STATISTICA

Regressione-2

Esempio

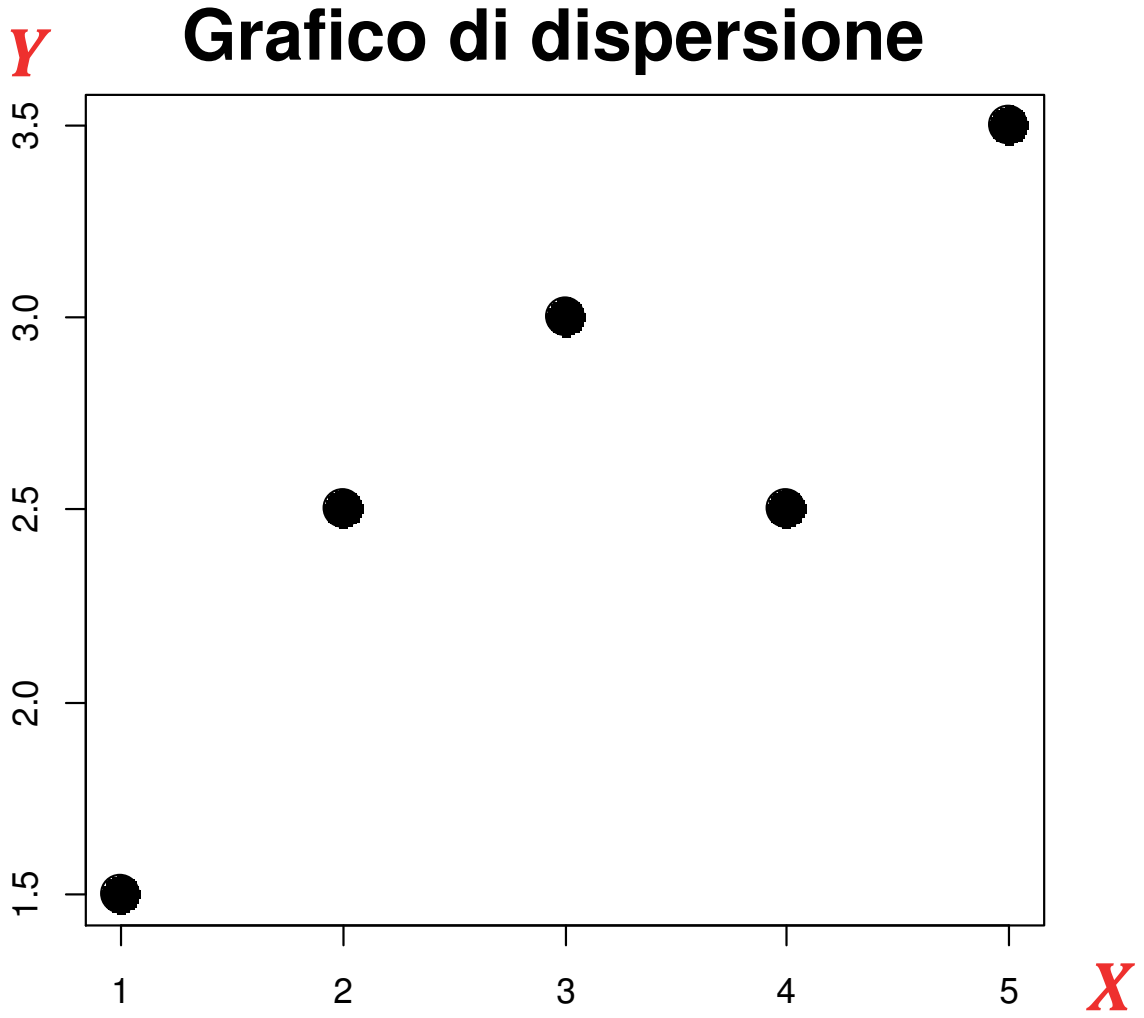
Su un campione di $n = 5$ unità sono state osservate due variabili, X ed Y:

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5

1. Rappresentare l'andamento congiunto **di Y in funzione di X** mediante un opportuno grafico: le due variabili sono correlate? Come e quanto?
2. Determinare le stime dei minimi quadrati dei parametri della retta interpolatrice dei dati

Esempio

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5



Esempio

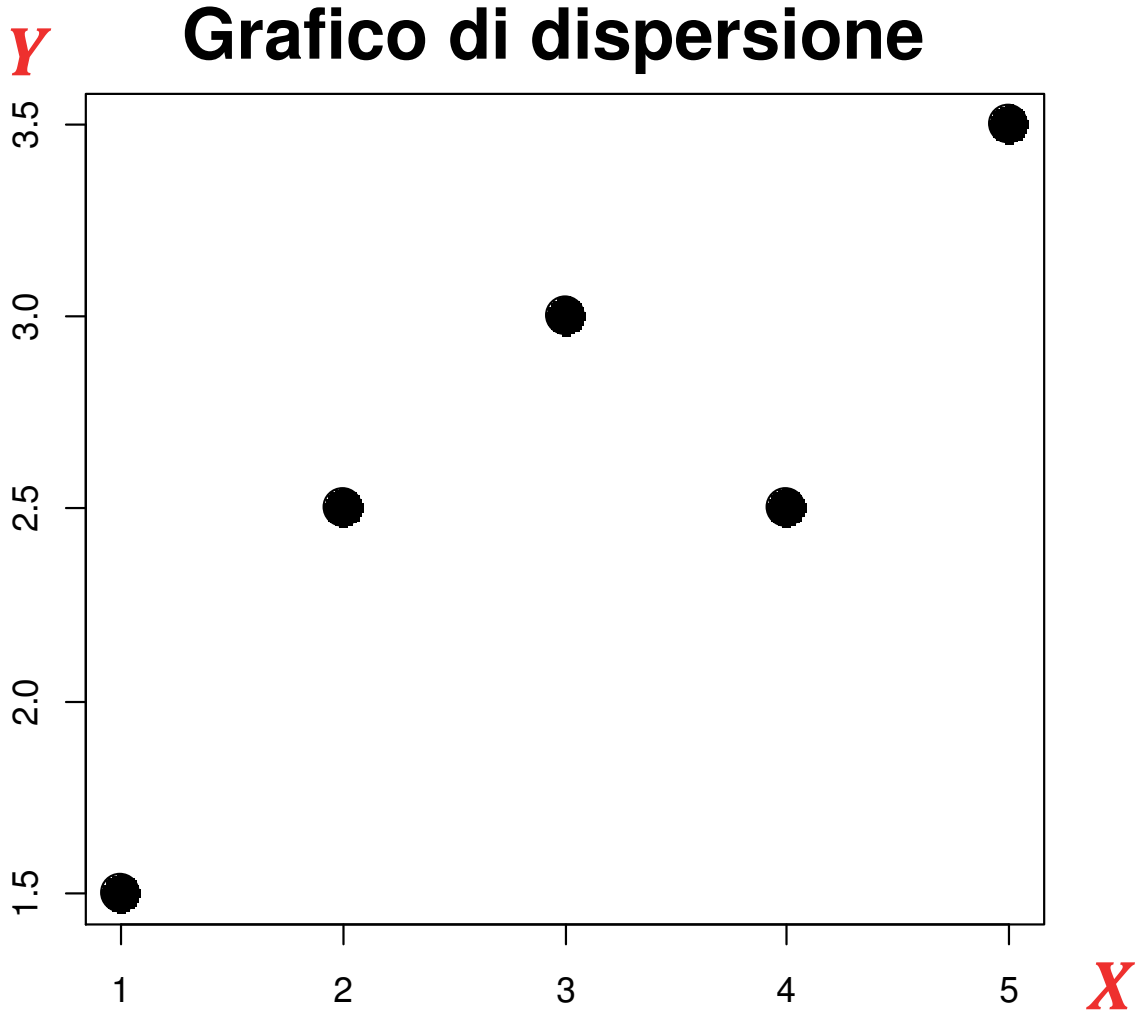
x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5

$\overline{x}, \overline{y}$

σ_x^2 , σ_y^2

σ_{xy}

ρ_{xy}



Esempio

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5
$x_i y_i$	1.5	5	9	10	17.5

$$\bar{x} = 3, \bar{y} = 2.6$$

$$\sigma_x{}^2 = 11 - 3^2 = 2\,,$$
$$\sigma_y{}^2 = 7.2 - 2.6^2 = 0.44$$

$$\sigma_x{}^2 = \left[\frac{1}{n}\sum x_i{}^2\right] - \bar{x}^2$$

$$\sigma_{xy} = \frac{1}{5}\sum_i x_i y_i - \bar{x}\,\bar{y} = \frac{43}{5} - 3 \times 2.6 = 0.8$$

$$\boldsymbol{\rho_{xy}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{0.8}{\sqrt{2 \times 0.44}} = \mathbf{0.85}$$

Esempio

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5

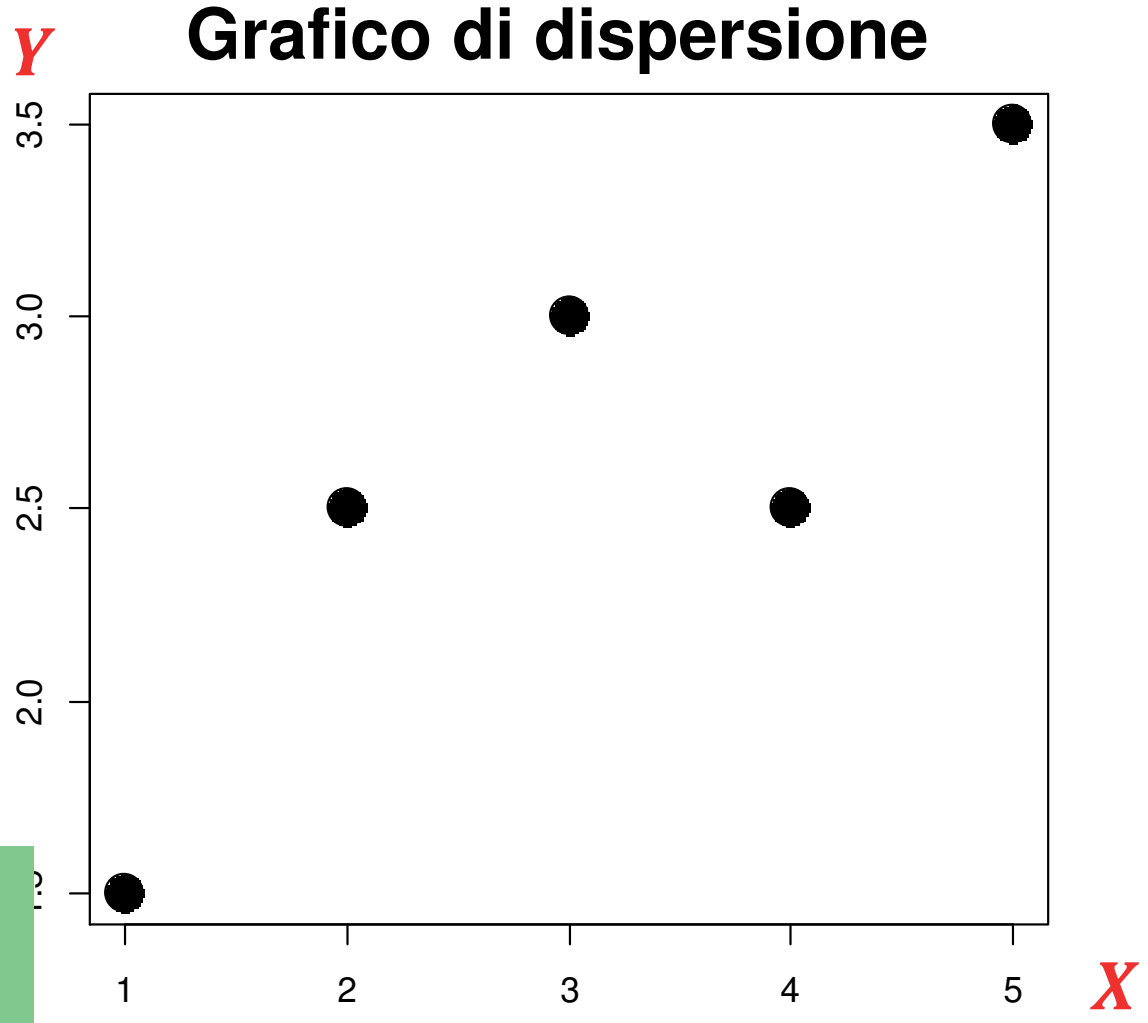
$\overline{x}, \overline{y}$

σ_x^2 , σ_y^2

σ_{xy}

$\rho_{xy} = 0.85$

Correlazione lineare positiva



Esempio

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5
$x_i y_i$	1.5	5	9	10	17.5

$$\bar{x} = 3, \bar{y} = 2.6$$

$$\sigma_x{}^2 = 11 - 3^2 = 2\,,$$
$$\sigma_y{}^2 = 7.2 - 2.6^2 = 0.44$$

$$\sigma_x{}^2 = \left[\frac{1}{n}\sum x_i{}^2\right] - \bar{x}^2$$

$$\sigma_{xy} = \frac{1}{5}\sum_i x_i y_i - \bar{x}\,\bar{y} = \frac{43}{5} - 3 \times 2.6 = \mathbf{0.8}$$

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x{}^2} = \frac{0.8}{2} = \mathbf{0.4}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 2.6 - 0.4 \times 3 = \mathbf{1.4}$$

