Correlação e Regressão

Correlação linear

Estudar a relação entre duas variáveis quantitativas

Ou seja, a força da relação entre elas, ou grau de associação linear.

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

Investigaremos a presença ou ausência de **relação linear** sob dois pontos de vista:

a) Quantificando a força dessa relação: correlação.
b) Explicitando a forma dessa relação: regressão.

Representação gráfica de duas variáveis quantitativas: **Diagrama de dispersão**

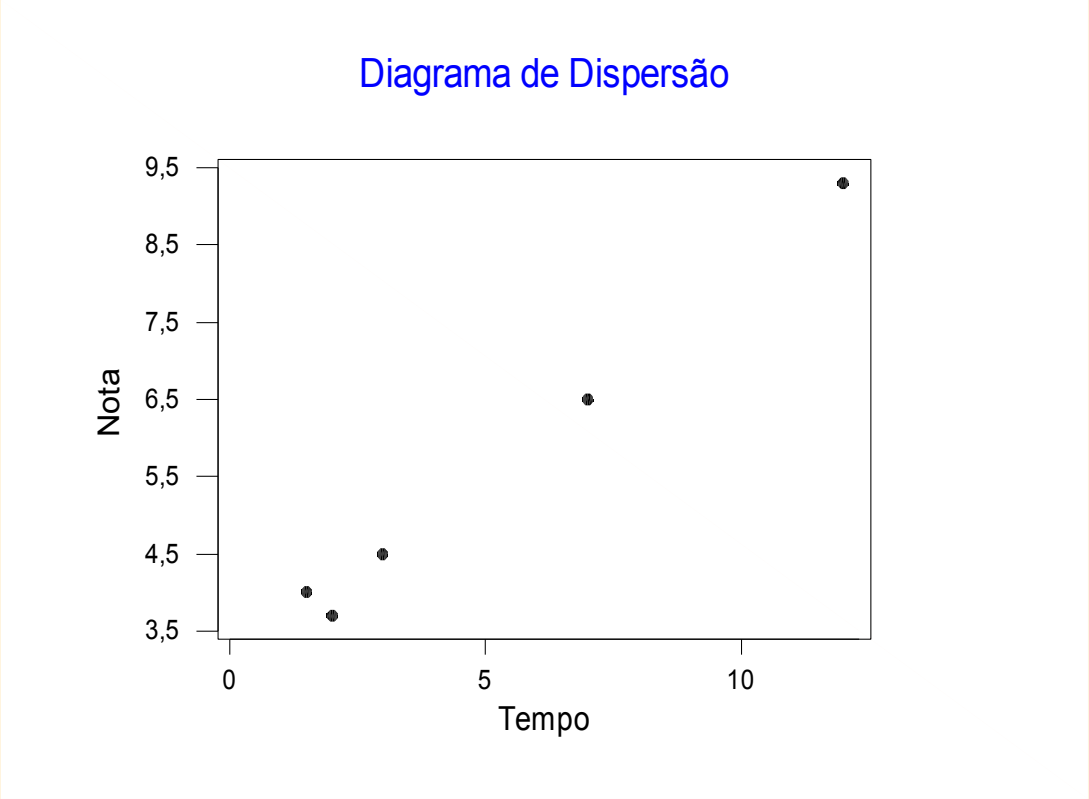
Exemplo 1: nota da prova e tempo de estudo

X : tempo de estudo (em horas)

Y : nota da prova

Pares de observações (X_i , Y_i) para cada estudante

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



Coeficiente de Correlação de Pearson.

- Definição.

Se os dados se referirem a:	
Uma População	Uma Amostra
$\rho = \frac{cov(X,Y)}{\sigma_X . \sigma_Y}$	$r = \frac{cov(X,Y)}{s_x . s_Y}$

Co-Variância
* Definição *

Uma População	$\text{cov}(X,Y) = \frac{\sum (x_i - \mu_x).(y_i - \mu_y)}{N}$
Uma Amostra	$\text{cov}(X,Y) = \frac{\sum (x_i - \bar{x}).(y_i - \bar{y})}{n-1}$

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:

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

sendo que,

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

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

No exemplo:

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

$\bar{X} = 5,1$ $\bar{Y} = 5,6$

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

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

Então,

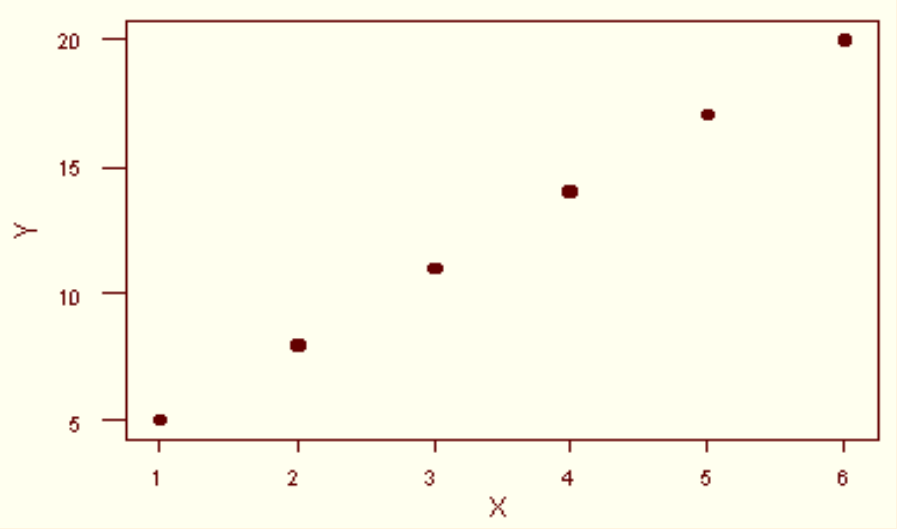
$$r = \frac{41,2}{4 \,.\, 4,42 \,.\, 2,34} = 0,9959$$

