# Esercitazione 6

## Statistica

Alfonso Iodice D'Enza
iodicede@unina.it

Università degli studi di Cassino

# Outline

# Outline

# Outline

# Outline

# Dipendenza lineare

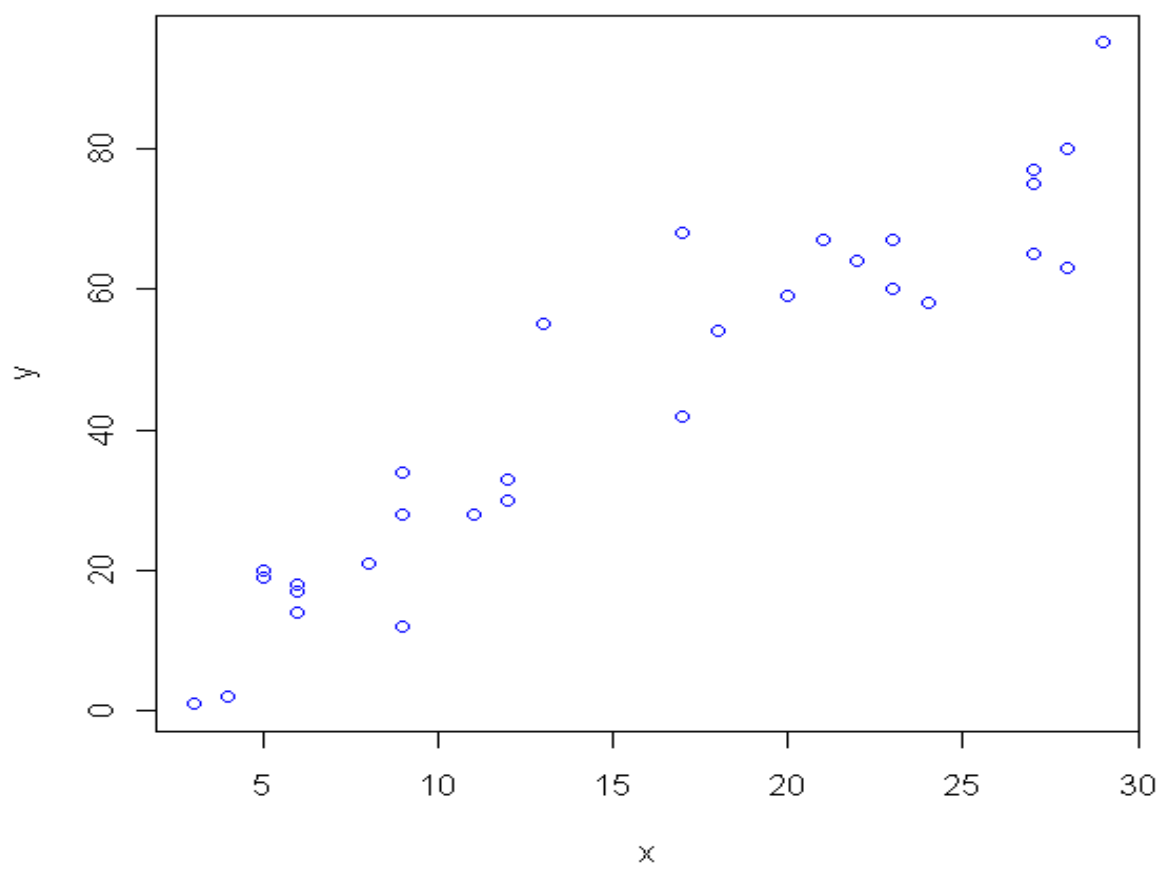


Lo studio della relazione tra caratteri statistici è, nel caso della interdipendenza, di tipo simmetrico: due caratteri quantitativi $X$ e $Y$ hanno lo stesso ruolo e si vuole studiare se essi siano indipendenti o meno. A questo scopo sono stati introdotti gli indici di covarianza $\sigma_{xy}$ e di correlazione lineare $\rho$. Si consideri di aver osservato due caratteri quantitativi $X$ ed $Y$. Si riportano i valori e il grafico di dispersione:

