# La Retta di Regressione II


M.M.I.B. 2019/20
corso di laurea in Scienze Biologiche
*Sapienza* Università di Roma

# La retta di regressione dei minimi quadrati

Schema di lavoro: partendo dalle osservazioni $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

1. ricaviamo le medie $\overline{x}$ e $\overline{y}$,

# La retta di regressione dei minimi quadrati

Schema di lavoro: partendo dalle osservazioni $\{(x_1, y_1), ..., (x_n, y_n)\}$

1. ricaviamo le medie $\overline{x}$ e $\overline{y}$,

2. ricaviamo le deviazioni standard campionarie dei due campioni $s_x$ e $s_y$,

# La retta di regressione dei minimi quadrati

Schema di lavoro: partendo dalle osservazioni $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

1. ricaviamo le medie $\overline{x}$ e $\overline{y}$,
2. ricaviamo le deviazioni standard campionarie dei due campioni $s_x$ e $s_y$,
3. calcoliamo la loro correlazione $r$,

# La retta di regressione dei minimi quadrati

Schema di lavoro: partendo dalle osservazioni $\{(x_1, y_1), ..., (x_n, y_n)\}$

1. ricaviamo le medie $\overline{x}$ e $\overline{y}$,

2. ricaviamo le deviazioni standard campionarie dei due campioni $s_x$ e $s_y$,

3. calcoliamo la loro correlazione $r$,

allora la retta di regressione dei minimi quadrati ha la seguente espressione

$$y = bx + a \qquad \text{dove} \qquad b = r\frac{s_y}{s_x} \qquad \text{e} \qquad a = \overline{y} - b\overline{x}$$

# Stimare la bontà di una regressione

Il coefficiente di determinazione $R^2$ è la frazione di varianza della variabile dipendente $y$ predicibile dalla variabile indipendente, o (se si preferisce) spiegata dalla retta di regressione dei minimi quadrati.

# Stimare la bontà di una regressione

Il coefficiente di determinazione $R^2$ è la frazione di varianza della variabile dipendente $y$ predicibile dalla variabile indipendente, o (se si preferisce) spiegata dalla retta di regressione dei minimi quadrati.

Chiamando $\{y_i\}$ i dati del campione, $\{bx_i + a\}$ i risultati previsti dalla retta dei minimi quadrati e $\bar{y}$ il valor medio del campione, introduciamo la **devianza totale**

$$S_{yy} = (y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2 = (n-1)S_y^2$$

# Stimare la bontà di una regressione

Il coefficiente di determinazione $R^2$ è la frazione di varianza della variabile dipendente $y$ predicibile dalla variabile indipendente, o (se si preferisce) spiegata dalla retta di regressione dei minimi quadrati.

Chiamando $\{y_i\}$ i dati del campione, $\{bx_i + a\}$ i risultati previsti dalla retta dei minimi quadrati e $\bar{y}$ il valor medio del campione, introduciamo la **devianza totale**

$$S_{yy} = (y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2 = (n-1)S_y^2$$

la **devianza spiegata dal modello**

$$SS_{\text{mod}} = (bx_1 + a - \bar{y})^2 + \ldots + (bx_n + a - \bar{y})^2$$

# Stimare la bontà di una regressione

Il coefficiente di determinazione $R^2$ è la frazione di varianza della variabile dipendente $y$ predicibile dalla variabile indipendente, o (se si preferisce) spiegata dalla retta di regressione dei minimi quadrati.

Chiamando $\{y_i\}$ i dati del campione, $\{bx_i + a\}$ i risultati previsti dalla retta dei minimi quadrati e $\bar{y}$ il valor medio del campione, introduciamo la **devianza totale**

$$S_{yy} = (y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2 = (n-1)S_y^2$$

la **devianza spiegata dal modello**

$$SS_{\text{mod}} = (bx_1 + a - \bar{y})^2 + \ldots + (bx_n + a - \bar{y})^2$$

e infine anche la **devianza residua**

$$SS_{\text{res}} = (bx_1 + a - y_1)^2 + \ldots + (bx_n + a - y_n)^2$$

# Stimare la bontà di una regressione

La quantità $(S_{yy} - SS_{\text{res}})$ misura l'entità della deviazione giustificata dalla relazione lineare

# Stimare la bontà di una regressione

La quantità $(S_{yy} - SS_{\text{res}})$ misura l'entità della deviazione giustificata dalla relazione lineare

$$R^2 = \frac{SS_{\text{res}}}{S_{yy}} = 1 - \frac{SS_{\text{mod}}}{S_{yy}}$$

è la proporzione della variazione delle variabili di risposta.

