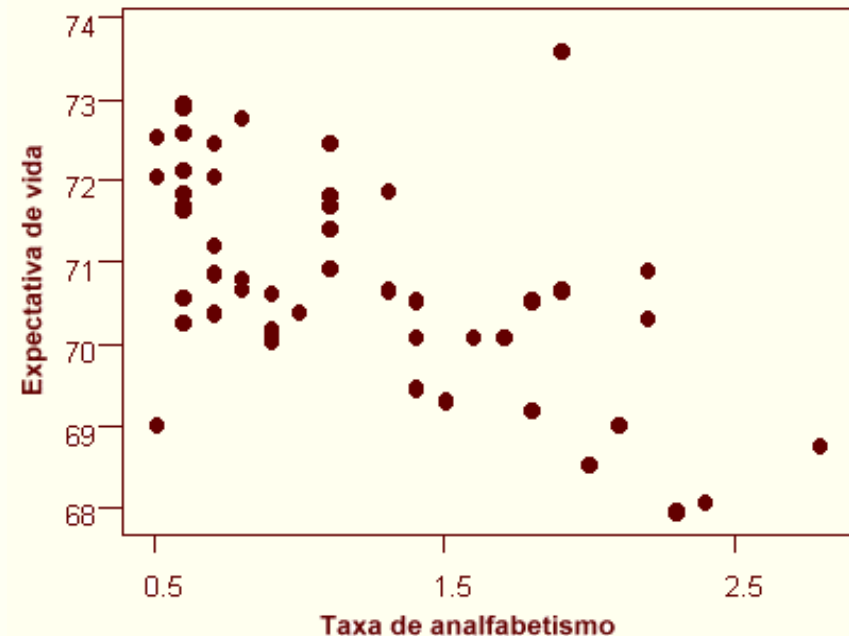


Como desenhar a reta no gráfico?

No Exemplo 4,

A reta ajustada fica dada por

$$\hat{Y} = 72,395 - 1,296 X$$



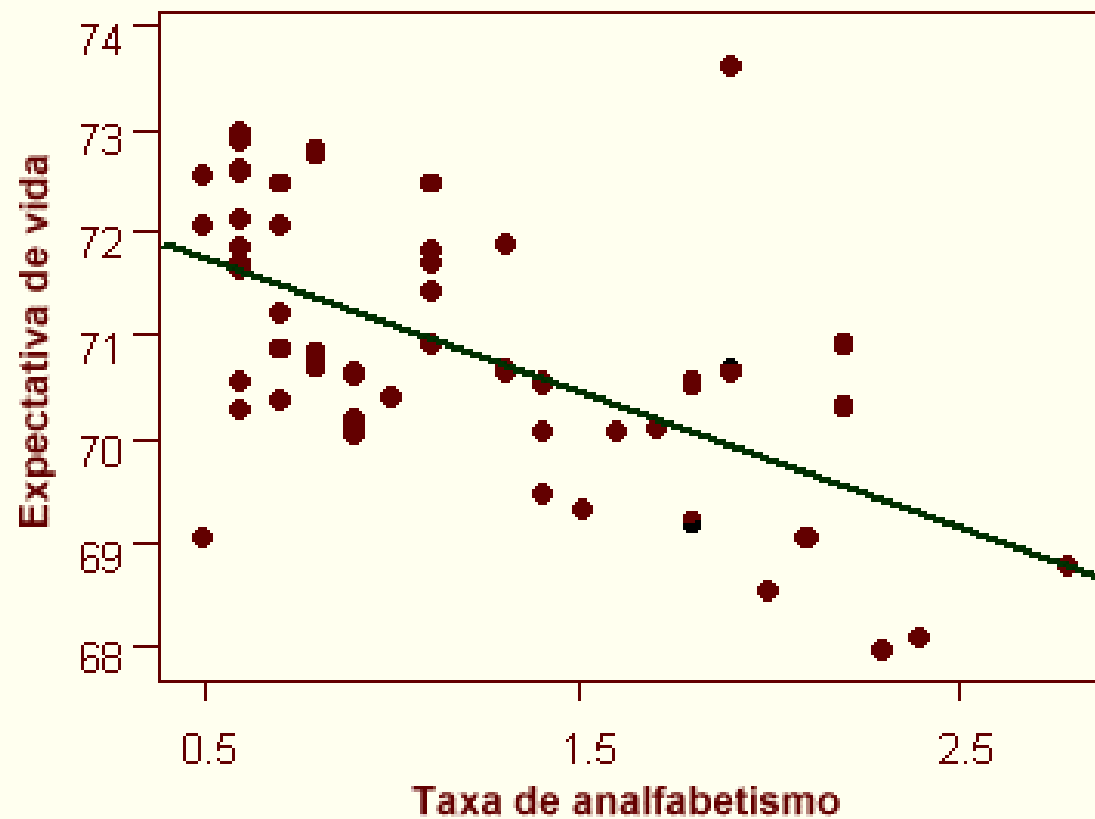
$\hat{Y}$: valor predito para a expectativa de vida

X : taxa de analfabetismo

Interpretação de b : Para um aumento de uma unidade na taxa do analfabetismo (X), a expectativa de vida (Y) diminui, em média, 1,296 anos.

