

13

La regressione lineare multipla

Introduzione 2

- 13.1 Il modello di regressione multipla 2
- 13.2 L'analisi dei residui nel modello di regressione multipla 9
- 13.3 Il test per la verifica della significatività del modello
di regressione lineare multipla 11
- 13.4 Inferenza sui coefficienti di regressione della popolazione 14
- 13.5 La verifica di ipotesi sulle proporzioni nel modello di regressione multipla 17
- 13.6 Il modello di regressione quadratica 23
- 13.7 I modelli con variabili Dummy 31
- 13.8 La multicollinearità 38
- 13.9 Costruzione del modello 39
- 13.10 Le trappole dell'analisi di regressione 48

Riepilogo del capitolo 48

- A13.1 L'uso di Microsoft Excel nei modelli di regressione multipla 54

OBIETTIVI DEL CAPITOLO

- ✓ *Sviluppare il modello di regressione multipla come estensione del modello di regressione semplice*
- ✓ *Valutare il contributo di ciascuna variabile indipendente*
- ✓ *Calcolare il coefficiente di determinazione parziale*
- ✓ *Sviluppare il modello di regressione quadratico*
- ✓ *Introdurre tra le variabili esplicative le variabili qualitative (dummy)*
- ✓ *Illustrare i metodi per la selezione automatica di un modello di regressione*

Introduzione

Nel Capitolo 12 abbiamo preso in considerazione il modello di regressione lineare semplice, in cui una sola variabile indipendente o esplicativa X viene usata per prevedere il valore della variabile dipendente o risposta Y . Spesso, tuttavia, si può ottenere un modello migliore prendendo in considerazione più di una variabile esplicativa. Per questo motivo, in questo capitolo intendiamo estendere l'analisi del capitolo precedente introducendo il modello **di regressione multipla** in cui si fa ricorso a più variabili esplicative per effettuare previsioni su una variabile dipendente.

◆ **APPLICAZIONE:** *Previsione delle vendite di Omnipower*

Il prezzo e la spesa in attività promozionali sono due dei fattori che determinano in maniera preponderante le vendite di un prodotto. Supponete che una grande catena di negozi alimentari operante su scala nazionale intenda introdurre una barretta energetica di basso prezzo, chiamata Omnipower. Le barrette energetiche contengono grassi, carboidrati e calorie e forniscono rapidamente energie ai corridori, agli scalatori e agli atleti in genere impegnati in lunghe ed estenuanti attività sportive. Le vendite delle barrette energetiche sono esplose negli ultimi anni e il grande magazzino ritiene che vi possa essere un buon mercato per la Omnipower. Prima di introdurre la barretta in tutti i magazzini, la divisione di marketing della catena intende stabilire l'effetto che il prezzo e le promozioni all'interno dei negozi possono avere sulle vendite. ◆

◆ 13.1 **SVILUPPARE IL MODELLO DI REGRESSIONE MULTIPLA**

Un campione di 34 negozi della catena viene selezionato per una ricerca di mercato sulla Omnipower. I negozi hanno tutti approssimativamente il medesimo volume di vendite mensili. Si prendono in considerazione due variabili indipendenti – il prezzo in centesimi di una barretta Omnipower (X_1) e la spesa mensile per le attività promozionali, espressa in dollari, (X_2). La spesa promozionale comprende la spesa per i cartelli pubblicitari, i tagliandi di sconto e i campioni gratuiti. La variabile dipendente Y è il numero di barrette di Omnipower vendute in un mese. Nella Tabella 13.1 si riportano i valori osservati per le tre variabili considerate.

Tabella 13.1 *Vendite mensili, prezzo e spese promozionali di Ominipower*

NEGOZIO	VENDITE	PREZZO	PROMOZIONE	NEGOZIO	VENDITE	PREZZO	PROMOZIONE
1	4141	59	200	18	2730	79	400
2	3842	59	200	19	2618	79	400
3	3056	59	200	20	4421	79	400
4	3519	59	200	21	4113	79	600
5	4226	59	400	22	3746	79	600
6	4630	59	400	23	3532	79	600
7	3507	59	400	24	3825	79	600
8	3754	59	400	25	1096	99	200
9	5000	59	600	26	761	99	200
10	5120	59	600	27	2088	99	200
11	4011	59	600	28	820	99	200
12	5015	59	600	29	2114	99	400
13	1916	79	200	30	1882	99	400
14	675	79	200	31	2159	99	400
15	3636	79	200	32	1602	99	400
16	3224	79	200	33	3354	99	600
17	2295	79	400	34	2927	99	600



DATASET
OMNI

I coefficienti della regressione

Al fine di tener conto di più di una variabile indipendente, estendiamo il modello di regressione lineare semplice dell'equazione (12.1) supponendo che tra la variabile dipendente e ciascuna delle variabili esplicative vi sia una relazione lineare. Nel caso di p variabili esplicative, il modello di regressione multipla assume la seguente espressione:

Il modello di regressione multipla con p variabili indipendenti

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_p X_{pi} + \epsilon_i \quad (13.1)$$

dove

β_0 = intercetta

β_1 = inclinazione di Y rispetto alla variabile X_1 tenendo costanti le variabili X_2, X_3, \dots, X_p .

β_2 = inclinazione di Y rispetto alla variabile X_2 tenendo costanti le variabili X_1, X_3, \dots, X_p .

β_3 = inclinazione di Y rispetto alla variabile X_3 tenendo costanti le variabili $X_1, X_2, X_4, \dots, X_p$.

β_p = inclinazione di Y rispetto alla variabile X_p tenendo costanti le variabili $X_1, X_2, X_3, \dots, X_{p-1}$.

ϵ_i = errore in corrispondenza dell'osservazione i .

Nel caso di due variabili esplicative, il modello di regressione multipla è espresso come segue

Il modello di regressione multipla con due variabili indipendenti

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \quad (13.2)$$

dove

β_0 = intercetta

β_1 = inclinazione di Y rispetto alla variabile X_1 tenendo costante la variabile X_2

β_2 = inclinazione di Y rispetto alla variabile X_2 tenendo costante la variabile X_1

ϵ_i = errore in corrispondenza dell'osservazione i

Confrontiamo questo modello con il modello di regressione lineare semplice dell'equazione (12.1) dato da:

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

Nel modello lineare semplice, l'inclinazione β_1 rappresenta la variazione che la variabile Y presenta in corrispondenza di una variazione unitaria di X . Non si prende in considerazione nessun'altra variabile oltre all'unica variabile indipendente inclusa nel modello. Nel modello di regressione multipla dell'equazione (13.2) l'inclinazione β_1 ci dice come varia Y in corrispondenza di una variazione unitaria della variabile X_1 , quando, tuttavia, si tiene conto anche degli effetti della variabile X_2 . Parleremo di **coefficiente netto di regressione**.

Come nella regressione semplice, i coefficienti di regressione campionari (b_0 , b_1 e b_2) vengono usati come stimatori dei corrispondenti parametri della popolazione (β_0 , β_1 e β_2). Pertanto, l'espressione campionaria dell'equazione di un modello di regressione multipla con due variabili esplicative ha la forma seguente.

L'equazione della regressione multipla con due variabili esplicative

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} \quad (13.3)$$

I valori dei coefficienti di regressione campionari si possono calcolare con il metodo dei minimi quadrati, ricorrendo a pacchetti statistici o a fogli elettronici come Microsoft Excel. Nella Figura 13.1 si riporta l'output parziale ottenuto da Excel per i dati relativi alle vendite della barretta Omnipower della Tabella 13.1.

RIQUADRO A

	A	B	C	D	E	F	G
1	Regression of Sales of OmniPower Bars						
2							
3	<i>Regression Statistics</i>						
4	Multiple R	0.870474549					
5	R Square	0.757725941					
6	Adjusted R Square	0.742095357					
7	Standard Error	638.0652881					
8	Observations	34					
9							
10	<i>ANOVA</i>						
11		df	SS	MS	F	Significance F	
12	Regression	2	39472730.77	19736365.39	48.47713433	2.86258E-10	
13	Residual	31	12620946.67	407127.3119			
14	Total	33	52093677.44				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	5837.520759	628.150225	9.29319218	1.79101E-10	4556.399214	7118.642304
18	Price	-53.21733631	6.852220559	-7.766436566	9.20016E-09	-67.19254007	-39.24213255
19	Promotion	3.613058036	0.685222056	5.27282799	9.82196E-06	2.215537659	5.010578412

FIGURA 13.1

Regressione per le vendite delle barrette energetiche Omnipower.

	A	B	C
23	RESIDUAL OUTPUT		
24			
25	<i>Observation</i>	<i>Predicted Errors</i>	<i>Residuals</i>
26	1	3420.309524	720.6904762
27	2	3420.309524	421.6904762
28	3	3420.309524	-364.3095238
29	4	3420.309524	98.69047619
30	5	4142.921131	83.07886905
31	6	4142.921131	487.078869
32	7	4142.921131	-635.921131
33	8	4142.921131	-388.921131
34	9	4865.532738	134.4672619
35	10	4865.532738	254.4672619
36	11	4865.532738	-854.5327381
37	12	4865.532738	149.4672619
38	13	2355.962798	-439.9627976
39	14	2355.962798	-1680.962798
40	15	2355.962798	1280.037202
41	16	2355.962798	868.0372024
42	17	3078.574405	-783.5744048
43	18	3078.574405	-348.5744048
44	19	3078.574405	-460.5744048
45	20	3078.574405	1342.425595
46	21	3801.186012	311.8139881
47	22	3801.186012	-55.1860119
48	23	3801.186012	-263.1860119
49	24	3801.186012	23.8139881
50	25	1291.616071	-195.6160714
51	26	1291.616071	-530.6160714
52	27	1291.616071	796.3839286
53	28	1291.616071	-471.6160714
54	29	2014.227679	99.77232143
55	30	2014.227679	-132.2276786
56	31	2014.227679	144.7723214
57	32	2014.227679	-412.2276786
58	33	2736.839286	617.1607143
59	34	2736.839286	190.1607143

RIQUADRO B

In base alla Figura 13.1, i valori dei coefficienti di regressione campionari sono:

$$b_0 = 5,837.52 \quad b_1 = -53.2173 \quad b_2 = 3.6131$$

Pertanto, il modello di regressione multipla stimato è:

$$\hat{Y}_i = 5837.52 - 53.2173X_{1i} + 3.6131X_{2i}$$

in cui:

\hat{Y}_i = vendite mensili medie di Omnipower previste per il negozio i

X_{1i} = prezzo (in centesimi) di Omnipower per il negozio i

X_{2i} = spesa (in dollari) per la promozione della Omnipower nel negozio i

L'intercetta campionaria b_0 , pari a 5.83752, rappresenta il numero di barrette di Omnipower che ci si aspetterebbe di vendere ogni mese se il prezzo e l'ammontare totale speso per l'attività promozionale fossero entrambi uguali a \$ 0.00. Tali valori tuttavia, al di fuori del range dei valori osservati sia per il prezzo che per la spesa promozionale, non hanno alcun senso.

L'inclinazione delle vendite di Omnipower rispetto al prezzo ($b_1 = -53.2173$) ci dice che, per un dato ammontare della spesa per l'attività promozionale, si dovrebbero vendere 53.2173 barrette in meno per ogni centesimo di aumento del prezzo. L'inclinazione delle vendite rispetto alla spesa per l'attività promozionale ($b_2 = 3.6131$) ci dice che, per un dato prezzo, si dovrebbero vendere 3.6131 barrette in più per ogni centesimo speso in più in attività promozionali. Tali stime permettono alla divisione di marketing di prevedere l'effetto che eventuali decisioni in merito al prezzo e all'attività promozionale possono avere sulle vendite della barretta Omnipower. Per esempio, in base al modello stimato, si ritiene che per un dato ammontare della spesa promozionale, una riduzione di 10 centesimi del prezzo

della barretta determinerebbe un aumento del numero di barrette vendute pari a 532.173. Dall'altro lato, per un dato prezzo, un aumento della spesa promozionale di \$ 100 determinerebbe un aumento del numero di barrette vendute pari a 361.31 barrette.

COMMENTO: Interpretazione delle inclinazioni nel modello di regressione multipla

Abbiamo visto che i coefficienti in un modello di regressione multipla si devono considerare come coefficienti di regressione netti: essi misurano la variazione della variabile risposta Y in corrispondenza della variazione di una delle variabili esplicative, quando si tengono costanti le altre. Per esempio, nello studio delle vendite della barretta Omnipower, abbiamo affermato che, per un dato negozio, in corrispondenza di una riduzione di un centesimo del prezzo si venderebbero 53.22 barrette in più, per un dato ammontare della spesa promozionale. Analogamente, i valori dei coefficienti di regressione si potrebbero interpretare prendendo in considerazione più negozi simili, tutti con un medesimo ammontare della spesa promozionale. Per tali negozi, si prevede che una riduzione del prezzo della barretta aumenterebbe le vendite di 53.22 barrette.

In maniera analoga, l'inclinazione delle vendite rispetto alla spesa promozionale, può essere interpretata nella prospettiva di diversi negozi simili, in cui la Omnipower ha un medesimo prezzo. Per questi negozi si ritiene che la vendita di barrette Omnipower aumenterebbe di 3.61 barrette al mese per ogni dollaro in più speso in attività promozionali.

La previsione

Il modello di regressione stimato può ora essere impiegato per la previsione dell'ammontare mensile delle vendite e per la costruzione di intervalli di confidenza per le quantità non note.

Supponete, ad esempio, di voler prevedere il numero di barrette di Omnipower vendute in un negozio nel quale per un mese si sia praticato il prezzo di 79 centesimi e si sia effettuata una spesa di 400\$ per l'attività promozionale. Il modello di regressione stimato ha la seguente forma:

$$\hat{Y}_i = 5837.52 - 53.2173X_{1i} + 3.6131X_{2i}$$

Pertanto ponendo $X_{1i} = 79$ e $X_{2i} = 400$, si ha

$$\hat{Y}_i = 5837.52 - 53.2173(79) + 3.6131(400)$$

da cui:

$$\hat{Y}_i = 3078.57$$

Stimiamo che in media in negozi in cui il prezzo della barretta è di 79 centesimi e che spendono \$400 in attività promozionali verrebbero vendute 3078.57 barrette.

I coefficienti di determinazione

Nel paragrafo 9.3 abbiamo visto che il coefficiente di determinazione consente di valutare la bontà del modello di regressione stimato. Nel modello di regressione multipla, dal momento che si è in presenza di almeno due variabili esplicative, il **coefficiente di determinazione** rappresenta la proporzione di variabilità della Y spiegata dalle variabili esplicative.

Il coefficiente di determinazione

Il coefficiente di determinazione è dato dal rapporto tra la somma dei quadrati della regressione e la somma totale dei quadrati.

$$r_{Y,12}^2 = \frac{SQR}{SQT} \tag{13.4}$$

dove

SQR = somma dei quadrati della regressione

SQT = somma totale dei quadrati

Nell'esempio relativo alla barretta Omnipower, in base alla Figura 13.1, $SQR = 39,472,730.77$ e $SQT = 52,093,677.44$. Pertanto:

$$r_{Y.12}^2 = \frac{SQR}{SQT} = \frac{39,472,730.77}{52,093,677.44} = 0.7577$$

Il coefficiente di determinazione è uguale a 0.7577 e, quindi, ci dice che il 75.77% della variabilità delle vendite di Omnipower è spiegato dal prezzo e dalle spese promozionali.

Tuttavia, alcuni ricercatori ritengono che quando si ricorre a un modello di regressione multipla, sia opportuno fare uso di un indice che tenga conto anche del numero di variabili esplicative incluse nel modello e dell'ampiezza del campione, l' r^2 **corretto**. Il ricorso a questo tipo di indice si rende necessario soprattutto qualora si vogliano confrontare modelli di regressione che intendono spiegare la medesima variabile dipendente, impiegando un numero diverso di variabili esplicative. L' r^2 **corretto** è dato dalla seguente espressione:

L' r^2 corretto

$$r_{\text{adj}}^2 = 1 - \left[(1 - r_{Y.12 \dots p}^2) \frac{n - 1}{n - p - 1} \right] \quad (13.5)$$

Dove p = numero delle variabili esplicative incluse nel modello.

