Capitolo I

Regressione e correlazione nel caso di tre variabili

1. *Introduzione*

Su ognuna delle N unità statistiche di una popolazione sono stati rilevati contemporaneamente i valori di tre caratteri quantitativi X_1, X_2 e X_3.

I valori ottenuti con la rilevazione sono rappresentabili nel prospetto che segue:

$$\begin{array}{ccc} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \\ \vdots & \vdots & \vdots \\ x_{1i} & x_{2i} & x_{3i} \\ \vdots & \vdots & \vdots \\ x_{1N} & x_{2N} & x_{3N} . \end{array}$$

La generica terna ordinata (x_{1i}, x_{2i}, x_{3i}) può essere rappresentata con un punto in uno spazio R^3 a tre dimensioni. L'insieme delle N terne ordinate sarà rappresentato da N punti in R^3.

Si supponga ora che la variabile X_1 "dipenda" congiuntamente dalle variabili X_2 e X_3. In altre parole si ha motivo di ritenere che i valori assunti da X_2 e da X_3 "influenzino" i valori della variabile X_1. Si è allora alla ricerca di un modello matematico

$$\hat{X}_1 = f(X_2, X_3) \tag{1.1}$$

in grado di approssimare i valori x_{1i} con i valori

$$\hat{x}_{1i} = f(x_{2i}, x_{3i}) \tag{1.2}$$

forniti dal modello.

Al variare di (X_2, X_3) la (1.1) descrive nello spazio R^3 una superficie.

Lo scopo principale di questo capitolo è quello di individuare un modello che descriva bene la variabile X_1 in funzione delle altre due.

In alcuni contesti si è invece interessati ad esaminare la correlazione che esiste fra due delle tre variabili al netto dell'influenza del terzo carattere. Si tratta di un aspetto molto importante nello studio di fenomeni economici in cui spesso la correlazione che si registra fra due variabili è influenzata dalla concomitante presenza di una terza variabile.

In questo capitolo si imparerà anche a eliminare l'effetto che la terza variabile ha sulla correlazione delle altre due.

2. *Alcuni modelli*

In questo paragrafo si analizzeranno alcuni modelli utili per rappresentare X_1 in funzione delle altre due variabili.

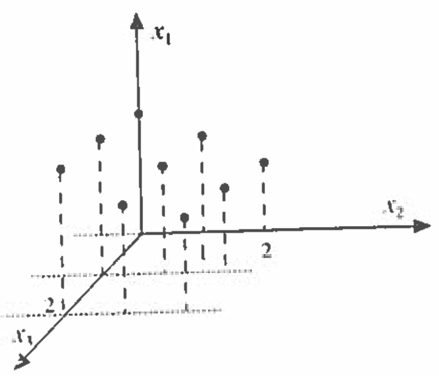
Esempio 2.1.

Si supponga che la (1.1) sia data da

$$\hat{X}_1 = 4 - X_2 + 0{,}5X_3,$$

che è l'equazione di un piano. Nel prospetto che segue sono riportate le coordinate di alcuni punti del predetto piano.

Figura 2.1. – *Rappresentazione di 9 punti del piano* $\hat{x}_1 = 4 - x_2 + x_3$



x_{2i}	x_{3i}	$\hat{x}_{1i}$
0	0	4
1	0	3
2	0	2
0	1	4,5
0	2	5,0
1	1	3,5
1	2	4,0
2	1	2,5
2	2	3,0

Dal grafico si osservi che per $x_2 = 0$ e $x_3 = 0$ si ha $\hat{x}_1 = 4$; inoltre, tenendo fisso il valore di X_2 ed incrementando di una unità il valore di X_3, il valore di $\hat{X}_1$ presenta sempre un incremento pari a 0,5 (indipendentemente dai valori di X_2 e X_3).

In modo analogo si osserva che, tenendo fisso il valore di X_3 ed incrementando di una unità il valore di X_2, il valore di $\hat{X}_1$ subisce una diminuzione pari ad 1.

È ora immediato rendersi conto che dato un generico piano di equazione

$$\hat{X}_1 = a + bX_2 + cX_3 \tag{2.1}$$

i) il parametro a indica il valore che assume $\hat{X}_1$ per $X_2 = 0$ e $X_3 = 0$ e viene detto intercetta del piano;

ii) il parametro b indica la variazione che subisce $\hat{X}_1$ allorquando, tenendo fisso X_3, si incrementa X_2 di una unità. Il parametro b è detto coefficiente angolare di $\hat{X}_1$ rispetto a X_2;

iii) il parametro c indica la variazione che subisce $\hat{X}_1$ allorquando, tenendo fisso X_2, si incrementa X_3 di una unita. Il parametro c è detto coefficiente angolare di $\hat{X}_1$ rispetto a X_3.

Esempio 2.2.

Si supponga ora che la (1.1) sia

$$\hat{X}_1 = 4 - X_2 + 0{,}5X_3 - 0{,}25X_2X_3.$$

Anche questo modello rappresenta in R^3 una superficie. La Figura 2.2 riporta le coordinate di alcuni punti della predetta superficie.

Si osservi che nel modello in esame se si tiene fisso X_3 la variabile $\hat{X}_1$ registra variazioni "costanti" facendo aumentare di una unità X_2. Tuttavia, queste variazioni di $\hat{X}_1$, pur essendo "costanti", dipendono dal valore assegnato ad X_3.

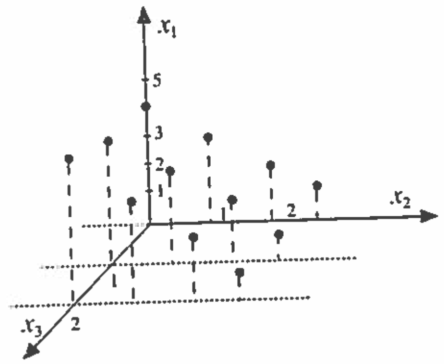
Si preciserà meglio quanto affermato considerando il modello

$$\hat{X}_1 = a + bX_2 + cX_3 + dX_2 \cdot X_3. \tag{2.2}$$

Tenendo fisso X_3 ed incrementando di una unità X_2, la variazione di $\hat{X}_1$ è data dal

$$\{a + b(X_2 + 1) + cX_3 + d(X_2 + 1) \cdot X_3\} +$$
$$- \{a + bX_2 + cX_3 + dX_2X_3\} = b + dX_3.$$

Figura 2.2. – *Rappresentazione di 12 punti del piano* $\hat{x}_1 = 4 - x_2 + 0{,}5 \cdot x_3 + -0{,}25 x_2 \cdot x_3$



x_2	x_3	$\hat{x}_1$
0	0	4,00
1	0	3,00
2	0	2,00
0	1	4,50
0	2	5,00
1	1	3,25
1	2	3,50
2	1	2,00
2	2	2,00
3	0	1,00
3	1	0,75
3	2	0,50

Analogamente la variazione di $\hat{X}_1$ ad un incremento unitario di X_3, tenendo fisso X_2, è data da

$$\{a + bX_2 + c(X_3 + 1) + dX_2 \cdot (X_3 + 1)\} + \\ - \{a + bX_2 + cX_3 + dX_2X_3\} = c + dX_2.$$

In altre parole, tenendo fissa una variabile, le variazioni di $\hat{X}_1$ ad incrementi unitari dell'altra, sono pari ad una costante (b o c) più una quantità che dipende dal valore della variabile che si tiene fissa (dX_3 o dX_2).

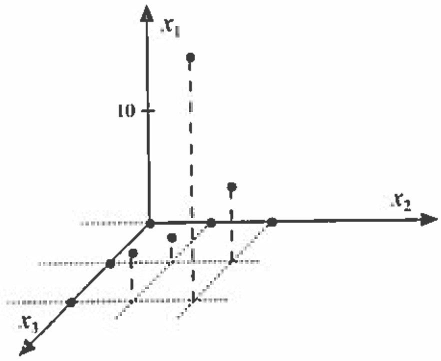
Esempio 2.3.

Si supponga ora che la (1.1) sia data da

$$\hat{X}_1 = 2 \cdot X_2^2 \cdot X_3^{1,5} \qquad (X_2 \geq 0;\ X_3 \geq 0).$$

Anche questo modello rappresenta in R^3 una superficie. Nel grafico che segue sono indicate le coordinate di alcuni punti della superficie.

Figura 2.3. – *Rappresentazione di 9 punti della superficie* $\hat{x}_1 = 2 \cdot x_2^2 \cdot x_3^{1,5}$



x_{2i}	x_{3i}	$\hat{x}_{1i}$
0	0	0
1	0	0
2	0	0
0	1	0
0	2	0
1	1	2
1	2	5,66
2	1	8,00
2	2	22,63

La funzione $\hat{X}_1 = 2 \cdot X_2^2 \cdot X_3^{1,5}$ è un caso particolare della funzione

$$\hat{X}_1 = a \cdot X_2^b \cdot X_3^c \quad (a > 0, X_2 > 0, X_3 > 0) \tag{2.3}$$

che trova vasti impieghi in economia.

Se nella (2.3) si pone $x_2 = 1$ e $x_3 = 1$ si ha $\hat{x}_1 = a$. Pertanto, il parametro a indica il valore di $\hat{x}_1$ per $x_2 = x_3 = 1$.

Il parametro b indica l'elasticità parziale di $\hat{X}_1$ rispetto a X_2. Cioè, se si tiene fissa la variabile X_3 e si fa variare X_2 allora b indica l'elasticità di $\hat{X}_1$ rispetto a X_2. In modo analogo si interpreta il parametro c. Infatti, esso indica l'elasticità parziale di $\hat{X}_1$ rispetto a X_2.

Si osservi che solitamente i valori delle elasticità parziali di un modello dipendono dal punto (X_2, X_3) in cui vengono calcolate. Fa eccezione il modello (2.3) le cui elasticità non variano al variare di (X_2, X_3). Per questo motivo il modello (2.3) è detto ad elasticità parziali costanti.

Esempio 2.4.

Rivestono una particolare importanza i modelli del tipo

$$\hat{X}_1 = f(X_2, X_3) = \sum_{j=1}^{k} \alpha_j \cdot g_j(X_1, X_3) \tag{2.4}$$

in cui $g_1(X_2, X_3) = 1$ e $g_j(X_2, X_3)$, $j = 2, 3, \ldots, k$, sono funzioni di X_2 e X_3.

Casi particolari della precedente sono

$$f(X_2, X_3) = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3$$
$$f(X_2, X_3) = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_2 X_3$$
$$f(X_2, X_3) = \alpha_1 + \alpha_2 X_2 + \alpha_3 \log X_2 + \alpha_4 X_3^2.$$

È immediato verificare che

$$f(X_2, X_3) = aX_2^b \cdot X_3^c$$

non è del tipo (2.4).

Le funzioni del tipo (2.4) sono lineari nei parametri. Ciò significa che, tenendo fissi sia i valori di X_2 e X_3 che quelli di tutti i parametri tranne uno, ad esempio α_j e, facendo aumentare quest'ultimo di una unità, la $f(X_2, X_3)$ fa riscontrare variazioni "costanti" pari a $g_j(X_2, X_3)$.

3. *Il metodo dei minimi quadrati*

Si supponga che per descrivere la variabile X_1 si sia scelto il modello

$$\hat{X}_1 = f(X_2, X_3;\ \alpha_1, \alpha_2, \ldots, \alpha_k). \tag{3.1}$$

Si tratta ora di ricavare quei particolari valori dei parametri $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_k$ che sostituiti nella (3.1) forniscono valori $\hat{x}_{1i} = f(x_{2i}, x_{3i};\ \hat{\alpha}_1, \ldots, \hat{\alpha}_k)$ assai prossimi ai valori reali x_{1i}.

Uno dei metodi più comunemente impiegati per ricavare $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_k$ è quello dei minimi quadrati, che consiste nel minimizzare la funzione

$$D(\alpha_1, \alpha_2, \ldots \alpha_k) = \sum_{i=1}^{N} \{x_{1i} - f(x_{2i}, x_{3i}; \alpha_1, \alpha_2, \ldots, \alpha_k)\}^2 \tag{3.2}$$

rispetto ai parametri $\alpha_1, \alpha_2, \ldots, \alpha_k$. Si tratta allora di ricavare innanzi tutto le k derivate parziali $\dfrac{\partial D}{\partial \alpha_j}$, $j = 1, \ldots, k$. Quindi le stesse vengono uguagliate a zero è si ottiene così un sistema di k equazioni nelle k incognite $\alpha_1, \alpha_2, \ldots, \alpha_k$:

$$\begin{cases} \dfrac{\partial D}{\partial \alpha_1} = 0 \\ \dfrac{\partial D}{\partial \alpha_2} = 0 \\ \vdots \\ \dfrac{\partial D}{\partial \alpha_k} = 0. \end{cases} \tag{3.3}$$

Si tratta poi di risolvere il sistema (3.3) e controllare che effettivamente la soluzione trovata minimizzi la (3.2).

La soluzione del sistema (3.3) è particolarmente semplice nel caso in cui $f(X_2, X_3;\ \alpha_1, \alpha_2, \ldots, \alpha_k)$ sia una funzione lineare nei parametri. Le derivate parziali in questo caso risultano

$$\begin{cases} \dfrac{\partial D}{\partial \alpha_1} = -2\sum\left(x_{1i} - f(x_{2i}, x_{3i}; \alpha_1, \ldots, \alpha_k)\right); \\ \dfrac{\partial D}{\partial \alpha_2} = -2\sum\left(x_{1i} - f(x_{2i}, x_{3i}; \alpha_1, \ldots, \alpha_k)\right) \cdot g_2(x_{2i}, x_{3i}); \\ \vdots \\ \dfrac{\partial D}{\partial \alpha_k} = -2\sum\left(x_{1i} - f(x_{2i}, x_{3i}; \alpha_1, \ldots, \alpha_k)\right) \cdot g_k(x_{2i}, x_{3i}). \end{cases}$$

Uguagliando a zero le derivate parziali si ha il seguente sistema

$$\begin{cases} \sum(x_{1i} - \hat{x}_{1i}) = 0 \\ \sum(x_{1i} - \hat{x}_{1i}) \cdot g_2(x_{2i}, x_{3i}) = 0 \\ \vdots \\ \sum(x_{1i} - \hat{x}_{1i}) \cdot g_k(x_{2i}, x_{3i}) = 0, \end{cases} \tag{3.4}$$

nel quale ovviamente è $\hat{x}_{1i} = f(x_{2i}, x_{3i}; \alpha_1, \ldots, \alpha_k)$.

Il sistema (3.4) è solitamente detto sistema normale.

