Appunti sulla regressione lineare semplice e multipla*

Germano Rossi†

9 aprile 2004
vers. 0.3.2

Indice

*I titoli fra parentesi quadre indicano paragrafi vuoti, ancora da scrivere

†Università degli Studi di Milano-Bicocca, Dipartimento di Psicologia (germano.rossi@unimib.it)

1 Appunti sulla regressione lineare semplice e multipla

1.1 Introduzione

1.1.1 Un piccolo ripasso di statistica descrittiva

Uno degli indici statistici caratteristici che si possono calcolare su una variabile misurata a livello intervallo o a rapporto, è la media, ovvero la somma di tutti i valori divisi per il loro numero:

$$\bar{X} = \frac{\sum X_i}{N}$$

Se la variabile si distribuisce normalmente, la media sara anche il valore più frequente (moda) e quello che occupa il posto centrale se i dati vengono ordinati in modo crescente (mediana).

Questo è uno dei motivi per cui la media è anche chiamata "speranza matematica" o valore atteso e viene indicata con $E(X)$.

Se la media è un indice della tendenza centrale della distribuzione di una variabile, la varianza e la deviazione standard (o scarto quadratico medio) sono indici di dispersione.

Ricordiamo che la varianza di una qualunque variabile X è la media degli scarti al quadrato. Le formule qui proposte includono quelle per il calcolo con i dati grezzi (ovvero le formule che permettono di effettuare i calcoli più velocemente):

$$var(X) = s_x^2 = s_{xx} = \frac{\sum (X - \bar{X})^2}{N} = \frac{\sum X^2}{N} - \bar{X}^2 = \sum X^2 - \frac{(\sum X)^2}{N} \tag{1.1}$$

La notazione s_{xx} (oppure anche σ_{xx}) viene solitamente usata nella notazione Lisrel per indicare la varianza.

La relativa deviazione standard è

$$s_x = \sqrt{s^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} = \sqrt{\frac{\sum X^2}{N} - \bar{X}^2} \tag{1.2}$$

Un indice statistico analogo alla varianza è la covarianza che può misurare la co-variazione di due variabili:

$$\begin{aligned} cov(X,Y) = s_{xy} = \sigma_{xy} &= \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N} \\ &= \frac{\sum XY}{N} - \bar{X}\bar{Y} = \frac{\sum XY}{N} - \frac{\sum X}{N}\frac{\sum Y}{N} \end{aligned} \tag{1.3}$$

Osserviamo come le formule della varianza e della covarianza siano molto simili, in particolare se scriviamo la prima parte in questo modo:

$$var(X) = \frac{\sum (X - \bar{X})(X - \bar{X})}{N}$$

La notazione s_{yx} (oppure σ_{yx}) viene usata in Lisrel come modo alternativo di indicare la covarianza.

Se i dati sono espressi semplicemente come scarti dalla media (cioè se $x = X - \bar{X}$), le formule della varianza e della covarianza si riducono a:

$$var(x) = s_x^2 = \frac{\sum(x-0)^2}{N} = \frac{\sum x^2}{N} \tag{1.4}$$

$$cov(x,y) = \frac{\sum(x-0)(y-0)}{N} = \frac{\sum xy}{N} \tag{1.5}$$

in quanto le somme degli scarti dalla media ($\sum(X_i - \bar{X})$) sono pari a 0 e così anche la media.

Con questa trasformazione, la distribuzione viene spostata in modo che la media coincida con il valore 0 e tutti i dati vengono spostati concordemente, restando sempre alla stessa distanza dalla media. Varianza e deviazione standard non cambiano affatto, perché non vengono modificate le distanze fra i valori della variabile.

Consideriamo anche che LISREL non utilizza varianza e covarianza calcolate sul campione, ma le relative stime della popolazione:

$$\widehat{var}(X) = \hat{s}_x^2 = \frac{\sum(X-\bar{X})^2}{N-1} = s_x^2 \frac{N}{N-1} \tag{1.6}$$

$$\widehat{cov}(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{N-1} = s_{xy} \frac{N}{N-1} \tag{1.7}$$

Vale la pena di ricordare che la deviazione standard è una distanza e che può essere usata per "standardizzare" le misurazioni di una variabile. I nuovi valori prendono il nome di "punti z":

$$z = \frac{X - \bar{X}}{s}$$

Dopo il processo di standardizzazione, la variabile avrà media zero e deviazione standard 1, perché ogni valore della variabile viene espresso come "numero di deviazioni standard comprese fra il valore e la media".

1.1.2 Correlazione

La correlazione lineare prodotto-momento di Pearson è

$$\begin{aligned} r = \frac{z_x z_y}{N} &= \frac{cov(X,Y)}{s_x s_y} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \\ &= \frac{N\sum XY - \sum X \sum Y}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} \end{aligned} \tag{1.8}$$

Questo indice esprime quanto le due variabili si muovono concordemente: oscilla fra -1 e +1, dove il segno indica l'andamento della relazione (positivo se le due variabili crescono o decrescono assieme; negativo se al crescere di una, l'altra decresce), e dove il valore 1 della correlazione (in valore assoluto) indica la correlazione perfetta, 0 la correlazione nulla e i valori compresi fra 0 e 1, gradi diversi di associazione fra le due variabili. In particolare la correlazione perfetta implica l'esistenza di una relazione matematica (lineare) che permette di calcolare esattamente una variabile conoscendo l'altra.

Se prendiamo in considerazione 3 variabili, sono possibili le seguenti situazioni (vedi Fig. 1):

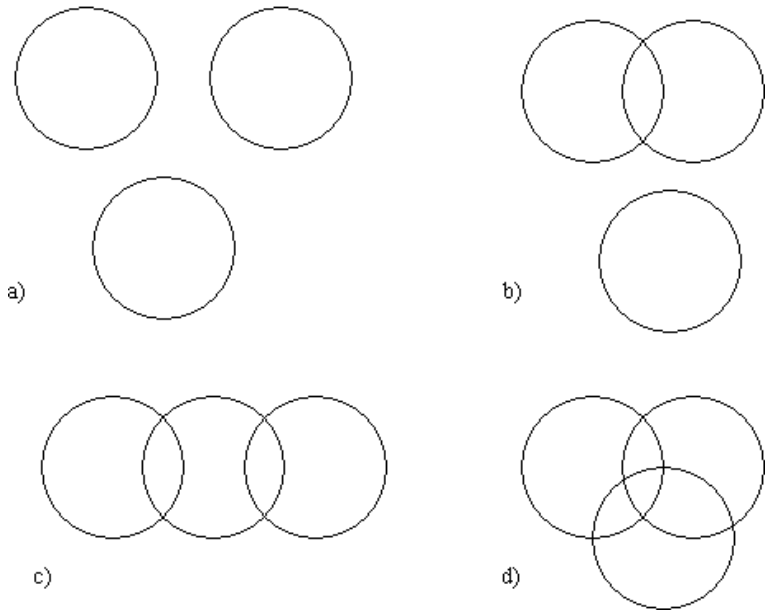


Figura 1: Correlazioni possibili fra 3 variabili

a) sono tutte fra loro non correlate;

b) due variabili sono fra loro correlate, ma non la terza;

c) una variabile è correlata con le altre due che però non sono correlate fra loro;

d) tutte le variabili sono fra loro correlate;

La situazione c), potrebbe corrispondere alla situazione c) della Fig. **??**, in quanto una delle tre variabili potrebbe essere in relazione con le altre due, senza che queste siano fra loro correlate. Tuttavia anche la situazione d) della Fig. 1 potrebbe essere spiegata nello stesso modo: una variabile è in relazione con le altre due e questo produce l'impressione che queste ultime siano fra loro correlate, mentre non lo sono.

La correlazione parziale serve proprio a calcolare la correlazione fra due variabili a cui viene "tolta" l'influenza di una terza variabile. In questo modo se la correlazione parzializzata sulla terza è nulla, sappiamo di trovarci in una situazione come quella rappresentata dalla situazione c).

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \tag{1.9}$$

La correlazione semi-parziale è ancora una correlazione fra due variabili, ma il contributo della terza viene tolto solo da una delle due.

$$r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{23}^2)}} \tag{1.10}$$

Infine la correlazione multipla è la correlazione

di una variabile simultaneamente con due o più variabili:

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{1.11}$$

1.1.3 Matrice di varianza/covarianza

Tabella 1: Esempio di tabella dati

x_1	x_2	x_3	x_4	y
1	2	1	1	1
2	3	1	3	2
4	3	2	2	3
5	5	2	3	4
2	4	2	2	2
14	17	8	11	12

A partire da un certo insieme di dati (ad esempio quello di Tab. 1), possiamo costruire una matrice di varianza/covarianza (Tab. 2) usando le formule appena ricordate (Eq. 1.1 e Eq. 1.3). In questa matrice (che è simmetrica) ogni riga e colonna rappresenta una delle variabili dei dati; lungo la diagonale principale troviamo la varianza di ogni variabile e nei due triangoli le covarianze. Per cui l'elemento in posizione 3,3 sarà la varianza della terza variabile, mentre quello in posizione 2,3 sarà la covarianza fra le variabili 2 e 3. Molto spesso la matrice di varianza/covarianza è chiamata, per semplicità, solo matrice delle varianze, ma si intende che le varianze sono solo quelle lungo la diagonale, mentre tutte le altre sono covarianze.

Tabella 2: Matrice di Varianza/Covarianza con N

	x1	x2	x3	x4	y
x1	2,16	1,08	0,52	0,64	1,48
x2	1,08	1,04	0,36	0,52	0,84
x3	0,52	0,36	0,24	0,08	0,36
x4	0,64	0,52	0,08	0,56	0,52
y	1,48	0,84	0,36	0,52	1,04

Usando le stime della varianza e della covarianza, si otterrà ancora una matrice di varianza/covarianza (Tab. 3). La maggior parte dei programmi di statistica (SPSS, Lisrel...) utilizzano questo tipo di matrice di varianze.

Tabella 3: Matrice di Varianza/Covarianza con N-1

	x1	x2	x3	x4	y
x1	2,70	1,35	0,65	0,80	1,85
x2	1,35	1,30	0,45	0,65	1,05
x3	0,65	0,45	0,30	0,10	0,45
x4	0,85	0,65	0,10	0,70	0,65
y	1,85	1,05	0,45	0,65	1,30

Analogamente è possibile costruire una matrice (o tabella) di correlazione fra queste variabili (Tab. 4); in tal caso la diagonale principale non conterrà nessun valore oppure il valore 1, cioè la correlazione della variabile con se stessa, correlazione positiva perfetta.

E' possibile passare da una matrice di varianza ad una di correlazione applicando ricorsivamente la formula:

$$r_{ab} = \frac{cov(a,b)}{\sqrt{var(a)var(b)}} \tag{1.12}$$

Tabella 4: Matrice di Correlazione

	x1	x2	x3	x4	y
x1	1	0,72	0,72	0,58	0,99
x2	0,72	1	0,72	0,68	0,81
x3	0,72	0,72	1	0,22	0,72
x4	0,58	0,68	0,22	1	0,68
y	0,99	0,81	0,72	0,68	1

dove a e b saranno di volta in volta le varie variabili.

Non è possibile passare da una matrice di correlazione ad una di varianza se non si conoscono almeno le deviazioni standard o le varianze, in quanto $cov(a, b) = r_{ab}s_a s_b$.

In tutti questi casi, visto che le matrici sono simmetriche, è facile imbattersi in matrici la cui metà superiore destra è stata lasciata vuota; non significa che quelle celle sono vuote, ma che è inutile ripetere il loro contenuto dal momento che è speculare lungo la diagonale principale.

1.2 Regressione lineare semplice

1.2.1 Che cos'è

Tabella 5: Dati fittizi

	Test	Voto
A	12	8
B	10	7
C	14	8
D	9	5
E	9	6
F	13	9
G	11	7
H	8	5

Facciamo un esempio numerico, ipotizzando di aver misurato 8 studenti con un test di apprendimento durante l'anno scolastico e di voler studiare la sua relazione con il voto finale della materia (Tab. 5). La correlazione fra il *test* e il *voto* è .91. Questo indice statistico, lo abbiamo già detto, non ci da informazioni sul tipo di relazione esistente. Osservando i dati possiamo vedere che a valori alti nel *test* corrispondono valori alti del *voto* e viceversa. Poiché la variabile *Test* e la variabile *Voto* sono separati nel tempo e successivi, è illogico pensare che il voto finale possa aver avuto un'azione retroattiva e aver influenzato il test, mentre è più logico immaginare che il risultato del test sia in relazione diretta con il voto. Ancora più logico è pensare che entrambe le variabili siano "influenzate" da altre variabili come il numero di ore passate a studiare, la facilità/difficoltà della materia, la predisposizione personale...

A scopo didattico, partiamo dal presupposto che il test possa essere la causa del voto. Se rappresentiamo graficamente le due variabili, usando l'asse X per il Test e l'asse Y per il Voto, otterremo il grafico di Fig. 2. Se lo osserviamo attentamente, possiamo immaginare una linea retta che passa più o meno in mezzo ai punti e che indica la tendenza della relazione fra le due

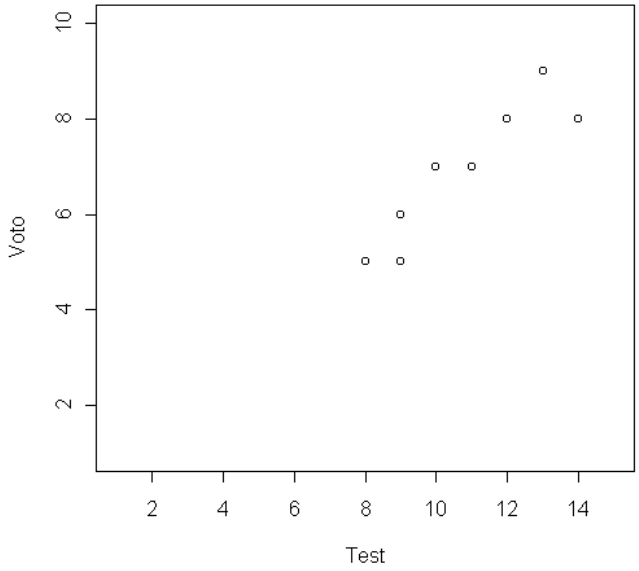


Figura 2: Grafico XY fra il voto e il test

variabili (il modello causale è in Fig. 3). Il grafico di Fig. 4 rappresenta gli stessi dati con la sovrapposizione della retta di tendenza. Vediamo che la retta non va a coprire esattamente tutti i punti del grafico, ma che ci sono dei punti abbastanza vicini e altri più lontani. Noi potremmo sovrapporre diverse rette e ciascuna rappresenterebbe una tendenza fra le due variabili.

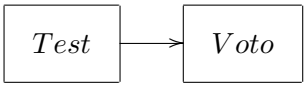


Figura 3: Modello causale

A cosa ci serve conoscere questa tendenza?

