CAPITOLO 4

INTERPOLAZIONE STATISTICA REGRESSIONE E CORRELAZIONE

Nello studio dei legami fra due variabili si cerca di determinare, partendo da un insieme di coppie $(x_i; y_i)$ di dati rilevati, una funzione $y = f(x)$ che rappresenti il fenomeno in modo da descrivere sinteticamente la relazione fra le due variabili osservate, determinare la legge di distribuzione dei dati statistici, ricavare eventuali dati intermedi mancanti. Per trovare tale funzione si può procedere in due modi:

- Determinare una funzione che assuma esattamente i valori $(x_i; y_i)$ rilevati; questa procedura viene detta interpolazione **per** punti noti o **interpolazione matematica**.
- Determinare una funzione il cui grafico si "avvicini" il più possibile ai punti del diagramma a dispersione; questa procedura è detta interpolazione (o perequazione) **fra** punti noti o **interpolazione statistica**.

Poiché in statistica si ha a che fare con un numero, in genere, molto elevato di punti, sarebbe molto difficile utilizzare l'interpolazione matematica e si preferisce usare l'interpolazione statistica, mediante il **metodo dei minimi quadrati**.

1. METODO DEI MINIMI QUADRATI

Si considerino due variabili X e Y sulle quali si sono effettuate n rilevazioni espresse dalle coppie di valori $(x_i; y_i)$. Tali coppie, rappresentate in un diagramma cartesiano mediante punti, formano il **diagramma a dispersione**, che può assumere varie forme. La funzione interpolante indica il cosiddetto movimento tendenziale o trend del fenomeno, cioè l'andamento di fondo, indipendentemente dalle oscillazioni di varia natura. E' necessario:

1. scegliere il tipo di funzione (lineare, quadratica, esponenziale, ecc.) che esprima meglio il legame tra X e Y;
2. calcolare i parametri della funzione scelta. Per questo il metodo più utilizzato è il **metodo dei minimi quadrati**.

Scelta la funzione perequatrice $y = f(x; a, b, c, \dots, k)$ e calcolati su di essa i valori teorici $\hat{y}_i$ corrispondenti ai valori x_i rilevati, sostituendo ai valori y_i i valori teorici $\hat{y}_i$, si commettono errori dati da:

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

che possono essere positivi, negativi o nulli. E' necessario minimizzare tali errori, ma non sarebbe corretto minimizzarne la loro somma poiché errori positivi potrebbero compensare quelli negativi, quindi per ovviare all'inconveniente si minimizza la somma dei loro quadrati, ossia

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

Per trovare il minimo di tale funzione si deve risolvere il sistema che rende nulle tutte le derivate parziali rispetto ai parametri $a, b, c, \dots, k$. La funzione che si ottiene con questo metodo è la migliore tra le funzioni del tipo scelto, cioè se dal grafico a dispersione risulta che il migliore accostamento si ottiene con una retta e si interpola con una parabola, questa non è la curva migliore, ma fra tutte le parabole è quella che più si accosta ai valori rilevati.

Una volta trovata la funzione interpolante è necessario verificare se sia accettabile. Per questo scopo si calcolano possono calcolare i seguenti indicatori.

ERRORE STANDARD

Dato dalla media quadratica delle differenze, quindi legato alle unità di misura dei dati:

$$S_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

INDICI DI SCOSTAMENTO

Indice lineare relativo

$$I_1 = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}\hat{y}_i}$$

Indice quadratico relativo

$$I_2 = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}}{\frac{\sum_{i=1}^{n}\hat{y}_i}{n}}$$

Fra questi due indici è preferibile il secondo poiché il metodo dei minimi quadrati lavora sui quadrati delle differenze. L'indice I_2 è il rapporto tra l'errore standard e il valore medio dei valori teorici ricavati. I valori ottenuti vanno considerati in relazione al fenomeno; in ogni caso per avere un buon accostamento non devono superare 0,1 (in certi casi 0,01); è chiaro che tanto più piccoli sono i valori degli indici di scostamento, tanto migliore è l'accostamento.

COEFFICIENTE DI DETERMINAZIONE

Si utilizza se lo scopo della ricerca della funzione è quello di avere un modello matematico del fenomeno, poiché tiene conto degli scarti dei valori delle y_i.

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

con $\bar{y}$ valore medio degli y_i. Quanto più δ si "avvicina" a 1 tanto più il modello rappresenta bene il fenomeno.