Per i dati relativi alle vendite della barretta Omnipower, poiché $r_{Y.12}^2 = 0.7577$, $n = 34$ e $p = 2$,

$$\begin{aligned} r_{\text{adj}}^2 &= 1 - \left[(1 - r_{Y.12}^2) \frac{34 - 1}{34 - 2 - 1} \right] \\ &= 1 - \left[(1 - 0.7577) \frac{33}{31} \right] \\ &= 1 - 0.2579 \\ &= 0.7421 \end{aligned}$$

Pertanto il 74.21% della variabilità delle vendite può essere spiegato dal modello proposto, tenuto conto delle numero di previsori e dell'ampiezza campionaria.

Esercizi del paragrafo 13.1

- 13.1 Prendete in considerazione il seguente modello stimato di regressione multipla:

$$\hat{Y}_i = 10 + 5X_{1i} + 3X_{2i} \quad \text{e} \quad r_{Y.12}^2 = 0.60$$

- Fornite una spiegazione delle inclinazioni della variabile dipendente rispetto a ciascuna delle variabili esplicative.
- Fornite una spiegazione dell'intercetta.
- Fornite una spiegazione del coefficiente di determinazione $r_{Y.12}^2$.

Nota: Risolvete i seguenti esercizi facendo uso di Microsoft Excel

- 13.2 Un ricercatore di mercato per un'impresa produttrice di scarpe deve valutare l'opportunità di produrre un nuovo tipo di scarpe da corsa. A tale scopo, intende stabilire quali variabili si possano impiegare per prevedere la resistenza delle scarpe. Il ricercatore decide di prendere in considerazione, come variabili esplicative, X_1 (FOREIMP), una misura della capacità di assorbimento degli shock nella parte anteriore della scarpa, e X_2 (MIDSOLE), una misura della capacità di assorbimento degli urti, mentre assume come variabile dipendente Y (LTIMP), una misura della capacità di assorbimento degli shock nel lungo periodo. Si seleziona per il test un campione di 15 tipi di scarpe da corsa attualmente prodotte dall'impresa. Con il ricorso a Excel si ottiene il seguente output

ANALISI					
VARIANZA	<i>GDL</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	SIGNIFICATIVITÀ <i>F</i>
Regressione	2	12.61020	6.30510	97.69	0.0001
Residuo	12	0.77453	0.06454		
Totale	14	13.38473			

VARIABILE	COEFFICIENTI	ERRORE STANDARD	STAT <i>t</i>	VALORE DI
				SIGNIFICATIVITÀ
Intercetta	-0.02686	0.06905	-0.39	
Foreimp	0.79116	0.06295	12.57	0.0000
Midsole	0.60484	0.07174	8.43	0.0000

- Supponendo che vi sia una relazione lineare tra la variabile dipendente e ciascuna delle variabili indipendenti, scrivete l'espressione del modello di regressione multipla.
- Fornite un'interpretazione delle inclinazioni della variabile dipendente rispetto a ciascuna delle variabili esplicative.
- Calcolate il coefficiente di determinazione $r^2_{Y,12}$ e interpretatene il significato.
- Calcolate l' r^2 corretto.

- 13.3 Una società di vendita per corrispondenza di computer, software e accessori per computer ha un deposito unico da cui vengono prelevati e distribuiti i prodotti ordinati. Il management intende esaminare il processo di distribuzione dei prodotti dal deposito per stabilire quali siano i fattori che ne determinano i costi. Infatti, attualmente viene applicata una tariffa di trasporto dall'importo limitato su tutti gli ordini, indipendentemente dal loro ammontare. Nella tabella seguente si riportano i dati raccolti negli ultimi 24 mesi in relazione ai costi di distribuzione, alle vendite e al numero di ordini ricevuti.

MESE	COSTI DI	VENDITE	ORDINI	MESE	COSTI DI	VENDITE	ORDINI
	DISTRIBUZIONE				DISTRIBUZIONE		
	(\$ 000)	(\$ 000)			(\$ 000)	(\$ 000)	
1	52.95	386	4015	13	62.98	372	3977
2	71.66	446	3806	14	72.30	328	4428
3	85.58	512	5309	15	58.99	408	3964
4	63.69	401	4262	16	79.38	491	4582
5	72.81	457	4296	17	94.44	527	5582
6	68.44	458	4097	18	59.74	444	3450
7	52.46	301	3213	19	90.50	623	5079
8	70.77	484	4809	20	93.24	596	5735
9	82.03	517	5237	21	69.33	463	4269
10	74.39	503	4732	22	53.71	389	3708
11	70.84	535	4413	23	89.18	547	5387
12	54.08	353	2921	24	66.80	415	4161



DATASET
WARECOST



DATASET
ADRADTV

Sulla base dei dati raccolti:

- Scrivete l'espressione del modello di regressione multipla.
- Fornite un'interpretazione delle inclinazioni della variabile dipendente rispetto a ciascuna delle variabili esplicative.
- Fornite una previsione dei costi di distribuzione per un ammontare delle vendite pari a \$ 400 000 e degli ordini pari a 4500.
- Calcolate il coefficiente di determinazione $r^2_{Y,12}$ e interpretatene il significato.
- Calcolate l' r^2 corretto.

- 13.4 Supponete che un'azienda produttrice di beni di largo consumo intenda valutare l'efficacia di diversi tipi di pubblicità nella promozione dei suoi prodotti. A tale scopo si prendono in considerazione due tipi di pubblicità: la pubblicità per radio e televisione e la pubblicità sui giornali. Un campione di 22 città con approssimativamente la medesima popolazione viene sottoposto a un test per un mese: in ciascuna città viene allocato un dato livello di spesa per la pubblicità mediante radio e televisione e per quella sui giornali e si raccolgono i dati relativi alle vendite dei prodotti. Nella seguente tabella si riportano i dati raccolti per un mese in relazione all'ammontare della spesa per la pubblicità mediante radio e televisione, di quella su giornali e alle vendite dei prodotti.

CITTÀ	PUBBLICITÀ PER RADIO E TELEVISIONE			CITTÀ	PUBBLICITÀ PER RADIO E TELEVISIONE		
	VENDITE (\$ 000)	PUBBLICITÀ (\$ 000)	SU GIORNALI (\$ 000)		VENDITE (\$ 000)	PUBBLICITÀ (\$ 000)	SU GIORNALI (\$ 000)
1	973	0	40	12	1577	45	45
2	1119	0	40	13	1044	50	0
3	875	25	25	14	914	50	0
4	625	25	25	15	1329	55	25
5	910	30	30	16	1330	55	25
6	971	30	30	17	1405	60	30
7	931	35	35	18	1436	60	30
8	1177	35	35	19	1521	65	35
9	882	40	25	20	1741	65	35
10	982	40	25	21	1866	70	40
11	1628	45	45	22	1717	70	40

Sulla base dei dati raccolti:

- Scrivete l'espressione del modello di regressione multipla.
- Fornite un'interpretazione delle inclinazioni con riferimento al problema in considerazione.
- Fornite una previsione delle vendite per una città in cui l'ammontare della spesa in pubblicità per radio e televisione è pari a \$ 20 000 e quello della spesa in pubblicità su giornali è pari a \$ 20 000.
- Calcolate il coefficiente di determinazione $r^2_{Y,12}$ e interpretatene il significato.
- Calcolate l' r^2 corretto.

13.2

L'ANALISI DEI RESIDUI NEL MODELLO DI REGRESSIONE MULTIPLA

Nel Paragrafo 12.5 abbiamo introdotto l'analisi dei residui come utile strumento per valutare se il modello di regressione impiegato è adeguato per l'analisi dell'insieme dei dati considerato. Nel riquadro 13.1 si riporta un elenco di grafici dei residui utili per la valutazione di un modello di regressione lineare con due variabili esplicative.

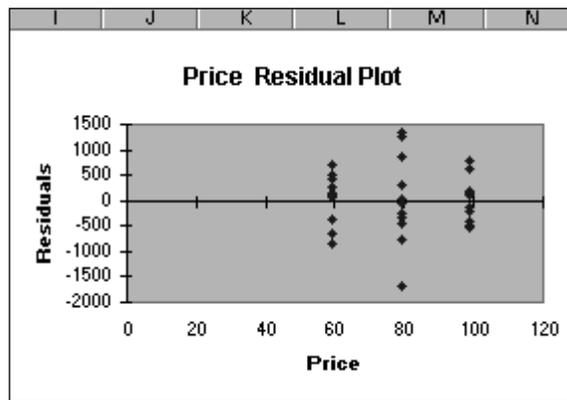


Riquadro 13.1 I grafici dei residui impiegati nella regressione multipla

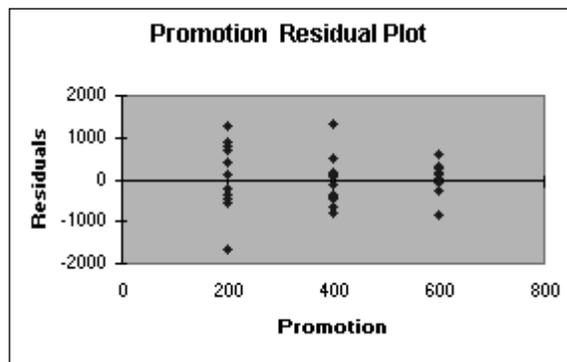
- ✓ 1. Il grafico dei residui rispetto a \hat{Y}_i
- ✓ 2. Il grafico dei residui rispetto a X_{1i}
- ✓ 3. Il grafico dei residui rispetto a X_{2i}
- ✓ 4. Il grafico dei residui rispetto al tempo

Il primo grafico dei residui consente di stabilire se i residui presentano un andamento rispetto ai valori previsti della Y con una struttura riconoscibile. Un andamento di questo genere fornisce la prova della presenza di un legame non lineare tra almeno una delle variabili esplicative e la variabile dipendente e/o della necessità di una trasformazione della variabile dipendente Y . Il secondo e il terzo grafico dei residui riguardano, invece, le variabili esplicative. Il riconoscimento di un andamento strutturato dei residui rispetto a una delle variabili indipendenti può rivelare l'esistenza di un legame non lineare tra tale variabile esplicativa e la Y oppure la necessità di una trasformazione della variabile stessa. Il quarto grafico viene impiegato per stabilire la presenza di una struttura nei residui quando i dati vengono raccolti nel corso del tempo. In tal caso, come abbiamo visto nel Paragrafo 9.6, si può procedere anche al calcolo della statistica di Durbin-Watson per accertare la presenza di una autocorrelazione positiva tra i residui.

I grafici dei residui sono calcolati da tutti i programmi di analisi statistica. Nella Figura 13.2 si riportano i grafici dei residui ottenuti con Excel per i dati relativi alle vendite della barretta energetica Omnipower. I grafici non rivelano la presenza di una struttura dei resi-



RIQUADRO A



RIQUADRO B

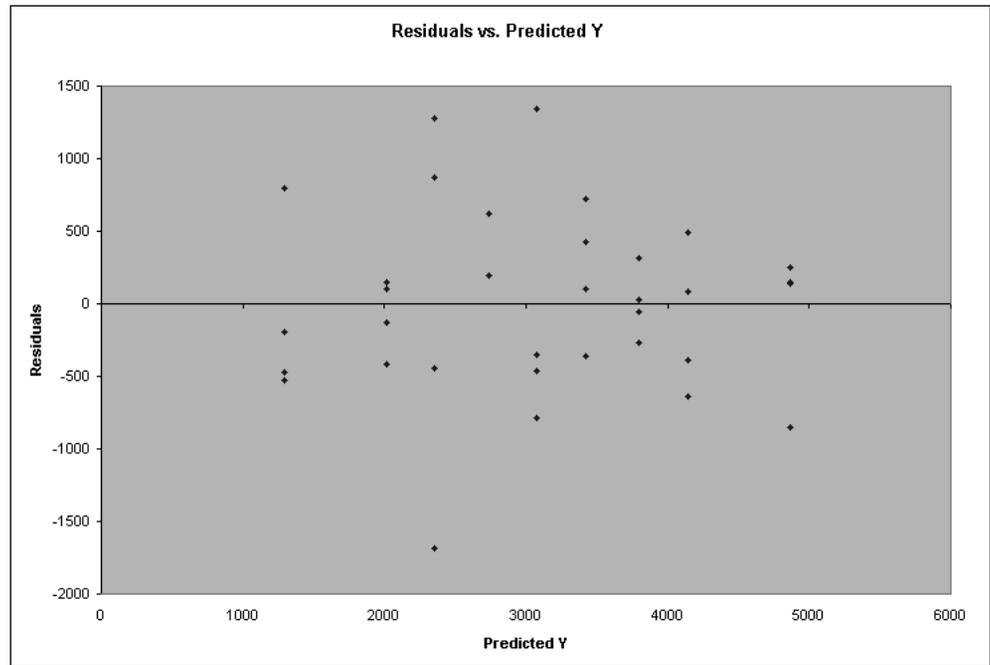


FIGURA 13.2 Grafici dei residui ottenuti con Excel per il modello di regressione relativo alle vendite della barretta Omnipower: riquadro A residui rispetto al prezzo, riquadro B residui rispetto alle spese promozionali, riquadro C residui rispetto ai valori previsti della Y.

dui rispetto ai valori previsti della Y, ai valori della X_1 (il prezzo) e a quelli della X_2 (la spesa per le promozioni). Possiamo allora concludere che il modello di regressione multipla è adeguato per la previsione delle vendite della Omnipower.

Esercizi del paragrafo 13.2



DATASET
WARECOST

13.5 Nell'Esercizio 13.3 sono state impiegate per prevedere la variabile costi di distribuzione le variabili "Vendite" e "Numero di ordini". Conducete un'analisi dei residui sulla base dei risultati ottenuti e valutate l'adeguatezza del modello impiegato.

Rappresentate graficamente i residui rispetto ai mesi. Il grafico evidenzia la presenza di una struttura nei residui? Commentate.



DATASET
ADRADTV

• 13.6 Nell'Esercizio 13.4 sono state impiegate le variabili "Spesa per la pubblicità per televisione e radio" e "Spesa per la pubblicità sui giornali" per prevedere la variabile vendite dei prodotti. Conducete un'analisi dei residui per valutare la bontà del modello impiegato.

13.3

IL TEST PER LA VERIFICA DELLA SIGNIFICATIVITÀ DEL MODELLO DI REGRESSIONE LINEARE MULTIPLA

Una volta valutata, sulla base dell'analisi dei residui, l'adeguatezza del modello di regressione lineare multipla, passiamo a verificare se ci sia una relazione significativa tra la variabile dipendente e l'insieme delle variabili esplicative. Dal momento che siamo in presenza di più di una variabile esplicativa, l'ipotesi nulla e quella alternativa vanno specificate nella maniera seguente:

$$H_0: \beta_1 = \beta_2 = 0 \text{ (Non vi è una relazione lineare tra la variabile dipendente e le variabili esplicative.)}$$

H_1 : Almeno un $\beta_j \neq 0$ (Vi è una relazione lineare tra la variabile dipendente e almeno una delle variabili esplicative.)

Come nel caso del modello di regressione lineare semplice (cfr. Paragrafo 9.7), tale problema di verifica di ipotesi viene risolto ricorrendo al test F , riassunto nella Tabella 13.2.

Il test F sull'intero modello nel modello di regressione multipla

La statistica F è data dal rapporto tra la media dei quadrati della regressione (MQR) e la media dei quadrati dell'errore (MQE)

$$F = \frac{MQR}{MQE} \quad (13.6)$$

dove

p = numero delle variabili esplicative nel modello di regressione

F = la statistica test F avente una distribuzione F con p e $n - p - 1$ gradi di libertà

La regola decisionale in questo caso è:

Rifiutare H_0 se $F > F_U$,

dove F_U è il valore critico sulla coda di destra di una distribuzione F con p e $n - p - 1$ gradi di libertà;

altrimenti accettare H_0 .