Ipotizziamo di accettare per vera una di queste linee di tendenza. Essendo una retta, esiste una funzione matematica che possiamo utilizzare: dato un certo valore di X, lo inseriamo nella retta e troviamo il corrispondente valore Y. Guardiamo il grafico di Fig. 4; mi chiedo: "se uno studente avesse avuto un punteggio di 15 sul test, che voto avrebbe presumibilmente ottenuto?" Presumibilmente perché la retta che abbiamo rappresentato è una delle possibili; presumibilmente perché sappiamo che la retta non coincide esattamente con i dati, ma "ci passa vicino".

A questo punto la domanda diventa: "qual è la retta migliore?"

Ribadiamo che la variabile Y è stimabile tramite una retta che utilizza i dati di X e la cui equazione generica è:

$$Y = a + bX$$

dove a è l'intercetta sull'asse delle Y (cioè il valore assunto da Y quando X=0) e b è la pendenza della retta (Fig. 5).

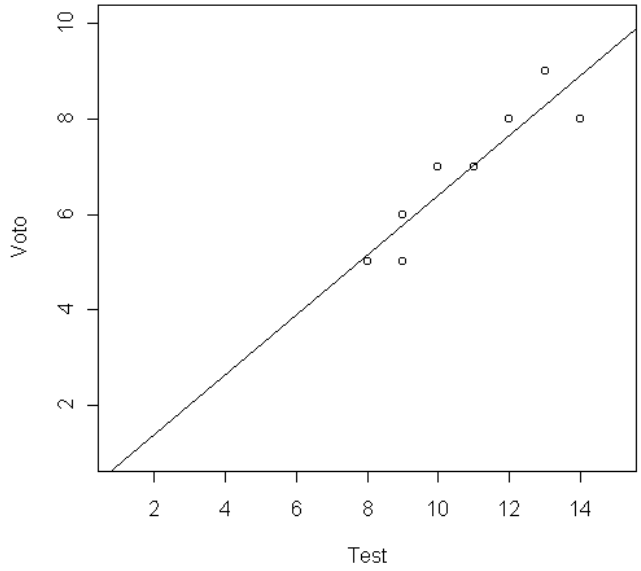


Figura 4: Grafico+retta

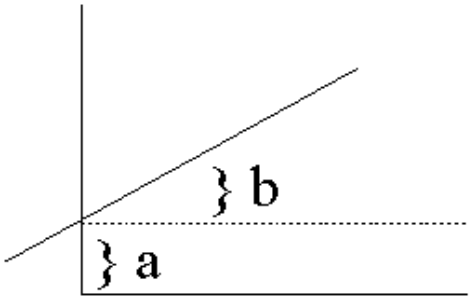


Figura 5: Modello della retta

Questa equazione sarebbe effettivamente vera se la retta si sovrapponesse perfettamente a tutti i punti osservati di Y, mentre invece abbiamo degli errori più o meno piccoli. Riscriviamo allora l'equazione in questi due modi:

$$Y = a + bX + e \qquad\qquad Y - e = a + bX \tag{1.13}$$

dove e indica il residuo (ovvero l'errore) che serve appunto a correggere i dati e ritrovare la Y osservata. In altre parole, dopo aver calcolato il valore di Y sulla base della retta di regressione (che chiameremo Y stimato, ovvero $\hat{Y}$ oppure Y'), otteniamo l'errore come differenza dal valore osservato:

$$e = Y - \hat{Y} \tag{1.14}$$

L'equazione di regressione può quindi essere scritta in due modi:

$$\begin{aligned} \hat{Y} &= a + bX \\ Y &= a + bX + e \end{aligned} \tag{1.15}$$

Nel primo caso, usiamo la formula della retta di regressione, ma otteniamo una stima di Y; nel secondo caso, Y è ri-calcolato correttamente, aggiungendo un errore di stima, definito come dall'equazione 1.14.

Tuttavia, poiché l'errore cambia per ogni valore di Y, mentre a e b restano uguali per l'intera retta, per essere corretti dovremmo scrivere:

$$\begin{aligned} \hat{Y_i} &= a + bX_i \\ Y_i &= a + bX_i + e_i \end{aligned}$$

in quanto ogni soggetto ha un suo Y, un suo X e un suo errore di stima, mentre i parametri a e b sono uguali per tutti. La seconda forma dell'equazione ci mette però in evidenza che la retta non stima esattamente e che ogni valore di Y ha un suo errore. La retta dipende dai valori di a e di b che sono rispettivamente l'intercetta sull'asse delle X e la pendenza della retta. Rette parallele a quella del grafico di Fig. 4 avranno la stessa pendenza ma intercette diverse, mentre il grafico di Fig. 6 mostra due diverse rette con una diversa pendenza.

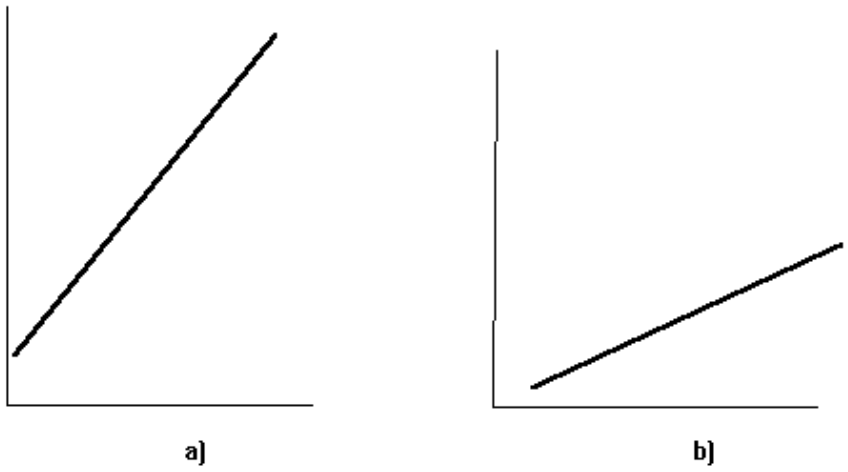


Figura 6: Rette con pendenze diverse

Quale sarà la migliore?

Ovviamente quella che si avvicina di più a tutti i punti osservati. E come possiamo stabilirlo? Ragioniamo: dopo l'uso di una retta, abbiamo degli errori di stima (i valori e). Dobbiamo trovare la retta che produce i valori di e più piccoli possibili. Se proviamo a sommare questi errori (in riferimento ad una determinata retta) scopriamo che si annullano: $\sum e_i = 0$. Usiamo allora il trucco del quadrato (già usato con gli scarti dalla media per giungere alla varianza):

$$\sum e_i^2 \neq 0$$

La somma degli errori al quadrato sarà tanto più grande quanto più grandi sono gli errori di partenza, ovvero la diversità fra il valore Y osservato e quello stimato ($\hat{Y}$) da una retta di

regressione. La somma al quadrato degli errori dovrebbe essere allora la più piccola possibile, ovvero dev'essere al minimo.

Ma come facciamo a trovare e? Per difetto!

Ovvero, prima usiamo l'equazione della retta per calcolare la stima di Y ($\hat{Y}$ o Y'), quindi calcoliamo l'errore come differenza dai valori osservati:

$$e_i = Y_i - \hat{Y}_i \tag{1.16}$$

e quindi:

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = min \tag{1.17}$$

1.2.2 Metodo dei minimi quadrati

La procedura che minimizza l'errore e si chiama "metodo dei minimi quadrati" (in inglese: OLS, "Ordinary Least Squares criterion") e trovate il procedimento completo in appendice A.1.1. Per i nostri scopi attuali basterà dire che la pendenza corrisponde a una delle seguenti formule:

$$\begin{aligned} b &= r\frac{s_y}{s_x} = r\frac{N s_x s_y}{N s_x^2} = \frac{cov(XY)}{var(X)} \\ &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{N\sum X_i Y_i - \sum X_i \sum Y_i}{N\sum X_i^2 - (\sum X_i)^2} \end{aligned} \tag{1.18}$$

Mentre l'intercetta si trova con:

$$a = \bar{Y} - b\bar{X} = \frac{\sum Y}{N} - b\frac{\sum X}{N} \tag{1.19}$$

Guardando queste formule notiamo alcune cose, relativamente alla pendenza:

- la correlazione viene ponderata sulla base delle deviazioni standard delle due variabili (prima forma dell'eq. 1.18)
- nella terza forma, il numeratore corrisponde alla covarianza fra X e Y e il denominatore alla varianza di X.
- l'ultima forma della formula usa i dati grezzi e usa 5 informazioni: N, $\sum X$, $\sum Y$, $\sum X^2$, $\sum XY$ che sono i dati solitamente ricavati per calcolare media e deviazione standard, tranne $\sum XY$ che si ricava per la correlazione.

Rivediamo le equazioni in parallelo e analizziamo ogni simbolo.

$Y_i = bX_i + a + e_i$	Y è la variabile dipendente osservata
	$\hat{Y}$ è la stima della variabile dipendente
$\hat{Y}_i = bX_i + a$	X è la variabile indipendente
	b è l'inclinazione della retta
$e_i = Y_i - \hat{Y}_i$	a è l'intercetta
	e è l'errore della stima

Applichiamo le equazioni 1.18 e 1.19 ai dati di Tabella 5 (vedi Tab. 6).

Dalla ultime righe della tabella 6 ricaviamo: N=8, $\sum X = 86$, $\sum Y = 55$, $\sum X^2 = 956$, $\sum XY = 611$ e se li sostituiamo nell'equazione 1.18 otteniamo:

Tabella 6: Dati per il calcolo delle stime

	X-Test	Y-Voto	X^2	Y^2	XY
A	12	8	144	64	96
B	10	7	100	49	70
C	14	8	196	64	112
D	9	5	81	25	45
E	9	6	81	36	54
F	13	9	169	81	117
G	11	7	121	49	77
H	8	5	64	25	40
Somma	86	55	956	393	611
Media	10,75	6,875			

$$b = \frac{8 \times 611 - 86 \times 55}{8 \times 956 - (86)^2} = \frac{4888 - 4730}{7648 - 7396} = \frac{158}{252} = 0,627$$

Quindi usiamo b nell'equazione 1.19 e le medie ricavabili dall'ultima riga della tabella 6 ($\bar{X} = 10.75$, $\bar{Y} = 6.875$):

$$a = 6.875 - 0,627 \times 10.75 = 0.135$$

Se proviamo ad usare le formule più semplici basate su varianza e covarianza otteniamo lo stesso valore della pendenza:

$$cov(X, Y) = \frac{611}{8} - 10.75 \times 6.875 = 76.375 - 73.906 = 2.469$$

$$var(X) = \frac{956}{8} - (10.75)^2 = 119.5 - 115.563 = 3.938$$

$$b = \frac{2.469}{3.938} = 0.627$$

E se proviamo tramite la formula della correlazione, avremo ancora:

$$r_{xy} = \frac{8 \times 611 - 86 \times 55}{\sqrt{(8 \times 956 - 86^2)(8 \times 393 - 55^2)}} = \frac{158}{\sqrt{252 \times 119}} = \frac{158}{173.17} = 0.912$$

$$s_x = \sqrt{3.938} = 1.984 \qquad s_y = \sqrt{\frac{393}{8} - 6.875^2} = 1.364$$

$$r\frac{s_y}{s_x} = 1.912\frac{1.364}{1.984} = 0.627$$

A questo punto abbiamo entrambi i parametri e possiamo costruire l'equazione completa:

$$\hat{Y}_i = 0,135 + 0,627X_i$$

Per ogni Y possiamo ri-scrivere l'equazione per arrivare a trovare le stime (colonne 4 e 5 della Tab. 7), quindi possiamo calcolare i residui, ovvero gli errori (ultima colonna).

Se rappresentassimo graficamente X e $\hat{Y}$ (la stima) vedremmo che i punti si dispongono in una perfetta linea retta.

Tabella 7: Stime e residui

	Test X	Voto Y	Eq.	Stimati $\hat{Y}$	Residui $Y - \hat{Y}$
A	12	8	0.135+0.627(12)	7.659	0.341
B	10	7	0.135+0.627(10)	6.405	0.595
C	14	8	0.135+0.627(14)	8.913	-0.913
D	9	5	0.135+0.627(9)	5.778	-0.778
E	9	6	0.135+0.627(9)	5.778	0.222
F	13	9	0.135+0.627(13)	8.286	0.714
G	11	7	0.135+0.627(11)	7.032	-0.032
H	8	5	0.135+0.627(8)	5.151	-0.151

1.2.3 Con dati standardizzati

Una b piccola equivale ad una piccola pendenza (Fig. 6b) mentre una retta come quella in Fig 6a dovrebbe avere una forte pendenza e quindi un valore elevato di b. Ma il parametro di regressione b (cioè il coefficiente di regressione o pendenza della retta) dipende dal modo in cui è espressa la variabile X e poiché è espressa su una propria gamma, non possiamo valutare se b è grande o piccola in modo diretto, senza fare riferimento alla media e alla deviazione standard di X. Se però trasformiamo i dati in punti z e se lavoriamo con quest'ultimi, allora X e Y sono espressi in una stessa scala di misura e le medie di X e Y standardizzate saranno 0 e la loro deviazione standard sarà 1. Per questo motivo la formula 1.18 si riduce a (usiamo b^* per indicare la pendenza standardizzata):

$$b^* = r \tag{1.20}$$

Infatti, se s_x e s_y (in forma standardizzata) valgono 1, allora l'eq. 1.18 diventa[1]:

$$b = r\frac{s_y}{s_x} \qquad b^* = r\frac{1}{1} = r$$

L'intercetta a sua volta si annulla perché, ripeto, le due medie (con dati standardizzati) da cui dipende valgono 0 e quindi:

$$a = \bar{Y} - b\bar{X} = 0 - b(0) = 0$$

Le due formule in parallelo sono:

$$\begin{aligned} &\hat{Y}_i = bX_i + a && (normale) \\ &z_{\hat{Y}_i} = bz_{X_i} && (standardizzata) \end{aligned} \tag{1.21}$$

Possiamo riscrivere queste formule come:

$$\hat{Y}_i = r\frac{s_y}{s_x}X_i + \left(\bar{Y} - \frac{s_y}{s_x}\bar{X}\right)$$

in cui se s_x e s_y valgono 1 e $\bar{X}$ e $\bar{Y}$ valgono 0, abbiamo:

$$\begin{aligned} &\hat{Y}_i = r\frac{s_y}{s_x}X_i + a \\ &z_{\hat{Y}_i} = rz_{X_i} \end{aligned} \tag{1.22}$$

[1]La stessa cosa succede quando la varianza (e quindi la deviazione standard) di X e Y sono uguali.

In queste due ultime equazioni possiamo notare come, dal momento che b si riduce a r ed r oscilla fra -1 e 1, possiamo pensare a $z_{\hat{Y}_i}$ come ad una proporzione (con segno) di $z_{\hat{X}_i}$, ovvero una parte di X standardizzato.