Esempio

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5

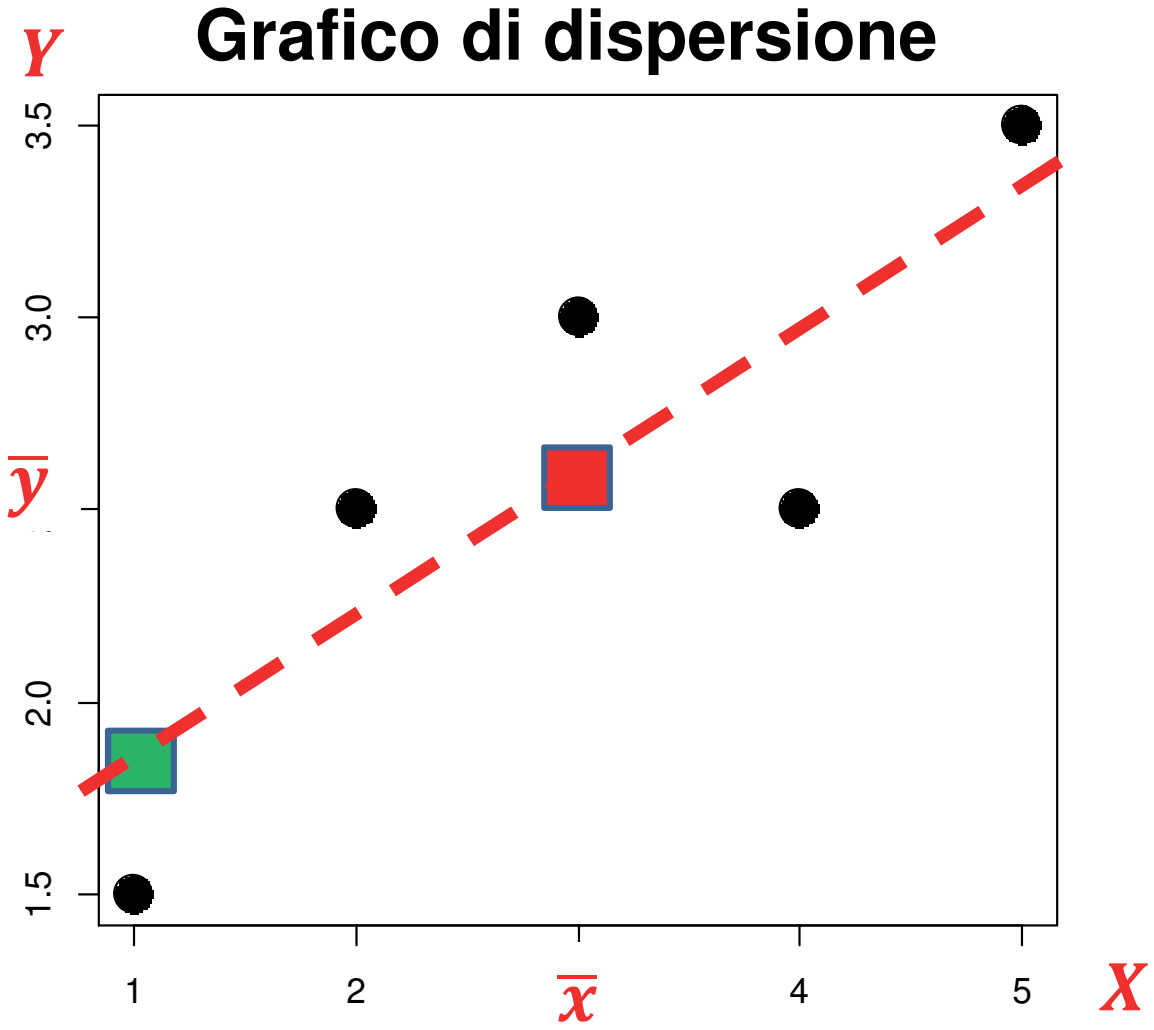
$\bar{x} = 3, \bar{y} = 2.6$

$\hat{b} = 0.4$

$\hat{a} = 1.4$

$\boldsymbol{y = 1.4 + 0.4x}$

$\boldsymbol{x = 1,}$
$\boldsymbol{y = 1.4 + 0.4 = 1.8}$



Esempio

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5

$\bar{x} = 3, \bar{y} = 2.6$

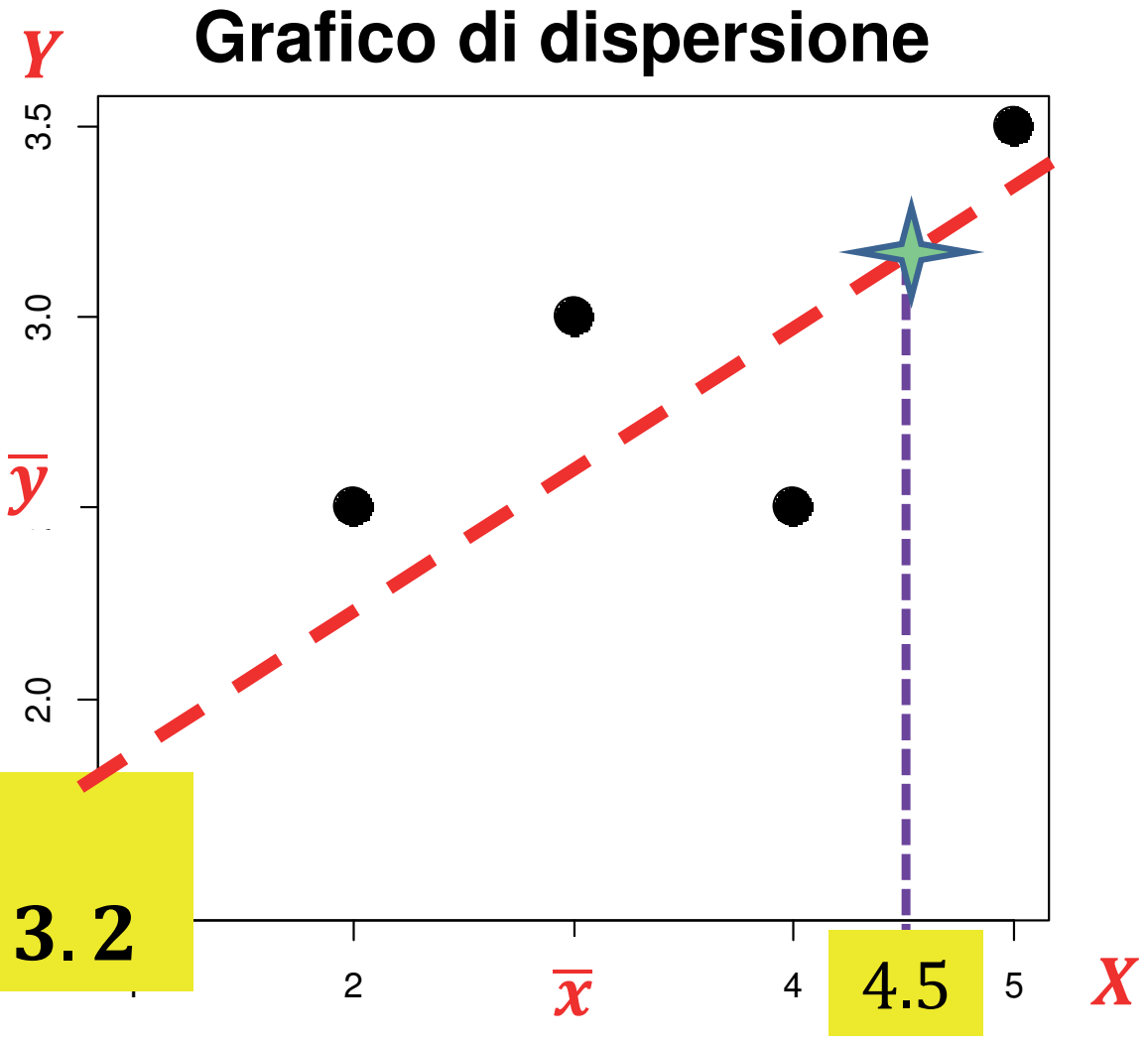
$\hat{b} = 0.4$

$\hat{a} = 1.4$

$$y = 1.4 + 0.4x$$

Previsione:

$$x = 4.5,$$
$$y = 1.4 + 0.4 \times 4.5 = 3.2$$



Esempio

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5

$\bar{x} = 3, \bar{y} = 2.6$

$\hat{b} = 0.4$

$\hat{a} = 1.4$

$y = 1.4 + 0.$

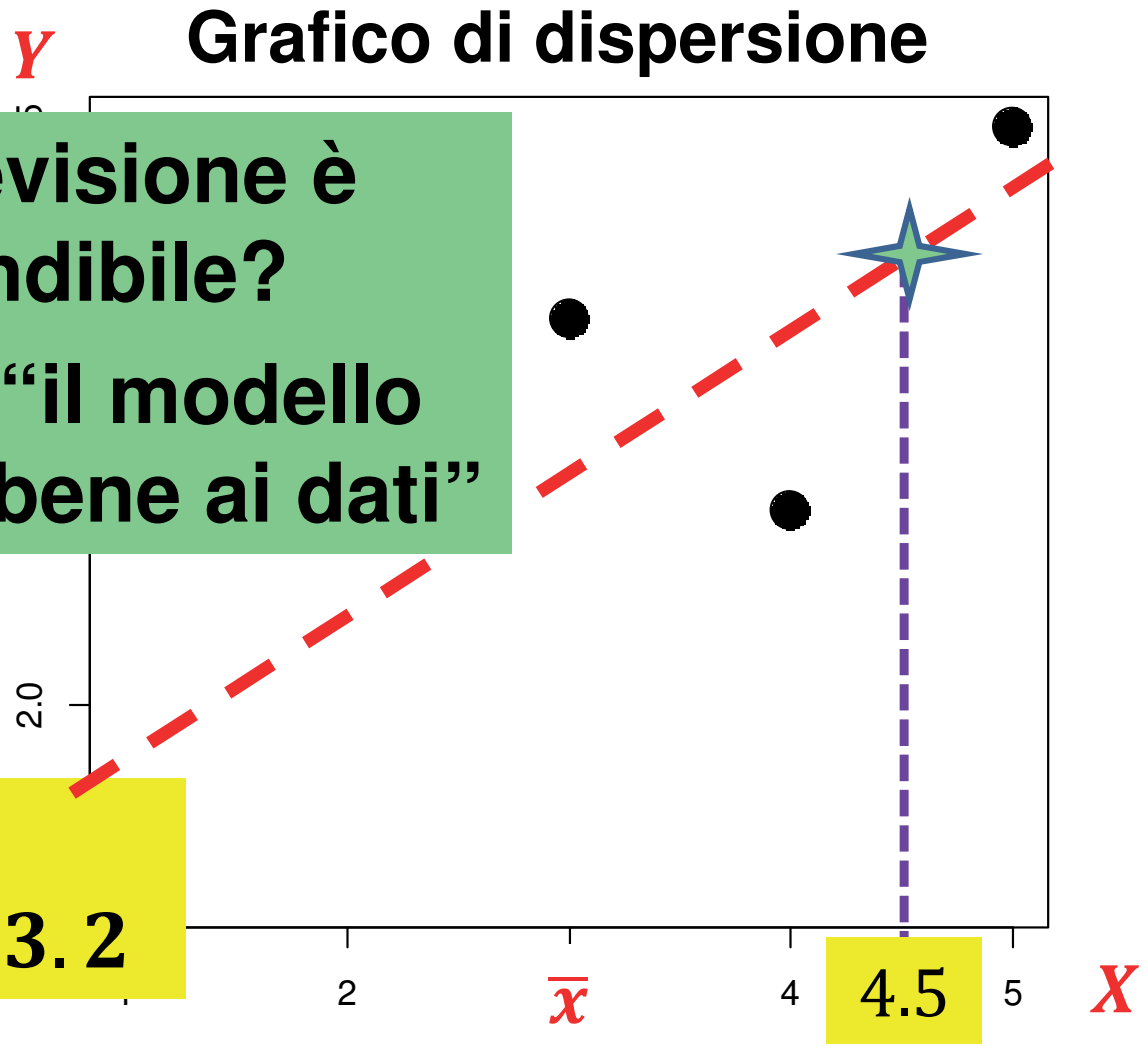
La previsione è attendibile?

solo se "il modello si adatta bene ai dati"

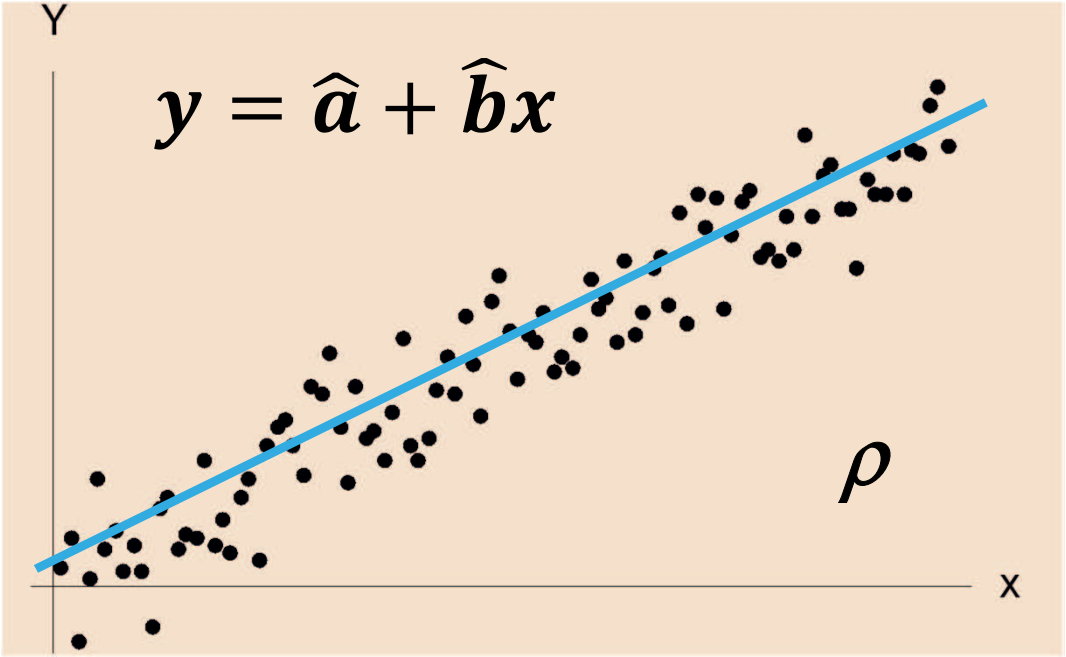
Previsione:

$\boldsymbol{x = 4.5,}$

$\boldsymbol{y = 1.4 + 0.4 \times 4.5 = 3.2}$



Regressione lineare



A. Valutazione preliminare se una retta possa essere una buona approssimazione

B. Stima dei parametri della retta.

C. Valutazione della bontà di adattamento del modello ai dati

Esempio

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5

COME LO VERIFICO?