4. *Il piano a minimi quadrati*

Nel caso in cui

$$\hat{x}_{1i} = \alpha_1 + \alpha_2 x_{2i} + \alpha_3 x_{3i} \tag{4.1}$$

il sistema normale diventa

$$\begin{cases} \sum(x_{1i} - \alpha_1 - \alpha_2 x_{2i} - \alpha_3 x_{3i}) = 0 \\ \sum(x_{1i} - \alpha_1 - \alpha_2 x_{2i} - \alpha_3 x_{3i}) \cdot x_{2i} = 0 \\ \sum(x_{1i} - \alpha_1 - \alpha_2 x_{2i} - \alpha_3 x_{3i}) \cdot x_{3i} = 0. \end{cases} \tag{4.2}$$

Riordinando opportunamente i valori si ottiene

$$\begin{cases} N\alpha_1 + \alpha_2 \sum x_{2i} + \alpha_3 \sum x_{3i} = \sum x_{1i} \\ \alpha_1 \sum x_{2i} + \alpha_2 \sum x_{2i}^2 + \alpha_3 \sum x_{3i} \cdot x_{2i} = \sum x_{1i} \cdot x_{2i} \\ \alpha_1 \sum x_{3i} + \alpha_2 \sum x_{2i} \cdot x_{3i} + \alpha_3 \sum x_{3i}^2 = \sum x_{1i} \cdot x_{3i}. \end{cases} \tag{4.3}$$

Si ha così un classico sistema lineare di tre equazioni nelle tre incognite α_1, α_2 e α_3 che può essere risolto ricorrendo ai consueti procedimenti della matematica.

Soluzione con il metodo dei determinanti

Si ricavano innanzi tutto il determinante del sistema Δ ed i determinanti dei parametri $\Delta_{\alpha_1}, \Delta_{\alpha_2}$ e Δ_{α_3}.

$$\Delta = \begin{vmatrix} N & \sum x_{2i} & \sum x_{3i} \\ \sum x_{2i} & \sum x_{2i}^2 & \sum x_{3i}x_{2i} \\ \sum x_{3i} & \sum x_{2i}x_{3i} & \sum x_{3i}^2 \end{vmatrix}; \qquad \Delta_{\alpha_1} = \begin{vmatrix} \sum x_{1i} & \sum x_{2i} & \sum x_{3i} \\ \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \sum x_{3i}x_{2i} \\ \sum x_{1i}x_{3i} & \sum x_{2i}x_{3i} & \sum x_{3i}^2 \end{vmatrix}$$

$$\Delta_{\alpha_2} = \begin{vmatrix} N & \sum x_{1i} & \sum x_{3i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{3i}x_{2i} \\ \sum x_{3i} & \sum x_{1i}x_{3i} & \sum x_{3i}^2 \end{vmatrix}; \qquad \Delta_{\alpha_3} = \begin{vmatrix} N & \sum x_{2i} & \sum x_{1i} \\ \sum x_{2i} & \sum x_{2i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{3i} & \sum x_{2i}x_{3i} & \sum x_{1i}x_{3i} \end{vmatrix}$$

Risolvendo i determinanti con la regola di Sarrus si ha:

$$\Delta = N\sum x_{2i}^2 \cdot \sum x_{3i}^2 + 2\sum x_{2i} \cdot \sum x_{2i}x_{3i} \sum x_{3i} - \left(\sum x_{3i}\right)^2 \cdot \sum x_{2i}^2 + \\ - N\left(\sum x_{2i}x_{3i}\right)^2 - \sum x_{3i}^2 \cdot \left(\sum x_{2i}\right)^2;$$

$$\Delta_{\alpha 1} = \sum x_{1i} \sum x_{2i}^2 \sum x_{3i}^2 + \sum x_{2i} \sum x_{3i}x_{2i} \cdot \sum x_{1i}x_{3i} + \sum x_{3i} \cdot \sum x_{1i}x_{2i} \cdot \sum x_{2i}x_{3i} + \\ - \sum x_{1i}x_{3i} \cdot \sum x_{2i}^2 \sum x_{3i} - \left(\sum x_{2i}x_{3i}\right)^2 \cdot \sum x_{1i} - \sum x_{3i}^2 \cdot \sum x_{1i}x_{2i} \cdot \sum x_{2i};$$

$$\Delta_{\alpha 2} = N\sum x_{1i}x_{2i} \sum x_{3i}^2 + \sum x_{1i} \cdot \sum x_{3i}x_{2i} \cdot \sum x_{3i} + \sum x_{3i} \sum x_{2i} \sum x_{1i}x_{3i} + \\ - \left(\sum x_{3i}\right)^2 \sum x_{1i}x_{2i} - N\sum x_{1i}x_{3i} \sum x_{3i}x_{2i} - \sum x_{3i}^2 \sum x_{2i} \cdot \sum x_{1i};$$

$$\Delta_{\alpha 3} = N\sum x_{2i}^2 \sum x_{1i}x_{3i} + \sum x_{2i} \sum x_{1i}x_{2i} \sum x_{3i} + \sum x_{1i} \sum x_{2i} \sum x_{2i}x_{3i} + \\ - \sum x_{3i} \sum x_{2i}^2 \sum x_{1i} - N\sum x_{2i}x_{3i} \sum x_{1i}x_{2i} - \sum x_{1i}x_{3i} \cdot \left(\sum x_{2i}\right)^2.$$

Nel caso in cui $\Delta \neq 0$ le soluzioni sono pari a

$$\hat{\alpha}_1 = \frac{\Delta_{\alpha_1}}{\Delta}; \quad \hat{\alpha}_2 = \frac{\Delta_{\alpha_2}}{\Delta}; \quad \hat{\alpha}_3 = \frac{\Delta_{\alpha_3}}{\Delta}. \tag{4.3}$$

Esempio 4.1.

Nel prospetto che segue sono riportati i valori di tre variabili X_2 = superficie in ettari, X_3 = numero di bovini e X_1 = reddito annuo in milioni di lire di 20 aziende agricole.

X_2	X_3	X_1	X_2^2	$X_2 \cdot X_3$	$X_1 \cdot X_2$	X_3^2	$X_1 \cdot X_3$	$\hat{X}_1$	$(X_1 - \hat{X}_1)$
6	18	96	36	108	576	324	1.728	100,0	− 4,0
22	0	83	484	0	1.826	0	0	75,6	+ 7,4
18	14	126	324	252	2.268	196	1.764	112,6	+ 13,4
8	6	61	64	48	488	36	366	65,2	− 4,2
12	1	59	144	12	708	1	59	57,5	+ 1,5
10	9	90	100	90	900	81	810	79,2	+ 10,8
17	6	82	289	102	1.394	36	492	84,1	− 2,4
11	12	88	121	132	968	144	1.056	91,2	− 3,2
16	7	86	256	112	1.376	49	602	85,6	+ 0,4
23	2	76	529	46	1.748	4	152	84,2	− 8,2
7	17	102	49	119	714	289	1.734	98,9	+ 3,1
12	15	108	144	180	1.296	225	1.620	103,1	+ 4,9
24	7	96	576	168	2.304	49	672	102,7	− 6,7
16	0	70	256	0	1.120	0	0	62,8	+ 7,2
9	12	80	81	108	720	144	960	86,9	− 6,9
11	16	113	121	176	1243	256	1808	104,2	+ 8,8
22	2	76	484	44	1672	4	152	82,1	− 6,1
11	6	74	121	66	814	36	444	71,6	+ 2,4
16	12	98	256	192	1568	144	1.176	101,8	− 3,8
8	15	80	64	120	640	225	1.200	94,5	− 14,5
279	177	1.744	4.499	2.075	24.343	2.243	16.795	1.744,1	− 60,0 + 59,9

Si devono ricavare i parametri del piano a minimi quadrati

$$\hat{X}_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3$$

Per determinare i parametri bisogna ricavare i determinanti Δ, Δ_{α_1}, Δ_{α_2} e Δ_{α_3} È allora necessario valutare le sommatorie

$$\sum x_{2i},\ \sum x_{3i},\ \sum x_{1i},\ \sum x_{2i}^2,\ \sum x_{3i}^2,\ \sum x_{3i}x_{2i},\ \sum x_{1i}x_{3i},\ \sum x_{1i}x_{2i}.$$

I valori di queste sommatorie sono riportati nel prospetto precedente. Si ha

$$\Delta = 20 \cdot 4.499 \cdot 2.243 + 2 \cdot 279 \cdot 2.075 \cdot 177 - 177^2 \cdot 4.499 - 20 \cdot 2.075^2 + \\ - 2.243 \cdot 279^2 = 5.105.556;$$

$$\Delta_{\alpha_1} = 1.744 \cdot 4.499 \cdot 2.243 + 279 \cdot 2.075 \cdot 16.795 + 177 \cdot 24.343 \cdot 2.075 + \\ - 16.795 \cdot 4.499 \cdot 177 - 2.075^2 \cdot 1.744 - 2.243 \cdot 24.343 \cdot 279 = \\ = 145.741.750;$$

$$\Delta_{\alpha_2} = 20 \cdot 24.343 \cdot 2.243 + 1.744 \cdot 2.075 \cdot 177 + 177 \cdot 279 \cdot 16.795 + \\ - (177)^2 \cdot 24.343 - 16.795 \cdot 2.075 \cdot 20 - 2.243 \cdot 279 \cdot 1.744 = \\ = 10.917.750;$$

$$\Delta_{\alpha_3} = 20 \cdot 4.499 \cdot 16.795 + 279 \cdot 24.343 \cdot 177 + 1.744 \cdot 279 \cdot 2.075 + \\ - 177 \cdot 4.499 \cdot 1.744 - 2.075 \cdot 24.343 \cdot 20 - 16.795 \cdot 279^2 = \\ = 16.628.262.$$

I valori dei tre parametri risultano

$$\hat{\alpha}_1 = 28,546; \quad \hat{\alpha}_2 = 2,138; \quad \hat{\alpha}_3 = 3,257.$$

L'equazione del piano interpolatore risulta

$$\hat{X}_1 = 28,546 + 2,138 X_2 + 3,257 X_3.$$

Esempio 4.2.

Si interpretano ora i parametri del piano interpolatore ricavato nell'Esempio 4.1. Sempre sui dati di tale esempio si valutano i redditi aziendali forniti dal piano interpolatore, nonché gli scarti residui.

L'intercetta $\hat{\alpha}_1 = 28,546$ milioni non ha un significato reale. In effetti $\hat{\alpha}_1$ indicherebbe il reddito di un'azienda agricola senza superficie e senza bovini.

Il parametro $\hat{\alpha}_2 = 2,138$ indica l'incremento di reddito che si consegue aumentando la superficie di un ettaro nell'ipotesi di tener fisso il numero di bovini. Analogamente il parametro $\hat{\alpha}_3 = 3{,}257$ indica l'incremento di reddito che si consegue aumentando di una unità il numero di bovini ipotizzando di tener fissa la superficie aziendale.

I redditi aziendali forniti dal piano interpolatore

$$\hat{x}_{1i} = 28{,}546 + 2{,}138 x_{2i} + 3{,}257 x_{3i}$$

nonché gli scarti $(x_{1i} - \hat{x}_{1i})$ sono riportati nel prospetto dell'esercizio precedente.

Da tale prospetto si deduce che l'ordine di grandezza dei residui è abbastanza limitato relativamente ai valori $\hat{X}_1$.

5. *Soluzione del sistema normale con il principio di riduzione*

(81)

È molto istruttivo risolvere il sistema (4.3) ricorrendo opportunamente al principio di riduzione.

Per rendere più snella la trattazione si farà uso della seguente simbologia.

Con $\bar{x}_1 = \frac{1}{N}\sum x_{1i}$, $\bar{x}_2 = \frac{1}{N}\sum x_{2i}$ e $\bar{x}_3 = \frac{1}{N}\sum x_{3i}$ si indicano le medie aritmetiche delle tre variabili.

Con

$$\sigma_{11} = \sigma^2(X_1) = \frac{1}{N}\sum (x_{1i} - \bar{x}_1)^2 = \frac{1}{N}\sum x_{1i}^2 - \bar{x}_1^2,$$

$$\sigma_{22} = \sigma^2(X_2) = \frac{1}{N}\sum (x_{2i} - \bar{x}_2)^2 = \frac{1}{N}\sum x_{2i}^2 - \bar{x}_2^2,$$

$$\sigma_{33} = \sigma^2(X_3) = \frac{1}{N}\sum (x_{3i} - \bar{x}_3)^2 = \frac{1}{N}\sum x_{3i}^2 - \bar{x}_3^2,$$

si indicano le varianze delle tre variabili.

Infine, con

$$\sigma_{12} = \sigma(X_1, X_2) = \frac{1}{N}\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = \frac{1}{N}\sum x_{1i}x_{2i} - \bar{x}_1\bar{x}_2,$$

$$\sigma_{13} = \sigma(X_1, X_3) = \frac{1}{N}\sum (x_{1i} - \bar{x}_1)(x_{3i} - \bar{x}_3) = \frac{1}{N}\sum x_{1i}x_{3i} - \bar{x}_1\bar{x}_3,$$

$$\sigma_{23} = \sigma(X_2, X_3) = \frac{1}{N}\sum(x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3) = \frac{1}{N}\sum x_{2i}x_{3i} - \bar{x}_2\bar{x}_3,$$

si indicano le covarianze fra i tre caratteri.

Si risolve ora il sistema (4.3) con il principio di riduzione. Dividendo per N sia i primi che i secondi membri delle equazioni del sistema (4.3) si ha

$$\begin{cases} \alpha_1 + \alpha_2\bar{x}_2 + \alpha_3\bar{x}_3 = \bar{x}_1 \\ \alpha_1\bar{x}_2 + \alpha_2 \cdot \frac{1}{N}\sum x_{2i}^2 + \alpha_3 \cdot \frac{1}{N}\sum x_{3i}x_{2i} = \frac{1}{N}\sum x_{1i}x_{2i} \\ \alpha_1\bar{x}_3 + \alpha_2 \cdot \frac{1}{N}\sum x_{2i}x_{3i} + \alpha_3 \cdot \frac{1}{N}\sum x_{3i}^2 = \frac{1}{N}\sum x_{1i}x_{3i}. \end{cases} \tag{5.1}$$

Moltiplicando la prima equazione del sistema (5.1) per $\bar{x}_2$ e sottraendo i valori trovati dalla seconda equazione si ottiene

$$\alpha_1(\bar{x}_2 - \bar{x}_2) + \alpha_2\left(\frac{1}{N}\sum x_{2i}^2 - \bar{x}_2^2\right) + \alpha_3\left(\frac{1}{N}\sum x_{3i}x_{2i} - \bar{x}_3\bar{x}_2\right) = \frac{1}{N}\sum x_{1i}x_{2i} - \bar{x}_1\bar{x}_2$$

ovvero

$$\alpha_2\sigma_{22} + \alpha_3\sigma_{32} = \sigma_{12}. \tag{5.2}$$

Analogamente, moltiplicando per $\bar{x}_3$ la prima equazione del sistema (5.1) e sottraendo i valori trovati dalla terza equazione si ottiene

$$\alpha_2\sigma_{23} + \alpha_3\sigma_{33} = \sigma_{13}. \tag{5.3}$$

Dalla prima equazione del sistema (5.1) si ricava

$$\alpha_1 = \bar{x}_1 - \alpha_2\bar{x}_2 - \alpha_3\bar{x}_3. \tag{5.4}$$

Mettendo assieme la (5.2), la (5.3) e la (5.4) si ha il seguente sistema

$$\begin{cases} \alpha_2\sigma_{22} + \alpha_3\sigma_{32} = \sigma_{12} \\ \alpha_2\sigma_{23} + \alpha_3\sigma_{33} = \sigma_{13} \\ \alpha_1 = \bar{x}_1 - \alpha_2\bar{x}_2 - \alpha_3\bar{x}_3. \end{cases} \tag{5.5}$$

Il sistema (5.5) è equivalente al sistema (4.3) e quindi le soluzioni fornite dagli stessi non possono che coincidere.

