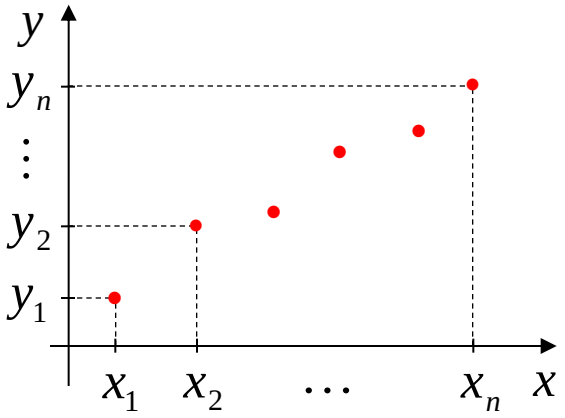
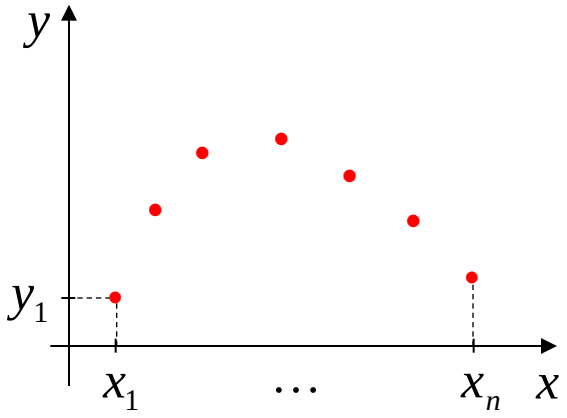
Dados bivariados

- ***Amostra bivariada***:
  - ***Def***: Uma ***amostra*** diz-se ***bivariada*** quando é constituída por pares ordenados de dados, isto é, por valores de duas variáveis.

- ***Diagrama de dispersão***:
  - ***Def***: Um ***diagrama de dispersão*** é uma representação gráfica de dados quantitativos bivariados num sistema de eixos ortogonais.
  - ***Objectivo***: Pôr em evidência a relação (linear, parabólica, etc.), caso exista, entre a variável independente, ***X***, e a variável dependente, ***Y***.
  - ***Exemplos***:

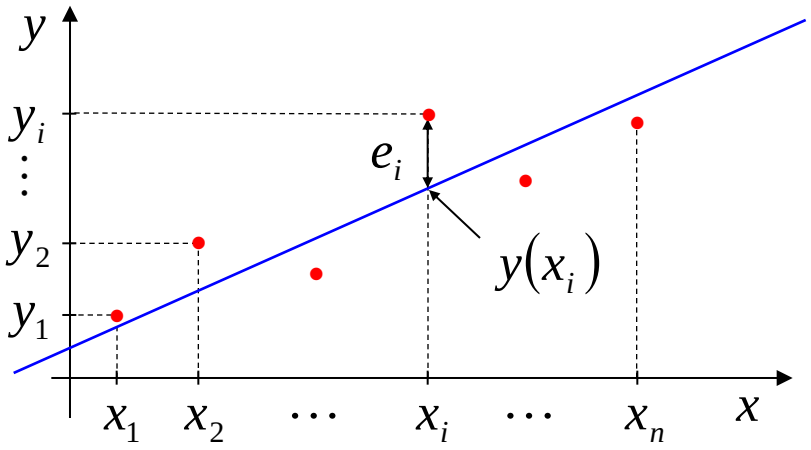
$R: plot(x, y)$





Dados bivariados

- ***Regressão linear***:
  - ***Def***: A ***regressão linear*** é um ajuste de uma recta, ***y = a+bx***, designada por recta de regressão, a um conjunto de dados bivariados quantitativos.
  - ***Validade***: Quando existe uma relação linear entre as duas variáveis.
  - ***Ex***:



$$y = a + bx$$

$$e_i = y_i - y(x_i)$$

e_i – erro (desvio) entre y_i e a recta de regressão, $y(x)$, no ponto de abcissa x_i