2. FUNZIONI INTERPOLANTI

2.1 FUNZIONE DI PRIMO GRADO: $y = a + bx$

Utilizzando il metodo dei minimi quadrati, si deve rendere minima la funzione

$$\varphi(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

annullandone le derivate parziali prime. Tralasciando i calcoli si ottiene:

$$\begin{cases} a = \dfrac{\sum y_i \cdot \sum x_i^2 - \sum y_i x_i \cdot \sum x_i}{n\sum x_i^2 - (\sum x_i)^2} \\ b = \dfrac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \end{cases}$$

L'equazione della **retta interpolante** (o perequatrice) fra punti noti, ottenuta con il metodo dei minimi quadrati, è $y - \bar{y} = b(x - \bar{x})$

con $\bar{x}$ e $\bar{y}$ le medie aritmetiche, rispettivamente, di x_i e y_i. Tale retta interpolante passa per il punto di coordinate $(\bar{x}; \bar{y})$, detto **baricentro della distribuzione**. Si perviene all'equazione della retta interpolante anche utilizzando le seguenti formule:

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \end{cases}$$

Osservazione: se si opera con valori molto grandi queste ultime formule permettono di ridurre i calcoli, ma nel casi di valori medi approssimati la propagazione dell'errore può dare risultati meno precisi e quindi sono preferibili le prime formule fornite.

ESEMPI

1. In un esperimento si sono misurate le lunghezze di una molla sottoposta a successivi carichi e si sono ottenute le seguenti rilevazioni:

Pesi (Kg)	1	2	3	4	5
Lunghezze (cm)	24	27	29,6	33	36,4

Determinare e rappresentare graficamente la retta interpolante fra punti noti.

Per il calcolo dei coefficienti della retta interpolante impostiamo la tabella seguente:

	pesi x	lunghezze y	x^2	xy	$\hat{y}$	$y - \hat{y}$	$\lvert y - \hat{y}\rvert$	$(y - \hat{y})^2$
	1	24	1	24	23,84	0,16	0,16	0,0256
	2	27	4	54	26,92	0,08	0,08	0,0064
	3	29,6	9	88,8	30	-0,4	0,4	0,16
	4	33	16	132	33,08	-0,08	0,08	0,0064
	5	36,4	25	182	36,16	0,24	0,24	0,0576
totali	15	150	55	480,8	150		0,96	0,256

Da cui si ottiene:

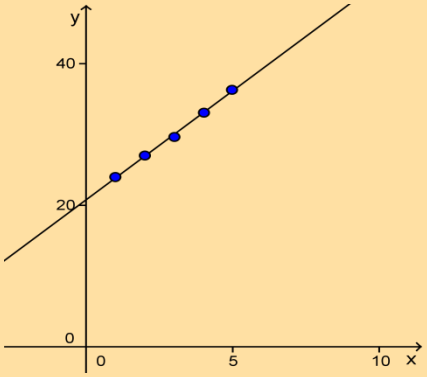
$$a = \frac{150 \cdot 55 - 480{,}8 \cdot 15}{5 \cdot 55 - 15^2} = 20{,}76 \qquad b = \frac{5 \cdot 480{,}8 - 15 \cdot 150}{5 \cdot 55 - 15^2} = 3{,}08$$

quindi la retta interpolante ha equazione: $y = 20{,}76 + 3{,}08x$

Ora si possono calcolare le ultime 4 colonne della tabella per trovare gli indici di scostamento:

$$I_1 = \frac{0{,}96}{150} = 0{,}0064 \qquad I_2 = \frac{\sqrt{\frac{0{,}256}{5}}}{\frac{150}{5}} = 0{,}0075$$

da cui si deduce che la retta trovata realizza un "ottimo accostamento". Ora su uno stesso piano cartesiano si rappresentano il diagramma a dispersione e la retta interpolante:



2. Data la seguente tabella della rilevazione dei movimenti sismici con magnitudo compresa tra 4.5 e 4.9 (ISTAT, *Annuario Statistico Italiano* 2012 Tav. 1.3):

Anni	2004	2005	2006	2007	2008	2009	2010	2011
N° movimenti sismici	2	5	3	3	1	7	1	4

determinare l'equazione della retta interpolante con il metodo dei minimi quadrati.

Per il calcolo dei coefficienti della retta interpolante impostiamo la tabella seguente, dopo aver calcolato i valori medi:

$$\bar{x} = \frac{36}{8} = 4{,}5 \qquad \bar{y} = \frac{26}{8} = 3{,}25$$

anni	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) \cdot (y - \bar{y})$	$(x - \bar{x})^2$
2004	1	2	-3,5	-1,25	4,375	12,25
2005	2	5	-2,5	1,75	-4,375	6,25
2006	3	3	-1,5	-0,25	0,375	2,25
2007	4	3	-0,5	-0,25	0,125	0,25
2008	5	1	0,5	-2,25	-1,125	0,25
2009	6	7	1,5	3,75	5,625	2,25
2010	7	1	2,5	-2,25	-5,625	6,25
2011	8	4	3,5	0,75	2,625	12,25
totali	36	26			2	42

