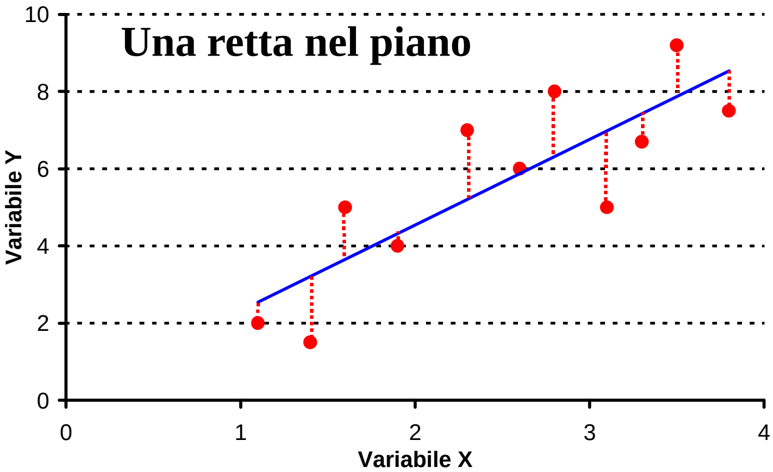
La regressione lineare semplice

Il modello di regressione lineare semplice - 1

$$y = \beta_0 + \beta x + \varepsilon$$



Il modello di regressione lineare semplice - 2

Parte sistematica del modello

Variabile di risposta (dipendente, response variable)

$$y = \beta_0 + \beta x + \varepsilon$$

Termine di errore

intercetta

coefficiente di regressione lineare

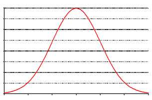
parametri ignoti del modello, stimati sulla base dei dati disponibili

Variabile esplicativa (predittiva, indipendente, explanatory)

Il modello di regressione lineare semplice - 3

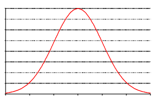
$$y = \beta_0 + \beta x + \varepsilon$$

Variabile di risposta (dipendente)



Predittore lineare, parte deterministica del modello, senza variabilità casuale

Termine di errore, parte probabilistica



L'errore, e quindi la variabile di risposta,

si distribuisce NORMALMENTE

Il modello di regressione lineare semplice - 4

Il peso (Y) dipende dalla statura (X_1)

$$E(y) = \beta_0 + \beta_1 x_1$$

E(y) = valore atteso (media) del peso degli individui che hanno quella determinata statura

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

y = peso di un determinato individuo, che dipende dalla statura, (parte sistematica del modello), ma anche da altre caratteristiche individuali (ε, parte probabilistica)

Il modello di regressione lineare semplice - 5

- **Modello teorico (ignoto)**

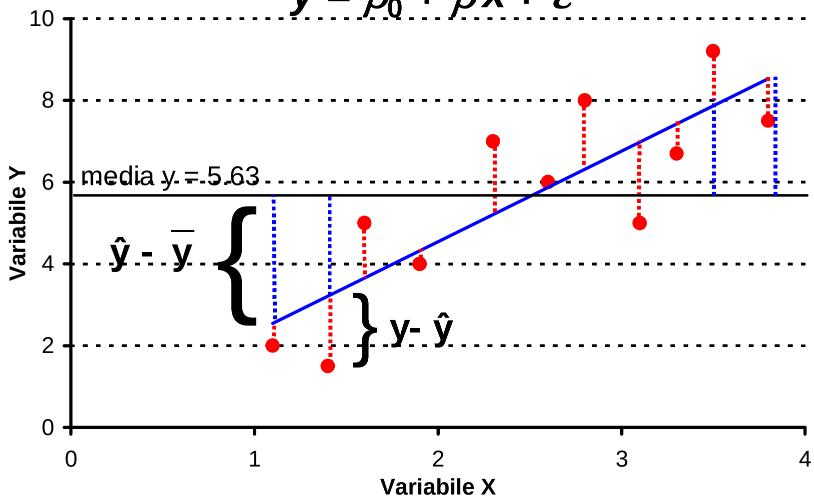
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- **Regressione Lineare stimata**

$$\hat{y} = b_0 + b_1 x$$

SCOMPOSIZIONE DELLA DEVIANZA nella Regressione lineare semplice - 1

$$y = \beta_0 + \beta x + \varepsilon$$



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

SCOMPOSIZIONE DELLA DEVIANZA nella Regressione lineare semplice - 2

Variabilità totale — **Variabilità residua**

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

Variabilità spiegata dalla regressione

Si può dimostrare che:

Devianza totale, SST — **Devianza residua, SSE**

$$\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2$$

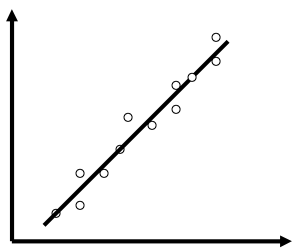
Devianza spiegata dalla regressione, SSR

Regressione lineare semplice

Si cerca di trovare la retta che meglio interpola, che meglio si adatta alla nuvola di punti.

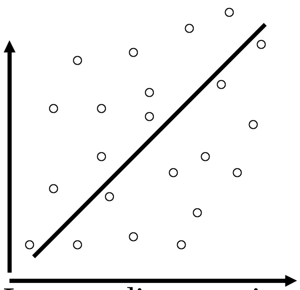
METODO DEI MINIMI QUADRATI

Si sceglie la retta che riduce al minimo la devianza residua, SSE, $\boldsymbol{\Sigma(y-\hat{y})^2}$



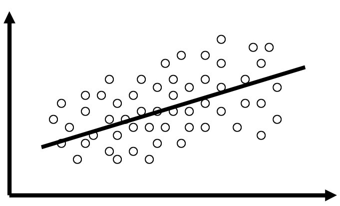
La retta di regressione approssima bene i dati.

devianza spiegata >dev.residua

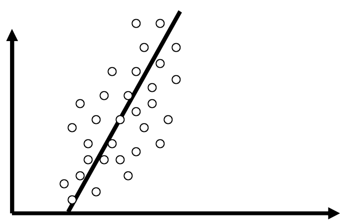


La retta di regressione approssima male i dati.

devianza spiegata <dev.residua



b prossimo a zero



b elevato

COEFFICIENTE
DI DETERMINAZIONE

$$r^2 = \Sigma (\hat{y} - \bar{y})^2 / \Sigma(y - \bar{y})^2$$

$$r^2 = 1\text{-}SSE/SST=SSR/SST$$

SST = Devianza (o Somma dei quadrati) totale
SSR = Devianza (o Somma dei quadrati) spiegata dalla Regressione

Proporzione di variazione totale della variabile dipendente *Y* **che è spiegata dalla variabile indipendente** *X*

Regressione = relazione di tipo asimmetrico: una variabile casuale (Y) dipende da una variabile fissa (X)

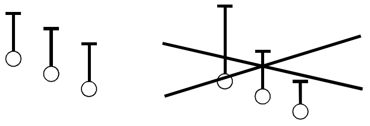
Correlazione = relazione di tipo simmetrico: le due variabili sono sullo stesso piano

Regressione lineare semplice

$y = \beta_0 + \beta x + \varepsilon$

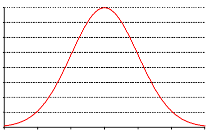
ASSUNZIONI

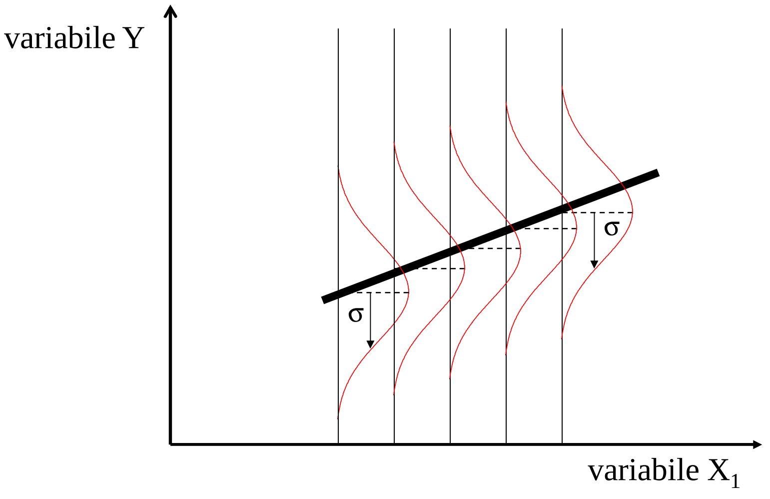
1) Il valore atteso degli errori E(ε) deve essere pari a ZERO
2) OMOSCEDASTICITA' (La varianza degli errori rimane costante)