Nota: Nesta definição de e_i assume-se que x_i não contém erros

- ***Método dos mínimos quadrados***:
  - ***Def***: O ***método dos mínimos quadrados*** consiste na minimização da soma dos quadrados dos erros (***sqe***) entre a recta de regressão, $y(x) = a + bx$, e os dados $(x_i, y_i), i = 1, 2, \ldots, n$, isto é, na minimização de

$$sqe = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - y(x_i)]^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

  - ***Objectivo***: Determinar os valores de ***a*** e ***b***.
  - ***Teor***: A minimização de ***sqe*** conduz aos seguintes resultados:

$$\begin{cases} b = \dfrac{s_{xy}}{s_{xx}} \\ a = \bar{y} - b\bar{x} \end{cases} ; \qquad s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

    R: $lm(y \sim x)$
  - ***Nota***: A recta de regressão passa pelo ponto $(\bar{x}, \bar{y})$.

Coeficiente de determinação

- *Def*: A ***soma dos quadrados dos erros*** (***sqe***)

$$sqe = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - y(x_i)]^2 = S_{yy} - bS_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$\text{onde} \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

mede a ***variabilidade de y <u>não explicada</u> pela recta de regressão***.

- *Def*: A ***soma dos quadrados dos desvios totais*** (***sqt***)

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

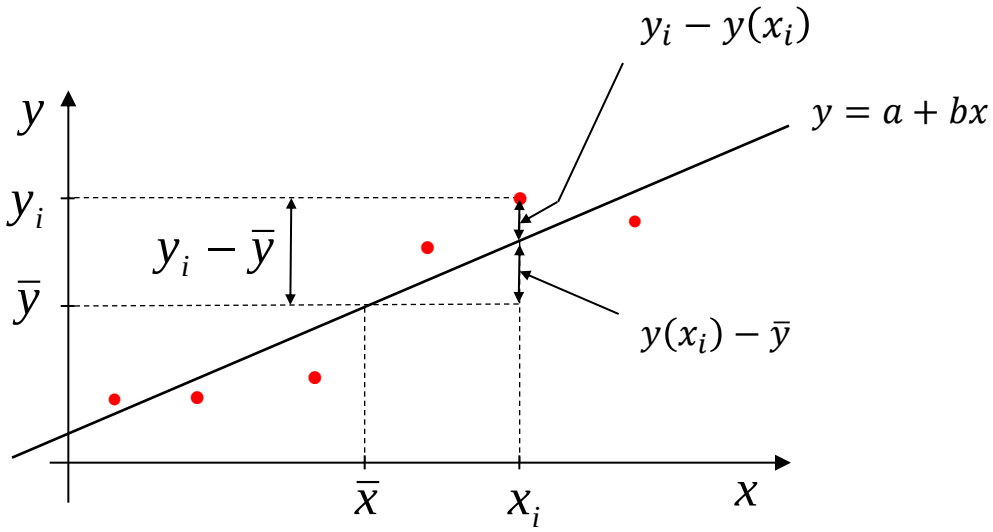
mede a ***variabilidade <u>total</u> de y*** em torno da sua média.

Coeficiente de determinação

- *Def*: A ***soma dos quadrados dos desvios explicados pela regressão*** (***sqr***)

$$sqr = \sum_{i=1}^{n} [y(x_i) - \bar{y}]^2 = \frac{S_{xy}^2}{S_{xx}} = bS_{xy} = sqt - sqe$$

mede a ***variabilidade de y <u>explicada</u> pela recta de regressão***.