I dati

```
> Data
    Y  X
1  28 11
2  21  8
3  63 28
4  42 17
5  28  9
6   2  4
7  80 28
8  19  5
9  33 12
10 60 23
11 14  6
12 58 24
13 54 18
14 67 21
15 18  6
16 64 22
17 65 27
18 68 17
19 77 27
20 17  6
21 95 29
22 12  9
23  1  3
24 30 12
25 34  9
26 67 23
27 20  5
28 75 27
29 59 20
30 55 13
```

scatter plot



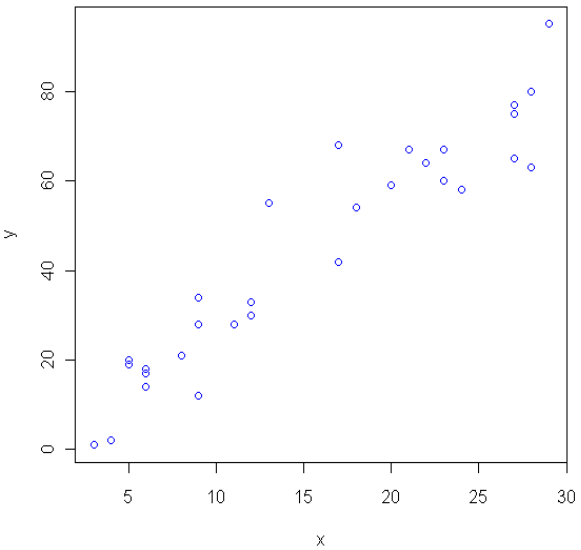
# Dipendenza lineare



### covarianza e coefficiente di correlazione

- $\mu_x = \frac{\sum_{i=1}^{30} x_i}{30} = 15.63$
- $\mu_y = \frac{\sum_{i=1}^{30} y_i}{30} = 44.2$
- $\sigma_x = \sqrt{\frac{\sum_{i=1}^{30} (x_i - \mu_x)^2}{30}} = 8.55$
- $\sigma_y = \sqrt{\frac{\sum_{i=1}^{30} (y_i - \mu_y)^2}{30}} = 25.35$
- $\sigma_{xy} = \frac{\sum_{i=1}^{30} (x_i - \mu_x)(y_i - \mu_y)}{30} = 205.04$
- $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{205.04}{216.7716} = 0.9458805$

### scatter plot




### Dipendenza funzionale lineare

Essendo il valore del coefficiente di correlazione lineare prossimo ad 1 esiste una forte relazione lineare tra $X$ ed $Y$. Come confermato dal grafico di dispersione, i dati sono approssimativamente allineati lungo una retta crescente. Ci si può dunque aspettare che sussista una relazione funzionale tra i dati del tipo

$$Y = f(X) = b_0 + b_1 X$$

che rappresenta l'equazione di una retta passante attraverso la nube di punti di coordinate $(x_i, y_i)$.