3) INDIPENDENZA degli errori
   se le provette tra un esame e l'altro non vengono lavate adeguatamente, una determinazione risente della determinazione precedente

4) Distribuzione NORMALE degli errori





Inferenza sui parametri: i dati “supportano” il modello proposto?

I punti hanno distanze dalla retta di regressione che sono sensibilmente minori di quelle dalla media.

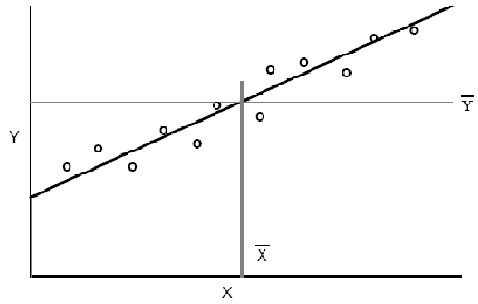


Figura A

Il valore stimato mediante la retta di regressione si avvicina molto all’osservazione reale.

Raccogliendo altri punti campionari, la retta calcolata resterebbe probabilmente immutata. La retta di regressione esprime **la relazione reale** che esiste tra X ed Y.

Inferenza sui parametri: i dati “supportano” il modello proposto?

La retta calcolata non rappresenta un miglioramento effettivo della distribuzione dei punti, rispetto alla loro media.

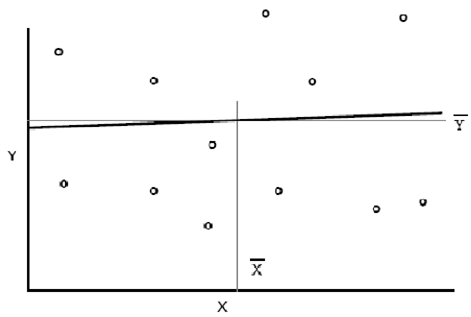


Figura B

La retta calcolata non è rappresentativa di una relazione reale tra X ed Y.

Inferenza sui parametri: i dati "supportano" il modello proposto?

Test t di Student, basato su b_1 (stima di β_1)

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases} \qquad b_1 \sim N\left(\beta_1 , \frac{\sigma^2_\varepsilon}{\Sigma(x-\bar{x})^2}\right)$$

Non conosco la quantità σ^2_ε , ne ottengo una stima:

$$\hat{\sigma}^2_\varepsilon = \frac{\Sigma(y-\hat{y})^2}{n-2}$$

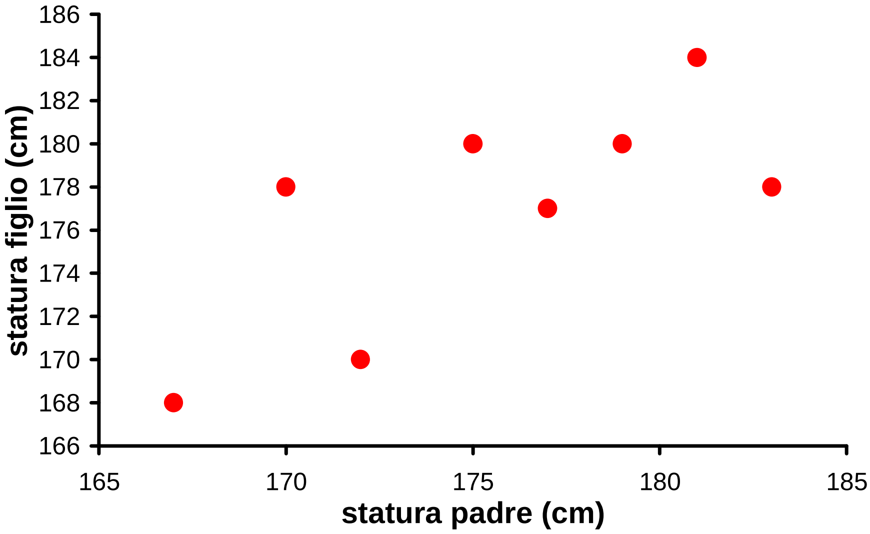
$$t = \frac{b_1-0}{ES_b} = \frac{b_1}{\sqrt{\frac{\hat{\sigma}^2_\varepsilon}{\Sigma(x-\bar{x})^2}}} = \frac{b_1}{\sqrt{\frac{\Sigma(y-\hat{y})^2}{(n-2)\,\Sigma(x-\bar{x})^2}}} \sim t_{\text{n-2 g.d.l.}}$$

