Statistica
A.A. 2019/2020
CdL Scienze Economiche

Prof. Massimiliano Ferrara
Dott. Bruno Antonio Pansera

Lezione n.3

Correlazione e regressione: Introduzione

Dall'analisi ed inferenza riguardante una singola variabile statistica passiamo alla relazione tra (due) variabili statistiche.

Le relazioni tra variabili importanti nell'analisi della realtà economico-aziendale possono essere matematicamente espresse come:

$$y=f(X)$$

dove la funzione f può assumere varie forme, lineari o non lineari, e può non essere conosciuta in modo preciso.

Correlazione e regressione: Introduzione

In molte situazioni interessa studiare se esiste una relazione tra due variabili misurate sulle stesse unità.

- “Le misurazioni del peso prima della terapia sono in relazione con le misurazioni dopo la terapia?”
- “il voto di maturità `è in relazione con la performance universitaria?”

Oppure si desidera prevedere il valore di una variabile conoscendo il valore di un’altra.

- “Conoscendo l’altezza del padre, è possibile prevedere l’altezza di un figlio?”
- “Conoscendo la durata della gravidanza, si può stimare il peso alla nascita?”

Correlazione e regressione: Introduzione

La statistica permette di rispondere a questo tipo di domande, con strumenti adatti alla natura delle variabili in esame. A tale scopo, per variabili quantitative, si tratteranno:

- La CORRELAZIONE, che misura la dipendenza lineare tra due variabili;
- La REGRESSIONE, che valuta la relazione lineare tra due variabili.

Correlazione

In statistica, una correlazione è una relazione tra due variabili tale che a ciascun valore della prima corrisponda un valore della seconda, seguendo una certa regolarità.

La correlazione, quindi, indica la tendenza che hanno due variabili (X e Y) a variare insieme, ovvero, a covariare. Quando si parla di correlazione bisogna prendere in considerazione due aspetti:

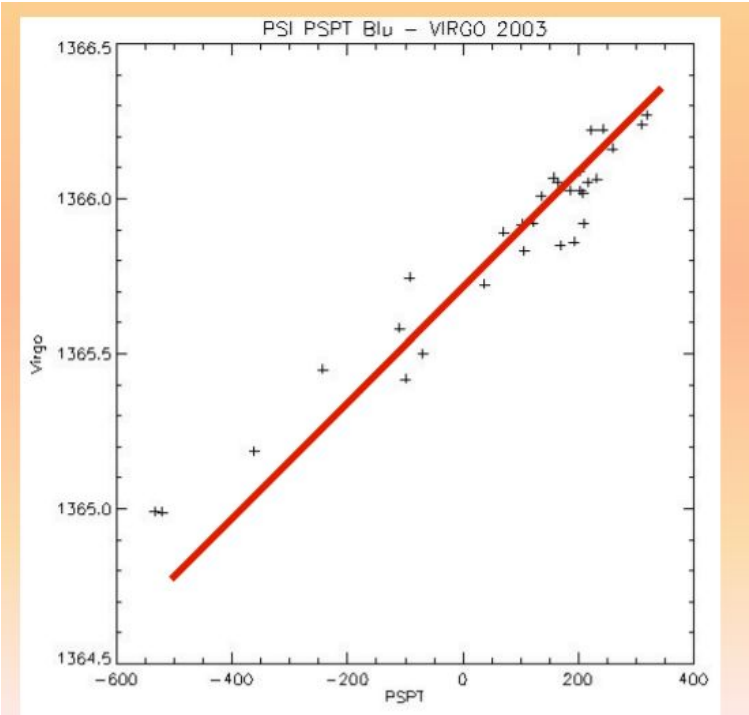
- il tipo di relazione esistente tra due variabili
- la forma della relazione.

