Statistica della Formazione

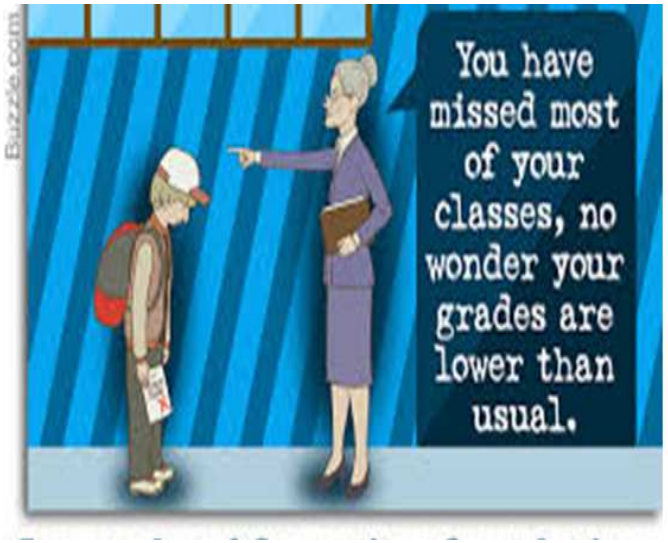
Slides 7

A.A. 2020-2021

Docente: ANNA LINA SARRA

Modulo 1: elementi di statistica descrittiva

- **Analisi delle distribuzioni doppie: analisi della regressione**



Analisi della dipendenza

Obiettivo:

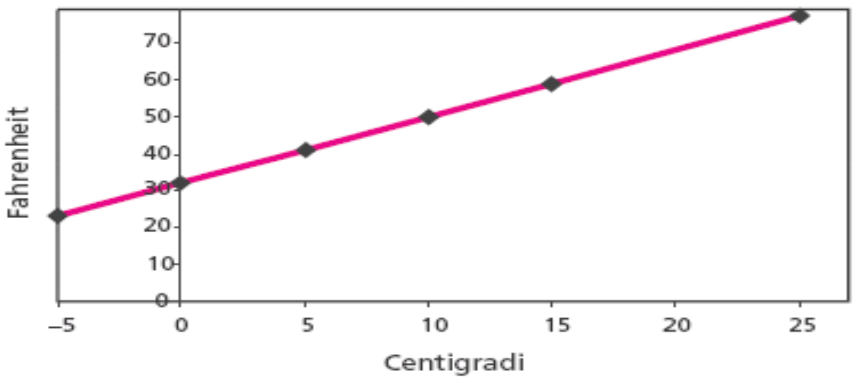
Date due variabili quantitative, X e Y, si è interessati a comprendere come la variabile Y (dipendente o risposta) sia influenzata dalla X (esplicativa o indipendente).

Y è funzione di X se ad ogni valore di X corrisponde un solo valore di Y. La relazione funzionale è lineare, se possiamo scrivere:

$$Y = \beta_0 + \beta_1 X$$

β_0=intercetta

β_1=coefficiente angolare



Relazione funzionale e statistica

Negli studi empirici, la relazione tra Y e X non è mai funzionale (a un valore X corrispondono più valori di Y).

Una relazione statistica tra la Y e la X può essere descritta da:

$$Y = f(X) + \varepsilon$$

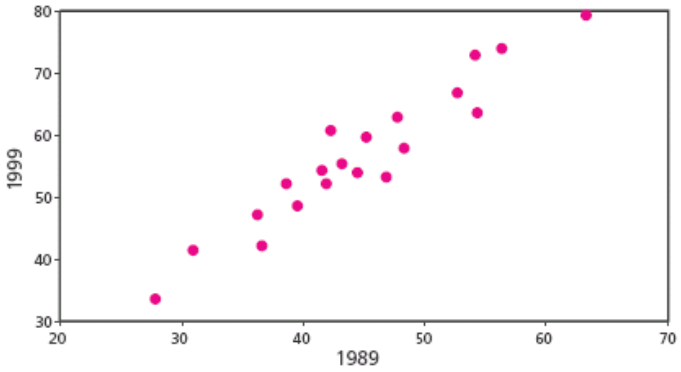
$f(X)$ definisce il contributo della X

ε rappresenta il contributo di tutti i fattori non osservati

$f(X)$ ⇒ è una componente deterministica

ε ⇒ è una componente stocastica

Y ⇒ è una variabile casuale.