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

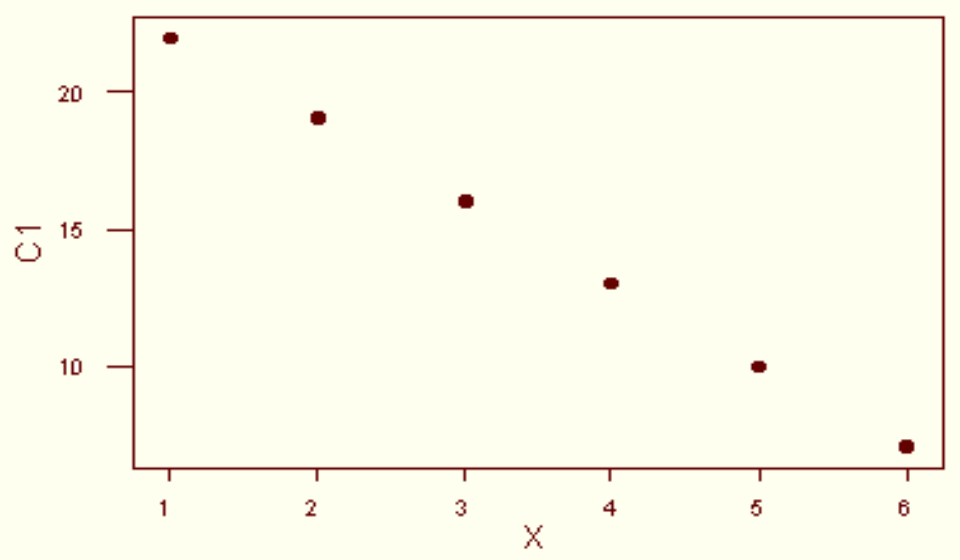
- **Casos particulares:**
-
- **$r = 1 \Rightarrow$ correlação linear positiva e perfeita**
- **$r = -1 \Rightarrow$ correlação linear negativa e perfeita**
- **$r = 0 \Rightarrow$ inexistência de correlação linear**

- **Correlação forte**
- **Correlação moderada**
- **Correlação fraca**

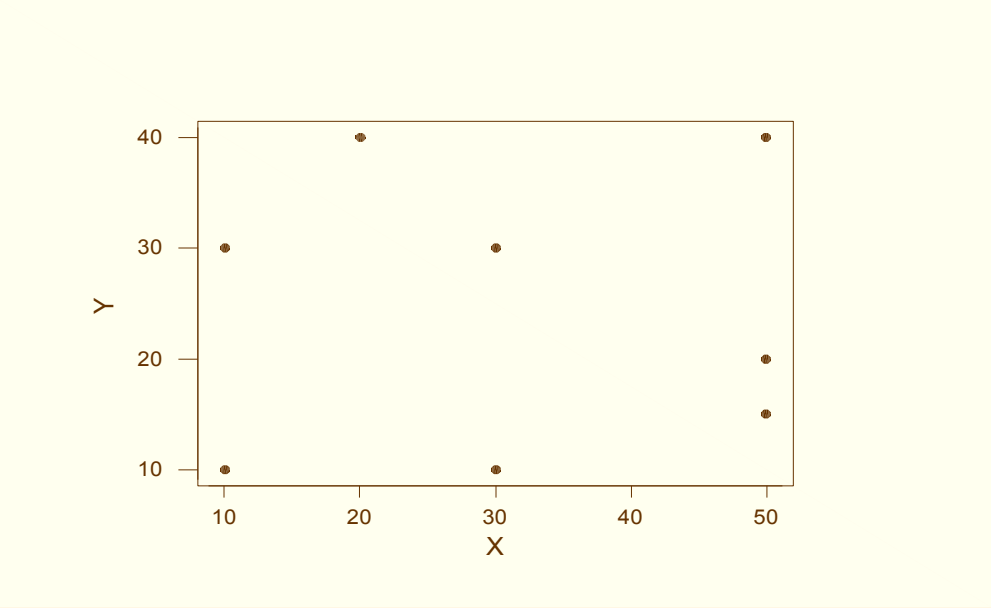
r = 1, correlação linear positiva e perfeita

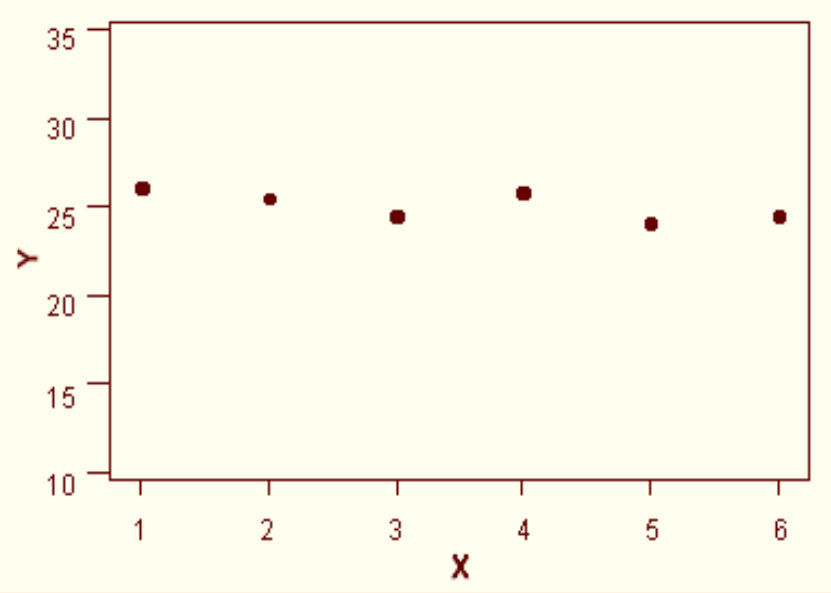


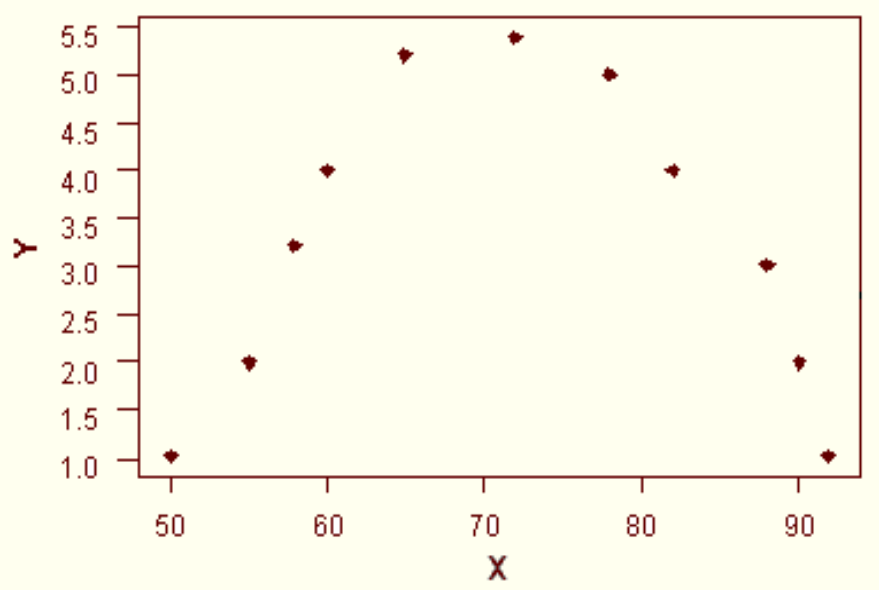
r = -1, correlação linear negativa e perfeita