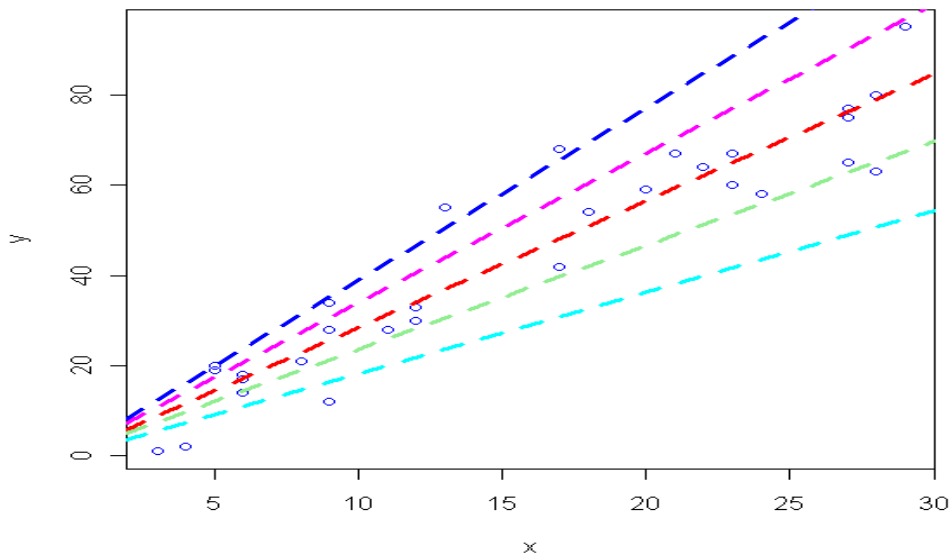
# La retta di regressione



### La retta di regressione

La retta di regressione fornisce una approssimazione della dipendenza dei valori di $Y$ dai valori di $X$. La relazione di dipendenza non è esattamente riprodotta dalla retta; i valori $y_i^* = b_0 + b_1 x_i$ sono dunque i valori teorici, ovvero i valori che la variabile Y assume, secondo il modello $Y = b_0 + b_1 X$, in corrispondenza dei valori $x_i$ osservati.

### rette passanti per la nube di punti




### Determinazione della retta di regressione

L'identificazione della retta avviene attraverso la determinazione dei valori di $b_0$, l'intercetta, e $b_1$, il coefficiente angolare o pendenza. La retta 'migliore' è quella che passa più 'vicina' ai punti osservati. In altre parole, si vuole trovare la retta per la quale le differenze tra i valori teorici $y_i^*$ e i valori osservati $y_i$ siano minime.

# La retta di regressione

## I residui

le differenze tra i valori teorici $y_i^*$ e i valori osservati $y_i$ vengono definite residui. La retta di regressione è tale che la somma dei residui al quadrato sia minima. Formalmente

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Il problema consiste dunque nel ricercare $b_0$ e $b_1$ che minimizzano la precedente espressione. Da un punto di vista operativo bisogna risolvere il seguente sistema di equazioni

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

## Ricerca dei parametri della retta di regressione:($b_0$)

$$-2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) =$$

$$\sum_{i=1}^{n} y_i - n * b_0 - b_1 \sum_{i=1}^{n} x_i = 0$$

$$b_0 = \mu_y - b_1 \mu_x$$

# La retta di regressione

## I residui

le differenze tra i valori teorici $y_i^*$ e i valori osservati $y_i$ vengono definite residui. La retta di regressione è tale che la somma dei residui al quadrato sia minima. Formalmente

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i^*)^2 =$$

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Il problema consiste dunque nel ricercare $b_0$ e $b_1$ che minimizzano la precedente espressione. Da un punto di vista operativo bisogna risolvere il seguente sistema di equazioni

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

## Ricerca dei parametri della retta di regressione:($b_1$)

$$-2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \left( \frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} \right)$$

$$b_1 \left( n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 \right) = n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

