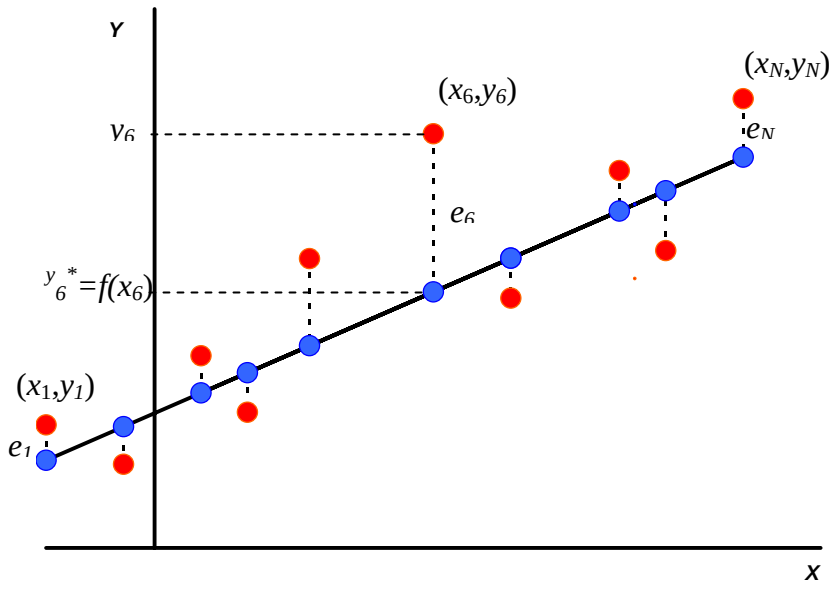
FACOLTA' DI INGEGNERIA

CORSO DI LAUREA IN INGEGNERIA PER L'AMBIENTE ED IL TERRITORIO

CORSO DI STATISTICA E CALCOLO DELLE PROBABILITA'

PROF. PASQUALE VERSACE



SCHEDA DIDATTICA N°6

ARGOMENTO:

REGRESSIONE E CORRELAZIONE

A.A. 2008-09

REGRESSIONE E CORRELAZIONE

Molto spesso nella pratica si rileva che tra due (o più) variabili rappresentanti fenomeni del mondo reale possa esistere una qualche relazione. Per esempio il peso dei maschi adulti dipende in qualche grado dalla loro altezza; la circonferenza dei cerchi dipende dal loro raggio.

La *regressione* è la tecnica per individuare un'equazione che descriva in termini matematici il legame fra le variabili, ed in particolare tra una o più variabili indipendenti ed una variabile dipendente a partire da un campione di osservazioni. Con il termine *correlazione* si indica, invece, il grado di relazione esistente tra le variabili e per mezzo di essa si cerca di determinare quanto bene una certa equazione descriva o spieghi tale relazione.

Quando si considerano solo due variabili si parla di *regressione e correlazione semplice*, quando invece si considerano più di due variabili si parla di *regressione e correlazione multipla*. Nel seguito verrà considerato solo il caso della regressione e della correlazione semplice indicando con Y la variabile dipendente e con X la variabile indipendente.

REGRESSIONE

Il primo passo nella determinazione di un legame funzionale $Y=f(X)$ che leghi le variabili in esame è la raccolta di dati che mostrino valori corrispondenti delle variabili considerate.

Ad esempio se X ed Y indicano rispettivamente l'altezza ed il peso di persone adulte, considerando un campione di N individui è possibile, per ciascuno di essi, misurare la coppia di valori (x_i,y_i), con $i = 1,\ldots N$, che ne rappresentano l'altezza ed il peso rispettivamente.

Il passo seguente consiste nel riportare i punti determinati dalle coppie di valori (x_1, y_1), (x_2, y_2), ..., (x_N, y_N) su un sistema di coordinate cartesiane ottenendo il cosiddetto *diagramma di dispersione* (fig.1).

Nei fenomeni reali è altamente improbabile che i dati si dispongano perfettamente lungo una curva seguendo una relazione esatta, è più corretto quindi considerare l'espressione

$$Y = f(X) + e \qquad (1)$$

in cui *e* rappresenta un termine di errore (o disturbo) che caratterizza le differenze tra i valori di *Y* osservati e quelli che si ottengono invece dalla relazione funzionale con la *X*.
In questo modo *Y* è una variabile casuale risultante dalla somma di una componente deterministica, *f*(*X*), e di una componente stocastica *e*.
Obiettivo dell'analisi di regressione è quello di esplicitare la forma funzionale della componente deterministica individuando la curva che "**meglio**" interpola la relazione ipotizzata tra *X* ed *Y* secondo criteri specificati nel paragrafo successivo.
L'analisi di regressione consente quindi di:

1) visualizzare la relazione tra due variabili;
2) effettuare operazioni di interpolazione ed estrapolazione per ricavare il valore della variabile dipendente in casi non osservati per un fissato valore della variabile indipendente.