CONFRONTANDO LE OSSERVAZIONI

y_i

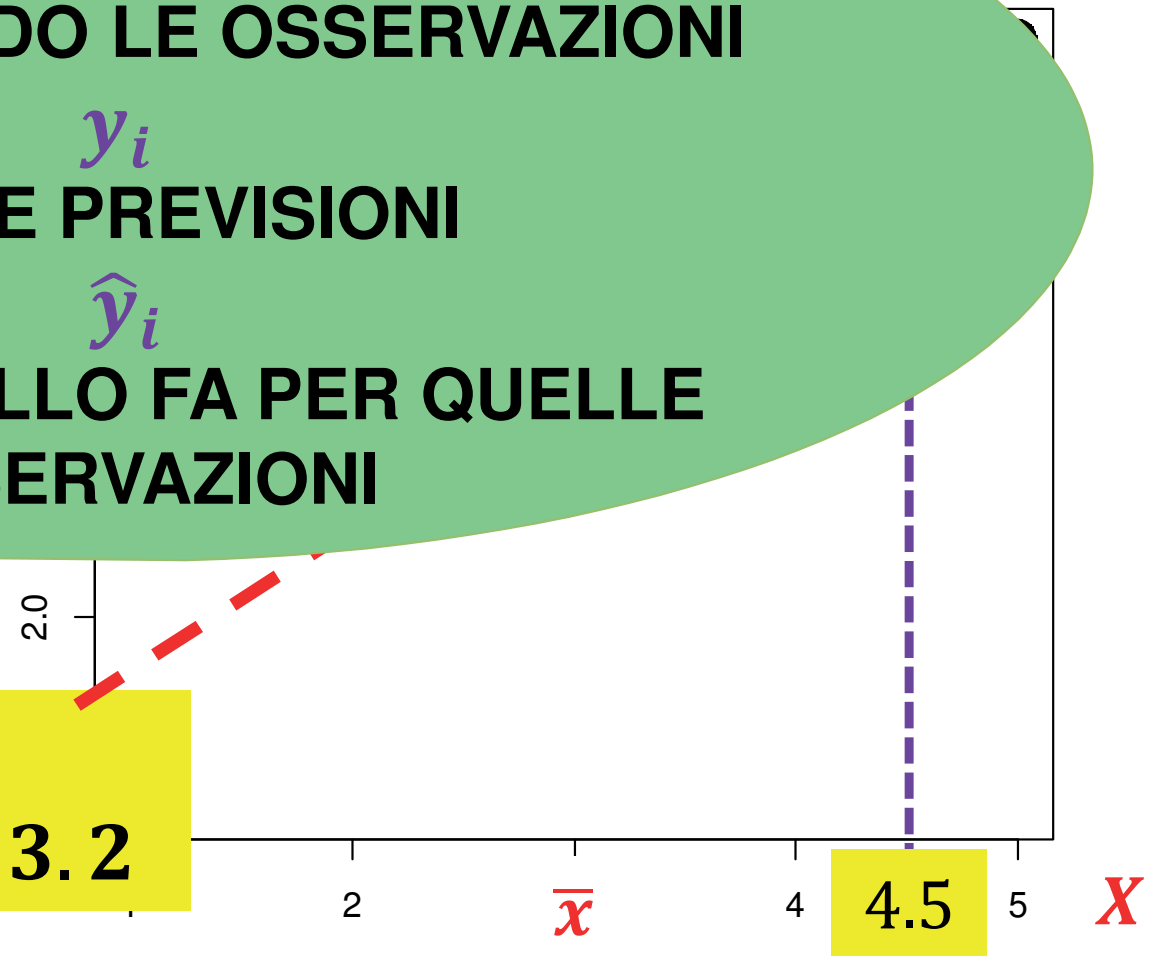
CON LE PREVISIONI

$\hat{y}_i$

CHE IL MODELLO FA PER QUELLE OSSERVAZIONI

$\bar{x} =$

$y = 1.4$

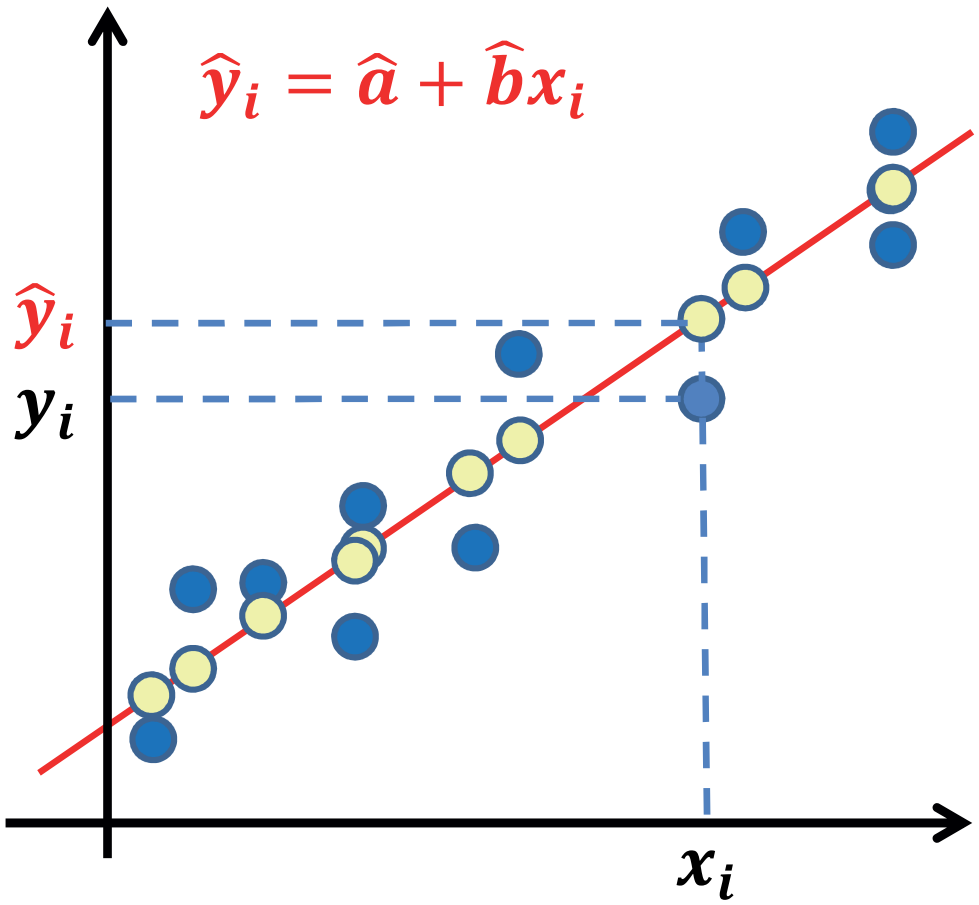


Previsione:

$x = 4.5,$

$y = 1.4 + 0.4 \times 4.5 = 3.2$

La varianza spiegata dalla retta

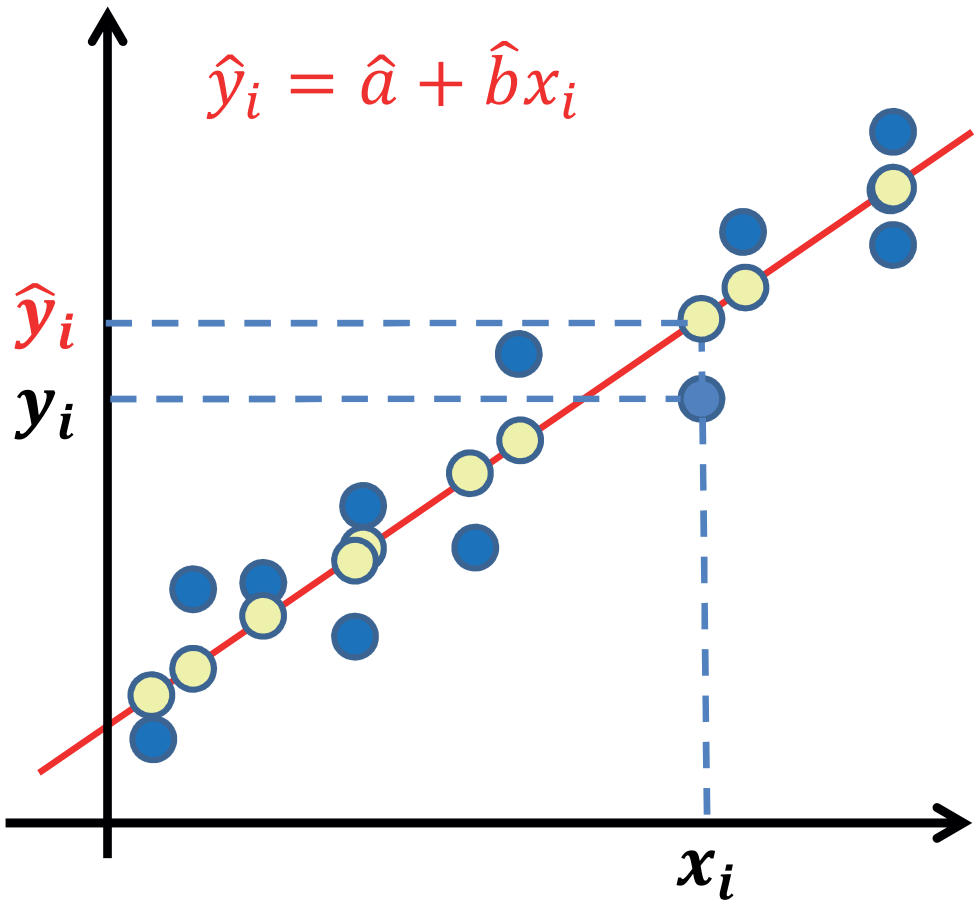


$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{a} + \hat{b}x_i) =$$

$$= \hat{a} + \hat{b}\bar{x} = (\bar{y} - \hat{b}\bar{x}) + \hat{b}\bar{x} = \bar{y}$$

la media delle previsioni
coincide con
la media delle osservazioni

La varianza spiegata dalla retta



$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{a} + \hat{b}x_i) =$$

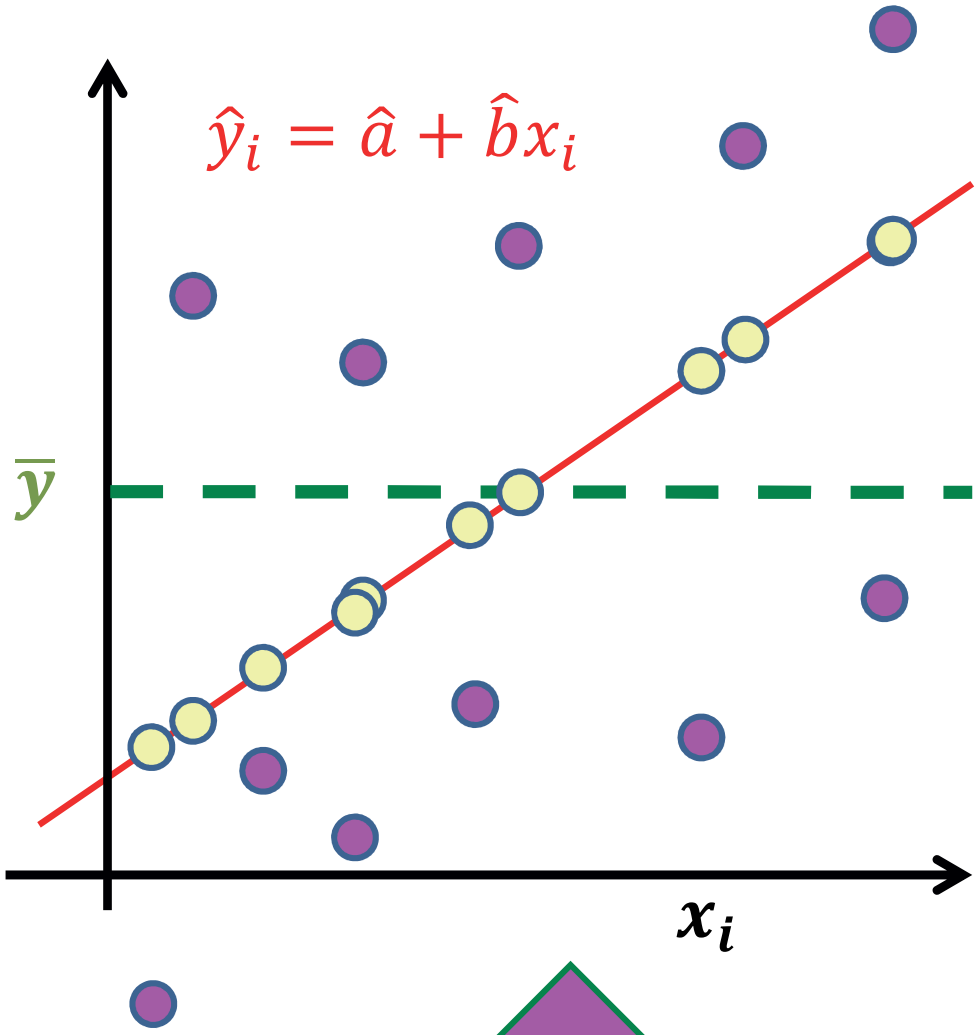
$$= \hat{a} + \hat{b}\bar{x} = (\bar{y} - \hat{b}\bar{x}) + \hat{b}\bar{x} = \bar{y}$$

la media delle previsioni
coincide con
la media delle osservazioni

se le previsioni sono molto vicine alle osservazioni anche le varianze saranno vicine!

$$\sigma_y{}^2 = \frac{1}{n}\Sigma(y_i - \bar{y})^2 \approx \sigma_{\hat{y}}{}^2 = \frac{1}{n}\Sigma(\hat{y}_i - \bar{y})^2$$

La varianza spiegata dalla retta



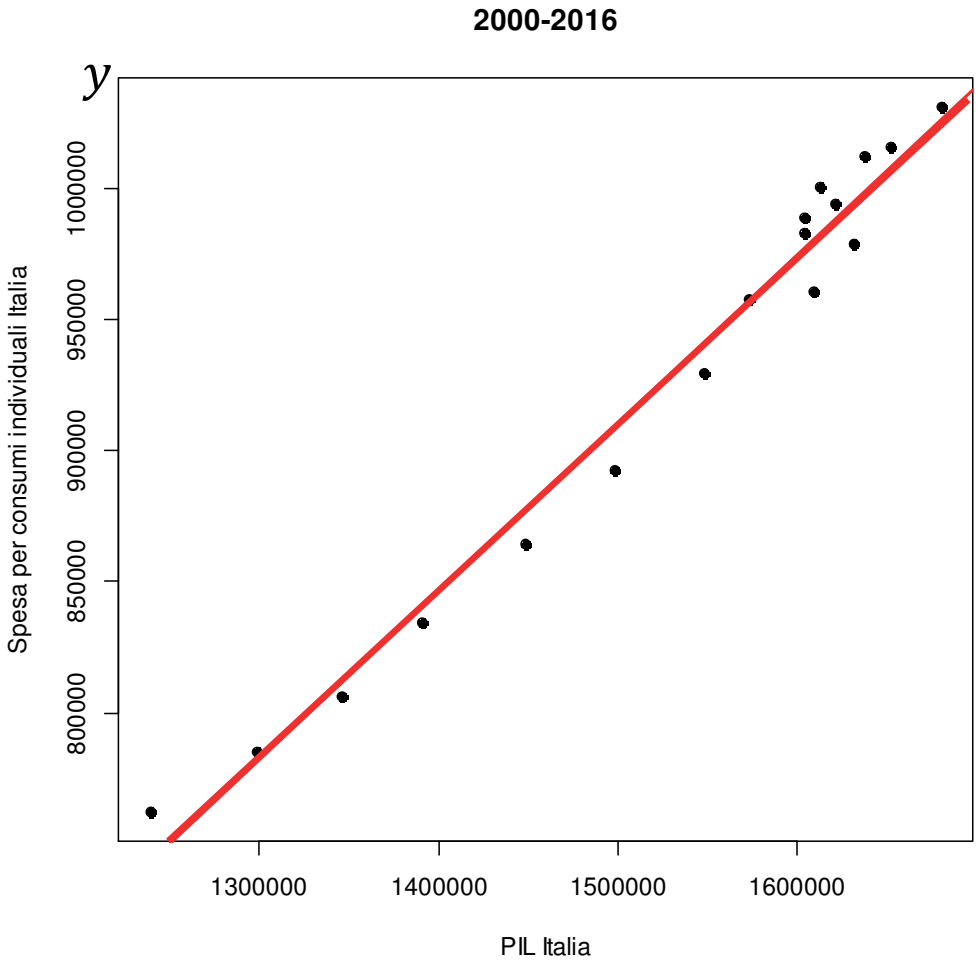
$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{a} + \hat{b}x_i) =$$

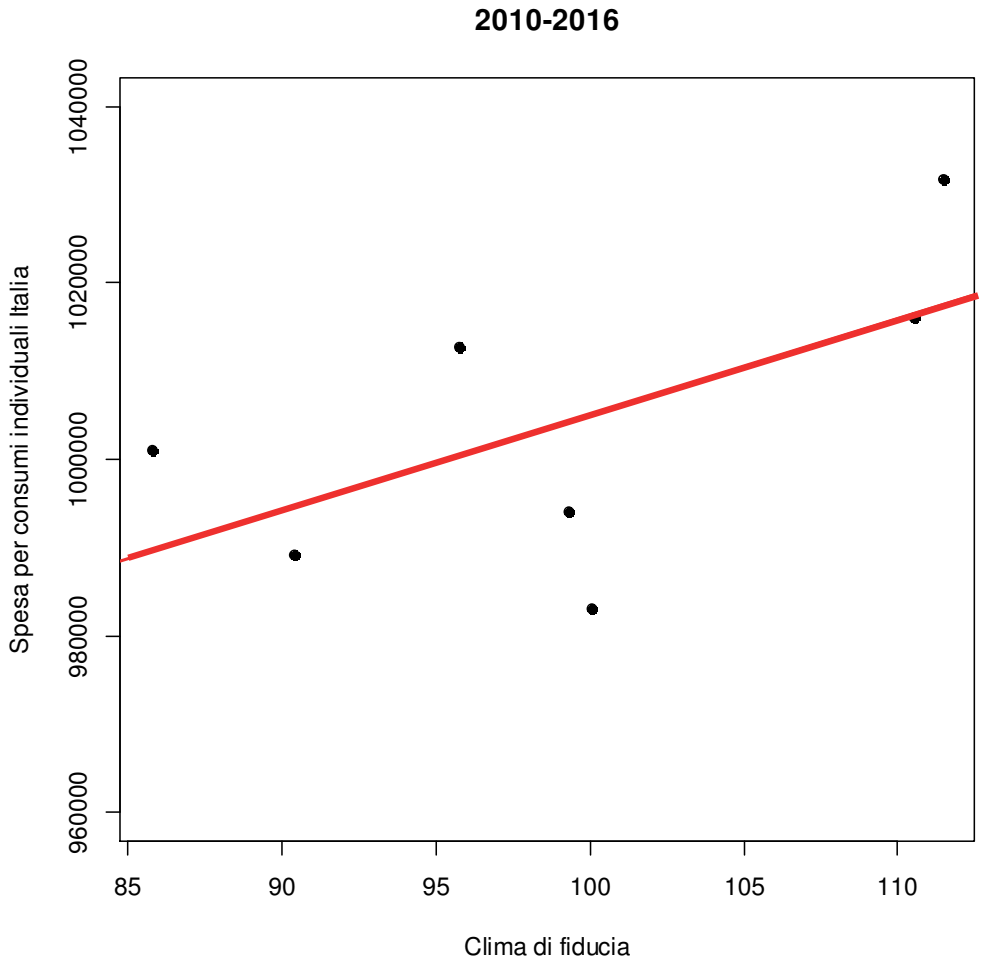
$$= \hat{a} + \hat{b}\bar{x} = (\bar{y} - \hat{b}\bar{x}) + \hat{b}\bar{x} = \bar{y}$$

la media delle previsioni
coincide con
la media delle osservazioni

se le previsioni **non** sono molto vicine alle osservazioni, la **varianza dei dati** sarà **maggiore** di quella delle previsioni

