Modelli lineari

REGRESSIONE LINEARE

- Molti problemi dell'ingegneria e della scienza vogliono **studiare le relazioni tra due o più insiemi di variabili**
- *Ad esempio, in un processo chimico è interessante studiare la dipendenza tra quantità di catalizzatore impiegato, temperatura e rendimento → obiettivo: predire rendimento per diversi valori di temperatura e quantità di catalizzatore*

Variabile di risposta Y e variabili $x_1, \ldots, x_n$ di ingresso

risposta dipendente

esplicative (indipendenti)

Il modello suppone che la risposta sia funzione degli ingressi

$$Y = h(x_1, x_2, \ldots, x_n)$$

➡ Y variabile dipendente, x_i variabili indipendenti

La relazione *più semplice* che è possibile immaginare è quella *lineare*

$$\boxed{Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon}$$

p = n° di variabili esplicative

$\beta_0, \beta_1, .. \beta_p$ costanti

$\boxed{\epsilon \text{ errore casuale}}$

o residuo

partiamo da p = 1

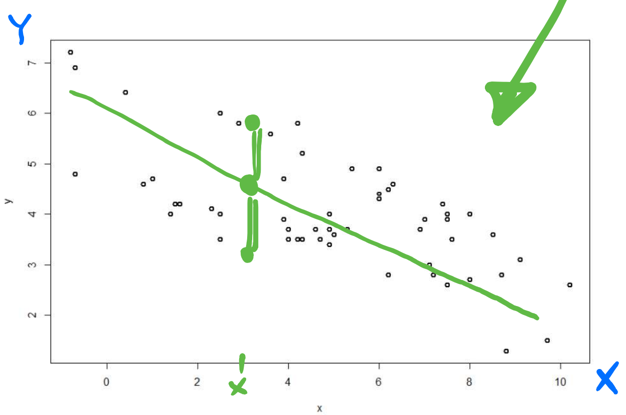
$$\boxed{Y = \beta_0 + \beta_1 x_1 + \varepsilon}$$

(Y,X)

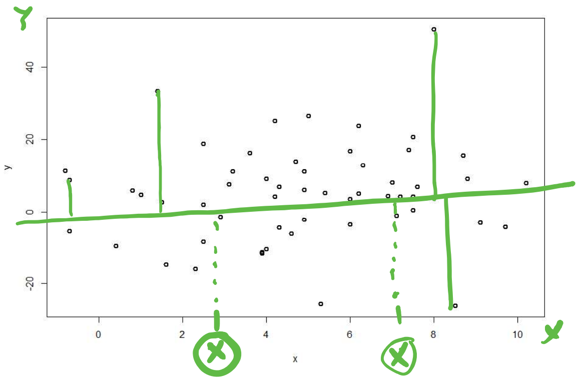
Partiamo dal caso più semplice (due variabili) e riprendiamo il concetto di correlazione lineare dalla statistica descrittiva

$$\frac{C_{XY}}{\sqrt{\sigma_X^2\,\sigma_Y^2}} = r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{Cov[X,Y]}{\sqrt{V[X]\cdot V[Y]}}$$

r=-0.68



r=0.1



Per quale dei due esempi provereste ad impostare un modello lineare? Altre osservazioni?

