Statistica: principi e metodi

Capitolo 10

Analisi delle distribuzioni doppie: regressione

Analisi delle relazioni fra i caratteri di una distribuzione doppia quantitativa

Data la distribuzione congiunta di due caratteri quantitativi, i quesiti a cui la statistica deve rispondere sono

- ✓esiste una relazione fra i due caratteri?
- ✓che tipo di relazione esiste?

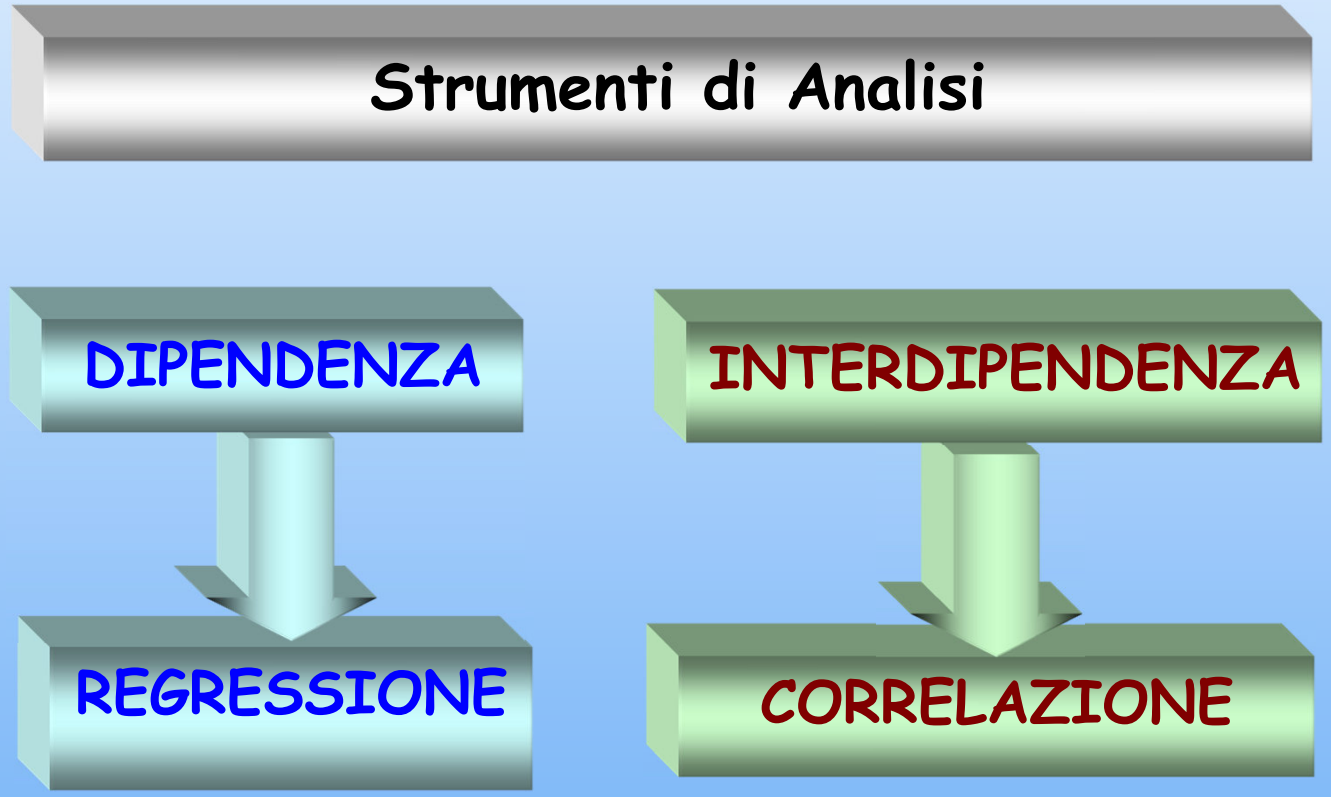
DIPENDENZA → Un carattere può essere considerato *antecedente* logico dell'altro

ESEMPIO: Età (antecedente) → Statura (conseguente)
Reddito(antecedente) → Consumo (conseguente)

INTERDIPENDENZA → Non si può stabilire quale sia il carattere *antecedente* e quale il *conseguente*

ESEMPIO: Voto in matematica ↔ Voto in statistica

Strumenti per l'analisi delle relazioni di una distribuzione doppia quantitativa



Modello di regressione

Una relazione statistica può essere descritta tramite l'equazione

$$y = f(x) + \varepsilon$$

dove la **variabile risposta**, *y*, è espressa come somma di due componenti:

- quella rappresentata dalla **funzione matematica *f(x)***, che fornisce il *contributo della variabile indipendente x al livello della variabile risposta y*
- quella "**residuale**", ε, che sintetizza il *contributo di tutti i fattori che potrebbero influire sulla variabile risposta y e che non vengono considerati.*