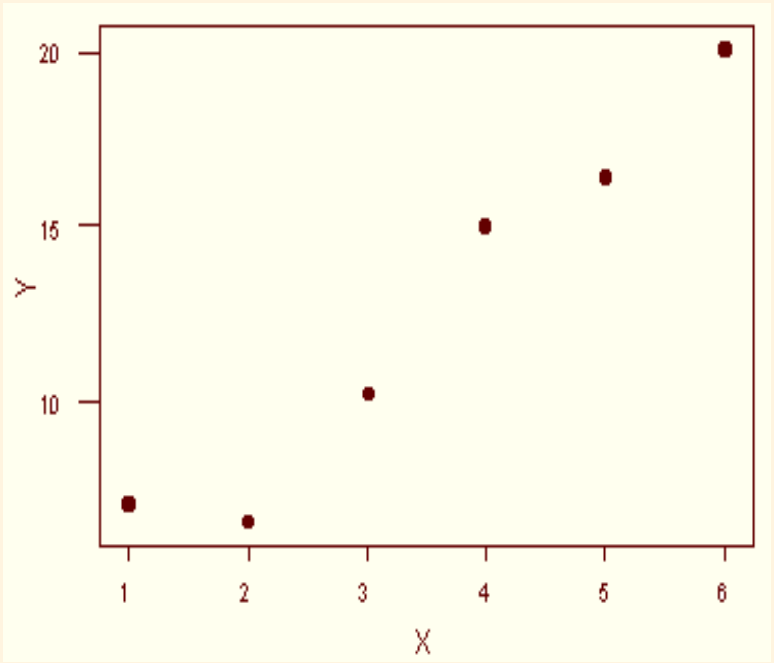
$r \cong 0$


40
30
20
10
Y
10
20
30
40
50
X

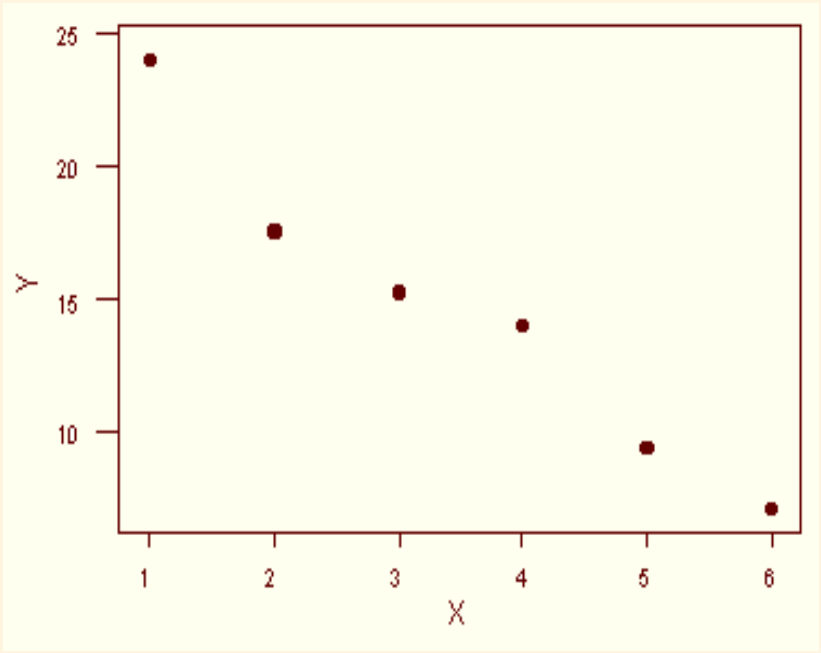

35
30
25
20
15
10
Y
1
2
3
4
5
6
X


5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Y
50
60
70
80
90
X

$r \cong 1$



$r \cong -1$



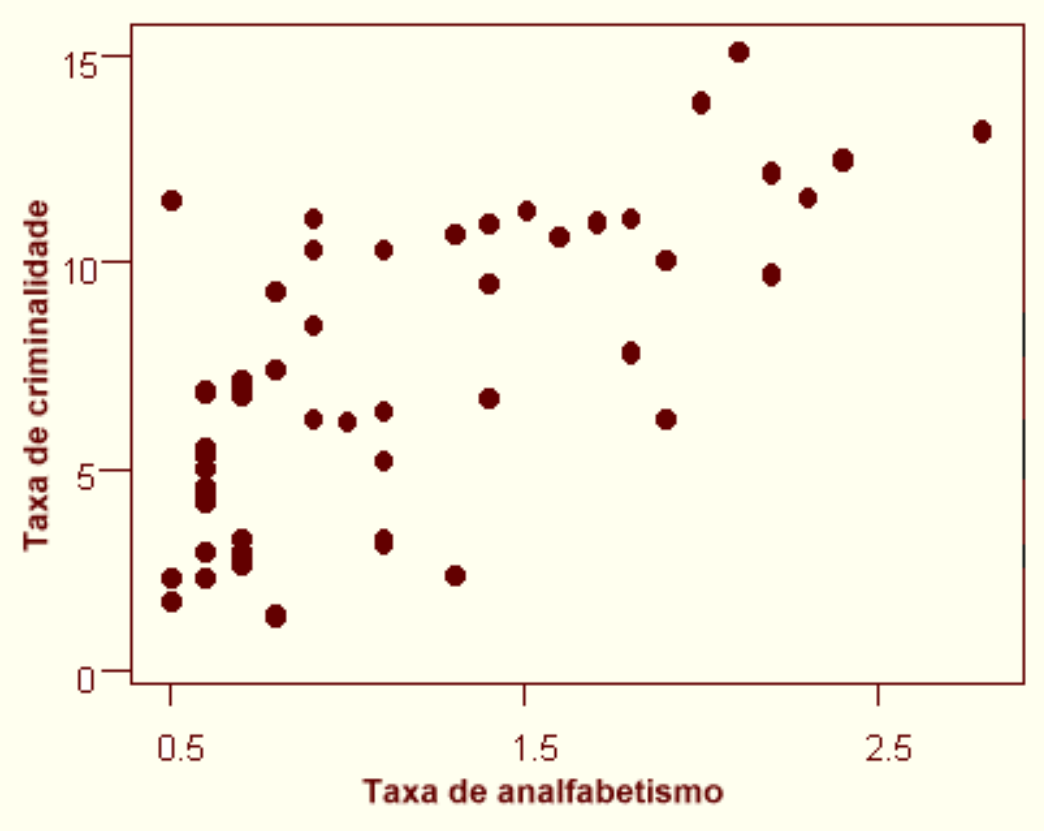
Exemplo 2: criminalidade e analfabetismo