La varianza spiegata dalla retta

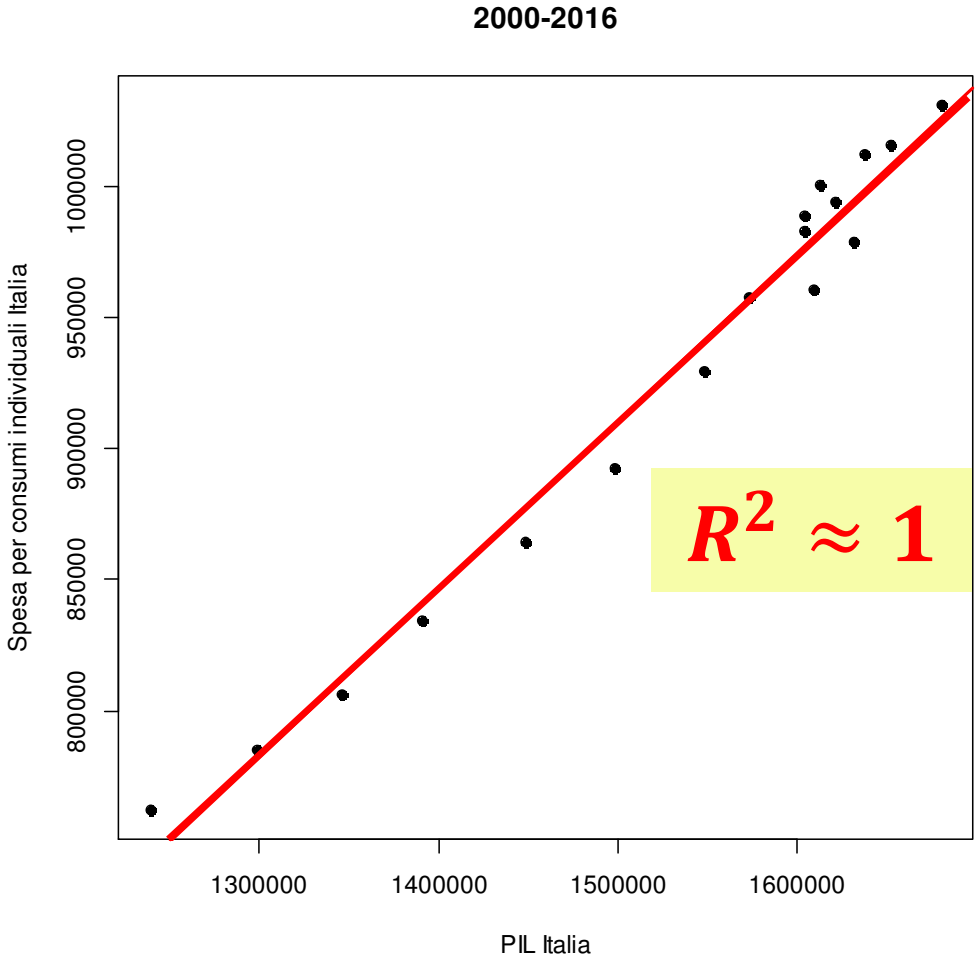




in quale grafico le previsioni e le osservazioni sono più vicine?

La varianza spiegata dalla retta

$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \bar{y}$$



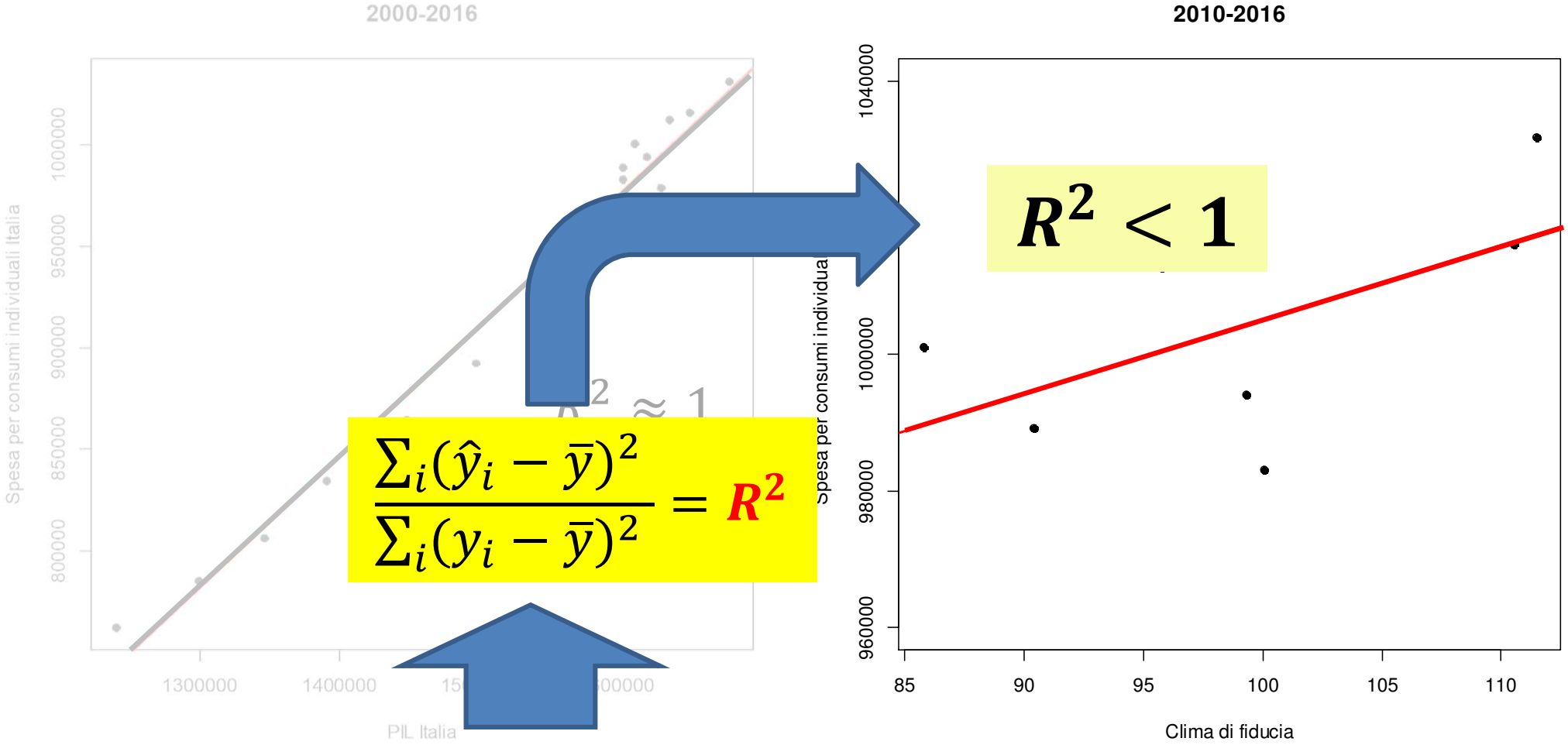
$$R^2 \approx 1$$

$$\frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2} = R^2$$

$$\sigma_y{}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2 \approx \sigma_{\hat{y}}{}^2 = \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2$$

La varianza spiegata dalla retta

$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \bar{y}$$



$$R^2 \approx 1$$

$$\frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2} = R^2$$

$$R^2 < 1$$

$$\sigma_y{}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2 \quad > \quad \sigma_{\hat{y}}{}^2 = \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2$$

Analisi della Varianza

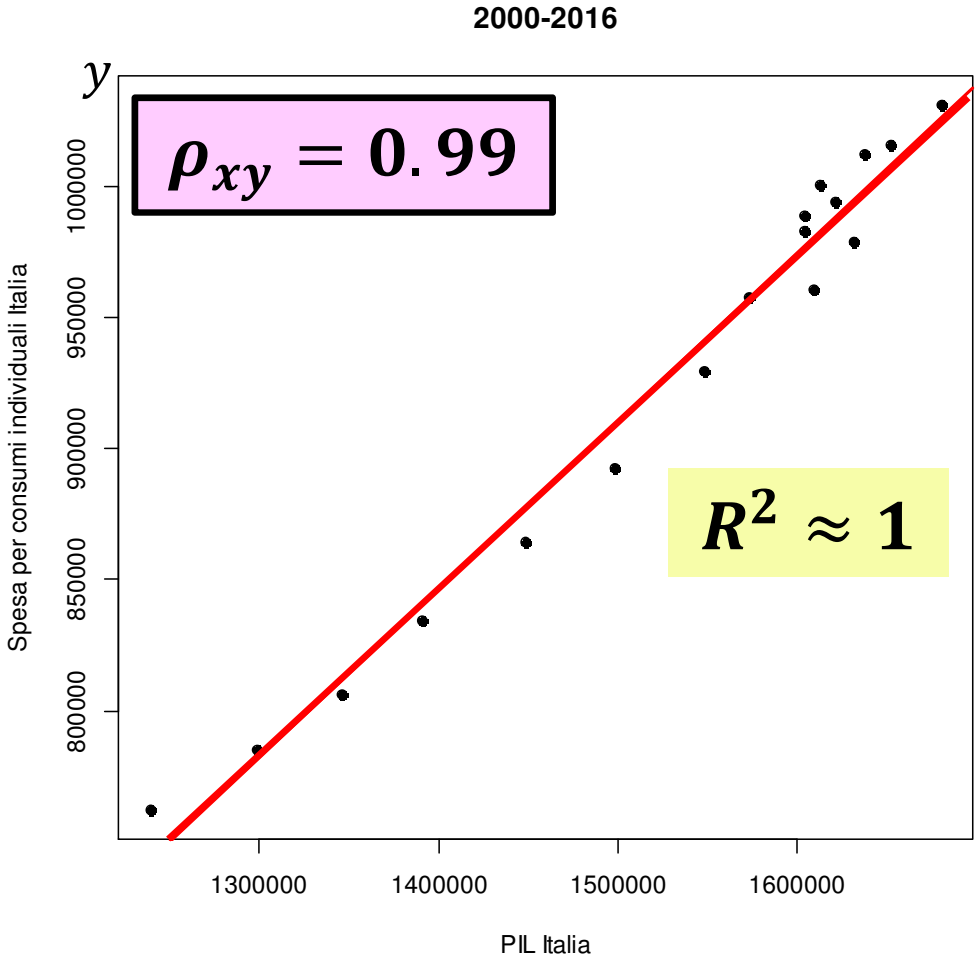
Fonte di variabilità	Gradi di libertà (gl)	SS (Sum of Squares)	Mean Square (SS/gl)
Retta di regressione	1	$\sum_i (\hat{y}_i - \bar{y})^2$	
Attorno alla retta	$n - 2$	$\sum_i (y_i - \hat{y}_i)^2$	$\frac{1}{n-2}\sum_{i=1}^{n} {e_i}^2$
Totale	$n - 1$	$\sum_i (y_i - \bar{y})^2$	

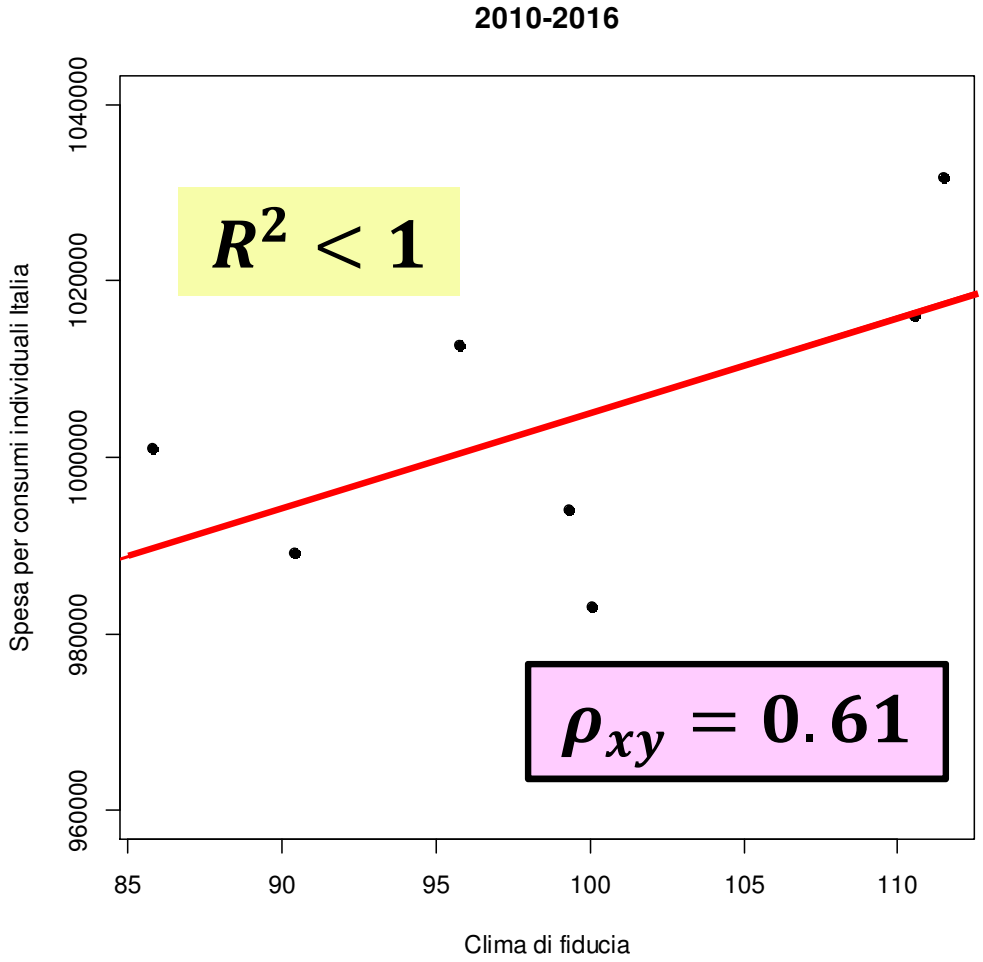
varianza ***spiegata***

varianza ***totale***

num. di parametri stimati (a e b)

La varianza spiegata dalla retta





in quale grafico le previsioni e le osservazioni sono più vicine?