Graficamente, temos

$$\hat{Y} = 72,395 - 1,296 X$$



Exemplo 5: consumo de cerveja e temperatura

**Y : consumo de cerveja diário por mil habitantes
(em litros);**

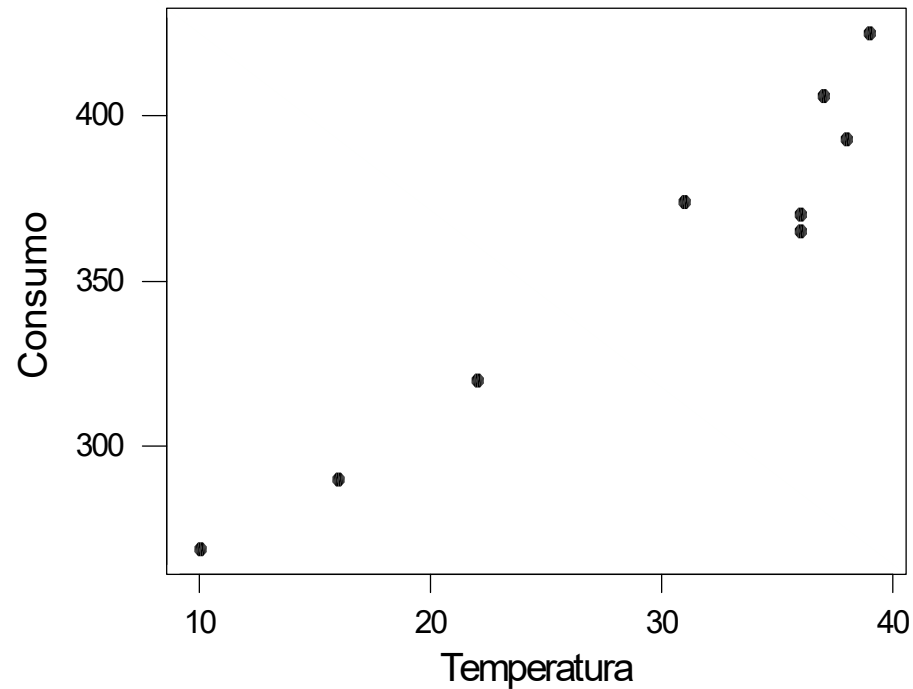
X : temperatura máxima (em $^{\circ}\text{C}$).

As variáveis foram observadas em nove localidades com as mesmas características demográficas e sócio-econômicas.

Exemplo 5: descrição dos dados

Localidade	Temperatura (X)	Consumo (Y)
1	16	290
2	31	374
3	38	393
4	39	425
5	37	406
6	36	370
7	36	365
8	22	320
9	10	269

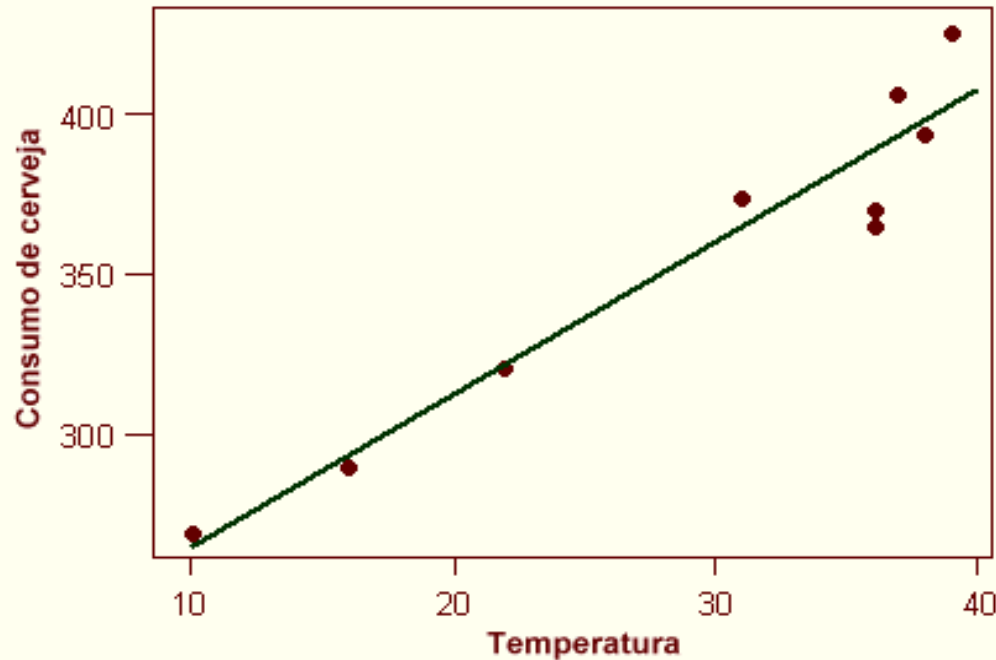
Diagrama de Dispersão



A correlação entre X e Y é $r = 0,962$.

A reta ajustada fica dada por:

$$\hat{Y} = 217,37 + 4,74 X$$



Qual é a interpretação de b ?

Aumentando-se um grau de temperatura (X), o consumo de cerveja (Y) aumenta, em média, 4,74 litros por mil habitantes.

Qual é o consumo previsto para uma temperatura de 25°C?

$$\hat{Y} = 217,37 + 4,74 (25) = 335,87 \text{ litros}$$