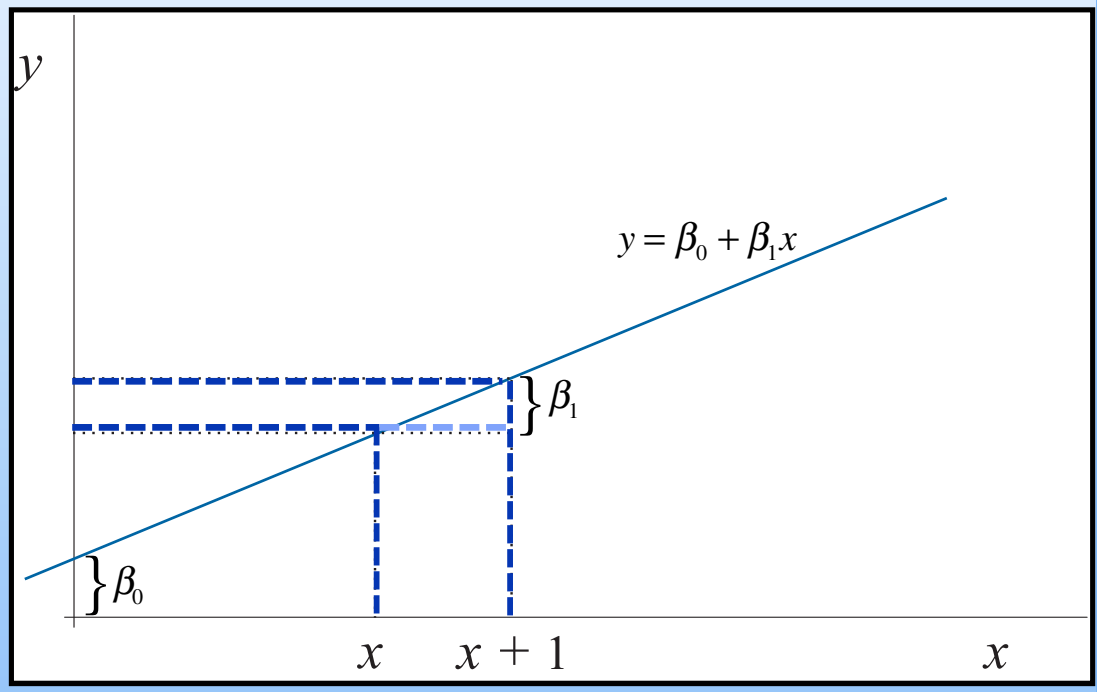
Tale relazione definisce il **modello di regressione di *Y* su *X***.

Regressione lineare

Se la funzione matematica $f(x)$, che descrive la dipendenza di Y da X, è l'equazione della **retta**

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

dove β_0 e β_1 sono i **parametri della funzione**, abbiamo la regressione lineare.



β_0 : intercetta
β_1 : coefficiente angolare

Esempio di relazione statistica

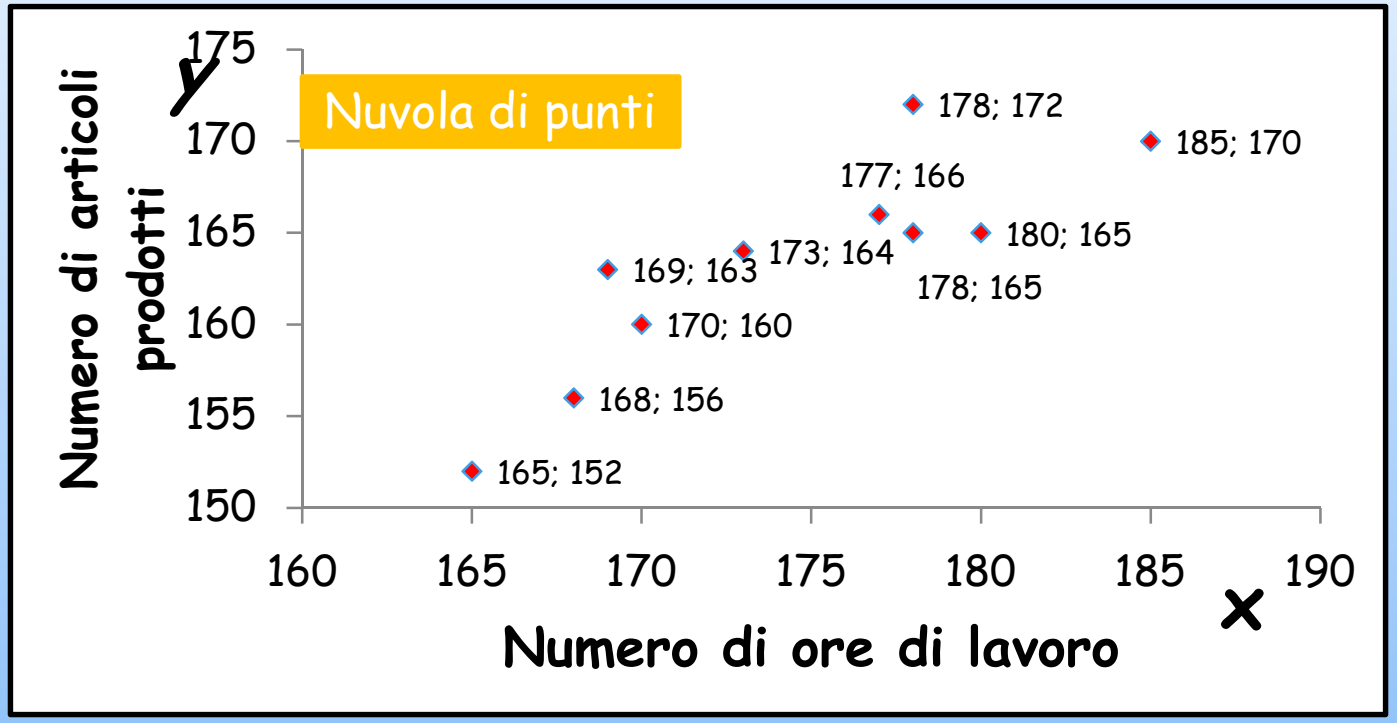
Nella tabella sono riportati il numero di ore lavorate ed il numero di articoli prodotti da 10 artigiani nel corso di un mese.