La bontà della regressione

$$\frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \boldsymbol{R^2}$$

coeff. di determinazione

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

$$R^2 = \rho_{xy}{}^2$$

BUON ADATTAMENTO :

$$\boldsymbol{R^2 > 0.7 \Leftrightarrow \rho_{xy} > 0.837} \text{ o } \boldsymbol{\rho_{xy} < -0.837}$$

(per tendenze crescenti)

(per tendenze decrescenti)

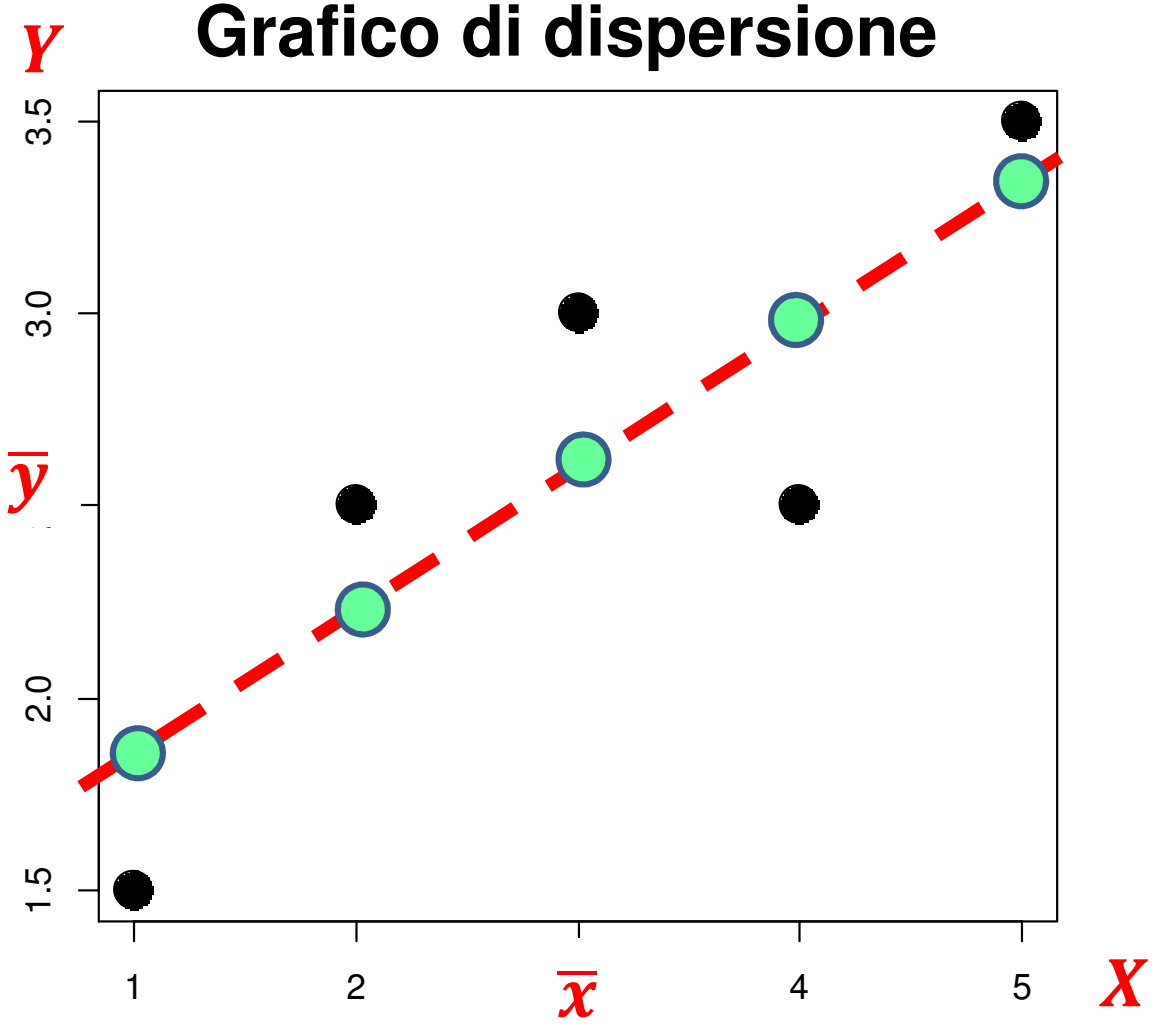
Esempio, Cont.

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5
$\hat{y}_i$	1.8	2.2	2.6	3.0	3.4

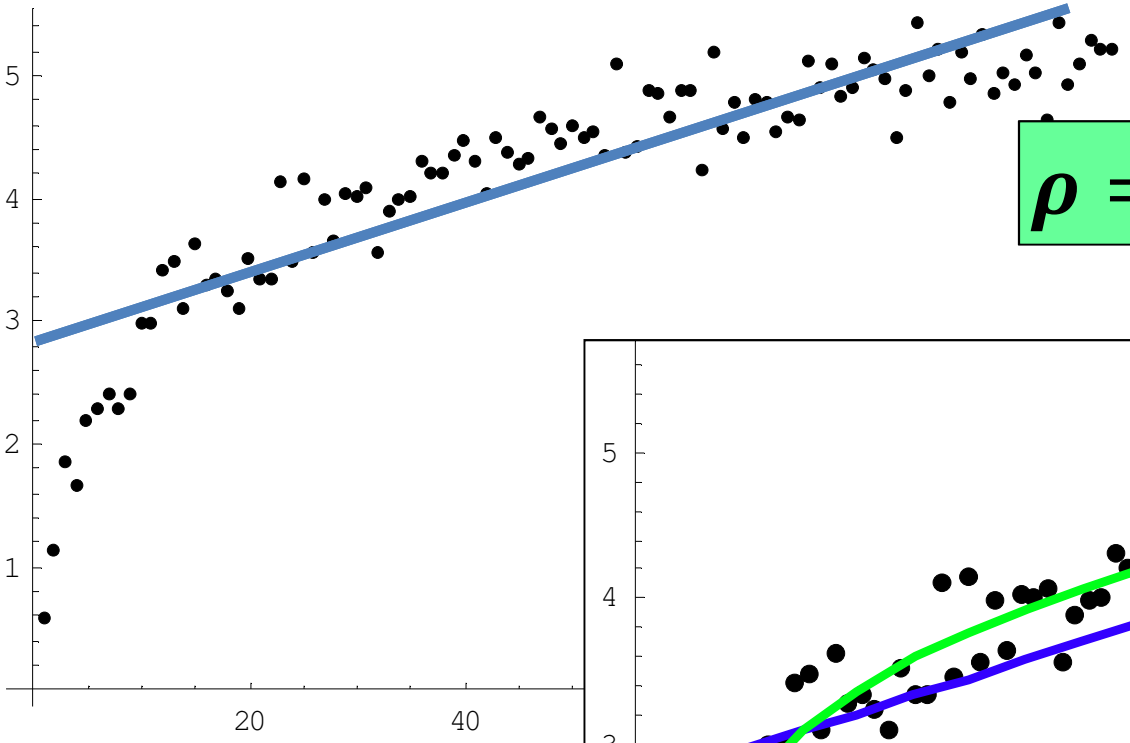
$$\hat{y}_i = 1.4 + 0.4x_i$$

$$\rho_{xy} = 0.85$$

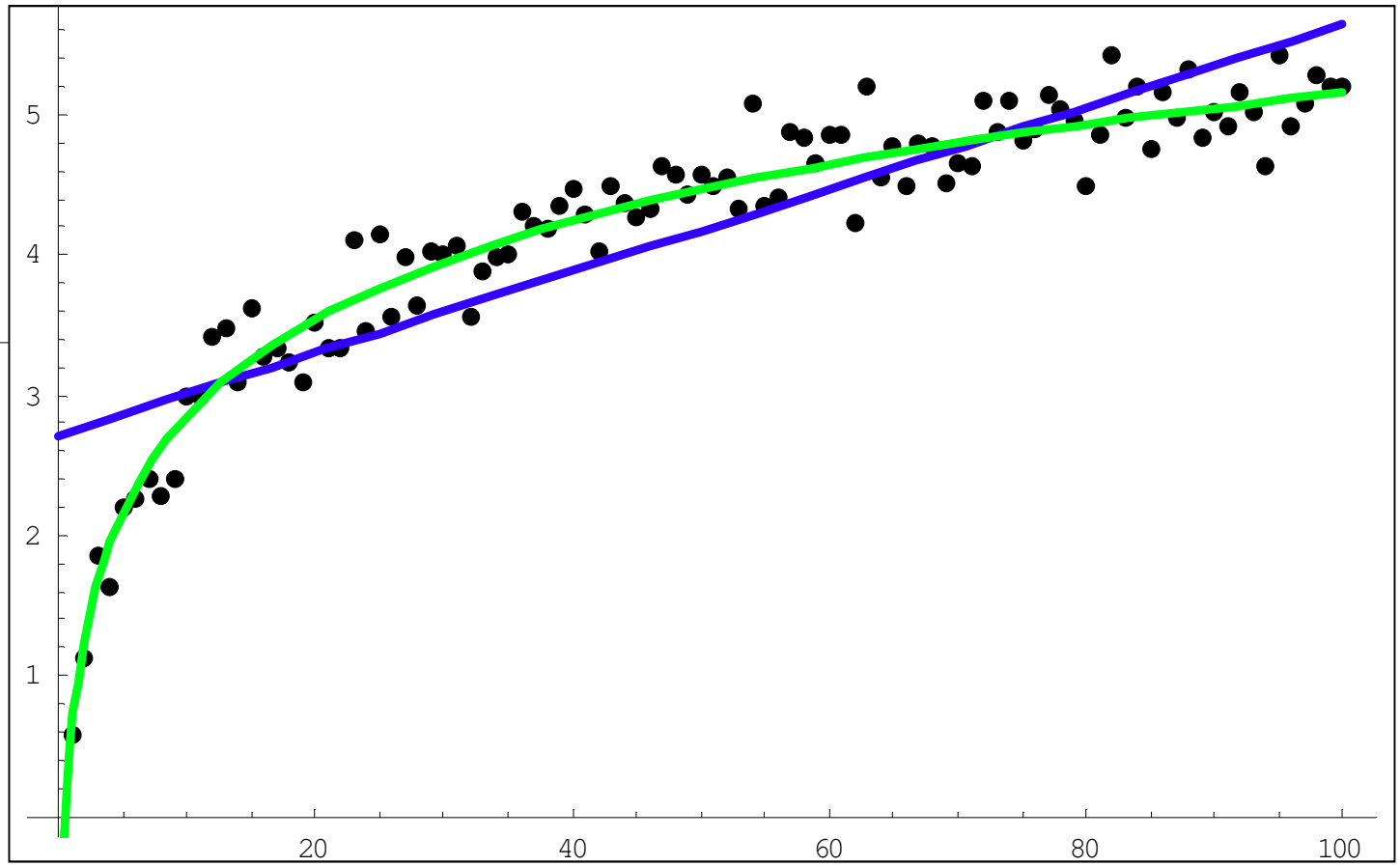
$$R^2 = \frac{\sigma(\hat{y})^2}{\sigma_y{}^2} = \rho^2 = 0.85^2 = 0.72$$



L’indice non basta!



$$\rho = 0.87 \Rightarrow R^2 = 0.76$$



Regressione lineare: sunto

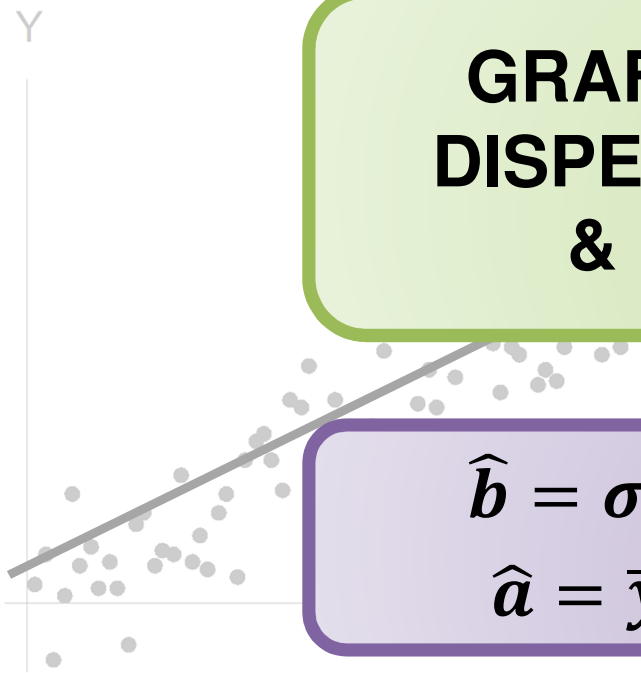


GRAFICO DI DISPERSIONE
& ρ_{xy}

A. Valutazione preliminare se una retta possa essere una buona approssimazione

$$\hat{b} = \sigma_{xy}/{\sigma_x}^2$$
$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

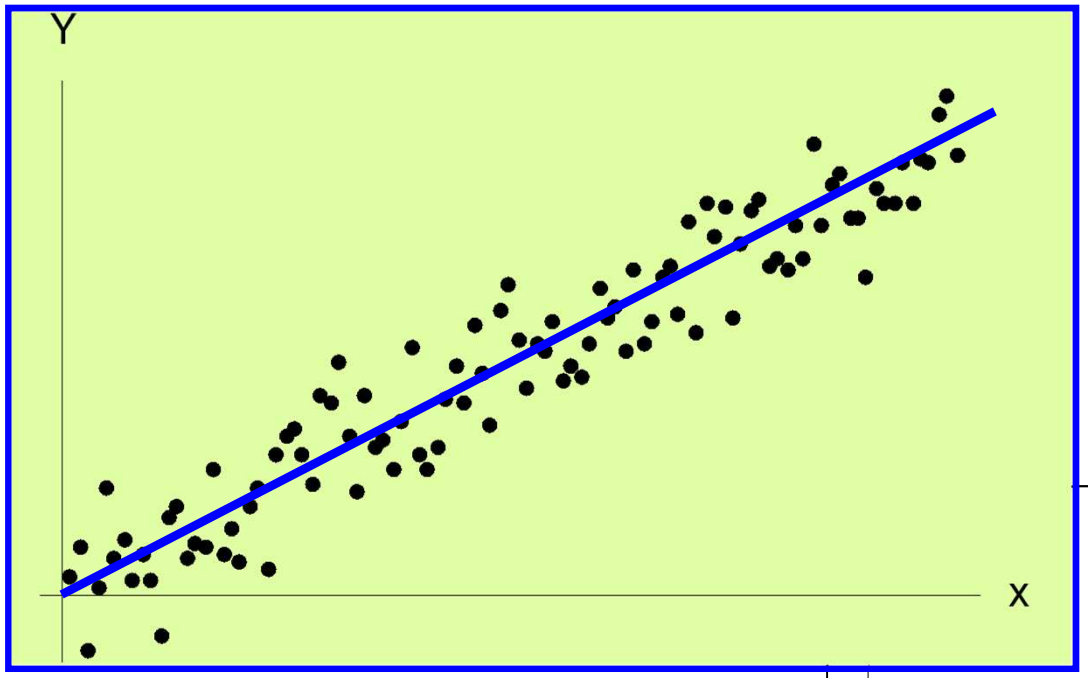
B. Stima dei parametri della retta.

$$R^2 = {\rho_{xy}}^2$$

C. Valutazione della bontà di adattamento del modello ai dati

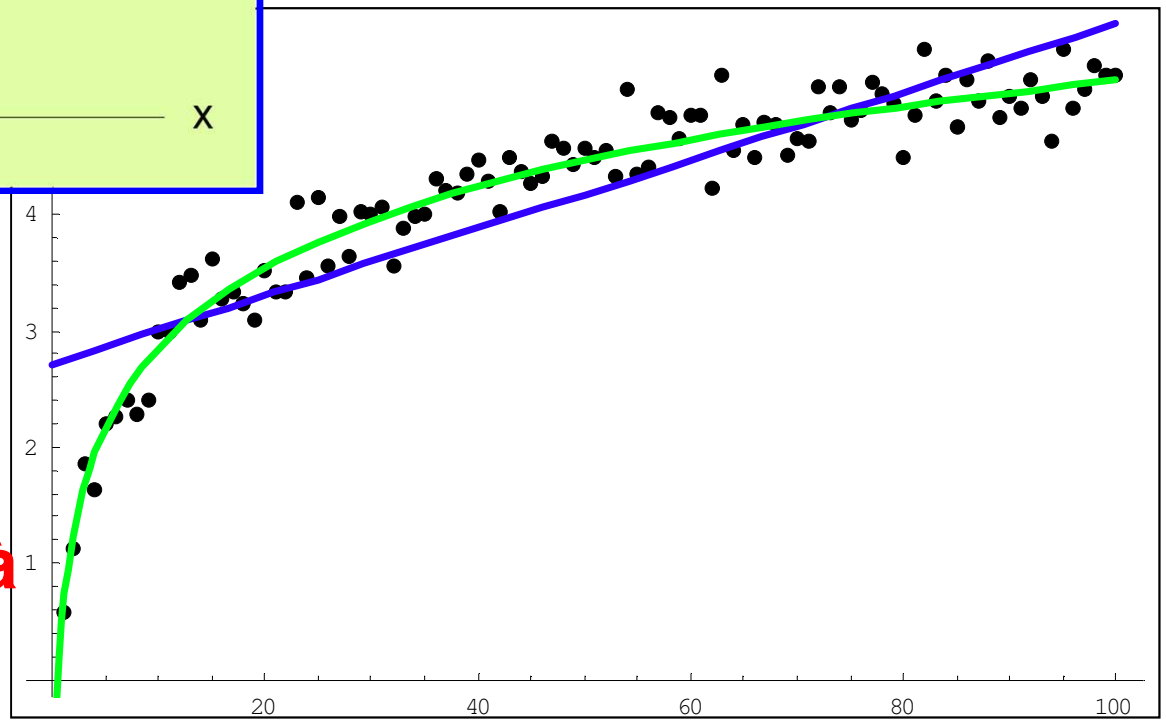
NON BASTA !