Questo è vero solo e soltanto nel caso della regressione lineare semplice, ma non vale con la regressione multipla che vedremo successivamente.

La pendenza standardizzata della regressione semplice è dunque uguale alla correlazione (cioè .912), ma se volessimo calcolarla a partire da quella non standardizzata, dovremmo usare la seguente formula (la differenza sul terzo decimale dipende dall'approssimazione):

$$b^* = r = b\frac{s_x}{s_y} \qquad b^* = .627 \times \frac{1.984}{1.363} = .913$$

Notate che le deviazioni standard sono invertite rispetto all'Eq. 1.18.

Tabella 8: Dati espressi come scarti dalla media

	x-Test	y-Voto	x^2	y^2	xy
A	1,25	1,125	1,56	1,27	1,41
B	-0,75	0,125	0,56	0,02	-0,09
C	3,25	1,125	10,56	1,27	3,66
D	-1,75	-1,875	3,06	3,52	3,28
E	-1,75	-0,875	3,06	0,77	1,53
F	2,25	2,125	5,06	4,52	4,78
G	0,25	0,125	0,06	0,02	0,03
H	-2,75	-1,875	7,56	3,52	5,16
Somma	0	0	31,5	15,0	19,8
Media	0	0			

Vediamo ora cosa succede in pratica quando usiamo i dati espressi come scarti dalla media e poi come punti z.

Trasformiamo i dati di Tab. 5 in scarti dalla media (useremo x e y minuscoli per indicare gli scarti). In questo modo abbiamo spostato l'origine dei dati sulle rispettive medie e quindi ci aspettiamo che la pendenza non cambi, ma che scompaia l'intercetta (perché diventa zero). Su questi dati applichiamo le formule 1.18 e 1.19 (il valore che troviamo è diverso a causa degli arrotondamenti nel calcolo degli scarti, dovrebbe essere 0.627).

$$b = \frac{8 \times 19.8 - 0 \times 0}{8 \times 31.5 - (0)^2} = \frac{158.4}{252.0} = 0.628$$

$$a = 0 - 0.628 \times 0 = 0$$

Se invece usiamo i dati espressi in punti z (quindi con una misura standardizzata, Tab. 9), ci aspettiamo che la pendenza calcolata coincida con la correlazione e che l'intercetta si annulli.

$$b = \frac{8 \times 7.3 - 0 \times 0}{8 \times 8 - (0)^2} = \frac{58.4}{64} = 0.913$$

1.2.4 Residui

Occupiamoci ora degli scarti. In Fig. 7 possiamo vedere una retta di regressione fra X e Y, la linea orizzontale rappresenta la media di Y ($\bar{Y}$) e il punto rappresenta un qualunque ipotetico

Tabella 9: Dati espressi come punti z

	z_x-Test	z_y-Voto	z_x^2	z_y^2	$z_x z_y$
A	0,63	0,83	0,40	0,68	0,52
B	-0,38	0,09	0,14	0,01	-0,03
C	1,64	0,83	2,68	0,68	1,35
D	-0,88	-1,38	0,78	1,89	1,21
E	-0,88	-0,64	0,78	0,41	0,57
F	1,13	1,56	1,29	2,43	1,77
G	0,13	0,09	0,02	0,01	0,01
H	-1,39	-1,38	1,92	1,89	1,91
Somma	0	0	8	8	7,3
Media	0	0			

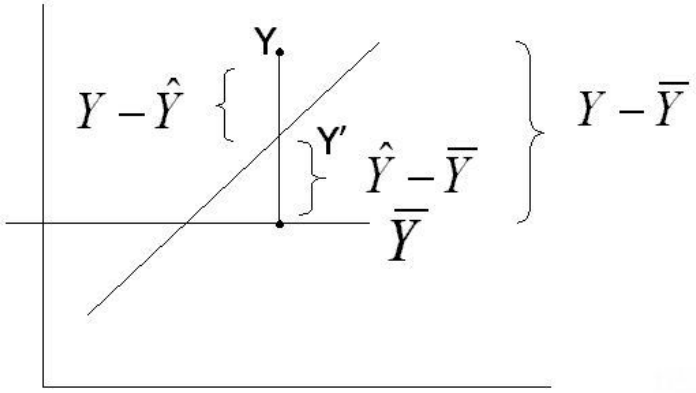


Figura 7: Scarti

valore realmente osservato di Y. La distanza fra Y e la media (ovvero $Y - \bar{Y}$) è lo scarto di Y dalla media e possiamo pensare ai valori Y come a delle deviazioni individuali dalla media, per ragioni ignote. Il punto in cui la retta di regressione interseca lo scarto dalla media, corrisponde al valore stimato $\hat{Y}$. Possiamo quindi dire che una parte dello scarto dalla media è, in qualche modo, spiegato dalla retta di regressione ed è il segmento $\hat{Y} - \bar{Y}$. Avanza una parte di segmento che non viene spiegato dalla retta $(Y - \hat{Y})$ e che è esattamente coincidente con il residuo o errore dell'equazione 1.14. Tutti questi valori ($Y - \bar{Y}$, $\hat{Y} - \bar{Y}$ e $Y - \hat{Y}$) stanno fra loro in una ben precisa relazione:

$$Y - \bar{Y} = (Y - \hat{Y}) + (\hat{Y} - \bar{Y})$$

Se vengono sommati per tutti i punti della variabile dipendente sommano a 0, per cui li eleviamo a quadrato. Se poi li dividessimo per N (la numerosità), potremo notare la somiglianza con la formula della varianza.

Per una serie di trasformazioni matematiche (vedi in Appendice A.1.2) possiamo scrivere che:

$$\underset{\text{totale}}{\sum (Y - \bar{Y})^2} = \underset{\text{non spiegata}}{\sum (Y - \hat{Y})^2} + \underset{\text{spiegata}}{\sum (\hat{Y} - \bar{Y})^2} \tag{1.23}$$

	Residui $Y - \hat{Y}$	Non sp. $(Y - \hat{Y})^2$	Totale $(Y - \bar{Y})^2$
A	0,341	0,116	1,266
B	0,595	0,354	0,016
C	-0,913	0,834	1,266
D	-0,778	0,605	3,516
E	0,222	0,049	0,766
F	0,714	0,510	4,516
G	-0,032	0,001	0,016
H	-0,151	0,023	3,516
Somma	0	2,492	14,875

Tabella 10: Dati per il calcolo di r^2

e possiamo calcolare la proporzione di varianza spiegata rispetto al totale della varianza. Questa proporzione coincide con il quadrato di r (cioè della correlazione fra X e Y):

$$r^2 = (r)^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{\sum(Y - \bar{Y})^2 - \sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2} \tag{1.24}$$

L'r^2 o proporzione di varianza spiegata è anche chiamata "coefficiente di determinazione" oppure varianza comune

$$r^2 = (r)^2 = \frac{cov(XY)^2}{s_X^2 s_Y^2} \tag{1.25}$$

E' importante ricordare che una proporzione è un valore compreso fra 0 e 1 e quindi che l'r^2 non potrà mai superare il valore di 1.

Usando i dati di Tab. 10, proviamo a calcolare l'r^2 tramite l'Eq. 1.24 e subito dopo con l'Eq. 1.25. Dopo aver fatto il quadrato dei residui, li sommiamo e otteniamo la varianza non spiegata (ovvero $Y - \hat{Y}$), quindi sommiamo gli scarti dalla media (ovvero $Y - \bar{Y}$). Con questi dati impostiamo il calcolo:

$$\frac{14.875 - 2.492}{14.875} = .832 \quad \text{oppure} \quad \frac{2.469^2}{3.937 \times 1.859} = .833$$

Per verifica, facciamo il quadrato della correlazione che avevamo ottenuto (.912), e dovremmo ottenere esattamente lo stesso risultato.

$$.912^2 = .832$$

Il concetto di varianza comune può essere rappresentato come l'intersezione fra due aree (v. Fig. 8): se non vi è intersezione l'r^2 sarà pari a 0, mentre se le due aree si sovrappongono completamente, l'r^2 sarà uguale a 1, in tutti gli altri casi l'r^2 oscillerà fra 0 e 1.

Parallelamente al coefficiente di determinazione esiste un coefficiente di indeterminazione che è il suo complemento a 1:

$$1 - r^2$$

1.2.5 Errore standard

Noi sappiamo che Y è una misura osservata casuale, così come lo è X. Poiché usiamo X per stimare Y, dobbiamo essere consapevoli che la stima non è "esatta" se riferita ad un campione

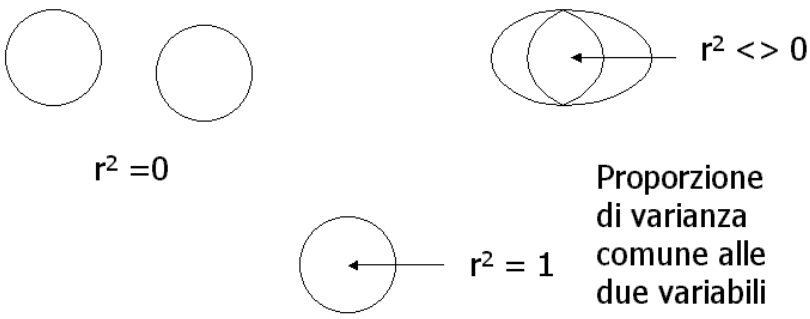


Figura 8: Rappresentazione grafica della varianza comune

estratto casualmente da una popolazione. Vale a dire che se in base all'equazione 1.15 noi volessimo stimare i voti corrispondenti a $X = 7$ e a $X = 15$, otterremmo rispettivamente $Y = 4.5$ e $Y = 9.5$ che non sono voti reali, ma stimati. Dobbiamo essere consapevoli che un eventuale studente che si sottopone al test e che ottiene come valore $X = 15$ potrebbe alla fine dell'anno ottenere un voto reale inferiore (come 8 o 9) oppure anche superiore (come 10).

La deviazione standard degli errori previsti ci può aiutare a capire la gamma di oscillazione del voto reale previsto (sempre nell'ipotesi che la variabile si distribuisca normalmente):

$$s_{y.x}^2 = \frac{\sum(Y - \hat{Y})^2}{N} \tag{1.26}$$

$$s_{y.x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{N}} = s_y\sqrt{1 - r^2} \tag{1.27}$$

Usando i dati trovati in precedenza, calcoliamo:

$$\sqrt{\frac{2.492}{8}} = \sqrt{.312} = .558 \quad \text{ovvero} \quad 1.363\sqrt{1 - .832} = .559$$

Osservando l'ultima forma dell'equazione 1.27, possiamo notare come la deviazione standard dell'errore previsto dipenda da r. Se la correlazione è perfetta ($r = 1$, ovvero esiste una funzione lineare che trasforma X in Y), la stima dell'errore va a 0 (non c'è nessun errore):

$$s_y\sqrt{1 - r^2} = s_y\sqrt{1 - 1^2} = 0$$

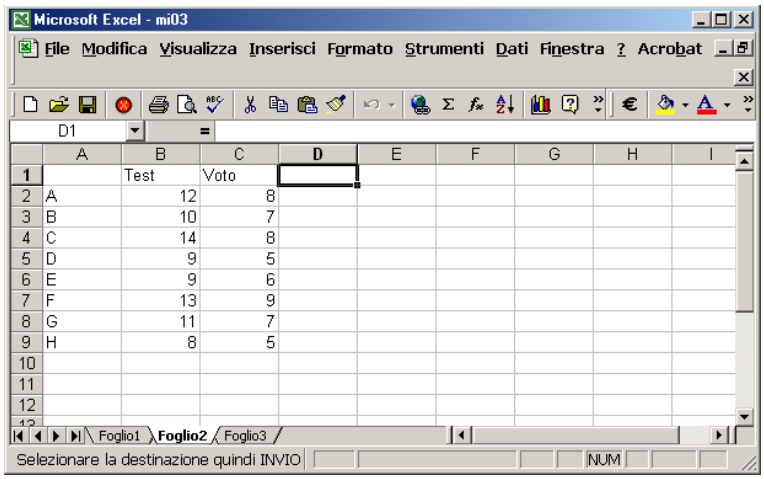
mentre se la correlazione è nulla ($r = 0$, ovvero non esiste alcun andamento comune fra le due variabili) l'errore standard è massimo e coincide con la deviazione standard della variabile Y:

$$s_y\sqrt{1 - r^2} = s_y\sqrt{1 - 0^2} = s_y$$

La deviazione standard degli errori previsti viene chiamata anche "errore stimato" o "errore standard delle stime".

Per conoscere le possibili oscillazioni della previsione di un qualunque Y possiamo usare l'intervallo di fiducia. Nell'ipotesi che X e Y si distribuiscano normalmente, l'intervallo di fiducia al 95% si calcola con:

$$\hat{Y} - 1.96s_{y.x} \qquad \text{e} \qquad \hat{Y} + 1.96s_{y.x}$$



	A	B	C	D	E	F	G	H	I
1		Test	Voto						
2	A	12	8						
3	B	10	7						
4	C	14	8						
5	D	9	5						
6	E	9	6						
7	F	13	9						
8	G	11	7						
9	H	8	5						
10									
11									
12									

Figura 9: Dati di esempio inseriti in Excel

dove $\hat{Y}$ è la stima di Y a partire da un certo X, 1.96 è il valore del punto z corrispondente ad un'area del 95% (.9500) attorno alla media della curva normale e $s_{y.x}$ è la deviazione standard dell'errore previsto.

Se, ad esempio, l'errore standard delle stime fosse pari a $s_{y.x} = 3.5$ e $\hat{Y} = 70$, l'intervallo di fiducia porterebbe ai due valori:

$$70 - 1.96(3.5) = 63.14 \qquad \text{e} \qquad 70 + 1.96(3.5) = 76.86$$

Vale a dire che il valore realmente osservato di Y potrà oscillare fra 63.14 e 76.86.

1.2.6 [Test di Significatività]

1.2.7 Regressione lineare semplice con Excel

1.2.7.1 Inserire i dati

In un tabellone vuoto di Excel, inserire i dati di X e di Y. Ad esempio i dati della Tabella 5 dovrebbero essere inseriti in Excel in modo che corrispondano alla Figura 9.

1.2.7.2 Calcolare i totali di colonna e le medie

Per calcolare i totali di colonna, andare in fondo alla colonna, lasciare una cella vuota e nella cella successiva scrivere:

`=SOMMA(cella_inizio:cella_fine)`

Nel caso specifico dell'esempio, posizionarsi sulla cella B11 e inserire (fig. 10):

`=SOMMA(B2:B9)`

Ci sono due modi per inserire un'area di celle:

1. dopo aver scritto l'uguale e la parola `somma`, aprire la parentesi tonda, quindi spostarsi con il cursore sulla prima cella dell'area che interessa, premere un punto e spostarsi sull'ultima cella dell'area. A questo punto, scrivere la parente tonda di chiusura;

2. scrivere direttamente l'area usando i riferimenti di cella separati da un due punti o da un punto.