Da cui si ottiene:

$$b = \frac{2}{42} = 0{,}05 \qquad a = 3{,}25 - 0{,}05 \cdot 4{,}5 = 3{,}036$$

quindi la retta interpolante ha equazione: $y = 3{,}036 + 0{,}05x$

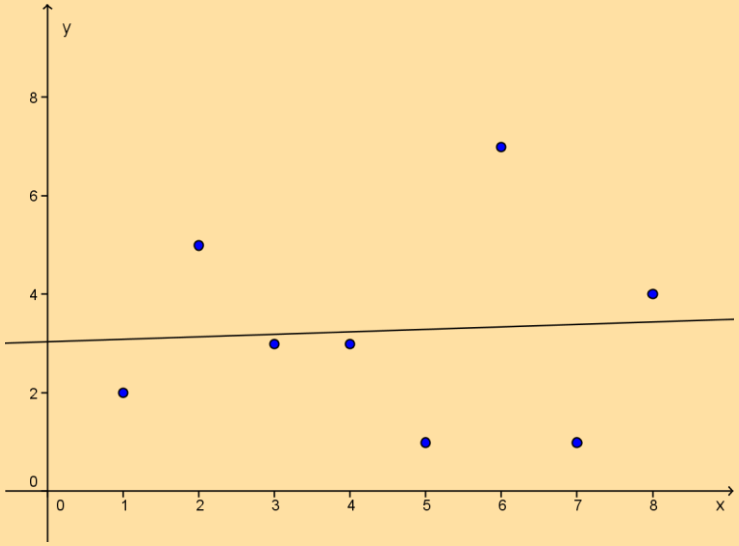
Gli indici di scostamento risultano:

$$I_1 = 0{,}47 \qquad I_2 = 0{,}59$$

da cui si deduce che la retta **non** ha un buon accostamento ai dati rilevati. In più, il coefficiente di determinazione è:

$$\delta = 0{,}0032$$

che indica che il modello lineare non è accettabile. Vediamo graficamente tale situazione:



2.2 FUNZIONE DI SECONDO GRADO: $y = a + bx + cx^2$

Utilizzando il metodo dei minimi quadrati, si deve rendere minima la funzione

$$\varphi(a,b,c) = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

annullandone le derivate parziali prime. Tralasciando i calcoli si ottiene il seguente sistema:

$$\begin{cases} na + b\sum x_i + c\sum x_i^2 = \sum y_i \\ a\sum x_i + b\sum x_i^2 + c\sum x_i^3 = \sum x_i y_i \\ a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 = \sum x_i^2 y_i \end{cases}$$

da cui si ricavano a, b, c.

Osservazione: con procedimento analogo si possono determinare i parametri di un polinomio di grado qualunque $y = a_0 + a_1 x + a_2 x^2 + \ldots + a_r x^r$.

ESEMPIO

Sono assegnati i punti $A(0;1)$, $B(1;4)$, $C(2;3)$, $D(3;0)$. Determinare l'equazione della parabola interpolante col metodo dei minimi quadrati.

Per l'impostazione del sistema si costruisce la tabella:

	x	y	x^2	x^3	x^4	xy	x^2y	$\hat{y}$	$y-\hat{y}$	$\lvert y-\hat{y}\rvert$	$(y-\hat{y})^2$
	0	1	0	0	0	0	0	1,1	-0,1	0,1	0,01
	1	4	1	1	1	4	4	3,7	0,3	0,3	0,09
	2	3	4	8	16	6	12	3,3	-0,3	0,3	0,09
	3	0	9	27	81	0	0	-0,1	0,1	0,1	0,01
totali	6	8	14	36	98	10	16	8		0,8	0,2

quindi si ottiene:

$$\begin{cases} 4a + 6b + 14c = 8 \\ 6a + 14b + 36c = 10 \\ 14a + 36b + 98c = 16 \end{cases}$$

Risolvendo il sistema, si trova $a = 1{,}1 \quad b = 4{,}1 \quad c = -1{,}5$.

L'equazione della parabola interpolante è $y = 1{,}1 + 4{,}1x - 1{,}5x^2$

Gli indici di scostamento sono:

$$I_1 = \frac{0{,}8}{8} = 0{,}1 \qquad I_2 = \frac{\sqrt{\frac{0{,}2}{4}}}{\frac{8}{4}} = 0{,}11$$

quindi l'accostamento è accettabile.

2.3 FUNZIONE ESPONENZIALE: $y = a \cdot b^x$ (con $a > 0,\ b > 0,\ b \neq 1$)

E' necessario, prima di applicare il metodo dei minimi quadrati, operare un cambio di variabili per rendere il sistema da risolvere lineare. Quindi, alla funzione esponenziale si applicano i logaritmi decimali e si ottiene:

$$Log\ y = Log\ a + x Log\ b$$

Posto $Log\ a = c$, $Log\ b = d$ e $Log\ y = z$ si ottiene:

$$z = c + dx$$

Ora applicando il metodo dei minimi quadrati, si trova la retta interpolante con:

$$d = \frac{n\sum(x_i Logy_i) - \sum x_i \cdot \sum Logy_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad c = \frac{\sum Logy_i}{n} - d \cdot \bar{x}$$

Conoscendo c e d, si può ora ritornare ad a e b.