Spesso è possibile individuare il tipo di legame in grado di approssimare i dati a partire dall'osservazione del diagramma di dispersione: se i dati sono bene interpolati da una retta si dice che esiste una *relazione lineare* altrimenti è una relazione *non-lineare*. Nel seguito considereremo solo il caso di relazione lineare.

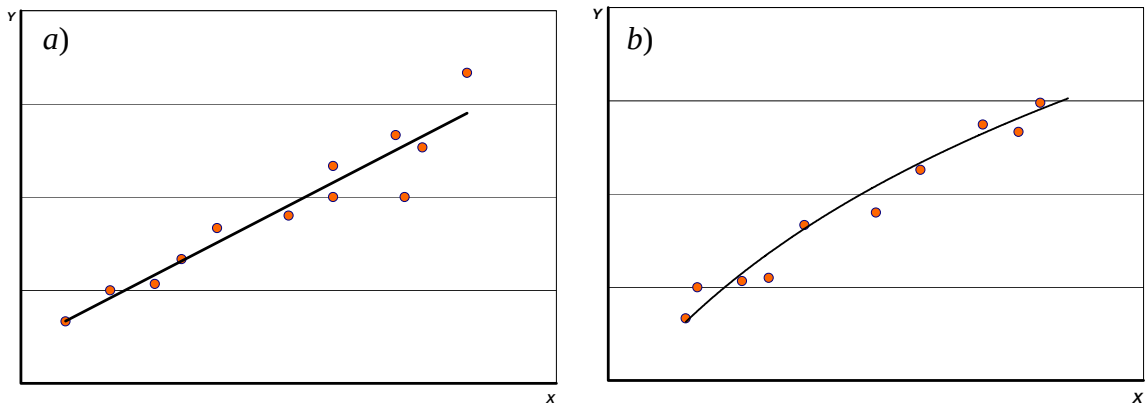


Figura 1 - Diagrammi a dispersione e curve interpolanti. *a*) legame lineare; *b*) legame non-lineare

LA RETTA DI REGRESSIONE CON IL METODO DEI MINIMI QUADRATI

Il tipo più semplice di curva interpolante è la retta. In questo caso il legame funzionale tra X ed Y (trascurando l'errore e) ha l'espressione generale:

$$Y = a_0 + a_1 X \qquad (2)$$

in cui a_0 ed a_1 che costituiscono rispettivamente l'intercetta ed il coefficiente angolare della retta di regressione.

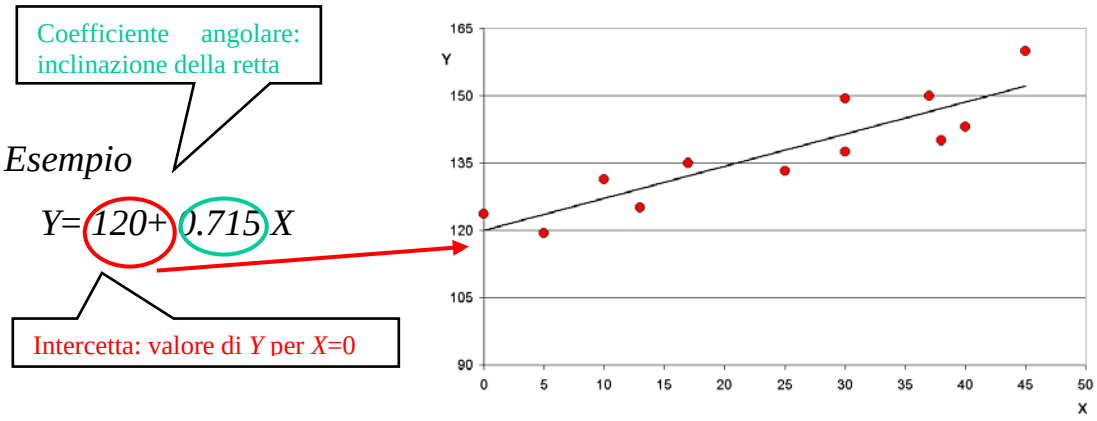


Figura 2 – Retta di regressione

Per determinare i valori dei parametri a_0 ed a_1 è necessario definire il criterio di "**migliore interpolante**" di un determinato insieme di osservazioni.

Siano le coppie (x_i, y_i) con i=1,…,N, l'insieme dei dati osservati dei quali vogliamo descrivere il comportamento mediante una funzione lineare.