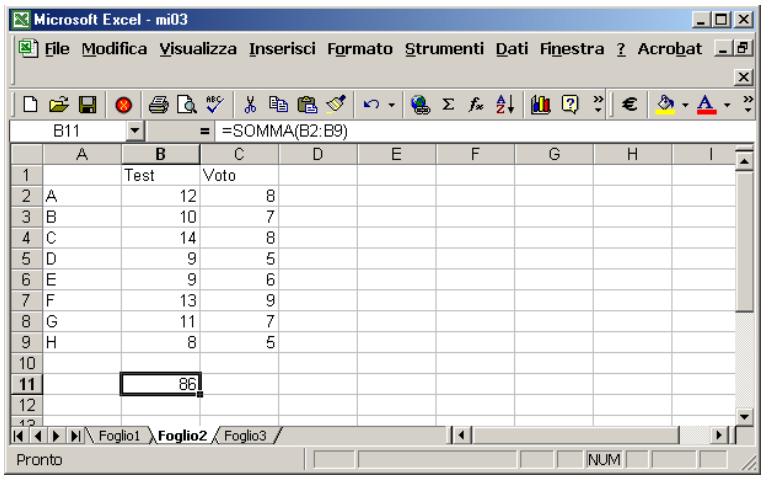


Figura 10: Comando somma

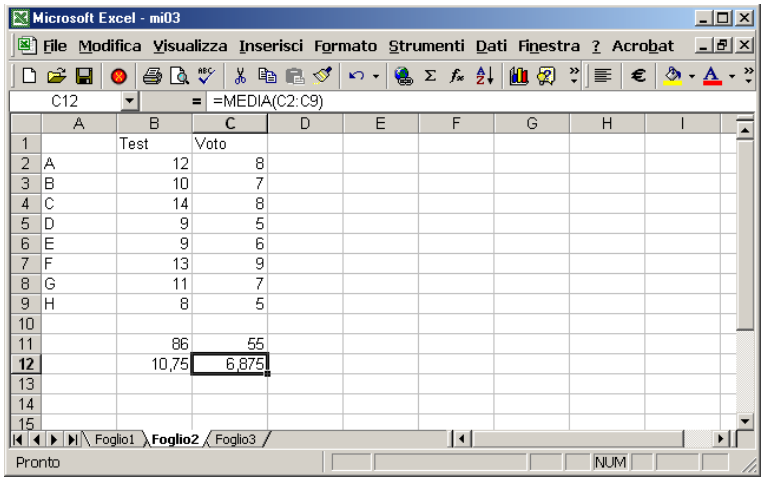
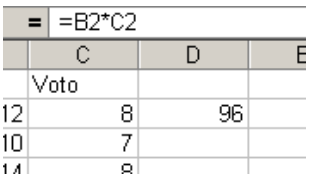


Figura 11: Comando media

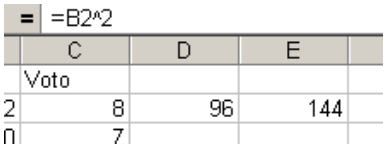
Nella cella immediatamente sotto scrivere (fig. 11):
`=MEDIA(B2:B9)`
Quindi evidenziare le due celle e copiarle nella colonna successiva; i riferimenti di cella si aggiusteranno automaticamente puntando all'area della variabile Y.

1.2.7.3 Calcolare il prodotto delle variabili e i quadrati

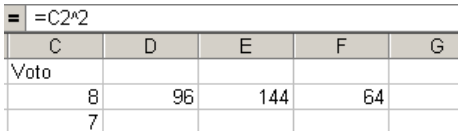
Nelle colonne successive, possiamo inserire il prodotto fra X e Y, il quadrato di X e quello di Y. In base al nostro esempio:
Andare sulla cella D2 e scrivere: `=B2*C2`



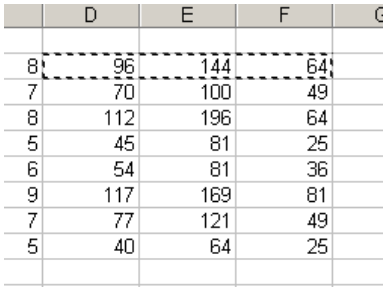
Andare alla cella E2 e scrivere: `=B2^2` (l'accento circonflesso indica l'elevazione a potenza, in Excel)



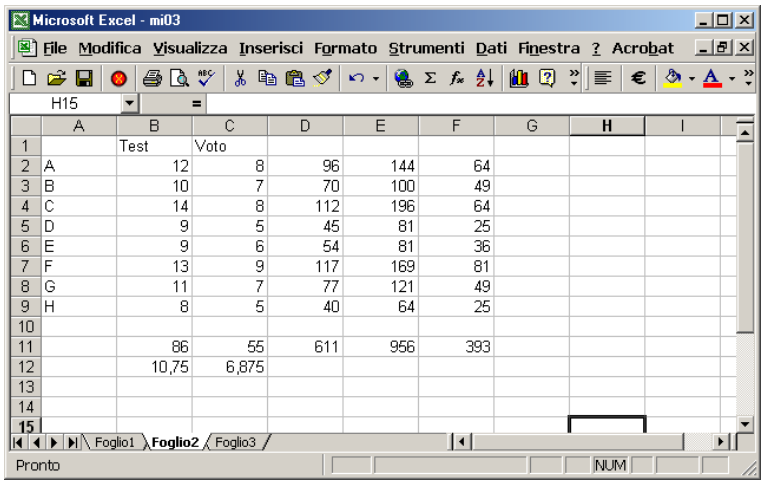
Andare alla cella F2 e scrivere. `=C2^2`



Quindi copiare le tre celle in verticale.



Infine copiare dalle colonne precedenti la formula della somma:

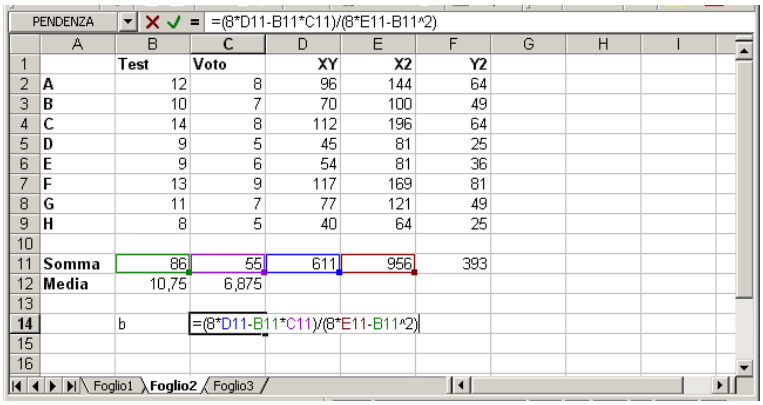


1.2.7.4 Calcolare la pendenza

Per calcolare la pendenza dobbiamo costruire in una cella la formula completa:

`=(8*D11-B11*C11)/(8*E11-B11^2)`

in cui, 8 è la numerosità (N), D11 il prodotto fra X e Y, B11 la somma di X, C11 la somma di Y e E11 è la somma dei quadrati di X.



	A	B	C	D	E	F
1		Test	Voto	XY	X2	Y2
2	A	12	8	96	144	64
3	B	10	7	70	100	49
4	C	14	8	112	196	64
5	D	9	5	45	81	25
6	E	9	6	54	81	36
7	F	13	9	117	169	81
8	G	11	7	77	121	49
9	H	8	5	40	64	25
10						
11	Somma	86	55	611	956	393
12	Media	10,75	6,875			
13						
14		b	=(8*D11-B11*C11)/(8*E11-B11^2)			
15						
16						

Se vogliamo essere sicuri di aver fatto i procedimenti giusti, possiamo andare in un'altra cella e scrivere:

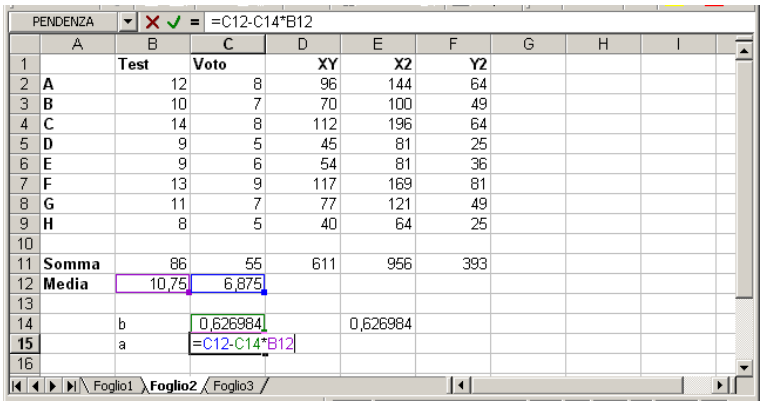
```
=PENDENZA(C2:C9;B2:B9)
```

Excel calcolerà la pendenza per noi e possiamo confrontare i risultati.

1.2.7.5 Calcolare l'intercetta

Lo stesso vale per l'intercetta, che si basa sulle medie di X (B12) e di Y (C12) e sull'intercetta (C14):

```
=C12-C14*B12
```



	A	B	C	D	E	F
1		Test	Voto	XY	X2	Y2
2	A	12	8	96	144	64
3	B	10	7	70	100	49
4	C	14	8	112	196	64
5	D	9	5	45	81	25
6	E	9	6	54	81	36
7	F	13	9	117	169	81
8	G	11	7	77	121	49
9	H	8	5	40	64	25
10						
11	Somma	86	55	611	956	393
12	Media	10,75	6,875			
13						
14		b	0,626984		0,626984	
15		a	=C12-C14*B12			
16						

Anche in questo caso per essere sicuri, in un'altra cella possiamo chiedere ad Excel di calcolare l'intercetta:

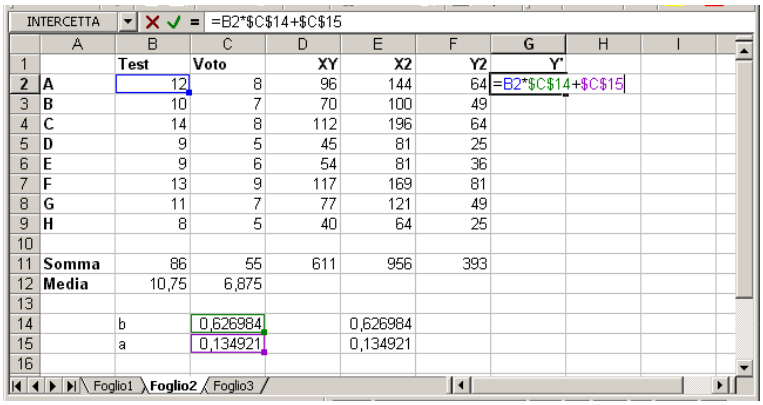
```
=INTERCETTA(C2:C9;B2:B9)
```

1.2.7.6 Calcolare le stime di y

Calcoliamo le stime di Y in una nuova colonna. In questo caso, nella prima cella scriviamo:

```
=B2*$C$14*$C$15
```

B2 è la cella di X di cui vogliamo calcolare la stima Y', C14 è il riferimento di cella al coefficiente angolare (b) e C15 è il riferimento all'intercetta (a). In entrambi i casi C14 e C15 sono scritti con dei dollari ($) davanti perché in questo modo Excel saprà che, quando copieremo la cella, non deve aggiustare questi indirizzi. Il dollaro davanti ad un riferimento di riga o di colonna, significa che quel riferimento dev'essere considerato come assoluto. In caso contrario viene considerato relativo e "aggiustato" durante le operazioni di copia.



	A	B	C	D	E	F	G	H	I
1		Test	Voto	XY	X2	Y2	Y'		
2	A	12	8	96	144	64	=B2*C14+C15		
3	B	10	7	70	100	49			
4	C	14	8	112	196	64			
5	D	9	5	45	81	25			
6	E	9	6	54	81	36			
7	F	13	9	117	169	81			
8	G	11	7	77	121	49			
9	H	8	5	40	64	25			
10									
11	Somma	86	55	611	956	393			
12	Media	10,75	6,875						
13									
14		b	0,626984		0,626984				
15		a	0,134921		0,134921				
16									

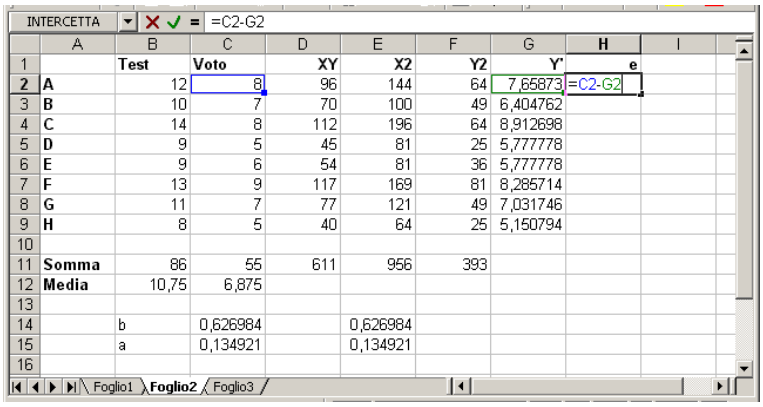
Copiare la cella in verticale per ogni valore di Y che vogliamo calcolare.

1.2.7.7 Calcolare l'errore

Per calcolare gli errori, dobbiamo fare la differenza fra il valore Y osservato e quello stimato con i parametri di regressione (Y'). Questo errore dev'essere poi elevato al quadrato per ottenere la somma degli errori al quadrato. Iniziamo ad inserire in una nuova colonna la formula per il calcolo degli errori:

```
=C2-G2
```

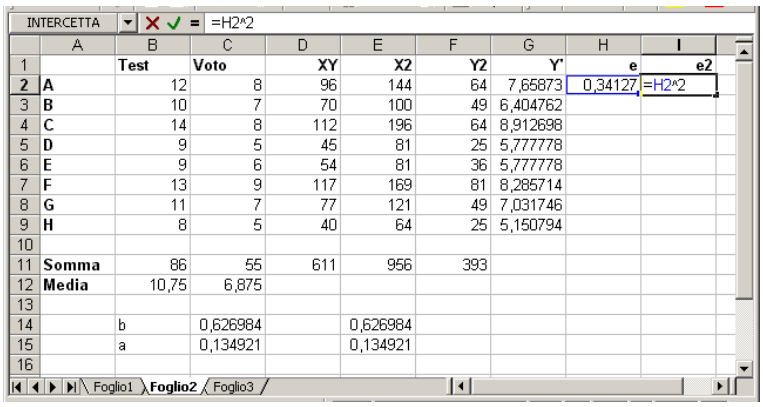
In questo esempio, C2 è il valore Y osservato per il soggetto A e G2 e il valore Y stimato, per lo stesso soggetto.