Vogliamo stabilire se la retta è una funzione adatta a esprimere il legame associativo tra il numero di articoli prodotti ed il numero di ore di lavoro.

Numero di ore di lavoro	Numero di articoli prodotti
173	164
178	172
169	163
170	160
177	166
178	165
180	165
185	170
165	152
168	156

Grafico di dispersione

Numero di ore di lavoro	Numero di articoli prodotti
173	164
178	172
169	163
170	160
177	166
178	165
180	165
185	170
165	152
168	156



L'andamento dei punti suggerisce che la relazione statistica che lega il numero di prodotti al numero di ore lavorate può essere espressa da una retta.

Regressione lineare

- $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ sono le coppie di valori osservati su N unità statistiche, dette **punti osservati** o **nuvola di punti**.
- Il problema è quello di **assegnare ai parametri** β_0 e β_1 della retta

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

  i valori che consentano di approssimare **nel miglior modo possibile** la nuvola dei punti. In altri termini, dobbiamo determinare quella **retta** - *tra le infinte del piano* -, **che meglio si adatta alla nuvola di punti**.
- La soluzione viene trovata utilizzando il **metodo dei minimi quadrati**.

Metodo dei minimi quadrati

Indicati con b_0 e b_1 due particolari valori di β_0 e β_1, siano

$$\hat{y}_i = b_0 + b_1 x_i, \; i = 1,2,\ldots,N$$

i **valori teorici** o **predizioni** di Y.

Con il **metodo dei minimi quadrati** si assegnano a b_0 e b_1 i valori che rendono minima la quantità S_q, data da

$$S_q = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2 = (y_1 - b_0 - b_1 x_1)^2 + \cdots + (y_N - b_0 - b_1 x_N)^2$$

Si tratta della **somma dei quadrati delle differenze tra i valori effettivi e i valori teorici di Y**, una misura del **grado di approssimazione dei valori osservati tramite le predizioni.**

Metodo dei minimi quadrati

Determinazione di b_0 e b_1 : valori che rendono minima la quantità

$$S_q = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2$$

Equazioni normali: derivate parziali della funzione rispetto ai parametri, poste uguali a zero

$$\begin{cases} \dfrac{\partial S_q}{\partial b_0} = -2\sum_{i=1}^{N}(y_i - b_0 - b_1 x_i) = 0 \\ \dfrac{\partial S_q}{\partial b_1} = -2\sum_{i=1}^{N}(y_i - b_0 - b_1 x_i)x_i = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} b_0 - \sum_{i=1}^{N} b_1 x_i = 0 \\ \sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} b_0 x_i - \sum_{i=1}^{N} b_1 x_i^2 = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{N} y_i - N b_0 - b_1 \sum_{i=1}^{N} x_i = 0 \\ \sum_{i=1}^{N} y_i x_i - b_0 \sum_{i=1}^{N} x_i - b_1 \sum_{i=1}^{N} x_i^2 = 0 \end{cases}$$

$$\begin{cases} b_0 = \dfrac{\sum_{i=1}^{N} y_i}{N} - b_1 \dfrac{\sum_{i=1}^{N} x_i}{N} = \mu_y - b_1 \mu_x \\ \sum_{i=1}^{N} y_i x_i - (\mu_y - b_1 \mu_x)\sum_{i=1}^{N} x_i - b_1 \sum_{i=1}^{N} x_i^2 = 0 \end{cases}$$

$$\sum_{i=1}^{N} y_i x_i - \mu_y \sum_{i=1}^{N} x_i + b_1 \mu_x \sum_{i=1}^{N} x_i - b_1 \sum_{i=1}^{N} x_i^2 = 0$$

$$\sum_{i=1}^{N} y_i x_i - N\mu_y \mu_x - b_1 \left(-N\mu_x^2 + \sum_{i=1}^{N} x_i^2 \right) = 0$$

$$b_1 = \frac{\sum_{i=1}^{N} y_i x_i - N\mu_y \mu_x}{\sum_{i=1}^{N} x_i^2 - N\mu_x^2} = \frac{C_{xy}}{D_x}$$

Stima dei parametri

Dalla soluzione del problema di minimo, si trovano le formule seguenti per b_1 e b_0:

codevianza

$$b_1 = \frac{C_{XY}}{D_X} = \frac{\sum_{i=1}^{N} x_i y_i - N\mu_X\mu_Y}{\sum_{i=1}^{N} x_i^2 - N\mu_X} = \frac{\sum_{i=1}^{N} (x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{N} (x_i - \mu_X)^2}$$

$$b_0 = \mu_Y - b_1\mu_X$$

devianza

- μ_X e μ_Y sono le medie aritmetiche delle distribuzioni marginali di X e di Y.