Osservazione: Se si riportassero su un grafico i punti $(x_i; Logy_i)$ si avrebbe una rappresentazione grafica in scala semilogaritmica, in tale scala la curva esponenziale diventa una retta.

Oltre alla funzione esponenziale con base variabile si utilizza anche la funzione

$$y = a \cdot e^{mx}$$

dove m rappresenta il **tasso di crescita o di diminuzione del fenomeno**. Si procede come indicato precedentemente solo che si lavora con i logaritmi naturali invece che decimali.

ESEMPIO

Si consideri la seguente tabella:

X	0	1	2	3	4	5
Y	24	22	20,4	18,5	16,8	15,5

Trovare la curva esponenziale perequatrice e rappresentare graficamente.

Per l'impostazione del sistema si costruisce la tabella:

	x	y	$z = Log\ y$	xz	x^2		$\hat{y}$	$y - \hat{y}$	$\lvert y - \hat{y}\rvert$	$(y - \hat{y})^2$	$(y - \bar{y})^2$
	0	24	1,380211	0	0		24,08733	-0,08733	0,087334	0,007627	19,95111
	1	22	1,342423	1,342423	1		22,05016	-0,05016	0,050157	0,002516	6,084444
	2	20,4	1,30963	2,61926	4		20,18527	0,214726	0,214726	0,046107	0,751111
	3	18,5	1,267172	3,801515	9		18,47811	0,021887	0,021887	0,000479	1,067778
	4	16,8	1,225309	4,901237	16		16,91533	-0,11533	0,115334	0,013302	7,471111
	5	15,5	1,190332	5,951658	25		15,48473	0,015274	0,015274	0,000233	16,26778
totali	15	117,2	7,715077	18,61609	55		117,2009	-0,00094	0,504712	0,070264	51,59333

con $\bar{x} = \frac{15}{6} = 2{,}5$ $\quad \bar{y} = \frac{117{,}2}{6} = 19{,}5\overline{3}$ $\quad \bar{z} = \frac{7{,}715077}{6} = 1{,}286$. Quindi:

$$d = \frac{6 \cdot 18{,}61609 - 15 \cdot 7{,}715077}{6 \cdot 55 - 15^2} = -0{,}03838 \qquad c = 1{,}286 + 0{,}03838 \cdot 2{,}5 = 1{,}382$$

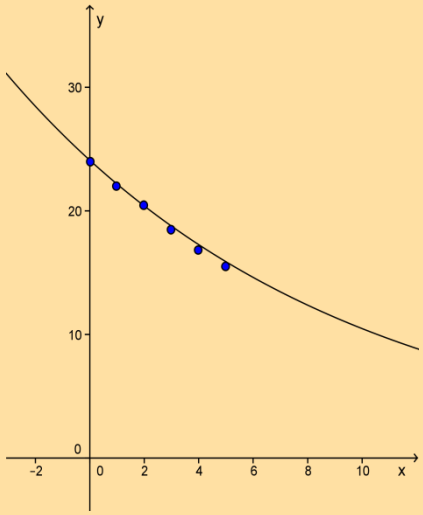
Tornando ai parametri a e b: $\quad a = 10^c = 10^{1{,}382} = 24{,}09 \qquad b = 10^d = 0{,}92$

da cui, la funzione esponenziale interpolante: $\quad y = 24{,}09 \cdot 0{,}92^x$

Gli indici di scostamento ed il coefficiente di determinazione sono:

$$I_1 = 0{,}0043 \qquad I_2 = 0{,}0055 \qquad \delta = 0{,}9986$$

quindi l'accostamento è ottimo. Graficamente:



2.4 FUNZIONE POTENZA: $y = a \cdot x^b$ (con $a > 0$)

Questa funzione è utilizzata spesso in economia per interpolare la relazione fra la domanda di un bene ed il relativo prezzo, in quanto tale funzione permette di determinare una curva di domanda ad elasticità costante. Anche in questo caso si passa ai logaritmi decimali:

$$Log\ y = Log\ a + bLog\ x$$

Posto $Log\ a = c$, $Log\ x = t$ e $Log\ y = z$ si ottiene:

$$z = c + bt$$

Ricavando b e c col metodo dei minimi quadrati, si ritorna alla funzione richiesta.