Modello di regressione lineare

Il **modello di regressione di *Y* su *X*** può essere descritto 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.*

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

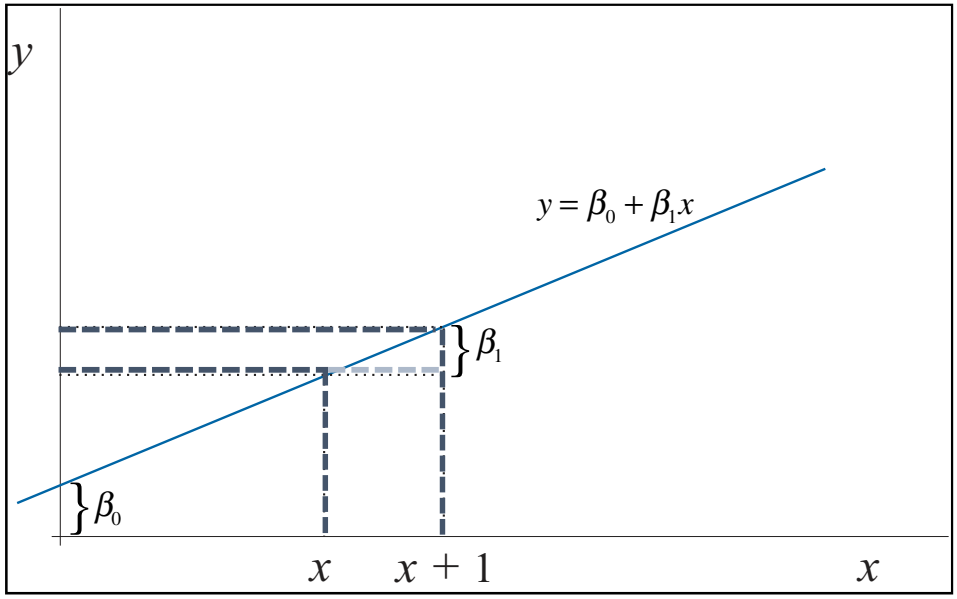
$$y = \beta_0 + \beta_1 x + \varepsilon,$$

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

Parametri del modello di regressione lineare

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

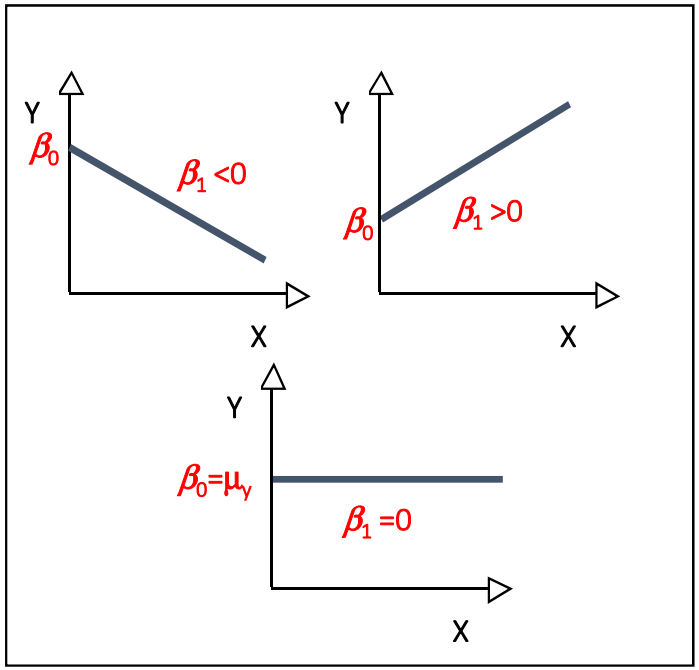
- $b_0 \rightarrow$ intercetta della retta: punto in cui la retta interseca l'asse verticale; valore della y per x=0:
- $b_1 \rightarrow$ coefficiente angolare della retta o coefficiente di regressione;



β_0 : intercetta
β_1 : coefficiente angolare

Pendenza della retta e coefficiente di regressione

- se β_1 **<0** la retta è inclinata negativamente: al crescere della variabile indipendente x la variabile dipendente y decresce;
- se β_1 **=0** la retta è parallela all'asse delle ascisse: al crescere della variabile x la y rimane costante (indipendenza lineare);
- se β_1 **>0** la retta è inclinata positivamente e al crescere della variabile x cresce anche la y