Parametri della retta di regressione: esempio di calcolo

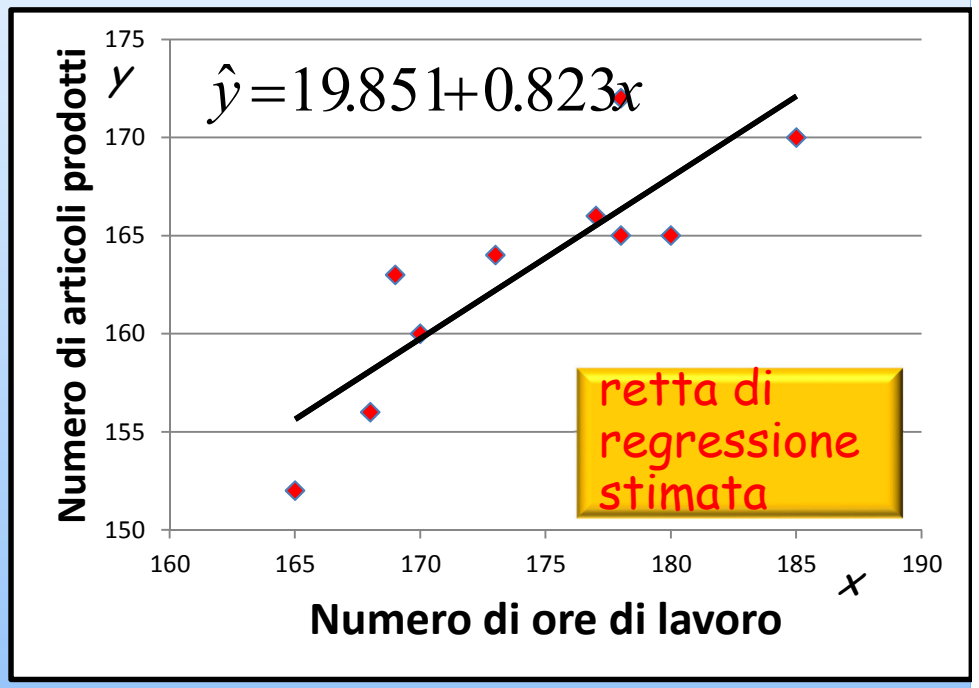
$$b_1 = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)}{\sum_{i=1}^{N}(x_i - \mu_X)^2}$$

$$b_0 = \mu_y - b_1\mu_X$$

μ_X	= 174.30
μ_y	= 163.30

$(x_i - \mu_X).(y_i - \mu_y)$

x_i	y_i	$x_i - \mu_X$	$y_i - \mu_y$	$(x_i - \mu_X)^2$	
173	164	-1.30	0.70	1.69	-0.91
178	172	3.70	8.70	13.69	32.19
169	163	-5.30	-0.30	28.09	1.59
170	160	-4.30	-3.30	18.49	14.19
177	166	2.70	2.70	7.29	7.29
178	165	3.70	1.70	13.69	6.29
180	165	5.70	1.70	32.49	9.69
185	170	10.70	6.70	114.49	71.69
165	152	-9.30	-11.30	86.49	105.09
168	156	-6.30	-7.30	39.69	45.99
			Totale	356.10	293.10



$$b_1 = \frac{C_{XY}}{D_X} = \frac{293.10}{356.10} = 0.823$$

$$b_0 = \mu_y - b_1\mu_x = 163.30 - 0.823 \cdot 174.30 = 19.851$$

Retta di regressione

Una volta calcolati b_1 e b_0 l'equazione che ne risulta

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

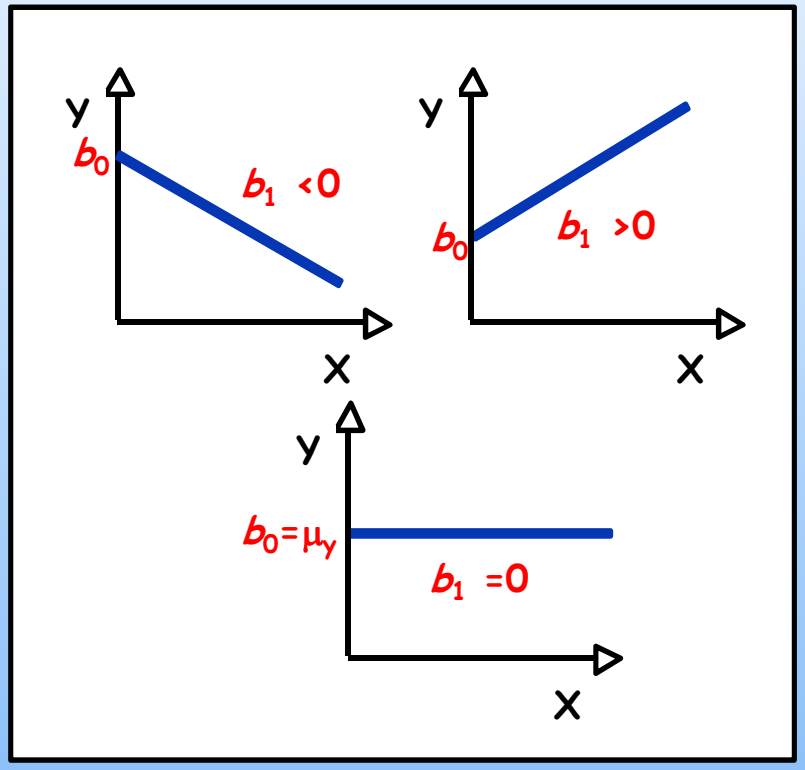
rappresenta la **retta di regressione stimata di Y su X**.

Il significato da attribuire al **coefficiente angolare** b_1 è il seguente. Poiché la retta rispecchia l'"andamento medio" dei dati osservati, b_1 indica la **variazione media** che subisce Y quando X aumenta di una unità.

La retta di regressione passa per il **baricentro** della distribuzione doppia, cioè per il punto (μ_X, μ_Y).