ESEMPIO

Si consideri la domanda di un bene al variare del prezzo:

prezzi	5	6	7	8	9
domanda	900	820	760	700	660

Trovare la funzione interpolante di equazione $y = a \cdot x^b$

Per l'impostazione del sistema si costruisce la tabella:

	x	y	$t = Logx$	$z = Log\ y$	tz	t^2
	5	900	0,69897	2,954243	2,064927	0,488559
	6	820	0,778151	2,913814	2,267388	0,605519
	7	760	0,845098	2,880814	2,43457	0,714191
	8	700	0,90309	2,845098	2,56938	0,815572
	9	660	0,954243	2,819544	2,690529	0,910579
totali	35	3840	4,179552	14,41351	12,02679	3,534419

con $\bar{t} = 0{,}83591$ $\bar{y} = 768$ $\bar{z} = 2{,}8827$. Quindi:

$$b = \frac{5 \cdot 12{,}02679 - 4{,}179552 \cdot 14{,}41351}{5 \cdot 3{,}534419 - 4{,}179552^2} = -0{,}53113$$

$$c = 2{,}8827 + 0{,}53113 \cdot 0{,}83591 = 3{,}3267$$

Tornando ai parametri iniziali: $a = 10^c = 10^{3{,}3267} = 2.121{,}672$ da cui, la funzione interpolante:

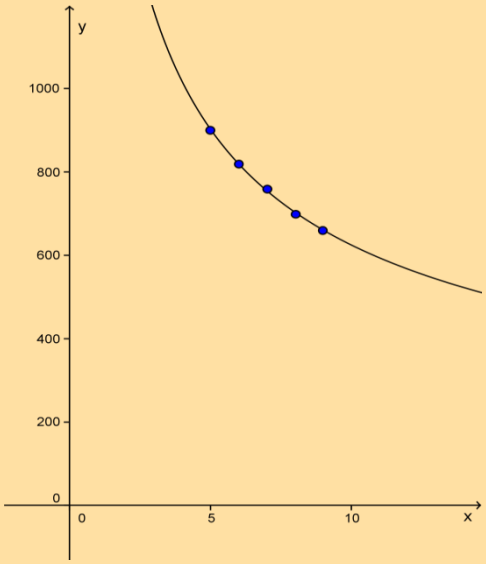
$$y = 2.121{,}672 \cdot x^{-0{,}53113}$$

Si completa la tabella per calcolare gli indici di scostamento ed il coefficiente di determinazione:

$\hat{y}$	$y-\hat{y}$	$\lvert y-\hat{y} \rvert$	$(y-\hat{y})^2$	$(y-y_m)^2$
902,4753	-2,47527	2,47527	6,126961	17424
819,181	0,818956	0,818956	0,670689	2704
754,7838	5,216242	5,216242	27,20918	64
703,1069	-3,10692	3,106918	9,652937	4624
660,4696	-0,46956	0,469563	0,22049	11664
3840,017	-0,01655	12,08695	43,88025	36480

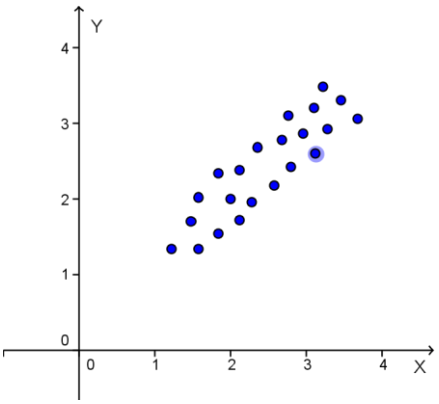
$$I_1 = 0{,}0031 \qquad I_2 = 0{,}0039 \qquad \delta = 0{,}9988$$

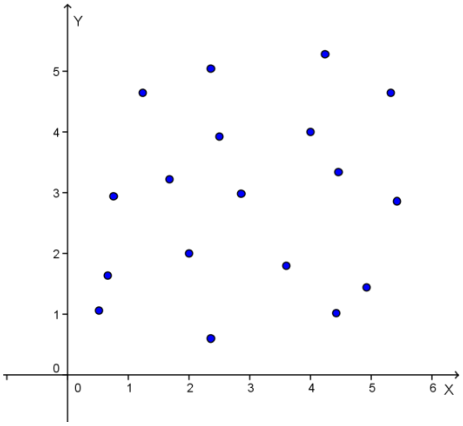
quindi l'accostamento è ottimo. Graficamente:



3. REGRESSIONE LINEARE

Date due variabili statistiche X e Y, lo studio della **regressione** consiste nel determinare una funzione matematica che esprima la relazione fra le due variabili. Per semplicità considereremo solo la **regressione lineare**. Sia X la variabile indipendente ed Y la variabile dipendente; se esistesse una relazione lineare, i punti del diagramma a dispersione si distribuirebbero vicino ad una retta, altrimenti se i punti fossero molto dispersi non esisterebbe alcuna relazione.





Applicando il metodo dei minimi quadrati si ottiene la **retta di regressione di Y rispetto a X**:

$$y = a_1 + b_1 x$$

$$\text{con} \qquad b_1 = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad a_1 = \bar{y} - b_1 \bar{x}$$

il coefficiente angolare b_1 è detto **coefficiente di regressione di Y rispetto a X** e indica di quanto varia la Y al variare di un'unità di X e se Y è crescente o decrescente.