Le prime due equazioni del sistema (5.5) costituiscono un sistema di due

SOLUZIONE X SOSTITUZIONE

equazioni nelle incognite α_2 e α_3. Il determinante di tale sistema e quelli delle incognite α_2 e α_3 sono rispettivamente pari a

$$\Delta = \begin{vmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{23} & \sigma_{33} \end{vmatrix} = \sigma_{22}\sigma_{33} - \sigma_{23}^2 \tag{5.6}$$

$$\Delta_{\alpha_2} = \begin{vmatrix} \sigma_{12} & \sigma_{32} \\ \sigma_{13} & \sigma_{33} \end{vmatrix} = \sigma_{33}\sigma_{12} - \sigma_{13}\sigma_{23} \tag{5.7}$$

$$\Delta_{\alpha_3} = \begin{vmatrix} \sigma_{22} & \sigma_{12} \\ \sigma_{23} & \sigma_{13} \end{vmatrix} = \sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23}. \tag{5.8}$$

Dalla diseguaglianza di Cauchy-Schwartz è noto che

$$\Delta = Var\,(X_2) \cdot Var\,(X_3) - Cov^2(X_2, X_3) \geq 0$$

avendosi $\Delta = 0$ solo nel caso in cui fra X_2 e X_3 esista una perfetta relazione lineare.

Nell'ipotesi che sia $\Delta > 0$ i parametri α_2 e α_3 sono pari a

$$\hat{\alpha}_2 = \frac{\sigma_{33}\sigma_{12} - \sigma_{13}\sigma_{23}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2}; \tag{5.9}$$

$$\hat{\alpha}_3 = \frac{\sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2}. \tag{5.10}$$

Dalla terza equazione del sistema (5.5) si ricava poi

$$\hat{\alpha}_1 = \bar{x}_1 - \hat{\alpha}_2\bar{x}_2 - \hat{\alpha}_3\bar{x}_3. \tag{5.11}$$

Prima di passare ad un esempio conviene soffermarsi sul caso $\Delta = 0$. Il fatto che vi sia una perfetta relazione lineare fra X_2 e X_3 significa che $X_3 = a + bX_2$.

Pertanto il piano interpolatore assumerà la forma

$$\begin{aligned} \hat{X}_1 &= \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 \\ &= \alpha_1 + \alpha_2 X_2 + \alpha_3[a + bX_2] \\ &= (\alpha_1 + \alpha_3 a) + (\alpha_2 + \alpha_3 b) \cdot X_2. \end{aligned}$$

Si vede allora che se $\Delta = 0$, significa che è sufficiente far riferimento ad una sola variabile esplicativa, ad esempio X_2 e che è ridondante impiegare anche l'altra variabile esplicativa. In altre parole, per valutare $\hat{X}_1$ è sufficiente far riferimento alla retta a minimi quadrati $\hat{X}_1 = p_0 + p_1 X_2$.

Esempio 5.1.

Si riconsiderino i dati dell'Esempio 4.1 per ricavare i parametri del piano interpolatore impiegando il procedimento di riduzione esaminato in questo paragrafo.

Si tratta di ricavare innanzi tutto le medie, le varianze e le covarianze delle variabili.

$$\bar{x}_1 = \frac{1}{20} \cdot 1.744 = 87,2; \quad \bar{x}_2 = \frac{1}{20} \cdot 279 = 13,95; \quad \bar{x}_3 = \frac{1}{20} \cdot 177 = 8,85;$$

$$\sigma_{22} = Var(X_2) = \frac{1}{N} \sum x_{2i}^2 - \bar{x}_2^2 = \frac{1}{20} \cdot 4.449 - 13,95^2 = 30,3475;$$

$$\sigma_{33} = Var(X_3) = \frac{1}{N} \sum x_{3i}^2 - \bar{x}_3^2 = \frac{1}{20} \cdot 2.243 - 8,85^2 = 33,8275;$$

$$\sigma_{23} = Cov(X_2, X_3) = \frac{1}{N} \sum x_{2i} x_{3i} - \bar{x}_2 \cdot \bar{x}_3 = \frac{1}{20} \cdot 2.075 - 13,95 \cdot 8,85 = -19,7075;$$

$$\sigma_{12} = Cov(X_1, X_2) = \frac{1}{N} \sum x_{1i} x_{2i} - \bar{x}_1 \cdot \bar{x}_2 = \frac{1}{20} \cdot 24.343 - 13,95 \cdot 87,2 = 0,71;$$

$$\sigma_{13} = Cov(X_1, X_2) = \frac{1}{N} \sum x_{1i} x_{3i} - \bar{x}_1 \cdot \bar{x}_3 = \frac{1}{20} 16.795 - 8,85 \cdot 87,2 = 68,03.$$

Pertanto i valori dei parametri risultano

$$\hat{\alpha}_2 = \frac{33,8275 \cdot 0,71 - (-19,7075) \cdot 68,03}{30,3475 \cdot 33,8275 - (-19.7075)^2} = \frac{1.364,71875}{638,1945} = 2,1384;$$

$$\hat{\alpha}_3 = \frac{30,3475 \cdot 68.03 - (-19,7075) \cdot 0,71}{638,1945} = \frac{2078,53275}{638,1945} = 3,2569;$$

$$\hat{\alpha}_1 = \bar{x}_1 - \hat{\alpha}_2 \cdot \bar{x}_2 - \hat{\alpha}_3 \cdot \bar{x}_3 = 87,2 - 2,1384 \cdot 13,95 - 3,2569 \cdot 8,85 = 28,546.$$

Si riottengono così gli stessi valori calcolati in precedenza.

6. *Proprietà dei residui*

Si riscrive per comodità il sistema (4.2).

$$\begin{cases} \sum(x_{1i}-\hat{x}_{1i})=0 \\ \sum(x_{1i}-\hat{x}_{1i})\cdot x_{2i}=0 \\ \sum(x_{1i}-\hat{x}_{1i})\cdot x_{3i}=0. \end{cases} \tag{6.1}$$

Dalla prima equazione si desume che la somma dei residui $(x_{1i}-\hat{x}_{1i})$ è nulla e quindi la variabile residuo $Z = X_1 - \hat{X}_1$ ha media aritmetica nulla. Inoltre, sempre dalla prima equazione si desume che $\sum x_{1i} = \sum \hat{x}_{1i}$, ovvero, la somma dei valori effettivi x_{1i} è uguale alla somma dei valori interpolati e quindi $\bar{x}_1 = M_1(X_1) = M_1(\hat{X}_1)$, avendo indicato con $M_1(\cdot)$ l'operatore media aritmetica. Come si è mostrato in precedenza, la prima equazione del sistema (6.1) si può scrivere anche come segue:

$$\alpha_1 + \alpha_2 \bar{x}_2 + \alpha_3 \bar{x}_3 = \bar{x}_1,$$

ovvero il valore di $\hat{x}_1$ è pari a $\bar{x}_1$ quando $x_2 = \bar{x}_2$ e $x_3 = \bar{x}_3$. In altre parole, il piano interpolatore a minimi quadrati passa per il punto dello spazio R^3 di coordinate $x_1 = \bar{x}_1$, $x_2 = \bar{x}_2$ e $x_3 = \bar{x}_3$.

Si dimostrerà ora che la variabile residuo $X_1 - \hat{X}_1 = Z$ è incorrelata sia con la variabile X_2 che con la variabile X_3. A tal proposito si osservi innanzi tutto che se almeno una delle variabili X e Y ha media aritmetica nulla allora

$$Cov(X,Y) = \frac{1}{N}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum x_i y_i - \bar{x}\,\bar{y} = \frac{1}{N}\sum x_i y_i.$$

Pertanto essendo la media aritmentica dei residui $\bar{z} = 0$ risulta

$$Cov(Z,X_2) = \frac{1}{N}\sum z_i x_{2i} = \frac{1}{N}\sum(x_{1i}-\hat{x}_{1i})x_{2i}$$

e

$$Cov(Z,X_3) = \frac{1}{N}\sum z_i x_{3i} = \frac{1}{N}\sum(x_{1i}-\hat{x}_{1i})x_{3i}.$$

Dividendo ora la seconda e la terza equazione del sistema (6.1) per N si ha

$$\begin{cases} Cov[Z,X_2]=0 \\ Cov[Z,X_3]=0, \end{cases}$$

come dovevasi dimostrare.

Si dimostrerà ora che la variabile residuo $Z = X_1 - \hat{X}_1$ è incorrelata anche con la variabile $\hat{X}_1$.

A tal fine moltiplicando la prima equazione del sistema (6.1) per α_1, la seconda per α_2 e la terza per α_3 si ottiene

$$\begin{cases} \sum z_i \cdot \alpha_1 = 0 \\ \sum z_i \cdot \alpha_2 \cdot x_{2i} = 0 \\ \sum z_i \cdot \alpha_3 \cdot x_{3i} = 0. \end{cases}$$

Sommando le tre equazioni, membro a membro, si ottiene

$$\sum z_i \cdot \{\alpha_1 + \alpha_2 x_{2i} + \alpha_3 x_{3i}\} = 0,$$

ovvero

$$\frac{1}{N} \sum z_i \cdot \hat{x}_{1i} = 0.$$

Ciò dimostra che effettivamente $Cov[Z, \hat{X}_1] = 0$.

Si può concludere questo paragrafo affermando che, quando si interpola a minimi quadrati con il piano $\hat{X}_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3$, i residui $z_i = x_{1i} - \hat{x}_{1i}$:

1. hanno media aritmetica nulla,
2. sono incorrelati sia con le variabili esplicative X_2, X_3 sia con la variabile $\hat{X}_1$.

7. *Varianza totale, varianza residua e varianza spiegata*

Si mostrerà ora che è possibile scomporre additivamente la varianza (devianza) di X_1 in varianza (devianza) spiegata dal piano interpolatore e varianza (devianza) residua. In effetti

$$\sum (x_{1i} - \bar{x}_1)^2 = \sum \{(x_{1i} - \hat{x}_{1i} + \hat{x}_{1i} - \bar{x}_1)\}^2 = \sum (x_{1i} - \hat{x}_{1i})^2 + \\ + \sum (\hat{x}_{1i} - \bar{x}_1)^2 + 2 \sum (x_{1i} - \hat{x}_{1i})(\hat{x}_{1i} - \bar{x}_1).$$

L'ultima sommatoria indica la codevianza fra la variabile residua $(X_1 - \hat{X}_1)$ e la variabile $\hat{X}_1$ che, come è stato dimostrato nel precedente paragrafo, è nulla. Consegue che

$$\underbrace{\sum (x_{1i} - \bar{x}_1)^2}_{\text{Devianza totale}} = \underbrace{\sum (\hat{x}_{1i} - \bar{x})^2}_{\text{Devianza spiegata}} + \underbrace{\sum (x_{1i} - \hat{x}_{1i})^2}_{\text{Devianza residua}} \qquad (7.1)$$

Ecco ora un'inaspettata ed utile rappresentazione della devianza (varianza) spiegata.

La devianza spiegata è data da

$$\sum(\hat{x}_{1i}-\bar{x}_1)^2=\sum(\hat{x}_{1i}-\bar{x}_1)(\hat{x}_{1i}-\bar{x}_1)=\sum(\hat{x}_{1i}-\bar{x}_1)\{(\hat{x}_{1i}-x_{1i})+(x_{1i}-\bar{x}_1)\}$$
$$=\sum(\hat{x}_{1i}-x_{1i})(\hat{x}_{1i}-\bar{x}_1)+\sum(\hat{x}_{1i}-\bar{x}_1)(x_{1i}-\bar{x}_1).$$

Si osservi che $\sum(\hat{x}_{1i}-x_{1i})(\hat{x}_{1i}-\bar{x}_1)=0$ in quanto l'espressione indica la codevianza fra Z e $\hat{X}_1$. Tenuto conto altresì che

$$(\hat{x}_{1i}-\bar{x}_1)=\alpha_1+\alpha_2 x_{2i}+\alpha_3 x_{3i}-\bar{x}_1=(\bar{x}_1-\alpha_2\bar{x}_2-\alpha_3\bar{x}_3)+\alpha_2 x_{2i}+\alpha_3 x_{3i}-\bar{x}_1=$$
$$=\alpha_2(x_{2i}-\bar{x}_2)+\alpha_3(x_{3i}-\bar{x}_3)$$

la devianza spiegata diventa

$$\sum(\hat{x}_{1i}-\bar{x}_1)^2=\alpha_2\sum(x_{2i}-\bar{x}_2)(x_{1i}-\bar{x}_1)+\alpha_3\sum(x_{3i}-\bar{x}_3)(x_{1i}-\bar{x}_1).$$

Infine dividendo per N si ha

$$\text{Varianza spiegata}=\alpha_2\cdot\sigma_{21}+\alpha_3\cdot\sigma_{31}. \tag{7.2}$$

Ricordando ora che $M_1(\hat{X}_1)=\bar{x}_1$ si ha che la varianza spiegata coincide con la varianza di $\hat{X}_1$, per cui

$$Var(\hat{X}_1)=\alpha_2\cdot\sigma_{21}+\alpha_3\cdot\sigma_{31}\ . \tag{7.3}$$

La (7.2) e la (7.3) troveranno diversi impieghi in seguito. Inoltre la (7.2) informa che per il calcolo della varianza spiegata non è necessario valutare preventivamente i valori $\hat{x}_{1i}$ essendo sufficiente conoscere α_2, α_3 e le covarianze σ_{21} e σ_{31}.

8. *Bontà di adattamento del piano interpolatore*

Dopo aver determinato i parametri α_1, α_2 e α_3 solitamente si determinano i valori $\hat{x}_{1i}=\alpha_1+\alpha_2 x_{2i}+\alpha_3 x_{3i}$. Bisogna ora valutare se il piano interpolatore approssima bene i dati osservati. Tale analisi si fonda essenzialmente sui residui $z_i=x_{1i}-\hat{x}_{1i}$. Misure della bontà di adattamento del piano si ottengono valutando:

1. l'ordine di grandezza dei residui;
2. la quota della variabilità di X_1 "attribuibile" al piano interpolatore.

Ragionevoli misure dell'ordine di grandezza dei residui sono gli indici:

$$A_1 = \frac{1}{N}\sum\left|x_{1i} - \hat{x}_{1i}\right| \quad \text{e} \quad A_2 = \left\{\frac{1}{N}\sum(x_{1i} - \hat{x}_{1i})^2\right\}^{\frac{1}{2}}.$$

Nei casi in cui ha senso si possono impiegare anche gli indici relativi:

$$A_1' = \frac{A_1}{\bar{x}_1} \quad \text{e} \quad A_2' = \frac{A_2}{\bar{x}_1}.$$

Un altro indice molto impiegato per valutare la bontà di adattamento del piano interpolatore si ottiene rapportando la devianza spiegata $\sum(\hat{x}_{1i} - \bar{x}_1)^2$ alla devianza totale $\sum(x_{1i} - \bar{x}_1)^2$. Si ha così l'indice di determinazione multiplo

$$I_{1.23}^2 = \frac{\sum(\hat{x}_{1i} - \bar{x}_1)^2}{\sum(x_{1i} - \bar{x}_1)^2} = \frac{\text{Varianza spiegata}}{\text{Varianza totale}}. \tag{8.1}$$

Tenuto conto della scomposizione (7.1), l'indice (8.1) si può scrivere come segue

$$I_{1.23}^2 = \frac{\sum(\hat{x}_{1i} - \bar{x}_1)^2}{\sum(\hat{x}_{1i} - \bar{x}_1)^2 + \sum(x_{1i} - \hat{x}_{1i})^2}. \tag{8.2}$$

L'indice $I_{1.23}^2$ è pari ad 1 solo se tutti i residui $z_i = x_{1i} - \hat{x}_{1i}$ sono nulli. Ovviamente ciò accade solo se i dati sperimentali si dispongono su un piano. L'indice è pari a zero solo se $\hat{x}_{1i} = \bar{x}_1$ per ogni $i = 1, 2, \ldots, N$. Ciò si realizza solo se $\alpha_2 = 0$ e $\alpha_3 = 0$ di modo che $\alpha_1 = \bar{x}_1 - \alpha_2 \cdot \bar{x}_2 - \alpha_3 \cdot \bar{x}_3 = \bar{x}_1$ e quindi $\hat{x}_{1i} = \bar{x}_1$.