Considere as 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)

$\Sigma X_i Y_i$ = 509,12

Correlação entre X e Y:

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

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

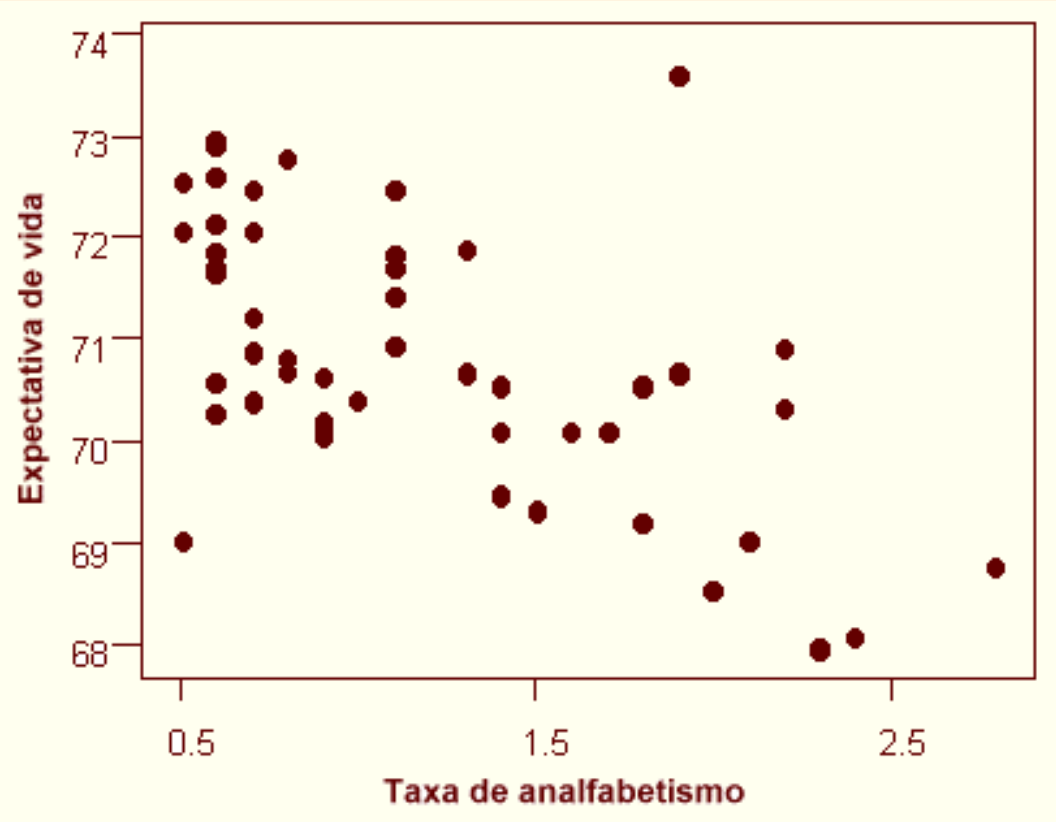
Exemplo 3: expectativa de vida e analfabetismo

Considere as 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)

$\Sigma X_i Y_i$ = 4122,8

Correlação entre X e Y:

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

$$r = \frac{4122{,}8 - 50 \cdot 70{,}88 \cdot 1{,}17}{49 \cdot 1{,}342 \cdot 0{,}609} = \frac{-23{,}68}{40{,}047} = -0{,}59$$

Regressão

- Estudo da forma do relacionamento entre variáveis quantitativas.
- Exemplos:
  - Peso e altura.
  - Renda familiar e número de filhos.
  - Renda e consumo.
  - Volume de produção e custos.
  - Risco e rentabilidade de ações.
  - Gastos com prevenção de defeitos e falhas nos produtos.

Regressão - Objetivos

- Predizer (estimar) uma variável dependente (Y) em função de uma variável independente (X).
- Conhecer o quanto variações de X podem afetar Y.

Exemplos

Variável independente, *X*		Variável dependente, *Y*
Temperatura ambiente (^{0}C)	→	Consumo eletricidade (Kwh)
Horas de estudo	→	Desempenho no vestibular
Renda (R$)	→	Consumo (R$)
Memória RAM do computador (Gb)	→	Tempo de resposta do sistema (s)
Área construída do imóvel (m^2)	→	Preço do imóvel (R$)

Regressão - Modelo

Y = [Predito por X, segundo uma função] + [Efeito aleatório]

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

Parâmetros

Regressão Linear Simples

Reta ajustada:

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

O que são a e b?

a: intercepto

b: inclinação

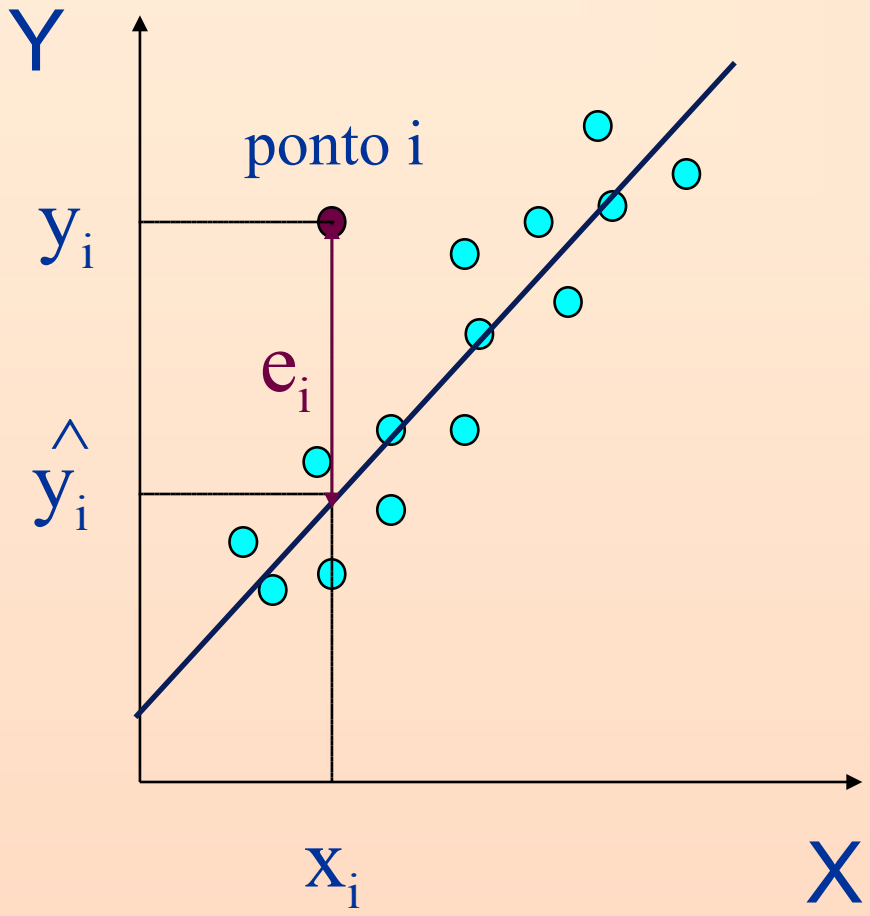
Interpretação de b:

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

Método dos Mínimos Quadrados

reta de regressão estimada:

$$y = a + b.x$$



O método dos míni-mos quadrados sele-ciona os valores de a e b de tal forma que o somatório dos quadrados dos erros (Σe_i^2) é minimizado.

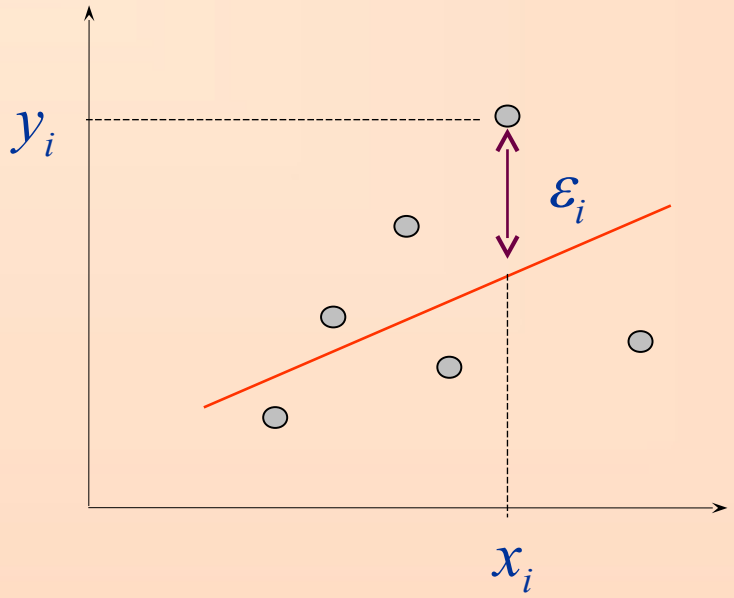
Método dos mínimos quadrados para estimar α e β

- Minimizar em relação a α e β :

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

$$\frac{\partial S}{\partial \alpha} = 0$$

$$\frac{\partial S}{\partial \beta} = 0$$



Método dos mínimos quadrados para estimar α e β

Os coeficientes a e b são calculados da seguinte maneira:

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

Estimativa de α : $$a = \frac{\sum y_i - b\sum x_i}{n}$$

Reta de regressão construída com os dados:

$$\hat{y} = a + bx$$

No exemplo 2,

a reta ajustada é:

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

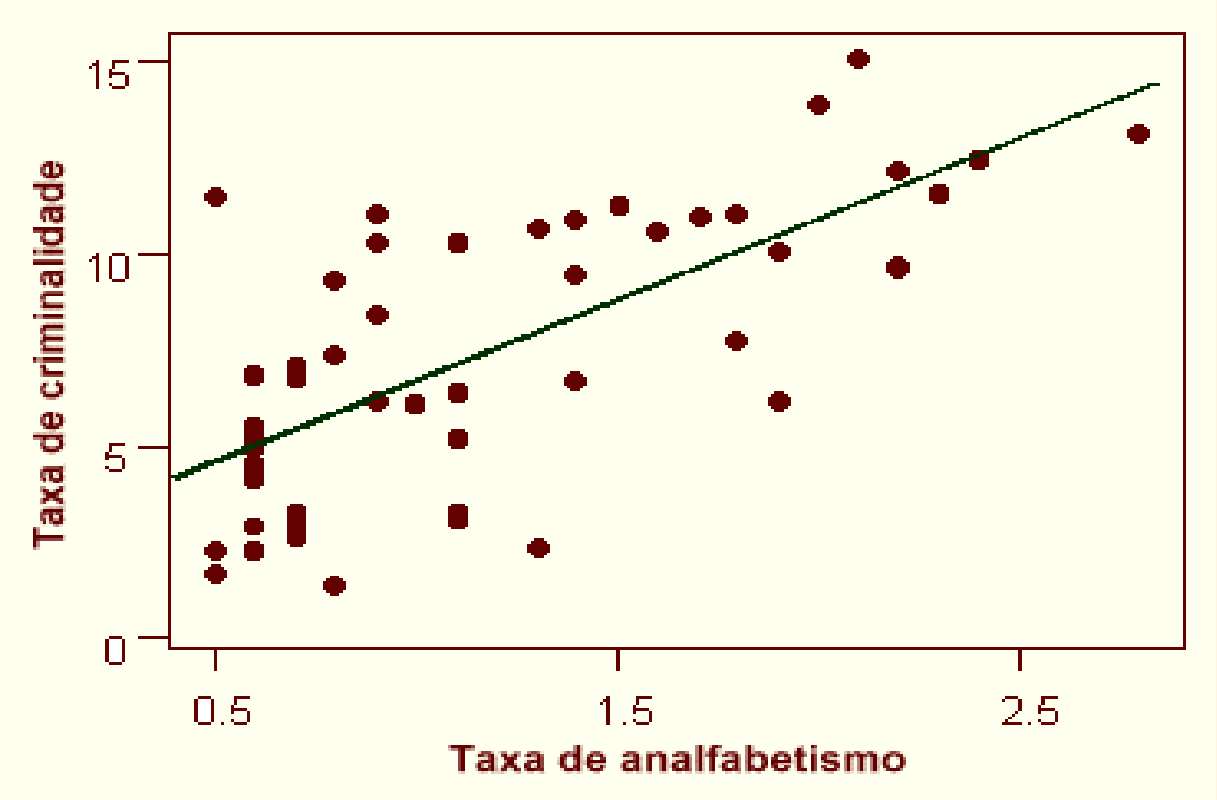
^
Y : valor predito para a taxa de criminalidade

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



Como desenhar a reta no gráfico?