# Stimare la bontà di una regressione

La quantità $(S_{yy} - SS_{\text{res}})$ misura l'entità della deviazione giustificata dalla relazione lineare

$$R^2 = \frac{SS_{\text{res}}}{S_{yy}} = 1 - \frac{SS_{\text{mod}}}{S_{yy}}$$

è la proporzione della variazione delle variabili di risposta.

$R^2$ viene detto **coefficiente di determinazione**

# 1. dati e computer

Supponiamo di voler studiare il legame che intercorre tra una certa mole di dati (misurati in MB) e tempo di lavoro (in secondi) necessario ad un calcolatore per elaborarli,

# 1. dati e computer

Supponiamo di voler studiare il legame che intercorre tra una certa mole di dati (misurati in MB) e tempo di lavoro (in secondi) necessario ad un calcolatore per elaborarli, quindi facciamo un certo numero di esperimenti ed otteniamo la seguente tabella di valori

| dati | 105 | 511 | 401 | 622 | 330 |
|---|---|---|---|---|---|
| tempo | 44 | 214 | 193 | 299 | 143 |

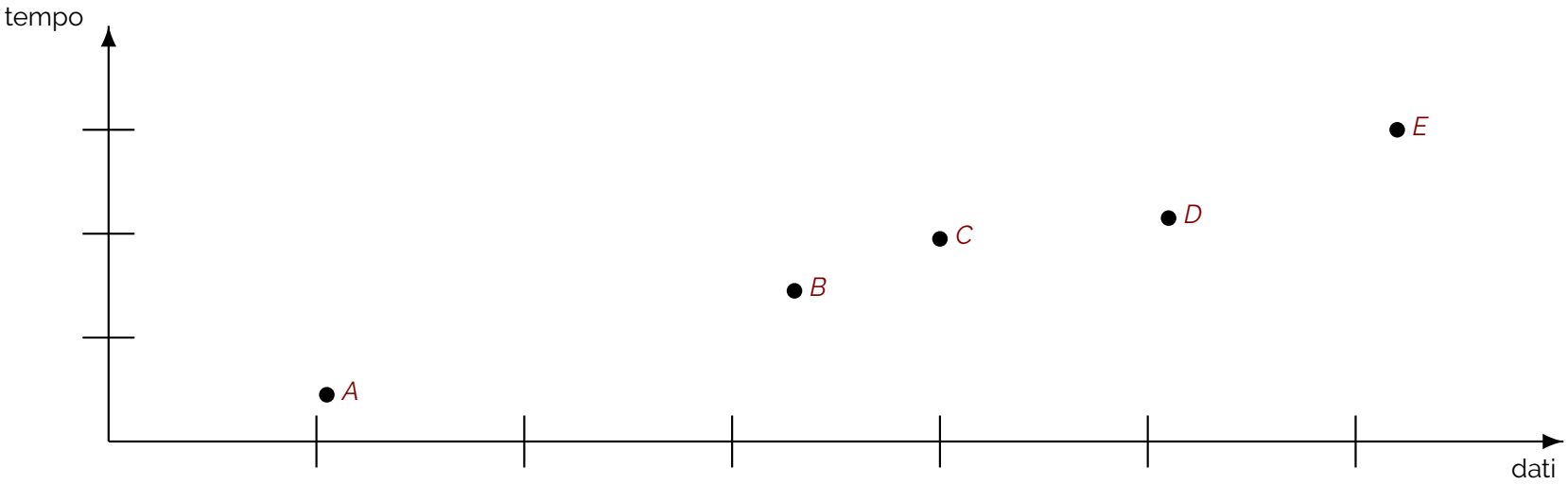
# 1. dati e computer

Supponiamo di voler studiare il legame che intercorre tra una certa mole di dati (misurati in MB) e tempo di lavoro (in secondi) necessario ad un calcolatore per elaborarli, quindi facciamo un certo numero di esperimenti ed otteniamo la seguente tabella di valori

| dati | 105 | 511 | 401 | 622 | 330 |
|---|---|---|---|---|---|
| tempo | 44 | 214 | 193 | 299 | 143 |

i. si tracci uno scatterplot dei dati,
ii. si calcoli il coefficiente di correlazione tra le variabili,
iii. si scriva l'equazione della retta di regressione,
iv. si stimi il tempo previsto per processare 250 MB di dati.

# scatterplot




dove i punti sono:
$A(105, 44)$,
$B(330, 143)$,
$C(401, 194)$,
$D(511, 214)$,
$E(622, 299)$.