Nella figura 3 per un dato valore x_i, sono stati evidenziati gli errori e_i , cioè le differenze tra il valore y_i osservato ed il corrispondente valore y_i^* determinato invece sulla retta di equazione:

$$y_i^* = a_0 + a_1 x_i \qquad (3)$$

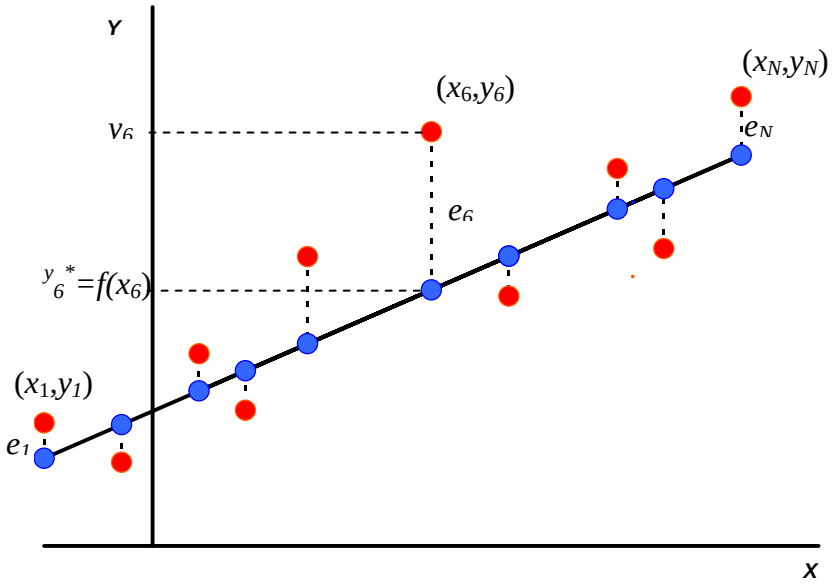


Figura 3 – Rappresentazione grafica del modello di regressione lineare

Una misura della bontà dell'interpolazione effettuata per mezzo della retta è fornita proprio dalla somma dei quadrati degli errori, $e_1^2 + e_2^2 + \ldots + e_N^2$: quanto più tale somma è piccola, tanto più l'interpolazione è buona.

Tra tutte le rette interpolanti un dato insieme di punti quella avente la proprietà di minimizzare la somma S dei quadrati degli errori

$$S = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - y_i^*\right)^2 \qquad (4)$$

viene definita come *migliore interpolante* o retta dei *minimi quadrati*.

I parametri a_0 ed a_1 della retta dei minimi quadrati devono essere tali che:

$$S = \sum_{i=1}^{N} \left(y_i - y_i^*\right)^2 = \sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_i\right)\right]^2 = \min \qquad (5)$$

S è minimo quando le derivate parziali di S rispetto ad a_0 ed a_1 valgono zero.

Allora imponendo le condizioni:

$$\frac{\partial S}{\partial a_0} = -2\sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_i\right) = 0 \qquad (6)$$

$$\frac{\partial S}{\partial a_1} = -2\sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_i\right) x_i = 0 \qquad (7)$$

e semplificando si ottiene un sistema di due equazioni lineari nelle due incognite a_0 ed a_1. Le equazioni precedenti semplificate sono usualmente indicate come *equazioni normali* della retta dei minimi quadrati:

$$a_0 N + a_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \tag{8}$$

$$a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i \tag{9}$$

Le due equazioni precedenti costituiscono un sistema, lineare che può essere rappresentato in forma matriciale come **XA** = **Y** , dove:

A=$\begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$:vettore con i parametri della retta che rappresentano le incognite;

X=$\begin{pmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{pmatrix}$: matrice dei coefficienti

Y=$\begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{pmatrix}$: vettore dei termini noti

Uno dei modi possibili di risolvere il problema è il seguente:

$$a_0 = \frac{\begin{vmatrix} \sum_{i=1}^{N} y_i & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} x_i^2 \end{vmatrix}}{\begin{vmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{vmatrix}} \qquad \text{e} \qquad a_1 = \frac{\begin{vmatrix} N & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i y_i \end{vmatrix}}{\begin{vmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{vmatrix}}$$

Risolvendo si ha:

$$a_0 = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{10}$$

$$a_1 = \frac{N\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{11}$$

La retta dei minimi quadrati è unica e passa attraverso il punto determinato dalla coppia di valori ($\bar{x}, \bar{y}$) costituita dalle medie delle osservazioni x_i ed y_i.

Inoltre, la retta dei minimi quadrati è tale per cui $\sum_{i=1}^{N} e_i = 0$ (tale proprietà discende dall'equazione (6)).

Le equazioni normali, infine, possono essere riscritte in modo da determinare i parametri in maniera operativamente più semplice:

$$a_1 = \frac{\sum_i \left(x_i - \bar{x}\right) \cdot \left(y_i - \bar{y}\right)}{\sum \left(x_i - \bar{x}\right)^2} \tag{12}$$

$$a_0 = \bar{y} - a_1 \bar{x} \tag{13}$$

Le precedenti equazioni (10-13) sono state derivate per il caso in cui *X* è la variabile indipendente ed *Y* è la variabile dipendente. In maniera del tutto analoga nel caso in cui *X* sia considerata variabile dipendente si possono ottenere le relazioni valide per la stima dei parametri della retta di regressione di *X* su *Y*

$$X = b_0 + b_1 Y \tag{14}$$

In tal caso vengono considerate le deviazioni orizzontali invece che verticali.