Ovviamente il piano interpolatore sarà tanto più adeguato quanto più $I_{1.23}^2$ è prossimo ad 1.

Esempio 8.1.

Si riconsiderino gli Esempi 4.1 e 5.1 e si valuti la bontà di adattamento dal piano interpolatore a minimi quadrati

$$\hat{x}_{1i} = 28{,}546 - 2{,}138x_{2i} + 3{,}257x_{3i}.$$

Nel prospetto dell'Esempio 4.1 erano già stati calcolati gli scarti

$(x_{1i}-\hat{x}_{1i})$. Si desume così che $\sum|x_{1i}-\hat{x}_{1i}|=60,0+59,9=119,9$ e quindi

$$A_1=\frac{119,9}{20}=5,995.$$

Essendo $\bar{x}_1=87,2$ (Esempio 5.1) si ricava

$$A_1'=\frac{5,995}{87,2}=0,069.$$

Dal valore di A_1' si desume che in media gli scarti residui sono pari a circa il 7% del valore della media della variabile dipendente X_1. Pertanto, almeno in base al valore di A_1', si deve ritenere che il piano interpolatore è adeguato a rappresentare i redditi aziendali X_1 in funzione della superficie X_2 ed in funzione del numero di bovini X_3.

Anche se il valore dell'indice A_1' dà già delle garanzie sull'adeguatezza del piano interpolatore a rappresentare i redditi aziendali, conviene calcolare anche l'indice $I_{1.23}^2$ in quanto considera l'accostamento da un altro punto di vista.

Dal prospetto della pagina seguente si ricava che la devianza totale è pari a 5.455,2 e quella residua è pari a 993,31.

Consegue che

$$I_{1.23}^2=\frac{5.455,2-993,31}{5.455,2}=\frac{4.467,89}{5.455,2}=0,818.$$

Dal valore di $I_{1.23}^2$ si desume che il piano interpolatore spiega circa l'81,8% della variabilità totale dei redditi aziendali.

Si ribadisce che, poiché gli indici A_1, A_1' e $I_{1.23}^2$ danno valutazioni del grado di adattamento secondo differenti punti di vista, è sempre preferibile, per una più completa analisi, impiegarli congiuntamente.

X_1	$X_1-\bar{x}_1$	$(X_1-\bar{x}_1)^2$	$X_1-\hat{X}_1$	$(X_1-\hat{X}_1)^2$	X_2	X_3	$(X_1-\hat{X}_1)\cdot X_2$	$(X_1-\hat{X}_1)\cdot X_3$
96	8,8	77,44	– 4	16	6	18	– 24,0	– 72
83	– 4,2	17,64	7,4	54,76	22	0	162,8	0
126	38,8	1.505,44	13,4	179,56	18	14	241,2	187,6
61	– 26,2	686,44	– 4,2	17,64	8	6	– 33,6	– 25,2
59	– 28,2	795,24	1,5	2,25	12	1	18,0	1,5
90	2,8	7,84	10,8	116,64	10	9	108,0	97,2
82	– 5,2	27,04	– 2,4	5,76	17	6	– 40,8	– 14,4
88	0,8	0,64	– 3,2	10,24	11	12	– 35,2	– 38,4
86	– 1,2	1,44	0,4	0,16	16	7	6,4	2,8
76	– 11,2	125,44	– 8,2	67,24	23	2	– 188,6	– 16,4
102	14,8	219,04	3,1	9,61	7	17	21,7	52,7
108	20,8	432,64	4,9	24,01	12	15	58,8	73,5
96	8,8	77,44	– 6,7	44,89	24	7	– 160,8	– 46,9
70	– 17,2	295,84	7,2	51,84	16	0	115,2	0
80	– 7,2	51,84	– 6,9	47,61	9	12	– 62,1	– 82,8
113	25,8	665,64	8,8	77,44	11	16	96,8	140,8
76	– 11,2	125,44	– 6,1	37,21	22	2	– 134,2	– 12,2
74	– 13,2	174,24	2,4	5,76	11	6	26,4	14,4
98	10,8	116,64	– 3,8	14,44	16	12	– 60,8	– 45,6
80	– 7,2	51,84	– 14,5	210,25	8	15	– 116,0	– 217,5
1.744	+ 132,2 – 132,2	5.455,20		993,31	279	177	+ 855,3 – 856,1	+ 570,5 – 571,4

Esempio 8.2.

Si vuole verificare la relazione (7.2)

$$\text{Varianza spiegata} = \alpha_2 \cdot \sigma_{21} + \alpha_3 \cdot \sigma_{31}$$

sui dati dell'Esempio 4.1.

I dati necessari al calcolo della (7.2) sono già stati ricavati per lo svolgimento dell'Esempio 5.1

$$\alpha_2 = 2{,}1384; \quad \alpha_3 = 3{,}2569; \quad \sigma_{21} = 0{,}71; \quad \sigma_{31} = 68{,}03.$$

Si ha così

$$2{,}1384 \cdot 0{,}71 + 3{,}2569 \cdot 68{,}03 = 223{,}09.$$

È possibile valutare la varianza spiegata, per via diretta, rapportando la devianza spiegata (5.455,2 – 993,31) = 4.461,89 al numero di osservazioni.

$-\hat{X}_1)\cdot X_3$
−72
0
187,6
−25,2
1,5
97,2
−14,4
−38,4
2,8
−16,4
52,7
73,5
−46,9
0
−82,8
140,8
−12,2
14,4
−45,6
217,5
570,5
571,4

lo svolgi-

ortando la
servazioni.

Si ha

$$\frac{4.461,89}{20} = 223,094 .$$

Si ha un valore praticamente uguale a quello ottenuto applicando la (7.2).

Esempio 8.3.

Si vuole verificare che non si siano commessi errori di calcolo nella determinazione dei parametri α_1, α_2 e α_3 dell'Esempio 4.1.

Occorre perciò esaminare se sono verificate le tre equazioni del sistema normale (4.2)

$$\begin{cases} \sum(x_{1i}-\hat{x}_{1i})=0 \\ \sum(x_{1i}-\hat{x}_{1i})\cdot x_{2i}=0 \\ \sum(x_{1i}-\hat{x}_{1i})\cdot x_{3i}=0. \end{cases}$$

In altri termini si tratta di controllare che i residui $Z=X_1-\hat{X}_1$ abbiano media nulla e siano incorrelati con le variabili esplicative X_2 e X_3.

Dal prospetto predisposto per l'Esempio 8.1 si ricava che effettivamente le tre sommatorie del sistema (4.2) sono praticamente nulle. Si può dunque ritenere che non si sono commessi errori nel calcolo dei parametri del piano interpolatore.

9. *Coefficiente di correlazione multiplo*

Dopo aver determinato i parametri α_1, α_2 e α_3 si possono ricavare i valori $\hat{x}_{1i}=\alpha_1+\alpha_2 x_{2i}+\alpha_3 x_{3i}$.

Per coefficiente di correlazione multiplo di X_1 rispetto alle variabili X_2 e X_3 si intende il coefficiente di correlazione lineare fra X_1 e $\hat{X}_1$.

Il coefficiente di correlazione multiplo si indica di solito con $R_{1.23}$. È quindi per definizione

$$R_{1.23}=\frac{Cov(X_1,\hat{X}_1)}{\sigma(X_1)\cdot\sigma(\hat{X}_1)}.$$

Si dimostrerà ora che $R^2_{1.23}=I^2_{1.23}$.

In effetti

$$\begin{aligned} Cov[X_1, \hat{X}_1] &= \frac{1}{N}\sum(x_{1i} - \bar{x}_1)(\hat{x}_{1i} - \bar{x}_1) = \\ &= \frac{1}{N}\sum(x_{1i} - \bar{x}_1)\{\alpha_2(x_{2i} - \bar{x}_2) + \alpha_3(x_{3i} - \bar{x}_3)\} = \\ &= \alpha_2\sigma_{12} + \alpha_3\sigma_{13} = Var(\hat{X}_1) \end{aligned} \tag{9.1}$$

per la (7.3). Si osservi che $Cov[X_1, \hat{X}_1] = Var(\hat{X}_1) \geq 0$ in quanto una varianza non può essere negativa. Si ha così

$$R_{1.23} = \frac{Var(\hat{X}_1)}{\sigma(X_1)\cdot\sigma(\hat{X}_1)} = \frac{\sigma(\hat{X}_1)}{\sigma(X_1)} \geq 0 \tag{9.2}$$

e quindi

$$R^2_{1.23} = \frac{Var(\hat{X}_1)}{Var(X_1)} = \frac{\alpha_2\sigma_{12} + \alpha_3\sigma_{13}}{\sigma_{11}} = \frac{\text{Var spiegata}}{\text{Var totale}} = I^2_{1.23}. \tag{9.3}$$

Si ricaverà ora un'ulteriore rappresentazione di $R^2_{1.23}$ che verrà impiegata in seguito.

Nella (9.3) al posto di α_2 e α_3 si sostituiscono i valori forniti dalla (5.9) e dalla (5.10). Si ha così

$$\begin{aligned} R^2_{1.23} &= \left\{\frac{\sigma_{33}\sigma_{12} - \sigma_{13}\sigma_{23}}{\sigma_{22}\sigma_{33} - \sigma^2_{23}}\cdot\frac{\sigma_{12}}{\sigma_{11}} + \frac{\sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23}}{\sigma_{22}\sigma_{33} - \sigma^2_{23}}\cdot\frac{\sigma_{13}}{\sigma_{11}}\right\} = \\ &= \left\{\frac{\dfrac{\sigma_{33}\sigma^2_{12}}{\sigma_{11}\sigma_{22}\sigma_{33}} - \dfrac{\sigma_{13}\sigma_{23}\sigma_{12}}{\sigma_{11}\sigma_{22}\sigma_{33}}}{1 - \dfrac{\sigma^2_{23}}{\sigma_{22}\sigma_{33}}} + \frac{\dfrac{\sigma_{22}\sigma^2_{13}}{\sigma_{11}\sigma_{22}\sigma_{33}} - \dfrac{\sigma_{12}\sigma_{23}\sigma_{13}}{\sigma_{11}\sigma_{22}\sigma_{33}}}{1 - \dfrac{\sigma^2_{23}}{\sigma_{22}\sigma_{33}}}\right\}. \end{aligned} \tag{9.4}$$

Indicando ora con $r_{ij} = \dfrac{\sigma_{ij}}{\sigma(X_i)\cdot\sigma(X_j)}$ il coefficiente di correlazione lineare fra le variabili X_i e X_j la (9.4) diventa

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}\cdot r_{13}\cdot r_{23}}{1 - r^2_{23}}. \tag{9.5}$$

La (9.5), tra l'altro, informa che nel caso in cui le variabili esplicative siano fra loro incorrelate risulta $R^2_{1.23} = I^2_{1.23} = r^2_{12} + r^2_{13}$. Cioè la quota della variabilità di X_1 spiegata dal piano interpolatore è pari alla somma della quota spiegata dalla retta X_1 rispetto a X_2 e della quota spiegata dalla retta X_1 rispetto a X_3.

Si consideri ora la matrice dei coefficienti di correlazione fra le tre variabili X_1, X_2 e X_3.

$$\underline{C} = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

nonché la matrice dei coefficienti di correlazione delle due variabili X_2 e X_3

$$\underline{C}_1 = \begin{vmatrix} r_{22} & r_{23} \\ r_{23} & r_{33} \end{vmatrix} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix}.$$

Applicando la regola di Sarrus si ricava che il determinante di $\underline{C}$ è

$$d(\underline{C}) = (1 - r_{23}^2) - r_{12}^2 - r_{13}^2 + 2r_{12}r_{13}r_{23}.$$

Il determinante di $\underline{C}_1$ risulta

$$d(\underline{C}_1) = 1 - r_{23}^2.$$

Si deduce ora facilmente dalla (9.5) che

$$R_{1.23}^2 = I_{1.23}^2 = 1 - \frac{d(\underline{C})}{d(\underline{C}_1)}. \tag{9.6}$$

10. *Miglioramento nella bontà di adattamento nel passaggio dalla retta al piano a minimi quadrati*

In questo paragrafo si valuterà l'aumento della varianza spiegata che si ha nel passaggio dalla retta a minimi quadrati al piano a minimi quadrati.

Nel caso dell'interpolazione con la retta a minimi quadrati $\hat{X}_1 = a + bX_2$, la devianza residua è pari a

$$\Sigma(x_{1i} - a - bx_{2i})^2. \tag{10.1}$$

Nel caso dell'interpolazione con il piano a minimi quadrati $\hat{X}_1 = \hat{\alpha}_1 + \hat{\alpha}_2 X_2 + \hat{\alpha}_3 X_3$, la devianza residua è pari a

$$\sum(x_{1i}-\hat{\alpha}_1-\hat{\alpha}_2x_{2i}-\hat{\alpha}_3x_{3i})^2. \tag{10.2}$$

Infine, la devianza residua dal generico piano $\hat{X}_1=\alpha_1+\alpha_2X_2+\alpha_3X_3$ risulta

$$\sum(x_{1i}-\alpha_1-\alpha_2x_{2i}-\alpha_3x_{3i})^2. \tag{10.3}$$

Ovviamente vale la seguente relazione

$$\sum(x_{1i}-\alpha_1-\alpha_2x_{2i}-\alpha_3x_{3i})^2 \geq \sum(x_{1i}-\hat{\alpha}_1-\hat{\alpha}_2x_{2i}-\hat{\alpha}_3x_{3i})^2. \tag{10.4}$$

La relazione (10.4) vale qualunque siano i valori di α_1, α_2 e α_3. In particolare la (10.4) vale anche per $\alpha_1=a$, $\alpha_2=b$ e $\alpha_3=0$. Si ha allora

$$\sum(x_{1i}-a-bx_{2i}-0\cdot x_{3i})^2 \geq \sum(x_{1i}-\hat{\alpha}_1-\hat{\alpha}_2x_{2i}-\hat{\alpha}_3x_{3i})^2,$$

ovvero

$$\sum(x_{1i}-a-bx_{2i})^2 \geq \sum(x_{1i}-\hat{\alpha}_1-\hat{\alpha}_2x_{2i}-\hat{\alpha}_3x_{3i})^2. \tag{10.5}$$

Si è così dimostrato che quando si passa dalla retta a minimi quadrati al piano a minimi quadrati la devianza residua diminuisce. Le due devianze sono uguali solo nel caso $\hat{\alpha}_3=0$.