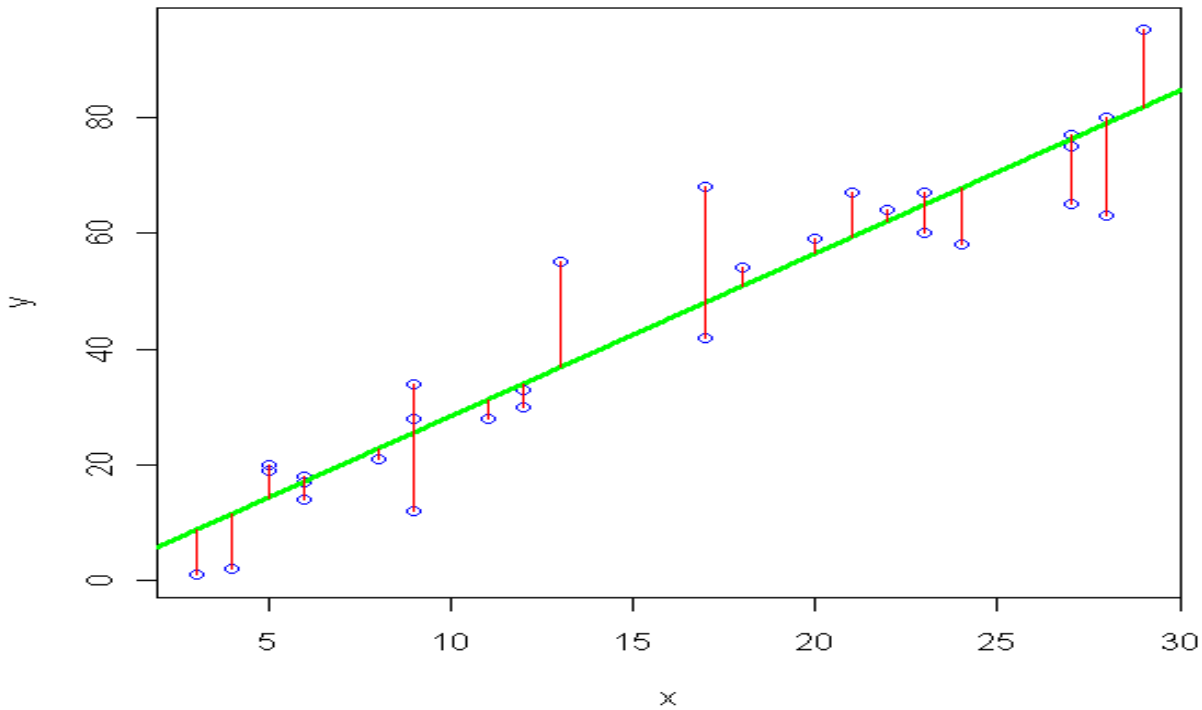
# Determinazione della retta di regressione

## Calcolo dei coefficienti

Richiamando le quantità calcolate in precedenza e le formule per il calcolo dei parametri si ha

- $b_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{205.04}{(8.55)^2} = \frac{205.04}{73.09889} = 2.804967$
- $b_0 = \mu_y - b_1 \mu_x = 44.2 - (2.804967 * 15.63) = 0.349$

## La retta 'migliore'



# Interpretazione dei valori dei coefficienti di regressione



- $b_0$ rappresenta l'intercetta della retta di regressione ed indica il valore della variabile di risposta $Y$ quando il predittore $X$ assume valore 0.
- $b_1$ rappresenta l'inclinazione della retta di regressione, ovvero la variazione della variabile di risposta $Y$ in conseguenza di un aumento unitario del predittore $X$.

# Bontà di adattamento

Esistono diversi strumenti grafici ed analitici per valutare la bontà dell'adattamento della retta di regressione ai dati

- Strumenti grafici: plot dei residui
- Strumenti analitici:coefficiente di determinazione lineare $R^2$