CORRELAZIONE LINEARE

Consideriamo due variabili *X* ed *Y* le cui osservazioni riportate in una diagramma a dispersione sembrano disporsi intorno ad una retta: in tali casi si è visto come per gli scopi della regressione sia appropriato considerare una relazione lineare; in questo caso si parla, quindi, di *correlazione lineare*.

E' possibile determinare in modo *qualitativo* la bontà dell'accostamento della retta di regressione per mezzo dell'osservazione diretta del diagramma a dispersione.

Se *Y* tende a crescere al crescere di *X* la correlazione è detta *positiva* (fig. 4*a*) , se *Y* tende a decrescere al crescere di *X* la correlazione è detta *negativa* o *inversa* (fig. 4*b*).

Se tutti i punti del diagramma si dispongono proprio su una retta vuol dire che le variabili soddisfano esattamente un'equazione diciamo che le variabili sono *perfettamente correlate* ovvero che tra loro esiste una correlazione perfetta (fig. 4*c*): ad esempio la circonferenza *C* ed il raggio *r* di ogni cerchio sono perfettamente correlati dal momento che $C=2\pi r$.

Se invece non c'è alcuna relazione tra le variabili si dice che non c'è correlazione tra esse cioè sono *incorrelate* (fig. 4*d*).

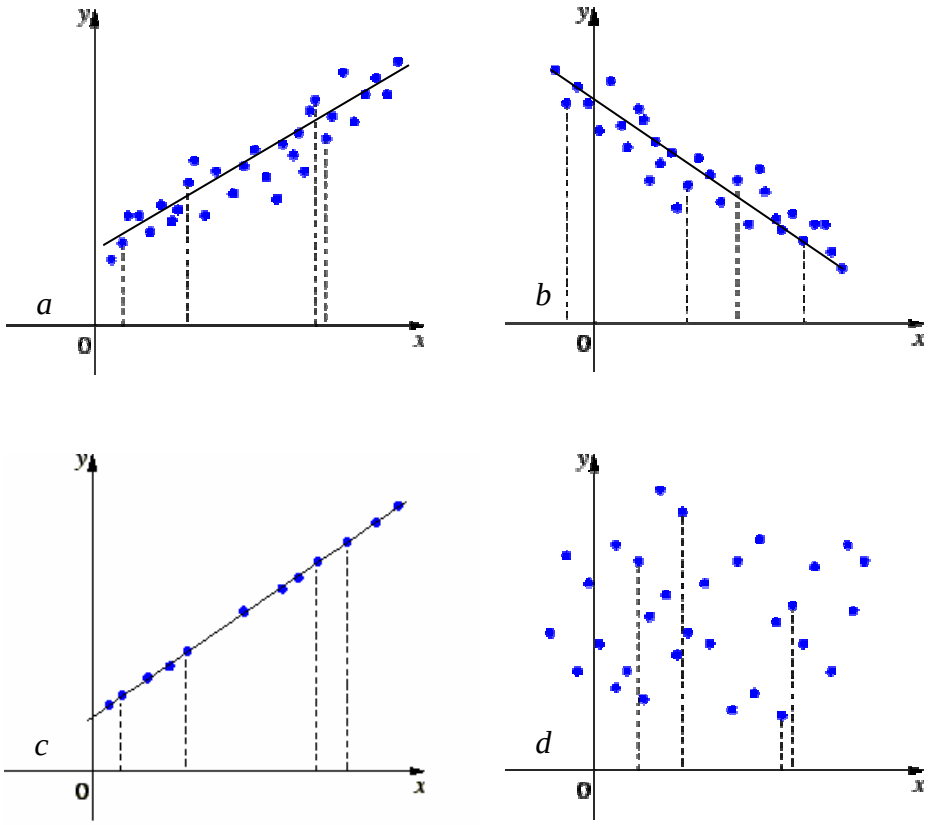


Figura 4 – *a*)correlazione lineare positiva; *b*) correlazione lineare negativa; *c*) correlazione perfetta; *d*) variabili incorrelate.

Per avere una valutazione *quantitativa* della bontà dell'accostamento è però opportuno fare riferimento ad indici sintetici.