Per valutare il miglioramento che si consegue nel passare dalla retta a minimi quadrati al piano a minimi quadrati si può sottrarre l'indice di determinazione della retta $I^2_{1.2}=r^2_{12}$ all'indice di determinazione del piano $I^2_{1.23}$ dato dalla (9.5). Si ha

MVT ≠ MVS

$$\begin{aligned} I^2_{1.23}-I^2_{1.2} &= \frac{r^2_{12}+r^2_{13}-2r_{12}r_{13}r_{23}}{1-r^2_{23}}-r^2_{12}= \\ &= \frac{r^2_{12}+r^2_{13}-2r_{12}r_{13}r_{23}-r^2_{12}(1-r^2_{23})}{1-r^2_{23}}= \\ &= \frac{\{r_{13}-r_{12}r_{23}\}^2}{1-r^2_{23}}\geq 0. \end{aligned} \tag{10.6}$$

La differenza indica la frazione della varianza totale che viene spiegata nel passare dalla retta al piano.

Un altro modo per valutare il miglioramento si ha operando opportunamente sulle seguenti varianze residue

$$\text{Varianza residua dalla retta} = \sigma_{11}\{1-I^2_{1.2}\} \tag{10.7}$$

$$\text{Varianza residua dal piano} = \sigma_{11}\left\{1 - r^2_{1.23}\right\}. \tag{10.8}$$

Si può quindi porre

$$\text{Grado di miglioramento} = \frac{\sigma_{11}\left\{1 - r^2_{1.2}\right\} - \sigma_{11}\left\{1 - r^2_{1.23}\right\}}{\sigma_{11}\left\{1 - r^2_{1.2}\right\}} = \tag{10.9}$$

$$\text{MVR} = \frac{r^2_{1.23} - r^2_{1.2}}{1 - r^2_{1.2}} = \frac{\left\{r_{13} - r_{12}r_{23}\right\}^2}{\left\{1 - r^2_{12}\right\}\left\{1 - r^2_{23}\right\}}.$$

La (10.9) indica la frazione della varianza residua dalla retta che viene spiegata nel passare dalla retta al piano.

11. *Coefficienti di regressione grezzi e parziali*

In questo paragrafo si esamineranno le relazioni fra il coefficiente angolare della retta a minimi quadrati e quelli del piano interpolatore.

Per evitare confusione sembra opportuno indicare i parametri del piano e delle rette come segue

$$\hat{X}_1 = a + \alpha_{12.3}X_2 + \alpha_{13.2} \cdot X_3 \tag{11.1}$$

$$\hat{X}_1 = b + \alpha_{12} \cdot X_2 \tag{11.2}$$

$$\hat{X}_1 = c + \alpha_{13} \cdot X_3.$$

Il parametro $\alpha_{12.3}$ del piano (11.1) indica la variazione di $\hat{X}_1$ che si ha in corrispondenza di un incremento unitario di X_2 nell'ipotesi che X_3 resti costante. Per tale motivo $\alpha_{12.3}$ è detto *coefficiente di regressione parziale o netto*.

Il parametro α_{12} della retta (11.2) indica la variazione di $\hat{X}_1$ in corrispondenza di un incremento unitario di X_2 (al lordo delle variazioni di X_3).

Significati analoghi hanno $\alpha_{13.2}$ e α_{13}.

Nelle scienze osservazionali quando si interpola con una retta X_1 in funzione di X_2, solitamente al variare di X_2 varia anche X_3 e quindi il valore di X_1 varia non solo perché è variato X_2 ma anche perché è variato X_3. Per tale motivo α_{12} è detto *coefficiente di regressione grezzo (totale)*. Ovviamente quanto detto per α_{12} vale per α_{13}.

Fra il valore dei coefficienti grezzi ed i valori di quelli parziali possono esservi notevoli differenze come indica l'esempio che segue.

Esempio 11.1.

Si riconsiderino i dati dell'Esempio 4.1. Si determinano ora i parametri delle rette a minimi quadrati

$$\hat{X}_1 = a + \alpha_{12} X_2$$
$$\hat{X}_1 = b + \alpha_{13} X_3 .$$

Com'è noto, i parametri richiesti sono forniti dalle formule

$$\alpha_{12} = \frac{\sigma_{12}}{\sigma_{22}}; \quad a = \overline{X}_1 - \alpha_{12}\overline{X}_2;$$

$$\alpha_{13} = \frac{\sigma_{13}}{\sigma_{33}}; \quad b = \overline{X}_1 - \alpha_{13}\overline{X}_3 .$$

Nello svolgimento dell'Esempio 5.1 le medie, le varianze e le covarianze, necessarie per la determinazione dei parametri delle rette, sono già state calcolate. In particolare si è ricavato

$$\overline{X}_1 = 87{,}2; \quad \overline{X}_2 = 13{,}95; \quad \overline{X}_3 = 8{,}85;$$

$$\sigma_{22} = 30{,}3475; \quad \sigma_{33} = 33{,}8275; \quad \sigma_{12} = 0{,}71; \quad \sigma_{13} = 68{,}03.$$

Si ha così

$$\alpha_{12} = \frac{0{,}71}{30{,}3475} = 0{,}023; \quad a = 87{,}2 - 0{,}023 \cdot 13{,}95 = 86{,}879;$$

$$\alpha_{13} = \frac{68{,}03}{33{,}8275} = 2{,}011; \quad b = 87{,}2 - 2{,}011 \cdot 8{,}85 = 69{,}403.$$

Le equazioni delle due rette interpolanti risultano

$$\hat{X}_1 = 86{,}879 + 0{,}023 \cdot X_2$$

$$\hat{X}_1 = 69{,}403 + 2{,}011 \cdot X_3 .$$

I valori dei coefficienti angolari delle due rette sono molto diversi dai valori dei coefficienti di regressione parziali del piano. In particolare dall'equazione della prima retta

$$\hat{X}_1 = 86{,}879 + 0{,}023X_2$$

sembrerebbe che la variabile X_2 (superficie aziendale) non abbia alcuna influenza sul reddito aziendale. Per spiegare perché ciò sia potuto accadere nonostante che $\alpha_{12.3} = 2{,}1384$, si sono riportati nel prospetto che segue i dati delle tre variabili X_1 (reddito aziendale), X_2 (superficie aziendale) e X_3 (numero di bovini) ordinati secondo la superficie aziendale X_2.

X_2	X_3	X_1
6	18	96
7	17	102
8	6	61
8	15	80
9	12	80
10	9	90
11	6	74
11	12	88
11	16	113
12	1	59
12	15	108
16	0	70
16	7	86
16	12	98
17	6	82
18	14	126
22	0	83
22	2	76
23	2	76
24	7	96

Dal prospetto si desume che all'aumentare di X_2 l'altra variabile esplicativa X_3 tende a diminuire. In effetti il coefficiente di correlazione fra queste ultime risulta

$$r_{23} = \frac{\sigma_{23}}{\sqrt{\sigma_{22}} \cdot \sqrt{\sigma_{33}}} = \frac{-19{,}7075}{\sqrt{30{,}3475} \cdot \sqrt{33{,}8275}} = \frac{-19{,}7075}{5{,}509 \cdot 5{,}816} = -0{,}615.$$

Dopo questa precisazione sul comportamento simultaneo (reale) delle due variabili esplicative si cercherà di spiegare l'apparente anomalia sopra evidenziata.

Aumentando la superficie X_2 e tenendo fisso il numero dei bovini X_3 il reddito aziendale X_1 dovrebbe aumentare. È quanto suggerisce il valore di $\alpha_{12.3} = 2{,}1384$ ed è quanto effettivamente si verifica nella realtà come mostra il prospetto che segue relativo alle aziende con almeno due valori uguali di X_3.

X_3	X_2	X_1
6	8	61
6	11	74
6	17	82
7	16	86
7	24	96
12	9	80
12	11	88
12	16	98
15	8	80
15	12	108

Il prospetto mostra che tenendo fisso X_3 ed aumentando X_2 anche X_1 aumenta.

Analogamente tenendo fissa la superficie X_2 e facendo aumentare il numero di bovini X_3 il reddito aziendale X_1 dovrebbe crescere. È, in effetti, quanto suggerisce il valore di $\alpha_{13.2} = 3{,}2569$ ed è quanto effettivamente si realizza nella realtà come mostra il seguente prospetto relativo alle aziende con almeno due valori uguali di X_2.

X_2	X_3	X_1
11	6	74
11	12	88
11	16	113
12	1	59
12	15	108
16	0	70
16	7	86
16	12	98
22	0	83
22	2	76

Nel complesso delle 20 aziende però, le due variabili esplicative hanno un andamento contrapposto e, quindi, tenuto conto dei valori positivi di $\alpha_{13.2}$ e di $\alpha_{12.3}$, anche il reddito X_1 è influenzato da due forze contrapposte: la prima dovuta all'aumento di X_2 che tende a far aumentare X_1 e la seconda dovuta alla contemporanea diminuzione di X_3 che tende a far diminuire X_1. Consegue che all'aumento di X_2 si contrappone una quasi stazionarietà di X_1; in effetti X_1 tende a non variare essendo $\alpha_{12} = 0{,}023$. Questo esempio mostra che bisogna usare molta cautela nell'interpretazione dei coefficienti grezzi (totali) di regressione.

È tempo ora di ricavare le relazioni fra i coefficienti di regressione parziali e quelli di regressione grezzi (totali).

In forza della (5.9) il coefficiente di regressione parziale risulta

$$\alpha_{12.3} = \frac{\sigma_{33}\sigma_{12} - \sigma_{13}\sigma_{23}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2}. \tag{11.3}$$

Dividendo numeratore e denominatore della (11.3) per $\sigma_{22} \cdot \sigma_{33}$ si ha

$$\alpha_{12.3} = \frac{\dfrac{\sigma_{12}}{\sigma_{22}} - \dfrac{\sigma_{13}}{\sigma_1\sigma_3} \cdot \dfrac{\sigma_{23}}{\sigma_2\sigma_3} \cdot \dfrac{\sigma_1}{\sigma_2}}{1 - r_{23}^2} = \frac{\alpha_{12} - r_{13}r_{23} \cdot \dfrac{\sigma_1}{\sigma_2}}{1 - r_{23}^2}. \tag{11.4}$$

Dalla (11.4) si ricava

$$\alpha_{12} = \alpha_{12.3} \cdot \left\{1 - r_{23}^2\right\} + r_{13}r_{23} \cdot \frac{\sigma_1}{\sigma_2}. \tag{11.5}$$

In modo analogo si dimostra che

$$\alpha_{13.2} = \frac{\alpha_{13} - r_{12} \cdot r_{23} \cdot \dfrac{\sigma_1}{\sigma_3}}{1 - r_{23}^2} \tag{11.6}$$

$$\alpha_{13} = \alpha_{13.2}\left\{1 - r_{23}^2\right\} + r_{12}r_{23} \cdot \frac{\sigma_1}{\sigma_3}. \tag{11.7}$$

Le formule ora ricavate informano che se vi è incorrelazione fra le variabili esplicative X_2 e X_3 allora $\alpha_{12} = \alpha_{12.3}$ e $\alpha_{13} = \alpha_{13.2}$. Negli altri casi fra i coefficienti di regressione grezzi e quelli parziali vi possono essere differenze rilevanti.

La formula (11.5) fa comprendere perché nell'Esempio 11.1 si sia riscontrata tanta diversità fra α_{12} e $\alpha_{12.3}$. Le grandezze non ancora determinate in precedenza e necessarie per l'applicazione della formula (11.5) sono

$$\sigma_{11} = \frac{1}{20} 5.455{,}2 = 272{,}76; \quad \sigma_1 = \sqrt{272{,}76} = 16{,}515;$$

$$\sigma_3 = \sqrt{33{,}8275} = 5{,}816; \quad \sigma_2 = \sqrt{30{,}3475} = 5{,}509;$$

$$r_{13} = \frac{68{,}03}{16{,}515 \cdot 5{,}816} = 0{,}708.$$

Si ha ora

$$\alpha_{12} = 2{,}1384\left[1-(-0{,}615)^2\right] + 0{,}708 \cdot (-0{,}615) \cdot \frac{16{,}515}{5{,}509} =$$
$$= 2{,}1384 \cdot 0{,}6218 - 1{,}3053 = 1{,}330 - 1{,}3053 = 0{,}0244$$

valore praticamente coincidente con quello ottenuto applicando la formula $\alpha_{12} = \dfrac{\sigma_{12}}{\sigma_{22}}$.

La verifica numerica eseguita conferma che nell'esempio considerato il valore di α_{12} risulta praticamente nullo in quanto la presenza di $\alpha_{12.3} = 2{,}1384$ è controbilanciata dalla presenza di un'elevata correlazione negativa fra le variabili esplicative.

12. *Coefficienti di correlazione parziali*

Per coefficiente di correlazione parziale fra due variabili X_1 e X_2 si intende il coefficiente di correlazione fra le stesse allorché si mantiene fissa la terza variabile X_3. Nelle scienze economiche non è quasi mai possibile tenere fissa la terza variabile e quindi la correlazione r_{12} sconta l'effetto che le variazioni di X_3 hanno su X_1 e su X_2. Tuttavia è possibile pervenire lo stesso ad una valutazione del coefficiente di correlazione parziale fra X_1 e X_2 al netto dell'influenza di X_3, che sarà indicata con $r_{12.3}$, impiegando opportunamente grandezze già incontrate nei precedenti paragrafi.

I) *Si sa che r_{12} può essere ricavato come media geometrica di α_{12} e α_{21}.* Analogamente $r_{12.3}$ può ottenersi calcolando la media geometrica fra $\alpha_{12.3}$ e $\alpha_{21.3}$.

Dalla (11.4) si sa che

$$\alpha_{12.3} = \frac{\alpha_{12} - r_{13}r_{23} \cdot \dfrac{\sigma_1}{\sigma_2}}{1 - r_{23}^2}$$

tenuto conto che $\alpha_{12} = r_{12} \cdot \dfrac{\sigma_1}{\sigma_2}$ diventa

$$\alpha_{12.3} = \frac{\sigma_1}{\sigma_2} \cdot \left\{ \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \right\}. \tag{12.1}$$

Scambiando le posizioni degli indici 1 e 2 si ha

$$\alpha_{21.3} = \frac{\sigma_2}{\sigma_1} \cdot \left\{ \frac{r_{21} - r_{23}r_{13}}{1 - r_{13}^2} \right\}. \tag{12.2}$$

Si osservi che il segno di $\alpha_{12.3}$ è lo stesso di quello di $\alpha_{21.3}$. Ne segue che è possibile calcolare la media geometrica fra $\alpha_{12.3}$ e $\alpha_{21.3}$ e che essa risulta

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\{1 - r_{23}^2\}} \cdot \sqrt{\{1 - r_{13}^2\}}}. \tag{12.3}$$

La (12.3) informa che se le variabili X_1 e X_2 sono incorrelate con X_3 allora la correlazione parziale $r_{12.3}$ è uguale a quella grezza r_{12}. Negli altri casi fra queste ultime può esservi notevole differenza.