	A	B	C	D	E	F	G	H	I
1		Test	Voto	XY	X2	Y2	Y'	e	
2	A	12	8	96	144	64	7,65873	=C2-G2	
3	B	10	7	70	100	49	6,404762		
4	C	14	8	112	196	64	8,912698		
5	D	9	5	45	81	25	5,777778		
6	E	9	6	54	81	36	5,777778		
7	F	13	9	117	169	81	8,285714		
8	G	11	7	77	121	49	7,031746		
9	H	8	5	40	64	25	5,150794		
10									
11	Somma	86	55	611	956	393			
12	Media	10,75	6,875						
13									
14		b	0,626984		0,626984				
15		a	0,134921		0,134921				
16									

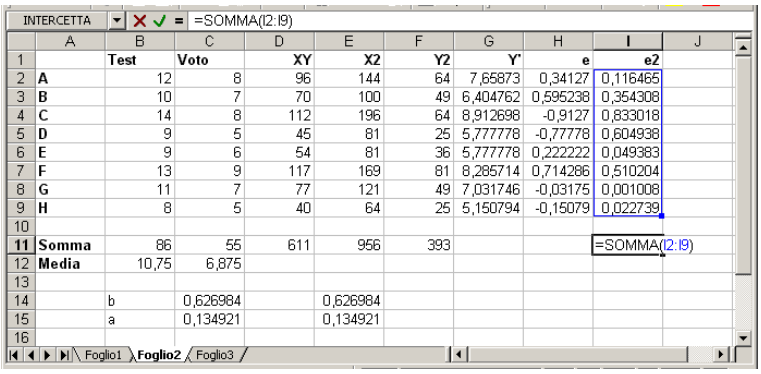
Nella cella a fianco eleviamo a quadrato:

```
=H2^2
```



	A	B	C	D	E	F	G	H	I
1		Test	Voto	XY	X2	Y2	Y'	e	e2
2	A	12	8	96	144	64	7,65873	0,34127	=H2^2
3	B	10	7	70	100	49	6,404762		
4	C	14	8	112	196	64	8,912698		
5	D	9	5	45	81	25	5,777778		
6	E	9	6	54	81	36	5,777778		
7	F	13	9	117	169	81	8,285714		
8	G	11	7	77	121	49	7,031746		
9	H	8	5	40	64	25	5,150794		
10									
11	Somma	86	55	611	956	393			
12	Media	10,75	6,875						
13									
14		b	0,626984		0,626984				
15		a	0,134921		0,134921				
16									

Quindi copiamo in verticale e facciamo la somma.



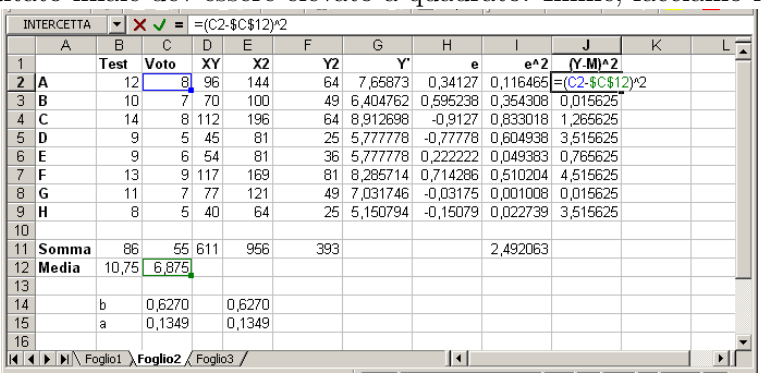
	A	B	C	D	E	F	G	H	I	J
1		Test	Voto	XY	X2	Y2	Y'	e	e2	
2	A	12	8	96	144	64	7,65873	0,34127	0,116465	
3	B	10	7	70	100	49	6,404762	0,595238	0,354308	
4	C	14	8	112	196	64	8,912698	-0,9127	0,833018	
5	D	9	5	45	81	25	5,777778	-0,77778	0,604938	
6	E	9	6	54	81	36	5,777778	0,222222	0,049383	
7	F	13	9	117	169	81	8,285714	0,714286	0,510204	
8	G	11	7	77	121	49	7,031746	-0,03175	0,001008	
9	H	8	5	40	64	25	5,150794	-0,15079	0,022739	
10										
11	Somma	86	55	611	956	393			=SOMMA(I2:I9)	
12	Media	10,75	6,875							
13										
14		b	0,626984		0,626984					
15		a	0,134921		0,134921					
16										

1.2.7.8 Calcolare la proporzione di varianza spiegata

Adesso ci serve di calcolare l'errore previsto, ovvero gli scarti di Y dalla media di Y. Con la somma degli errori spiegati, possiamo calcolare la proporzione di varianza spiegata.

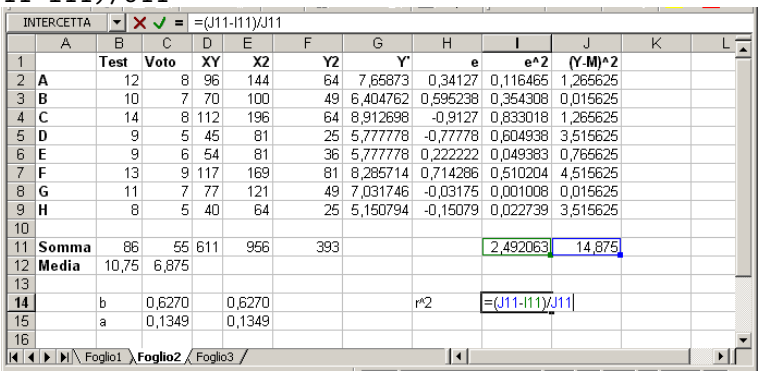
`=(C2-$C$12)^2`

In questo caso, C2 è il valore Y del primo soggetto, mentre C12 punta alla media di Y ed è scritta in forma assoluta perché Excel non dovrà aggiustarla durante la copia della cella. Il risultato finale dev'essere elevato a quadrato. Infine, facciamo la somma: `=SOMMA(J2:J9)`



	A	B	C	D	E	F	G	H	I	J	K	L
1		Test	Voto	XY	X2	Y2	Y'	e	e^2	(Y-M)^2		
2	A	12	8	96	144	64	7,65873	0,34127	0,116465	=(C2-C12)^2		
3	B	10	7	70	100	49	6,404762	0,595238	0,354308	0,015625		
4	C	14	8	112	196	64	8,912698	-0,9127	0,833018	1,265625		
5	D	9	5	45	81	25	5,777778	-0,77778	0,604938	3,515625		
6	E	9	6	54	81	36	5,777778	0,222222	0,049383	0,765625		
7	F	13	9	117	169	81	8,285714	0,714286	0,510204	4,515625		
8	G	11	7	77	121	49	7,031746	-0,03175	0,001008	0,015625		
9	H	8	5	40	64	25	5,150794	-0,15079	0,022739	3,515625		
10												
11	Somma	86	55	611	956	393			2,492063			
12	Media	10,75	6,875									
13												
14		b	0,6270		0,6270							
15		a	0,1349		0,1349							
16												

Adesso possiamo fare la divisione che ci produce la proporzione di varianza spiegata: `=(J11-I11)/J11`



	A	B	C	D	E	F	G	H	I	J	K	L
1		Test	Voto	XY	X2	Y2	Y'	e	e^2	(Y-M)^2		
2	A	12	8	96	144	64	7,65873	0,34127	0,116465	1,265625		
3	B	10	7	70	100	49	6,404762	0,595238	0,354308	0,015625		
4	C	14	8	112	196	64	8,912698	-0,9127	0,833018	1,265625		
5	D	9	5	45	81	25	5,777778	-0,77778	0,604938	3,515625		
6	E	9	6	54	81	36	5,777778	0,222222	0,049383	0,765625		
7	F	13	9	117	169	81	8,285714	0,714286	0,510204	4,515625		
8	G	11	7	77	121	49	7,031746	-0,03175	0,001008	0,015625		
9	H	8	5	40	64	25	5,150794	-0,15079	0,022739	3,515625		
10												
11	Somma	86	55	611	956	393			2,492063	14,875		
12	Media	10,75	6,875									
13												
14		b	0,6270		0,6270			r^2	=(J11-I11)/J11			
15		a	0,1349		0,1349							
16												

Riepilogando, la nostra regressione ha prodotto i seguenti risultati:

- pendenza = b = 0.6270 (arrotondato)
- intercetta = a = 0.1349 (arrotondato)
- r^2= .83 (arrotondato)

La retta di regressione dovrebbe essere scritta come:
$Y' = 0.1349 + 0.6270 * X$

1.2.8 Espressione matriciale

Per la forma matriciale, teniamo presente che l'equazione

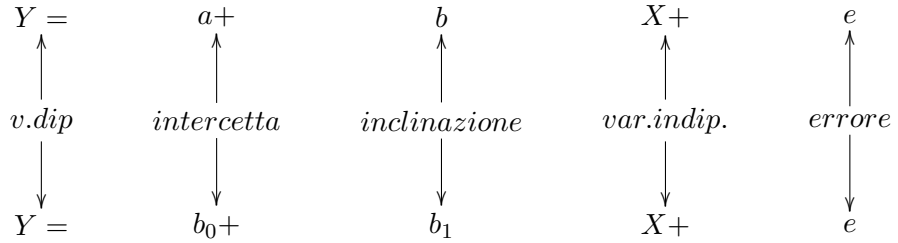
$$Y = a + bX + e$$

viene invece scritta come

$$Y = b_0 + b_1 X + e \tag{1.28}$$

(v. Fig. 12 per un confronto)

Figura 12: Confronto con la notazione matriciale



Immaginiamo di avere due variabili con 5 osservazioni ciascuna (prime 2 colonne di Tab. 11). Quelle 5 osservazioni, se vengono sostituite nell'equazione 1.28 diventeranno (terza colonna di Tab. 11): la prima colonna corrisponde alla variabile Y, la seconda colonna contiene l'intercetta (b_0) moltiplicata per 1 (cioè una costante unitaria), la terza colonna contiene la pendenza (b_1) moltiplicata per la variabile X, infine l'ultima colonna contiene l'errore da aggiungere ad ogni stima per ricostruire esattamente il valore Y osservato.

Tabella 11: Dati grezzi

Y	X	$Y = b_0 + b_1 X + e$
3	2	$3 = b_0 + b_1 3 + e_1$
2	3	$2 = b_0 + b_1 2 + e_2$
4	5	$4 = b_0 + b_1 5 + e_3$
5	7	$5 = b_0 + b_1 7 + e_4$
8	8	$8 = b_0 + b_1 8 + e_5$

Mentre b_1 è un moltiplicatore per un valore X, b_0 sembra essere un valore a se stante. Possiamo però ipotizzare che anche b_0 sia un moltiplicatore di una costante, il valore 1.

In tal modo il blocco centrale può essere pensato come il risultato del prodotto di una matrice ($\mathbf{X}$) per un vettore ($\mathbf{b}$), cioè una combinazione lineare:

$$\begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1b_0 + 2b_1 \\ 1b_0 + 3b_1 \\ 1b_0 + 5b_1 \\ 1b_0 + 7b_1 \\ 1b_0 + 8b_1 \end{bmatrix}$$

Immaginiamo a questo punto di trasformare queste combinazioni lineari in matrici o vettori. La variabile Y diventerà un vettore $\mathbf{y}$ di ordine 5, i pesi b_0 e b_1 un vettore $\mathbf{b}$ di ordine 2, l'errore

un vettore **e** di ordine 5 ed infine la variabile X e la costante unitaria vanno a formare la matrice **X** di ordine 5×2. Possiamo scriverle tutte assieme così (forma esplicita o espansa e forma compatta):

$$\begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$\begin{array}{ccccccc} 5 \times 1 & & 5 \times 2 & & 2 \times 1 & & 5 \times 1 \\ \mathbf{y} & = & \mathbf{X} & & \mathbf{b} & + & \mathbf{e} \end{array}$$

Al posto di **b** usiamo β e riscriviamo l'equazione in forma matriciale compatta:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \tag{1.29}$$

A questo punto, sappiamo che il vettore **e** verrà calcolato alla fine come sottrazione di $\hat{\mathbf{y}}$ da **y**, per cui l'unica altra incognita è il vettore β dei parametri di regressione che si trova con:

$$\beta = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y} \tag{1.30}$$

Se applichiamo questa formula ai dati di Tab. 11 avremo le formule dell'equazione precedente che vengono esplicitate:

$$\left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix}$$

Risolvendo, calcoliamo le due parti separatamente:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 5 & 25 \\ 25 & 151 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 22 \\ 131 \end{bmatrix}$$

Calcoliamo l'inversa di **X'X**. Partiamo dal determinante:

$$\begin{vmatrix} 5 & 25 \\ 25 & 151 \end{vmatrix} = 5 \times 151 - 25 \times 25 = 130$$

quindi i cofattori

(la trasposta non serve perché è simmetrica)

e la divisione per il determinante:

$$\begin{bmatrix} 151 & -25 \\ -25 & 5 \end{bmatrix} \Longrightarrow \begin{bmatrix} 151/130 & -25/130 \\ -25/130 & 5/130 \end{bmatrix}$$

Infine troviamo i parametri:

$$\frac{1}{130}\begin{bmatrix} 151 & -25 \\ -25 & 5 \end{bmatrix}\begin{bmatrix} 22 \\ 131 \end{bmatrix} = \begin{bmatrix} \frac{151 \times 22 - 25 \times 131}{130} \\ \frac{-25 \times 22 + 5 \times 131}{130} \end{bmatrix} = \begin{bmatrix} 0.362 \\ 0.807 \end{bmatrix}$$

Il vettore dei parametri β ci indica rispettivamente l'intercetta (b_0=0.362) e la pendenza ($b_1 = 0.807$).

Notate come il prodotto **X'X** e il prodotto **X'y** implicano tutti i dati che ci interessano (fate riferimento all'equazione 1.18 e al paragrafo **??**):

$$\mathbf{X'X} = \begin{bmatrix} N & \sum X \\ \sum X & \sum X^2 \end{bmatrix} = \begin{bmatrix} 5 & 25 \\ 25 & 151 \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix} = \begin{bmatrix} 22 \\ 131 \end{bmatrix}$$

Se applichiamo le formule non matriciali a questi stessi dati, otteniamo gli stessi valori:

$$b = b_1 = \frac{5 \times 131 - 25 \times 22}{5 \times 151 - 25^2} = \frac{655 - 550}{755 - 625} = \frac{105}{130} = 0.807$$

e (esprimendo le medie come *somma/N*)

$$a = b_0 = \frac{22}{5} - 0.807 \times \frac{25}{5} = 0.362$$

che sono gli stessi valori calcolati con le formule matriciali.