Tabella 13.2 *Tabella ANOVA per il test per la verifica della significatività dell'insieme dei coefficienti di regressione nel modello di regressione multipla con $p = 2$ variabili esplicative*

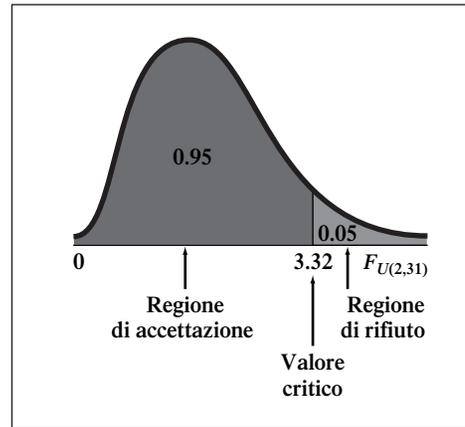
Fonte	GDL	Somma dei quadrati	Media dei quadrati (varianza)	F
Regressione	p	SQR	$MQR = \frac{SQR}{p}$	$F = \frac{MQR}{MQE}$
Residuo	$n - p - 1$	SQE	$MQE = \frac{SQE}{n - p - 1}$	
Totale		$n - 1$	STQ	

La Figura 13.1 riporta tutti i calcoli necessari per la costruzione del test F per l'esempio relativo alle vendite della Omnipower.

Se il livello di significatività scelto è 0.05, dalla Tabella E.5 ricaviamo che il valore critico (per una distribuzione F con 2 e 31 gradi di libertà) è approssimativamente uguale a 3.32, come illustrato nella Figura 13.3. Il valore di F può essere calcolato in base all'equazione (13.6) e ai valori riportati nella Figura 13.1. Poiché $F = 48.48 > F_U = 3.32$ (cfr. Figura 13.3) o ancora poiché il p -value = 0.000 < 0.05, possiamo rifiutare H_0 e quindi concludere che vi è una relazione lineare tra almeno una variabile esplicative (il prezzo e/o le spese di promozione) e le vendite.

FIGURA 13.3

Verifica della significatività dell'insieme dei coefficienti di regressione con un livello di significatività pari a 0.05 e 2 e 31 gradi di libertà.



Esercizi del paragrafo 13.3

13.7 La seguente tabella di analisi della varianza si riferisce a un modello di regressione lineare multipla con due variabili indipendenti.

FORTE	GDL	SOMMA DEI QUADRATI	MEDIA DEI QUADRATI (VARIANZA)	F
Regressione	2	60		
Residuo	18	120		
Totale	20	180		

- Calcolate la media dei quadrati della regressione e la media dei quadrati dell'errore.
- Calcolate la statistica F .
- Verificate se vi sia una relazione significativa tra Y e le due variabili esplicative, per un livello di significatività pari a 0.05.

- 13.8 Riprendete l'Esercizio 13.2. L'output di Excel riportato contiene la seguente tabella di analisi della varianza:

FORTE	GDL	SOMMA DEI QUADRATI	MEDIA DEI QUADRATI (VARIANZA)	F	F
Regressione	2	12.61020	6.30510	97.69	0.0001
Residuo	12	0.77453	0.06454		
Totale	14	13.38473			

- Per un livello di significatività pari a 0.05 stabilite se vi è una relazione lineare significativa tra l'impatto di lungo periodo e le due variabili esplicative (FOREIMP e MIDSOLE)
- Fornite un'interpretazione del p -value con riferimento al caso preso in considerazione.

13.9 Tornate all'Esercizio 13.3 e in base all'output di Excel ottenuto:

- Per un livello di significatività pari a 0.05 stabilite se vi è una relazione lineare significativa tra i costi di distribuzione e le due variabili esplicative (vendite e numero di ordini).
- Fornite un'interpretazione del p -value con riferimento al caso preso in considerazione.

13.10 Tornate all'Esercizio 12.4 e in base all'output di Excel ottenuto:

- Per un livello di significatività pari a 0.05, stabilite se vi è una relazione lineare significativa tra le vendite e le due variabili esplicative (pubblicità alla radio e televisione e pubblicità su giornali).
- Fornite un'interpretazione del p -value con riferimento al caso preso in considerazione



DATASET
WARECOST



DATASET
ADRADTV

INFERENZA SUI COEFFICIENTI DI REGRESSIONE DELLA POPOLAZIONE

Nel Paragrafo 12.7 abbiamo introdotto un test di ipotesi sull'inclinazione della retta di regressione per la verifica della significatività della relazione tra X e Y e abbiamo costruito un intervallo di confidenza per la stima dell'inclinazione. In questo paragrafo intendiamo estendere tali procedure al modello di regressione multipla.

Test di ipotesi

La statistica test per la verifica dell'ipotesi $\beta_1 = 0$ nel modello di regressione semplice è per l'equazione (12.16):

$$t = \frac{b_1}{S_{b_1}}$$

Generalizzando al caso del modello di regressione multipla, otteniamo la seguente espres-

Il test t per la verifica di ipotesi sull'inclinazione nel modello di regressione multipla

$$t = \frac{b_k}{S_{b_k}} \quad (13.7)$$

dove

p = numero di variabili esplicative

b_k = inclinazione di Y rispetto alla variabile k tenendo costanti le altre variabili

S_{b_k} = errore standard del coefficiente di regressione b_k

t = statistica test con distribuzione t con $n - p - 1$ gradi di libertà.

sione:

I risultati del test t per ciascuna delle variabili esplicative sono riportati nell'output di Excel della Figura 13.1.

Pertanto, se vogliamo stabilire se la variabile X_2 (ammontare delle spese promozionali) ha un effetto significativo sulle vendite, tenendo conto del prezzo della barretta OmniPower, l'ipotesi nulla e quella alternativa sono:

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

In base all'equazione (13.7) abbiamo:

$$t = \frac{b_2}{S_{b_2}}$$

e con riferimento ai dati relativi all'esempio considerato:

$$b_2 = 3.6131 \quad \text{e} \quad S_{b_2} = 0.6852$$

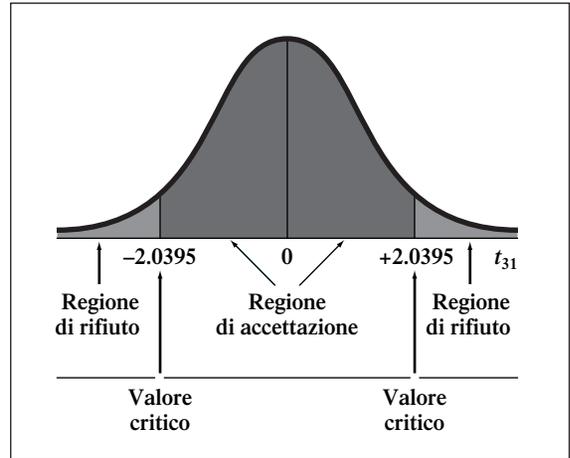
Pertanto:

$$t = \frac{3.6131}{0.6852} = 5.27$$

Per un livello di significatività pari a 0.05, dalla Tabella E.3 ricaviamo che i valori della statistica t per 31 gradi di libertà sono -2.0395 e $+2.0395$ (cfr. Figura 13.4). Dalla Figura 13.1 osserviamo inoltre che il p -value è pari a 0.00000982 (9.2E-06 in notazione scientifica).

FIGURA 13.4

Verifica della significatività del coefficiente di regressione con un livello di significatività pari a 0.05 e 31 gradi di libertà.



Poiché $t = 5.27 > t_{31} = 2.0395$ o ancora poiché $p\text{-value} = 0.00000982 > 0.05$, rifiutiamo H_0 e possiamo concludere che vi è una relazione significativa tra la variabile X_2 (spese promozionali) e le vendite, tenendo conto del prezzo X_1 .

Costruiamo il test sulla significatività di b_1 , l'inclinazione delle vendite rispetto al prezzo.

Esempio 13.1 *Test sulla significatività dell'inclinazione delle vendite rispetto al prezzo*

Si può ritenere, per un livello di significatività pari a 0.05, che l'inclinazione delle vendite rispetto al prezzo sia diversa da zero?

SOLUZIONE

In base alla Figura 13.1, $t = -7.766 < -2.0395$ e $p\text{-value} = 0.000000092 < 0.05$. Pertanto, vi è una relazione lineare significativa tra il prezzo (X_1) e le vendite, dato l'ammontare della spesa promozionale (X_2).

Verificare la significatività di un particolare coefficiente di regressione equivale a verificare la significatività dell'inserimento della variabile corrispondente nel modello di regressione, date le variabile già presenti. Pertanto, il test t su un coefficiente di regressione equivale al test sul contributo della variabile esplicativa corrispondente.

La stima per intervalli di confidenza

Si può essere interessati a stimare uno dei coefficienti di regressione, anziché a valutarne la significatività. Nel caso del modello di regressione multipla, l'intervallo di confidenza per il generico coefficiente di regressione β_k assume la seguente espressione

Stima per intervallo di confidenza per l'inclinazione

$$b_k \pm t_{n-p-1} S_{b_k} \quad (13.8)$$

Per esempio, l'intervallo di confidenza per il coefficiente β_1 in base all'equazione (13.8) e alla Figura 13.1 è dato dalla seguente espressione:

$$b_1 \pm t_{31} S_{b_1}$$

Poiché il valore critico di t per un livello di significatività pari a 0.95 e 31 gradi di libertà è uguale a 2.0395 (cfr. Tavola E.3), si ha:

$$-53.2173 \pm (2.0395)(6.8522)$$

$$-53.2173 \pm 13.9752$$

$$-67.1925 \leq \beta_1 \leq -39.2421$$

Pertanto riteniamo che, dato l'effetto della spesa promozionale, l'aumento di un centesimo del prezzo della barretta Omnipower determini una riduzione del numero delle barrette vendute compresa tra 67.2 e 39.2. Riteniamo che, per un livello di confidenza pari a 95%, questo intervallo stimi correttamente la vera relazione esistente tra le variabili considerate.

D'altro canto, poiché l'intervallo trovato non comprende lo zero, possiamo concludere che X_1 abbia un effetto significativo sulla variabile dipendente.

Esempio 13.2 *L'intervallo di confidenza per l'inclinazione delle vendite rispetto alle spese promozionali*

Costruite un intervallo di confidenza del 95% per l'inclinazione delle vendite rispetto alle spese promozionali.

SOLUZIONE

Poiché il valore critico di t per un livello di confidenza del 95% e 31 gradi di libertà è uguale a 2.0395 (cfr. Tabella E.3), abbiamo:

$$3.6131 \pm (2.0395)(0.6852)$$

$$3.6131 \pm 1.3975$$

$$2.2156 \leq \beta_2 \leq 5.0106$$

Pertanto riteniamo che, dato l'effetto del prezzo, per ciascun dollaro di aumento delle spese promozionali il numero delle barrette vendute aumenti di un ammontare compreso tra 2.2 e 5 barrette. Riteniamo che, per un livello di confidenza pari a 95%, questo intervallo stimi correttamente la vera relazione esistente tra le variabili considerate. Poiché l'intervallo trovato non comprende lo zero, possiamo concludere che X_2 abbia un effetto significativo sulla variabile dipendente.

Esercizi del paragrafo 13.4

- **13.11** Supponete che vi vengano fornite le seguenti informazioni in relazione a un modello di regressione lineare multipla:

$$n = 25, b_1 = 5, b_2 = 10, S_{b_1} = 2, S_{b_2} = 8$$

- (a) Quale variabile presenta un'inclinazione maggiore?
 - (b) Costruite un intervallo di confidenza del 95% per l'inclinazione di X_1 .
 - (c) Per un livello di significatività pari a 0.05, stabilite se ciascuna delle variabili esplicative contribuisce in maniera significativa al modello di regressione. Alla luce dei risultati ottenuti, indicate le variabili esplicative da includere nel modello.
- **13.12** Con riferimento all'Esercizio 13.2, prendete in considerazione il seguente output:

VARIABILE	COEFFICIENTI	ERRORE STANDARD	STAT t	VALORE DI SIGNIFICATIVITÀ
Intercetta	-0.02686	0.06905	-0.39	
Foreimp	0.79116	0.06295	12.57	0.0000
Midsole	0.60484	0.07174	8.43	0.0000

- (a) Costruite un intervallo di confidenza del 95% per l'inclinazione della variabile dipendente (impatto di lungo periodo) rispetto alla variabile FOREIMP.
- (b) Per un livello di significatività pari a 0.05, stabilite se le variabili esplicative contribuiscono in maniera significativa al modello. Sulla base dei risultati ottenuti, indicate quali variabili esplicative dovrebbero essere incluse nel modello.



DATASET
WARECOST

13.13 Tornate all'Esercizio 13.3 e in base all'output di Excel ottenuto:

- (a) Costruite un intervallo di confidenza del 95% per l'inclinazione dei costi rispetto alle vendite.
- (b) Per un livello di significatività pari a 0.05, stabilite se le variabili esplicative contribuiscono in maniera significativa al modello. Sulla base dei risultati ottenuti, indicate quali variabili esplicative dovrebbero essere incluse nel modello.



DATASET
ADRADTV

13.14 Tornate all'Esercizio 13.4 e in base all'output di Excel ottenuto:

- (a) Costruite un intervallo di confidenza del 95% per l'inclinazione delle vendite rispetto alla pubblicità per radio e televisione.
- (b) Per un livello di significatività pari a 0.05, stabilite se le variabili esplicative contribuiscono in maniera significativa al modello. Sulla base dei risultati ottenuti, indicate quali variabili esplicative dovrebbero essere incluse nel modello.

13.5

LA VERIFICA DI IPOTESI SULLE PROPORZIONI NEL MODELLO DI REGRESSIONE MULTIPLA

Nella costruzione di un modello di regressione multipla, intendiamo inserire solo le variabili esplicative che possono essere utili per la previsione della variabile dipendente. Una variabile che non risulta utile in tal senso, dovrebbe essere eliminata dal modello e dovrebbe essere utilizzato un modello con un numero minore di variabili.

Un metodo alternativo per la valutazione del contributo di ciascuna variabile esplicativa è il cosiddetto **criterio del test F parziale**. Tale metodo comporta il calcolo del contributo che ciascuna variabile esplicativa dà alla somma dei quadrati dopo che tutte le altre variabili esplicative siano state incluse nel modello. La nuova variabile è inclusa nel modello solo se il modello ne risulta significativamente migliorato.

Per applicare il criterio del test F parziale ai dati relativi alle vendite della barretta Omnipower, dobbiamo calcolare il contributo della variabile spese promozionali (X_2) dopo che la variabile prezzo (X_1) sia stata inclusa nel modello e vice versa il contributo della variabile prezzo (X_1) dopo che la variabile spese promozionali (X_2) sia stata inclusa nel modello.

In generale, in presenza di più di una variabile esplicativa, il contributo di ciascuna di essa si può valutare sulla base della somma dei quadrati della regressione per un modello che comprende tutte le variabili esplicative eccetto quella presa in considerazione, SQR (tutte le variabili tranne la k -esima). Pertanto il contributo della variabile k supponendo che tutte le altre variabili siano incluse nel modello può essere valutato sulla base della seguente quantità:

Valutazione del contributo di una variabile indipendente al modello di regressione

$$\begin{aligned}
 SQR(X_k \mid \text{tutte le variabili tranne la } k\text{-esima}) \\
 &= SQR(\text{tutte le variabili inclusa la } k\text{-esima}) \\
 &\quad - SQR(\text{tutte le variabili esclusa la } k\text{-esima})
 \end{aligned}
 \tag{13.9}$$

In presenza di due variabili esplicative (come nell'esempio relativo alla Omnipower) il contributo di ciascuna delle variabili esplicative può essere determinato in base alle equazioni (13.10a) e (13.10b).

Valutazione del contributo di X_1 e di X_2 al modello di regressione

Il contributo della variabile X_1 quando X_2 è inclusa nel modello

$$SQR(X_1 | X_2) = SQR(X_1 \text{ e } X_2) - SQR(X_2) \quad (13.10a)$$

Il contributo della variabile X_2 quando X_1 è inclusa nel modello

$$SQR(X_2 | X_1) = SQR(X_1 \text{ e } X_2) - SQR(X_1) \quad (13.10b)$$

I termini $SQR(X_2)$ e $SQR(X_1)$ rappresentano rispettivamente la somma dei quadrati della regressione per il modello che ha come sola variabile esplicativa X_2 (le spese promozionali) e la somma dei quadrati della regressione per il modello che ha come sola variabile esplicativa X_1 (il prezzo). Nelle Figure 13.6 e 13.7 sono riportati gli output ottenuti con Excel relativi a questi due modelli.