Per fare un buon modello lineare serve:

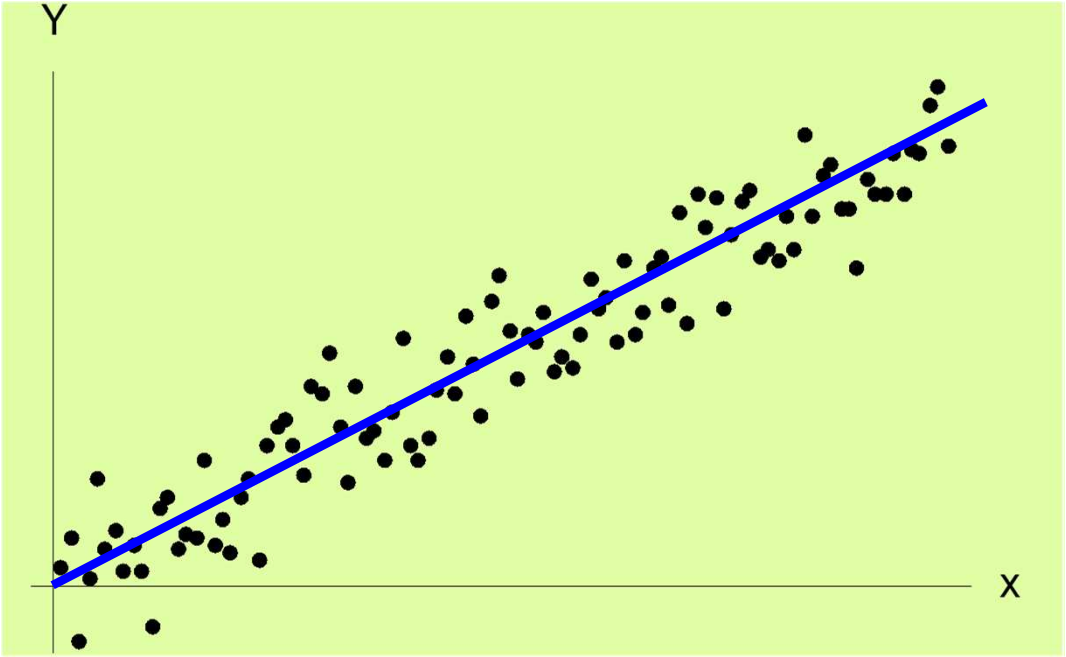


✓ una **correlazione alta**

✓ alcune ipotesi che garantiscono la **linearità del modello**



Inferenza



Il modello della
regressione lineare semplice:

$$f(x_i) = a + bx_i$$
$$\approx \Leftrightarrow \ \varepsilon_i \sim N(0, \sigma^2)$$
$$\varepsilon_i \text{ indipendenti}$$

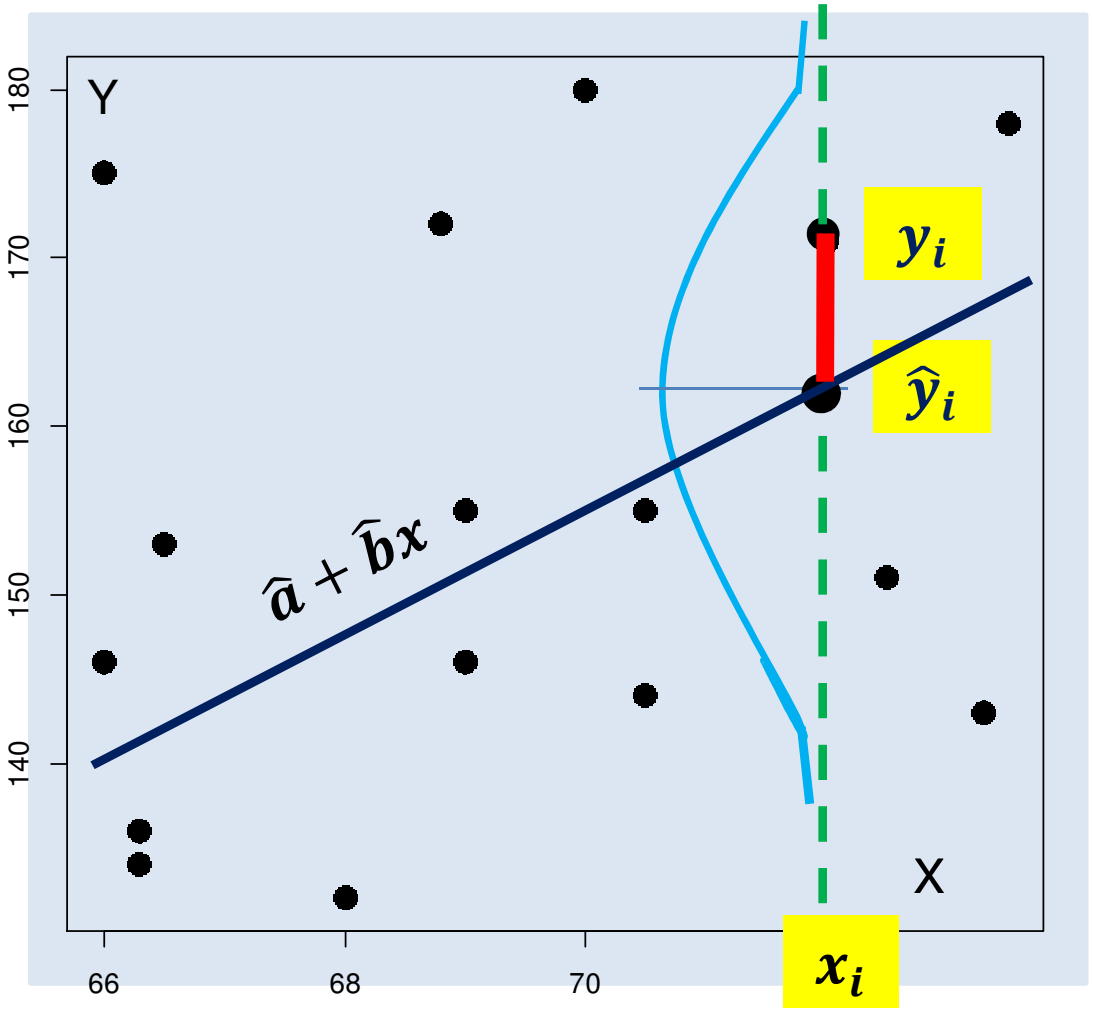
$$Y_i = a + bx_i + \varepsilon_i$$

$$Y_i \sim N(a + bx_i, \sigma^2)$$

Il **modello** di regressione lineare

$$Y_i = a + bx_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \Rightarrow Y_i \sim N(a + bx_i, \sigma^2)$$

In questo modello, **mi aspetto** di osservare il valore $\hat{y}_i = \hat{a} + \hat{b}x_i$ **(sulla retta),** ma **l'incertezza** del fenomeno può produrre **un'osservazione y_i che non sta sulla retta**. Questo *errore*, $e_i = y_i - \hat{y}_i$,è supposto **gaussiano**, quindi non può essere troppo grande ("$-3\sigma, 3\sigma$"), e deve essere simmetrico, nel senso che l'istogramma degli e_i deve dare una «campana» simmetrica.

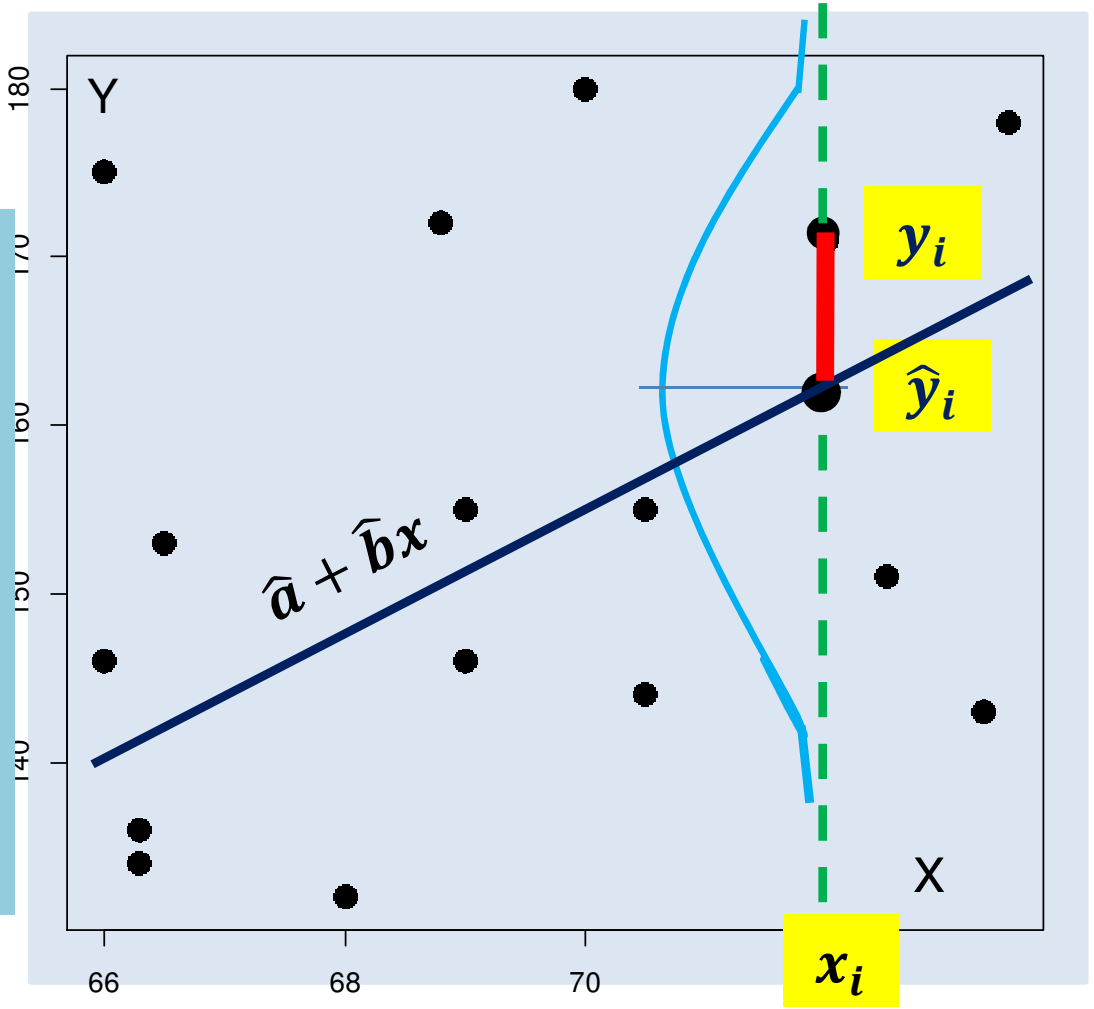


Il **modello** di regressione lineare

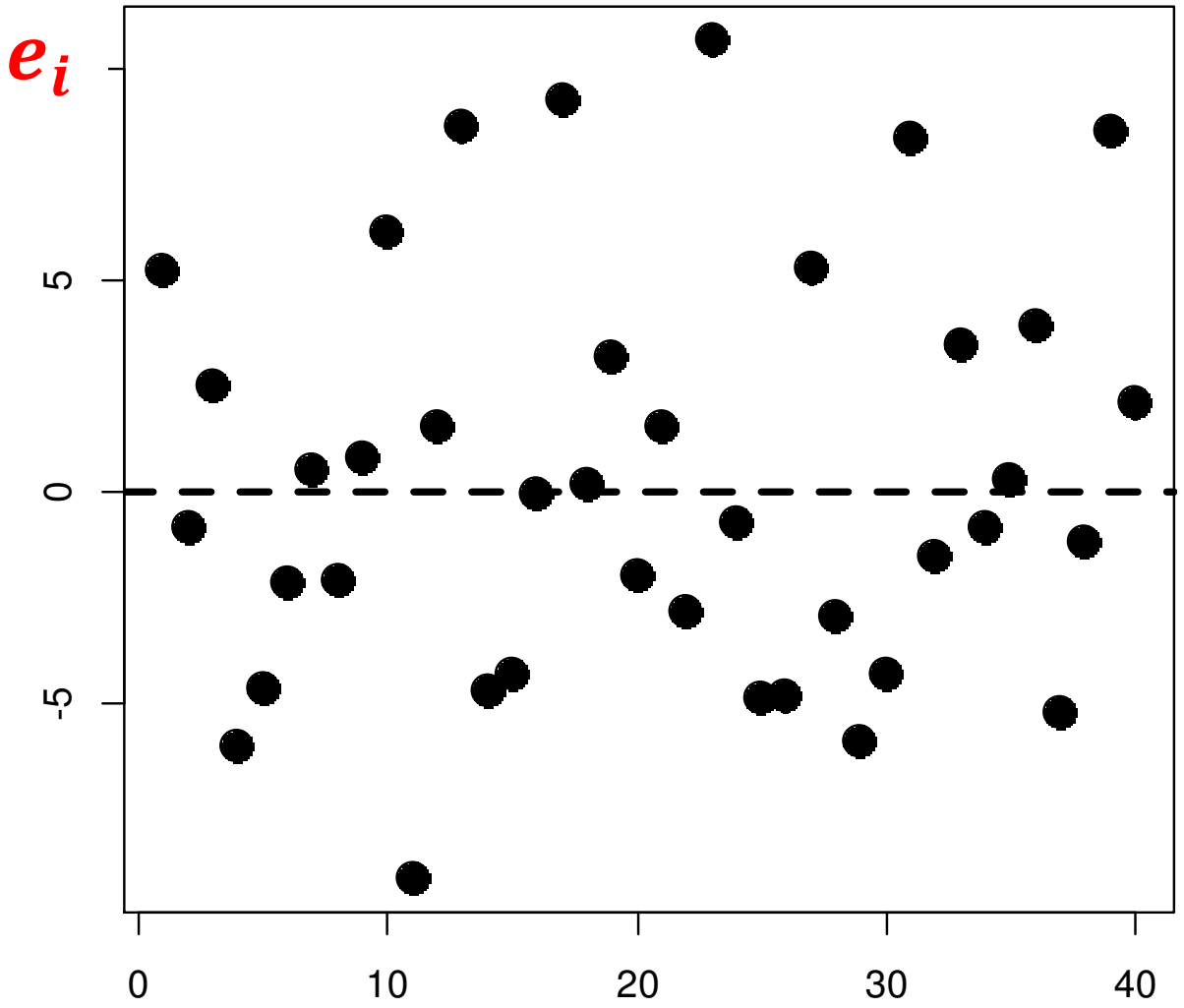
$$Y_i = a + bx_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \Rightarrow Y_i \sim N(a + bx_i, \sigma^2)$$