Applichiamo le formule matriciali ai dati di Tab. 6. Dal momento che abbiamo già calcolato le sommatorie, possiamo usare direttamente **X'X** e **X'y**.

$$\begin{bmatrix} 8 & 86 \\ 86 & 956 \end{bmatrix}^{-1}\begin{bmatrix} 55 \\ 611 \end{bmatrix}$$

Per l'inversione della matrice ci serve di calcolare il determinante, $8 \times 956 - 86 \times 86 = 252$. Facciamo poi il cofattore:

$$\frac{1}{252}\begin{bmatrix} 956 & -86 \\ -86 & 8 \end{bmatrix}^{-1}\begin{bmatrix} 55 \\ 611 \end{bmatrix}$$

e risolviamo:

$$\begin{bmatrix} \frac{956 \times 55 - 86 \times 611}{252} \\ \frac{-86 \times 55 + 8 \times 611}{252} \end{bmatrix} = \begin{bmatrix} 0,135 \\ 0,627 \end{bmatrix}$$

Osservate come il determinante coincide con il denominatore delle formule per il calcolo di b e di come coincidono perfettamente il parametro b_0 e b_1.

Possiamo applicare le stesse formule ai dati espressi come scarti dalla media (Tab. 8) e come punti z (Tab. 9) otteniamo rispettivamente i valori di b e b^*, mentre l'intercetta è 0 in entrambi i casi.

Quando i dati sono scarti dalla media, avremo che

$$\begin{bmatrix} N & \sum X \\ \sum X & \sum X^2 \end{bmatrix} = \begin{bmatrix} 8 & 0 \\ 0 & 31,5 \end{bmatrix} \qquad \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix} = \begin{bmatrix} 0 \\ 19,8 \end{bmatrix}$$

da cui

$$\begin{bmatrix} 8 & 0 \\ 0 & 31,5 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 19,8 \end{bmatrix} = \frac{1}{252} \begin{bmatrix} 31,5 & 0 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} 0 \\ 19,8 \end{bmatrix} = \begin{bmatrix} 0 \\ (8 \times 19,8)/251 \end{bmatrix} = \begin{bmatrix} 0 \\ 0,628 \end{bmatrix}$$

Mentre con i dati espressi come punti z

$$\begin{bmatrix} N & \sum X \\ \sum X & \sum X^2 \end{bmatrix} = \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix} \qquad \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix} = \begin{bmatrix} 0 \\ 7,3 \end{bmatrix}$$

$$\begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 7,3 \end{bmatrix} = \frac{1}{64} \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} 0 \\ 7,3 \end{bmatrix} = \begin{bmatrix} 0 \\ \dfrac{8 \times 7,3}{64} \end{bmatrix} = \begin{bmatrix} 0 \\ 0,913 \end{bmatrix}$$

1.2.9 Riassunto terminologico

La *regressione lineare semplice* implica due variabili e per questo motivo è chiamata anche *regressione bivariata*. Poiché uno dei suoi scopi è quello di prevedere dei valori ancora sconosciuti sulla base di un'esperienza passata, è anche chiamata *previsione bivariata*.

Le due variabili implicate vengono generalmente formalizzare come X o *variabile indipendente* o *predittiva* e Y o *variabile dipendente* o *predetta* o *stimata* o *criterio*.

I simboli Y' oppure $\hat{Y}$ vengono usati per indicare il valore di Y stimato (o predetto) di un certo valore X sulla base di un'equazione di regressione. La retta di regressione possiede un'intercetta (anche chiamata costante) per cui si usa il simbolo a (oppure b_0) e una pendenza (o coefficiente di regressione o coefficiente angolare) per cui si usa il simbolo b (o b_1). Se la pendenza è espressa in termini standardizzati, la si trova indicata con $b*$ o con β.

1.3 Regressione lineare multipla

La regressione lineare multipla è un'estensione di quella semplice: usa una sola variabile dipendente (Y) ma due o più variabili indipendenti ($X_1, X_2, \cdots X_n$). L'equazione generica diventa, allora:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni} + \epsilon_i$$

Ampliando i dati di Tab. 11, facciamo un esempio con due variabili indipendenti (v. Tab. 12).

Tabella 12: Dati per regressione multipla

Y	X_1	X_2	$Y = b_0 + b_1X_1 + b_2X_2 + e$
3	2	1	$3 = 1\beta_0 + 2\beta_1 + 1\beta_2 + e_1$
2	3	5	$2 = 1\beta_0 + 3\beta_1 + 5\beta_2 + e_2$
4	5	3	$4 = 1\beta_0 + 5\beta_1 + 3\beta_2 + e_3$
5	7	6	$5 = 1\beta_0 + 7\beta_1 + 6\beta_2 + e_4$
8	8	7	$8 = 1\beta_0 + 8\beta_1 + 7\beta_2 + e_5$

$$\begin{bmatrix}3\\2\\4\\5\\8\end{bmatrix} = \begin{bmatrix}1&2&1\\1&3&5\\1&5&3\\1&7&6\\1&8&7\end{bmatrix} \begin{bmatrix}\beta_0\\\beta_1\\\beta_2\end{bmatrix} + \begin{bmatrix}e_1\\e_2\\e_3\\e_4\\e_5\end{bmatrix}$$

$$\begin{matrix}5\times 1 & & 5\times 3 & & 3\times 1 & & 5\times 1\\ \mathbf{y} & = & \mathbf{X} & & \beta & + & \mathbf{e}\end{matrix}$$

In questo caso il vettore β conterrà l'intercetta, il parametro di regressione della prima X e quello della seconda X. Poiché in una regressione multipla non siamo più in un piano, ma in un iperpiano (spazio a più dimensioni), l'intercetta viene chiamata *costante* e le pendenze, semplicemente *parametri di regressione*. La costante è il valore che assume Y stimato quando tutte le X valgono zero.

Se sviluppiamo l'equazione, otterremo

$$\left(\begin{bmatrix}1&1&1&1&1\\2&3&5&7&8\\1&5&3&6&7\end{bmatrix}\begin{bmatrix}1&2&1\\1&3&5\\1&5&3\\1&7&6\\1&8&7\end{bmatrix}\right)^{-1}\begin{bmatrix}1&1&1&1&1\\2&3&5&7&8\\1&5&3&6&7\end{bmatrix}\begin{bmatrix}3\\2\\4\\5\\8\end{bmatrix}$$

$$\begin{bmatrix}5&25&22\\25&151&130\\22&130&120\end{bmatrix}^{-1}\begin{bmatrix}22\\131\\111\end{bmatrix} = \begin{bmatrix}0.50\\1\\-0.25\end{bmatrix}\begin{matrix}b_0\\b_1\\b_2\end{matrix}$$

Quindi, la nostra equazione di regressione andrà scritta come:

$$Y_i = 0.50 + 1X_{1_i} - 0.25X_{2_i} + e_i$$

Con due indipendenti i prodotti **X'X** e **X'y** diventano:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix}N & \sum X_1 & \sum X_2\\ \sum X_1 & \sum X_1^2 & \sum X_1X_2\\ \sum X_2 & \sum X_1X_2 & \sum X_2^2\end{bmatrix} \qquad \mathbf{X}'\mathbf{y} = \begin{bmatrix}\sum Y\\ \sum X_1Y\\ \sum X_2Y\end{bmatrix}$$

Più in generale, date n variabili indipendenti, **X'X** e **X'y** diventeranno:

$$\mathbf{X'X} = \begin{bmatrix} N & \sum X_1 & \cdots & \sum X_n \\ \sum X_1 & \sum X_1^2 & \cdots & \sum X_1 X_n \\ \cdots & \cdots & \cdots & \cdots \\ \sum X_n & \sum X_1 X_n & \cdots & \sum X_n^2 \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \cdots \\ \sum X_n Y \end{bmatrix}$$

1.3.1 Parametri standardizzati

Nella regressione multipla, i parametri standardizzati si ottengono con la formula

$$b^*_{yx_i} = b_{yx_i} \frac{s_{x_i}}{s_y}$$

dove x_i è di volta in volta una variabile indipendente diversa (X).

Se standardizziamo i parametri dell'esercizio precedente ($b_1 = 1$ e $b_2 = -0.25$) avremo:

$$b^*_1 = 1 \times \frac{2.54}{2.3} = 1.109 \qquad b^*_2 = -.25 \times \frac{2.408}{2.3} = -.262$$

Dovrebbe essere evidente che, nell'ambito della regressione lineare multipla, la pendenza standardizzata di una X non coincide con la correlazione fra quella indipendente e la dipendente.

I parametri standardizzati, essendo espressi come misura di s_y possono servire per vedere quale dei parametri calcolati è il più importante nello spiegare la dipendente. Nel fare questo confronto ignoriamo il segno perché questo indica soltanto l'orientamento della retta e confrontiamo invece i valori fra loro: il parametro standardizzato più elevato (in valore assoluto) è anche quello che partecipa maggiormente nella costruzione di Y ed è quindi il più importante all'interno dell'equazione.

1.3.2 Formule alternative

Se osserviamo le matrici X'X e X'y vediamo che gli elementi implicati servono anche per calcolare varianze e covarianze oppure le correlazioni. Esistono delle formule che permettono di calcolare i parametri a partire dalle matrici di varianza e covarianza o di correlazione. Sono:

$$\beta = C_{xx}^{-1} c_{xy} \quad \text{e} \quad \beta^* = R_{xx}^{-1} r_{xy}$$

dove C_{xx} indica la matrice di covarianza fra le variabili indipendenti, c_{xy} il vettore delle covarianze fra dipendente e indipendenti, R_{xx} indica la matrice di correlazione fra le variabili indipendenti e r_{xy} il vettore delle correlazioni fra dipendente e indipendenti. Nel paragrafo 1.3.3 vedremo da dove derivano queste formule.

Usando le covarianze otterremo le stime dei parametri non standardizzati, mentre usando le correlazioni otterremo la stima dei parametri standardizzati. In entrambi i casi, non verrà stimata l'intercetta o costante: nel primo caso si può calcolare successivamente con la formula $b_0 = \bar{Y} - \sum b_i \bar{X}_i$; nel secondo caso la costante è zero perché le correlazioni sono misure standardizzate e quindi, quando i dati sono espressi in punti z, la retta passa sempre per l'origine degli assi dell'iperpiano e l'intercetta è uguale a zero.

Proviamo ad esprimere i dati di Tab. 11 come matrice di varianza e covarianza e come matrice di correlazione. Per poter applicare le Eq. 1.1 e 1.3, ci servono alcuni valori già usati in precedenza:

$\sum Y =$	22	$\sum Y^2 =$	118	$\sum X_1 X_2 =$	130
$\sum X_1 =$	25	$\sum X_1^2 =$	151	$\sum Y X_1 =$	131
$\sum X_2 =$	22	$\sum X_2^2 =$	120	$\sum Y X_2 =$	111

che ci servono per calcolare

$$var(Y) = \frac{118}{5} - \left(\frac{22}{5}\right)^2 = 5.2 \qquad cov(Y, X_1) = \frac{131}{5} - \frac{22}{5}\frac{25}{5} = 4.2$$

$$var(X_1) = \frac{151}{5} - \left(\frac{25}{5}\right)^2 = 4.64 \qquad cov(Y, X_2) = \frac{111}{5} - \frac{22}{5}\frac{22}{5} = 2.84$$

$$var(X_2) = \frac{120}{5} - \left(\frac{22}{5}\right)^2 = 2.154 \quad cov(X_1, X_2) = \frac{130}{5} - \frac{25}{5}\frac{22}{5} = 4.0$$

Con l'Eq. 1.8 calcoliamo la matrice di correlazione (uso 1 per indicare X_1 e 2 per indicare X_2)

$$\begin{aligned} r_{y,1} &= 4.20/\sqrt{5.20 \times 4.24} &= .894 \\ r_{y,2} &= 2.84/\sqrt{4.64 \times 4.24} &= .640 \\ r_{1,2} &= 4.00/\sqrt{5.20 \times 4.64} &= .814 \end{aligned}$$

con cui costruiamo le due matrici, di varianza/covarianza e di correlazione:

Matrice covarianze

	Y	X_1	X_2
Y	4.24		
X_1	4.20	5.20	
X_2	2.84	4.00	4.64

Matrice correlazioni

	Y	X_1	X_2
Y	1		
X_1	.894	1	
X_2	.640	.814	1

A partire dalla prima tabella, costruiamo le due matrici che ci servono:

$$\mathbf{C}_{xx} = \begin{bmatrix} 5.20 & 4.00 \\ 4.00 & 4.64 \end{bmatrix} \qquad \mathbf{c}_{xy} = \begin{bmatrix} 4.20 \\ 2.84 \end{bmatrix}$$

e analogamente con i dati della seconda

$$\mathbf{R}_{xx} = \begin{bmatrix} 1 & .814 \\ .814 & 1 \end{bmatrix} \qquad \mathbf{r}_{xy} = \begin{bmatrix} .894 \\ .640 \end{bmatrix}$$

Usando queste matrici per calcolare i valori che ci servono, otterremo:

$$\begin{bmatrix} 5.20 & 4.00 \\ 4.00 & 4.64 \end{bmatrix}^{-1} \begin{bmatrix} 4.20 \\ 2.84 \end{bmatrix} = \frac{1}{8.128} \begin{bmatrix} 4.64 & -4.00 \\ -4.00 & 5.20 \end{bmatrix} \begin{bmatrix} 4.20 \\ 2.84 \end{bmatrix} =$$

$$\begin{bmatrix} (4.64 \times 4.2 - 4 \times 2.84)/8.128 \\ (-4 \times 4.2 + 5.2 \times 2.84)/8.128 \end{bmatrix} = \begin{bmatrix} 1.00 \\ -0.25 \end{bmatrix} \begin{matrix} b_1 \\ b_2 \end{matrix}$$

$$\begin{bmatrix} 1 & .814 \\ .814 & 1 \end{bmatrix}^{-1} \begin{bmatrix} .894 \\ .640 \end{bmatrix} = \frac{1}{.337} \begin{bmatrix} 1 & -.814 \\ -.814 & 1 \end{bmatrix} \begin{bmatrix} .894 \\ .640 \end{bmatrix}$$

$$\begin{bmatrix} (1 \times .894 - .814 \times .64)/.337 \\ (-.814 \times .894 + 1 \times .64)/.337 \end{bmatrix} = \begin{bmatrix} 1.107 \\ -0.260 \end{bmatrix} \begin{matrix} b_1^* \\ b_2^* \end{matrix}$$

In questi risultati possiamo notare che uno dei parametri standardizzati supera il valore 1. Dobbiamo ricordare che, nella regressione multipla, il parametro standardizzato non coincide con la correlazione e quindi non vi è nessun motivo per cui non possa superare l'unità.

1.3.3 Percorsi causali

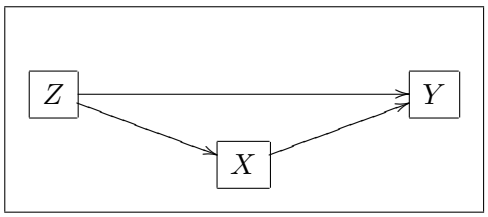


Figura 13: Percorsi causali fra 3 variabili osservate

In un grafico di modello causale, la freccia che collega una variabile ad un'altra è chiamata "influenza diretta" o *effetto diretto*, perché è l'influenza che la prima variabile ha direttamente sull'altra. Nel grafico di Fig. 13, le frecce fra Z e Y, fra Z e X e fra X e Y indicano tutte delle influenze dirette. Un'influenza diretta corrisponde al parametro di regressione calcolato sulla variabile dipendente.