FIGURA 13.5

Output parziale ottenuto con Excel per il modello di regressione delle vendite rispetto alle spese promozionali.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.5350951					
5	R Square	0.28632676					
6	Adjusted R Square	0.26402448					
7	Standard Error	1077.87208					
8	Observations	34					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	14915814.1	14915814	12.83845	0.001111494	
13	Residual	32	37177863.34	1161808			
14	Total	33	52093677.44				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	1496.01613	483.978853	3.091077	0.004111	510.1843006	2481.84796
18	Promotion	4.12806452	1.152100113	3.583078	0.001111	1.781315368	6.47481366

FIGURA 13.6

Output parziale ottenuto con Excel per il modello di regressione delle vendite rispetto al prezzo.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.735145971					
5	R Square	0.540439599					
6	Adjusted R Square	0.526078336					
7	Standard Error	864.9456503					
8	Observations	34					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	28153486.15	28153486	37.63176	7.35855E-07	
13	Residual	32	23940191.29	748131			
14	Total	33	52093677.44				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	7512.347984	734.6188701	10.22618	1.31E-11	6015.97958	9008.71639
18	Price	-56.71384409	9.245104279	-6.13447	7.36E-07	-75.54548931	-37.882195

Dalla Figura 13.5 osserviamo che:

$$SQR(X_2) = 14\,915\,814.10$$

e pertanto, in base all'equazione (13.10a),

$$SQR(X_1 | X_2) = SQR(X_1 \text{ e } X_2) - SQR(X_2)$$

Otteniamo

$$SQR(X_1 | X_2) = 39\,472\,730.77 - 14\,915\,814.1 = 24\,556\,916.67$$

Per stabilire se X_1 migliora in maniera significativa il modello, in cui X_2 sia già stata inclusa, possiamo suddividere la somma dei quadrati della regressione in due componenti come illustrato nella Tabella 13.3.

L'ipotesi nulla e l'ipotesi alternativa per verificare la significatività del contributo di X_1 al modello sono rispettivamente:

H_0 : La variabile X_1 non migliora in maniera significativa il modello in cui la variabile X_2 sia stata inclusa.

H_1 : La variabile X_1 migliora in maniera significativa il modello in cui la variabile X_2 sia stata inclusa.

Il test F parziale è allora dato dalla seguente espressione:

Il test F parziale per la valutazione del contributo di una variabile indipendente

$$F = \frac{SQR(X_k | \text{tutte le variabili eccetto } k)}{MQE} \quad (13.11)$$

Nell'equazione (13.11) con F si indica la statistica F che ha una distribuzione F con 1 e $n - p - 1$ gradi di libertà.

Pertanto in base alla Tabella 13.3 abbiamo

$$F = \frac{24,556,916.67}{407,127.31} = 60.32$$

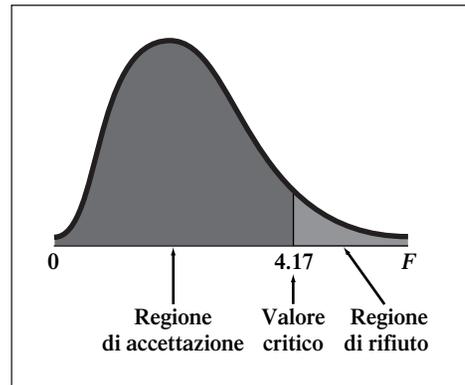
Dal momento che vi sono 1 e 31 gradi di libertà, se il livello di significatività scelto è 0.05, in base alla Tabella E.5 il valore critico è uguale approssimativamente a 4.17 (cfr. Figura 13.7)

Tabella 13.3 *Tavola della ANOVA in cui la somma dei quadrati della regressione viene divisa in componenti per valutare il contributo della variabile X_1*

Fonte	GDL	Somma dei quadrati	Media dei quadrati (varianza)	F
Regressione	2	39 472 730.77	19 736 365.39	
$\left\{ \begin{array}{l} X_2 \\ X_1 X_2 \end{array} \right\}$	$\left\{ \begin{array}{l} 1 \\ 1 \end{array} \right\}$	$\left\{ \begin{array}{l} 14\,915\,814.10 \\ 24\,556\,916.67 \end{array} \right\}$	24 556 916.67	60.32
Residuo	31	12 620 946.67	407 127.31	
Totale	33	52 093 677.44		

FIGURA 13.7

Verifica della significatività del contributo al modello del coefficiente di regressione con un livello di significatività pari a 0.05 e 1 e 31 gradi di libertà.



Poiché il valore osservato della statistica F è maggiore del valore critico ($60.32 > 4.17$) decidiamo di rifiutare H_0 e concludiamo che l'inserimento della variabile X_1 (il prezzo) migliora in maniera significativa un modello di regressione che già contenga la variabile X_2 (spese promozionali).

Per valutare il contributo della variabile X_2 (spese promozionali) al modello in cui X_1 è già stata inclusa, ricorriamo all'equazione (13.10b):

$$SQR(X_2 | X_1) = SQR(X_1 \text{ e } X_2) - SQR(X_1)$$

In base alla Figura 13.6 abbiamo

$$SQR(X_1) = 28,153,486.15$$

Pertanto in base alle Figure 13.1, 13.2 e 13.7

$$SQR(X_2 | X_1) = 39\,472\,730.77 - 28\,153\,486.15 = 11\,319\,244.62$$

Per stabilire se X_2 migliora in maniera significativa il modello che include già X_1 , possiamo suddividere la somma dei quadrati della regressione in due componenti come illustrato nella Tabella 13.4.

L'ipotesi nulla e l'ipotesi alternativa per verificare la significatività del contributo di X_2 al modello sono rispettivamente:

H_0 : La variabile X_2 non migliora in maniera significativa il modello in cui la variabile X_1 sia già stata inclusa.

H_1 : La variabile X_2 migliora in maniera significativa il modello in cui la variabile X_1 sia già stata inclusa.

In base all'equazione (13.11), otteniamo:

$$F = \frac{11\,319\,244.62}{407\,127.31} = 27.80$$

Come illustrato dalla Tabella 13.4

Dal momento che vi sono 1 e 31 gradi di libertà, per un livello di significatività pari a 0.05, in base alla Tabella E.5 il valore critico è uguale approssimativamente a 4.17. Poiché il valore osservato della statistica F è maggiore del valore critico ($27.80 > 4.17$) decidiamo di rifiutare H_0 e concludiamo che l'inserimento nel modello della variabile X_2 (spese promozionali) migliora in maniera significativa un modello di regressione che già contenga la variabile X_1 (il prezzo).

In base ai risultati ottenuti sottoponendo a verifica la significatività del contributo di ciascuna variabile possiamo concludere che entrambe le variabili migliorano significativamente il modello e pertanto entrambe vi dovrebbero essere incluse.

Tabella 13.4 *Tavola della ANOVA in cui la somma dei quadrati della regressione viene divisa in componenti per valutare il contributo della variabile X_2*

Fonte	gdl	Somma dei quadrati	Media dei quadrati (varianza)	F
Regressione	2	39 472 730.77	19 736 365.39	
$\left\{ \begin{array}{l} X_1 \\ X_2 X_1 \end{array} \right\}$	$\left\{ \begin{array}{l} 1 \\ 1 \end{array} \right\}$	$\left\{ \begin{array}{l} 28 153 486.15 \\ 11 319 244.62 \end{array} \right\}$	11 319 244.62	27.80
Residuo	31		407 127.31	
Totale	33	52 093 677.44		

È possibile individuare una relazione tra i valori della statistica test t , ottenuti in base all'equazione (13.7) per valutare la significatività dei coefficienti corrispondenti alle variabili esplicative X_1 e X_2 , e quelli della statistica test F ottenuti in base all'equazione (13.11) per valutare il contributo al modello di regressione delle variabili esplicative X_1 e X_2 . I valori trovati della statistica t sono -7.77 e $+5.27$ e i corrispondenti valori della F sono 60.32 e 27.80 . Si può quindi individuare la seguente relazione¹ tra t e F :

¹La relazione tra t e F individuata dall'equazione (13.12) vale solo se la statistica t viene usata in un test a due code.

La relazione tra la statistica t e la statistica F

$$t_v^2 = F_{1,v} \quad (13.12)$$

Dove v = numero di gradi di libertà

Il coefficiente di determinazione parziale

Nel paragrafo 13.1 abbiamo introdotto il coefficiente di determinazione ($r_{Y,12}^2$), che misura la proporzione della variabilità di Y spiegata dalla variabilità delle variabili esplicative. Dopo aver discusso i metodi che consentono di valutare il contributo di ciascuna delle variabili esplicative al modello di regressione lineare, possiamo introdurre anche i **coefficienti di determinazione parziali** ($r_{Y1,2}^2$ e $r_{Y2,1}^2$). Tali coefficienti misurano la proporzione della variabilità della variabile dipendente che è spiegata da ciascuna variabile esplicativa quando si mantengono costanti le altre variabili esplicative. Pertanto in un modello di regressione con due variabili esplicative, avremo:

I coefficienti di determinazione parziale per un modello con due variabili esplicative

$$r_{Y1,2}^2 = \frac{SQR(X_1 | X_2)}{SQT - SQR(X_1 \text{ e } X_2) + SQR(X_1 | X_2)} \quad (13.13a)$$

e

$$r_{Y2,1}^2 = \frac{SQR(X_2 | X_1)}{SQT - SQR(X_1 \text{ e } X_2) + SQR(X_2 | X_1)} \quad (13.13b)$$

dove

$SQR(X_1 | X_2)$ = parte della somma dei quadrati della regressione dovuta al contributo di X_1 quando X_2 è già inclusa nel modello

(continua)

I coefficienti di determinazione parziale per un modello con due variabili esplicative (seguito)

SQT = somma totale dei quadrati

$SQR(X_1 \text{ e } X_2)$ = somma dei quadrati della regressione quando sia X_1 che X_2 sono incluse nel modello

$SQR(X_2 | X_1)$ = parte della somma dei quadrati della regressione dovuta al contributo di X_2 quando X_1 è già inclusa nel modello

In un modello contenente p variabili esplicative, per la k -esima variabile avremo:

I coefficienti di determinazione parziali per un modello di regressione contenente p variabili indipendenti

$$r_{Yk}^2 \text{ (tutte le variabili tranne la } k\text{-esima)} \tag{13.14}$$

$$= \frac{SQR(X_k | \text{ tutte le variabili tranne la } k\text{-esima})}{SQT - SQR(\text{tutte le variabili compresa la } k\text{-esima}) + SQR(X_k | \text{ tutte le variabili tranne la } k\text{-esima})}$$

Per i dati relativi alle vendite della barretta Omnipower, otteniamo:

$$r_{Y1.2}^2 = \frac{24\,556\,916.67}{52\,093\,677.44 - 39\,472\,730.77 + 24\,556\,916.67} = .6605$$

e

$$r_{Y2.1}^2 = \frac{11\,319\,244.62}{52\,093\,677.24 - 39\,472\,730.77 + 11\,319\,244.62} = 0.4728$$

Il coefficiente di determinazione parziale corrispondente alla variabile X_1 , quando X_2 è tenuta costante, ci dice che per un dato ammontare delle spese promozionali, il 66.05% della variabilità delle vendite della Omnipower può essere spiegata dalla variabilità del prezzo nei negozi. Il coefficiente di determinazione parziale corrispondente alla variabile X_2 , quando X_1 è tenuta costante, ci dice che per un dato ammontare delle spese promozionali, il 47.28% della variabilità delle vendite della Omnipower può essere spiegata dalla variabilità del prezzo nei negozi.

Esercizi del paragrafo 13.5

- **13.15** Supponete che la seguente tabella di analisi della varianza sia stata ottenuta da un modello di regressione multipla con due variabili esplicative.

Fonte	GDL	Somma dei quadrati	Media dei quadrati (varianza)	F
Regressione	2	60		
Residuo	18	120		
Totale	20	180		



DATASET
WARECOST



DATASET
ADRADTV

$$SQR(X_1)=45 \quad SQR(X_2)=25$$

- (a) Per un livello di significatività pari a 0.05, stabilite se vi sia una relazione significativa tra Y e ciascuna delle variabili esplicative.
- (b) Calcolate i coefficienti di determinazione parziali $r_{Y1.2}^2$ e $r_{Y2.1}^2$, e interpretatene il significato.
- 13.16** Tornate all'Esercizio 13.3 e in base all'output di Excel ottenuto:
- (a) Per un livello di significatività pari a 0.05 stabilite se ciascuna delle variabili esplicative apporta un contributo significativo al modello di regressione considerato. Alla luce dei risultati ottenuti indicate quale modello di regressione dovrebbe essere impiegato.
- (b) Calcolate i coefficienti di determinazione parziali $r_{Y1.2}^2$ e $r_{Y2.1}^2$ e interpretatene il significato.
- **13.17** Tornate all'Esercizio 13.4 e in base all'output di Excel ottenuto:
- (a) Per un livello di significatività pari a 0.05 stabilite se ciascuna delle variabili esplicative apporta un contributo significativo al modello di regressione considerato. Alla luce dei risultati ottenuti indicate quale modello di regressione dovrebbe essere impiegato.
- (b) Calcolate i coefficienti di determinazione parziali $r_{Y1.2}^2$ e $r_{Y2.1}^2$ e interpretatene il significato.

13.6

IL MODELLO DI REGRESSIONE QUADRATICA

Sino a ora, come nel capitolo precedente, abbiamo supposto che tra la variabile dipendente Y e ciascuna delle variabile esplicative vi fosse una relazione di tipo lineare. Tuttavia, nel Paragrafo 12.1, abbiamo discusso dei diversi tipi di relazione che possono sussistere tra variabili quantitative. Tra le relazioni non lineari più comuni, vi è quella quadratica (cfr. Figura 12.2, riquadri C-E) in base alla quale Y aumenta o si riduce a un tasso diverso al variare dei valori assunti dalla X . In questo paragrafo introduciamo il modello conseguente all'assunzione di una relazione di tipo quadratico tra Y e X .

Il modello di regressione quadratica

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \epsilon_i \quad (13.15)$$

dove

β_0 = Y intercetta

β_1 = coefficiente relativo all'effetto lineare su Y

β_2 = coefficiente relativo all'effetto quadratico su Y

ϵ_i = l'errore casuale in Y corrispondente all' i -esima osservazione

Il **modello di regressione quadratica** è simile al modello di regressione multipla con due variabili esplicative introdotto nel paragrafo 13.2, con l'unica differenza che la seconda variabile esplicativa X_2 è il quadrato della prima X_1 .

Come nel modello di regressione multipla, i coefficienti di regressione (β_0 , β_1 e β_2) possono essere stimati mediante gli analoghi campionari (b_0 , b_1 e b_2) e di conseguenza l'espressione campionaria del modello quadratico con una sola variabile esplicativa (X_1) e una sola variabile dipendente (Y) assume la forma seguente.

Espressione campionaria del modello di regressione quadratica

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2 \quad (13.16)$$

Nell'equazione precedente, il primo coefficiente campionario b_0 rappresenta l'intercetta, il secondo b_1 il coefficiente lineare e il terzo b_2 l'effetto quadratico.

I coefficienti di regressione e la previsione di Y

A illustrazione del modello di regressione quadratica, supponete che una catena di supermercati intenda studiare l'elasticità del prezzo di rasoi usa e getta. Si preleva un campione di 15 negozi con un medesimo volume d'affari e una medesima localizzazione del prodotto (nel senso del bancone in cui sono esposti i rasoi). Si considerano 3 livelli di prezzo (79, 99 e 119 centesimi) e ciascuno di essi viene assegnato casualmente a 5 negozi. Nella Tabella 13.5 si riporta il numero di pacchi di rasoi venduti in una settimana nei negozi corrispondenti.