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

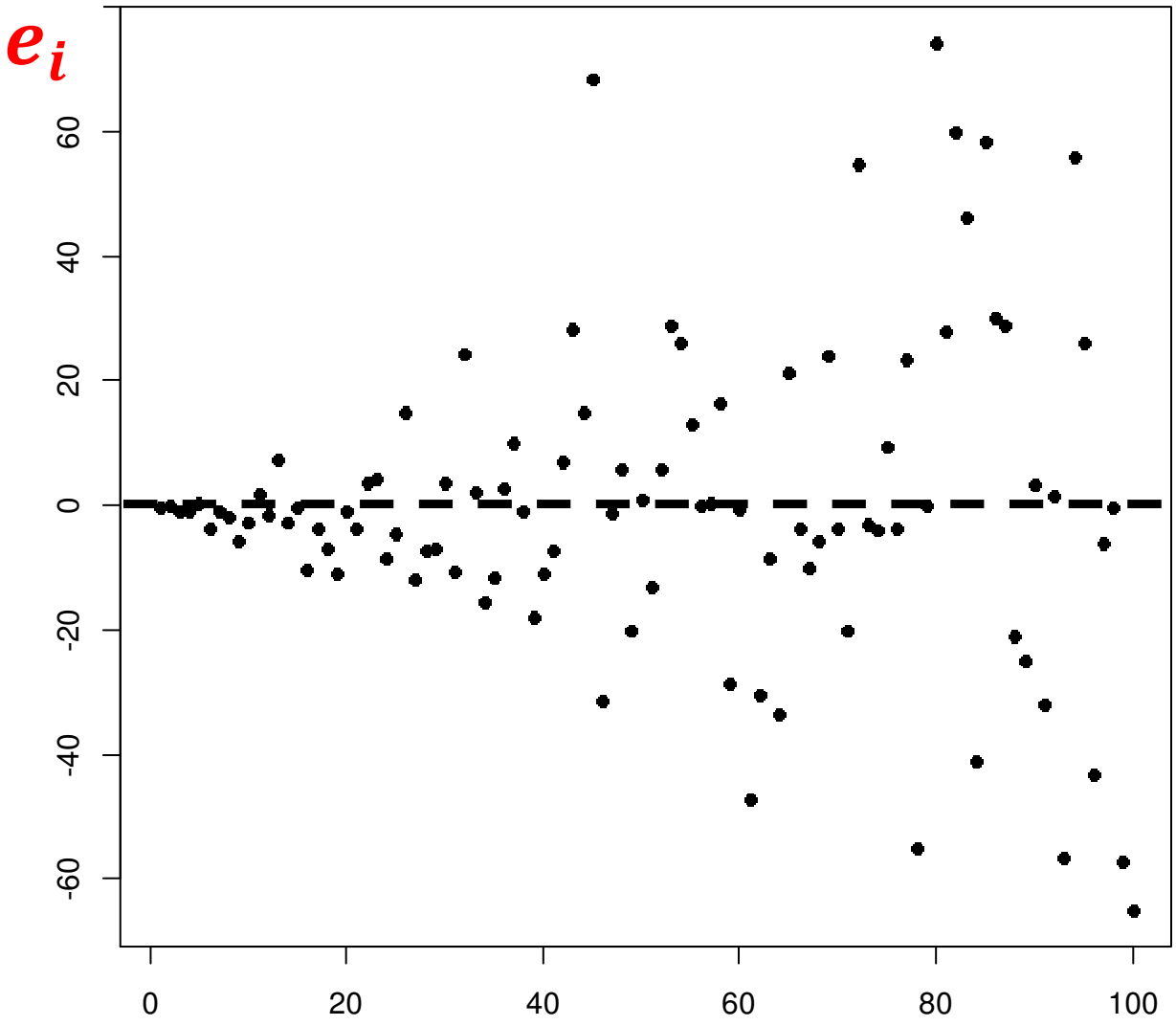
- non sono «troppo grandi»: $(-3\sigma, +3\sigma)$;
- sono in parte positivi e in parte negativi;
- il loro grafico è "sparpagliato".



Verifica della Gaussianità



Verifica della Gaussianità



La varianza non è costante

$$\varepsilon_i \sim N(0, \sigma^2)$$

Esempio, Cont.

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5

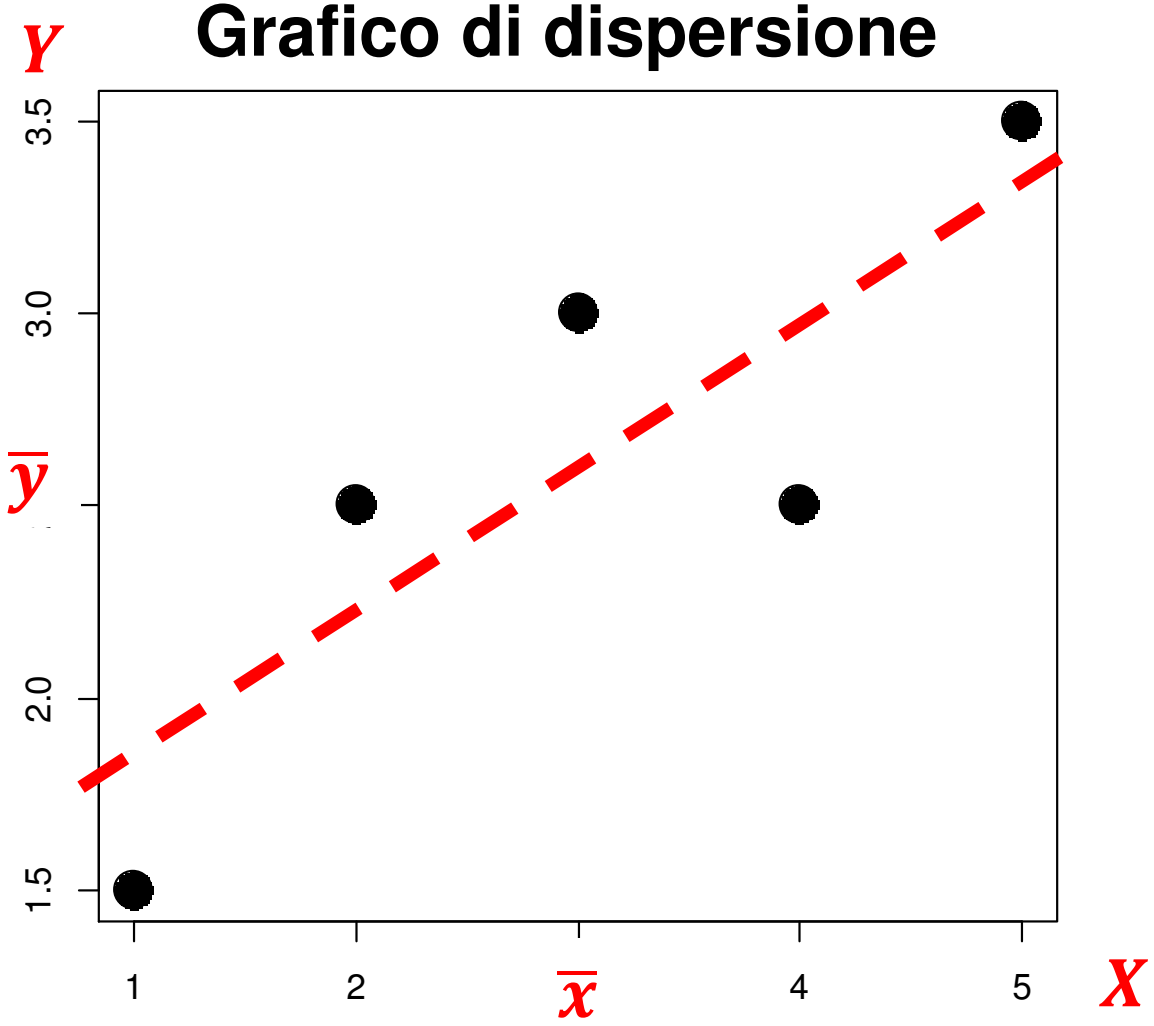
$\bar{x} = 3, \bar{y} = 2.6$

$\hat{b} = 0.4$

$\hat{a} = 1.4$

$$\boldsymbol{y = 1.4 + 0.4x}$$

$$\boldsymbol{\rho_{xy} = 0.85}$$



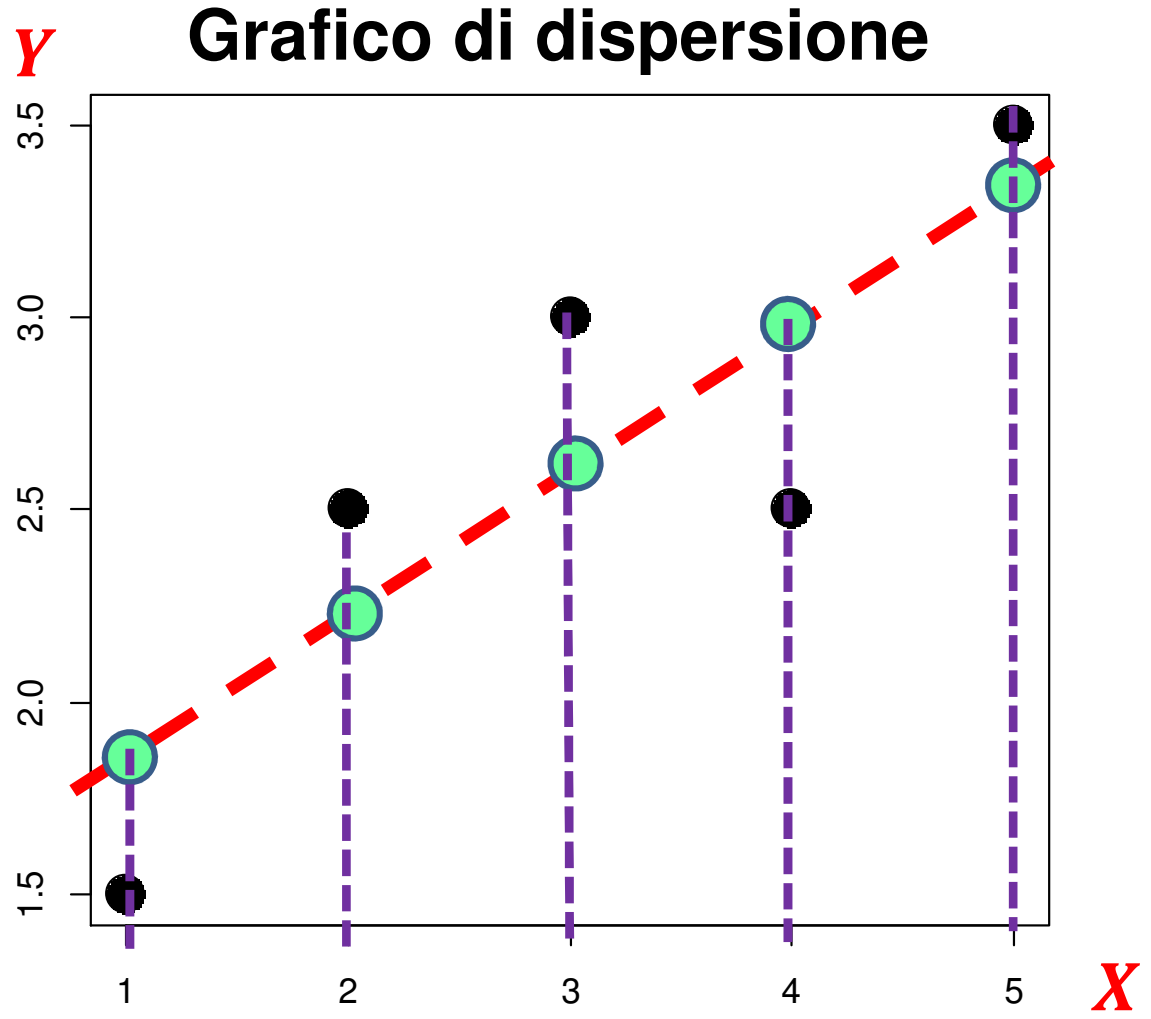
Esempio, Cont.

x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5
$\hat{y}_i$	1.8	2.2	2.6	3.0	3.4

$$\hat{y}_i = 1.4 + 0.4x_i$$

$$\rho_{xy} = 0.85$$

$$R^2 = \rho^2 = 0.72$$



Esempio, Cont.

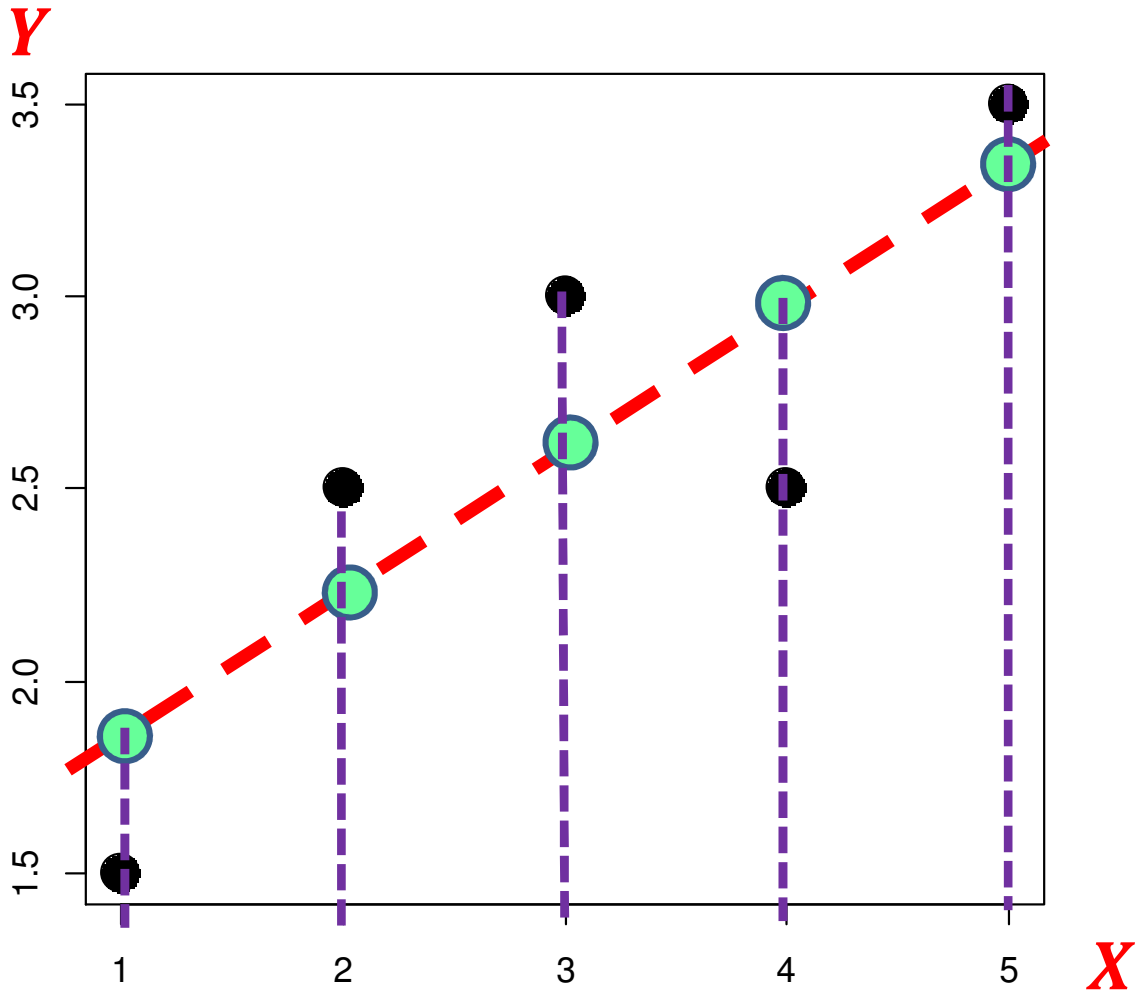
x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5
$\hat{y}_i$	1.8	2.2	2.6	3.0	3.4
e_i	-0.3	0.3	0.4	-0.5	0.1

$$\hat{y}_i = 1.4 + 0.4x_i$$

$$\rho_{xy} = 0.85$$

$$R^2 = \rho^2 = 0.72$$

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



Esempio, Cont.