Se l'influenza di una variabile su un'altra è mediata da una terza variabile, si parla invece di "influenza indiretta" e di *effetto indiretto*; ad es. nel grafico, Z ha anche un'influenza indiretta su Y tramite X. Un'influenza indiretta è pari al prodotto delle influenze semplici.

Il percorso fra due variabili è chiamato "percorso semplice" se è composto da un'influenza diretta ed è chiamato "percorso composto" se è formato da un'influenza indiretta.

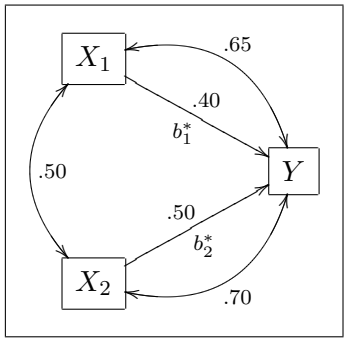


Figura 14: Relazioni causali fra 3 variabili osservate

La correlazione fra due variabili è la somma di tutte le influenze dirette e indirette che collegano fra loro le due variabili. Consideriamo la Fig. 14. In base a quello che abbiamo detto, la correlazione fra X_1 e Y dovrebbe essere uguale alla somma di tutti gli effetti diretti ed indiretti fra le due variabili. Quindi dobbiamo sommare l'effetto diretto fra X_1 e Y con l'effetto indiretto che passa tramite la correlazione che esiste tra X_1 e X_2. Lo stesso discorso vale per la correlazione fra X_2 e Y, quindi possiamo scrivere le due equazioni (dove 1 e 2 indicano rispettivamente X_1 e X_2):

$$r_{y1} = b_1^* + b_2^* r_{12} \qquad .65 = .40 + .50 \times .50$$
$$r_{y2} = b_2^* + b_1^* r_{12} \qquad .70 = .50 + .45 \times .50$$

Riscriviamo le due equazioni precedenti in un modo leggermente diverso:

$$r_{y1} = b_1^* r_{11} + b_2^* r_{12}$$
$$r_{y2} = b_1^* r_{12} + b_2^* r_{22}$$

che può essere pensato come una combinazione lineare fra la matrice delle correlazioni fra le indipendenti e i parametri standardizzati.

$$\begin{bmatrix} r_{y1} \\ r_{y2} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{bmatrix} \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix}$$

ossia, in forma compatta, $\mathbf{r}_{yx} = \mathbf{R}_{xx}\mathbf{b}_{yx}^*$, da cui si ricava $\mathbf{b}_{yx}^* = \mathbf{R}_{xx}^{-1}\mathbf{r}_{yx}$

1.3.4 Proporzione di varianza spiegata

Nell'ambito della regressione lineare multipla, l'r^2 (spesso anche indicato come R^2) ha lo stesso significato che aveva nella regressione semplice, ma non corrisponde semplicemente al quadrato della correlazione fra X e Y; corrisponde invece al quadrato della correlazione multipla fra Y e tutte le X o al quadrato della correlazione fra Y e $\hat{Y}$.

$$r^2 = (r_{y\hat{y}})^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{\sum(Y - \bar{Y})^2 - \sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2} = \sum b_i^* r_{yx_i} \tag{1.31}$$

Da questa equazione vediamo che l'R^2 si può calcolare anche facendo la sommatoria dei prodotti di ogni parametro standardizzato con la correlazione relativa. Ovvero, per ogni variabile indipendente (X), si moltiplica il parametro standardizzato di quella X con la correlazione fra quella X e la Y, infine si sommano tutti i prodotti.

Quindi, se abbiamo due variabili indipendenti (X_1 e X_2), avremo:

$$R^2 = b_1^* r_{y1} + b_2^* r_{y2} \tag{1.32}$$

Ma anche facendo il quadrato della correlazione multipla (Eq. 1.11):

$$R^2_{y.12} = \frac{r_{y1}^2 + r_{y2}^2 - 2 r_{y1} r_{y2} r_{12}}{1 - r_{12}^2} \tag{1.33}$$

Applichiamo queste formule ai dati di Tab. 12[12], usando i dati ottenuti al paragrafo 1.3.2, vale a dire:

$$\begin{array}{lll} b_1^* = 1.107 & & r_{y1} = .894 \\ b_2^* = -0.26 & \text{e} & r_{y2} = .640 \\ & & r_{12} = .814 \end{array}$$

Usando l'eq. 1.32, avremo:

$$R^2 = 1.107 \times .894 + (-.26) \times .64 = .823$$

mentre usando l'eq. 1.33 avremo:

$$R^2 = \frac{.894^2 + .64^2 - 2 \times .894 \times .64 \times .814}{1 - .814^2} = .823$$

Poiché R^2 tende ad aumentare con il numero delle variabili indipendenti, è possibile utilizzare un valore aggiustato:

$$adjR^2 = 1 - (1 - R^2)^2 \frac{N-1}{N-K-1}$$

dove N è la numerosità del campione e K indica il numero di variabili indipendenti.

E' importante ricordare che l'R^2 ci fornisce informazioni sulla percentuale di varianza spiegata dall'intera equazione, ovvero l'effetto combinato di tutte le variabili indipendenti presenti nell'equazione, ma non ci dà informazioni sul contributo di ogni singola X. Per questo dovremo confrontare fra loro i parametri standardizzati o effettuare dei test di significatività.

1.3.5 Test di Significatività

Nella regressione semplice, la proporzione di varianza spiegata ci dava informazioni sull'unica variabile inclusa nell'equazione, mentre in una regressione multipla abbiamo almeno due variabili indipendenti. Ci serve quindi un metodo per stabilire se l'apporto delle X nello spiegare Y sia statisticamente significativo.

Innanzitutto operiamo un test globale, che include tutte le variabili indipendenti. Ipotizzando di lavorare con due sole X, un test globale porrebbe la seguente ipotesi nulla:

$$H_0 : b_1 = b_2 = 0$$

Sulla base di questa ipotesi, l'equazione generale della regressione con due variabili indipendenti, si ridurrebbe alla sola costante. Per cui il confronto avviene fra:

Ipotesi	Equazione relativa	modello	gradi di libertà
$H_0 : b_1 = b_2 = 0$	$Y = b_0 + e$	ristretto o nullo	$df_r = N - 1$
$H_1 : b_1 \neq b_2 \neq 0$	$Y = b_0 + b_1X_1 + b_2X_2 + e$	completo (full)	$df_f = N - 3$

I gradi di libertà si calcolano togliendo il numero di parametri di regressione alla numerosità del campione. Nell'esempio sopra, il modello ristretto usa un solo parametro (b_0), mentre nel modello completo ce ne sono 3 (b_0, b_1 e b_2).

Dovendo confrontare il modello nullo con un modello completo, possiamo usare una formula che si distribuisce come la statistica F di Fisher:

$$F = \frac{R_f^2/k}{(1 - R_f^2)/(N - k - 1)} \tag{1.34}$$

dove k indica il numero delle indipendenti. La statistica in questione si può usare solo quando operiamo a livello globale, ovvero quando l'R_r^2 è uguale a zero. Altrimenti si può usare la formula più generale

$$F = \frac{(R_f^2 - R_r^2)/(d_r - d_f)}{(1 - R_f^2)/d_f} \tag{1.35}$$

dove gli indici f e r fanno riferimento al modello full o al modello ristretto e d indica i gradi di libertà.

Una volta calcolata la statistica F, dovremmo consultare le tavole della distribuzione di F per vedere se, dati quei gradi di libertà e un prefissato livello α, la F è o no, significativa. In

realtà, quando si fanno regressioni multiple, si usa un programma statistico e questo fornisce normalmente in automatico la probabilità associata alla F, quindi non è necessario consultare le tavole.

Se la statistica F è significativa, significa che l'apporto delle variabili indipendenti prese in considerazione è consistente e quindi c'è una relazione fra le X e la Y, ovvero la regressione ha senso.

Applichiamo le formule 1.34 e 1.35 ai dati già usati in precedenza per verificare il modello globale:

$$F = \frac{.823/2}{(1-.823)/(5-2-1)} = \frac{.412}{.323/2} = .637$$

$$F = \frac{(.823-0)/(4-2)}{(1-.823)/2} = \frac{.412}{.323/2} = .637$$

Se il modello globale è significativo, si può passare a verificare l'apporto di ogni singola indipendente ovvero si può fare un test per ciascuna delle X. Infatti, anche se il modello globale è significativo, questo non significa che tutte le X presenti nell'equazione siano a loro volta significativamente associate a Y.

Per questa verifica, si può usare sia la statistica F, sia un confronto tramite t di Student. La maggior parte dei programmi statistici utilizza proprio un semplice t-test. Se la statistica è significativa, la X_i può stare nel modello, altrimenti si dovrebbe toglierla in quanto non contribuisce a spiegare la dipendente.

Se applichiamo questo meccanismo (tramite la statistica F) alle singole variabile indipendenti dell'esempio precedente. Costruiamo le ipotesi nulle e l'equazione ridotta per ognuna delle due X.

X_1	X_2
$H_0 : b_1 = 0$	$H_0 : b_2 = 0$
$Y = b_0 + b_2X_2 + e$	$Y = b_0 + b_1X_1 + e$

Adesso, consideriamo che le equazioni ridotte corrispondono a delle regressioni semplici e quindi l'R^2 corrisponde al quadrato della regressione (oppure, il parametro standardizzato moltiplicato per la correlazione, che è la stessa cosa) e quindi:

$R_r^2 = .640^2$	$R_r^2 = .894^2$
$d_r = 5-2 = 3$	$d_r = 5-2 = 3$
$F = \frac{(.823-.64^2)/(3-2)}{(1-.823)/2}$	$F = \frac{(.823-.894^2)/(3-2)}{(1-.823)/2}$

1.3.6 [Residui]

1.4 Esercizi

1. Uno psicologo dello sport, lavorando con atleti in un particolare sport ha trovato che il punteggio su un test di conoscenza della fisiologia correla -.40 con il numero di incidenti nell'anno successivo. A questo punto lo psicologo pensa di usare una regressione semplice per predire il numero di incidenti sulla base del punteggio del test.

   a) Qual è la variabile predittrice?

b) Qual è la variabile dipendente?

c) Qual è la pendenza standardizzata?

d) Scrivi il modello predittivo standardizzato

e) Un atleta che ha un punteggio standard sul test di fisiologia pari a -2, quanti incidenti dovrebbe subire?

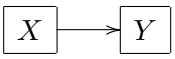
2. Con i dati di Tab. 5 calcola i parametri di regressione usando la procedura matriciale.

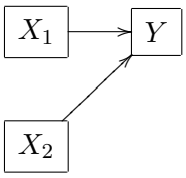
3. Usando i dati della tabella che segue, calcola la pendenza, l'intercetta, la pendenza standardizzata e la percentuale di varianza spiegata.

X	2	3	4	6	7	8	9	10	11	12	13
Y	3	6	8	4	10	14	8	12	14	12	16

4. Quale sarebbe il valore stimato di Y per $X = 5$? Quale sarebbe la sua deviazione standard?

5. Quale delle due rappresentazioni si può utilizzare per indicare una regressione multipla?

a) 

b) 

6. Cosa rappresenta R^2 in una regressione?

7. Cosa significa $\mathbf{a}'\mathbf{a}$?

8. Che cosa è $\sum(\hat{Y} - \bar{Y})^2$?

9. Nella formula che segue c'è un errore. Quale?

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + e$$

10. Qual è la differenza fra le due equazioni seguenti?

$$Y = a + bX + e \quad \hat{Y} = a + bX$$

11. In una regressione abbiamo ottenuto i seguenti pesi standardizzati. Qual è il più importante e quale il meno importante?

	Beta
X_1	.653
X_2	.219
X_3	.266

12. Usando la matrice di varianza/covarianza che segue, calcola i parametri della regressione lineare multipla

	Y	X_1	X_2
Y	15.8	10.16	12.43
X_1	10.16	11.02	9.23
X_2	12.43	9.23	15.37

13. Con la stessa matrice dell'esercizio precedente, genera la matrice di correlazione.

14. Con la matrice di correlazione calcolata all'esercizio precedente, stima i parametri standardizzati della regressione.

15. Con i dati che seguono, scrivi la sequenza di equazioni matriciali necessarie per calcolare i coefficienti di regressione:

Y	X_1	X_2
2	1	1
3	2	6
2	3	3
2	4	2
4	5	2
3	1	1

16. Usando i seguenti valori dei coefficienti di regressione, scrivi l'equazione di regressione:

	B
costante	−12.30
X_1	5.73
X_2	2.04
X_3	−0.12

17. Nell'equazione che segue indica il significato di ogni componente:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

18. Usando la matrice **X** che segue, e facendo **X'X**, indica in formule il contenuto della matrice che risulta.

$$\begin{bmatrix} 1 & 3 & 7 \\ 1 & 2 & 12 \\ 1 & 4 & 25 \\ 1 & 3 & 16 \end{bmatrix}$$

19. Se in una regressione abbiamo 3 variabili X e una variabile Y, a cosa corrisponderanno le celle delle matrici/vettori risultanti da **X'X** e da **X'y**?

20. Dopo aver effettuato una regressione, abbiamo trovato che:

$$var_{tot} = 25,15 \qquad var_{non\,spieg} = 3.12$$

Calcola la proporzione di varianza spiegata.

1.5 Soluzioni

1. Dal momento che lo psicologo intende usare il punteggio al test per prevedere il numero di incidenti:

   a) la variabile predittrice (o X o indipendente) è il punteggio al test

   b) la variabile dipendente (o Y) è il numero di incidenti

   c) La pendenza standardizzata (in una regressione semplice) coincide con la correlazione fra X e Y, quindi la pendenza è: $b^* = r = -.40$

   d) Il modello predittivo standardizzato diventa: $z_{\hat{Y}} = -.40 z_X$

e) Per sapere quanti incidenti dovrebbe subire un atleta che raggiunge un punteggio standardizzato al test pari a -2, dobbiamo completare l'equazione precedente ponendo $z_X = -2$. Quindi:

$$z_{\hat{Y}} = -.40(-2) = 0.8$$

Non avendo i dati grezzi, non possiamo esattamente sapere il numero di incidenti, ma solo il punto z corrispondente.