Nella Figura 13.8 si riporta il grafico di dispersione delle variabili prezzo e vendite. Il grafico mostra come le vendite diminuiscano all'aumentare del prezzo, ma il tasso a cui tale diminuzione si verifica si riduce da un certo punto in poi. Le vendite corrispondenti al prezzo di 99 centesimi sono decisamente inferiori alle vendite corrispondenti alla prezzo di 79 centesimi, ma le vendite corrispondenti a \$ 1.19 sono leggermente inferiori alle vendite corrispondenti a 99 centesimi. Pertanto, un modello di regressione quadratica sembra più appropriato di un modello di regressione lineare per la stima delle vendite sulla base del prezzo.

Nella Figura 13.9 sono riportati i valori dei coefficienti di regressione b_0 , b_1 e b_2 calcolati con Excel:

$$b_0 = 729.8665 \quad b_1 = -10.887 \quad b_2 = 0.0465$$

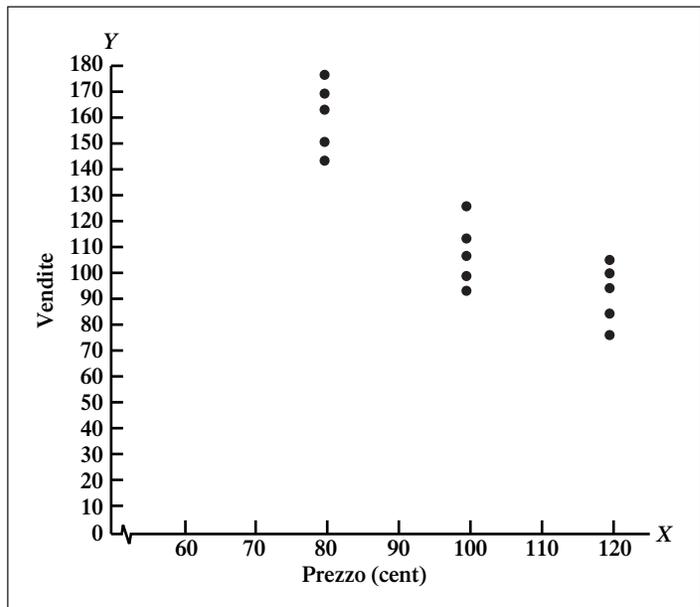


Tabella 13.5 *Vendite e prezzi di rasoi usa e getta per un campione di 15 negozi*

VENDITE	PREZZO (CENTS)	VENDITE	PREZZO (CENTS)
142	79	115	99
151	79	126	99
163	79	77	119
168	79	86	119
176	79	95	119
91	99	100	119
100	99	106	119
107	99		

FIGURA 13.8

Diagramma di dispersione delle variabili prezzo (X) e vendite (Y).



	A	B	C	D	E	F	G
1	Regression Analysis for Price Elasticity						
2							
3	<i>Regression Statistics</i>						
4	Multiple R	0.928581176					
5	R Square	0.862263					
6	Adjusted R Square	0.839306834					
7	Standard Error	12.86986143					
8	Observations	15					
9							
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	2	12442.8	6221.4	37.56127994	6.82816E-06	
13	Residual	12	1987.6	165.6333333			
14	Total	14	14430.4				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	729.8665	169.2575176	4.312165924	0.001009972	361.0860554	1098.646945
18	Price	-10.887	3.495239703	-3.114807831	0.008940633	-18.50247298	-3.271527022
19	Price^2	0.0465	0.017622784	2.638629696	0.0216284	0.008103254	0.084896746

FIGURA 13.9 Output parziale ottenuto con Excel per i dati relativi alle vendite di rasoi.

Pertanto il modello di regressione stimato assume la seguente espressione:

$$\hat{Y}_i = 729.8665 - 10.887X_{1i} + 0.0465X_{1i}^2$$

dove

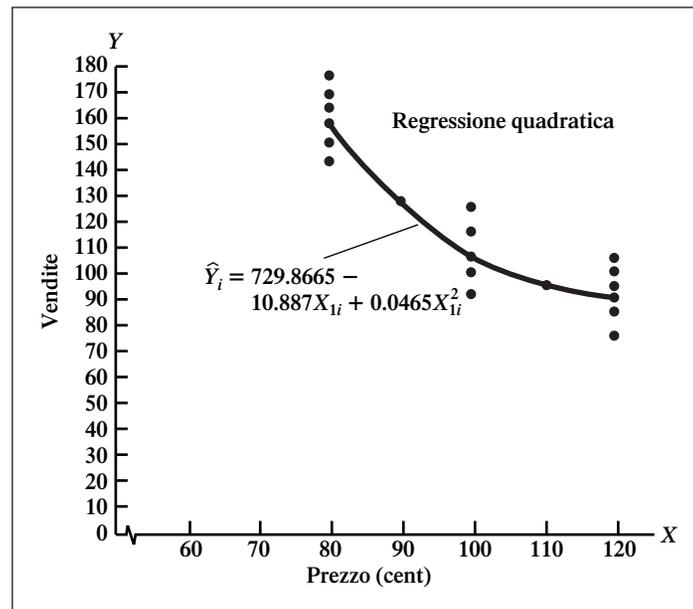
$$\hat{Y}_i = \text{vendite previste per il negozio } i$$

$$X_{1i} = \text{prezzo dei rasoi nel negozio } i$$

Nella Figura 13.10 riportiamo la rappresentazione grafica del modello di regressione stimato. L'intercetta b_0 non ha in questo caso un'interpretazione, ma deve semplicemente essere intesa come un punto di partenza. In maniera analoga il termine che si riferisce alla componente lineare, b_1 non ha un'interpretazione diretta nell'ambito del modello di regressione quadratica. Per cogliere il significato del coefficiente b_1 , osserviamo dalla Figura 13.11 che le vendite diminuiscono all'aumentare del prezzo, ma con un tasso di diminuzione che si riduce all'aumentare del prezzo.

FIGURA 13.10

Diagramma di dispersione delle variabili prezzo (X) e vendite (Y) dove la linea di tendenza raffigura la relazione quadratica tra le variabili (dati sulle vendite dei rasoi).



Per cogliere tale andamento delle vendite rispetto al prezzo, prevediamo l'ammontare delle vendite in corrispondenza dei prezzi 79, 99 e 119. In base al modello di regressione stimato

$$\hat{Y}_i = 729.8665 - 10.887X_{1i} + 0.0465X_{1i}^2$$

per $X_{1i} = 79$, abbiamo:

$$\hat{Y}_i = 729.8665 - 10.887(79) + 0.0465(79)^2 = 160$$

per $X_{1i} = 99$, abbiamo:

$$\hat{Y}_i = 729.8665 - 10.887(99) + 0.0465(99)^2 = 107.8$$

per $X_{1i} = 119$, abbiamo:

$$\hat{Y}_i = 729.8665 - 10.887(119) + 0.0465(119)^2 = 92.8$$

Pertanto, ci aspettiamo che un negozio che pratica il prezzo di 79 centesimi venda 160 pacchi di rasoi in più rispetto a un negozio in cui il prezzo è di 99 centesimi. Tuttavia, ci aspettiamo che un negozio che pratica il prezzo di 99 centesimi venda solo 15 pacchi in più rispetto a un negozio in cui il prezzo dei rasoi è di \$ 1.19 (119 centesimi).

Verifica della significatività di un modello quadratico

Introduciamo ora i metodi che ci consentono di stabilire se sussiste una relazione significativa tra le vendite Y e il prezzo X . Come per il modello di regressione lineare, l'ipotesi nulla e l'ipotesi alternativa sono specificate nella maniera seguente:

$$H_0: \beta_1 = \beta_2 = 0 \text{ (Non vi è una relazione significativa tra } X_1 \text{ e } Y)$$

$$H_1: \beta_2 \neq 0 \text{ o } \beta_1 \neq 0 \text{ (Vi è una relazione significativa tra } X_1 \text{ e } Y)$$

L'ipotesi nulla può essere verificata usando l'equazione (13.6):

$$F = \frac{MQR}{MQE}$$

Dall'output di Excel riportato nella Figura 13.9, abbiamo:

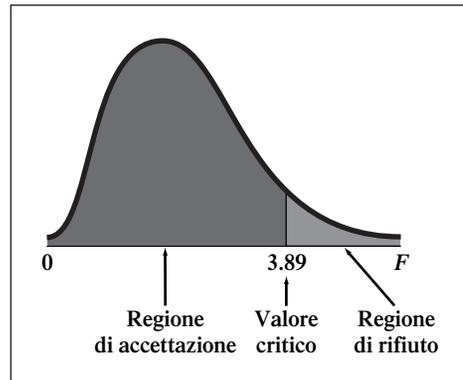
$$F = \frac{MQR}{MQE} = \frac{6221.4}{165.63} = 37.56$$

Per un livello di significatività pari a 0.05, in base alla Tavola E.5 il valore critico della statistica F per 2 e 12 gradi è uguale a 3.89 (vedi Figura 13.11).

Poiché $F = 37.56 > 3.89$ o poiché in base alla Figura 13.9 il p -value = 0.000006828 < 0.05, rifiutiamo l'ipotesi nulla H_0 e concludiamo che vi è una relazione significativa tra le vendite e il prezzo dei rasoi.

FIGURA 13.11

Verifica dell'esistenza di una relazione complessiva con un livello di significatività pari a 0.05 e 2 e 12 gradi di libertà.



Verifica dell'effetto quadratico

Quando ricorriamo a un modello di regressione per studiare la relazione tra due variabili, vorremmo non solo stimare il modello più accurato, ma anche quello più semplice. Pertanto risulta importante stabilire se vi sia una differenza significativa tra il modello di regressione quadratica

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \epsilon_i$$

e il modello lineare

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

I due modelli possono essere confrontati valutando l'effetto dell'aggiunta del termine quadratico a un modello in cui sia già stato inserito un termine lineare [$SQR(X_1^2 | X_1)$].

Nel Paragrafo 13.4 siamo ricorsi al test t sui coefficienti per stabilire se ciascuna variabile dà un contributo significativo al modello di regressione. Dal momento che l'output di Excel (cfr. Figura 13.9) riporta il valore dell'errore standard e della statistica t per ciascun coefficiente, possiamo verificare se il termine quadratico dia un contributo significativo specificando le seguenti ipotesi nulla e alternativa

H_0 : l'inclusione del termine quadratico non migliora in maniera significativa il modello ($\beta_2 = 0$)

H_1 : l'inclusione del termine quadratico migliora in maniera significativa il modello ($\beta_2 \neq 0$)

Per dati relativi alle vendite di rasoi

$$t = \frac{b_2}{S_{b_2}}$$

per cui

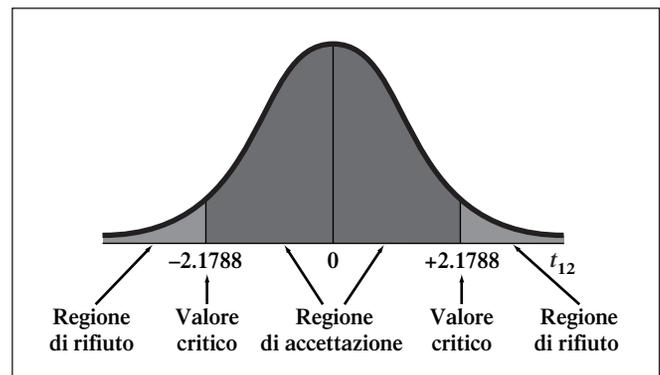
$$t = \frac{0.0465}{0.01762} = 2.64$$

Per un livello di significatività pari a 0.05, in base alla Tavola E.3 i valori critici della statistica t per 12 gradi sono +2.1788 e -2.1788 (vedi Figura 13.12).

Poiché $t = 2.64 > t_{12} = 2.1788$ o poiché il p -value = 0.0216 < 0.05, rifiutiamo l'ipotesi nulla H_0 e concludiamo che il modello quadratico è significativamente migliore del modello di regressione lineare per rappresentare la relazione tra le vendite e il prezzo. Rimandiamo all'Esempio 13.3 per un'ulteriore illustrazione del modello quadratico.

FIGURA 13.12

Verifica del contributo dell'effetto quadratico con un livello di significatività pari a 0.05 e 12 gradi di libertà.



Esempio 13.3 *Studio dell'effetto quadratico in un modello di regressione multipla*

Supponete che un costruttore intenda stabilire l'effetto che la temperatura esterna e il grado di isolamento dell'appartamento hanno sulla quantità di combustibile per riscaldamento utilizzato. Viene preso in considerazione un campione di 15 case monofamiliari e viene rilevato il consumo di combustibile per il mese di gennaio. I dati raccolti sono riportati nel file HTNGOIL. Si stima un modello di regressione lineare e i risultati sono riportati nel seguente tabulato

Output parziale ottenuto con Excel per i dati relativi al consumo mensile di combustibile per riscaldamenti.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
4	Multiple R	0.982654757					
5	R Square	0.965610371					
6	Adjusted R Square	0.959878766					
7	Standard Error	26.01378323					
8	Observations	15					
9							
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	2	228014.6263	114007.3132	168.4712028	1.65411E-09	
13	Residual	12	8120.603016	676.716918			
14	Total	14	236135.2293				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	562.1510092	21.09310433	26.65093769	4.77868E-12	516.1930837	608.1089348
18	Temperature	-5.436580588	0.336216167	-16.16989642	1.64178E-09	-6.169132673	-4.704028503
19	Insulation	-20.01232067	2.342505227	-8.543127434	1.90731E-06	-25.11620102	-14.90844031

Il grafico dei residui per la variabile isolamento evidenzia la presenza di un effetto di tipo non lineare. Stimate un modello di regressione in cui il quadrato della variabile isolamento viene impiegato come variabile esplicativa. Per un livello di significatività pari a 0.05 vi è prova di un effetto di tipo non lineare della variabile isolamento sull'ammontare di combustibile consumato?

SOLUZIONE

	A	B	C	D	E	F	G
1	Regression Analysis: Curvilinear Effect for Attic Insulation?						
2							
3	<i>Regression Statistics</i>						
4	Multiple R	0.986157697					
5	R Square	0.972507004					
6	Adjusted R Square	0.965008914					
7	Standard Error	24.29377937					
8	Observations	15					
9							
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	3	229643.1645	76547.72149	129.7006349	7.26403E-09	
13	Residual	11	6492.064875	590.1877159			
14	Total	14	236135.2293				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	624.5864209	42.43515952	14.71860664	1.39085E-08	531.1872173	717.9856245
18	Temperature	-5.362603095	0.317128467	-16.90987611	3.20817E-09	-6.060598498	-4.664607691
19	Insulation	44.58678859	14.9546884	2.981458884	0.012486902	77.50185248	-11.67172469
20	Insulation^2	1.866704651	1.123755228	1.661131004	0.124891755	-0.606665181	4.340074482

Output ottenuto con Excel relativo al modello di regressione con componente non lineare nella variabile isolamento

Il modello di regressione stimato è:

$$\hat{Y}_i = 624.5864 - 5.3626X_{1i} - 44.5868X_{2i} + 1.8667X_{2i}^2$$

Per verificare la significatività dell'effetto non lineare, consideriamo le seguenti ipotesi:

H_0 : l'inclusione del termine quadratico non migliora in maniera significativa il modello ($\beta_3 = 0$)

H_1 : l'inclusione del termine quadratico migliora in maniera significativa il modello ($\beta_3 \neq 0$)

In base all'output ottenuto $t = 1.661$ e $p\text{-value} = 0.1249 > 0.05$. Possiamo decidere allora di non rifiutare l'ipotesi nulla e concludere che non vi è prova sufficiente per ritenere che il l'effetto non lineare della variabile isolamento sia diverso da 0. Pertanto se ricerchiamo il modello più semplice, dovremmo ricorrere al modello $\hat{Y}_i = 562.151 - 5.43658X_{1i} - 20.0123X_{2i}$.