Caso p = 1: regressione lineare semplice

Modello: $Y = \beta_0 + \beta_1 x + \varepsilon$ ← errore casuale

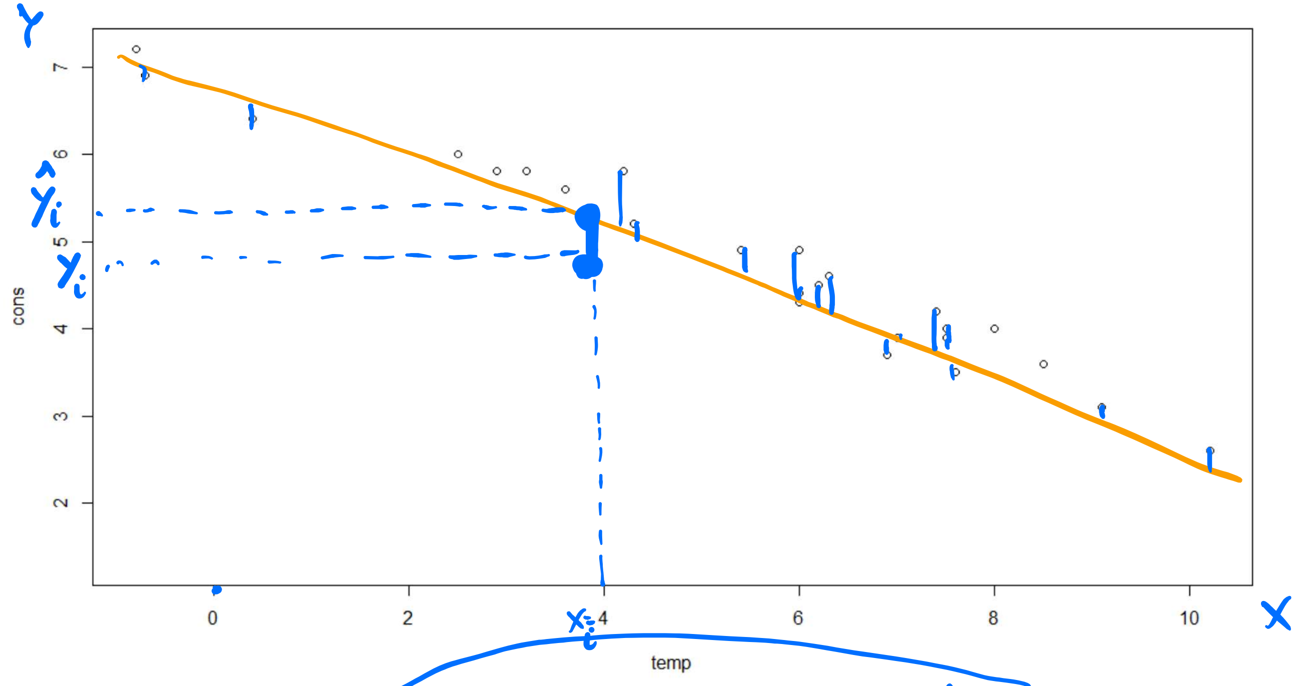
ε é una v.c.

$E[\varepsilon]=0$

Cosa prevede: $\mathbb{E}[Y] = \beta_0 + \beta_1 x$

Dati n punti sperimentali (x_i, y_i), si vuole determinare una stima $\hat{y} = b_0 + b_1 x$ della retta di regressione $Y = \beta_0 + \beta_1 x + \varepsilon$, <u>minimizzando la somma dei quadrati degli errori casuali $\varepsilon_i = Y_i - \hat{Y}_i$ tra i valori della variabile casuale Y e i valori previsti con la retta in corrispondenza del valore della variabile indipendente x_i.</u>

Esempio: relazione tra temperatura esterna (in gradi Celsius) X e consumo di gas Y (in ft^3) per il riscaldamento di un appartamento.