MISURE DI CORRELAZIONE

Devianza spiegata e residua

Si consideri l'identità, valida per i = 1, 2, …, N, illustrata nella figura 5,

$$\left(y_i - \bar{y}\right) = \left(y_i - y_i^*\right) + \left(y_i^* - \bar{y}\right) = e_i + \left(y_i^* - \bar{y}\right) \qquad (15)$$

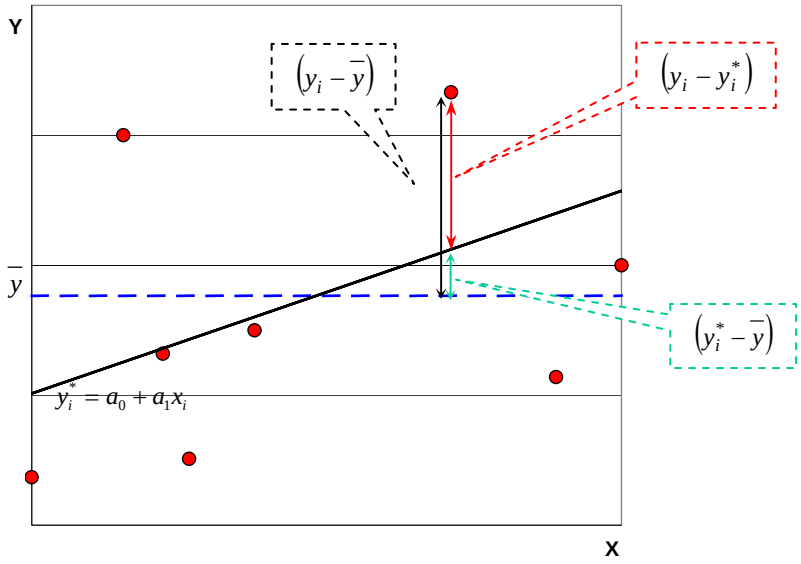


Figura 5

L'identità precedente elevata al quadrato e sommata per i = 1, 2, …, N, diventa:

$$\sum\left(y_i - \bar{y}\right)^2 = \sum\left(y_i - y_i^*\right)^2 + \sum\left(y_i^* - \bar{y}\right)^2 + 2\sum\left(y_i - y_i^*\right)\left(y_i^* - \bar{y}\right) \qquad (16)$$

nella quale è possibile definire:

$$\begin{cases} \sum\left(y_i - \bar{y}\right)^2 = \text{devianza totale di Y} \\ \sum\left(y_i - y_i^*\right)^2 = \text{devianza residua} \\ \sum\left(y_i^* - \bar{y}\right)^2 = \text{devianza spiegata} \end{cases}$$

Il doppio prodotto $\sum\left(y_i - y_i^*\right)\left(y_i^* - \bar{y}\right)$ nella precedente espressione è nullo.

Difatti:

$$y_i^* - \bar{y} = a_0 + a_1 x_i - \left(a_0 + a_1 \bar{x}\right) = a_1 (x_i - \bar{x}) \tag{17}$$

e quindi:

$$\begin{aligned} &\sum \left(y_i - y_i^*\right)\left(y_i^* - \bar{y}\right) = \sum e_i \left(y_i^* - \bar{y}\right) = \sum (y_i - a_0 - a_1 x_i) a_1 (x_i - \bar{x}) = \\ &= a_1 \sum (y_i - a_0 - a_1 x_i) x_i - a_1 \bar{x} \sum (y_i - a_0 - a_1 x_i) = a_1 \cdot 0 - a_1 \bar{x} \cdot 0 = 0 \end{aligned} \tag{18}$$

per le equazioni (6-7) ricavate in precedenza.

Il calcolo della *devianza totale* di Y , quindi, può essere ricondotto alla seguente espressione:

$$\sum \left(y_i - \bar{y}\right)^2 = \sum \left(y_i - y_i^*\right)^2 + \sum \left(y_i^* - \bar{y}\right)^2 \tag{19}$$

ovvero

devianza totale = devianza residua + devianza spiegata.

Tale decomposizione afferma che la variabilità totale del fenomeno Y che si cerca di spiegare tramite una relazione lineare con il fenomeno X, è sempre in parte attribuibile alla retta di regressione (devianza spiegata) ed in parte è dovuta agli errori e (devianza residua). Tanto maggiore sarà il contributo della devianza spiegata tanto più la relazione lineare ipotizzata riuscirà a spiegare la variabilità di Y (retta più vicina ai valori campionari).

Coefficiente di determinazione

Un indice della bontà di accostamento della retta di regressione ai dati campionari può essere definito attraverso il rapporto tra la devianza spiegata e devianza totale, indicato come *coefficiente di determinazione*.

$$r^2 = \frac{devianzaspiegata}{devianzatotale} = \frac{\sum \left(y_i^* - \bar{y}\right)^2}{\sum \left(y_i - \bar{y}\right)^2} \tag{20}$$

Il coefficiente di determinazione è una grandezza adimensionale e può variare tra zero ed uno. Se la devianza spiegata vale zero, cioè i valori stimati sono tutti costanti e pari alla media dei valori osservati, tale coefficiente vale zero. Il coefficiente è, invece, pari ad uno quando la devianza residua vale zero, cioè i valori osservati e stimati coincidono. Nei casi reali naturalmente si ha una situazione intermedia che indica la percentuale di variabilità totale spiegata dalla retta di regressione.