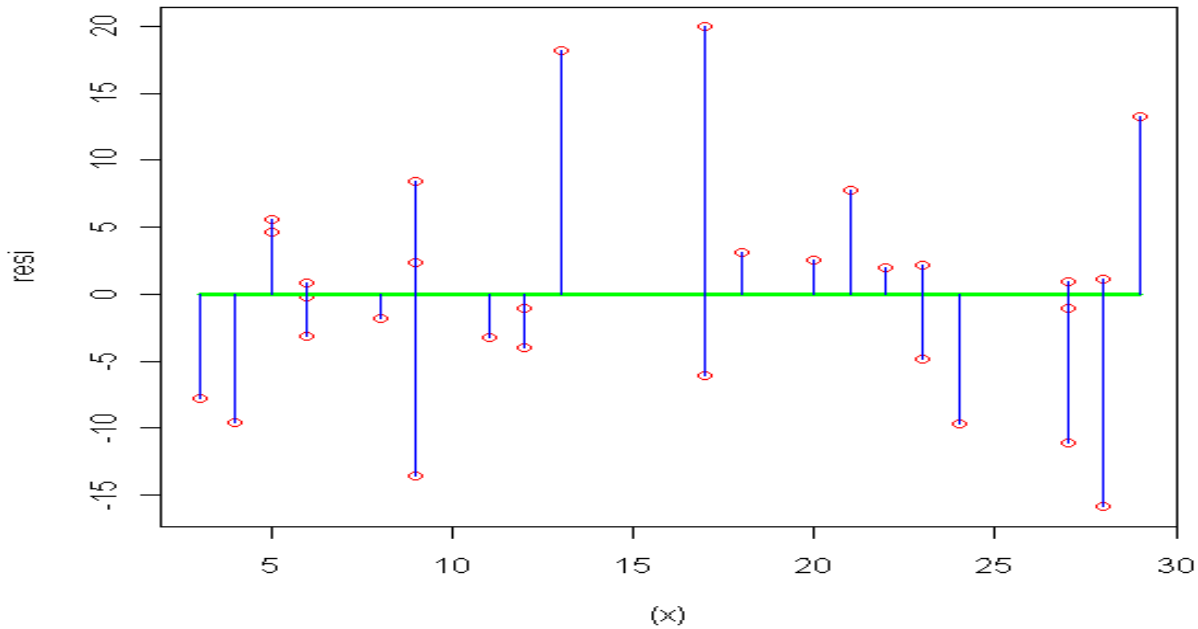
# Plot dei residui



Perchè la retta possa essere considerata una buona approssimazione della relazione che intercorre tra $Y$ ed $X$ è necessario che i residui abbiano un andamento casuale rispetto ai valori della $X$. Se, ad esempio, all'aumentare dei valori della $X$ aumentassero sistematicamente anche i residui, allora la relazione potrebbe non essere non lineare: la retta di regressione ne sarebbe dunque una cattiva approssimazione.

## Plot dei residui

Per verificare che l'andamento dei residui sia effettivamente casuale rispetto ad $X$, è possibile utilizzare un diagramma di dispesione tra i valori $x_i$ ed i corrispondenti residui $e_i (i = 1, \ldots, n)$



# coefficiente di determinazione lineare $R^2$

Ricordando che la devianza il numeratore della varianza...

$$Dev_y = \sum_{i=1}^{n} (y_i - \mu_y)^2 = \sum_{i=1}^{n} (y_i - y_i^* + y_i^* - \mu_y)^2 =$$
$$= \sum_{i=1}^{n} (y_i - y_i^*)^2 + \sum_{i=1}^{n} (y_i^* - \mu_y)^2 + 2\sum_{i=1}^{n} (y_i - y_i^*)(y_i^* - \mu_y)$$
$$= \sum_{i=1}^{n} (y_i - y_i^*)^2 + \sum_{i=1}^{n} (y_i^* - \mu_y)^2 + 2(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i^*)(\sum_{i=1}^{n} y_i^* - n\mu_y)$$

Il metodo dei minimi quadrati assicura che $\sum_{i=1}^{n} y_i^* = \sum_{i=1}^{n} y_i$, quindi

$$Dev(y) = \sum_{i=1}^{n} (y_i - y_i^*)^2 + \sum_{i=1}^{n} (y_i^* - \mu_y)^2 + 2 * 0 * (\sum_{i=1}^{n} y_i^* - n\mu_y)$$
$$= \sum_{i=1}^{n} (y_i^* - \mu_y)^2 + \sum_{i=1}^{n} (y_i - y_i^*)^2 = Dev_r + Dev_e$$

