Esercitazioni di statistica

Covarianza, Regressione e Correlazione

Stefania Spina

Universitá di Napoli Federico II

stefania.spina@unina.it

28 Ottobre, 2014

La covarianza

La quantità che rappresenta la media aritmetica del prodotto degli scarti di X e di Y si chiama **covarianza** e questo indice è di solito indicato con $cov(X, Y), \sigma_{xy}, E[(X - \mu_x)(Y - \mu_y)]$, e con s_{xy} nel caso di rilevazioni campionarie.
Essa misura come X ed Y covariano ed è un indice espresso nel **prodotto delle unità di misura** usate per rilevare X ed Y e quindi non può essere utilizzato per stabilire quanto è forte l'eventuale legame lineare esistente fra le due variabili.
E' definita da:

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} (x_i - \mu_x)(y_j - \mu_y)$$

Nel caso di distribuzioni di frequenza:

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} (x_i - \mu_x)(y_j - \mu_y) n_{ij} = \sum_{i=1}^{r} \sum_{j=1}^{c} (x_i - \mu_x)(y_j - \mu_y) f_{ij}$$

ove con μ_x è indicata la media aritmetica della marginale X

$$\mu_x = \frac{1}{N} \sum_{i=1}^{c} x_i ni$$

e con μ_y la media aritmetica della marginale Y

$$\mu_y = \frac{1}{N} \sum_{i=1}^{r} y_j nj$$

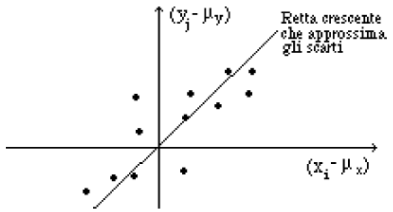
La covarianza

La covarianza è un indice che può teoricamente assumere qualsiasi valore da $-\infty$ *a* $+\infty$.
Più precisamente:

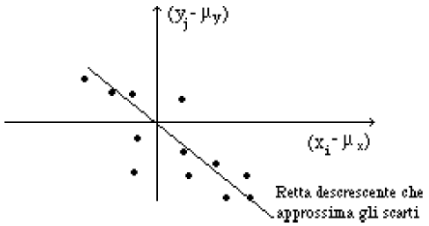
- se è $\sigma_{xy} > 0$ allora fra X ed Y esiste un **legame lineare positivo**;
- se è $\sigma_{xy} < 0$ allora fra X ed Y esiste un **legame lineare negativo**;
- se è $\sigma_{xy} = 0$ allora X ed Y sono **incorrelate** (non esiste legame lineare).

La covarianza

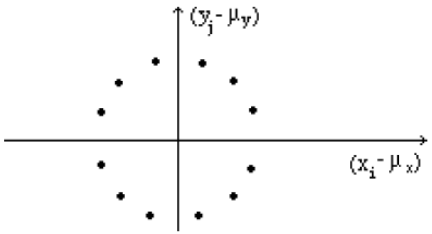
$\sigma_{xy} > 0$



$\sigma_{xy} < 0$



$\sigma_{xy} = 0$



Proprietà della covarianza

Per la covarianza è disponibile una formula calcolatoria, equivalente alla definizione, che permette di non dover calcolare i singoli scostamenti:

$$\sigma_{xy} = M(XY) - \overline{x} * \overline{y} = \frac{1}{N} \sum_{i=1}^{N} x_i y_i - \overline{x} * \overline{y}$$

La correlazione lineare

Quando entrambi i caratteri della distribuzione doppia sono delle variabili quantitative è possibile elaborare un indice capace di misurare l'eventuale **legame lineare** esistente fra X ed Y.
Questo legame, oltre a permettere una semplice ed immediata interpretazione, può rappresentare una prima approssimazione di legami più complessi.
Nella ricerca di un legame lineare esistono **due casi limite** che servono come termine di paragone per poter stabilire il grado del legame lineare esistente fra due variabili:

- il **perfetto legame lineare** quando al crescere della X la Y cresce o decresce esattamente come una retta, questo caso si ha se X = a+ bY con a, b costanti reali e $b \neq 0$;
- **l'incorrelazione** quando al crescere o decrescere della X la Y, in media, rimane costante.