No exemplo 3,

a reta ajustada é:

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

^

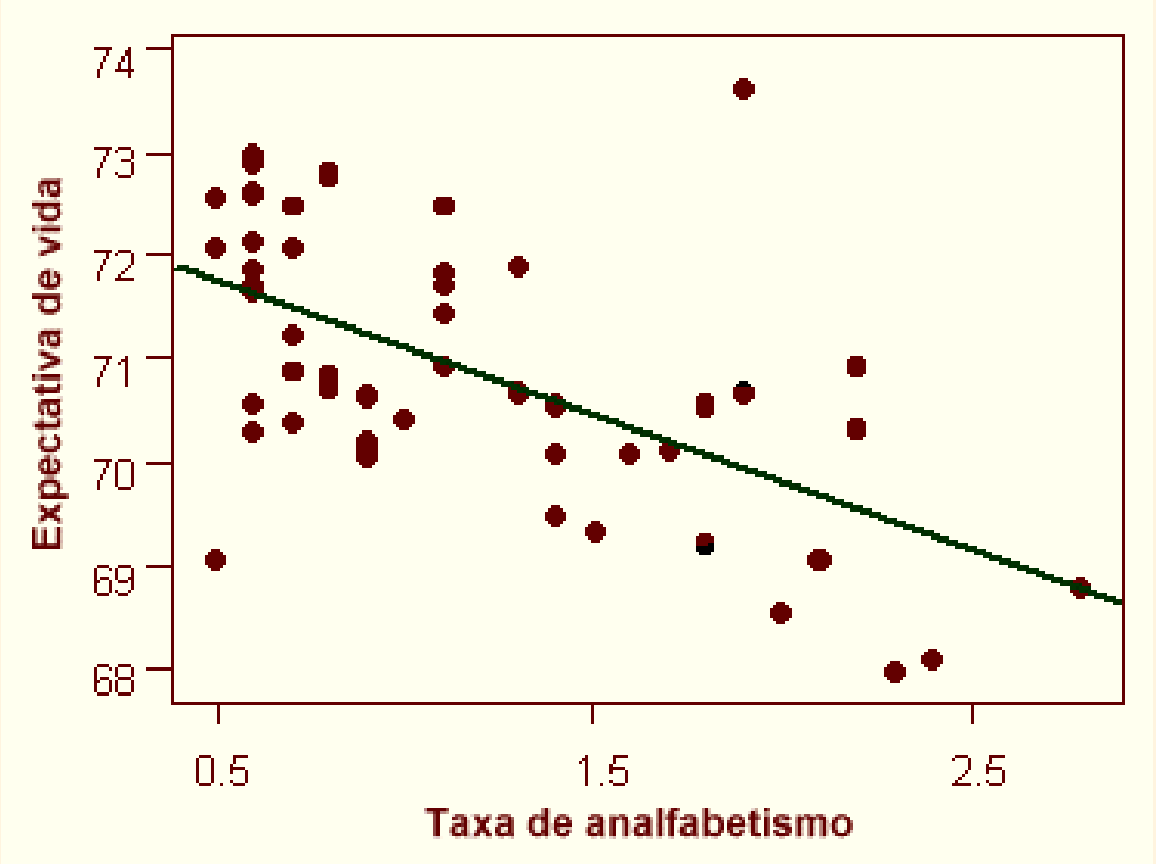
Y : valor predito para a expectativa de vida

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



Exemplo 4: consumo de cerveja e temperatura

Y: consumo de cerveja diário por mil habitantes, em litros.

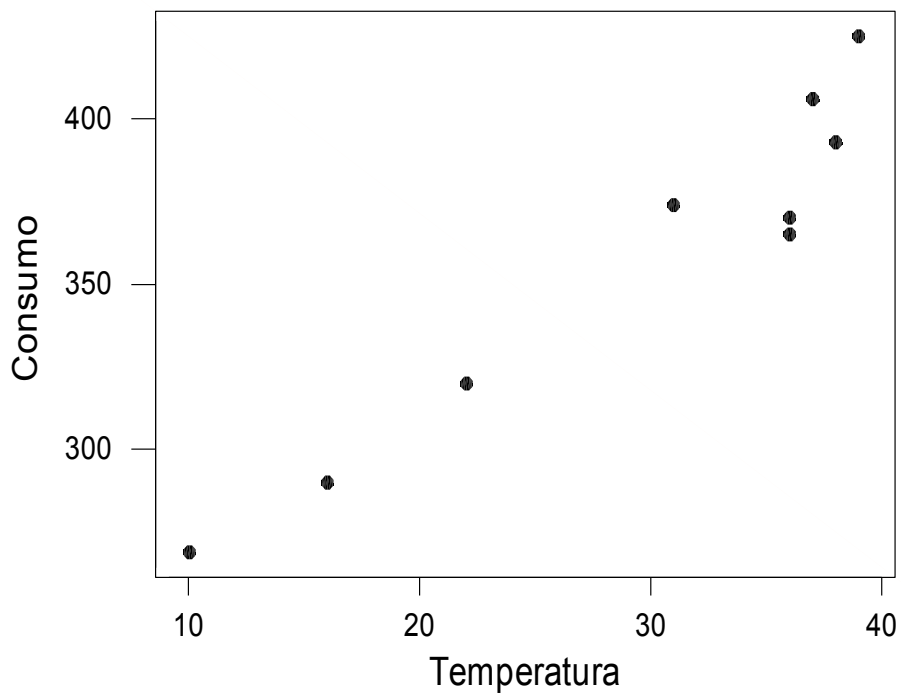
X: temperatura máxima (em ºC).

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

Dados:

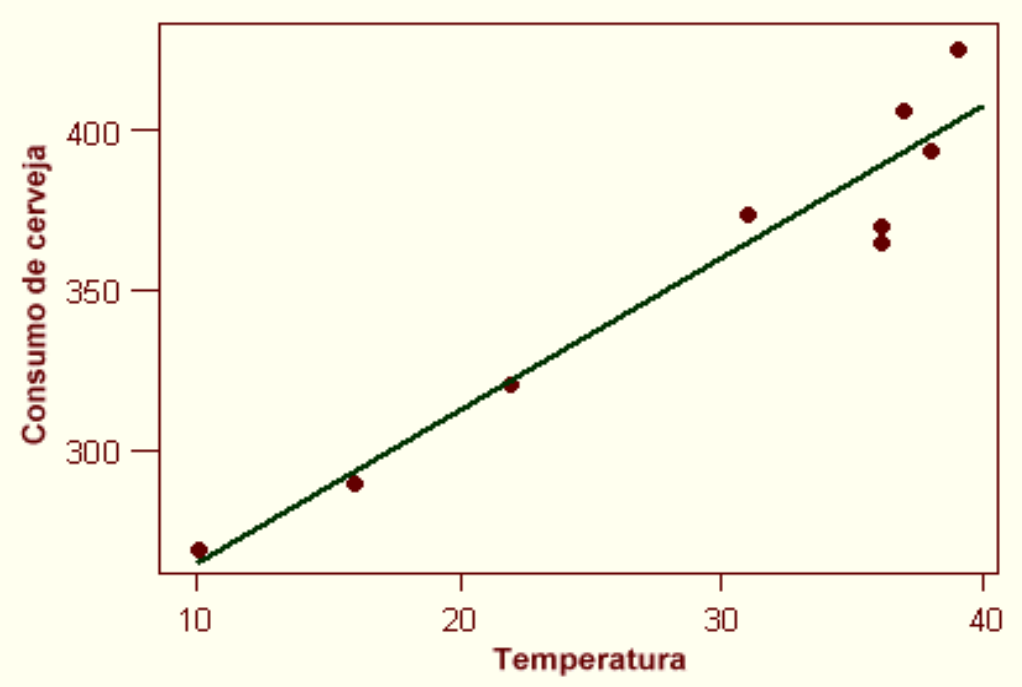
Localidade	Temperatura (X)	Consumo (Y)
1	16	290
2	31	374
3	38	393
4	39	425
5	37	406
6	36	370
7	36	365
8	22	320
9	10	269

Diagrama de dispersão



A correlação entre X e Y é r = 0,962.

A reta ajustada é: $\hat{Y} = 217{,}37 + 4{,}74\,X$



Qual 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 o consumo previsto para uma temperatura de 25ºC?

$$\hat{Y} = 217{,}37 + 4{,}74.25 = 335{,}87 \text{ litros}$$

Exercício: uma empresa opera estúdios fotográficos para crianças em 12 cidades. A empresa deseja expandir seus estúdios para outras cidades semelhantes e deseja investigar se as vendas (Y) podem ser estimadas através do número de pessoas com 16 anos ou menos (X_1) e a renda per capita na cidade (X_2). Os resultados foram:

1 OBSERVAC	2 NUMERO	3 RENDA	4 VENDAS
1,000	68,000	17,000	174,000
2,000	45,000	16,000	164,000
3,000	91,000	18,000	244,000
4,000	48,000	16,000	154,000
5,000	47,000	17,000	182,000
6,000	66,000	18,000	208,000
7,000	50,000	17,000	163,000
8,000	52,000	17,000	145,000
9,000	49,000	17,000	145,000
10,000	38,000	16,000	137,000
11,000	88,000	18,000	242,000
12,000	73,000	17,000	191,000

Modelo de regressão de primeira ordem com duas variáveis preditoras

O modelo de regressão linear é dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad (1)$$

Onde Y_i é a resposta no i-ésimo ensaio, X_{i1} e X_{i2} são os valores das duas variáveis preditoras no i-ésimo ensaio. Os parâmetros do modelo são β_0, β_1, β_2 e o termo do erro é ε_i.

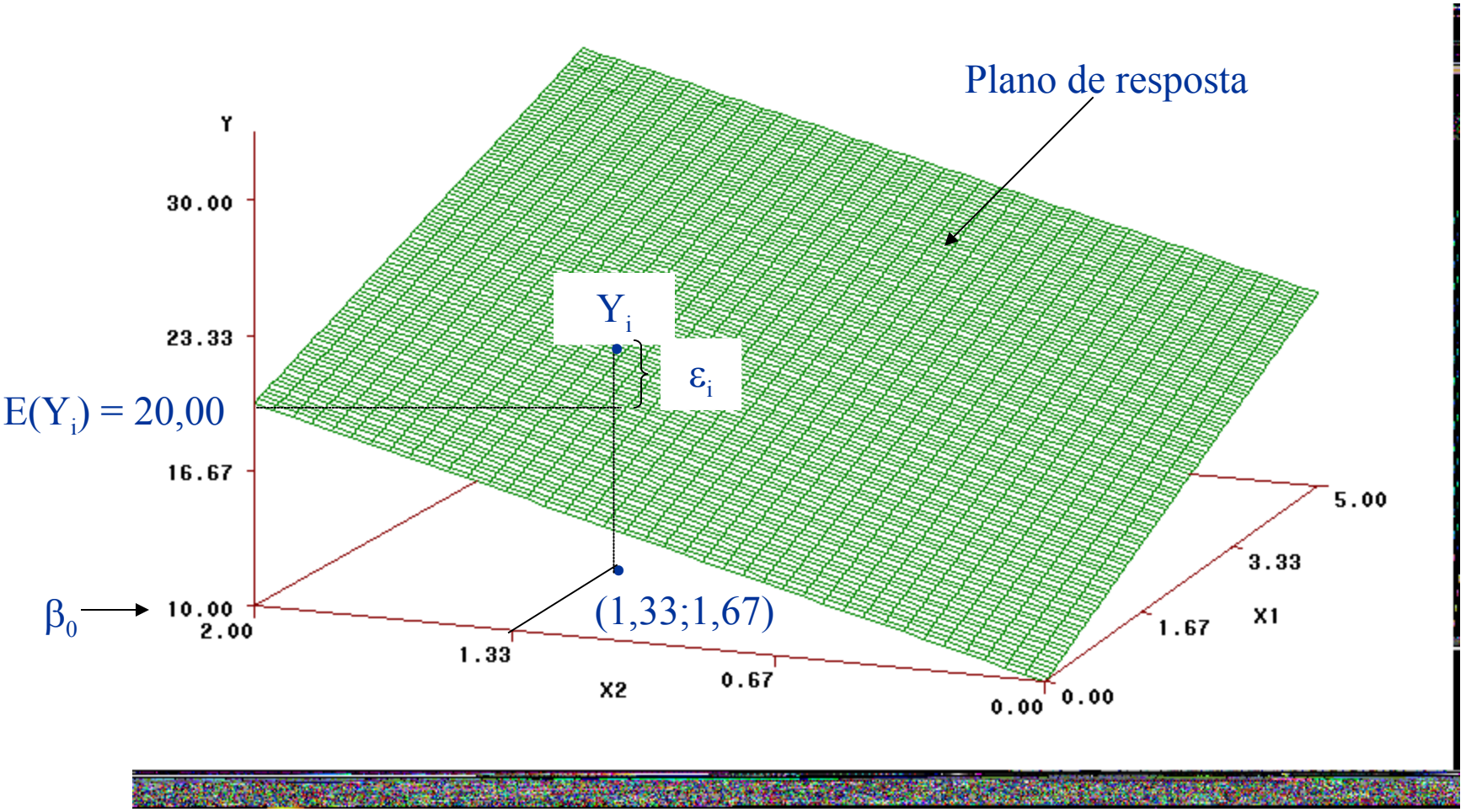
Vamos assumir que $E(\varepsilon_i)=0$, portanto, a função de regressão do modelo de primeira ordem é:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (2)$$