La relazione può essere valutata tramite:

- Un grafico (grafico di dispersione)
- Un indice che quantifica il grado di correlazione (coefficiente di correlazione)

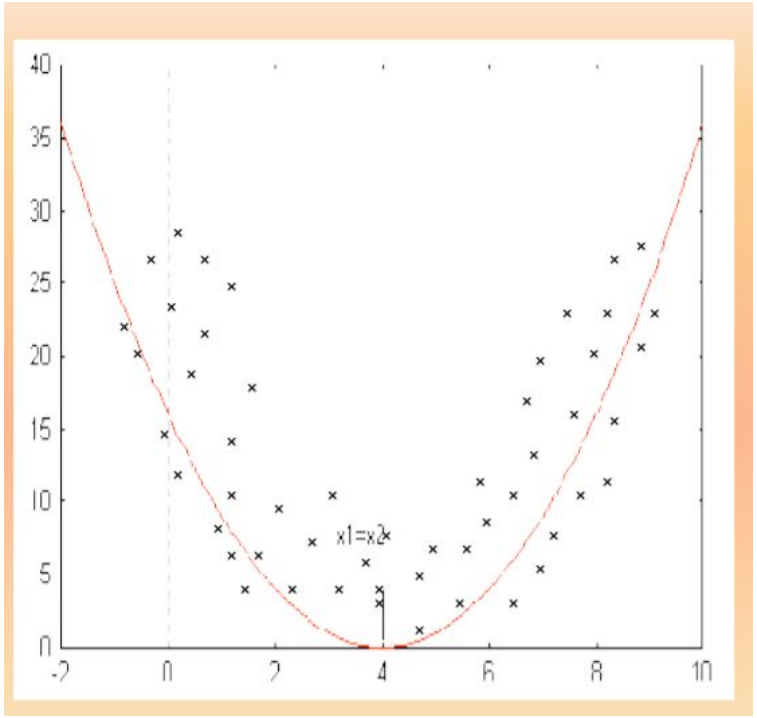
Correlazione Lineare

La relazione è di tipo lineare se, rappresentata su assi cartesiane, si avvicina alla forma di una retta.



Correlazione non Lineare

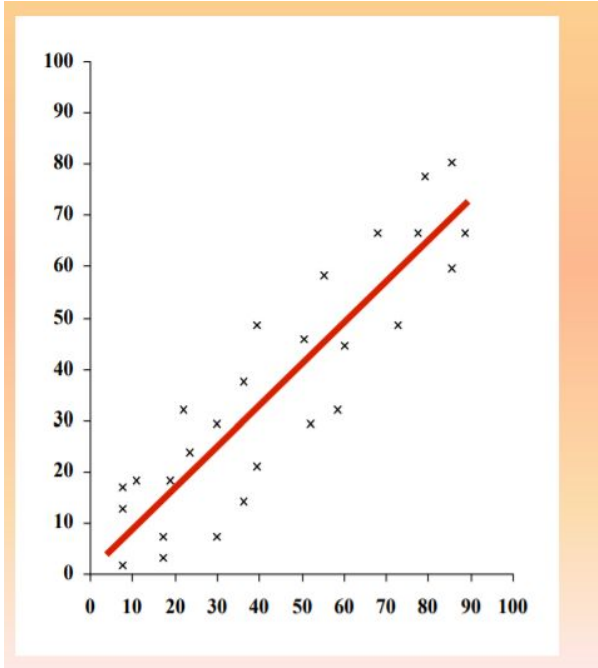
La relazione è di tipo non lineare, se rappresentata su assi cartesiane, ha un andamento curvilineo (parabola o iperbole).