Se la scelta della variabile indipendente non è fissata dal problema, si può calcolare anche la **retta di regressione di X rispetto a Y** la cui equazione è:

$$x = a_2 + b_2 y$$

$$\text{con} \qquad b_2 = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum y_i^2 - (\sum y_i)^2} \qquad a_2 = \bar{x} - b_2\bar{y}$$

b_2 (che non è il coefficiente angolare, ma il suo reciproco) è detto **coefficiente di regressione di X rispetto a Y** e indica di quanto varia la X al variare di un'unità di Y.

Osservazioni:

- Se b_1 e b_2 sono positivi, quando cresce una variabile cresce anche l'altra, se sono negativi al crescere di una variabile l'altra diminuisce.
- Entrambe le rette di regressione passano per il baricentro della distribuzione, ossia il punto $(\bar{x}; \bar{y})$ le cui coordinate sono le medie aritmetiche rispettivamente dei valori di X e di Y.
- Le rette di regressione coincidono quando tutti i punti del diagramma a dispersione appartengono ad una retta; invece quanto maggiore è la dispersione, tanto maggiore è l'angolo formato dalle due rette; quindi più piccolo è l'angolo tra le due rette, migliore è l'approssimazione che le rette danno alla distribuzione.
- Se $b_1 = b_2 = 0$, le rette hanno equazione $x = \bar{x}$ e $y = \bar{y}$, cioè sono parallele agli assi cartesiani. Questo significa che non esiste regressione lineare tra le due variabili, ma potrebbero essere legate da una relazione di tipo parabolico , esponenziale, ecc.

ESEMPI

1. Data la tabella della rilevazione in 5 aziende del n° dei dipendenti e del fatturato (in milioni di euro), determinare le rette di regressione e rappresentare graficamente i risultati ottenuti.

Azienda	n° dipendenti	Fatturato
A	27	240
B	30	416
C	39	304
D	54	440
E	60	360

Si costruisce la tabella:

	x	y	x^2	xy	y^2
	27	240	729	6480	57600
	30	416	900	12480	173056
	39	304	1521	11856	92416
	54	440	2916	23760	193600
	60	360	3600	21600	129600
totali	210	1760	9666	76176	646272

$$\bar{x} = \frac{210}{5} = 42 \qquad \bar{y} = \frac{1.760}{5} = 352$$

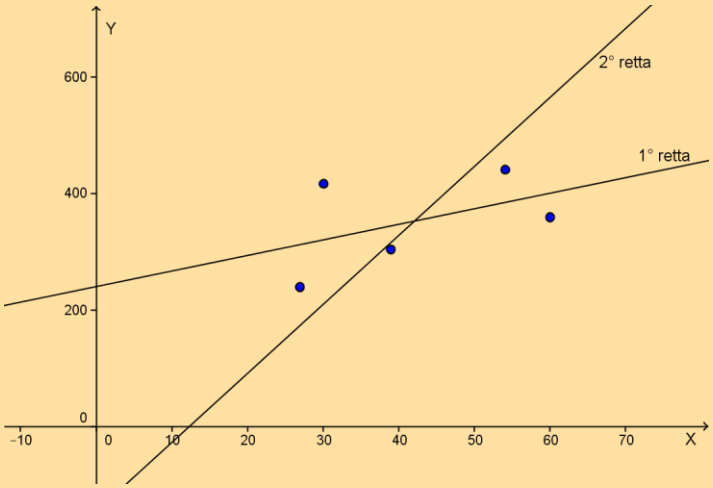
Si calcolano i coefficienti di regressione:

$$b_1 = \frac{5 \cdot 23.760 - 210 \cdot 1.760}{5 \cdot 9.666 - 210^2} = 2.\bar{6} \qquad b_2 = \frac{5 \cdot 23.760 - 210 \cdot 1.760}{5 \cdot 646.272 - 1.760^2} = 0{,}08433$$

Le rette di regressione hanno equazione:

$$y = 240 + 2{,}67x \qquad x = 12{,}316 + 0{,}08433y$$

Dal momento che i coefficienti di regressione sono positivi al crescere di una variabile cresce anche l'altra. Graficamente:



2. Si consideri la richiesta di un bene secondo la seguente tabella:

Prezzo (euro)	5	5,3	5,6	5,7	6,4	6,5
Quantità (Kg)	220	208	196	192	164	160

determinare le rette di regressione e rappresentare graficamente i risultati ottenuti.

Si costruisce la tabella:

	x	y	x^2	xy	y^2
	5	220	25	1100	48400
	5,3	208	28,09	1102,4	43264
	5,6	196	31,36	1097,6	38416
	5,7	192	32,49	1094,4	36864
	6,4	164	40,96	1049,6	26896
	6,5	160	42,25	1040	25600
totali	34,5	1140	200,15	6484	219440

$$\bar{x} = \frac{34{,}5}{6} = 5{,}75 \qquad \bar{y} = \frac{1.140}{6} = 190$$

Si calcolano i coefficienti di regressione:

$$b_1 = \frac{6 \cdot 6.484 - 34{,}5 \cdot 1.140}{6 \cdot 200{,}15 - 34{,}5^2} = -40 \qquad b_2 = \frac{6 \cdot 6.484 - 34{,}5 \cdot 1.140}{6 \cdot 219.440 - 1.140^2} = -0{,}025$$