# indici dei dati

$$\bar{x} = \frac{1}{5}(105 + 330 + 401 + 511 + 622) = 393,8$$

$$\bar{y} = \frac{1}{5}(44 + 143 + 193 + 214 + 299) = 178,6$$

$$s_x = \left[\frac{1}{4}\Big((105 - 393,8)^2 + (330 - 393,8)^2 + (401 - 393,8)^2 + (511 - 393,8)^2 + (622 - 393,8)^2\Big)\right]^{1/2} = 195,79$$

$$s_y = \left[\frac{1}{4}\Big((44 - 178,6)^2 + (143 - 178,6)^2 + (193 - 178,6)^2 + (214 - 178,6)^2 + (299 - 178,6)^2\Big)\right]^{1/2} = 94$$

$$r = \frac{1}{4}\left[\frac{(105 - 393,8)(44 - 178,6)}{94 \cdot 195,79} + \ldots\right] = 0,98987 = 0,99$$

# retta dei minimi quadrati

Ricordando che la retta di regressione dei minimi quadrati ha la seguente espressione

$$y = bx + a \qquad \text{dove} \qquad b = r\frac{s_y}{s_x} \qquad \text{e} \qquad a = \overline{y} - b\overline{x}$$

# retta dei minimi quadrati

Ricordando che la retta di regressione dei minimi quadrati ha la seguente espressione

$$y = bx + a \qquad \text{dove} \qquad b = r\frac{s_y}{s_x} \qquad \text{e} \qquad a = \overline{y} - b\overline{x}$$

otteniamo

$$b = 0,99 \cdot \frac{94}{195,79} = 0,475 \qquad a = 178,6 - 0,475 \cdot 393,8 = -8,455$$

# retta dei minimi quadrati

Ricordando che la retta di regressione dei minimi quadrati ha la seguente espressione

$$y = bx + a \qquad \text{dove} \qquad b = r\frac{s_y}{s_x} \qquad \text{e} \qquad a = \overline{y} - b\overline{x}$$

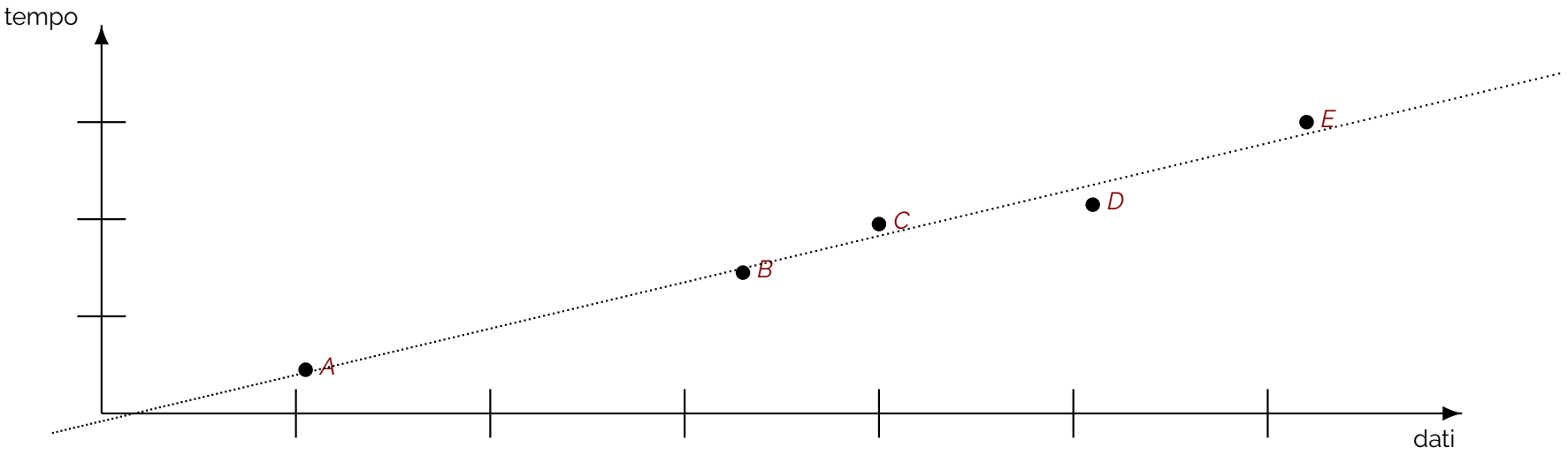
otteniamo

$$b = 0,99 \cdot \frac{94}{195,79} = 0,475 \qquad a = 178,6 - 0,475 \cdot 393,8 = -8,455$$