Calcolo del coefficiente di determinazione multiplo

Il coefficiente di determinazione nel modello di regressione quadratico si può calcolare in base all'equazione (13.4):

$$r_{Y.12}^2 = \frac{SQR}{SQT}$$

Dalla Figura 13.9

$$SQR = 12,442.8 \quad SQT = 14\,430.4$$

Pertanto

$$r_{Y.12}^2 = \frac{SQR}{SQT} = \frac{12\,442.8}{14\,430.4} = 0.862$$

Un valore del coefficiente di determinazione uguale a 0.862 ci dice che l'82.2% della variabilità nelle vendite può essere spiegata dalla relazione di tipo non lineare tra le vendite e i prezzi. È possibile calcolare anche l' r^2 corretto con cui si tiene conto del numero di variabili esplicative e dei gradi di libertà. Nel modello di regressione quadratico considerato, $p = 2$ dal momento che vi sono due variabili esplicative, X_1 e X_1^2 . Pertanto in base all'equazione (13.5) per i dati relativi alla vendite di rasoi:

$$\begin{aligned} r_{\text{adj}}^2 &= 1 - \left[(1 - r_{Y.12}^2) \frac{15 - 1}{15 - 2 - 1} \right] \\ &= 1 - \left[(1 - 0.862) \frac{14}{12} \right] \\ &= 1 - 0.161 \\ &= 0.839 \end{aligned}$$

Esercizi del paragrafo 13.6

- **13.18** Supponete che sia stato stimato il seguente modello di regressione quadratica per un campione di ampiezza $n = 25$

$$\hat{Y}_i = 5 + 3X_{1i} + 1.5X_{1i}^2$$

- Prevedete il valore di Y per $X = 2$
- Supponete che il valore della statistica t per il termine non lineare sia uguale a 2.35. Per un livello di significatività pari a 0.05 si può ritenere che il modello di regressione quadratica sia migliore del modello di regressione lineare?



**DATASET
SPEED**

- (c) Supponete che il valore della statistica t per il termine non lineare sia uguale a 1.17. Per un livello di significatività pari a 0.05 si può ritenere che il modello di regressione quadratica sia migliore del modello di regressione lineare?
- (d) Supponete che il coefficiente di regressione per la componente lineare sia -3.0 . Prevedete il valore di Y per $X = 2$.
- **13.19** Il ricercatore di una compagnia petrolifera intende sviluppare un modello per prevedere le miglia percorse da un'automobile con un gallone di benzina sulla base della velocità. Si esegue un esperimento in cui una macchina test è guidata a velocità comprese tra le 10 miglia all'ora e le 75 miglia all'ora con incrementi di 5 miglia ogni due prove.

OSSERVAZIONI	MIGLIA PER GALLONE	VELOCITÀ (MIGLIA ALL'ORA)	OSSERVAZIONI	MIGLIA PER GALLONE	VELOCITÀ (MIGLIA ALL'ORA)
1	4.8	10	15	21.3	45
2	5.7	10	16	22.0	45
3	8.6	15	17	20.5	50
4	7.3	15	18	19.7	50
5	9.8	20	19	18.6	55
6	11.2	20	20	19.3	55
7	13.7	25	21	14.4	60
8	12.4	25	22	13.7	60
9	18.2	30	23	12.1	65
10	16.8	30	24	13.0	65
11	19.9	35	25	10.1	70
12	19.0	35	26	9.4	70
13	22.4	40	27	8.4	75
14	23.5	40	28	7.6	75

Nell'ipotesi che sussista una relazione di tipo quadratico tra la velocità e le miglia, con l'uso di Excel:

- (a) Rappresentate il diagramma di dispersione della velocità e delle miglia.
- (b) Fornite l'espressione del modello di regressione quadratica.
- (c) Prevedete le miglia percorse da una macchina con un gallone che viaggia a 55 miglia all'ora.
- (d) Conducete un'analisi dei residui sui risultati ottenuti e stabilite se il modello proposto è adeguato.
- (e) Stabilite se vi sia una relazione quadratica tra le miglia percorse con un gallone di benzina e la velocità per un livello di significatività pari a 0.05.
- (f) Per un livello di significatività pari a 0.05, stabilite se il modello di regressione quadratica è più adeguato del modello di regressione lineare.
- (g) Fornite un'interpretazione del coefficiente di determinazione $r^2_{Y,12}$.
- (h) Calcolate r^2 aggiustato.
- **13.20** Il revisore dei conti dell'amministrazione di una contea intende sviluppare un modello per prevedere le tasse della contea sulla base dell'età delle case monofamiliari. Nella tabella seguente si riportano i dati relativi a un campione di 19 case monofamiliari.

TASSE DELLA CONTEA		TASSE DELLA CONTEA	
(\$)	ETÀ	(\$)	ETÀ
925	1	480	20
870	2	486	22
809	4	462	25
720	4	441	25
694	5	426	30



**DATASET
TAXES**

TASSE DELLA CONTEA		TASSE DELLA CONTEA	
(\$)	ETÀ	(\$)	ETÀ
630	8	368	35
626	10	350	40
562	10	348	50
546	12	322	50
523	15		

Nell'ipotesi che tra l'ammontare delle tasse e l'età delle case sussista una relazione di tipo quadratico, con l'uso di Excel:

- Rappresentate il diagramma di dispersione dell'età e delle tasse.
- Fornite l'espressione del modello di regressione quadratica.
- Prevedete l'ammontare delle tasse per una casa di 20 anni.
- Conducete un'analisi dei residui sui risultati ottenuti e stabilite se il modello proposto è adeguato.
- Stabilite se vi sia una relazione tra le tasse e l'età delle case per un livello di significatività pari a 0.05.
- Quale è il valore del p -value in (e)? Fornite un'interpretazione del suo significato.
- Per un livello di significatività pari a 0.05, stabilite se modello di regressione quadratica è più adeguato del modello di regressione lineare.
- Quale è il valore del p -value in (g)? Fornite un'interpretazione del suo significato.
- Fornite un'interpretazione del coefficiente di determinazione $r^2_{y.12}$.
- Calcolate l' r^2 aggiustato.

13.7

I MODELLI CON VARIABILI DUMMY

Sino a ora abbiamo supposto che le variabili esplicative dei modelli di regressione considerati fossero variabili quantitative. Tuttavia, in molti casi si rende necessaria l'introduzione anche di variabili qualitative. Nell'esempio relativo alle vendite della barretta energetica Omnipower, abbiamo assunto come variabili esplicative il prezzo e l'ammontare delle spese promozionali. Potrebbe essere utile inserire nel modello una variabile che rifletta l'effetto del posto nel negozio in cui sono esposte le barrette (per esempio un espositore alla fine di un corsia o un espositore non alla fine di una corsia).

Variabili di tipo qualitativo possono essere inserite in un modello di regressione mediante l'uso delle **variabili dummy**. Se la variabile qualitativa presa in considerazione assume solo due valori, sarà necessario inserire una sola variabile dummy, X_d , così definita:

$$X_d = 0 \text{ se si osserva il primo valore}$$

$$X_d = 1 \text{ se si osserva il secondo valore}$$

A illustrazione dell'uso delle variabili dummy nella regressione, prendete in considerazione un modello in cui si intende prevedere il valore di alcune abitazioni sulla base della superficie riscaldata (in migliaia di piedi al quadrato) e di una variabile con cui si tiene conto della presenza o meno di un caminetto. Si estrae un campione di 15 casi e la Tabella 13.6 riporta per ciascuna casa il valore accertato, la superficie riscaldata e se vi è o meno un caminetto.

In questo caso si introduce una variabile dummy X_2 così definita:

$$X_2 = 0 \text{ se la casa non ha caminetto}$$

$$X_2 = 1 \text{ se la casa ha caminetto}$$

Nell'ipotesi che l'inclinazione del valore delle case (variabile dipendente) rispetto alla superficie riscaldata sia uguale indipendentemente dalla presenza o meno di un caminetto nelle case, proponiamo il seguente modello di regressione lineare:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Tabella 13.6 Prevedere il valore sulla base della superficie riscaldata e della presenza di un caminetto

ABITAZIONE	SUPERFICIE RISCALDATA (IN MIGLIAIA DI PIEDI)			ABITAZIONE	SUPERFICIE RISCALDATA (IN MIGLIAIA DI PIEDI)		
	VALORE ACCERTATO (\$ 000)	AL QUADRATO	CAMINETTO		VALORE ACCERTATO (\$ 000)	AL QUADRATO	CAMINETTO
1	84.4	2.00	Yes	9	78.5	1.59	Yes
2	77.4	1.71	No	10	79.2	1.50	Yes
3	75.7	1.45	No	11	86.7	1.90	Yes
4	85.9	1.76	Yes	12	79.3	1.39	Yes
5	79.1	1.93	No	13	74.5	1.54	No
6	70.4	1.20	Yes	14	83.8	1.89	Yes
7	75.8	1.55	Yes	15	76.8	1.59	No
8	85.9	1.93	Yes				



DATASET
HOUSE3

dove

Y_i = valore accertato della i -esima casa in migliaia di dollari

β_0 = intercetta

β_1 = inclinazione del valore delle case rispetto alla superficie riscaldata se si tiene costante l'effetto della presenza del caminetto

β_2 = effetto addizionale della presenza del caminetto tenuto costante l'effetto della superficie riscaldata

ϵ_i = errore corrispondente alla casa i

Nella Figura 13.13 si riporta l'output di Excel

Dall'output ricaviamo la seguente espressione per il modello di regressione stimato;

$$\hat{Y}_i = 50.09 + 16.186X_{1i} + 3.853X_{2i}$$

che per le case senza caminetto si riduce a

$$\hat{Y}_i = 50.09 + 16.186X_{1i}$$

poiché $X_2 = 0$; mentre per le case con caminetto otteniamo la seguente espressione:

$$\hat{Y}_i = 53.943 + 16.186X_{1i}$$

FIGURA 13.13

Output di Excel per il modello di regressione che include la superficie riscaldata e la presenza del caminetto.

	A	B	C	D	E	F	G
1	Regression Analysis for Fireplace Prediction						
2							
3	Regression Statistics						
4	Multiple R	0.900587177					
5	R Square	0.811057264					
6	Adjusted R Square	0.779566808					
7	Standard Error	2.262595954					
8	Observations	15					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	2	263.7039146	131.8519573	25.75565321	4.54968E-05	
13	Residual	12	61.4320854	5.11934045			
14	Total	14	325.136				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	50.09048899	4.351657945	11.5106678	7.67943E-08	40.60904099	59.57193699
18	Heating	16.18583395	2.574441705	6.28712389	4.02437E-05	10.57660743	21.79506047
19	Fireplace	3.852982483	1.241222689	3.10418309	0.009118854	1.148590609	6.557374357

poiché $X_2 = 1$, di modo che 3.853 si aggiunge a 50.09. In questo caso i coefficienti di regressione vanno interpretati nella maniera seguente:

1. Mantenendo costante l'effetto dovuto alla presenza o meno di un caminetto, ci si aspetta che il valore delle case aumenti di \$ 16 186 in corrispondenza di un aumento di 1000 piedi al quadrato della superficie da riscaldare.
2. Mantenendo costante la superficie riscaldata, prevediamo che la presenza di un caminetto aumenti il valore delle case di \$ 3853.

Dalla Figura 13.13 ricaviamo che il valore della statistica t riferita all'inclinazione del valore delle case rispetto all'area riscaldata è 6.29, mentre il p -value è all'incirca 0.000; il valore della statistica t che si riferisce invece alla presenza o meno di un caminetto nelle case ha valore 3.10 e il corrispondente p -value è uguale 0.009. Pertanto per un livello di significatività pari a 0.01 entrambe le variabili danno un contributo significativo al modello di regressione. Inoltre l'81.1% della variabilità totale nel valore delle vendite è spiegato dalla variabilità nella superficie riscaldata e nella presenza o meno di un caminetto.

Tuttavia, per poter fare uso del modello preso in considerazione dobbiamo verificare che l'inclinazione del valore delle case rispetto alla superficie di riscaldamento non dipenda dalla presenza o meno di un caminetto. A tale scopo introduciamo una nuova variabile, chiamata **termine di interazione**, data dal prodotto della variabile esplicativa X_1 e della dummy X_2 e verifichiamo se questa nuova variabile dà un contributo significativo al modello considerato. Se il termine di interazione è significativo, non possiamo ricorrere al modello descritto per effettuare delle previsioni.

L'output riportato nella Figura 13.14 si riferisce a un modello di regressione in cui figura tra le variabili esplicative assieme alle variabili X_1 e X_2 già considerate, il termine di interazione:

$$X_3 = X_1 \times X_2$$

Per verificare l'ipotesi $H_0: \beta_3 = 0$ contro l'ipotesi $H_1: \beta_3 \neq 0$, dalla Figura 13.14 ricaviamo che il valore della statistica t corrispondente all'interazione tra la superficie riscaldata e la presenza di un caminetto è 1.48. Poiché il p -value = 0.166 > 0.05, non rifiutiamo l'ipotesi nulla. Concludiamo che il termine di interazione non dà un contributo significativo al modello una volta che la superficie riscaldata e la presenza di un caminetto siano già state incluse come variabili esplicative.

FIGURA 13.14

Output di Microsoft Excel per il modello di regressione che include la superficie riscaldata, la presenza del caminetto e il termine di interazione tra la superficie e la presenza del caminetto.

	A	B	C	D	E	F	G
1	Regression Analysis for Fireplace Prediction						
2							
3	Regression Statistics						
4	Multiple R	0.917906505					
5	R Square	0.842552352					
6	Adjusted R Square	0.799612084					
7	Standard Error	2.157268864					
8	Observations	15					
9							
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	3	273.9441015	91.3147005	19.62149745	0.000100865	
13	Residual	11	51.19189849	4.653808954			
14	Total	14	325.136				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	62.9521815	9.612176928	6.54921169	4.13993E-05	41.79591203	84.10845098
18	Heating	8.362420012	5.817298426	1.437509201	0.178405557	-4.441373971	21.16621399
19	Fireplace	-11.84036371	10.64550326	-1.1224086	0.289751611	-35.27097026	11.59024285
20	X1*X2	9.518000219	6.416468169	1.483370597	0.166052624	-4.604558143	23.64055858

Esempio 13.4 *Studio di un modello di regressione contenente una variabile dummy*

Nell'Esempio 13.3 abbiamo costruito un modello di regressione in cui il consumo di petrolio da combustibile viene spiegato in base alla temperatura atmosferica X_1 e al grado di isolamento dell'appartamento. Ora supponete che delle 15 case che formano il campione considerato, le case 1, 4, 6, 7, 8, 10 siano delle abitazioni agricole. Proponete e stimate un opportuno modello di regressione multipla.

SOLUZIONE

Definiamo la seguente variabile dummy X_3 :

$$X_3 = 0 \text{ se non si tratta di una casa agricola}$$

$$X_3 = 1 \text{ se si tratta di una casa agricola}$$

Nell'ipotesi che l'inclinazione del consumo di combustibile rispetto alla temperatura e l'inclinazione rispetto all'isolamento dell'appartamento siano uguali per le case di tipo agricolo e quelle di tipo non agricolo, prendiamo in considerazione il seguente modello di regressione:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

dove

Y_i = consumo mensile di combustibile per riscaldamento

β_0 = intercetta

β_1 = inclinazione del consumo di combustibile rispetto alla temperatura atmosferica se si tiene costante l'effetto dell'isolamento dell'appartamento;

β_2 = inclinazione del consumo di combustibile rispetto all'isolamento dell'appartamento tenuto costante l'effetto della temperatura atmosferica;

β_3 = effetto addizionale del tipo di casa (se agricola o meno) dati gli effetti della temperatura atmosferica e dell'isolamento;

ϵ_i = errore corrispondente alla casa i

Nella figura seguente si riporta l'output di Excel relativo al modello preso in considerazione.

Output di Microsoft Excel di un modello di regressione contenente una variabile dummy.