La correlazione lineare

Fra X ed Y esiste un legame lineare se al variare di una delle due variabili l'altra cresce o decresce, in media, secondo una retta.

Se al crescere di X l'altra variabile, in media, cresce come una retta si dice che fra X ed Y esiste un **legame lineare positivo**.

Se al crescere di X l'altra variabile decresce, in media, come una retta si dice che fra X ed Y esiste un **legame lineare negativo**.

Il coefficiente di correlazione lineare

Il coefficiente di correlazione, di solito indicato con ρ_{xy}, $corr(X, Y)$, r_{xy}, è dato da:

$$\rho_{xy} = corr(X, Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{cov(X, Y)}{\sqrt{var(X) var(Y)}}$$

è un indice normalizzato che varia nell'intervallo [-1, 1] e misura, oltre all'esistenza dei legami lineari fra X ed Y, anche la loro intensità.
Più in particolare:

1. più ρ_{xy} assume un valore vicino a - 1 più **il legame lineare è forte e negativo**;
2. più ρ_{xy} assume un valore vicino a 1 più **il legame lineare è forte e positivo**;
3. più ρ_{xy} assume un valore vicino a zero più il **legame lineare è trascurabile**.

Esercizio 1

Il responsabile commerciale di un'azienda paga alcune stazioni radio locali per mandare in onda per una settimana un messaggio pubblicitario relativo all'immissione sul mercato di un nuovo prodotto. Poichè le stazioni richiedono compensi diversi, esiste una variabilità nel numero di messe in onda del messaggio pubblicitario.

Stazioni radio	Messaggi al giorno	Vendite (in milioni)
Fox	4	15
FXZ	2	8
Power	5	21
Lizard	6	24
Rodeo	3	17

Si determini una misura dell'eventuale associazione tra la frequenza dei messaggi pubblicitari e le vendite del prodotto.

Esercizio 1

Stazioni radio	X	Y	XY
Fox	4	15	60
FXZ	2	8	16
Power	5	21	105
Lizard	6	24	144
Rodeo	3	17	51
Totale	*20*	*85*	*376*

$$\rho_{xy} = corr(X, Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

$$\sigma_{xy} = M(XY) - \overline{x} * \overline{y}$$

$$\Downarrow$$

$$\sigma_{xy} = \frac{376}{5} - \left(\frac{20}{5}\right) * \left(\frac{85}{5}\right) = 7.2$$

Esercizio 1

Stazioni radio	X	Y	XY	X^2
Fox	4	15	60	16
FXZ	2	8	16	4
Power	5	21	105	25
Lizard	6	24	144	36
Rodeo	3	17	51	9
Totale	*20*	*85*	*376*	*90*

$$\rho_{xy} = corr(X, Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

$$\sigma_x = \sqrt{M(X^2) - [M(X)]^2}$$

$$\sigma_x = \sqrt{\frac{90}{5} - \left[\frac{20}{5}\right]^2} = 1.4$$

Esercizio 1

Stazioni radio	X	Y	XY	X^2	Y^2
Fox	4	15	60	16	225
FXZ	2	8	16	4	64
Power	5	21	105	25	441
Lizard	6	24	144	36	576
Rodeo	3	17	51	9	289
Totale	*20*	*85*	*376*	*90*	*1595*

$$\rho_{xy} = corr(X, Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

$$\sigma_y = \sqrt{M(Y^2) - [M(Y)]^2}$$

$$\sigma_y = \sqrt{\frac{1595}{5} - \left[\frac{85}{5}\right]^2} = 5.5$$

Esercizio 1

$$\sigma_{xy} = \frac{376}{5} - \left(\frac{20}{5}\right) * \left(\frac{85}{5}\right) = 7.2$$