ovvero

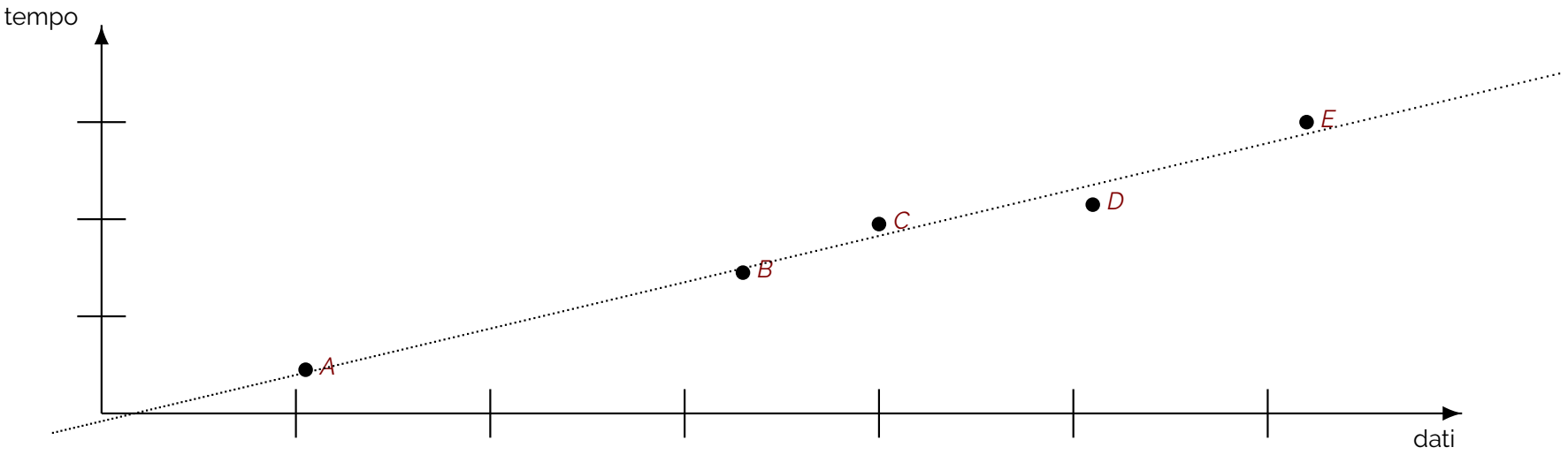
$$y = 0,475x - 8,455$$

# scatterplot e retta




$S_{yy} = 35.091,7$

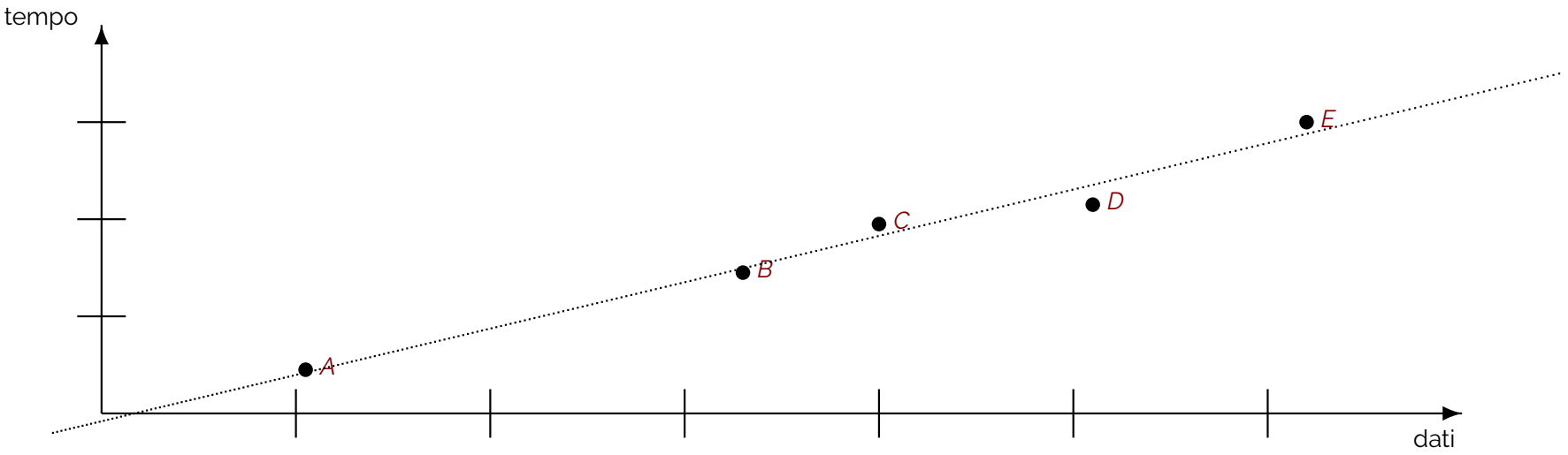
# scatterplot e retta




$S_{yy} = 35.091,7$

$SS_{res} = 34.354,77$

# scatterplot e retta

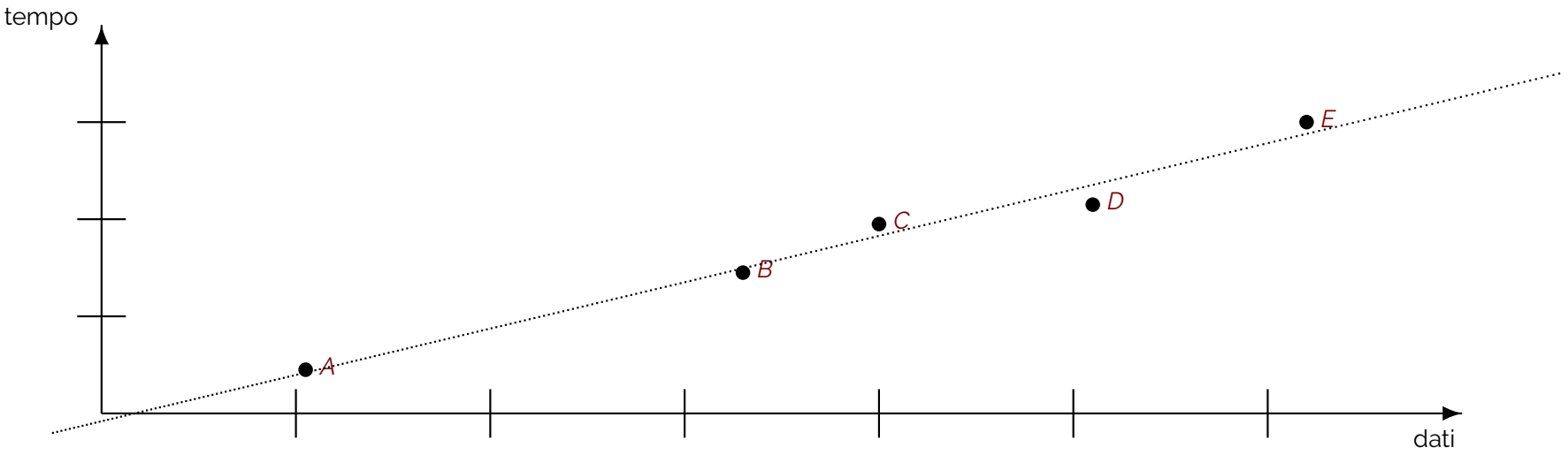



$S_{yy} = 35.091,7$

$SS_{\text{res}} = 34.354,77$

$R^2 = SS_{\text{res}}/S_{yy} = 0,979$

# scatterplot e retta




$S_{yy} = 35.091,7$

$SS_{\text{res}} = 34.354,77$

$R^2 = SS_{\text{res}}/S_{yy} = 0,979$

Se $x = 250MB$ la retta stima $y = 0,475 \cdot 250 - 8,455 = 110,3s$ di tempo necessario all'elaborazione.