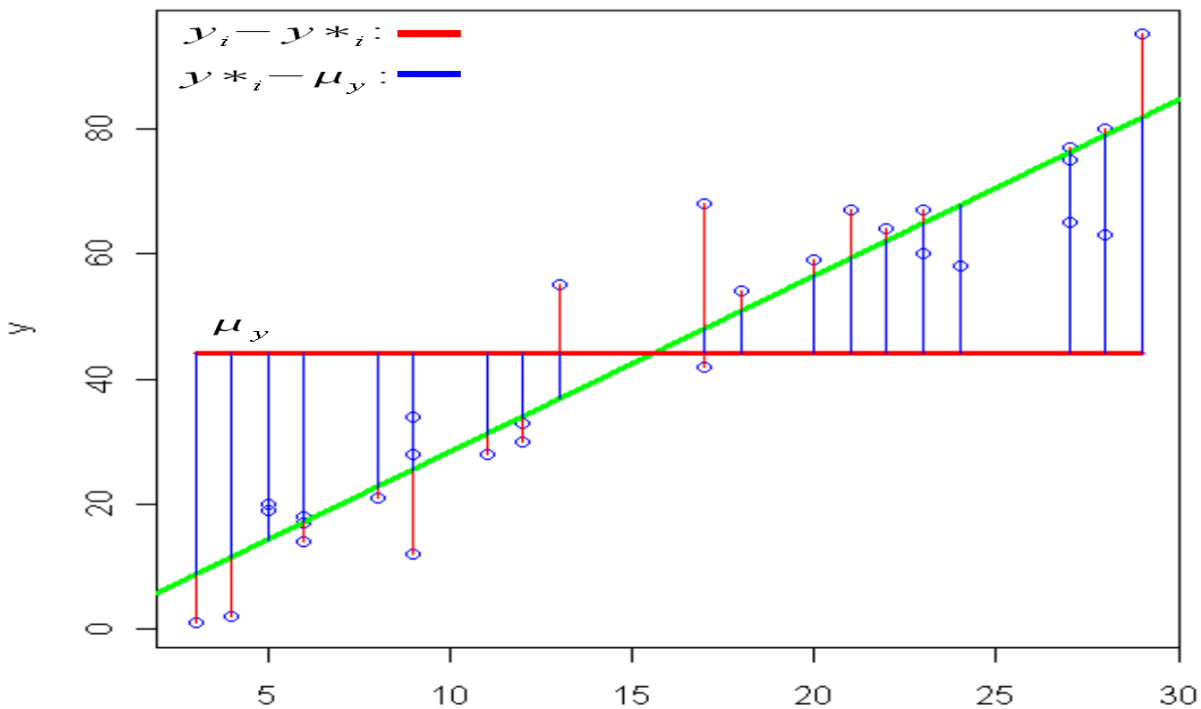
# Decomposizione della devianza



La devianza può essere decomposta dunque nelle seguenti quantità $Dev_y = Dev_r + Dev_e$

- $Dev_y = \sum_{i=1}^{n} (y_i - \mu_y)^2$ devianza totale
- $Dev_r = \sum_{i=1}^{n} (y_i^* - \mu_y)^2$ devianza di regressione
- $Dev_e = \sum_{i=1}^{n} (y_i - y_i^*)^2$ devianza dei residui

Interpretazione grafica



# Bontà dell'adattamento

Intituitivamente, l'adattamento della retta è migliore quanto maggiore sarà proporzione di variabilità totale che la retta di regressione riesce a spiegare; ovvero, l'adattamento della retta è migliore quanto minore sarà la variabilità residua. Una misura di come il modello approssima i dati osservati è data dal coefficiente di determinazione lineare $R^2$, dato da

$$R^2 = \frac{Dev_r}{Dev_y} = \frac{\sum_{i=1}^{n} (y_i^* - \mu_y)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

ovvero

$$R^2 = 1 - \frac{Dev_e}{Dev_y} = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^*)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

esempio di calcolo $R^2$

- $Dev_y = \sum_{i=1}^{n} (y_i - \mu_y)^2 = 19284.8$
- $Dev_r = \sum_{i=1}^{n} (y_i^* - \mu_y)^2 = 17253.92$
- $Dev_e = \sum_{i=1}^{n} (y_i - y_i^*)^2 = 2030.885$

$$R^2 = \frac{Dev_r}{Dev_y} = \frac{17253.92}{19284.8} = 0.8947$$

ovvero

$$R^2 = 1 - \frac{Dev_e}{Dev_y} = 1 - \frac{203.885}{19284.8} = 1 - 10.53 = 0.8947$$