$$\sigma_x = \sqrt{\frac{90}{5} - \left[\frac{20}{5}\right]^2} = 1.4$$

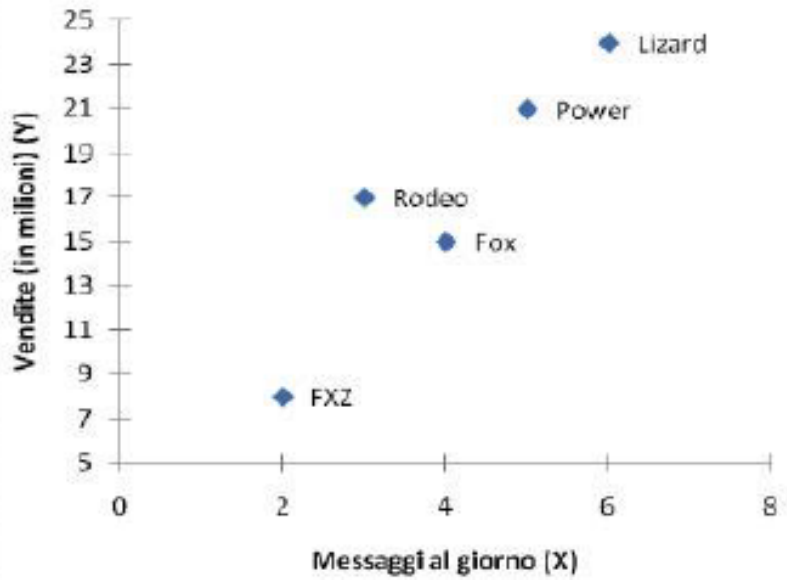
$$\sigma_y = \sqrt{\frac{1595}{5} - \left[\frac{85}{5}\right]^2} = 5.5$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{7.2}{1.4 * 5.5} = 0.9$$

Esercizio 1

Stazioni radio	X	Y	XY	X^2	Y^2
Fox	4	15	60	16	225
FXZ	2	8	16	4	64
Power	5	21	105	25	441
Lizard	6	24	144	36	576
Rodeo	3	17	51	9	289
Totale	*20*	*85*	*376*	*90*	*1595*

$r = 0{,}9$



Esercizio 2

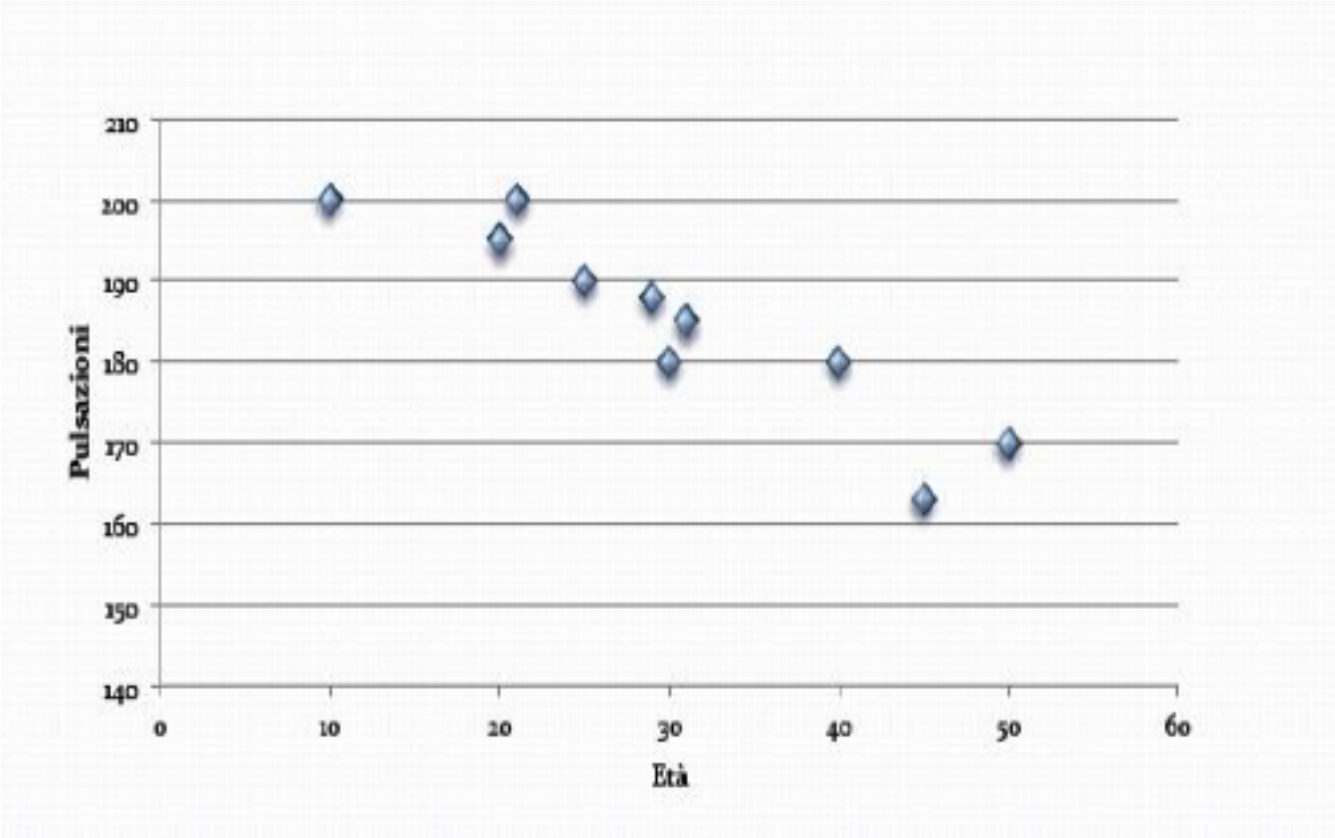
In un esperimento diretto allo studio della relazione tra il numero di pulsazioni sotto sforzo (per minuto) e l'età (in anni) sono stati rilevati i seguenti dati su 10 soggetti di sesso maschile:

Pulsazioni	200	195	200	190	188	180	185	180	163	170
Età	10	20	21	25	29	30	31	40	45	50

a. Si disegni e si commenti il diagramma di dispersione
b. Si calcoli e si commenti l'indice di correlazione
c. Sapendo che pulsazioni = 213.17 - 0.93 * Età calcolare le pulsazioni di una persona di: 9,33, 70 anni.

Esercizio 2

a. Si disegni e si commenti il diagramma di dispersione



Esercizio 2

b. Si calcoli e si commenti l'indice di correlazione

Pulsazioni	Età	Pulsazioni * Età	Pulsazioni^2	Età^2
200	10	2000	40000	100
195	20	3900	38025	400
200	21	4200	40000	441
190	25	4750	36100	625
188	29	5452	35344	841
180	30	5400	32400	900
185	31	5735	34225	961
180	40	7200	32400	1600
163	45	7335	26569	2025
170	50	8500	28900	2500
1851	**301**	**54472**	**343963**	**10393**

$$cov(X, Y) = M(XY) - M(X)M(Y) = \frac{54472}{10} - \frac{1851}{10} * \frac{301}{10} = 5447.2 - 185.1 * 30.1 = 5447.2 - 5571.51 = -124.31$$

$$\sigma_X = \sqrt{M(X^2) - [M(X)]^2} = \sqrt{\frac{343963}{10} - \left(\frac{1851}{10}\right)^2} = \sqrt{34396.3 - 34262.01} = 11.59$$

$$\sigma_y = \sqrt{M(Y^2) - [M(Y)]^2} = \sqrt{\frac{10393}{10} - \left(\frac{301}{10}\right)^2} = \sqrt{1039.3 - 906.01} = 11.54$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{-124.31}{11.59 * 11.54} = \frac{-124.31}{133.75} = -0.93$$

Esercizio 2

c. Sapendo che pulsazioni = 213.17 - 0.93 * Età calcolare le pulsazioni di una persona di: 9,33, 70 anni.