	A	B	C	D	E	F	G
1	Regression Model with Ranch-Style						
2							
3	Regression Statistics						
4	Multiple R	0.994206187					
5	R Square	0.988445943					
6	Adjusted R Square	0.985294836					
7	Standard Error	15.7489393					
8	Observations	15					
9							
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	3	233406.9094	77802.30312	313.6821708	6.21548E-11	
13	Residual	11	2728.319981	248.0290892			
14	Total	14	236135.2293				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	592.5401168	14.33698425	41.32948091	2.02317E-13	560.9846112	624.0956223
18	Temperature	-5.525100884	0.204431228	-27.0266971	2.07188E-11	-5.975051212	-5.075150557
19	Insulation	-21.37612794	1.448019304	-14.76232249	1.34816E-08	-24.56319855	-18.18905733
20	Ranch-Style	-38.97266608	8.358437237	-4.662673772	0.000690709	-57.3694717	-20.57586045

Dall'output ricaviamo la seguente espressione per il modello stimato:

$$\hat{Y}_i = 592.5401 - 5.5251X_{1i} - 21.3761X_{2i} - 38.9726X_{3i}$$

Per le case che non sono agricole, la precedente espressione diventa:

$$\hat{Y}_i = 592.5401 - 5.5251X_{1i} - 21.3761X_{2i}$$

poiché $X_3 = 0$; per le case agricole il modello stimato avrà la seguente forma:

$$\hat{Y}_i = 553.5674 - 5.5251X_{1i} - 21.3761X_{2i}$$

poiché $X_3 = 1$. Inoltre osserviamo che:

1. Mantenendo costante l'effetto dell'isolamento dell'appartamento e del tipo di casa, ci aspettiamo che il consumo di combustibile si riduca di 5.525 galloni per ogni grado F° di aumento della temperatura atmosferica;
2. Mantenendo costante l'effetto della temperatura atmosferica e del tipo di casa, ci aspettiamo che il consumo di combustibile si riduca di 21.376 galloni per ogni pollice in più di isolamento;
3. b_3 rappresenta l'effetto che il tipo di casa (agricola $X_3 = 1$, non agricola $X_3 = 0$) ha sul consumo di combustibile. Pertanto riteniamo che mantenendo costante l'effetto della temperatura atmosferica e del grado di isolamento dell'appartamento, un casa di tipo agricolo consumi 38.973 galloni in meno di combustibile per riscaldamento rispetto a una casa agricola.

Dall'output di Excel ricaviamo che le tre statistiche t riferite alla temperatura, al grado di isolamento e al tipo di casa hanno rispettivamente valori -27.03 , -14.76 e -4.66 . I corrispondenti p -value sono tutti molto piccoli e comunque inferiori a 0.001. Pertanto possiamo concludere che tutte e tre le variabili danno un contributo significativo al modello di regressione. Inoltre il 98.8% della variabilità totale del consumo di combustibile è spiegato dalla variabilità delle tre variabili considerate.

Prima di poter fare uso del modello introdotto nell'Esempio 13.4, è, tuttavia, necessario verificare se l'inclinazione del consumo di combustibile per riscaldamento rispetto alla temperatura sia la stessa per le case di tipo agricolo e quelle non agricole e che lo stesso valga per l'inclinazione rispetto all'isolamento.

Esempio 13.5 *Valutazione di un modello lineare con due termini di interazione*

Con riferimento ai dati dell'Esempio 13.4, stabilite se i due termini di interazione danno un contributo significativo al modello di regressione.

SOLUZIONE

Come abbiamo visto, quando si introduce una variabile dummy in un modello di regressione semplice, l'indipendenza dell'inclinazione di Y rispetto X dall'effetto della variabile dummy può essere valutata introducendo un termine di interazione dato dal prodotto di X e della dummy e verificando se tale termine dà un contributo significativo al modello di regressione. Nel caso che qui prendiamo in considerazione, vi sono due variabili esplicative per cui dovremo valutare il contributo proveniente da due termini di interazione, ciascuno il prodotto di una delle variabili esplicative per la variabile dummy:

$$X_4 = X_1 \times X_3 \text{ e } X_5 = X_2 \times X_3$$

Riportiamo di seguito l'output di Excel per il modello di regressione che include la temperatura X_1 , l'isolamento X_2 , il tipo di casa X_3 , il termine di interazione X_1 con X_3 (X_4) e il termine di interazione X_2 con X_3 (X_5).

Output di Microsoft Excel per il modello di regressione che include la temperatura X_1 , l'isolamento X_2 , il tipo di casa X_3 , il termine di interazione tra temperatura e tipo di casa X_4 , e l'interazione tra isolamento e tipo di casa (X_5).

	A	B	C	D	E	F	G
1	Regression Model with Two Interactions						
2							
3	Regression Statistics						
4	Multiple R	0.994961389					
5	R Square	0.989948166					
6	Adjusted R Square	0.984363813					
7	Standard Error	16.23984198					
8	Observations	15					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	5	233761.6371	46752.32743	177.2717932	1.04608E-08	
13	Residual	9	2373.592207	263.7324675			
14	Total	14	236135.2293				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	599.9782434	16.66954901	35.99247004	4.88276E-11	562.2690749	637.6874118
18	Temperature	-5.624853951	0.374378066	-15.02452857	1.11221E-07	-6.47175662	-4.777951282
19	Insulation	-21.84863672	1.955643913	-11.17209354	1.41221E-06	-26.27261397	-17.42465946
20	Ranch-Style	-76.85563866	33.87148579	-2.269036532	0.049440882	-153.4783213	-0.232956068
21	X1*X3	0.379785059	0.499241778	0.760723714	0.466284201	-0.749579165	1.509149284
22	X2*X3	3.924142467	3.954160702	0.992408443	0.346928526	-5.020797306	12.86908224

L'ipotesi nulla e l'ipotesi alternativa sono:

$$H_0: \beta_4 = \beta_5 = 0 \text{ (non vi è interazione tra } X_1 \text{ o } X_2 \text{ e } X_3.)$$

$$H_1: \beta_4 \text{ e/o } \beta_5 \neq 0 \text{ (vi è interazione tra } X_1 \text{ o } X_2 \text{ e } X_3.)$$

Per verificare l'ipotesi nulla dobbiamo confrontare la somma dei quadrati della regressione per il modello che contiene entrambe le interazioni con quella corrispondente al modello senza interazioni. Dall'output dell'Esempio 13.4 e da quello di sopra riportato ricaviamo:

$$SQR(X_1, X_2, X_3, X_4, X_5) = 233\,761.6371 \text{ con 5 gradi di libertà}$$

e

$$SQR(X_1, X_2, X_3) = 233\,406.9094 \text{ con 3 gradi di libertà}$$

Pertanto

$$SQR(X_1, X_2, X_3, X_4, X_5) - SQR(X_1, X_2, X_3) = 233\,761.6371 - 233\,406.9094 = 354.7277$$

e la differenza dei gradi di libertà è $5 - 3 = 2$.

Ricorriamo al test F per verificare l'ipotesi nulla e quindi in base all'Equazione (13.11) avremo

$$F = \frac{[SQR(X_1, X_2, X_3, X_4, X_5) - SQR(X_1, X_2, X_3)] / \text{differenza tra i gradi di libertà}}{MQE(X_1, X_2, X_3, X_4, X_5)}$$

$$F = \frac{354.7277 / 2}{263.7325}$$

$$= 0.673$$

Per un livello di significatività pari a 0.05, poiché $F = 0.673 < 4.26$, concludiamo che nessuno dei termini di interazione dà un contributo significativo al modello in cui sono già incluse come variabili esplicative la temperatura X_1 , l'isolamento X_2 , il tipo di casa X_3 . A questo punto dovremmo valutare il contributo di ciascun termine di interazione separatamente per stabilire se dovrebbe essere incluso nel modello.

Esercizi del paragrafo 13.7

- **13.21** Supponete che X_1 sia una variabile quantitativa e X_2 una variabile dummy. Per un campione di ampiezza $n = 20$ si ottiene la seguente stima di un modello di regressione:

$$\hat{Y}_i = 6 + 4X_{1i} + 2X_{2i}$$



DATASET
PETFOOD

- (a) Fornite una spiegazione dell'inclinazione rispetto alla variabile X_1 .
- (b) Interpretate il significato dell'inclinazione rispetto alla variabile X_2 .
- (c) Supponete che il valore della statistica t per la variabile X_2 sia 3.27. Per un livello di significatività pari a 0.05 ritenete che la variabile X_2 dia un contributo significativo al modello di regressione?

- **13.22** Il responsabile del marketing di una catena di supermercati vuole stabilire l'effetto che possono avere sulle vendite di cibo per animali lo spazio dedicato sullo scaffale alle confezioni e la posizione all'interno della corsia (davanti o dietro). Si estrae un campione di 12 negozi di uguali dimensioni e si ottengono i seguenti risultati

NEGOZIO			SPAZIO SULLO SCAFFALE,	POSIZIONE	VENDITE SETTIMANALI, Y (CENTINAIA DI DOLLARI)	NEGOZIO			SPAZIO SULLO SCAFFALE,	POSIZIONE	VENDITE SETTIMANALI, Y (CENTINAIA DI DOLLARI)
X (PIEDI)						X (PIEDI)					
1	5	Dietro			1.6	7	15	Dietro			2.3
2	5	Fronte			2.2	8	15	Dietro			2.7
3	5	Dietro			1.4	9	15	Fronte			2.8
4	10	Dietro			1.9	10	20	Dietro			2.6
5	10	Dietro			2.4	11	20	Dietro			2.9
6	10	Fronte			2.6	12	20	Fronte			3.1

Con l'uso di Microsoft Excel risolvete i seguenti punti:

- (a) Formulate l'espressione del modello di regressione multipla.
- (b) Fornite un'interpretazione delle inclinazioni.
- (c) Prevedete l'ammontare delle vendite di cibo per animali per un negozio in cui le confezioni occupano uno spazio di 8 piedi su uno scaffale posizionato nella parte posteriore della corsia.
- (d) Conducete un'analisi dei residui e valutate l'adeguatezza del modello.
- (e) Per un livello di significatività pari a 0.05 stabilite se vi è una relazione significativa tra le vendite e le due variabili esplicative (spazio sullo scaffale e posizione nella corsia).
- (f) Per un livello di significatività pari a 0.05 valutate il contributo di ciascuna variabile esplicativa al modello di regressione. Sulla base dei risultati ottenuti indicate quale modello andrebbe usato.
- (g) Costruite un intervallo di confidenza del 95% per la relazione tra le vendite e lo spazio sullo scaffale e tra le vendite e la posizione nella corsia.
- (h) Confrontate il valore dell'inclinazione ottenuto in (b) con quello ottenuto per il modello di regressione semplice nell'Esercizio 9.2. Spiegate le differenze eventualmente riscontrate.
- (i) Interpretate il significato del coefficiente di determinazione $r_{Y,12}^2$.
- (j) Calcolate l' r^2 corretto.
- (k) Confrontate $r_{Y,12}^2$ con r^2 calcolato nell'Esercizio 9.7(a).
- (l) Calcolate i coefficienti di correlazione parziali e spiegate il significato.
- (m) Quale ipotesi circa l'inclinazione delle vendite rispetto allo spazio sullo scaffale deve essere fatta in questo caso?
- (n) Includete nel modello un termine di interazione e, per un livello di significatività pari a 0.05, valutatene il contributo al modello.
- (o) Sulla base dei risultati ottenuti ai punti (e) e (n) quale modello ritenete sia più appropriato?

- **13.23** Una società immobiliare intende studiare la relazione tra la dimensione di una casa monofamiliare (misurata in numero di locali) e il prezzo di vendita. Lo studio viene condotto in due diverse aree della città, Est e Ovest. Si estrae un campione di 20 casi e si ottengono i seguenti risultati:



DATASET
NEIGHBOR

PREZZO DI VENDITA	NUMERO DI LOCALI	AREA	PREZZO DI VENDITA	NUMERO DI LOCALI	AREA
109.6	7	Est	108.5	6	Ovest
107.4	8	Est	181.3	13	Ovest
140.3	9	Est	137.4	10	Ovest
146.5	12	Est	146.2	10	Ovest
98.2	6	Est	142.4	9	Ovest
137.8	9	Est	123.7	8	Ovest
124.1	10	Est	129.6	8	Ovest
113.2	8	Est	143.6	9	Ovest
127.8	9	Est	160.7	11	Ovest
125.3	8	Est	148.3	9	Ovest

Con l'uso di Microsoft Excel risolvete i seguenti punti:

- Formulate l'espressione del modello di regressione multipla.
- Fornite un'interpretazione delle inclinazioni.
- Prevedete il prezzo di vendita di una casa di 9 locali situata nell'area Est.
- Conducete un'analisi dei residui e valutate l'adeguatezza del modello.
- Per un livello di significatività pari a 0.05 stabilite se vi è una relazione significativa tra il prezzo di vendita e le due variabili esplicative (numero di locali e area).
- Per un livello di significatività pari a 0.05 valutate il contributo di ciascuna variabile esplicativa al modello di regressione. Sulla base dei risultati ottenuti indicate quale modello andrebbe usato.
- Costruite un intervallo di confidenza del 95% per la relazione tra il prezzo di vendita e il numero di locali e tra il prezzo di vendita e l'area.
- Interpretate il significato del coefficiente di determinazione $r^2_{Y.12}$.
- Calcolate l' r^2 corretto.
- Calcolate i coefficienti di correlazione parziali e spiegate il significato.
- Quale ipotesi potete fare circa l'inclinazione del prezzo rispetto al numero di locali?
- Includete nel modello un termine di interazione e, per un livello di significatività pari a 0.05, valutatene il contributo al modello.
- Sulla base dei risultati ottenuti ai punti (e) e (n) quale modello ritenete sia più appropriato?

13.8

LA MULTICOLLINEARITÀ

Uno dei problemi che si può presentare nell'analisi di un modello di regressione multipla è la **multicollinearità** delle variabili esplicative, che consiste nella presenza di una elevata correlazione tra le variabili esplicative. In questo caso, le variabili collineari non forniscono delle informazioni aggiuntive e risulta difficile individuare l'effetto che ciascuna di esse ha sulla variabile risposta. I valori dei coefficienti di regressione per queste variabili potrebbero variare in maniera elevata a seconda di quali delle variabili indipendenti sono incluse nel modello.

Un metodo per la misurazione della multicollinearità si basa sul **variance inflationary factor (VIF)**, che si può calcolare per ciascuna delle variabili esplicative. Il VIF_j corrispondente alla variabile j , è di seguito definito.

Variance Inflationary Factor

$$VIF_j = \frac{1}{1 - R_j^2} \quad (13.17)$$

dove R_j^2 è il coefficiente di determinazione che caratterizza il modello in cui la variabile dipendente è X_j e tutte le altre variabili esplicative sono incluse nel modello.

In presenza di due sole variabili esplicative, R_1^2 è il coefficiente di determinazione della regressione di X_1 su X_2 ed è identico a R_2^2 , il coefficiente di determinazione della regressione di X_2 su X_1 . Se, ad esempio, vi sono tre variabili esplicative, R_1^2 è il coefficiente di determinazione della regressione di X_1 su X_2 e X_3 ; R_2^2 il coefficiente di determinazione della regressione di X_2 su X_1 e X_3 e R_3^2 il coefficiente di determinazione di X_3 con X_1 e X_2 .

Se le variabili esplicative non sono correlate, il VIF_j è uguale a 1. Se le variabili esplicative sono altamente correlate tra di loro, il VIF_j è elevato e potrebbe eccedere 10. Marquardt (riferimento bibliografico 3) sostiene che se VIF_j è maggiore di 10, vi è un'elevata correlazione tra X_j e le altre variabili esplicative. Altri studiosi hanno una posizione più prudente e suggeriscono di correre a metodi di stima diversi dai minimi quadrati quando si è in presenza di un VIF_j maggiore di 5 (riferimento bibliografico 6).

Tornando ai dati relativi alle vendite della barretta Omnipower, la correlazione tra le due variabili esplicative, prezzo e spese promozionali, è uguale a 0.0968. Pertanto, in base all'equazione (13.17):

$$\begin{aligned} VIF_1 = VIF_2 &= \frac{1}{1 - (0.0968)^2} \\ &= 1.009 \end{aligned}$$

Concludiamo che non vi è prova della presenza di multicollinearità tra le variabili.