Modello di regressione lineare semplice: assunzioni

Assunzione 1: la funzione f(x) è lineare $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Assunzione 2: le $\boldsymbol{\varepsilon_i}$ sono variabili casuali indipendenti con valore atteso nullo,

$\boldsymbol{E(\varepsilon_i) = 0}$, e varianza costante, $\boldsymbol{Var(\varepsilon_i) = \sigma^2}$, per ogni i=1,…,n

Assunzione 3: i valori della variabile esplicativa X sono noti senza errore

x_i

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.

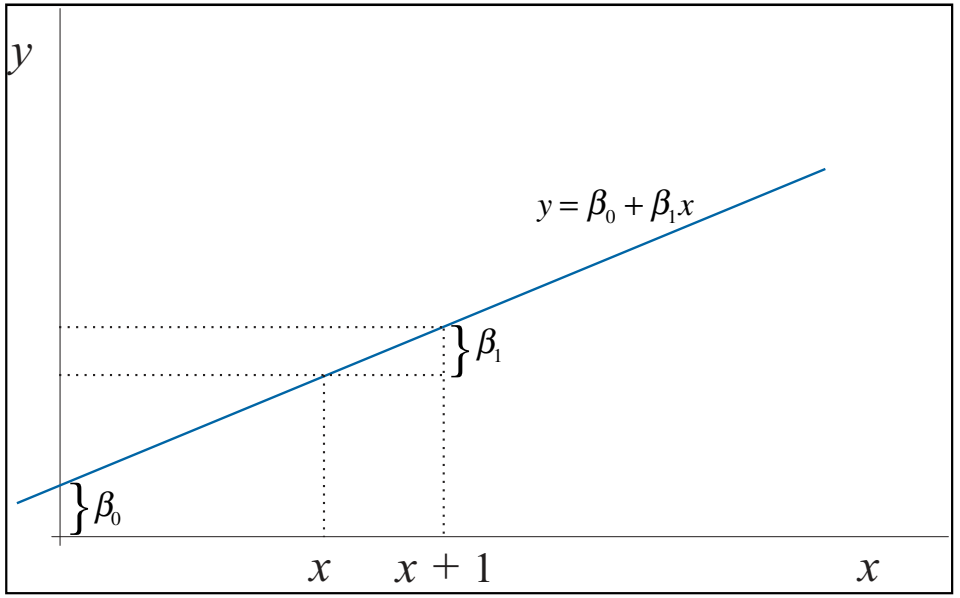
Vogliamo stabilire se la retta è una funzione adatta a esprimere il legame associativo tra il numero di articoli prodotti ed il numero di ore di lavoro.

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

Regressione lineare

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

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



Regressione lineare

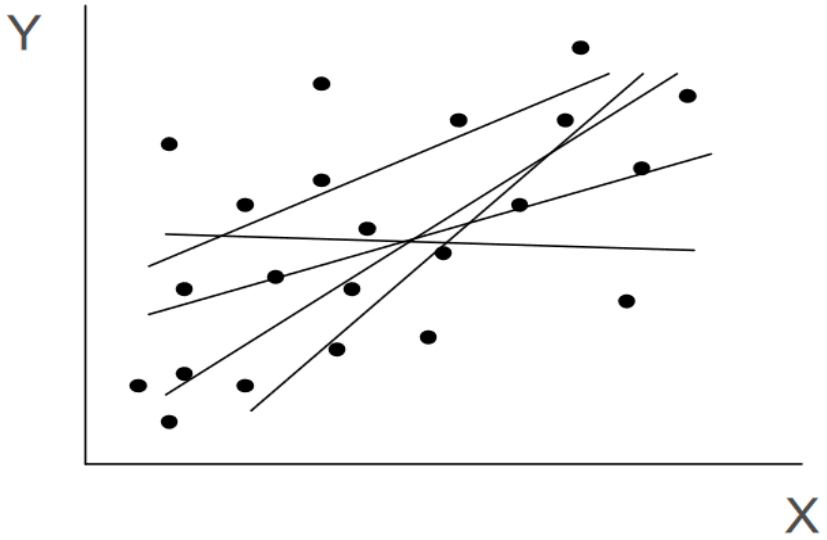
- $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ sono le coppie di valori osservati su N unità statistiche, dette **punti osservati** o **nuvola di punti**.