Età = 9 Pulsazioni = 204.8
Età = 33 Pulsazioni = 154.3
Età = 70 Pulsazioni = 147.9

Introduzione

Un modello che mette in relazione una variabile X con un'altra variabile Y, ossia che studia la dipendenza lineare di una variabile di risposta (o dipendente) da una variabile indipendente (regressore, predittore) è

il modello di regressione lineare semplice

tale modello, stabilisce, a meno di variazioni casuali, una **relazione lineare tra risposta e predittore**.

La regressione lineare semplice

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

dove:

- i è il pedice che varia tra le osservazioni i=1, . . . , n;
- Y_i è la **variabile dipendente** o da spiegare;
- X_i è la **variabile indipendente** o il regressore;
- β_0 è **l'intercetta** della retta di regressione della popolazione;
- β_1 è la **pendenza** della retta di regressione della popolazione;
- $\beta_0 + \beta_1 X_i$ è la **componente deterministica**;
- ϵ_i è la **componente erratica o casuale**, cioè le variabili casuali ϵ_i rappresentano l'**errore che si commette nella spiegazione dele v.c.** Y_i **mediante una funzione lineare di** X_i.

La retta di regressione

Determinazione della retta di regressione

L'identificazione della retta avviene attraverso la determinazione dei valori di $\hat{\beta}_0$ *e* $\hat{\beta}_1$, stime dell'intercetta e del coefficiente angolare o pendenza, rispettivamente.

La retta migliore è quella che passa più vicina ai punti osservati.

$$y_i - \hat{y}_i = \textit{minime}$$

La stima dei parametri: il metodo dei minimi quadrati

La retta di regressione è tale che la somma dei residui al quadrato sia minima.
Formalmente:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Il problema consiste dunque nel ricercare $\hat{\beta}_0$ *e* $\hat{\beta}_1$ che minimizzano la precedente espressione.
Da un punto di vista operativo bisogna risolvere il seguente sistema di equazioni (condizioni del primo ordine o stazionarietà).

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

Nota: si tratta di punti di minimo perchè le derivate seconde $\partial \hat{\beta}_0 \hat{\beta}_0 f(\hat{\beta}_0, \hat{\beta}_1) = -2(-n)$, $\partial \hat{\beta}_1 \hat{\beta}_1 f(\hat{\beta}_0, \hat{\beta}_1) = -2 \sum_i^n (-x_i^2)$ sono sempre non negative.

Stimatori dei parametri della retta di regressione: $(\hat{\beta}_0)$

$$-2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} y_i - n * \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\Downarrow$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Stimatori dei parametri della retta di regressione: $(\hat{\beta}_1)$

$$-2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \left(\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} \right)$$

$$\hat{\beta}_1 \left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \right) = n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

La Valutazione dell'adattamento

Una misura della bontà dell'adattamento della retta di regressione ai dati può essere data dal rapporto tra la devianza spiegata e la devianza totale.

$$R^2 \rightarrow \textit{indice di determinazione}$$

$$R^2 = \frac{Dev(\hat{y})}{Dev(y)} = \frac{\sum_i (\hat{y} - \overline{y})^2}{\sum_i (y - \overline{y})^2}$$

$$0 \leq R^2 \leq 1$$

La Valutazione dell'adattamento

- Quando è $\mathbf{R^2 = 0}$, la **devianza spiegata è pari a zero**. Questo vuol dire che l'osservazione della variabile X non ha aggiunto nulla a quanto già si sapeva dalla sola osservazione della Y. Dal punto di vista geometrico, la retta di regressione coincide con la retta M(Y); dal punto di vista interpretativo, le **variabili X e Y sono incorrelate**;
- Quando è $\mathbf{R^2 = 1}$, la **devianza spiegata è uguale alla devianza totale**. Questo vuol dire che l'osservazione della variabile X spiega perfettamente la variabile Y, e ne rende possibile la previsione senza possibilità di errore. Dal punto di vista geometrico, tutti i punti sono allineati e la retta di regressione passa per tutti i punti (siamo quindi nel caso di una dipendenza funzionale, deterministica, esatta); dal punto di vista interpretativo, **le variabili X e Y sono massimamente correlate**;
- Quando è $0 \leq R^2 \leq 1$, **la devianza spiegata è pari a una quota della devianza totale**. L'osservazione della variabile X migliora quindi la previsione della variabile Y, con una quota di errore residua dovuta in parte alle variabili non osservate, in parte alla sempre presente quota di imponderabilità dei fenomeni osservati.