# 2. inquinamento

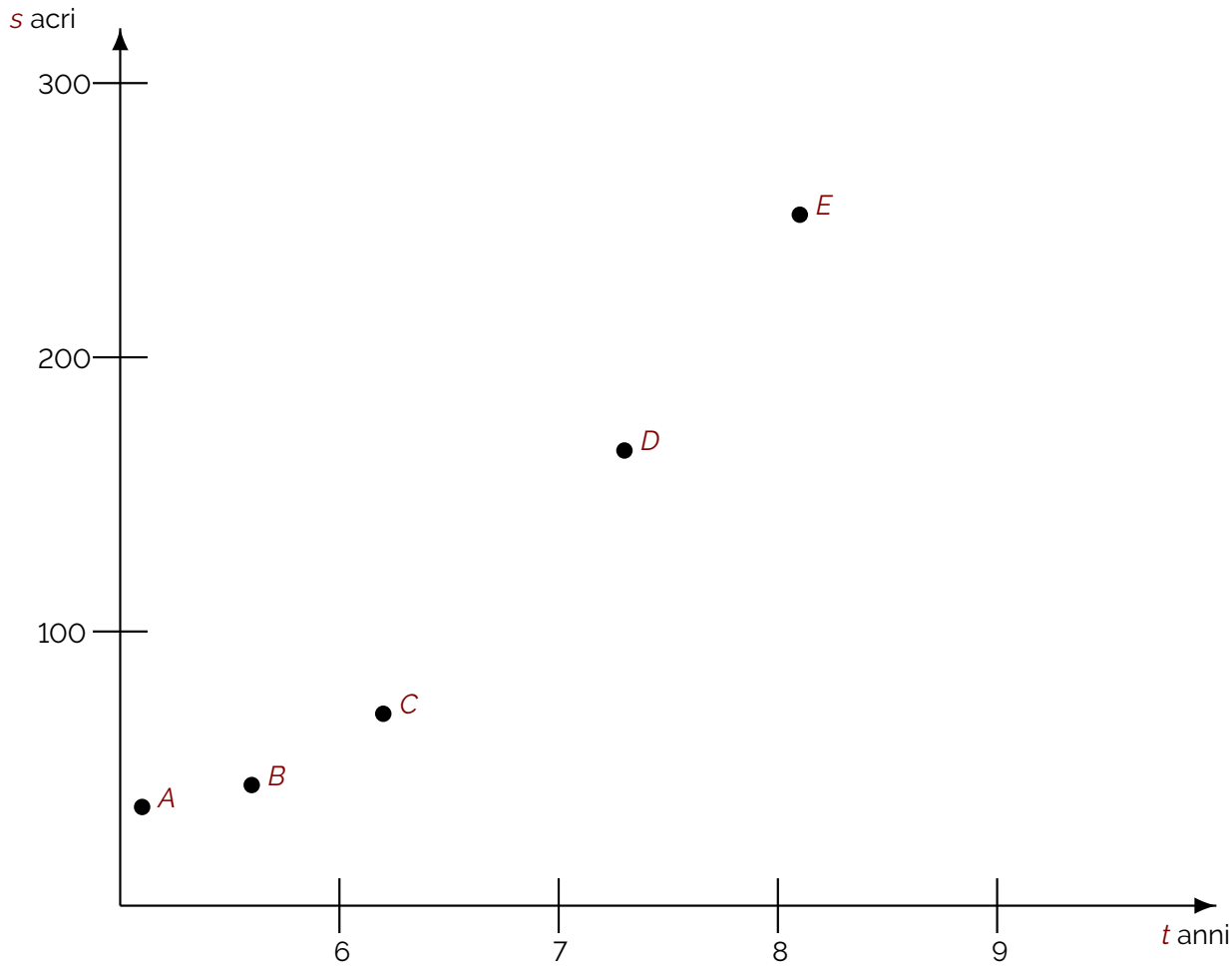
Nella seguente tabella possiamo vedere l'ampiezza dell'area di contaminazione di acqua (misurata in acri, in seguito ad una fuga di un agente tossico), con il passare del tempo (misurato in anni)

# 2. inquinamento

Nella seguente tabella possiamo vedere l'ampiezza dell'area di contaminazione di acqua (misurata in acri, in seguito ad una fuga di un agente tossico), con il passare del tempo (misurato in anni)

| tempo | 5,1 | 5,6 | 6,2 | 7,3 | 8,1 |
|---|---|---|---|---|---|
| area | 35,8 | 44,5 | 68,7 | 165,6 | 253,4 |

# scatterplot




$A(5,1;35,8)$, $B(5,6;44,5)$, $C(6,2;68,7)$, $D(7,3;165,6)$, $E(8,1;253,4)$.