Dalla (12.3) si ricava agevolmente che

$$r_{12} = r_{12.3}\{1 - r_{23}^2\}^{\frac{1}{2}} \cdot \{1 - r_{13}^2\}^{\frac{1}{2}} + r_{13}r_{23}. \tag{12.4}$$

II) *Alla formula (12.3) si può pervenire anche considerando le variabili* *residuo che si ottengono dopo aver interpolato a minimi quadrati la X_1 rispetto a X_3 e la X_2 rispetto a X_3.*

Siano $\hat{X}_1 = a + \alpha_{13} X_3$ e $\hat{X}_2 = b + \alpha_{23} \cdot X_3$ le interpolanti rispetto alla variabile X_3.

Il residuo di X_1 dopo aver eliminato l'influenza di X_3 risulta

$$\begin{aligned} Y &= X_1 - (a + \alpha_{13} X_3) = X_1 - (\overline{X}_1 - \alpha_{13}\overline{X}_3 + \alpha_{13} X_3) \\ &= (X_1 - \overline{X}_1) - \alpha_{13}(X_3 - \overline{X}_3). \end{aligned} \tag{12.5}$$

Analogamente il residuo di X_2 dopo aver eliminato l'influenza di X_3 risulta

$$V = X_2 - (b + \alpha_{23} X_3) = (X_2 - \overline{X}_2) - \alpha_{23}(X_3 - \overline{X}_3). \tag{12.6}$$

La correlazione parziale fra X_1 e X_2, al netto di X_3, non è altro che la correlazione fra le variabili residuo Y e V.

Essendo $\overline{Y} = 0$ e $\overline{V} = 0$ risulta $Cov(Y,V) = \frac{1}{N} \sum y_i v_i$. Tenuto conto ora della (12.5) e della (12.6) si ha

$$\begin{aligned} y_i \cdot v_i &= (x_{1i} - \overline{x}_1) \cdot (x_{2i} - \overline{x}_2) - \alpha_{23} \cdot (x_{1i} - \overline{x}_1)(x_{3i} - \overline{x}_3) + \\ &\quad - \alpha_{13}(x_{3i} - \overline{x}_3)(x_{2i} - \overline{x}_2) + \alpha_{13}\alpha_{23}(x_{3i} - \overline{x}_3)^2. \end{aligned}$$

Pertanto

$$Cov\ (Y, V) = \sigma_{12} - \alpha_{23} \cdot \sigma_{13} - \alpha_{13} \cdot \sigma_{23} + \alpha_{13} \cdot \alpha_{23} \cdot \sigma_{33}. \tag{12.7}$$

Ricordando infine che $\sigma_{ij} = r_{ij} \cdot \sigma_i \cdot \sigma_j$ e che $\alpha_{ij} = r_{ij} \cdot \frac{\sigma_i}{\sigma_j}$ si ha

$$\begin{aligned} Cov(Y,V) &= r_{12}\sigma_1\sigma_2 - r_{23}\frac{\sigma_2}{\sigma_3} \cdot r_{13}\sigma_1\sigma_3 - r_{13}\frac{\sigma_1}{\sigma_3} r_{23}\sigma_2\sigma_3 + \\ &\quad + r_{13}\frac{\sigma_1}{\sigma_3} r_{23}\frac{\sigma_2}{\sigma_3}\sigma_{33} = \sigma_1\sigma_2\{r_{12} - r_{13}r_{23}\}. \end{aligned} \tag{12.8}$$

La varianza di Y non è altro che la varianza residua dopo aver interpolato X_1 in funzione di X_3. Pertanto ricordando che

$$(1 - r_{13}^2) = \frac{\text{Varianza residua}}{\text{Varianza di } X_1}$$

si ha

$$Var\,(Y) = \sigma_{11}\left\{1 - r_{13}^2\right\}. \tag{12.9}$$

Analogamente

$$Var\,(V) = \sigma_{22}\left\{1 - r_{23}^2\right\}. \tag{12.10}$$

In conclusione, risulta ancora

$$r_{12.3} = \frac{Cov(Y,V)}{\sigma(Y)\cdot\sigma(V)} = \frac{r_{12} - r_{13}\cdot r_{23}}{\sqrt{1-r_{13}^2}\,\sqrt{1-r_{23}^2}}.$$

Può valer la pena, a questo punto, mostrare che $\alpha_{12.3}$ non è altro che il coefficiente angolare della retta a minimi quadrati che interpola i residui Y in funzione dei residui V. Tale coefficiente angolare risulta pari a

$$\frac{Cov(Y,V)}{Var(V)} = \frac{\sigma_1\sigma_2\{r_{12} - r_{13}r_{23}\}}{\sigma_{22}\{1 - r_{23}^2\}} = \frac{\alpha_{12} - r_{13}r_{23}\cdot\dfrac{\sigma_1}{\sigma_2}}{1 - r_{23}^2}.$$

Si riottiene così la (11.4) che fornisce proprio $\alpha_{12.3}$.

Esempio 12.1.

Nel prospetto a pagina seguente si riporta la spesa settimanale totale X_1, la spesa settimanale per la carne X_2 e la spesa settimanale per il pesce X_3 relative a $N = 15$ famiglie. Le spese sono espresse in sterline.

Si vogliono determinare:
- i coefficienti di correlazione grezzi r_{12}, r_{13} e r_{23};
- il coefficiente di correlazione parziale $r_{23.1}$;
- i parametri del piano interpolatore

$$\hat{X}_1 = \alpha_1 + \alpha_{12.3}X_2 + \alpha_{13.2}\cdot X_3;$$

- l'indice di determinazione multiplo $I_{1.23}^2$.

Le medie aritmetiche dei tre caratteri risultano

X_1	X_2	X_3	$\hat{X}_1$	$X_1-\hat{X}_1$	$(X_1-\hat{X}_1)^2$
100	9	4,0	104,99	−4,99	24,9001
100	10	3,5	101,78	−1,78	3,1684
100	11	3,0	98,56	+1,44	2,0736
120	10	4,1	119,40	+0,60	0,3600
120	11	3,7	119,12	+0,88	0,7744
120	12	3,3	118,85	+1,15	1,3225
140	11	4,2	133,81	+6,19	38,3161
140	12	4,0	139,41	+0,59	0,3481
140	13	3,8	145,01	−5,00	25,1001
160	12	4,5	154,10	+5,90	34,8100
160	13	4,3	159,70	+0,30	0,0900
160	14	4,1	165,30	−5,30	28,0900
180	13	5,1	183,20	−3,20	10,2400
180	14	4,6	179,99	+0,01	0,0001
180	15	4,1	176,78	+3,22	10,3684
2.100	180	60,3	2.100,00	−20,28 +20,28	179,9618

$\overline{X}_1 = 140$ sterline; $\overline{X}_2 = 12$ sterline; $\overline{X}_3 = 4{,}02$ sterline.

Le devianze e le codevianze necessarie per la determinazione delle grandezze richieste si ottengono con i consueti procedimenti

$$Dev\,(X_2) = 40;\; Dev\,(X_3) = 3{,}844;\; Dev\,(X_1) = 12.000;$$

$$Cod\,(X_1, X_2) = 600;\; Cod\,(X_1, X_3) = 168;\; Cod\,(X_2, X_3) = 4{,}8.$$

I coefficienti di correlazione grezzi risultano

$$r_{12} = \frac{600}{\sqrt{12.000}\sqrt{40}} = +0{,}866;$$

$$r_{13} = \frac{168}{\sqrt{12.000}\sqrt{3{,}844}} = +0{,}782;$$

$$r_{23} = \frac{4{,}80}{\sqrt{40}\sqrt{3{,}844}} = +0{,}387.$$

I valori dei primi due coefficienti di correlazione sono conformi alle attese. Infatti, all'aumentare della spesa totale tende ad aumentare sia la spesa per la carne, sia la spesa per il pesce. La correlazione positiva fra spesa per il consumo di carne e spesa per il consumo di pesce sembra piuttosto sorprendente perché trattandosi di beni "succedanei", all'aumentare della spesa per uno di essi dovrebbe diminuire quella per l'altro. È bene però precisare che quest'ultimo comportamento presuppone la costanza della spesa totale. In effetti il valore $r_{23} = +\,0{,}387$ sconta anche il fatto che X_1 varia ed il suo valore influenza positivamente sia X_2 sia X_3. Per valutare correttamente il vicendevole legame che vi è fra X_2 e X_3 bisogna fare allora ricorso alla correlazione parziale $r_{23.1}$.

Il valore di $r_{23.1}$ risulta

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\cdot\sqrt{1-r_{13}^2}} = \frac{0{,}387 - 0{,}866 \cdot 0{,}782}{\sqrt{1-0{,}866^2}\,\sqrt{1-0{,}788^2}} =$$

$$= \frac{-0{,}2902}{\sqrt{0{,}25004}\,\sqrt{0{,}388}} = \frac{-0{,}2902}{0{,}50004 \cdot 0{,}623} = -\,0{,}932.$$

Il coefficiente di correlazione $r_{23.1}$ informa che, al netto della influenza della spesa totale X_1, all'aumentare della spesa per la carne X_2 tende a diminuire quella per il pesce X_3.

I coefficienti di regressione parziali risultano in forza delle (5.9) e (5.10)

$$\alpha_{12.3} = \frac{\sigma_{33}\sigma_{12} - \sigma_{13}\sigma_{23}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2} \tag{12.11}$$

$$\alpha_{13.2} = \frac{\sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2}. \tag{12.12}$$

Si ha così

$$\sigma_{33} = \frac{3.844}{15} = 0{,}25627; \quad \sigma_{22} = \frac{40}{15} = 2{,}667; \quad \sigma_{12} = \frac{600}{15} = 40;$$

$$\sigma_{13} = \frac{168}{15} = 11{,}2; \quad \sigma_{23} = \frac{4{,}8}{15} = 0{,}32.$$

Sostituendo i valori sopra calcolati nella (12.11) e nella (12.12) si ha

$$\alpha_{12.3} = 11{,}475 \quad \text{e} \quad \alpha_{13.2} = 29{,}376.$$

Il valore dell'intercetta è pari a

$$\alpha_1 = \overline{X}_1 - \alpha_{12.3}\overline{X}_2 - \alpha_{13.2}\overline{X}_3 =$$

$$= 140 - 11{,}475 \cdot 12 - 29{,}376 \cdot 4{,}02 = -115{,}792.$$

L'equazione del piano interpolatore a minimi quadrati risulta così

$$\hat{X}_1 = -115{,}792 + 11{,}475 \cdot X_2 + 29{,}376 \cdot X_3.$$

Il parametro $\alpha_{12.3} = 11{,}475$ informa che, tenendo fissa la spesa per il pesce, la spesa totale settimanale aumenta di 11,475 sterline per ogni incremento unitario di spesa per la carne. Analogamente $\alpha_{13.2}$ informa che a parità di spesa per la carne, la spesa totale settimanale aumenta di 29,376 sterline per ogni incremento di una sterlina nella spesa per il pesce. In questo caso il parametro $\alpha_1 = -115{,}792$ è privo di significato interpretativo reale.

Per ricavare l'indice di determinazione multiplo conviene calcolare innanzi tutto i valori $\hat{x}_{1i}$ e poi i residui $x_{1i} - \hat{x}_{1i}$. Detti valori sono stati riportati nella quarta e nella quinta colonna dell'ultimo prospetto. Nella sesta colonna sono riportati i quadrati dei residui da cui si ricava la devianza residua $\sum(x_{1i} - \hat{x}_{1i})^2 = 179{,}9618$. Si ha allora

$$I^2_{1.23} = 1 - \frac{\text{Dev. Residua}}{\text{Dev. totale}} = 1 - \frac{179{,}9618}{12.000} = 1 - 0{,}015 = 0{,}985.$$

In altre parole il piano interpolatore spiega quasi tutta la variabilità di X_1.

III) *All'espressione che fornisce il quadrato del coefficiente di correlazione parziale si può pervenire anche considerando il grado di miglioramento, inteso come riduzione relativa della varianza residua, che si ha nel passare dalla retta al piano a minimi quadrati.* In effetti la formula (10.9) coincide con $r^2_{13.2}$.

Le espressioni che forniscono i quadrati dei tre coefficienti di correlazione parziali sono

$$r^2_{12.3} = \frac{\{r_{12} - r_{13}r_{23}\}^2}{\{1 - r^2_{13}\}\{1 - r^2_{23}\}} = \frac{I^2_{1.23} - I^2_{1.3}}{1 - I^2_{1.3}} \qquad (12.13)$$

$$r^2_{13.2} = \frac{\{r_{13} - r_{12}r_{23}\}^2}{\{1 - r^2_{12}\}\{1 - r^2_{23}\}} = \frac{I^2_{1.23} - I^2_{1.2}}{1 - I^2_{1.2}} \qquad (12.14)$$

$$r^2_{23.1} = \frac{\{r_{23} - r_{12}r_{13}\}^2}{\{1 - r^2_{12}\}\{1 - r^2_{13}\}} = \frac{I^2_{2.31} - I^2_{2.1}}{1 - I^2_{2.1}}. \qquad (12.15)$$

Esempio 12.2.

Sui dati dell'Esempio 12.1 si vuole calcolare $r^2_{23.1}$ con la formula (12.15). Per ricavare $I^2_{2.31}$ conviene impiegare la formula

$$I^2_{2.31} = \frac{\alpha_{21.3} \cdot \sigma_{21} + \alpha_{23.1} \cdot \sigma_{23}}{\sigma_{22}}.$$

Si verifica immediatamente che

$$\alpha_{21.3} = \frac{\sigma_{33}\sigma_{21} - \sigma_{23}\sigma_{13}}{\sigma_{11}\sigma_{33} - \sigma^2_{13}}; \quad \alpha_{23.1} = \frac{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{31}}{\sigma_{11}\sigma_{33} - \sigma^2_{13}}.$$

Sostituendo alle varianze e covarianze sopra indicate i corrispondenti valori si ricavano

$$\alpha_{21.3} = 0{,}08378; \quad \alpha_{23.1} = -\,2{,}413,$$

e sostituendo questi valori nell'espressione di $I^2_{2.31}$ si ottiene

$$I^2_{2.31} = 0{,}96714.$$

Per ricavare $I^2_{2.1} = r^2_{21}$ si può applicare la formula

$$r^2_{21} = \frac{\{Cod\,(X_1, X_2)\}^2}{Dev(X_1) \cdot Dev(X_2)} = \frac{600^2}{12.000 \cdot 40} = 0{,}75.$$

In conclusione

$$r^2_{23.1} = \frac{0{,}96714 - 0{,}75}{1 - 0{,}75} = 0{,}8686.$$

Il valore trovato, che coincide (praticamente) con il quadrato di $r_{23.1} = -0{,}932$ ricavato in precedenza, indica che la varianza residua si riduce di circa l'87% nel passare dalla retta $\hat{X}_2$ verso X_1 al piano $\hat{X}_2$ verso X_1 e X_3.

13. *Linearizzazione*

Come si è evidenziato nelle pagine precedenti il metodo dei minimi quadrati è particolarmente semplice nel caso di interpolanti riconducibili a funzioni del tipo (2.8), cioè a funzioni lineari nei parametri.

In taluni casi interessa però impiegare modelli non lineari nei parametri. In economia viene spesso impiegato il modello

$$\hat{X}_1 = \alpha_1 X_2^{\alpha_2} X_3^{\alpha_3} \qquad (\alpha_1 > 0, X_2 > 0, X_3 > 0) \tag{13.1}$$

già introdotto nei paragrafi precedenti.