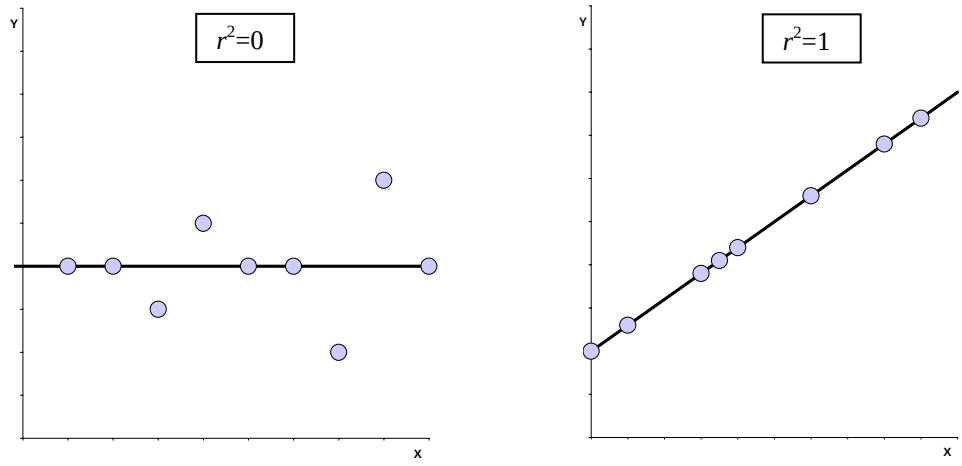


Figura 6- Estremi di variazione dell'indice r^2

Coefficiente di correlazione

La radice quadrata del coefficiente di determinazione viene indicata come *coefficiente di correlazione* che assume valori compresi tra –1 ed 1.

$$r = \pm\sqrt{\frac{devianzaspiegata}{devianzatotale}} \qquad (21)$$

Il segno + o – (da determinare in base ad un'analisi del diagramma a dispersione) indica rispettivamente il caso di correlazione lineare positiva o negativa.

Il coefficiente di correlazione può essere calcolato anche attraverso la seguente espressione, indicata come formula dei *momenti misti*:

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}} \qquad (22)$$

In tal caso ad r viene automaticamente associato il segno corretto.

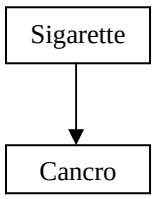
Le proprietà del coefficiente di correlazione sono le seguenti

1) *simmetria* $\Rightarrow r(X,Y) = r(Y,X)$
2) *se X ed Y sono indipendenti* $\Rightarrow r(X,Y) = 0$
3) $r(X,Y) = \pm 1 \Rightarrow Y = a_0 \pm a_1 X$ *cioè tra X ed Y esiste un perfetto legame lineare.*

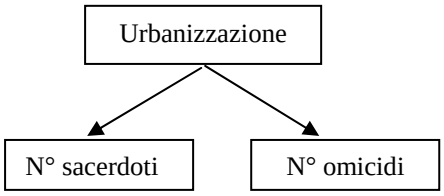
Rischio dell'interpretazione

Non necessariamente una certa correlazione implica un rapporto di causa ed effetto tra le variabili. Ad esempio ci può essere un'alta correlazione tra il numero di omicidi ed il

numero di sacerdoti in una comunità. In tal caso si parla di rapporto di correlazione *spurio* .



Rapporto causa-effetto



Correlazione spuria

ESEMPIO (SCHAUM)

La tabella riporta i pesi X ed Y di un campione di 12 padri e dei loro rispettivi figli primogeniti.

1) Costruire il diagramma a dispersione;
2) Trovare la retta dei minimi quadrati di Y su X;
3) Calcolare il coefficiente di correlazione ed il coefficiente di determinazione.

Risoluzione

Il lavoro necessario per i calcoli può essere organizzato come nella tabella che segue.

	Dati		Elaborazioni		
	X	Y	$(X-\overline{X})^2$	$(Y-\overline{Y})^2$	$(X-\overline{X})(Y-\overline{Y})$
	65	68	2.778	0.174	-0.694
	63	66	13.444	2.507	5.806
	67	68	0.111	0.174	0.139
	64	65	7.111	6.674	6.889
	68	69	1.778	2.007	1.889
	62	66	21.778	2.507	7.389
	70	68	11.111	0.174	1.389
	66	65	0.444	6.674	1.722
	68	71	1.778	11.674	4.556
	67	67	0.111	0.340	-0.194
	69	68	5.444	0.174	0.972
	71	70	18.778	5.840	10.472
somma	800	811	84.667	38.917	40.333
media	66.667	67.583			

1) Il diagramma a dispersione si ottiene riportando i punti (X,Y) su un sistema di coordinate cartesiane (fig.7).