- Il problema è quello di **assegnare ai parametri** β_0 e β_1 della retta i valori che consentano di approssimare **nel miglior modo possibile** la nuvola dei punti.

In altri termini, dobbiamo determinare quella retta - tra le infinite del piano -, che meglio si adatta alla nuvola di punti.

La soluzione viene trovata utilizzando il metodo **dei minimi quadrati.**

Come scegliere la retta migliore?



Per un insieme di punti possono passare infinite rette!

Come scegliere la retta "migliore"?

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.*

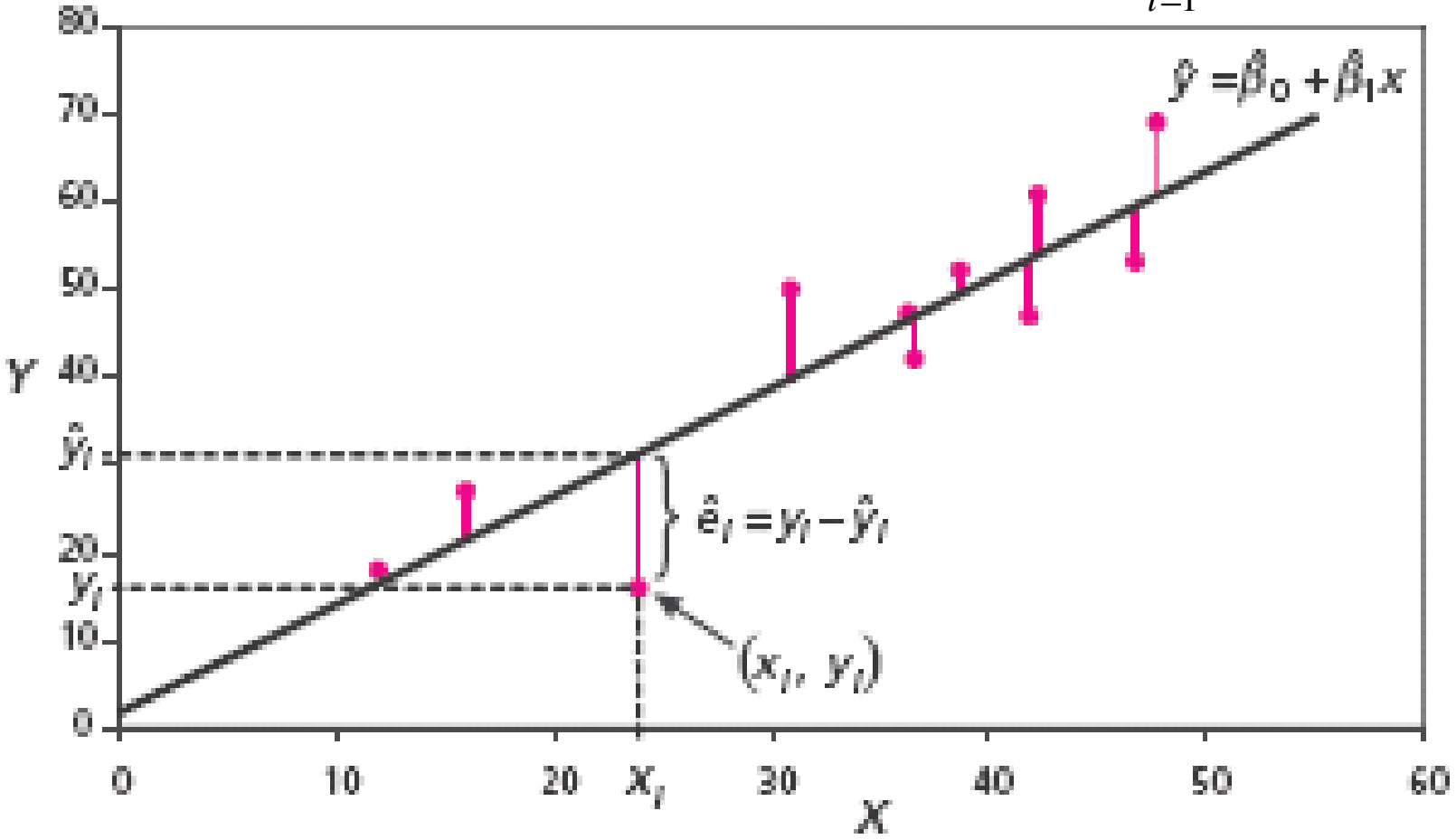
Con il **metodo dei minimi quadrati** si assegnano a b_0 e b_1 i valori che rendono minima la quantità S_q, data da