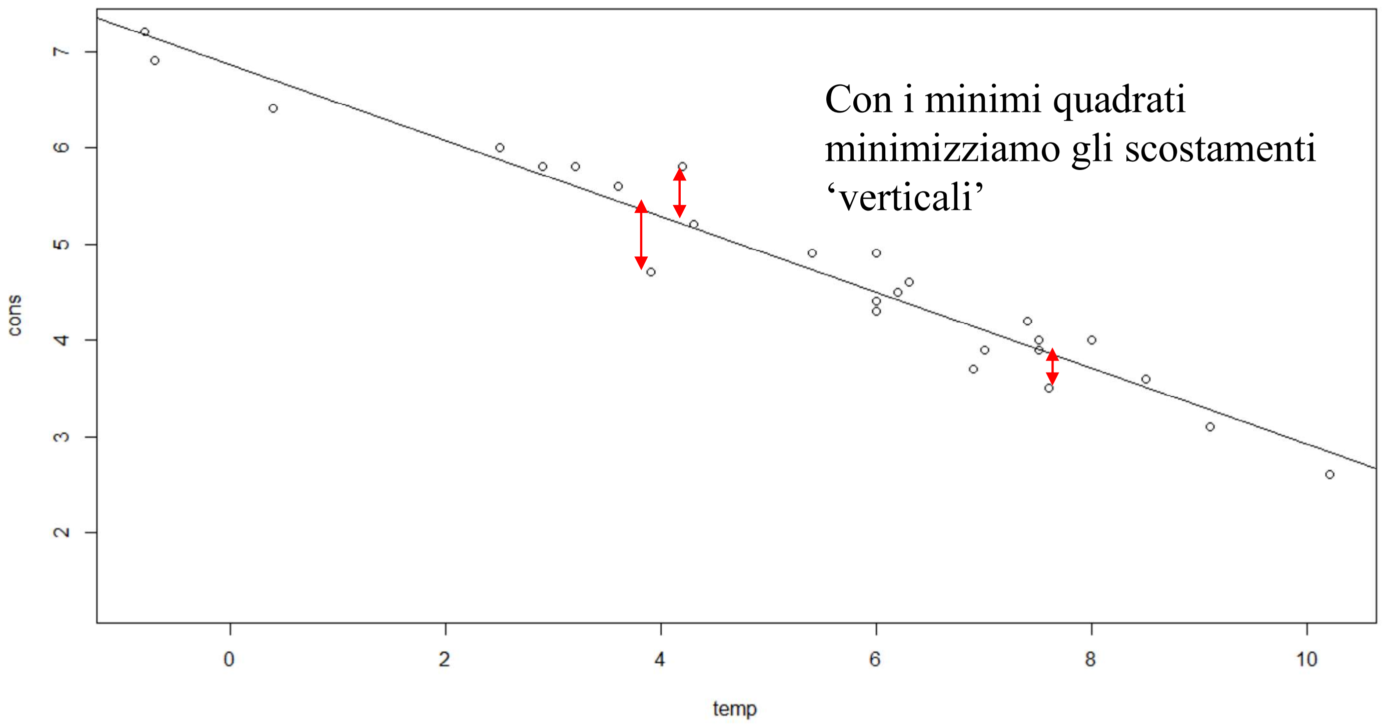


$$\min \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

r= -0.97

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Retta di regressione lineare → Metodo dei minimi quadrati



Stime dei parametri di regressione e proprietà

$$SS(\beta_0,\beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

SS Somma dei quadrati degli scarti tra risposte stimate e reali

Il metodo dei minimi quadrati consiste nello scegliere come stimatori di β_0 e β_1 i due valori che minimizzano SS

$$\begin{cases} \dfrac{\partial SS}{\partial \beta_0} = \cancel{-2} \displaystyle\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{\partial SS}{\partial \beta_1} = \cancel{-2} \displaystyle\sum_{i=1}^{n} x_i (Y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\mathrm{B}_0 + \mathrm{B}_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i Y_i = \mathrm{B}_0 \sum_{i=1}^{n} x_i + \mathrm{B}_1 \sum_{i=1}^{n} x_i^2 \end{cases}$$

Stimatori ai minimi quadrati

$$\begin{cases} B_0 = \overline{Y} - B_1 \overline{x} \\ B_1 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \end{cases}$$

con:

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$\bar{x} = \frac{1}{n} \sum x_i$$

$$\begin{cases} b_0 = \overline{y} - b_1 \overline{x} \\ b_1 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \end{cases}$$

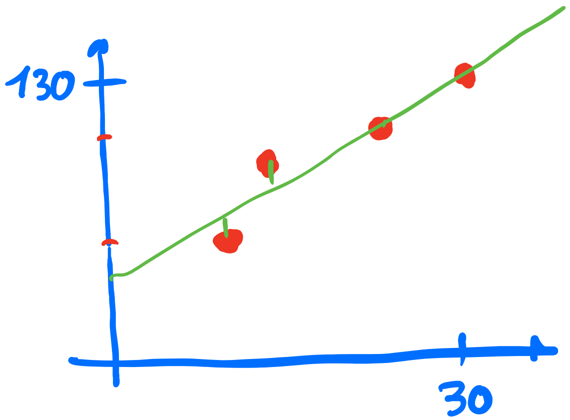
con $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

o in forma equivalente ma più adatta al calcolo:

$$\begin{cases} b_0 = \overline{y} - b_1 \overline{x} \\ b_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - n \overline{x}\, \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \overline{x}^2} \end{cases}$$