2) La retta di regressione di Y su X è data da $Y = a_0 + a_1 X$ dove a_0 ed a_1 vengono ottenuti a partire dalle equazioni normali.

$$a_1 = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{40.333}{84.667} = 0.476$$

$$a_0 = \bar{y} - a_1 \bar{x} = 67.583 - 0.476 \cdot 66.667 = 35.82$$

La retta ha pertanto equazione Y= 35.82 + 0.476 X (fig.7).

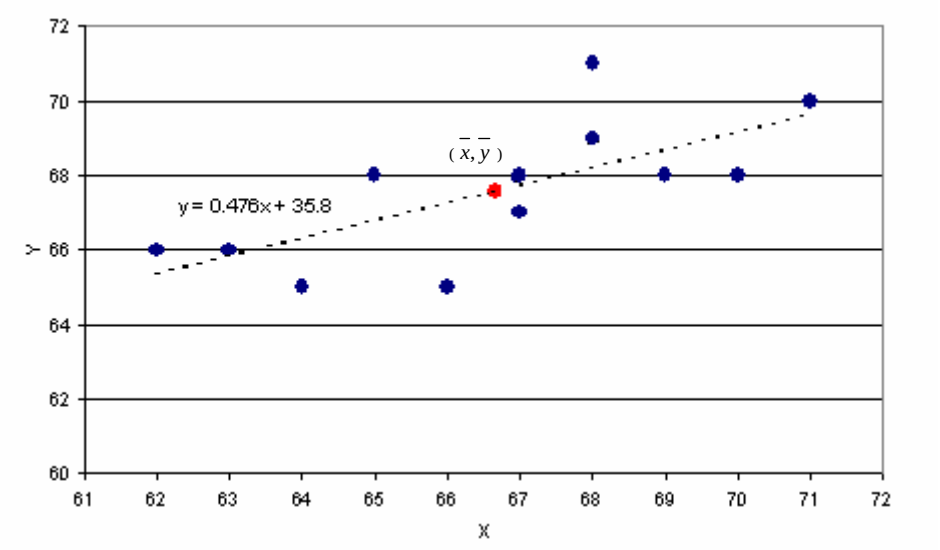


Figura 7

3) Il calcolo del coefficiente di correlazione lineare può essere effettuato attraverso la formula dei momenti misti:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{40.333}{\sqrt{84.667 \cdot 38.917}} = 0.702$$

La correlazione è positiva, infatti, la stima di Y con la retta di regressione aumenta all'aumentare di X.

Il coefficiente di determinazione pertanto risulta:

$r^2 = 0.494$

APPENDICE

A.1

Nel caso in cui la curva interpolante sia una parabola di espressione:

$$Y = a_0 + a_1 X + a_2 X^2 \tag{23}$$

si considerano le coppie (x_i, y_i) con i=1,…,N, e i corrispondenti valori teorici y_i^*:

$$y_i^* = a_0 + a_1 x_i + a_2 x_i^2 \tag{24}$$

e si definisce, come nel caso della regressione lineare, la funzione obiettivo S come:

$$S = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - y_i^* \right)^2 \tag{25}$$

I parametri a_0, a_1 ed a_2 della parabola dei minimi quadrati devono essere tali che:

$$S = \sum_{i=1}^{N} \left(y_i - y_i^* \right)^2 = \sum_{i=1}^{N} \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 \right) \right]^2 = \min \tag{26}$$

S è minimo quando le derivate parziali rispetto ad a_0, a_1 ed a_2 valgono zero.

Allora imponendo le condizioni:

$$\frac{\partial S}{\partial a_0} = -2 \sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \right) = 0 \tag{27}$$

$$\frac{\partial S}{\partial a_1} = -2 \sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \right) x_i = 0 \tag{28}$$

$$\frac{\partial S}{\partial a_2} = -2 \sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \right) x_i^2 = 0 \tag{29}$$

si ottiene il seguente sistema lineare da risolvere nelle incognite a_0, a_1 ed a_2:

$$a_0 N + a_1 \sum_{i=1}^{N} x_i + a_2 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i$$

$$a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + a_2 \sum_{i=1}^{N} x_i^3 = \sum_{i=1}^{N} y_i x_i$$

$$a_0 \sum_{i=1}^{N} x_i^2 + a_1 \sum_{i=1}^{N} x_i^3 + a_2 \sum_{i=1}^{N} x_i^4 = \sum_{i=1}^{N} y_i x_i^2$$

che può essere riscritto in forma matriciale nel modo seguente:

$$\underline{\underline{M}}\,\underline{A} = \underline{B}$$

con

$$\underline{\underline{M}} = \begin{bmatrix} N & \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i^3 \\ \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^4 \end{bmatrix} \qquad \underline{A} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} \qquad \underline{B} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} y_i x_i \\ \sum_{i=1}^{N} y_i x_i^2 \end{bmatrix}$$