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

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

Rappresentazione grafica

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



Metodo dei minimi quadrati: stima dei parametri

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

codevianza

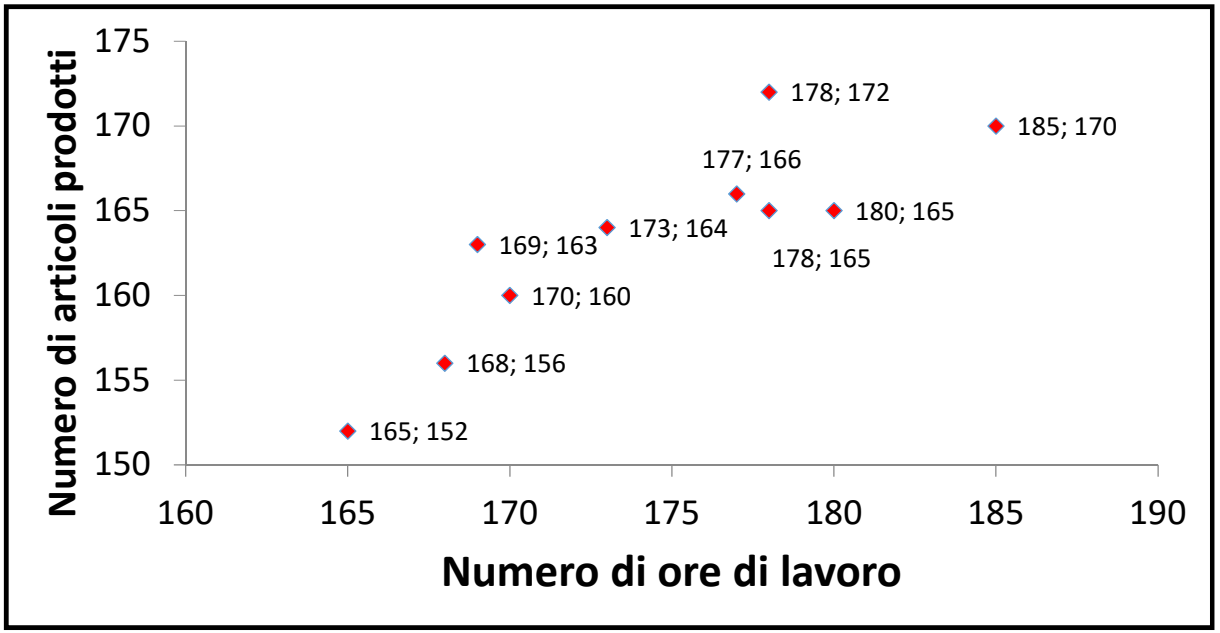
$$b_1 = \frac{C_{XY}}{D_X} = \frac{\sum_{i=1}^{N} x_i y_i - N\mu_X\mu_Y}{\sum_{i=1}^{N} x_i^2 - N\mu_X} = \frac{\sum_{i=1}^{N} (x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{N} (x_i - \mu_X)^2}$$

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

devianza

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.