A representação gráfica desta função é um plano no espaço. A figura, na página seguinte, mostra este plano para a função:

$$E(Y) = 10 + 2X_1 + 5X_2 \qquad (3)$$

A função de regressão na regressão múltipla é chamada de ***superfície de resposta.***


Plano de resposta
Y
30.00
23.33
16.67
10.00
Y_i
ε_i
$E(Y_i) = 20,00$
β_0
(1,33;1,67)
2.00
1.33
0.67
0.00
X2
0.00
1.67
3.33
5.00
X1

Significado dos coeficientes de regressão:

O parâmetro β_0 é o intercepto do plano de regressão. Se a abrangência do modelo inclui $X_1=0$ e $X_2=0$ então $\beta_0=10$ representa a resposta média $E(Y)$ neste ponto. Em outras situações, β_0 não tem qualquer outro significado como um termo separado no modelo de regressão.

O parâmetro β_1 indica a mudança na resposta média $E(Y)$ por unidade de acréscimo em X_1 quando X_2 é mantido constante. Da mesma forma β_2 indica a mudança na resposta média por unidade de aumento em X_2 quando X_1 é mantido constante.

Neste modelo, o efeito de X_1 sobre a resposta média não depende de X_2 e vice-versa, assim, dissemos que as variáveis preditoras tem efeito aditivo ou não interagem. Temos um modelo de primeira ordem sem interação.

Qualidade do ajuste

- Ajustou-se uma equação de regressão entre X e Y. E a qualidade do ajuste?
  - análise de variância do modelo
  - análise dos resíduos

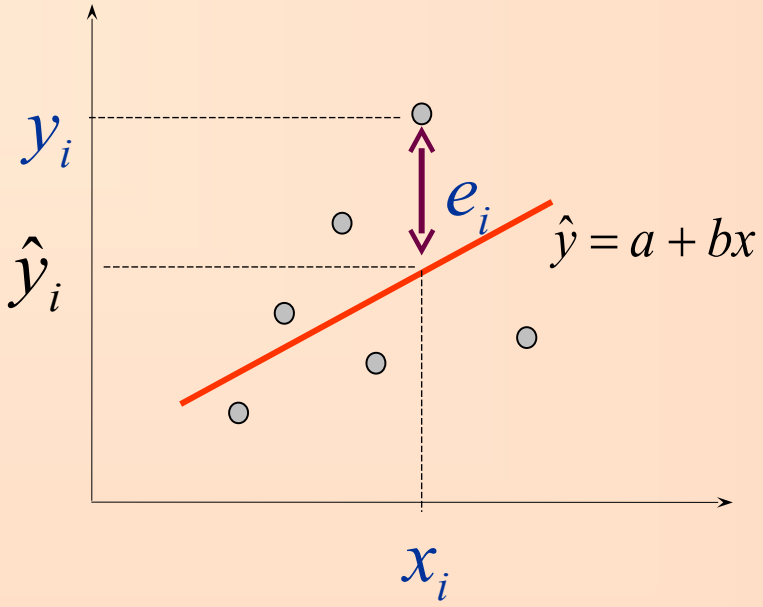
Reta de regressão e resíduos

- **Valores preditos:**

$$\hat{y}_i = a + bx_i$$

Resíduos:

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



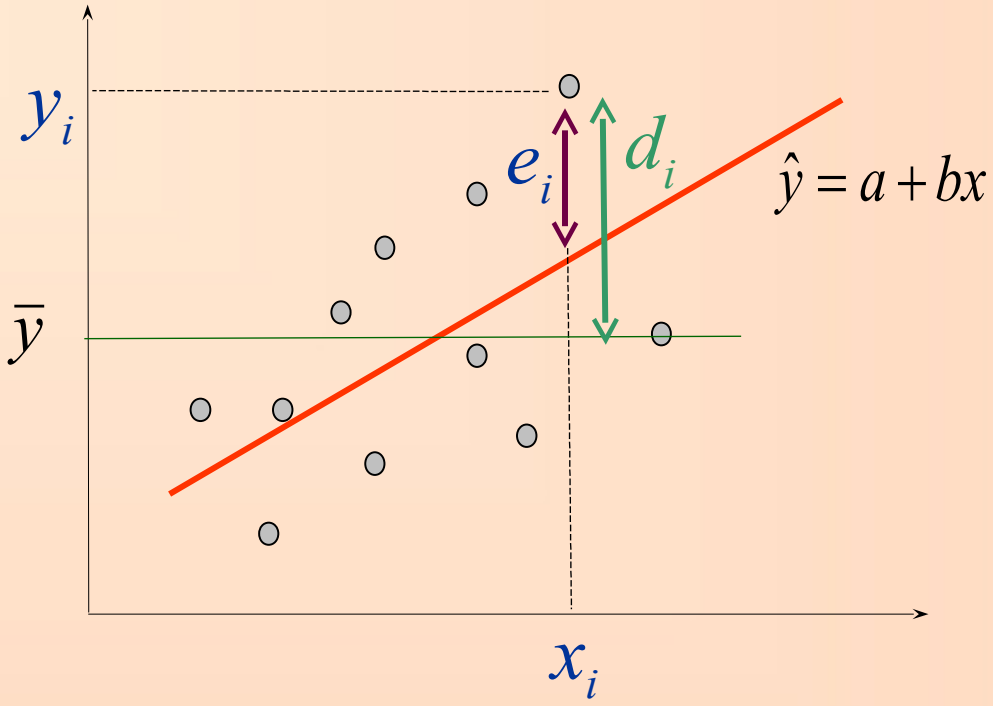
Análise de variância do modelo

Desvio em relação à média aritmética:

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

Desvio em relação à reta de regressão (resíduo da regressão):

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



Somas de quadrados

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

SQT	SQR	SQE
variação total	variação explicada pela equação de regressão	variação não explicada

Medida da qualidade do ajuste:

Coeficiente de determinação (R^2)

$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}} = \frac{\Sigma (\hat{y}_i - \bar{y})^2}{\Sigma (y_i - \bar{y})^2}$$

$$0 \leq R^2 \leq 1$$

Matematicamente, R^2 é o quadrado do Coef. de Correlação de Pearson.