2. Per calcolare i parametri di regressione con le formule matriciali a partire dai dati grezzi, usiamo

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Cominciamo calcolando $\mathbf{X}'\mathbf{X}$:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 12 & 10 & 14 & 9 & 9 & 13 & 11 & 8 \end{bmatrix} \begin{bmatrix} 1 & 12 \\ 1 & 10 \\ 1 & 14 \\ 1 & 9 \\ 1 & 9 \\ 1 & 13 \\ 1 & 11 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 8 & 86 \\ 86 & 956 \end{bmatrix}$$

Per calcolare l'inversa, ci serve il determinante:

$$\begin{vmatrix} 8 & 86 \\ 86 & 956 \end{vmatrix} = 8 \times 956 - 86 \times 86 = 252$$

Quindi calcoliamo i cofattori, facciamo la trasposta e poi dividiamo per il determinante:

$$\begin{bmatrix} 956 & -86 \\ -86 & 8 \end{bmatrix} = \begin{bmatrix} 956/252 & -86/252 \\ -86/252 & 8/252 \end{bmatrix}$$

E ora calcoliamo $\mathbf{X}'\mathbf{y}$:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 12 & 10 & 14 & 9 & 9 & 13 & 11 & 8 \end{bmatrix} \begin{bmatrix} 8 \\ 7 \\ 8 \\ 5 \\ 6 \\ 9 \\ 7 \\ 5 \end{bmatrix} = \begin{bmatrix} 55 \\ 611 \end{bmatrix}$$

E infine la moltiplicazione finale:

$$\begin{bmatrix} 956/252 & -86/252 \\ -86/252 & 8/252 \end{bmatrix} \begin{bmatrix} 55 \\ 611 \end{bmatrix} = \begin{bmatrix} \dfrac{956 \times 55 - 86 \times 611}{252} \\ \dfrac{-86 \times 55 + 8 \times 611}{252} \end{bmatrix} = \begin{bmatrix} 0.135 \\ 0.627 \end{bmatrix}$$

3. Partiamo calcolando i vari totali che ci servono.

	X	Y	X^2	Y^2	XY
	2	3	4	9	6
	3	6	9	36	18
	4	8	16	64	32
	6	4	36	16	24
	7	10	49	100	70
	8	14	64	196	112
	9	8	81	64	72
	10	12	100	144	120
	11	14	121	196	154
	12	12	144	144	144
	13	16	169	256	208
$\sum$	85	107	793	1225	960

Adesso, con questi dati, calcoliamo la pendenza:

$$b = \frac{11 \times 960 - 85 \times 107}{11 \times 793 - (85)^2} = \frac{1465}{1498} = 0.978$$

l'intercetta

$$a = \frac{107}{11} - 0.978 \times \frac{85}{11} = 2.170$$

la pendenza standardizzata si può calcolare sia tramite la correlazione, sia tramite il rapporto delle deviazioni standard.

$$b^* = r = \frac{11 \times 960 - 85 \times 107}{\sqrt{[11 \times 793 - (85)^2][11 \times 1225 - (107)^2]}} = \frac{1465}{1742.11} = 0.841$$

Le deviazioni standard sono:

$$s_x = \sqrt{\frac{793}{11} - \frac{85}{11}} = 3.518 \qquad s_y = \sqrt{\frac{1225}{11} - \frac{107}{11}} = 4.092$$

e quindi poi

$$b^* = 0.978 \times \frac{3.518}{4.092} = 0.841$$

Infine il quadrato di r ci da la proporzione di varianza spiegata

$$r^2 = .841^2 = .707$$

4. Se $X = 5$ allora applichiamo la formula generale della regressione lineare facendo le dovute sostituzioni: $Y = 2.17 + 0.978 \times 5 = 7.06$ Per la deviazione standard dell'errore

$$S_{y.x} = 4.092\sqrt{1 - .707} = 2.215$$

5. Una regressione multipla è una regressione che utilizza più variabili indipendenti. Quindi il grafico da utilizzare è quello b).

6. R^2 rappresenta la proporzione di varianza spiegata, è il quadrato della correlazione fra X e Y in una regressione semplice, è il quadrato della correlazione multipla in una regressione multipla.

7. **a'a** è il prodotto scalare di un vettore con se stesso ed è uguale a $\sum a_i^2$. Infatti

$$\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} = a_1a_1 + a_2a_2 + \cdots + a_na_n = \sum_{i=1}^{N} a_i^2$$

8. $\sum(\hat{Y} - \bar{Y})^2$ è la somma al quadrato degli scarti fra la stima e la media e viene considerata come "varianza spiegata" (andrebbe divisa per N).

9. Se consideriamo $\hat{Y}$ allora bisogna eliminare la e finale dell'equazione. Se consideriamo la e finale, allora bisogna sostituire la $\hat{Y}$ con Y.

10. La prima è l'equazione di regressione completa per la variabile osservata Y, mentre la seconda è l'equazione per la stima di Y.

11. Dal momento che i pesi b^* sono già espressi in forma standardizzata, è sufficiente scegliere il valore più grande e quello più piccolo. La variabile più importante nell'equazione è la X_1 con un peso di .653, mentre la meno importante è la X_2 che ha un peso di .219

12. E' sufficiente usare la formula $\beta = \mathbf{C^{-1}c}$:

$$C = \begin{bmatrix} s_{x_1x_1} & s_{x_1x_2} \\ s_{x_2x_1} & s_{x_2x_2} \end{bmatrix} = \begin{bmatrix} 11.02 & 9.23 \\ 9.23 & 15.37 \end{bmatrix} \qquad c = \begin{bmatrix} s_{yx_1} \\ s_{yx_2} \end{bmatrix} = \begin{bmatrix} 10.16 \\ 12.43 \end{bmatrix}$$

$$\begin{bmatrix} 11.02 & 9.23 \\ 9.23 & 15.37 \end{bmatrix}^{-1} \begin{bmatrix} 10.16 \\ 12.43 \end{bmatrix} = \frac{1}{84.185} \begin{bmatrix} 15.37 & -9.23 \\ -9.23 & 11.02 \end{bmatrix} \begin{bmatrix} 10.16 \\ 12.43 \end{bmatrix}$$

ovvero

$$\begin{bmatrix} \frac{(15.37 \times 10.16) + (-9.23 \times 12.43)}{84.185} \\ \frac{(-9.23 \times 10.16) + (11.02 \times 12.43)}{84.185} \end{bmatrix} = \begin{bmatrix} 0.492 \\ 0.513 \end{bmatrix}$$

I due parametri trovati corrispondono a b_1 e a b_2; b_0 non si può calcolare, in questo caso, perché non conosciamo le medie delle variabili.

13. Ricordando che

$$r = \frac{cov(x, y)}{\sqrt{var(x)var(y)}}$$

possiamo calcolare le correlazioni:

$$r_{yx_1} = \frac{10.16}{\sqrt{11.02 \times 15.8}} = \frac{10.16}{13.195} = .770$$

$$r_{yx_2} = \frac{12.43}{\sqrt{15.8 \times 15.37}} = \frac{12.43}{15.584} = .798$$

$$r_{x_1x_2} = \frac{9.23}{\sqrt{11.02 \times 15.37}} = \frac{9.23}{13.015} = .709$$

e quindi scrivere la matrice

$$\mathbf{R} = \begin{bmatrix} 1 & .770 & .798 \\ .770 & 1 & .709 \\ .798 & .709 & 1 \end{bmatrix}$$

14. E' sufficiente usare la formula $\beta = \mathbf{R^{-1}r}$, quindi:

$$\begin{bmatrix} r_{x_1x_1} & r_{x_1x_2} \\ r_{x_2x_1} & r_{x_2x_2} \end{bmatrix}^{-1} \begin{bmatrix} r_{yx_1} \\ r_{yx_2} \end{bmatrix} = \begin{bmatrix} 1 & .709 \\ .709 & 1 \end{bmatrix}^{-1} \begin{bmatrix} .770 \\ .798 \end{bmatrix}$$

Il determinante è pari a

$$1 - .709^2 = .497$$

, per cui l'inversa diventa:

$$\frac{1}{0.497} \begin{bmatrix} 1 & -.709 \\ -.709 & 1 \end{bmatrix}$$

Risolvendo i calcoli otteniamo

$$\begin{bmatrix} \dfrac{1 \times .77 - .709 \times .798}{.497} \\ \dfrac{-.709 \times .770 + 1 \times .798}{.497} \end{bmatrix} = \begin{bmatrix} .411 \\ .507 \end{bmatrix}$$

Se i parametri cercati sono quelli standardizzati, non esiste b_0.

Per verifica, partiamo dai parametri non standardizzati dell'esercizio 12 e usiamo la formula di standardizzazione:

$$b^*_{yi} = b_{yi} \frac{s_i}{s_y}$$

Dal momento che conosciamo le varianze, ma non le deviazioni standard, facciamo la radice quadrata delle varianze:

$$b^*_1 = .492 \frac{\sqrt{11.02}}{\sqrt{15.8}} = .411$$

$$b^*_2 = .513 \frac{\sqrt{15.37}}{\sqrt{15.8}} = .506$$

che, salvo il terzo decimale, sono analoghi a quelli calcolati in precedenza.

15. Ricordiamo che l'equazione di regressione per il calcolo dei pesi beta in forma matriciale (a partire dai dati grezzi) è: $\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Impostiamo le sequenze di matrici.

$$\left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 1 \\ 1 & 6 & 3 & 2 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 6 \\ 1 & 3 & 3 \\ 1 & 4 & 2 \\ 1 & 5 & 2 \\ 1 & 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 1 \\ 1 & 6 & 3 & 2 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 2 \\ 2 \\ 4 \\ 3 \end{bmatrix}$$

16.

$$\hat{Y} = -12.3 + 5.73X_1 + 2.04X_2 - 0.12X_3$$

17. $\mathbf{y}$ è il vettore della variabile dipendente; $\mathbf{X}$ è la matrice dei dati con in aggiunta il vettore unitario davanti; β è il vettore dei pesi da stimare, uno per ogni variabile indipendente più uno per la costante; ε è il vettore degli errori di stima.

18.

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 3 & 2 & 4 & 3 \\ 7 & 12 & 25 & 16 \end{bmatrix} \begin{bmatrix} 1 & 3 & 7 \\ 1 & 2 & 12 \\ 1 & 4 & 25 \\ 1 & 3 & 16 \end{bmatrix} = \begin{bmatrix} N & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 \end{bmatrix}$$

19.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} N & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 \end{bmatrix} \qquad \mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum y \\ \sum x_1 y \\ \sum x_2 y \\ \sum x_3 y \end{bmatrix}$$

20.

$$\frac{25.15 - 3.12}{25.15} = 0,876$$

1.6 Fonti

Runyon e Haber (1976, cap. 9), Zuliani (1976, pp. 110-123, 268-285), Achen (1982), Schroeder, Sjoquist, e Stephan (1986), Tabachnick e Fidell (1989, cap. 5), Visauta Vinacua e Batallé Descals (1991, capp. 16-17, 19), Stevens (1992, cap. 3), Areni, Ercolani, e Scalisi (1994, cap. 6), Ortalda (1998, pp. 202-230), Luccio (1996, pp. 140-146), Allen (1997), Tacq (1997, pp. 99-139), Minium, Clarke, e Coladarci (1999, cap. 8).

A Appendice

A.1 [Dimostrazioni]

A.1.1 Derivazione di b e a]

Ci sono diverse dimostrazioni, la più adeguata richiede l'uso delle derivate. Quella che utilizzerò qui, utilizza un'altra strada. Il punto di partenza è l'equazione (Eq. 1.14) dell'errore come differenza fra il valore reale e quello stimato integrata con l'equazione della retta.

$$Y - \hat{Y} = Y - (a + bX)$$

Al secondo membro dell'eguaglianza abbiamo due incognite (a e b) e quindi servono due equazioni per poter risolvere il sistema. Iniziamo con l'equazione $Y = a + bX$ in cui moltiplichiamo entrambi i membri per X,

$$XY = (a + bX)X = aX + bX^2$$

. Successivamente sommiamo entrambe le equazioni per ogni valore di X e Y.

$$\sum Y = aN + B\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

Usando la prima delle due equazioni, ricaviamo a:

$$a = \frac{\sum Y - b\sum X}{N} \tag{A.1}$$

che sostituiamo nella seconda

$$\sum XY = \frac{\sum Y - b\sum X}{N}\sum X + b\sum X^2$$

e sviluppando

$$\sum XY = \frac{\sum X\sum Y - b(\sum X)^2}{N} + b\sum X^2$$

moltiplichiamo tutto per N (eliminando così il denominatore)

$$N\sum XY = -b(\sum X)^2 + Nb\sum X^2$$

raccogliamo b e spostiamo un elemento

$$N\sum XY - \sum X\sum Y = b[N\sum X^2 - (\sum X)^2]$$

e adesso ricaviamo b

$$b = \frac{N\sum XY - \sum X\sum Y}{N\sum X^2 - (\sum X)^2}$$

Se torniamo all'Eq. A.1 possiamo spezzare le due parti dell'equazione e sostituire con la media

$$a = \frac{\sum Y - b\sum X}{N} = \frac{\sum Y}{N} - \frac{b\sum X}{N} = \bar{Y} - b\bar{X}$$

A.1.2 [Dimostrazione 2]

[Inserire la dimostrazione della varianza]

Poiché $\hat{Y} = a + bX$ e $\bar{Y} = a + b\bar{X}$ allora:

$$(\hat{Y} - \bar{Y}) = a + bX - (a + b\bar{X}) = b(X - \bar{X})$$

Esprimendo X e Y come scarti dalla media (usando x e y, per convenzione) e considerando che, in tal modo,

$$b = \frac{\sum xy}{\sum x2}$$

Bibliografia

Achen, C. H. (1982). *Interpreting and using regression.* Newbury Park-London: Sage Publications.

Allen, M. P. (1997). *Understanding regression analysis.* New York-London: Plenum Press.

Areni, A., Ercolani, A. P., & Scalisi, T. G. (1994). *Introduzione all'uso della statistica in psicologia.* Milano: LED.

Luccio, R. (1996). *Tecniche di ricerca e analisi dei dati in psicologia.* Bologna: Il Mulino.

Minium, E., Clarke, R. C., & Coladarci, T. (1999). *Element of statistical reasoning* (2nd ed.). New York: Wiley & Sons.

Ortalda, F. (1998). *La survey in psicologia.* Roma: Carocci.

Runyon, R. P., & Haber, A. (1976). *Fundamentals of behavioral statistics* (3rd ed.). Reading, MA: Addison-Wesley. (Trad. it. Fondamenti di statistica per le scienze del comportamento. Amsterdam: InterEuropean Editions.)

Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). *Understanding regression analysis: An introductory guide.* Newbury Park-London: Sage Publications.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: HarperCollins Publishers.

Tacq, J. (1997). *Multivariate analysis technique in social science research. Fromn problem to analysis.* London: Sage Publications.

Visauta Vinacua, B., & Batallé Descals, P. (1991). *Métodos estadísticos aplicados. Tomo I: Estadística descriptiva.* Barcelona: PPU.

Zuliani, A. (1976). *Statistica per la ricerca educativa.* Torino: SEI.