Forma della relazione

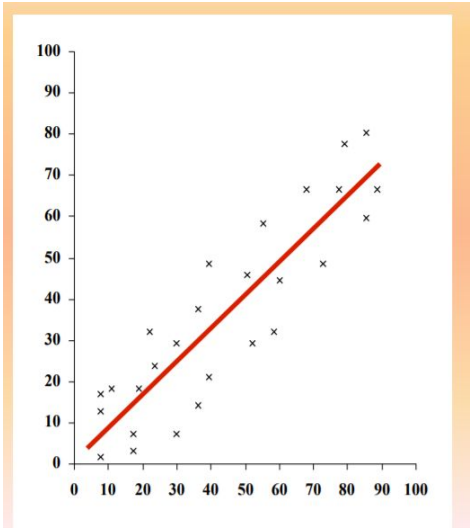
Per quanto riguarda la forma della relazione, si distinguono l'entità e la direzione

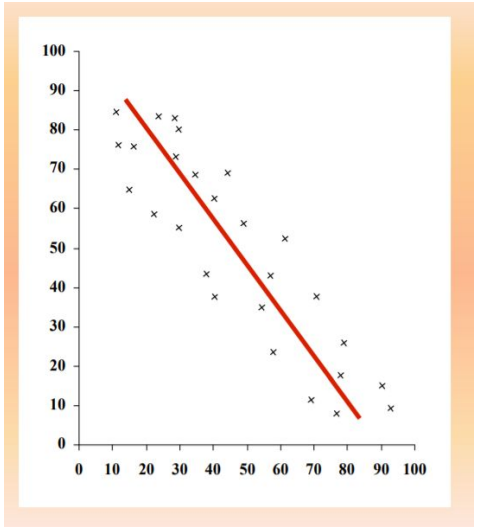
- L'entità si riferisce alla forza della relazione esistente tra due variabili. Quanto più i punteggi sono raggruppati attorno ad una retta, tanto più forte è la relazione tra due variabili.



Forma della relazione

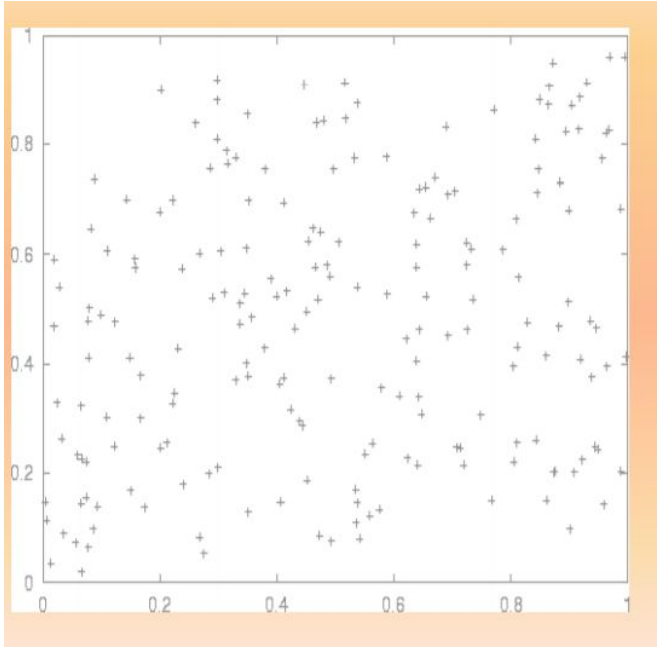
- La direzione può essere: positiva, se all'aumentare di una variabile aumenta anche l'altra.
- La direzione è negativa se all'aumentare di una variabile diminuisce l'altra.





Forma della relazione

- Se i punteggi sono dispersi in maniera uniforme, invece, tra le due variabili non esiste alcuna relazione.



Media e Varianza

La media aritmetica e la varianza di X sono

$$m_x = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \,,$$

e

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m_x^2 \,.$$

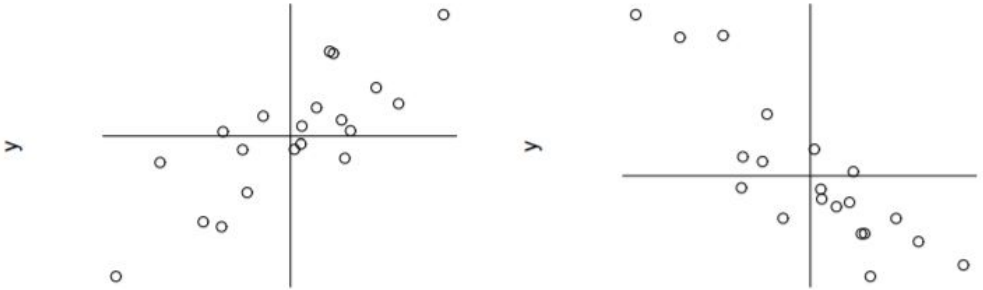
Analogamente, si indicano con m_y e S_y^2 media e varianza di Y.