Interpretazione geometrica dei parametri

- b_0 → intercetta della retta: punto in cui la retta interseca l'asse verticale; valore della y per x=0:
- b_1 → coefficiente angolare della retta o coefficiente di regressione;
  - se b_1 **<0** la retta è inclinata negativamente: al crescere della variabile indipendente x la variabile dipendente y decresce;
  - se b_1 **=0** la retta è parallela all'asse delle ascisse: al crescere della variabile x la y rimane costante (indipendenza lineare);
  - se b_1 **>0** la retta è inclinata positivamente e al crescere della variabile x cresce anche la y;



Residui

Le differenze tra i valori effettivi e i valori teorici di *Y*,

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i), \quad i = 1, 2, \ldots, N$$

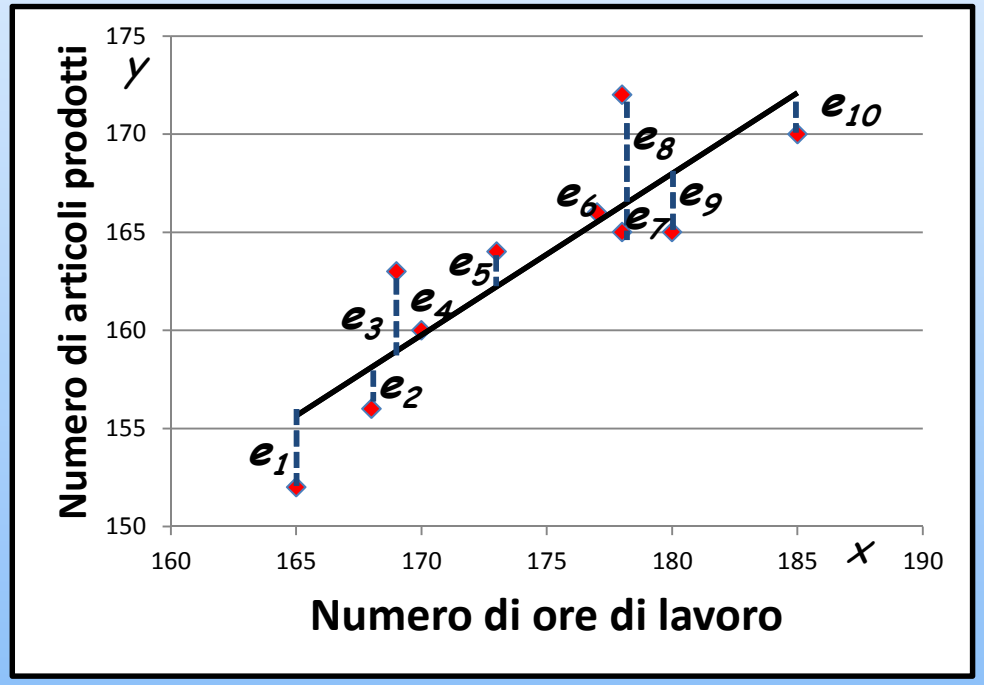
sono dette **residui**.

La loro somma è uguale a 0:

$$\sum_{i=1}^{N} e_i = 0$$

Dalla prima equazione normale del metodo dei minimi quadrati si ha

$$\frac{\partial S_q}{\partial b_0} = -2\sum_{i=1}^{N}(y_i - b_0 - b_1 x_i) = 0 \Rightarrow \sum_{i=1}^{N}(y_i - \hat{y}_i) = 0$$



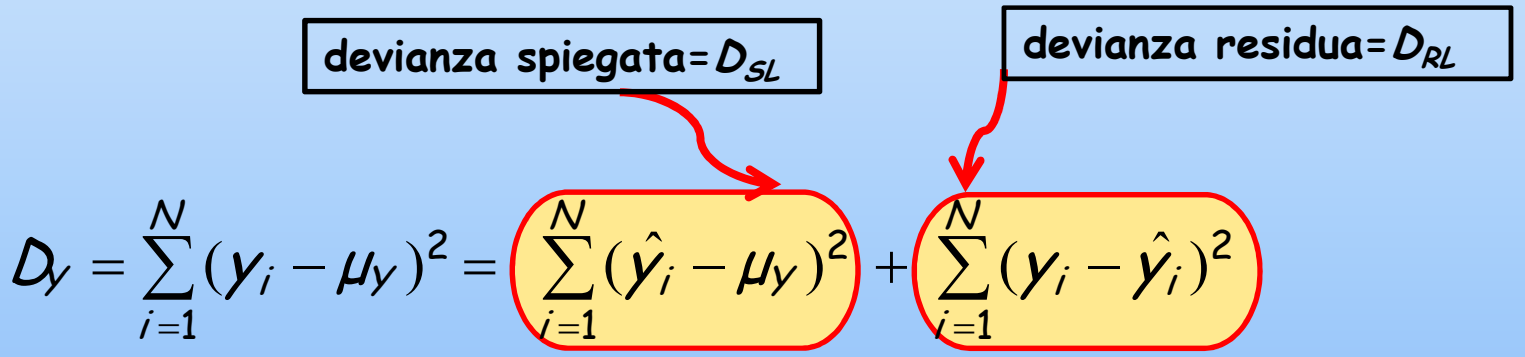
Adattamento della retta di regressione ai dati

L'analisi di regressione include la verifica dell'**idoneità del modello** a rappresentare la relazione statistica tra le variabili *Y* e *X*.

A questo fine, viene introdotto un apposito indice che misura la bontà dell'adattamento della retta di regressione ai punti osservati, per la cui costruzione ci si avvale della **scomposizione della devianza**.

Scomposizione della devianza nel modello di regressione

Data una distribuzione doppia disaggregata, la devianza della distribuzione marginale del carattere Y può essere così scomposta:



dove $\hat{y}_i$ sono le predizioni fornite dalla retta di regressione.