La regressione lineare semplice: Esercizio 1

Il responsabile del marketing di un'impresa vuole stabilire l'effetto delle *Spese pubblicitarie* (in centinaia di euro) sul rispettivo *Fatturato* (in migliaia di euro). Si estrae un campione di 5 unità locali dell'impresa e si ottengono i seguenti risultati:

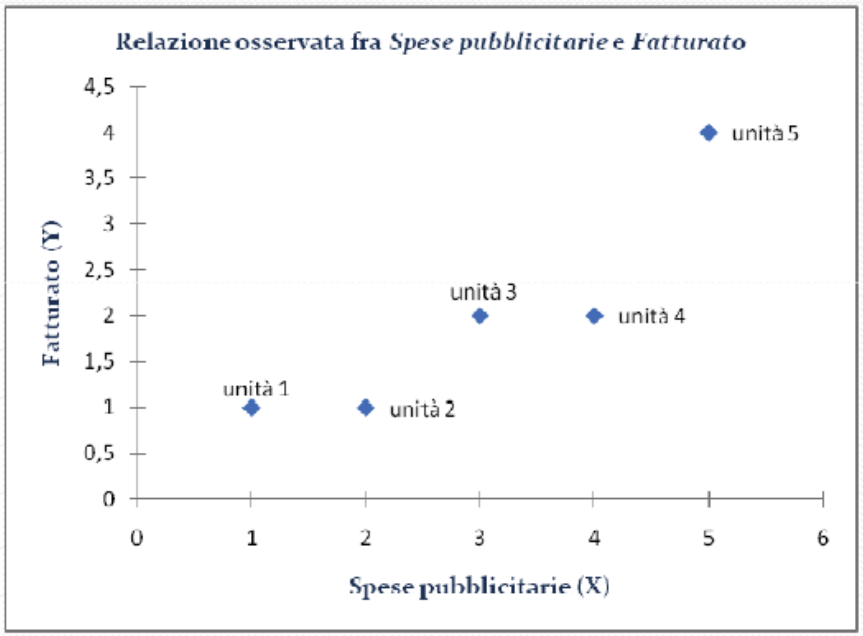
	Spese pubblicitarie (x 100 euro)	Fatturato (x 1000 euro)
Unità 1	1	1
Unità 2	2	1
Unità 3	3	2
Unità 4	4	2
Unità 5	5	4

a. Rappresentare mediante grafico a dispersione i valori osservati
b. Determinare i coefficienti della retta di regressione $Y_i = \beta_0 + \beta_1 X_i$ che esprima la dipendenza del *Fatturato (Y)* dalle *Spese pubblicitarie (X)*
c. Valutare la bontà di adattamento della retta ai dati.

Esercizio 1

a. Rappresentare mediante grafico a dispersione i valori osservati

	Spese Pub.	Fatturato
Unità 1	1	1
Unità 2	2	1
Unità 3	3	2
Unità 4	4	2
Unità 5	5	4



Esercizio 1

b. Determinare i coefficienti della retta di regressione Y= a+bX che esprima la dipendenza del *Fatturato (Y)* dalle *Spese pubblicitarie (X)*

	Spese pubblicitarie (x 100 euro)	Fatturato (x 1000 euro)
Unità 1	1	1
Unità 2	2	1
Unità 3	3	2
Unità 4	4	2
Unità 5	5	4

Parametri dell'interpolante lineare ricavati con il metodo dei **Minimi Quadrati**.