Covarianza

Per avere una valutazione analitica del grado di associazione tra due variabili quantitative, esiste un indice che misura la dispersione nel piano dei punti dal proprio centro: la COVARIANZA.

Il nome lascia intuire che si tratta di un'estensione al caso di due variabili della varianza. La covarianza si basa infatti sugli scarti delle x_i dalla propria media, $(x_i - m_x)$, e delle y_i dalla propria media, $(y_i - m_y)$.

La covarianza, a differenza della varianza che è sempre positiva, misura l'eventuale direzione del legame, ovvero se le due variabili si muovono nella stessa direzione o in direzioni opposte. Il segno della covarianza riflette il senso crescente o decrescente dell'allineamento tendenziale.



Covarianza

La covarianza segnala una concordanza (sia che X e Y decrescono o crescono) con un segno $+$ e una discordanza (quando X cresce e Y decresce, o viceversa) con il segno -. Formalmente, l'indice è

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y) \ .$$

Una formula alternativa per il calcolo della covarianza è

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_x m_y$$

Si noti che $S_{xx} = S_x^2$, ossia la covarianza tra X e X coincide con la varianza di X.

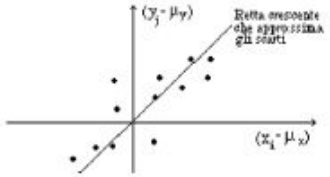
Covarianza

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

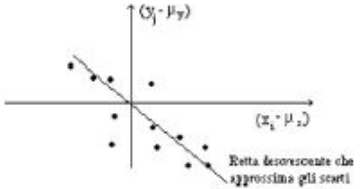
- se è $\sigma_{xy} > 0$ allora fra X ed Y esiste un **legame lineare positivo**;
- se è $\sigma_{xy} < 0$ allora fra X ed Y esiste un **legame lineare negativo**;
- se è $\sigma_{xy} = 0$ allora X ed Y sono **incorrelate** (non esiste legame lineare).

Covarianza

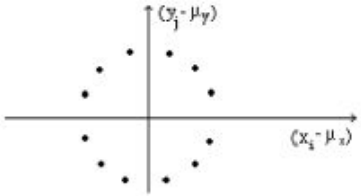
$\sigma_{xy} > 0$



$\sigma_{xy} < 0$



$\sigma_{xy} = 0$



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 diverso da 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 $\rho_{xy}, corr(X, Y), r_{xy}$, 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 n. 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 n. 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 n. 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 n. 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 n. 1

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

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

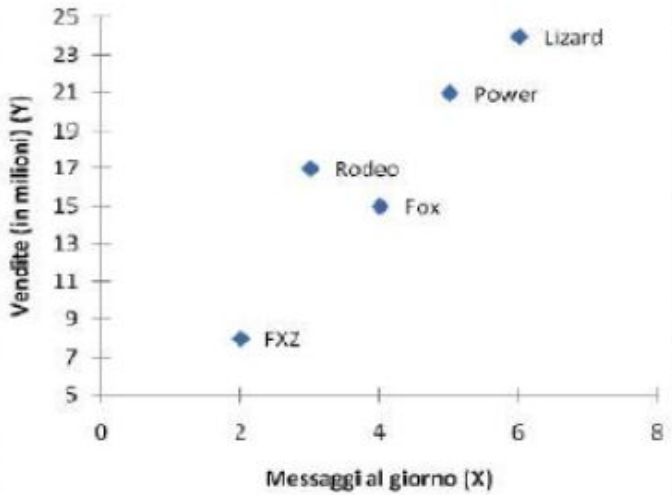
$$\sigma_y = \sqrt{\frac{1595}{5} - \left[\frac{85}{5}\right]^2} = 5.5$$