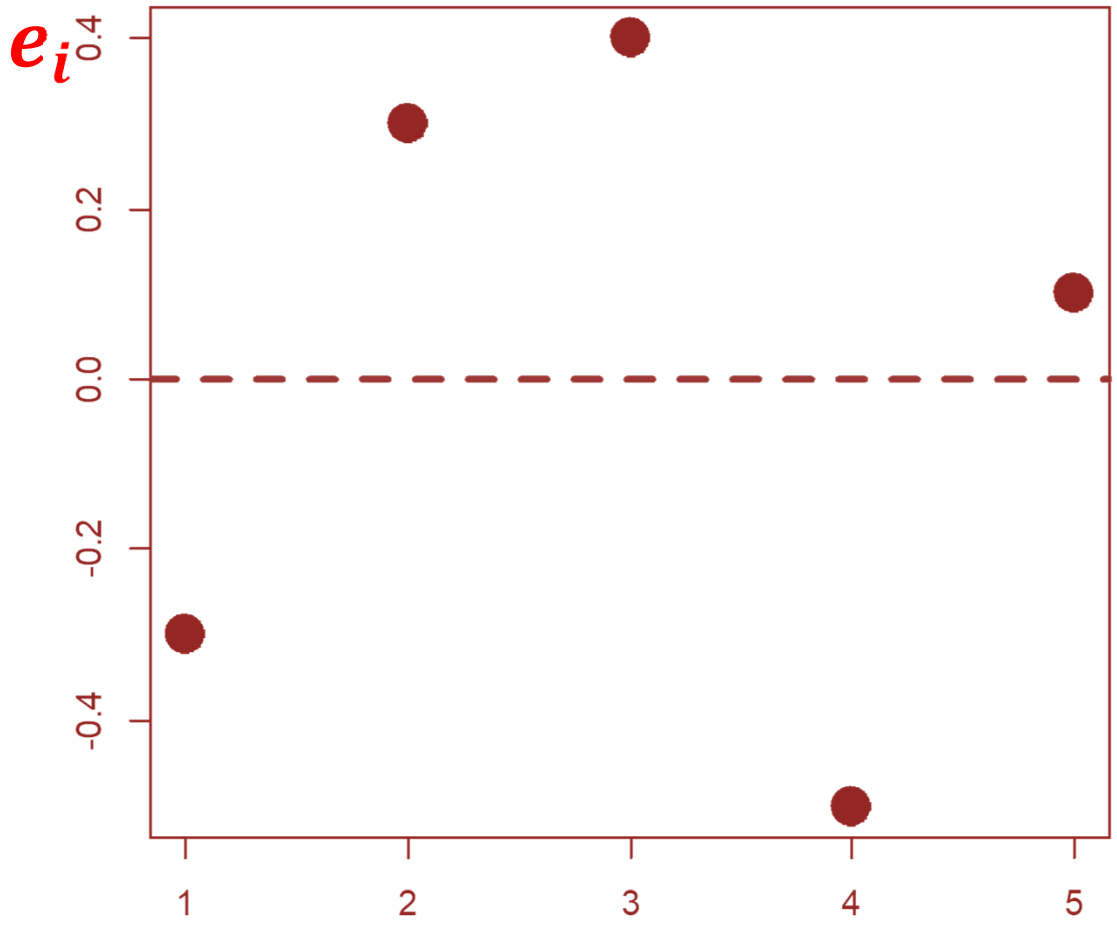
x_i	1	2	3	4	5
y_i	1.5	2.5	3	2.5	3.5
$\hat{y}_i$	1.8	2.2	2.6	3.0	3.4
e_i	-0.3	0.3	0.4	-0.5	0.1

$$\hat{y}_i = 1.4 + 0.4x_i$$

$$\rho_{xy} = 0.85$$

$$\boldsymbol{R^2 = \rho^2 = 0.72}$$

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



Esempio, Cont

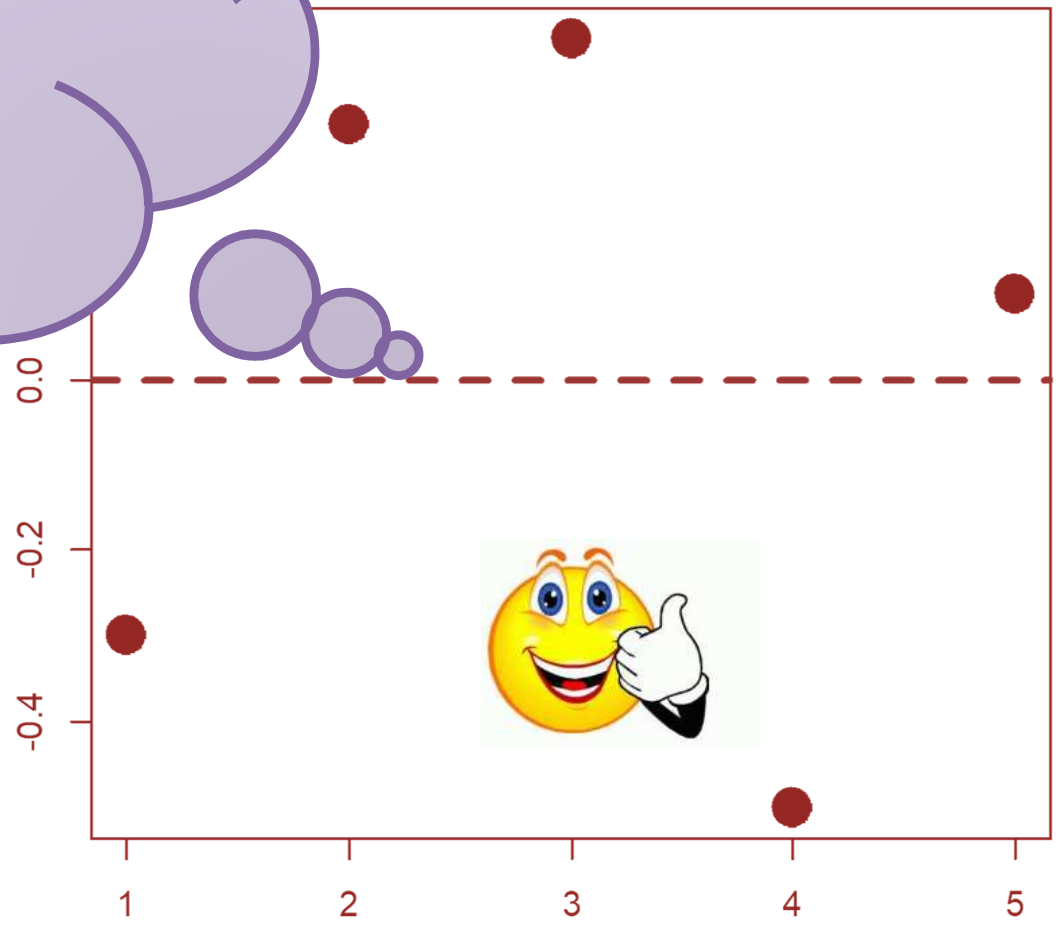
x_i	1	2	3	4	5
y	1.5	2.5	3	2.5	3.5
		2.2	2.6	3.0	3.4
			0.4	-0.5	0.1

Sparpagliati attorno alla retta tratteggiata; abbastanza simmetrici

ρ_{xy}

$$R^2 = \rho^2 = 0.72$$

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



Regressione lineare:

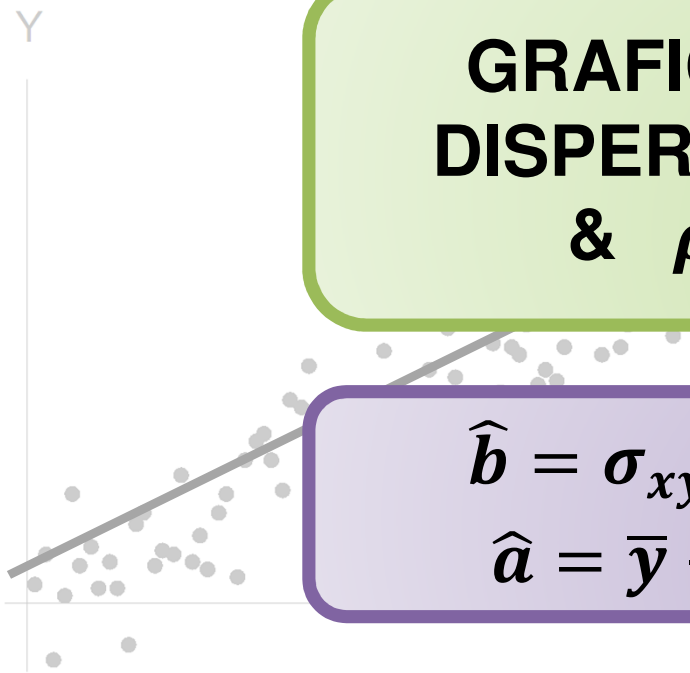


GRAFICO DI DISPERSIONE
& ρ_{xy}

A. Valutazione preliminare se una retta possa essere una buona approssimazione

$$\hat{b} = \sigma_{xy}/{\sigma_x}^2$$
$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

B. Stima dei parametri della retta.

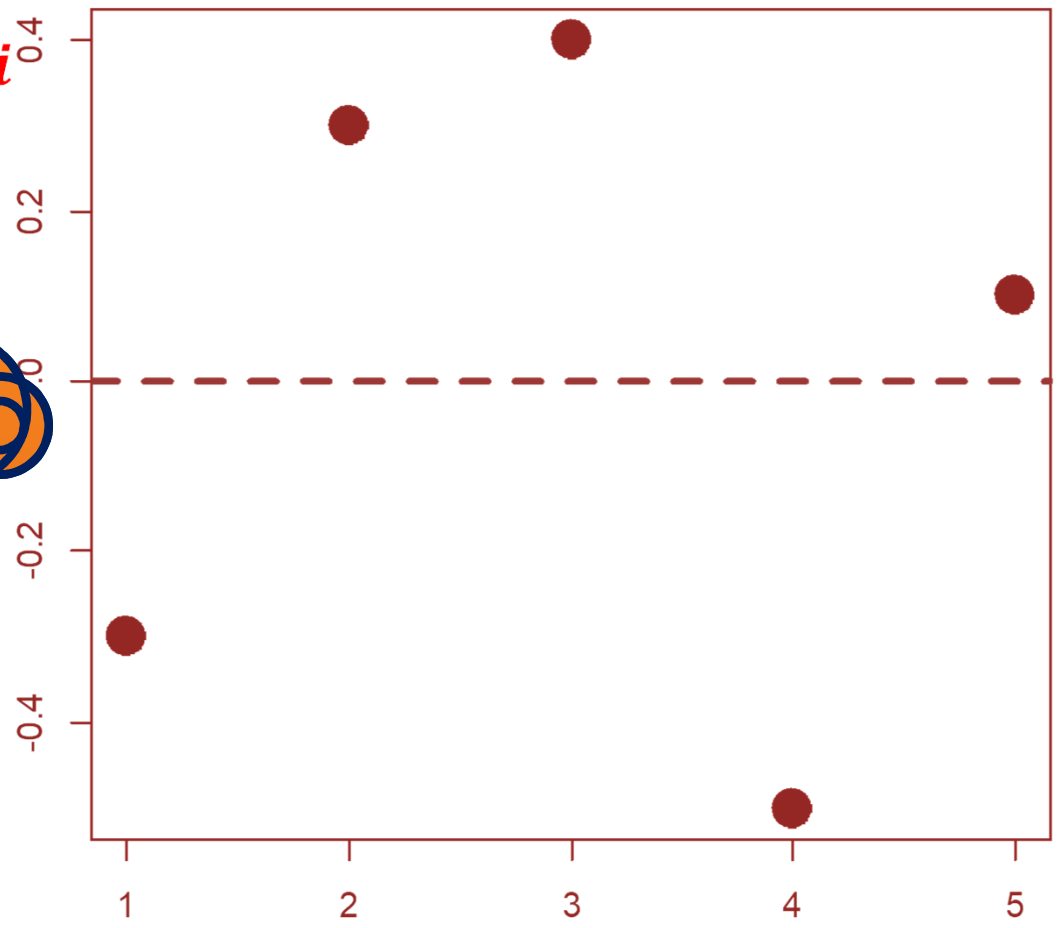
$$R^2 = {\rho_{xy}}^2$$
& analisi residui

C. Valutazione della bontà di adattamento del **modello** ai dati

Esempio C

x_i	1	2	3	4	5
	1.5	2.5	3	2.5	3.5
	1.8	2.2	2.6	3.0	3.4
	-0.3	0.3	0.4	-0.5	0.1

I DATI SONO TROPPO POCHI PER FARE UN'ANALISI SERIA DEI RESIDUI !



Regressione lineare:

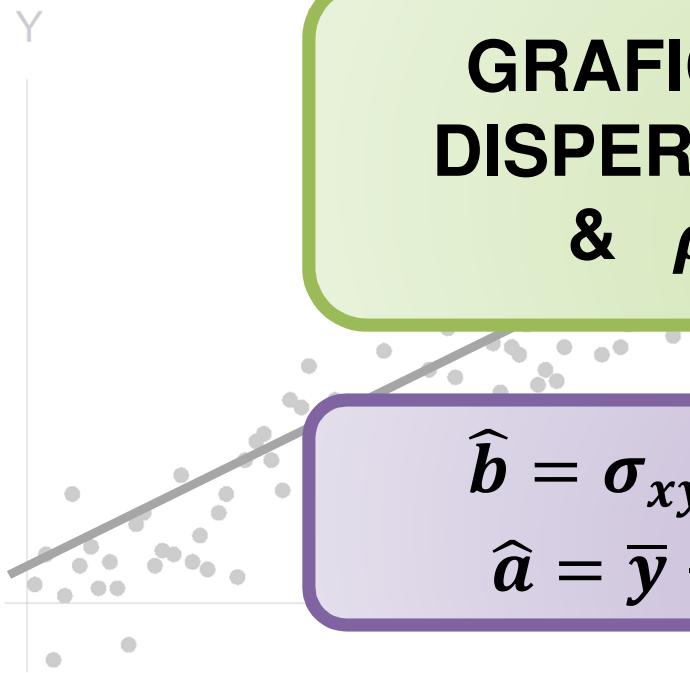


GRAFICO DI DISPERSIONE
& ρ_{xy}

$$\hat{b} = \sigma_{xy}/{\sigma_x}^2$$
$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$${R^2} = {\rho_{xy}}^2$$
& analisi residui

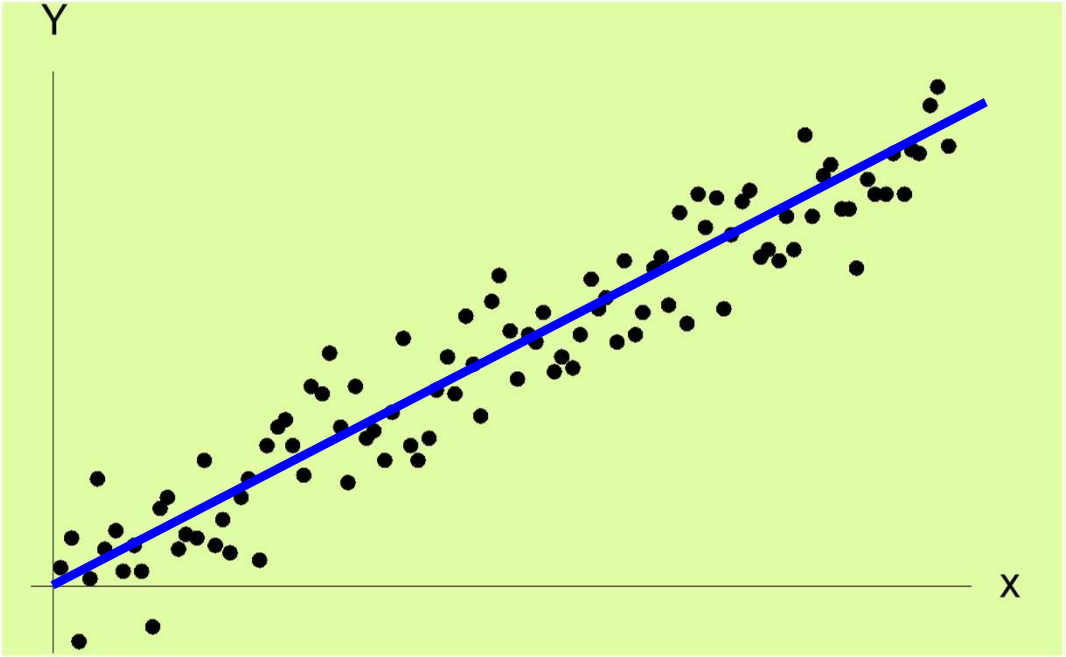
A. Valutazione preliminare se una retta possa essere una buona approssimazione

B. Stima dei parametri della retta.

C. Valutazione della bontà di adattamento del **modello** ai dati

D. Significatività della regressione

Inferenza



Il modello della
regressione lineare semplice:

$$f(x) = a + bx$$
$$\approx \Leftrightarrow \ \varepsilon_i \sim N(0, \sigma^2)$$
$$\varepsilon_i \text{ indipendenti}$$

$$Y_i = a + bx_i + \varepsilon_i$$

Verificare se il vero valore della pendenza nella popolazione di riferimento è davvero diverso da zero ($\Leftrightarrow$ previsioni!) oppure no:

$$H_0 : b = 0, \quad H_1 : b \neq 0$$