# Influenza di un outlier sulla soluzione



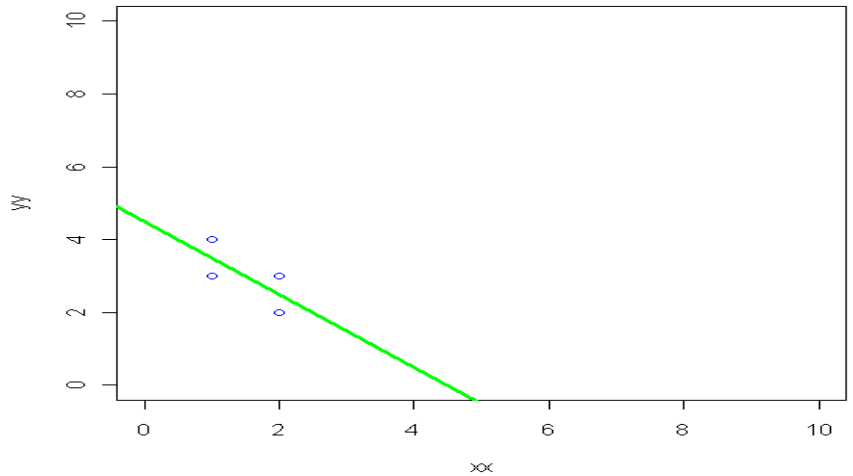
## Un piccolo esempio

Si considerino le seguenti osservazioni

| xx | yy |
|---|---|
| 1 | 4 |
| 1 | 3 |
| 2 | 3 |
| 2 | 2 |

## Retta di regressione

La soluzione induce a concludere che vi sia una relazione di proporzionalità inversa: poichè la retta è decrescente si deduce che all'aumentare di $X$, la variabile dipendente $Y$ diminuisce.



# Influenza di un outlier sulla soluzione

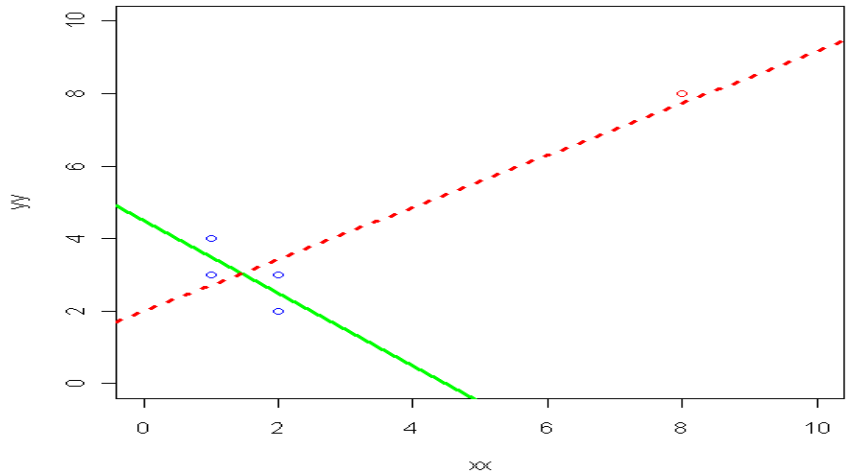


## Un (altro) piccolo esempio

Si considerino le osservazioni precedenti a cui è aggiunta un unica coppia di valori $(8, 8)$. I dati sono

| xx2 | yy2 |
|---|---|
| 1 | 4 |
| 1 | 3 |
| 2 | 3 |
| 2 | 2 |
| 8 | 8 |

## Retta di regressione




In questo caso, la sola presenza della nuova osservazione conduce all'identificazione di una retta di regressione diversa dalla prima: l'inclinazione positiva della retta indica una relazione di diretta proporzionalità. Tuttavia tale soluzione è unicamente dovuta dalla presenza dell' osservazione $(8, 8)$ che pertanto induce a valutare la relazione di dipendenza tra $Y$ ed $X$ in maniera errata. L'osservazione $(8, 8)$ si definisce pertanto un outlier. L'identificazione e la conseguente eliminazione degli eventuali outlier è un elemento molto importante nello studio della dipendenza tra fenomeni.