Coeficiente de determinação

- *Def*: O ***coeficiente de determinação, r^2,*** é uma medida da qualidade do ajuste e é dado por:

$$r^2 = \frac{sqr}{sqt} = 1 - \frac{sqe}{sqt} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = b^2 \frac{S_{xx}}{S_{yy}}$$

- Propriedades de r^2:
  - Mede a ***qualidade do ajuste*** da recta pois corresponde à proporção da variação de *y* explicada pela regressão.
  - $0 \leq r^2 \leq 1$
  - A qualidade do ajuste aumenta quando r^2 se aproxima de 1.
  - $r^2 = 1 \Rightarrow$ O ***ajuste é perfeito***. (Observações sobre a recta de regressão.)

$$Demo: \quad r^2 = 1 \Leftrightarrow 1 - \frac{sqe}{sqt} = 1 \Leftrightarrow sqe = 0 \Leftrightarrow \sum_{i=1}^{n} [y_i - y(x_i)]^2 = 0 \Leftrightarrow y_i = y(x_i)$$

  - $r^2 = 0 \Rightarrow$ O ***ajuste é inútil***. (Recta de regressão horizontal.)

$$Demo: \quad r^2 = 0 \Leftrightarrow sqr = 0 \Leftrightarrow sqe = sqt \Leftrightarrow S_{yy} - bS_{xy} = S_{yy} \Leftrightarrow b = 0 \ \therefore \ y = \bar{y}$$

Coeficiente de correlação

- *Def*: O ***coeficiente de correlação***, ***r***, *é:*

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = b\sqrt{\frac{S_{xx}}{S_{yy}}}$$

$R: cor(x, y)$

e também é usado para medir a qualidade do ajuste.

- *Desvantagem*: *r* não tem uma interpretação estatística tão conveniente como r^2.

- Propriedades de *r*:
  - $-1 \leq r \leq 1$
  - A ***qualidade do ajuste*** aumenta quando $|r|$ se aproxima de 1.
  - O ***sinal*** de *r* é igual ao sinal de *b* (o ***declive da recta*** de regressão).
  - $|r| = 1 \Rightarrow$ O ajuste é perfeito (pontos sobre a recta)
  - $r = 0 \Rightarrow$ O ajuste é inútil (recta horizontal)

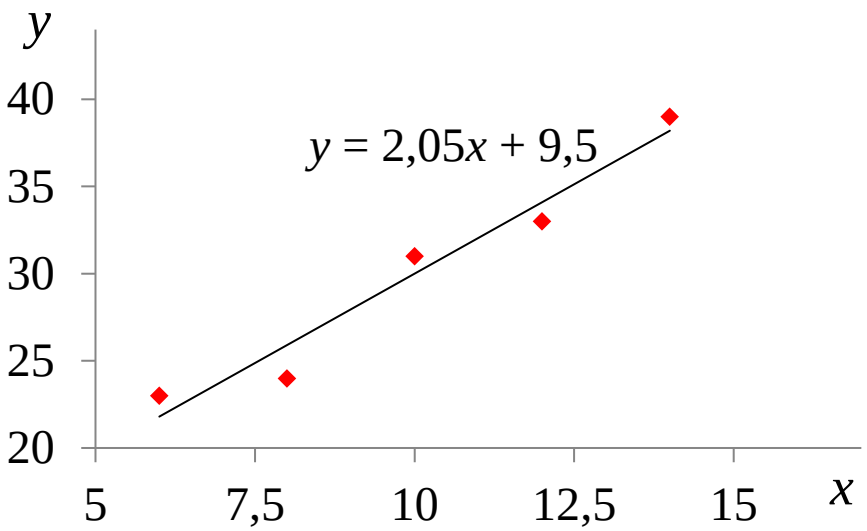
Exercício

Dada a seguinte amostra

x	12	8	14	10	6
y	33	24	39	31	23

a) Represente o diagrama de dispersão e a recta de regressão correspondentes.
b) Calcule o coeficiente de correlação.

a)



b) $r = b\sqrt{\frac{s_{xx}}{s_{yy}}} = 2.05\sqrt{\frac{40}{176}} \approx 0.977$