Volendo applicare il principio dei minimi quadrati si tratta di minimizzare la quantità

$$D(\alpha_1, \alpha_2, \alpha_3) = \sum (x_{1i} - \alpha_1 x_{2i}^{\alpha_2} x_{3i}^{\alpha_3})^2 \tag{13.2}$$

rispetto ai parametri α_1, α_2 e α_3.

Uguagliando a zero le derivate parziali $\frac{\delta D}{\delta \alpha_j}$ ($j = 1, 2, 3$) si ottiene il seguente sistema nelle incognite α_1, α_2 e α_3

$$\begin{cases} \sum (x_{1i} - \alpha_1 x_{2i}^{\alpha_2} x_{3i}^{\alpha_3}) \cdot x_{2i}^{\alpha_2} x_{3i}^{\alpha_3} = 0 \\ \sum (x_{1i} - \alpha_1 x_{2i}^{\alpha_2} x_{3i}^{\alpha_3}) \cdot \alpha_1 x_{2i}^{\alpha_2} x_{3i}^{\alpha_3} \log x_{2i} = 0 \\ \sum (x_{1i} - \alpha_1 x_{2i}^{\alpha_2} x_{3i}^{\alpha_3}) \cdot \alpha_1 x_{2i}^{\alpha_2} x_{3i}^{\alpha_3} \log x_{3i} = 0. \end{cases} \tag{13.3}$$

Si tratta di un sistema non lineare che si può risolvere iterativamente mediante il "calcolo numerico". Il seguente artificio permette di trovare un'interpolante assai prossima a quella a minimi quadrati.

Si considerino i logaritmi di entrambi i membri del modello (13.1)

$$\log \hat{X}_1 = \log \alpha_1 + \alpha_2 \log X_2 + \alpha_3 \log X_3.$$

Ponendo: $\log \hat{X}_1 = \hat{Y}_1$, $\log X_2 = Y_2$, $\log X_3 = Y_3$, $\log \alpha_1 = p_1$, $\alpha_2 = p_2$, $\alpha_3 = p_3$, la relazione precedente diventa

$$\hat{Y}_1 = p_1 + p_2 Y_2 + p_3 Y_3. \tag{13.4}$$

Per determinare i parametri p_1, p_2 e p_3 bisogna minimizzare la quantità

$$G(p_1, p_2, p_3) = \sum (y_{1i} - p_1 - p_2 y_{2i} - p_3 y_{3i})^2 \tag{13.5}$$

essendo $y_{1i} = \log x_{1i}$, $y_{2i} = \log x_{2i}$ e $y_{3i} = \log x_{3i}$.

Si tratta allora di ricavare i parametri del piano interpolatore relativo alle nuove variabili Y_1, Y_2 e Y_3. In forza delle formule (5.9), (5.10) e (5.11) i valori dei parametri che minimizzano la (13.5) sono

$$\hat{p}_2 = \frac{\sigma_{33}\sigma_{12} - \sigma_{13}\sigma_{23}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2} \tag{13.6}$$

$$\hat{p}_3 = \frac{\sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2} \tag{13.7}$$

$$\hat{p}_1 = \overline{y}_1 - \hat{p}_2\overline{y}_2 - \hat{p}_3\overline{y}_3; \tag{13.8}$$

essendo

$$\sigma_{33} = Var\ (Y_3) = Var\ (\log X_3);$$
$$\sigma_{22} = Var\ (Y_2) = Var\ (\log X_2);$$
$$\sigma_{12} = Cov\ (Y_1, Y_2) = Cov\ (\log X_1, \log X_2);$$
$$\sigma_{13} = Cov\ (Y_1, Y_3) = Cov\ (\log X_1, \log X_3);$$
$$\sigma_{23} = Cov\ (Y_2, Y_3) = Cov\ (\log X_2, \log X_3);$$
$$\overline{y}_1 = M_1(Y_1) = M_1(\log X_1);$$
$$\overline{y}_2 = M_1(Y_2) = M_1(\log X_2);$$
$$\overline{y}_3 = M_1(Y_3) = M_1(\log X_3).$$

Prima di passare ad un esempio numerico si tenga presente che i valori dei parametri p_1, p_2, p_3 sono tali da minimizzare la (13.5) e non la (13.2), conseguentemente le proprietà dei residui riguardano le variabili Y_1, Y_2 e Y_3 e non le variabili originarie X_1, X_2 e X_3. In particolare risulta

$$\begin{cases} \sum(\log x_{1i} - \log \hat{x}_{1i}) = 0 \\ \sum(\log x_{1i} - \log \hat{x}_{1i}) \log x_{2i} = 0 \\ \sum(\log x_{1i} - \log \hat{x}_{1i}) \log x_{3i} = 0. \end{cases} \tag{13.9}$$

Dai valori dei parametri p_1, p_2, p_3 si passa ai valori dei parametri richiesti α_1, α_2, α_3 con le relazioni

$$\alpha_1 = e^{p_1}, \quad \alpha_2 = p_2, \quad \alpha_3 = p_3.$$

Esempio 13.1.

Nel prospetto che segue è riportato il prodotto interno lordo delle industrie manifatturiere X_1, la quantità di lavoro X_2 e la quantità di capitale X_3. I dati riguardano l'economia italiana nel periodo 1951-62. Le variabili X_1 e X_3 sono espresse in miliardi di lire ai prezzi del 1954. La variabile X_2 è espressa in termini di giornate-operaio.

Anno	X_1	X_2	X_3	$\log X_3$	$(\log X_1)^2$	$\log X_2$	$(\log X_2)^2$
1951	2.802	5.651,1	11.368	3,44747	11,88504	3,75213	14,07850
1952	2.872	5.709,6	11.853	3,45818	11,95904	3,75661	14,11209
1953	3.153	5.834,0	12.304	3,49872	12,24107	3,76597	14,18250
1954	3.503	6.066,8	12.791	3,54444	12,56306	3,78296	14,31078
1955	3.818	6.188,1	13.306	3,58184	12,82955	3,79156	14,37591
1956	4.068	6.365,6	13.904	3,60938	13,02763	3,80384	14,46919
1957	4.344	6.454,2	14.539	3,63789	13,23424	3,80984	14,51490
1958	4.508	6.327,6	15.103	3,65398	13,35160	3,80124	14,44942
1959	5.041	6.444,9	15.746	3,70252	13,70863	3,80922	14,51013
1960	5.755	6.902,4	16.682	3,76005	14,13794	3,83900	14,73792
1961	6.301	7.347,3	17.902	3,79941	14,43551	3,86613	14,94694
1962	6.855	7.582,4	19.289	3,83601	14,71495	3,87981	15,05290
Totale	53.020	--	--	43,52989	158,08826	45,65831	173,74118

Secondo gli economisti Cobb e Douglas la produzione X_1 è legata alla quantità di lavoro X_2 ed al capitale X_3 secondo il modello (detto appunto di Cobb-Douglas)

$$\hat{X}_1 = \alpha_1 X_2^{\alpha_2} X_2^{\alpha_3}.$$

Si determinano, relativamente ai dati del prospetto, i valori dei parametri α_1, α_2 e α_3.

Occorre innanzi tutto linearizzare il modello passando alle variabili logaritmo

$$\log \hat{X}_1 = p_1 + p_2 \log X_2 + p_3 \log X_3.$$

Quindi bisogna ricavare

$$\Sigma \log x_{1i}; \quad \Sigma \log x_{2i}; \quad \Sigma \log x_{3i}; \quad \Sigma (\log x_{3i})^2; \quad \Sigma (\log x_{2i})^2;$$

$$\Sigma \log x_{1i} \cdot \log x_{2i}; \quad \Sigma \log x_{1i} \cdot \log x_{3i}; \quad \Sigma \log x_{2i} \cdot \log x_{3i}.$$

Tali sommatorie sono state determinate in parte nel prospetto precedente ed in parte in quello che segue.

Anno	$\log X_3$	$(\log X_2)^2$	$\log X_1 \cdot \log X_2$	$\log X_1 \cdot \log X_3$	$\log X_2 \cdot \log X_3$	$\hat{X}_1$	$X_1 - \hat{X}_1$
1951	4,05568	16,44857	12,93536	13,98184	15,21747	2.800	+ 2
1952	4,07383	16,59608	12,99104	14,08805	15,30377	3.000	− 128
1953	4,09005	16,72848	13,17608	14,30994	15,40298	3.207	− 54
1954	4,10690	16,86666	13,40847	14,55668	15,53625	3.465	+ 38
1955	4,12405	17,00777	13,58074	14,77166	15,63656	3.714	+ 104
1956	4,14314	17,16561	13,72951	14,95417	15,75984	4.024	+ 44
1957	4,16253	17,32669	13,85979	15,14284	15,85860	4.336	+ 8
1958	4,17906	17,46457	13,88967	15,27023	15,88562	4.556	− 48
1959	4,19717	17,61624	14,10369	15,54009	15,98793	4.897	+ 144
1960	4,22225	17,82738	14,43481	15,87584	16,20921	5.518	+ 237
1961	4,25290	18,08717	14,68900	16,15851	16,44226	6.325	− 24
1962	4,28531	18,36388	14,88297	16,43848	16,62617	7.193	− 338
Totale	49,89287	207,49910	165,68113	181,08833	189,86666	53.035	+ 577 − 592

Le medie $\bar{y}_1$, $\bar{y}_2$ e $\bar{y}_3$ dei logaritmi delle variabili risultano

$$\bar{y}_1 = \frac{1}{12} 43,52989 = 3,62749; \qquad \bar{y}_2 = \frac{1}{12} 45,65831 = 3,80486;$$

$$\bar{y}_3 = \frac{1}{12} 49,89287 = 4,15774.$$

Le varianze dei logaritmi delle variabili esplicative sono

$$\sigma_{33} = Var\,(\log X_3) = \frac{1}{12} \cdot 207,49910 - 4,15774^2 = 0,004797;$$

$$\sigma_{22} = Var\,(\log X_2) = \frac{1}{12} \cdot 173,74118 - 3,80486^2 = 0,001478.$$

Le covarianze fra i logaritmi delle variabili risultano

$$\sigma_{12} = Cov(\log X_1, \log X_2) = \frac{1}{12} \cdot 165,68113 - 3,62749 \cdot 3,80486 = 0,0046691;$$

$$\sigma_{13} = Cov(\log X_1, \log X_3) = \frac{1}{12} \cdot 181{,}08833 - 3{,}62749 \cdot 4{,}15774 = 0{,}0085334$$

$$\sigma_{23} = Cov(\log X_2, \log X_3) = \frac{1}{12} \cdot 189{,}86666 - 3{,}80486 \cdot 4{,}15774 = 0{,}00261.$$

I parametri p_2 e p_3 risultano

$$\hat{p}_2 = \frac{0{,}004797 \cdot 0{,}0046691 - 0{,}0085334 \cdot 0{,}00261}{0{,}001478 \cdot 0{,}004797 - 0{,}00261^2} = 0{,}4516518$$

$$\hat{p}_3 = \frac{0{,}001478 \cdot 0{,}0085334 - 0{,}0046691 \cdot 0{,}00261}{0{,}001478 \cdot 0{,}004797 - 0{,}00261^2} = 1{,}533164.$$

Infine, per il parametro p_1 si ha

$$\hat{p}_1 = 3{,}62749 - 0{,}4516518 \cdot 3{,}80486 - 1{,}533164 \cdot 4{,}15774 = -4{,}46547.$$

$$\alpha_1 = 10^{-4{,}46547} = 0{,}00003424.$$

L'equazione del modello Cobb-Douglas risulta

$$\hat{X}_1 = 0{,}00003424 \cdot X_2^{0{,}451618} \cdot X_3^{1{,}533164}.$$

Esempio 13.2.

Si vogliano ora interpretare i parametri del modello Cobb-Douglas ricavati nell'Esempio 13.1 e valutare il prodotto interno lordo delle industrie manifatturiere $\hat{X}_1$ fornito dal modello Cobb-Douglas, nonché gli scarti $X_1 - \hat{X}_1$.

Il parametro $\hat{\alpha}_2 \cong 0{,}452$ indica l'elasticità parziale del prodotto interno lordo $\hat{X}_1$ rispetto alla quantità di lavoro. In altre parole, nel periodo considerato le industrie manifatturiere italiane hanno fatto registrare un aumento dello 0,452% del prodotto interno lordo per un aumento dell'1% della quantità di lavoro, fermo restando il capitale X_3.

In modo analogo si interpreta il parametro $\hat{\alpha}_3$: tenendo fissa la quantità di lavoro, ad un aumento dell'1% del capitale è corrisposto un aumento dell'1,5% del prodotto interno lordo.

Il parametro $\hat{\alpha}_1 = 0{,}00003424$ non ha un particolare significato essendo un parametro di scala (di dimensionamento, secondo il linguaggio degli economisti).

I valori di $\hat{X}_1$ forniti dalla equazione

$$\hat{X}_1 = 0,00003424 X_2^{0,4516} X_3^{1,5331}$$

sono riportati nell'ultimo prospetto dell'Esempio 13.1.

Si osservi che $\sum x_{1i} = 53020$ risulta diversa da $\sum \hat{x}_{1i} = 53035$ non per effetto degli arrotondamenti effettuati nei calcoli, quanto per il fatto che la procedura impiegata garantisce che $\sum \hat{y}_{1i} = \sum y_{1i}$ (si veda la prima equazione del sistema (13.9)). Tuttavia, non si può non rilevare che la discrepanza fra $\sum \hat{y}_{1i}$ e $\sum y_{1i}$ risulta del tutto trascurabile.

Gli scarti $X_1 - \hat{X}_1$ sono anch'essi riportati nel secondo prospetto dell'Esempio 13.1. La loro somma risulta pari a – 15.

Esempio 13.3.

Si riprendano gli Esempi 13.1 e 13.2 per valutare la bontà di adattamento del modello Cobb-Douglas.

Si può considerare innanzi tutto la successione del segno degli scarti. L'analisi non indica nessuna "tendenziosità".

Si possono poi valutare gli indici A_1 e A_1'. Il numeratore di A_1 risulta

$$\sum |x_{1i} - \hat{x}_{1i}| = 577 + 592 = 1.169.$$

Pertanto

$$A_1 = \frac{1}{12} 1.169 = 97,42.$$

Il valore medio aritmetico $\overline{X}_1$ risulta

$$\overline{X}_1 = \frac{1}{12} 53.020 = 4.418,3$$

da cui

$$A_1' = \frac{97,42}{4.418,3} = 0,022.$$

8. *Coefficiente di correlazione multiplo*

Per coefficiente di correlazione multiplo si intende il coefficiente di correlazione lineare fra le variabili X_1 e $\hat{X}_1 = \alpha_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k$. Il coefficiente di correlazione multiplo si indica con $R_{1.23\ldots k}$ ed è per definizione dato da

$$R_{1.23\ldots k} = \frac{Cov(X_1, \hat{X}_1)}{\sigma(X_1)\cdot\sigma(\hat{X}_1)}. \qquad (8.1)$$

Si dimostra facilmente, seguendo la stessa procedura vista nel Capitolo I, che

$$\begin{aligned} Cov(X_1, \hat{X}_1) &= \alpha_2\sigma_{12} + \ldots + \alpha_k\sigma_{1k} = Var(\hat{X}_1) = \\ &= \text{Varianza spiegata dall'iperpiano.} \end{aligned} \qquad (8.2)$$

Pertanto

$$R_{1.23\ldots k} = \frac{Var(\hat{X}_1)}{\sigma(X_1)\cdot\sigma(\hat{X}_1)} = \frac{\sigma(\hat{X}_1)}{\sigma(X_1)} \geq 0. \qquad (8.3)$$

È immediato verificare che anche nel caso generale $k > 3$ vale l'uguaglianza

$$R^2_{1.23\ldots k} = \frac{\{Var(\hat{X}_1)\}^2}{Var(X_1)\cdot Var(\hat{X}_1)} = \frac{Var(\hat{X}_1)}{Var(X_1)} = \frac{\text{Varianza spiegata}}{\text{Varianza totale}} = I^2_{1.23\ldots k}.$$

9. *Analisi grafiche sui residui*

Utili informazioni sull'adeguatezza dell'interpolante

$$\hat{X}_1 = \alpha_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k$$

a descrivere la variabile X_1 si possono avere esaminando alcuni grafici dei residui $X_1 - \hat{X}_1$.