Scomposizione della devianza nel modello di regressione: dimostrazione

Sottraiamo e addizioniamo il valore teorico di y_i

Sviluppo del quadrato del binomio

$$D_y = \sum_{i=1}^{N}(y_i-\mu_y)^2 = \sum_{i=1}^{N}(y_i-\hat{y}_i+\hat{y}_i-\mu_y)^2 = \underbrace{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}_{D_{RL}} + \underbrace{\sum_{i=1}^{N}(\hat{y}_i-\mu_y)^2}_{D_{SL}} + 2\underbrace{\sum_{i=1}^{N}(y_i-\hat{y}_i)(\hat{y}_i-\mu_y)}_{=0}$$

Sviluppo il prodotto

$$\sum_{i=1}^{N}(y_i-\hat{y}_i)(\hat{y}_i-\mu_y) = \sum_{i=1}^{N}(y_i-\hat{y}_i)\hat{y}_i - \sum_{i=1}^{N}(y_i-\hat{y}_i)\,\mu_y = 0$$

Portando fuori dalla sommatoria la costante si ha

$$\mu_y \sum_{i=1}^{N}(y_i-\hat{y}_i) = 0$$

Sostituendo al valore teorico la sua espressione si ha

$$\sum_{i=1}^{N}(y_i-\hat{y}_i)\hat{y}_i = \sum_{i=1}^{N}e_i(b_0+b_1x_i) = \sum_{i=1}^{N}e_ib_0 + \sum_{i=1}^{N}e_ib_1x_i = b_0\underbrace{\sum_{i=1}^{N}e_i}_{=0} + b_1\underbrace{\sum_{i=1}^{N}e_ix_i}_{=0} = 0$$

Dalla seconda equazione normale dei minimi quadrati si ha

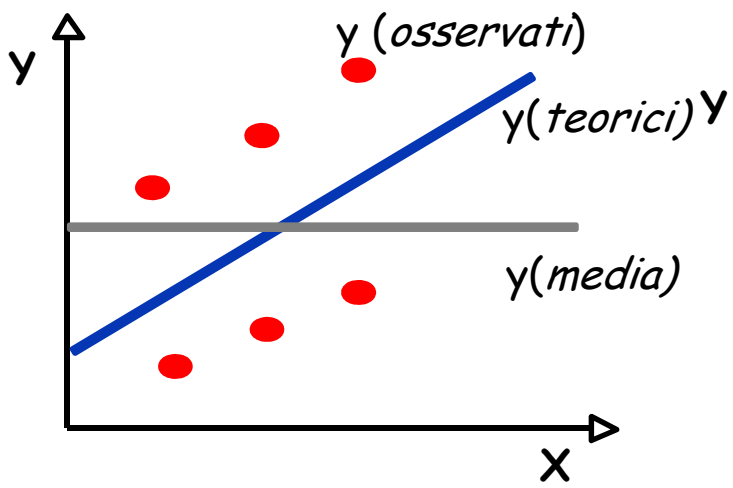
$$\frac{\partial S_q}{\partial b_1} = -2\sum_{i=1}^{N}(y_i-b_0-b_1x_i)x_i = 0$$

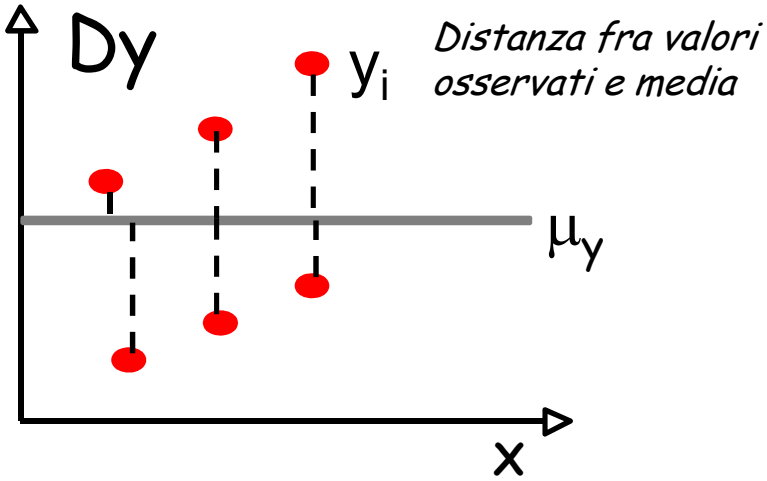
$$\Rightarrow \sum_{i=1}^{N}(y_i-\hat{y}_i)x_i = 0$$

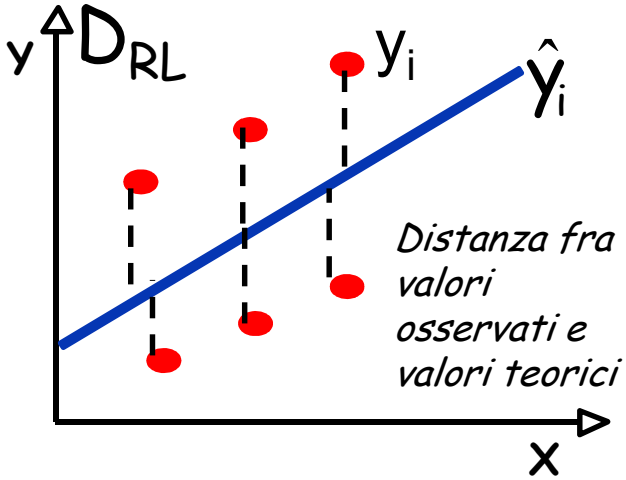
$$\sum_{i=1}^{N}(y_i-\mu_y)^2 = \sum_{i=1}^{N}(\hat{y}_i-\mu_y)^2 + \sum_{i=1}^{N}(y_i-\hat{y}_i)^2$$