A.2

Nel caso di regressione lineare multipla con 2 variabili indipendenti X e Y e con la variabile dipendente Z, si ha la seguente espressione teorica:

$$Z = a_0 + a_1 X + a_2 Y \tag{30}$$

Si considerano le terne (x_i, y_i, z_i) con i=1,…,N, e i corrispondenti valori teorici z_i^*:

$$z_i^* = a_0 + a_1 x_i + a_2 y_i \tag{31}$$

Si definisce la funzione obiettivo S come:

$$S = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(z_i - z_i^*\right)^2 \tag{32}$$

I parametri a_0, a_1 ed a_2 devono essere tali che:

$$S = \sum_{i=1}^{N} \left(z_i - z_i^*\right)^2 = \sum_{i=1}^{N} \left[z_i - \left(a_0 + a_1 x_i + a_2\, y_i\right)\right]^2 = \min \tag{33}$$

S è minimo quando le derivate parziali rispetto ad a_0, a_1 ed a_2 valgono zero.

Allora imponendo le condizioni:

$$\frac{\partial S}{\partial a_0} = -2\sum_{i=1}^{N} \left(z_i - a_0 - a_1 x_i - a_2 y_i\right) = 0 \tag{34}$$

$$\frac{\partial S}{\partial a_1} = -2\sum_{i=1}^{N} \left(z_i - a_0 - a_1 x_i - a_2 y_i\right) x_i = 0 \tag{35}$$

$$\frac{\partial S}{\partial a_2} = -2\sum_{i=1}^{N} \left(z_i - a_0 - a_1 x_i - a_2 y_i\right) y_i = 0 \tag{36}$$

si ottiene il seguente sistema lineare da risolvere nelle incognite $a_{0,}$ a_1 ed a_2:

$$\begin{aligned} & a_0 N + a_1 \sum_{i=1}^{N} x_i + a_2 \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} z_i \\ & a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + a_2 \sum_{i=1}^{N} y_i x_i = \sum_{i=1}^{N} z_i x_i \\ & a_0 \sum_{i=1}^{N} y_i + a_1 \sum_{i=1}^{N} y_i x_i + a_2 \sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} z_i y_i \end{aligned} \qquad (37)$$

che può essere riscritto in forma matriciale nel modo seguente:

$$\underline{\underline{M}}\,\underline{A} = \underline{B} \qquad (38)$$

con

$$\underline{\underline{M}} = \begin{bmatrix} N & \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} y_i x_i \\ \sum_{i=1}^{N} y_i & \sum_{i=1}^{N} y_i x_i & \sum_{i=1}^{N} y_i^2 \end{bmatrix} \qquad \underline{A} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} \qquad \underline{B} = \begin{bmatrix} \sum_{i=1}^{N} z_i \\ \sum_{i=1}^{N} z_i x_i \\ \sum_{i=1}^{N} z_i y_i \end{bmatrix}$$

A.3

Nel caso di regressione semplice non lineare in cui si utilizzi una legge di potenza del tipo:

$$Y = a_0 X^{a_1} \qquad (39)$$

ci si può ricondurre al caso della regressione semplice lineare applicando ad ambo i membri il logaritmo naturale:

$$\ln Y = \ln\left(a_0 X^{a_1}\right) = \ln a_0 + \ln X^{a_1} = \ln a_0 + a_1 \ln X \Rightarrow LY = A_0 + a_1 LX \qquad (40)$$

con $A_0 = \ln a_0$, $LY = \ln Y$ e $LX = \ln X$

Si considerano dunque le coppie (ln x_i, ln y_i) con i=1,…,N, e si segue la procedura descritta dalle espressioni (10)-(11) o (12)-(13) al fine di stimare A_0 e a_1; successivamente si porrà $a_0 = e^{A_0}$

A.4

Nel caso di regressione multipla non lineare in cui si utilizzi una legge di potenza del tipo:

$$Z = a_0 \, X^{a_1} Y^{a_2} \tag{41}$$

ci si può ricondurre al caso della regressione multipla lineare, descritta nella sezione A.3 della presente appendice, applicando ad ambo i membri il logaritmo naturale:

$$\begin{aligned} \ln Z &= \ln\left(a_0 \, X^{a_1} Y^{a_2}\right) = \ln a_0 + \ln X^{a_1} + \ln Y^{a_2} = \\ &= \ln a_0 + a_1 \ln X + a_2 \ln Y \Rightarrow LZ = A_0 + a_1 LX + a_2 LY \end{aligned} \tag{42}$$

con $A_0 = \ln a_0$, $LZ = \ln Z$, $LX = \ln X$ e $LY = \ln Y$

Si considerano dunque le terne (ln x_i, ln y_i, ln z_i) con i=1,…,N, e risolve un sistema lineare analogo a quello descritto nella sezione A.3 al fine di stimare A_0, a_1 e a_2; successivamente si porrà $a_0 = e^{A_0}$