Parametri della retta di regressione

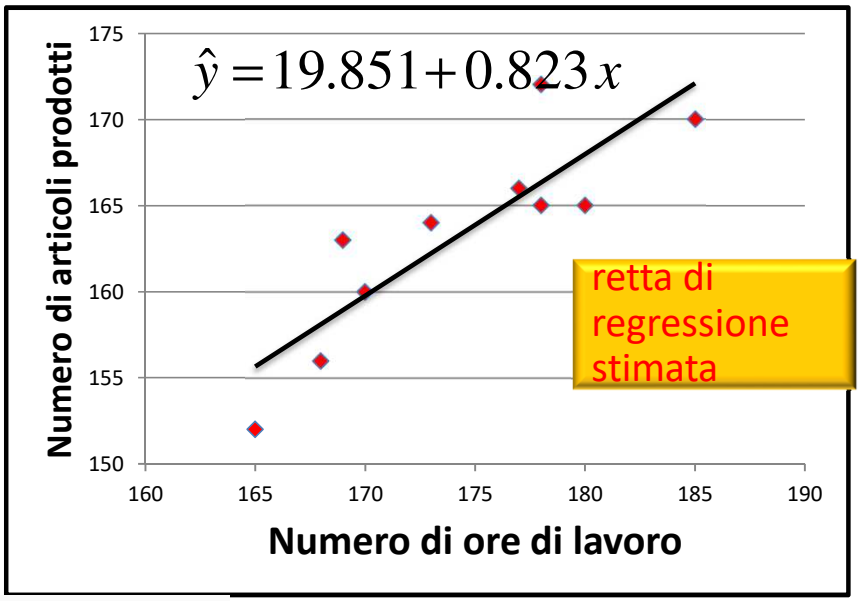
$$b_1 = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{N}(x_i - \mu_X)^2}$$

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

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

μ_X	= 174.30
μ_Y	= 163.30

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



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

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

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.

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

$$SQT = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Somma totale dei quadrati

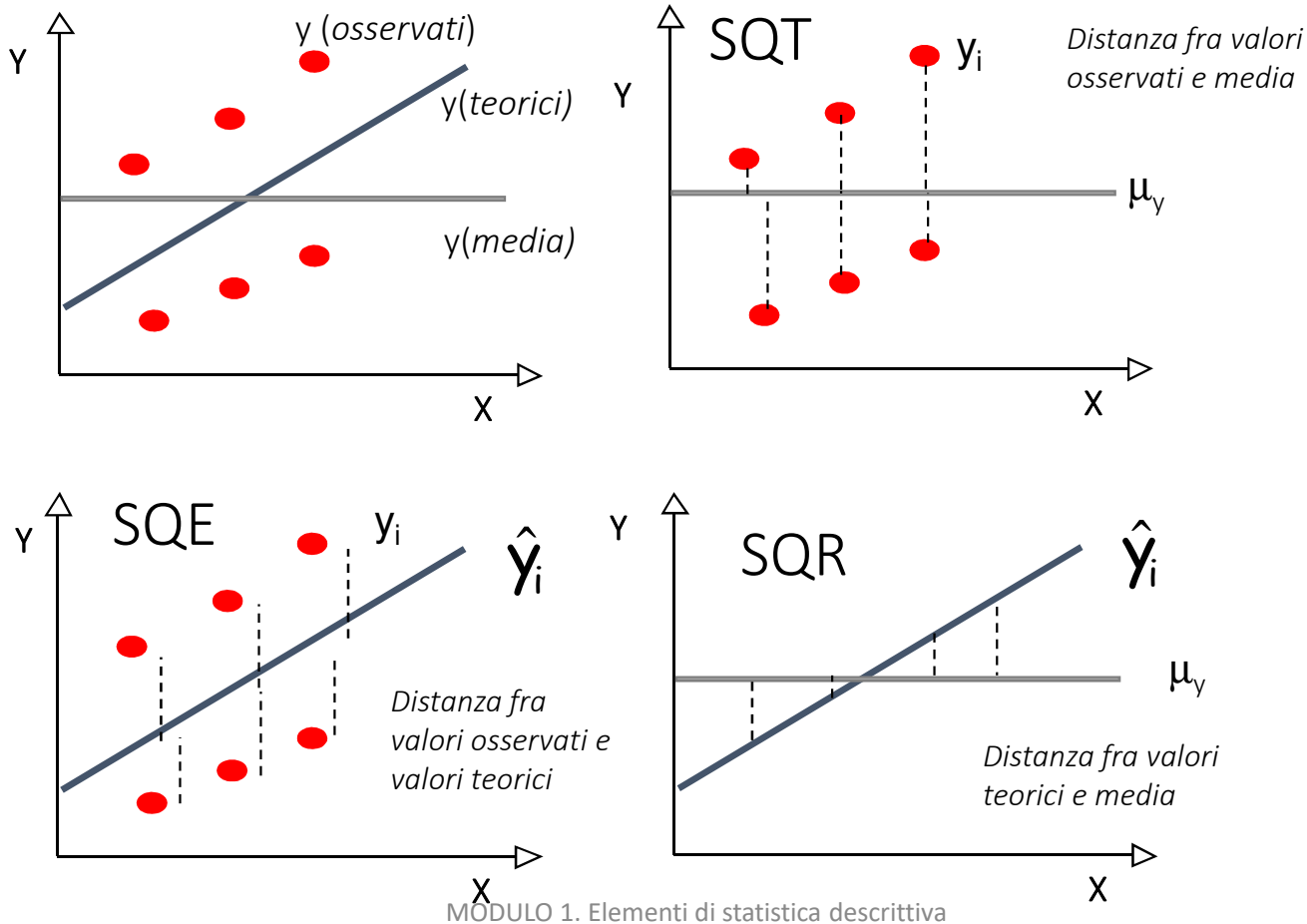
$$SQR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Somma dei quadrati della regressione