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

Esercizio n. 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*

$r = 0{,}9$



La Regressione Lineare Semplice: Esempi

- Il presidente di una ditta di materiali da costruzione ritiene che la Quantità media annua di piastrelle (Q) venduta sia una funzione (lineare) del Valore complessivo dei permessi edilizi rilasciati (V) nell'anno passato: Q=f(V) .
- Un grossista di cereali vuole conoscere l'effetto della produzione produzione annua Complessiva (Complessiva) sul prezzo di vendita a tonnellata (P): Q=f(P).
- L'area marketing di un'azienda ha necessità di sapere come il prezzo della Benzina influenzi la quantità venduta: ricorrendo alla serie storica dei prezzi settimanali e dei dati di vendita intendono sviluppare un modello (lineare) che indichi di quanto variano le vendite al variare del prezzo: Q=f(P).

La Regressione Lineare Semplice

Obiettivo:

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

La Regressione Lineare Semplice

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

- Quando dall'analisi di un diagramma di dispersione emerge un particolare andamento della nuvola di punti di X e Y , è naturale chiedersi se esiste una qualche relazione statistica del tipo

$$Y = f(X) + \text{errore tra X e Y}$$

- Il problema è lo stesso di prima: si vuole studiare una relazione tra le variabili. La relazione non è pi`u simmetrica!! Perché si vuole comprendere come la variabile risposta Y sia influenzata dalla variabile esplicativa X.
- Se la relazione che emerge `e di tipo lineare, si pu`o esprimere la relazione statistica tra X e Y usando un modello molto semplice: **l'equazione della retta**

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.

Determinazione di una 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.

La stima dei parametri: il metodo dei minimi quadrati

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

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

La Valutazione dell'adattamento

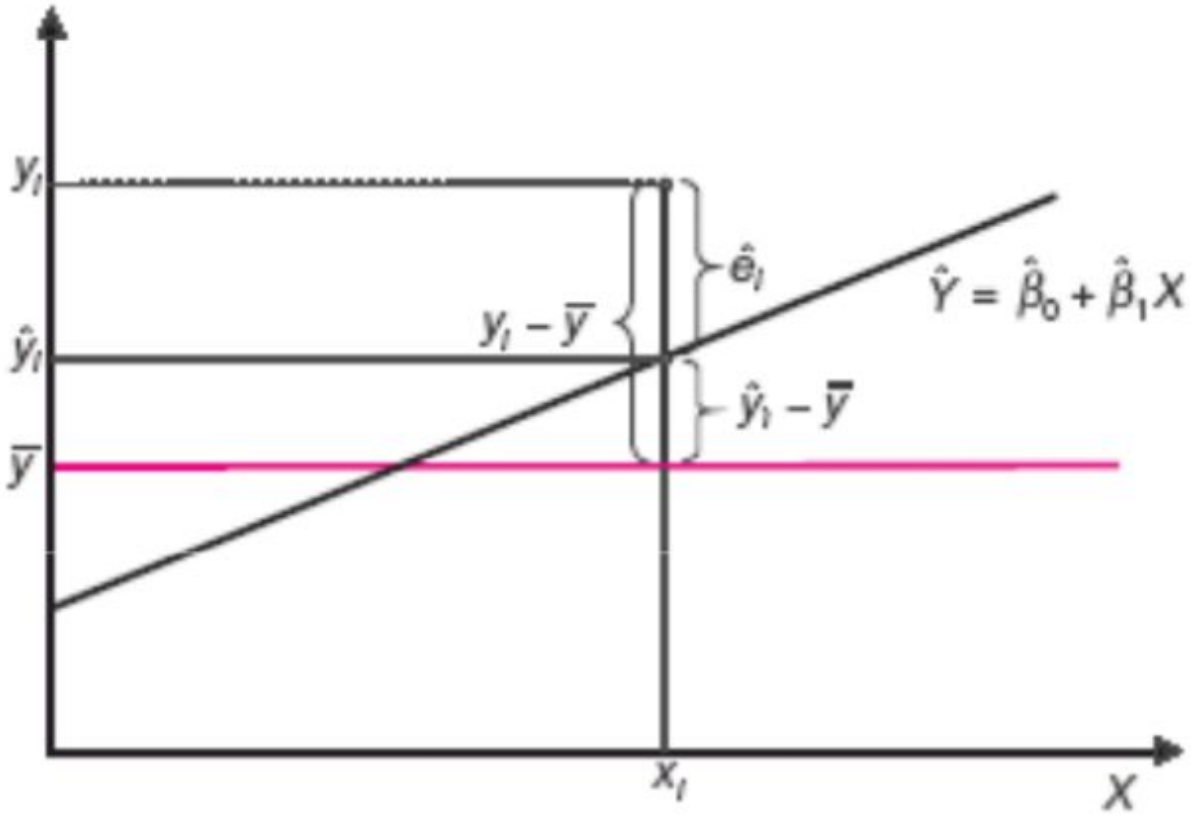
Una misura della bontà dell'adattamento della retta di regressione ai dati può essere data dal rapporto tra la devianza spiegata e la devianza totale.