ESEMPIO:

Esiste una relazione tra altezza dei padri e altezza dei figli maschi?
(A padri più bassi corrispondono figli più bassi?
A padri più alti corrispondono figli più alti?)

Padre	Figlio
167 cm	168 cm
175 cm	180 cm
183 cm	178 cm
170 cm	178 cm
181 cm	184 cm
172 cm	170 cm
177 cm	177 cm
179 cm	180 cm

I passo: rappresentazione grafica mediante diagramma di dispersione (scatterplot)



II passo: si ipotizza un modello statistico, che possa essere utile ad interpretare i dati

Ipotizziamo un modello lineare del tipo: $y = \beta_0 + \beta x + \varepsilon$

(***altezza figli***) = $\beta_0 + \beta$ ***(altezza padri)*** + ε

(i figli di uno stesso padre hanno statura abbastanza simile, ma non necessariamente uguale, anche se ομοιομετρός, cioè figli della stessa madre)

IV passo: Stima dei parametri del modello con il metodo dei minimi quadrati

$$b_1 = \frac{\textbf{codevianza}_{xy}}{\textbf{devianza}_x} = 150{,}5 / 216 = 0{,}697 \text{ cm/cm}$$

$$b_0 = \bar{y} - b_1 \bar{x} = 176{,}875 - 0{,}697 * 175{,}5 = 54{,}59 \text{ cm}$$

Retta di regressione:

altezza figlio (cm)= 54,6 cm + 0,697 cm/cm *altezza padre (cm)

Quando la statura del padre cresce di 1 cm, la statura del figlio cresce in media di 7 mm.

VI passo: Inferenza sui parametri: i dati "supportano" il modello proposto?

Test t di Student, basato su b_1 (stima di β_1)

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

test a due code

Livello di significatività = 5%

Gradi di libertà = n - 2 = 8 - 2 = 6

Soglia critica = $t_{6,\,0{,}025}$ = 2,447

$$t = \frac{b-0}{ES_b} = \frac{b}{\sqrt{var_{res} / dev_x}} = \frac{0{,}697}{\sqrt{15{,}67 / 216}} = 2{,}588$$

$dev_{res} = dev_y - codev_{xy}^2/dev_x = 198{,}9 - 150{,}5^2/ 216 = 94{,}01$

$var_{res} = dev_{res} / (n-2) = 94{,}01 / 6 = 15{,}67$

VII passo: Previsione sui valori della variabile Y

Per x = 185 cm, qual è il valore atteso di Y?

Retta di regressione:

$\hat{y}$ = 54,6 cm + 0,697cm/cm *185 cm = 183,49 cm

$$ES_{\hat{y}} = \sqrt{var_{res}\,[1/n + (x-\bar{x})^2 / dev_x\,]} =$$

$$= \sqrt{15{,}67\,[1/8 + (185-175{,}5)^2 / 216\,]} = 2{,}916$$

$$IC_{95\%} = \hat{y} \pm t_{\nu,\alpha/2} * ES_{\hat{y}} = 183{,}49 \pm 2{,}447 * 2{,}916 = \left[\begin{array}{l} 190{,}63 \\ 176{,}36 \end{array}\right.$$