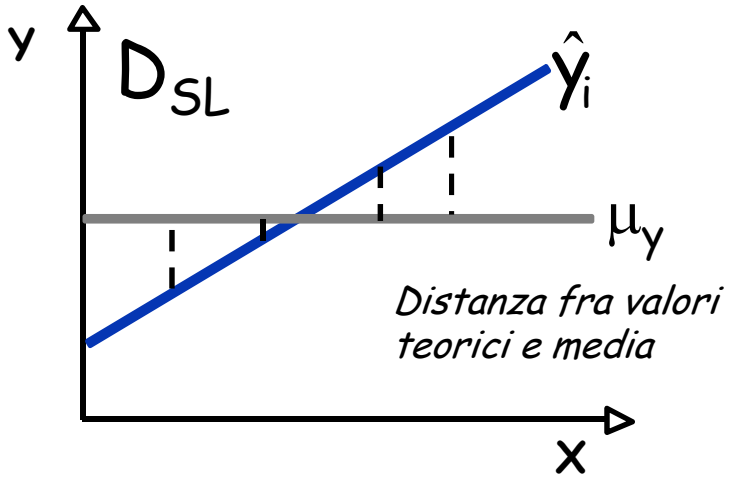
$$D_y = D_{SL} + D_{RL}$$

Scomposizione della devianza nel modello di regressione: interpretazione grafica









Scomposizione della devianza: esempio

$$\hat{y} = 19.851 + 0.823x$$

μ_Y = 163.30

Valori teorici

x_i	y_i	$\hat{y}_i$	$(y_i - \mu_Y)^2$	$(\hat{y}_i - \mu_Y)^2$	$(y_i - \hat{y}_i)^2$
173	164	162.23	0.49	1.145	3.13
178	172	166.35	75.69	9.27	31.98
169	163	158.94	0.09	19.03	16.50
170	160	159.76	10.89	12.52	0.06
177	166	165.52	7.29	4.94	0.23
178	165	166.35	2.89	9.27	1.81
180	165	167.99	2.89	22.01	8.95
185	170	172.11	44.89	77.55	4.44
165	152	155.65	127.69	58.58	13.29
168	156	158.12	53.29	26.88	4.47
		Totale	**326.10**	**241.20**	**84.85**
			D_Y	D_{SL}	D_{RL}

Verifichiamo empiricamente la scomposizione della devianza di cui sopra:

$$326.10 \approx 241.20 + 84.85 = 326.05$$

Possiamo anche verificare che la somma dei valori teorici di *Y* è uguale alla somma dei valori osservati.

Indice di determinazione

La misura della bontà dell'adattamento della retta ai punti osservati, è rappresentata dal rapporto:

$$r^2 = \frac{D_{SL}}{D_y} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \mu_y)^2}{\sum_{i=1}^{N}(y_i - \mu_y)^2}$$

Dato che $D_y = D_{SL} + D_{RL}$ l'indice di determinazione si può calcolare anche come

$$r^2 = 1 - \frac{D_{RL}}{D_y} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \mu_y)^2}$$

Espressione alternativa:

$$r^2 = \frac{\left[\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)\right]^2}{\sum_{i=1}^{N}(x_i - \mu_X)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2} = \frac{C_{xy}^2}{D_x D_y}$$

Proprietà dell'indice r^2

$$r^2 = \frac{D_{SL}}{D_y} = 1 - \frac{D_{RL}}{D_y}$$

- Assume valori nell'intervallo [0, 1].
- Raggiunge il **minimo** se e solo se $D_{SL} = 0$, cioè se e solo se la **retta di regressione** è **parallela all'asse delle ascisse.**
- Raggiunge il **massimo** se e solo se $D_{RL} = 0$, circostanza che si verifica se e solo se i **punti osservati giacciono su una retta.**
- In base alla prima espressione, rappresenta la **frazione della variabilità totale di Y spiegata dalla retta di regressione.**

Indice di determinazione: calcolo

$$r^2 = \frac{D_{SL}}{D_y} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \mu_y)^2}{\sum_{i=1}^{N}(y_i - \mu_y)^2} \qquad r^2 = 1 - \frac{D_{RL}}{D_y} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \mu_y)^2}$$

L'indice di determinazione con le tre formule:

$$r^2 = \frac{D_{SL}}{D_Y} = \frac{241.20}{326.10} = 0.74$$

$$r^2 = 1 - \frac{D_{RL}}{D_Y} = 1 - \frac{84.85}{326.10} = 0.74$$

$$r^2 = \frac{C_{XY}^2}{D_X D_Y} = \frac{293.10^2}{356.10 \cdot 326.10} = 0.74$$

codevianza calcolata a p. 9

x_i	y_i	$\hat{y}_i$	$(y_i - \mu_Y)^2$	$(\hat{y}_i - \mu_Y)^2$	$(y_i - \hat{y}_i)^2$
173	164	162.23	0.49	1.145	3.13
178	172	166.35	75.69	9.27	31.98
169	163	158.94	0.09	19.03	16.50
170	160	159.76	10.89	12.52	0.06
177	166	165.52	7.29	4.94	0.23
178	165	166.35	2.89	9.27	1.81
180	165	167.99	2.89	22.01	8.95
185	170	172.11	44.89	77.55	4.44
165	152	155.65	127.69	58.58	13.29
168	156	158.12	53.29	26.88	4.47
		Totale	**326.10**	**241.20**	**84.85**