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

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

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

La Valutazione dell'adattamento



La Valutazione dell'adattamento

- **$R^2 = 0$**, la devianza spiegata è pari a zero, ovvero l'osservazione della variabile X non ha aggiunto nulla a quanto già si sapeva dalla sola osservazione della Y. Dal punto di vista interpretativo, **le variabili X e Y sono incorrelate;**
- **$R^2 = 1$**, la devianza spiegata è uguale alla devianza totale, ovvero 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**;
- **$0 < R^2 < 1$**, 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.

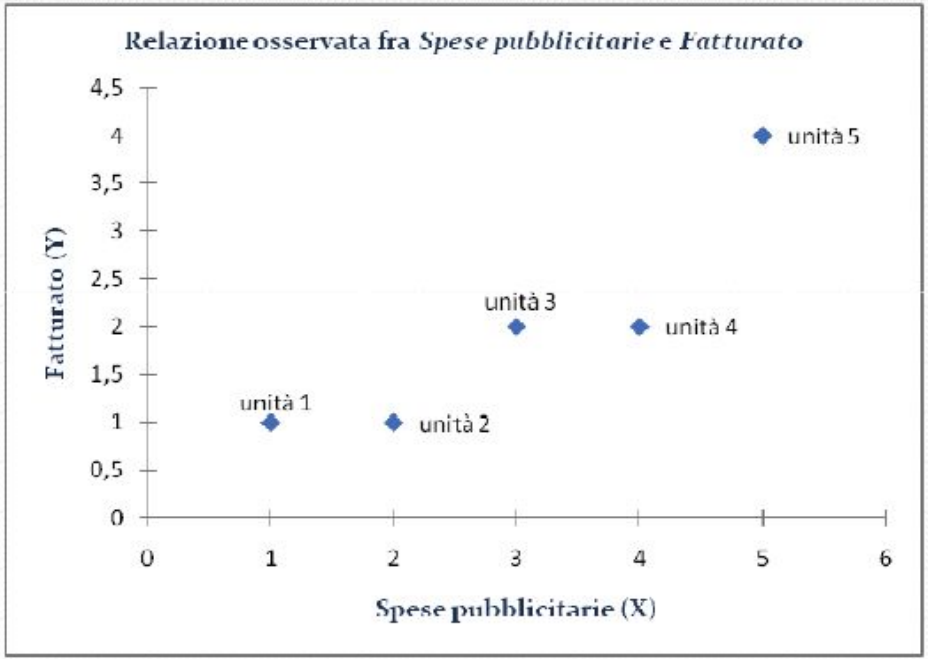
Esercizio

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.

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



Esercizio

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**.

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

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

	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

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

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

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

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

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

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

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

	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

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

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

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

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