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{Cov(XY)}{Var(X)}$$

Esercizio 1

	X	Y	X-M(X)	Y-M(Y)	(X-M(X))*(Y-M(Y))	(X-M(X))²
Unità 1	1	1	-2	-1	2	4
Unità 2	2	1	-1	-1	1	1
Unità 3	3	2	0	0	0	0
Unità 4	4	2	1	0	0	1
Unità 5	5	4	2	2	4	4
Totale	15	10	0	0	7	10

$$\overline{x} = \frac{15}{3} = 3$$

$$\overline{y} = \frac{10}{5} = 2$$

$$Cov(XY) = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{N} = \frac{7}{5} = 1.4$$

$$Var(X) = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N} = \frac{10}{5} = 2$$

$$\hat{\beta}_1 = \frac{cov(XY)}{var(X)} = \frac{1.4}{2} = 0.70$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 2 - 0.70 * 3 = -0.10$$

$$\hat{y} = -0.10 + 0.70x$$

Esercizio 1

	X	Y	$\hat{Y}$	$(\hat{y}-\bar{y})^2$	$(y-\bar{y})^2$
Unità 1	1	1	0,60	2,0	1
Unità 2	2	1	1,30	0,5	1
Unità 3	3	2	2,00	0,0	0
Unità 4	4	2	2,70	0,5	0
Unità 5	5	4	3,40	2,0	4
Totale	15	10		4,9	6

$$R^2 = \frac{Dev(\hat{Y})}{Dev(Y)} = \frac{\sum_i (\hat{y}-\bar{y})^2}{\sum_i (y-\bar{y})^2}$$

$$\hat{Y}_1 = -0{,}10 + 0{,}70 * (1) = 0{,}60$$

$$\hat{Y}_2 = -0{,}10 + 0{,}70 * (2) = 1{,}30$$

$$R^2 = \frac{4{,}9}{6} = 0{,}81$$

Esercizio 2

Recenti dati estratti dal Statistical Abstract of the United States per i 50 stati e il District of Columbia relativi a Tasso di criminalità (Y) e Tasso di povertà (X), hanno permesso di ottenere la retta di regressione Y = 209,9 + 25,5 X. Il Tasso di criminalità corrisponde al numero di omicidi per 100000 abitanti, mentre il Tasso di povertà corrisponde alla percentuale di popolazione con reddito sotto il livello di povertà.

a. Interpretare l'intercetta e il coefficiente angolare.
b. Determinare il Tasso di criminalità previsto e il relativo residuo per il Massachusetts (dove x = 10,7 e y = 805). Interpretare i risultati.
c. Determinare il segno della correlazione fra queste variabili.

Esercizio 2

a) Interpretare l'intercetta e il coefficiente angolare della retta $Y = 209{,}9 + 25{,}5\,X$.

Parametri della interpolante lineare ricavati con il metodo dei **Minimi Quadrati** (MQ).

$$a = \bar{y} - b\bar{x} = 209{,}9$$

$$b = \frac{\mathrm{cov}(XY)}{Var(X)} = 25{,}5$$

Può essere interpretato come il valore di Y per X=0 (quando ciò ha senso).

Esprime quanto varia la variabile Y al variare unitario della variabile X.

209,9 = tasso di criminalità per uno stato con un tasso di povertà = 0

25,5 = aumento nel tasso di criminalità previsto per un aumento di 1 nella percentuale di popolazione con reddito al di sotto della soglia di povertà.

Esercizio 2

b) Determinare il *Tasso di criminalità previsto* e il relativo *residuo* per il Massachusetts (dove x = 10,7 e y = 805). Interpretare i risultati.

$$\mathbf{Y = 209{,}9 + 25{,}5\ X}$$

Massachusetts (x = 10,7 e y = 805)

$$\hat{Y} = 209{,}9 + 25{,}5 * (10{,}7) = 482{,}8$$

$$\hat{e} = y - \hat{y} = 805 - 482{,}8 = 322{,}2$$

Il tasso di criminalità è molto più alto di quanto previsto per questo tasso di povertà.

Esercizio 2

c) Determinare il segno della correlazione fra queste variabili.

$$Y = 209{,}9 + 25{,}5\,X$$

Positivo perché ha lo stesso segno del coefficiente angolare.