D_Y D_{SL} D_{RL}

Il caso delle distribuzioni doppie di frequenze

Considerando che (x_i, y_j) si presenta con una frequenza n_{ij} e che alle le modalità x_i e y_j vanno associate le frequenze marginali n_{i0} e n_{0j}, abbiamo:

Carattere X	Carattere Y						Totale
	y_1	y_2	...	y_j	...	y_t	
x_1	n_{11}	n_{12}	⋮	n_{1j}	⋮	n_{1t}	n_{10}
x_2	n_{21}	n_{22}	⋮	n_{2j}	⋮	n_{2t}	n_{20}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	n_{i1}	n_{i2}	⋮	n_{ij}	⋮	n_{it}	n_{i0}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_s	n_{s1}	n_{s2}	⋮	n_{sj}	⋮	n_{st}	n_{s0}
Totale	n_{01}	n_{02}	⋮	n_{0j}	⋮	n_{0t}	N

$$\mu_X = \frac{1}{N}\sum_{i=1}^{s} x_i n_{i0}; \mu_Y = \frac{1}{N}\sum_{i=1}^{t} y_j n_{0j} \Rightarrow$$ Medie delle distribuzioni marginali

$$D_X = \sum_{i=1}^{s} (x_i - \mu_X)^2 n_{i0} \Rightarrow$$ Devianza di X

$$C_{XY} = \sum_{i=1}^{s}\sum_{j=1}^{t} (x_i - \mu_X)(y_j - \mu_Y) n_{ij} \Rightarrow$$ Codevianza

Il caso delle distribuzioni doppie di frequenze

$$D_Y = \sum_{j=1}^{t} (y_j - \mu_Y)^2 n_{0j}$$ ⇒ Devianza totale

Definite le quantità D_Y, D_X, C_{XY}, D_{SL} e D_{RL}, è possibile calcolare l'indice di determinazione con una delle tre formule di p. 15

$$b_1 = \frac{\sum_{i=1}^{s}\sum_{j=1}^{t}(x_i - \mu_X)(y_j - \mu_Y)n_{ij}}{\sum_{i=1}^{s}(x_i - \mu_X)^2 n_{i0}}$$

$$b_0 = \mu_Y - b_1\mu_X$$

⇒ Stima dei parametri

$$D_{SL} = \sum_{i=1}^{s} (\hat{y}_i - \mu_Y)^2 n_{i0}$$ ⇒ Devianza spiegata

$$D_{RL} = \sum_{i=1}^{s}\sum_{j=1}^{t} (y_j - \hat{y}_i)^2 n_{ij}$$ ⇒ Devianza residua

Carattere X	Carattere Y						Totale
	y_1	y_2	...	y_j	...	y_t	
x_1	n_{11}	n_{12}	⋮	n_{1j}	⋮	n_{1t}	n_{10}
x_2	n_{21}	n_{22}	⋮	n_{2j}	⋮	n_{2t}	n_{20}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	n_{i1}	n_{i2}	⋮	n_{ij}	⋮	n_{it}	n_{i0}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_s	n_{s1}	n_{s2}	⋮	n_{sj}	⋮	n_{st}	n_{s0}
Totale	n_{01}	n_{02}	⋮	n_{0j}	⋮	n_{0t}	N

Regressione nel caso delle distribuzioni di frequenze: calcoli

Distribuzione doppia di frequenze di un gruppo di pazienti per età e massa muscolare:

I numeri in rosso sono i valori centrali

Età (X)	Massa muscolare (Y)				Totale
	51-65 58	66-80 73	81-95 88	96-120 108	
41-51 46	0	0	2	13	15
51-61 56	0	5	7	5	17
61-71 66	2	8	4	0	14
71-81 76	6	6	2	0	14
Totale	8	19	15	18	60

$$\mu_X = \frac{46\cdot 15 + 56\cdot 17 + 66\cdot 14 + 76\cdot 14}{60} = 60.50$$

$$\mu_Y = \frac{53\cdot 8 + 73\cdot 19 + 88\cdot 15 + 108\cdot 18}{60} = 84.58$$

$$D_X = \frac{1}{60}\left[(46-60.50)^2\cdot 15 + (56-60.50)^2\cdot 17\right.$$

$$\left. = (66-60.50)^2\cdot 14 + (76-60.50)^2\cdot 14\right]$$

$$= 7.285.0$$

$$C_{XY} = \frac{1}{60}\left[(46-60.50)(88-85.25)\cdot 2 + (46-60.50)(108-85.25)\cdot 13 + \cdots\right.$$

$$\left. + (76-60.50)(88-85.25)\cdot 2 = -9057.50\right.$$

$$b_1 = \frac{-9057.50}{7285.0} = -1.24;\ b_0 = 85.25 - (-1.24\cdot 60.50) = 160.47$$

Regressione nel caso delle distribuzioni di frequenze: calcoli

$$D_Y = (58-84.58)^2 \cdot 8 + (73-84.58)^2 \cdot 19$$
$$= (88-84.58)^2 \cdot 15 + (108-84.58)^2 \cdot 18$$
$$= 18.221.25$$

I numeri in rosso sono i valori centrali

Età (X)	Massa muscolare (Y)				Totale
	51-65 58	66-80 73	81-95 88	96-120 108	
41-51 46	0	0	2	13	15
51-61 56	0	5	7	5	17
61-71 66	2	8	4	0	14
71-81 76	6	6	2	0	14
Totale	8	19	15	18	60

L'indice di determinazione assume il valore

$$r^2 = \frac{(-9057.50)^2}{7285.0 \cdot 18221.25} = 0.62$$

Dunque, la retta di regressione spiega il 62% della variabilità totale della massa muscolare