Scopo di queste analisi non è quello di valutare l'ordine di grandezza dei residui o di valutare la parte di variabilità di X_1 spiegata dal modello, bensì quello di scoprire se i residui presentano o meno "tendenziosità".

In altre parole gli scarti $(X_1 - \hat{X}_1)$ rappresentano il 2,2% dell'ordine di grandezza di X_1.

In questo tipo di analisi può essere utile valutare anche gli scarti relativi

$$\rho_i = \frac{x_{1i} - \hat{x}_{1i}}{\hat{x}_{1i}}.$$

I valori ρ_i sono riportati nel prospetto che segue.

Anno	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962
ρ_i	+0,001	−0,043	−0,017	+0,011	+0,028	+0,011	+0,002	−0,011	+0,029	+0,043	−0,004	−0,047

La successione degli scarti relativi indica che il valore assoluto degli stessi non tende ad aumentare all'aumentare di $\hat{X}_1$.

In conclusione, nel periodo considerato, si può ritenere il modello Cobb-Douglas idoneo a descrivere il prodotto interno lordo delle industrie manifatturiere in funzione della quantità di capitale e della quantità di lavoro.

Esempio 13.4.

Si consideri la superficie generata dal modello Cobb-Douglas (13.1)

$$\hat{X}_1 = \alpha_1 X_2^{\alpha_2} X_3^{\alpha_3}.$$

Si valuti il valore di $\hat{X}_1$ per $X_2 = M_0(X_2)$ (media geometrica di X_2) e $X_3 = M_0(X_3)$ (media geometrica di X_3).

Lo svolgimento dell'esempio viene lasciato al lettore, col suggerimento di considerare i logaritmi della (13.1).

14. *L'indipendenza in media ed il piano a minimi quadrati nel caso di una tabella a triplice entrata*

Nel caso in cui il numero delle osservazioni N sia molto elevato i dati della rilevazione statistica si possono presentare in una tabella a triplice entrata. A tal proposito, si supponga per semplicità che

Un *primo grafico* si può ottenere ponendo in ascissa la variabile $\hat{X}_1$ ed in ordinata i residui $X_1 - \hat{X}_1$. Nel valutare questo grafico si tenga comunque sempre presente che la media dei residui è nulla e che gli scarti sono incorrelati con la variabile $\hat{X}_1$.

Ecco ora due situazioni: nella prima gli scarti sembrano mostrare una certa "tendenziosità" (Figura 9.1), nell'altra invece sembrano mostrare una assenza di tendenziosità (Figura 9.2).

Figura 9.1. – *Residui con tendenziosità*

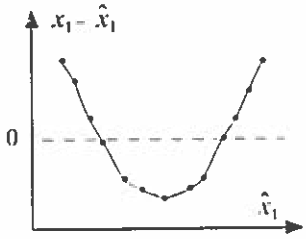
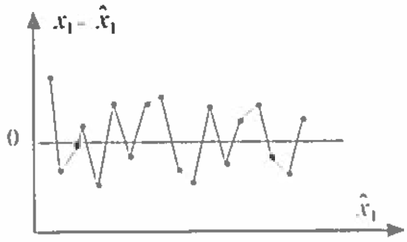


Figura 9.2. – *Residui senza tendenziosità*



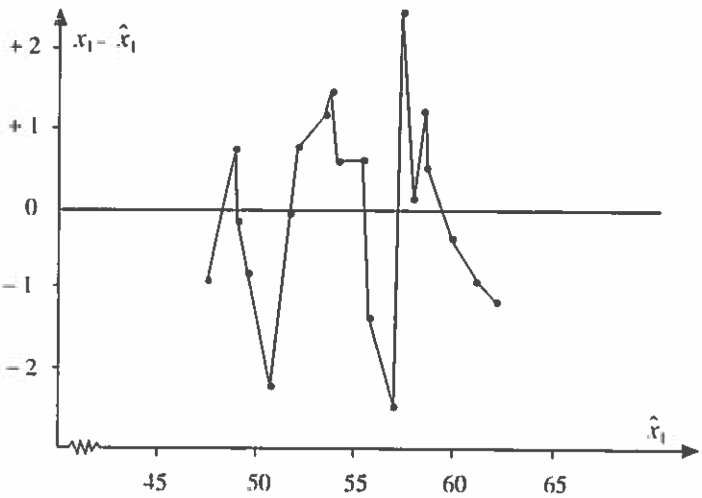
Esempio 9.1.

Si riconsiderano ora i dati dell'Esempio 7.1 e mediante il grafico $\{\hat{x}_{1i}, (x_{1i} - \hat{x}_{1i})\}$ verrà valutato se i residui presentino "tendenzosità".

Ordinando le coppie $\{\hat{x}_{1i}, (x_{1i} - \hat{x}_{1i})\}$ secondo il valore di $\hat{x}_{1i}$ si ottengono il prospetto che segue e la successiva Figura 9.3.

$\hat{X}_1$	$X_1 - \hat{X}_1$	Anni
47,57	− 0,87	1932
48,98	+ 0,72	1929
49,05	− 0,15	1930
49,30	− 0,60	1928
50,80	− 2,20	1931
51,60	− 0,10	1933
52,12	+ 0,78	1935
53,57	+ 1,33	1940
53,68	+ 1,52	1937
54,11	+ 0,59	1939
55,26	+ 0,64	1934
55,70	− 1,30	1938
56,95	− 2,45	1927
57,03	+ 2,47	1924
57,91	+ 0,19	1936
58,37	+ 1,23	1923
58,41	+ 0,69	1922
59,94	− 0,44	1925
61,17	− 0,87	1926
62,08	− 1,18	1941

Figura 9.3. – *Residui $x_1 - \hat{x}_1$ rispetto a $\hat{x}_1$*



Il prosp
fra residui

Esempi

Anche
Capitolo I
sentino ter
Il grafi
totale $\hat{X}_1$ e
esso non s

Figura 9.4.

Nei casi
essere utile
posti alla v

Esempio

Riconsic
si valuterà
tempo.

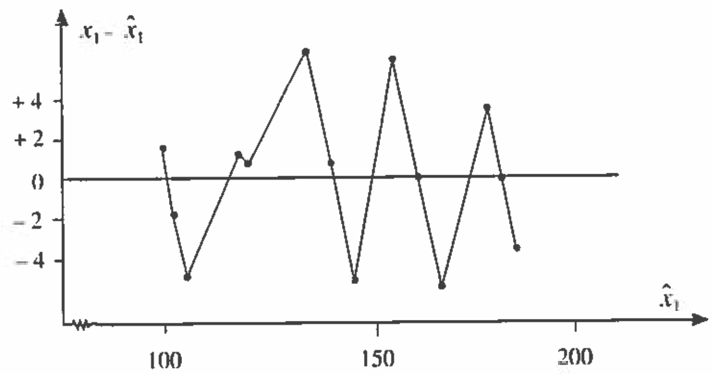
Il prospetto e la Figura 9.3 sembrano indicare che non vi sia tendenziosità fra residui e valori interpolati.

Esempio 9.2.

Anche in questo caso vengono riconsiderati i dati dell'Esempio 12.1 del Capitolo I e mediante il grafico $\{\hat{x}_{1i},(x_{1i}-\hat{x}_{1i})\}$ si valuterà se i residui presentino tendenziosità.

Il grafico che segue (Figura 9.4) riporta in ascissa la spesa settimanale totale $\hat{X}_1$ ed in ordinata i residui fra spesa totale effettiva X_1 e spesa totale $\hat{X}_1$; esso non sembra mostrare particolare tendenziosità nei residui.

Figura 9.4. – *Residui* $(x_1-\hat{x}_1)$ *rispetto a* $\hat{x}_1$



Nei casi in cui i dati facciano riferimento a rilevazioni temporali, può essere utile l'analisi di un *secondo grafico* in cui i residui vengano contrapposti alla variabile tempo.

Esempio 9.3.

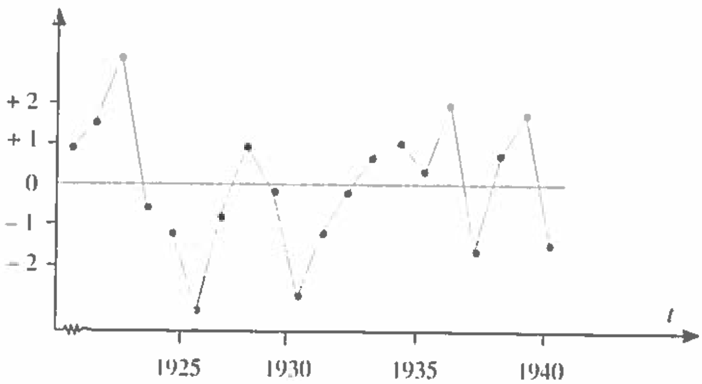
Riconsiderando i dati dell'Esempio 9.1, mediante il grafico $\{t,(x_{1t}-\hat{x}_{1t})\}$, si valuterà ora se i residui presentino tendenziosità rispetto alla variabile tempo.

Ordinando gli scarti $X_t - \hat{X}_t$ secondo la variabile tempo t si ha il prospetto che segue.

t	$X_t - \hat{X}_t$
1922	+ 0,69
1923	+ 1,23
1924	+ 2,47
1925	− 0,44
1926	− 0,87
1927	− 2,45
1928	− 0,60
1929	+ 0,72
1930	− 0,15
1931	− 2,20
1932	− 0,87
1933	− 0,10
1934	+ 0,60
1935	+ 0,78
1936	+ 0,19
1937	+ 1,52
1938	− 1,30
1939	+ 0,59
1940	+ 1,33
1941	− 1,18

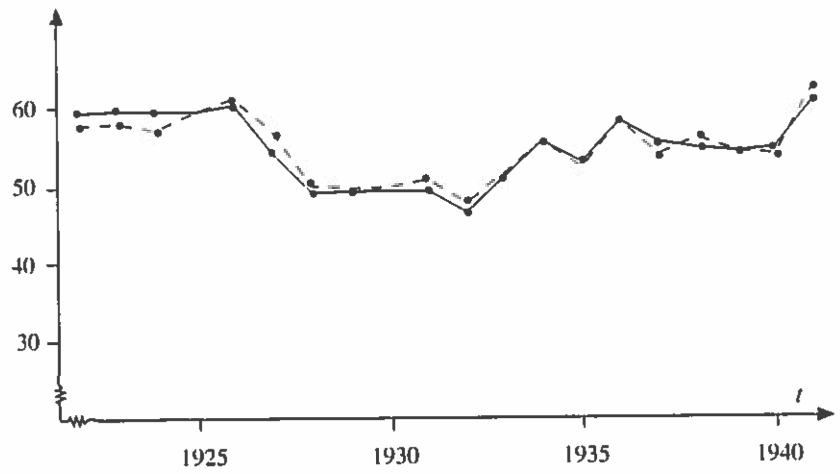
La Figura 9.5 sembra indicare un andamento "ciclico" dei residui rispetto al tempo; ovvero, i residui sembrano succedersi, rispetto al tempo, secondo qualche "regola".

Figura 9.5. – *Residui* $(x_t - \hat{x}_t)$ *rispetto al tempo* t



Per meglio valutare l'importanza di quest'ultimo aspetto può essere utile riportare in ordinata sia i consumi reali di carne bovina X_1 che quelli forniti dal modello $\hat{X}_1$.

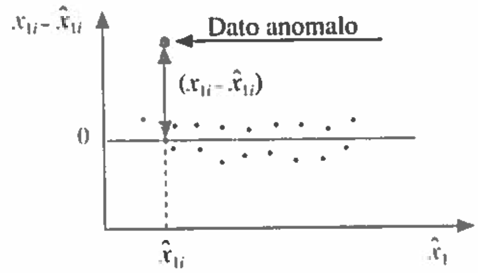
Figura 9.6. – *Consumi reali X_1 (—) di carne bovina e consumi teorici $\hat{X}_1$ (– –) forniti dal modello* $\hat{X}_1 = 90{,}8122 - 1{,}85\ X_2 + 0{,}0832\ X_3 - 0{,}4151\ X_4$



La Figura 9.6 pur evidenziando qualche sistematicità nei residui, specie nei primi anni del periodo osservazionale, conferma la limitata entità dei residui rispetto all'ordine di grandezza del fenomeno e quindi suggerisce, contrariamente a quanto poteva sembrare dal precedente grafico, che il modello ottenuto con la regressione multipla è adeguato a rappresentare i consumi di carne bovina che si sono avuti negli Stati Uniti nel periodo considerato.

10. *Ricerca di dati anomali*

Dopo aver determinato i parametri dell'iperpiano si calcolano gli indici che misurano la bontà dell'adattamento e spesso si effettuano anche analisi grafiche sui residui. Da queste analisi può accadere di individuare qualche dato che sembra particolarmente lontano dal modello. Sorge il problema di valutare se si tratta di dati "anomali" o meno.



Per dato "anomalo" si intende un dato che nel grafico è rappresentato da un punto con uno scarto $x_{1i}-\hat{x}_{1i}$ molto elevato. Se un dato è riconosciuto come "anomalo" viene rimosso e sui dati rimanenti si rideterminano i parametri del modello.

In questo paragrafo si fa solo un cenno alla problematica della ricerca di dati anomali rimandando a testi specializzati l'approfondimento della metodologia specifica.

Un modo elementare per valutare se un dato è anomalo è quello di confrontare il suo residuo $(x_{1i}-\hat{x}_{1i})$ con A_2, oppure di confrontare il residuo relativo $\dfrac{x_{1i}-\hat{x}_{1i}}{\hat{x}_{1i}}$ con A_2'. Ovviamente un dato è tanto più sospettato di essere anomalo quanto più $\left|\dfrac{x_{1i}-\hat{x}_{1i}}{A_2}\right|$ è elevato, o quanto più $\left\{\left|\dfrac{x_{1i}-\hat{x}_{1i}}{\hat{x}_{1i}}\right|\right\}/A_2'$ è elevato.

Esempio 10.1.

Riconsiderando l'Esempio 8.1 del Capitolo I si verificherà se sono presenti dati anomali.

Dai dati dell'Esempio 8.1 si ricava innanzitutto il valore dell'indice

$$A_2=\left\{\frac{1}{20}992,44\right\}^{\frac{1}{2}}=7,044.$$

Quindi si considera lo scarto assoluto massimo che risulta $\max\limits_{i=1,\dots,20}|x_{1i}-\hat{x}_{1i}|=14,6$ e lo si confronta con A_2. Si ha