	X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$\hat{y}_i$
	100	27	10	2	20	100	4	
	60	18	-30	-7	210	900	21	
	130	30	40	5	200	1600	25	
	70	25	-20	0	0	40	0	
medie	90	25	0	0	107.5	8850	19.5	

$y = b_0 + b_1 X$



$$\sigma_{xy} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n} = 107.5 \qquad \frac{\sum (x_i-\bar{x})^2}{n} = \sigma_x^2 = 8850 \qquad \sigma_Y^2 = 19.5$$

$$r = \frac{C_{xy}}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}} = \frac{107.5}{\sqrt{195 \cdot 8850}} \cong 0.89$$

$$b_1 = \frac{C_{xy}}{\sigma_x^2} = \frac{107.5}{8850} \cong 0.14$$

$$b_0 = \bar{y} - b_1 \bar{x} = 25 - 0.14 \cdot 90 \cong 12$$

$$\boxed{y = 0.14 \cdot x + 12}$$

$$x_0 = 80 \longrightarrow y_0 = 0.14 \cdot 80 + 12 \cong 11.2 + 12 \cong 23$$

$$y_0 = 23 + \varepsilon \qquad \varepsilon = ?$$

Distribuzione degli stimatori

$B_0, B_1,$ $\varepsilon = ?$

B_0 B_1

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ per $x = x_i$

ε_i (errore indiv. i-simo)

- $\mathbb{E}[\varepsilon_i] = 0$ e $var[\varepsilon_i] = \sigma^2$
- $cov[\varepsilon_i, \varepsilon_j] = 0$ per ogni $i \neq j$

⇓

- $\mathbb{E}[Y_i] = \mathbb{E}[\beta_0 + \beta_1 x_i + \varepsilon_i] = \beta_0 + \beta_1 x_i$ per ogni $i = 1, 2, \ldots, n$
- $var[Y_i] = var[\beta_0 + \beta_1 x_i + \varepsilon_i] = \sigma^2$ per ogni $i = 1, 2, \ldots, n$
- le Y_i sono non correlate, cioè $cov[Y_i, Y_j] = 0$ per ogni $i \neq j$
- la variabile x è nota senza errore ed è osservata per *almeno* due valori distinti

+ *Normalità* $\longleftrightarrow$ $\varepsilon_i \sim N(0, \sigma^2)$

$B_0, B_1 \sim N$

$\mathbb{E}[B_0] = \beta_0$

$$Var(B_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

$\mathbb{E}[B_1] = \beta_1$

$$var[B_1] = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$cov[B_0, B_1] = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \sum (\varepsilon_i - 0)^2$$

Sia $SS_R = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_i)^2$

Si può dimostrare che $\boxed{\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}}$

$$E\left[\frac{SS_R}{\sigma^2}\right] = n - 2$$

$$\frac{(1-r^2)\sum(y_i-\bar{y})^2}{n-2} = \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

stimatore non distorto di σ^2

Questa è la stima della varianza dell'errore. La sua radice è chiamata ERRORE STANDARD (o deviazione standard dell'errore)

Alcune definizioni utili per dopo

$$S_{xY} = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y - n\bar{x}\bar{Y}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

Inferenza statistica sui parametri di regressione

$$\frac{B_0 - E[\beta_0]}{\sqrt{\mathrm{var}[B_0]}} = \frac{B_0 - \beta_0}{\sigma / \sqrt{\frac{nS_{xx}}{\sum_{i=1}^{n} x_i^2}}} \sim N(0,1)$$

t-student

$\hat{\sigma}$

Intervallo di fiducia

Test di ipotesi

$$\frac{B_1 - E[\beta_1]}{\sqrt{\mathrm{var}[B_1]}} = \frac{B_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0,1)$$

Intervallo di fiducia

Test di ipotesi