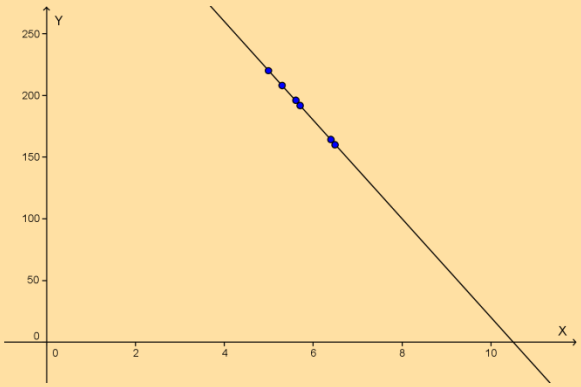
Le rette di regressione hanno equazione:

$$y = 420 - 40x \qquad x = 10{,}5 - 0{,}025y$$

Considerando la seconda retta ed esplicitando rispetto ad y, si ottiene:

$$0{,}025y = 10{,}5 - x \quad \Rightarrow \quad y = \frac{10{,}5}{0{,}025} - \frac{1}{0{,}025}x \quad \Rightarrow \quad y = 420 - 40x$$

che coincide perfettamente con la prima retta trovata, cioè esprime la legge della domanda in modo esatto. Graficamente:



3. Date le seguenti variabili:

X	6	7	8	9	10
Y	25	29	36	31	24

determinare le rette di regressione e rappresentare graficamente i risultati ottenuti.

Si costruisce la tabella:

	x	y	x^2	xy	y^2
	6	25	36	150	625
	7	29	49	203	841
	8	36	64	288	1296
	9	31	81	279	961
	10	24	100	240	576
totali	40	145	330	1160	4299

$$\bar{x} = \frac{40}{5} = 8 \qquad \bar{y} = \frac{145}{5} = 29$$

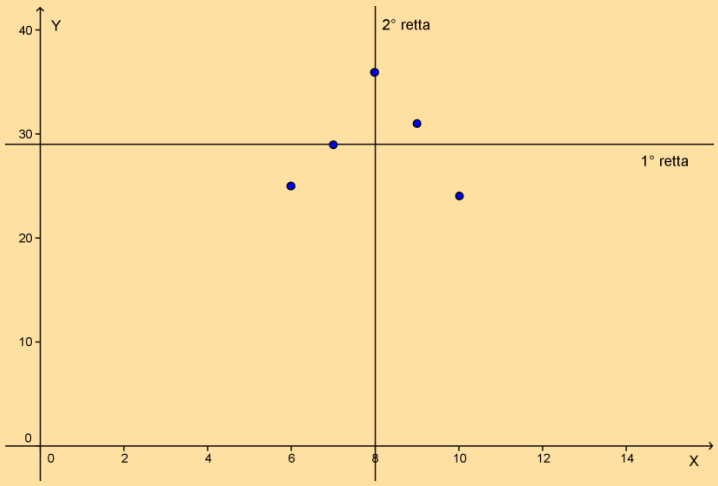
Si calcolano i coefficienti di regressione:

$$b_1 = \frac{5 \cdot 1.160 - 40 \cdot 145}{5 \cdot 330 - 40^2} = 0 \qquad b_2 = \frac{5 \cdot 1.160 - 40 \cdot 145}{5 \cdot 4.299 - 145^2} = 0$$

Le rette di regressione hanno equazione:

$$y = 29 \qquad x = 8$$

Le rette sono parallele agli assi e non esiste dipendenza lineare, ma dal diagramma a dispersione



si rileva che potrebbe esistere una relazione di tipo parabolico. Con il metodo dei minimi quadrati si trova l'equazione della parabola interpolante:

$$y = -121{,}571 + 38{,}857x - 2{,}429x^2$$

detta **parabola di regressione**. L'indice di scostamento è $I_2 = 0{,}052$.

4. CORRELAZIONE LINEARE

La regressione esprime un legame di dipendenza di una variabile da un'altra, ma non sempre si può evidenziare questo legame, in quanto le due variabili possono dipendere entrambe da una terza variabile. Si esamina quindi la **correlazione** tra le due variabili, che esprime la "forza", o "intensità", del loro legame. Per studiare la variabilità di una variabile rispetto al valore medio si utilizza, come visto nel capitolo 1, lo scarto quadratico medio:

$$\sigma_X = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \qquad \sigma_Y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}}$$

Per misurare la variabilità congiunta di due variabili X e Y si introduce la **covarianza di X e Y**, ossia il valore medio del prodotto degli scarti corrispondenti di X e Y:

$$\sigma_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

La covarianza può essere positiva, negativa o nulla, ma non viene assunta come indice poiché il suo valore varia enormemente a seconda delle distribuzioni. La correlazione si misura mediante indici, il più importante e utile dei quali è il **coefficiente di correlazione lineare di Bravais-Pearson**, che esprime il rapporto fra la covarianza di X e Y ed il prodotto degli scarti quadratici medi di X e di Y:

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Quindi:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Se gli scarti sono valori approssimati, per ridurre gli errori di approssimazione si utilizza l'espressione che utilizza i dati grezzi:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

OSSERVAZIONI

- r è un valore senza dimensioni e quindi non dipende dalle unità di misura delle variabili X e Y.
- $-1 \leq r \leq 1$
- Se $r > 0$ la correlazione è **diretta**, o **positiva**, cioè all'aumentare dei valori di una variabile, anche i valori dell'altra in media aumentano.
- Se $r < 0$ la correlazione è **inversa**, o **negativa**, cioè all'aumentare dei valori di una variabile, i valori dell'altra in media diminuiscono.
- Se $r = 1$ la correlazione è **perfetta positiva**.
- Se $r = -1$ la correlazione è **perfetta negativa.**
- Se $r = 0$ non esiste correlazione lineare (potrebbe però esistere una correlazione curvilinea).
- $r = \pm\sqrt{b_1 b_2}$ cioè r è la media geometrica dei due coefficienti di regressione; col segno "+" se i due coefficienti sono positivi, col segno "−" se i due coefficienti sono negativi.
- Formule inverse della precedente:

$$b_1 = r\frac{\sigma_Y}{\sigma_X} \qquad b_2 = r\frac{\sigma_X}{\sigma_Y}$$

- Per una funzione lineare $\delta = r^2$, indica cioè quanto il modello della regressione lineare è aderente al fenomeno in studio; quanto più r^2 è prossimo a 1, tanto maggiore è la "bontà" del modello lineare.

ESEMPIO

Data la seguente tabella:

DITTE	PROFITTO	SPESE
A	25	10
B	30	20
C	15	7
D	42,5	25
E	47,5	30
F	20	13

determinare le rette di regressione e calcolare il coefficiente di correlazione lineare di Bravais-Pearson. Rappresentare graficamente i risultati ottenuti.

Si costruisce la tabella:

	x	y	x^2	xy	y^2
	25	10	625	250	100
	30	20	900	600	400
	15	7	225	105	49
	42,5	25	1806,25	1062,5	625
	47,5	30	2256,25	1425	900
	20	13	400	260	169
totali	180	105	6212,5	3702,5	2243

$$\bar{x} = \frac{180}{6} = 30 \qquad \bar{y} = \frac{105}{6} = 17{,}5$$

Si calcolano i coefficienti di regressione:

$$b_1 = \frac{6 \cdot 3.702{,}5 - 180 \cdot 105}{6 \cdot 6.212{,}5 - 180^2} = 0{,}68 \qquad b_2 = \frac{6 \cdot 3.702{,}5 - 180 \cdot 105}{6 \cdot 2.243 - 105^2} = 1{,}36$$

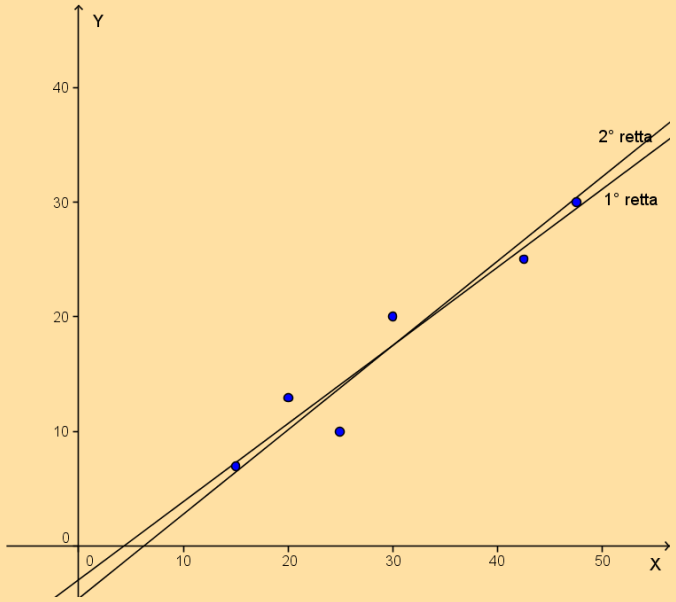
Le rette di regressione hanno equazione:

$$y = -2{,}9 + 0.68x \qquad x = 6{,}16 + 1{,}36y$$

L'indice di correlazione lineare di Bravais-Pearson è:

$$r = \sqrt{0{,}68 \cdot 1{,}36} \cong 0{,}96$$

da cui si deduce che fra profitti e spese vi è buona correlazione diretta. Graficamente:



Inoltre poiché $r^2 = 0{,}9265$, si può dire che il 92,65% della varianza di Y è spiegato dalla dipendenza della Y dalla X, perciò il modello della regressione lineare esprime bene il legame fra le due variabili.