# indici e retta

I principali indici campionari sono

$$\bar{t} = \frac{1}{5}(5,1 + 5,6 + 6,2 + 7,3 + 8,1) = 6,46$$

$$\bar{s} = \frac{1}{5}(35,8 + 44,5 + 68,7 + 165,6 + 253,4) = 178,6$$

$$s_t = \ldots = 1,23$$

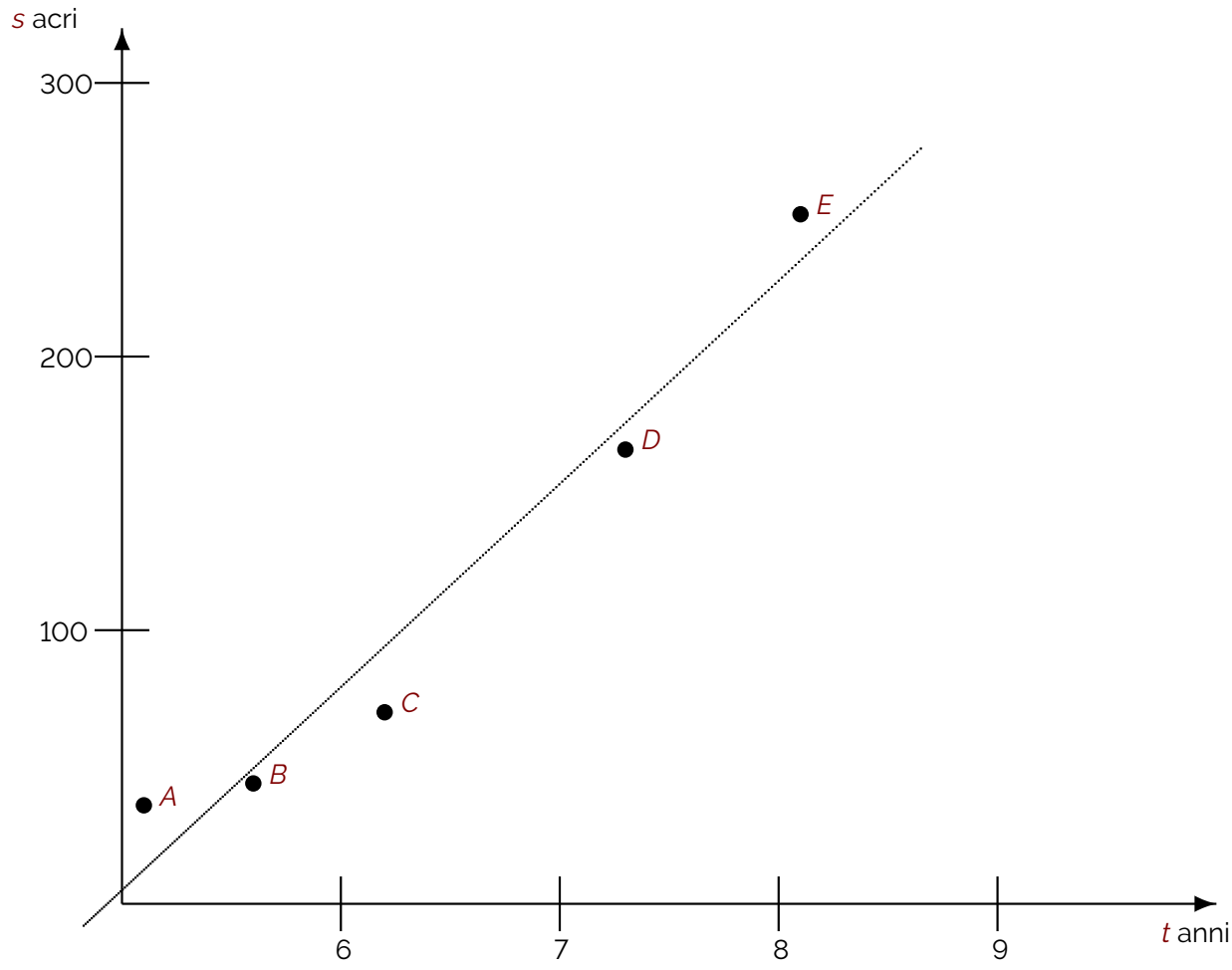
$$s_s = \ldots = 93,66$$

$$r = = 0,9762$$

di conseguenza la retta di regressione ha equazione

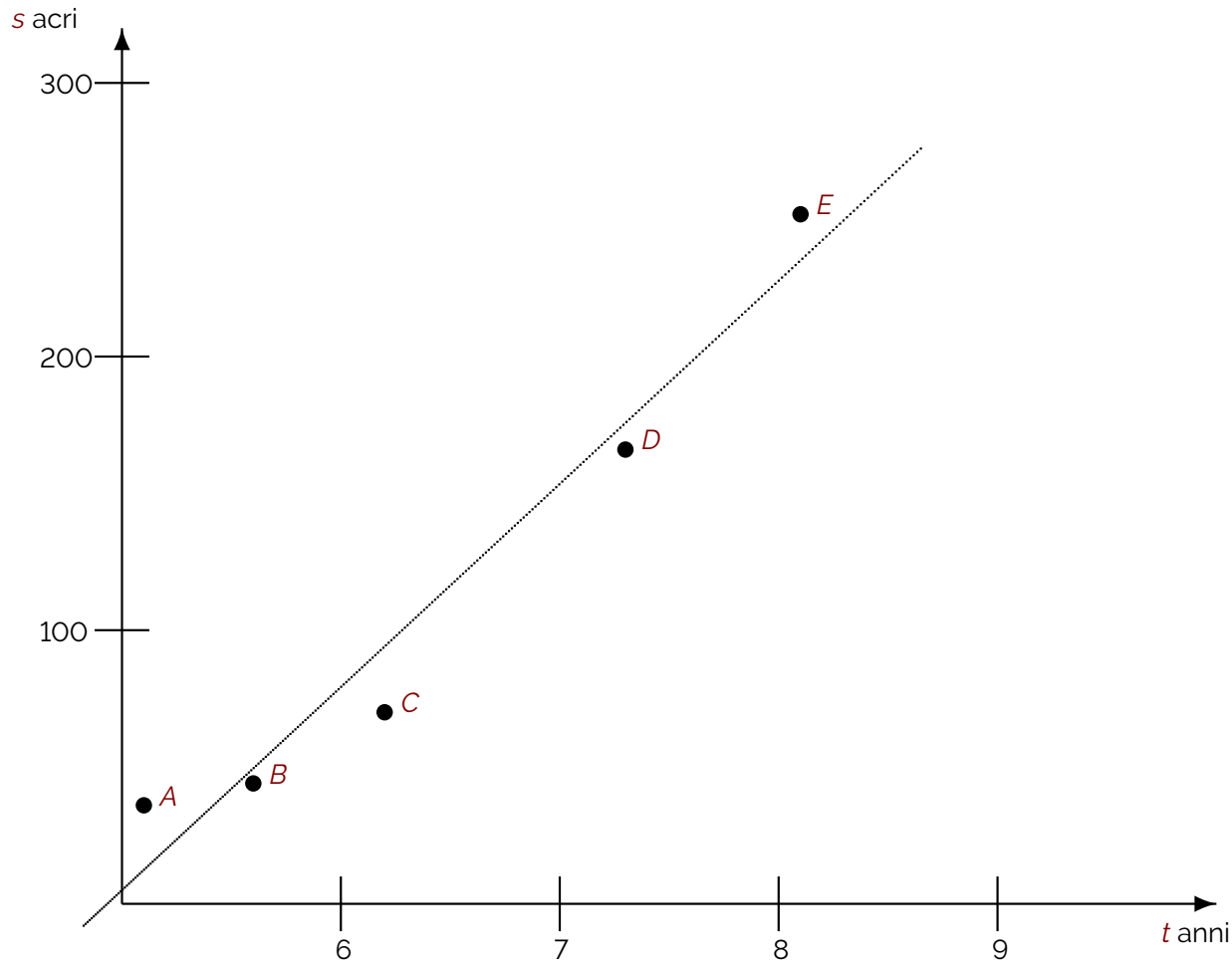
$$s = 74,33 \cdot t - 366,59$$

# scatterplot e retta



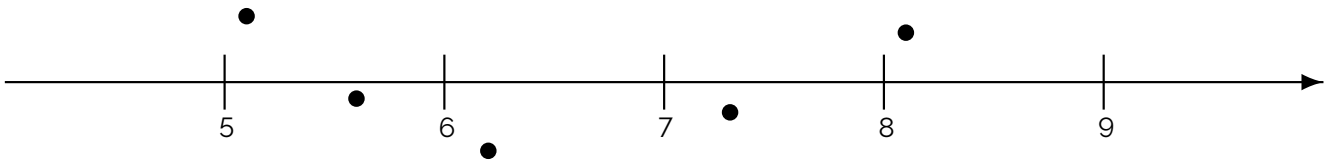

$S_{SS} = 35.421,158; \quad SS_{\text{res}} = 35.341,2;$

# scatterplot e retta

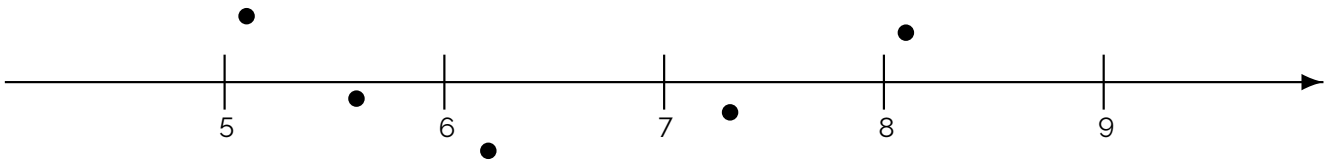



$S_{SS} = 35.421,158; \quad SS_{\text{res}} = 35.341,2; \quad R^2 = 0,998$

# residui



# residui




Tracciando un grafico dei residui, possiamo rendere evidente che, nonostante l'adattamento sembri buono, i punti sono disposti lungo una specie di parabola, di fatto la disposizione dei residui non è aleatoria. Questo suggerisce che la relazione in realtà non sia lineare.