$$SQE = \sum_{i=1}^{n}\hat{e}_i^2$$

Somma dei quadrati degli errori

Scomposizione della devianza nel modello di regressione: interpretazione grafica



Coefficiente di determinazione

Dalla relazione SQT=SQR+SQE si può definire un indice che misura la bontà di adattamento della retta di regressione.

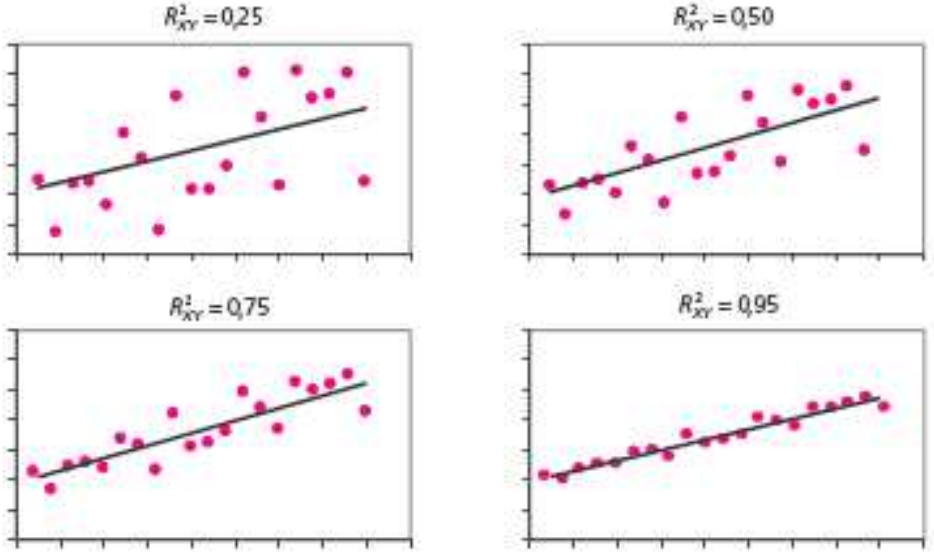
Il rapporto

$$R^2_{XY} = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

è detto coefficiente di determinazione e indica la proporzione di variabilità di Y spiegata dalla variabile esplicativa X, attraverso il modello di regressione.

Si può dimostrare che il coefficiente di determinazione corrisponde al quadrato del coefficiente di correlazione lineare:

$$R^2_{XY} = \rho^2_{XY}$$



Proprietà del coefficiente di determinazione

$$R^2_{XY} = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

- Assume valori nell'intervallo [0, 1].
- Raggiunge il **minimo** se e solo se *SQR* = 0, cioè se e solo se la **retta di regressione** è **parallela all'asse delle ascisse.**
- Raggiunge il **massimo** se e solo se *SQE* = 0, circostanza che si verifica se e solo se i **punti osservati giacciono su una retta.**
- Rappresenta la frazione della variabilità totale di Y spiegata dalla retta di regressione.

Indice di determinazione calcolo

$$R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$R^2 = 1 - \frac{SQE}{SQT} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

L'indice di determinazione con le tre formule:

$$R^2 = \frac{SQR}{SQT} = \frac{241.20}{326.10} = 0.74$$

$$R^2 = 1 - \frac{SQE}{SQT} = 1 - \frac{84.85}{326.10} = 0.74$$

$$R^2 = r^2 = \frac{C_{XY}^2}{D_X D_y} = \frac{293.10^2}{356.10 * 326.10} = 0.74$$

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**
			SQT	*SQR*	*SQE*