Esercizi del paragrafo 13.8

- **13.24** Se il coefficiente di correlazione tra due variabili è uguale a 0.20 quanto vale il VIF_1 ?
- **13.25** Tornando all'Esercizio 13.3 calcolate il VIF per ciascuna variabile esplicativa. Si può sospettare la presenza di multicollinearità tra le variabili?
- **13.26** Tornando all'Esercizio 13.4 calcolate il VIF per ciascuna variabile esplicativa. Si può sospettare la presenza di multicollinearità tra le variabili?

13.9

COSTRUZIONE DEL MODELLO

In questo paragrafo proseguiamo lo studio del modello lineare trattando i metodi di costruzione del modello stesso che consentono di pervenire a modelli di regressione in cui è preso in considerazione solo un sottoinsieme di variabili esplicative.

Di seguito ci riferiremo a un esempio in cui si prendono in considerazione quattro variabili esplicative (staff totale, ore di lavoro a distanza, ore Dubner, numero totale di ore di lavoro) per spiegare la variabile numero totale di ore di attesa di grafici iscritti a un sindacato. La Tabella 13.7 riporta i dati rilevati.

Tabella 13.7 *Previsione del numero di ore di attesa basato su staff totale, ore di lavoro a distanza, ore Dubner e numero totale di ore di lavoro*

WEEK	STANDBY HOURS	TOTAL STAFF PRESENT	REMOTE HOURS	DUBNER HOURS	TOTAL LABOR HOURS
1	245	338	414	323	2,001
2	177	333	598	340	2,030
3	271	358	656	340	2,226
4	211	372	631	352	2,154
5	196	339	528	380	2,078
6	135	289	409	339	2,080

(continua)



DATASET
STANDBY

Tabella 13.7 *Previsione del numero di ore di attesa basato su staff totale, ore di lavoro a distanza, ore Dubner e numero totale di ore di lavoro (seguito)*

7	195	334	382	331	2073
8	118	293	399	311	1758
9	116	325	343	328	1624
10	147	311	338	353	1889
11	154	304	353	518	1988
12	146	312	289	440	2049
13	115	283	388	276	1796
14	161	307	402	207	1720
15	274	322	151	287	2056
16	245	335	228	290	1890
17	201	350	271	355	2187
18	183	339	440	300	2032
19	237	327	475	284	1856
20	175	328	347	337	2068
21	152	319	449	279	1813
22	188	325	336	244	1808
23	188	322	267	253	1834
24	197	317	235	272	1973
25	261	315	164	223	1839
26	232	331	270	272	1935

Nella costruzione di un modello il criterio principale da seguire è la *parsimonia*. Tale criterio impone di inserire in un modello il numero minimo di variabili indipendenti che consentano di spiegare la variabile risposta. I modelli di regressione con poche variabili esplicative sono di più facile interpretazione, soprattutto perché sono meno esposti al rischio di multicollinearità tra le variabili esplicative.

Inoltre è opportuno sottolineare che la selezione di un modello di regressione appropriato in presenza di un elevato numero di variabili esplicative, risulta in genere più difficoltosa. Innanzitutto la considerazione di tutti i modelli possibili risulta più complessa dal punto di vista computazionale. Inoltre, sebbene modelli concorrenti possano essere valutati quantitativamente, è possibile che non vi sia un modello migliore, ma piuttosto più modelli *egualmente adeguati*.

Tornando al data set STANDBY cominciamo col valutare se vi sia o meno multicollinearità tra le variabili esplicative, mediante il calcolo del *VIF* per ciascuna variabile esplicativa. Nella Figura 13.15 riportiamo l'output di Excel con i *VIF* e la stima del modello di regressione.

Osserviamo che i valori del *VIF* sono tutti bassi: il valore più alto è 2.0 relativo al numero totale di ore e il più basso è 1.2, relativo al numero di ore di lavoro a distanza. Pertanto sulla base del criterio di Snee (riferimento bibliografico 6) non vi è prova della presenza di multicollinearità tra le variabili esplicative.

	A	B
1	Model for X1 and all other X	
2		
3	Regression Statistics	
4	Multiple R	0.643681
5	R Square	0.414325
6	Adjusted R Square	0.334461
7	Standard Error	16.47151
8	Observations	26
9	VIF	1.707433

RIQUADRO A

	A	B
1	Model for X2 and all other X	
2		
3	Regression Statistics	
4	Multiple R	0.434898
5	R Square	0.189136
6	Adjusted R Square	0.078564
7	Standard Error	124.9392
8	Observations	26
9	VIF	1.233253

RIQUADRO B

	A	B
1	Model for X3 and all other X	
2		
3	Regression Statistics	
4	Multiple R	0.560992
5	R Square	0.314712
6	Adjusted R Square	0.221263
7	Standard Error	57.55254
8	Observations	26
9	VIF	1.45924

RIQUADRO C

	A	B
1	Model for X4 and all other X	
2		
3	Regression Statistics	
4	Multiple R	0.70698
5	R Square	0.49982
6	Adjusted R Square	0.431614
7	Standard Error	114.4118
8	Observations	26
9	VIF	1.999281

RIQUADRO D

FIGURA 13.15

Output di Microsoft Excel per il modello di regressione per la previsione delle ore di stand-by sulla base di quattro variabili esplicative.

	A	B	C	D	E	F	G
1	Regression Analysis for Standby Hours						
2							
3	Regression Statistics						
4	Multiple R	0.78935216					
5	R Square	0.623076833					
6	Adjusted R Square	0.551281944					
7	Standard Error	31.83500743					
8	Observations	26					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	4	35181.79373	8795.448432	8.678568098	0.000268015	
13	Residual	21	21282.82166	1013.467698			
14	Total	25	56464.61538				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	330.8318447	110.8953572	-2.983279491	0.007087351	561.4514047	-100.2122847
18	Total Staff	1.245629161	0.412059739	3.022933435	0.006473169	0.388703875	2.102554447
19	Remote	-0.118417979	0.054324392	-2.179830739	0.040795357	-0.231391756	-0.005444202
20	Dubner	-0.297058588	0.1179313	-2.518912186	0.019945311	-0.542310195	-0.051806982
21	Total Labor	0.130534912	0.059322942	2.200411993	0.039106915	0.00716608	0.253903744

RIQUADRO E

L'approccio stepwise alla costruzione del modello di regressione

Intendiamo ora individuare quell'insieme di variabili esplicative che consente di costruire un modello adeguato senza dover ricorrere all'uso di tutte le variabili considerate. Cominciamo con l'introdurre un metodo di selezione del modello ampiamente usato che prende il nome di **regressione stepwise** e che mira all'individuazione del modello "migliore" senza dover tuttavia considerare tutti i modelli possibili. Una volta individuato il modello migliore, si conduce un'analisi dei residui per valutarne l'adeguatezza.

Nel Paragrafo 13.5, abbiamo introdotto il test F parziale per valutare le proporzioni in un modello di regressione multiplo. Con il metodo *stepwise* si estende il test F parziale al caso di un modello di regressione con più di due variabili esplicative. Una peculiarità di tale procedimento è che una variabile introdotta a un certo punto nel modello può essere

successivamente eliminata. Nella regressione *stepwise* infatti a ciascuno stadio del processo di costruzione del modello le variabili esplicative possono essere sia aggiunte che eliminate dal modello stesso. Il processo di selezione termina quando nessuna altra variabile può essere eliminata e nessuna può essere aggiunta.

L'origine di tale approccio alla costruzione del modello risale a più di 30 anni fa, a un'epoca in cui l'analisi di regressione veniva condotta con il ricorso a mainframe, cosa che comportava lunghi e quindi costosi tempi di elaborazione. In un contesto di questo genere possiamo capire perché una procedura come quella *stepwise* che conduce alla valutazione di un numero limitato di modelli alternativi abbia avuto un'ampia diffusione. Ora, invece, grazie al ricorso a computer con processori velocissimi, la valutazione di molti modelli alternativi può essere condotta in poco tempo e con costi limitati. Per questo motivo si è assistito allo diffusione di approcci alternativi alla costruzione di un modello di regressione, tra cui l'approccio *Best-Subsets*.

L'approccio Best-Subsets

Con questo approccio possiamo valutare tutti i possibili modelli di regressione dato un insieme di variabili esplicative o i sottoinsiemi migliori dei modelli con dato numero di variabili indipendenti.

Nella Figura 13.17 riportiamo l'output ottenuto con l'aggiunta PHStat, contenente tutti modelli di regressione che è possibile costruire.

I modelli di regressione che si possono ottenere per un dato insieme di variabili esplicative possono essere valutati e quindi confrontati facendo ricorso a criteri diversi.

Il primo criterio utilizzabile è quello dell' r^2 corretto, con cui l'indice di determinazione r^2 viene corretto tenendo conto del numero di variabili esplicative inserite nel modello e dell'ampiezza del campione. Risulta utile ricorrere a tale misura dal momento che intendiamo porre a confronto modelli aventi un diverso numero di variabili esplicative.

Con riferimento alla Figura 13.17, osserviamo che l' r^2 corretto raggiunge valore massimo 0.551 quando tutte le variabili esplicative e l'intercetta sono incluse nel modello.

FIGURA 13.16

Output dell'aggiunta PHStat relativo alla regressione Best-Subsets per i dati del numero di ore di attesa.

	A	B	C	D	E	F	G
1	Best Subsets Analysis for Standby Hours Models						
2							
3	R ² T	0.623077					
4	1 - R ² T	0.376923					
5	n	26					
6	T	5					
7	n - T	21					
8							Consider
9		Cp	p+1	R Square	Adj. R Square	Std. Error	This Model?
10	X1	13.32152	2	0.366024	0.339608215	38.6206	No
11	X1X2	8.41933	3	0.489909	0.445553638	35.38734	No
12	X1X2X3	7.841813	4	0.536172	0.47292327	34.50286	No
13	X1X2X3X4	5	5	0.623077	0.551281944	31.83501	Yes
14	X1X2X4	9.344919	4	0.509194	0.442265513	35.49212	No
15	X1X3	10.64856	3	0.449898	0.402062546	36.74905	No
16	X1X3X4	7.751662	4	0.537791	0.474762012	34.44263	No
17	X1X4	14.79818	3	0.375417	0.321105605	39.15789	No
18	X2	33.20781	2	0.00909	-0.032197524	48.28359	No
19	X2X3	32.30673	3	0.061161	-0.020477044	48.00868	No
20	X2X3X4	12.13813	4	0.459059	0.385294475	37.26076	No
21	X2X4	23.24809	3	0.223752	0.156252135	43.65405	No
22	X3	30.38835	2	0.059696	0.020516668	47.03452	No
23	X3X4	11.82309	3	0.428816	0.379148026	37.44658	No
24	X4	24.1846	2	0.171045	0.136505715	44.16192	No

Un secondo criterio spesso utilizzato per confrontare diversi modelli di regressione si basa su una statistica sviluppata da Mallows (riferimento bibliografico 5). Questa statistica, chiamata C_p , misura la differenza tra il modello di regressione stimato e il modello vero. La statistica C_p è definita come segue

La statistica C_p

$$C_p = \frac{(1 - R_p^2)(n - T)}{1 - R_T^2} - [n - 2(p + 1)] \quad (13.18)$$

dove

p = numero di variabili esplicative inserite nel modello di regressione

T = numero totale di parametri (inclusa l'intercetta) da stimare nel modello di regressione completo

R_p^2 = coefficiente di regressione multipla per un modello di regressione contenente p variabili esplicative

R_T^2 = coefficiente di regressione multipla per il modello di regressione completo.

In base all'equazione (13.18) calcoliamo in valore della statistica C_p per il modello contenente come variabili esplicative lo staff presente e il numero di ore di lavoro a distanza:

$$n = 26 \quad p = 2 \quad T = 4 + 1 = 5 \quad R_p^2 = 0.490 \quad R_T^2 = 0.623$$

di modo che

$$C_p = \frac{(1 - 0.49)(26 - 5)}{1 - 0.623} - [26 - 2(2 + 1)]$$

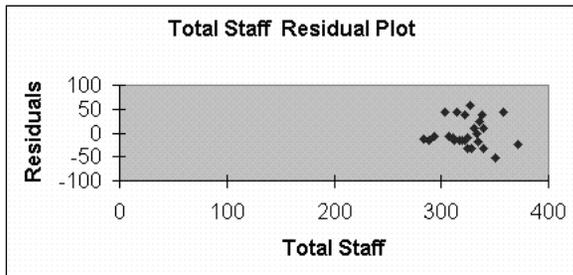
$$C_p = 8.42$$

Se un modello di regressione con p variabili esplicative differisce dal modello vero solo per gli errori casuali, il valore medio della statistica C_p è $(p + 1)$, il numero dei parametri. Pertanto, si tratta di individuare quei modelli con un valore di C_p minore o uguale a $p + 1$.

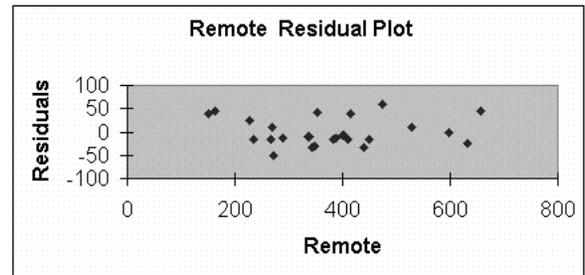
Dalla Figura 13.17 ricaviamo che solo il modello completo, contenente vale a dire tutte e quattro le variabili indipendenti ha un valore di C_p minore o uguale di $p + 1$. Pertanto dovremmo scegliere questo modello. Spesso, la statistica C_p ci porta a individuare diversi modelli, che quindi devono essere confrontati sulla base di altri criteri, quali il criterio della parsimonia, della facilità di interpretazione, della valutazione delle differenze dalle assunzioni su cui si fonda il modello stesso (analisi dei residui). Osserviamo anche che il modello selezionato con l'approccio stepwise ha un valore di C_p uguale a 8.4, sostanzialmente superiore a $p + 1 = 3$.

Una volta individuate le variabili da inserire nel modello, dobbiamo procedere all'analisi dei residui per valutare la bontà del modello così costruito. Nella Figura 13.18 è riportato l'output parziale ottenuto con l'uso di Excel.

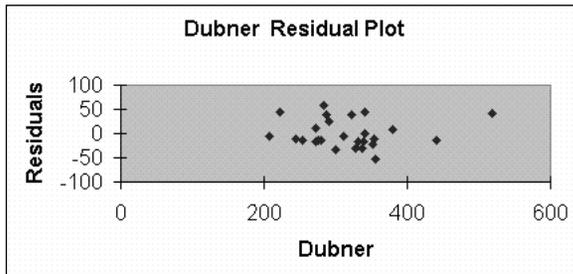
Osserviamo che i grafici dei residui rispetto alla variabile staff totale, rispetto alla variabile ore di lavoro a distanza, rispetto alla variabile ore Dubner e rispetto alla variabile ore totali di lavoro non rivelano alcuna struttura particolare. Inoltre l'istogramma dei residui (che qui non viene mostrato) mostra solo un leggero scostamento dal caso di normalità.



RIQUADRO A



RIQUADRO B



RIQUADRO C



RIQUADRO D

FIGURA 13.18 Grafici dei residui ottenuti con Excel per i dati relativi alle ore di standby.

Pertanto in base alla Figura 13.15 ricaviamo l'espressione della stima del modello di regressione:

$$\hat{Y}_i = -330.83 + 1.2456X_{1i} - 0.1184X_{2i} - 0.2917X_{3i} + 0.1305X_{4i}$$

In base a questo modello, possiamo concludere che mantenendo costante l'effetto del numero di ore di lavoro a distanza e del numero totale di ore di lavoro, se allo staff totale si aggiunge una persona il numero totale di ore di standby dovrebbe aumentare di 1.2456 ore. Mantenendo costante lo staff totale, le ore Dubner e il numero totale di ore di lavoro, prevediamo che per ogni ora in più di lavoro a distanza, il numero di ore di standby dovrebbe aumentare di 0.2971 ore. Mantenendo costante lo staff totale, le ore di lavoro a distanza e le ore Dubner, prevediamo che per ogni ora di lavoro totale in più, il numero di ore di standby dovrebbe diminuire di 0.1305 ore.

Nell'Esempio 13.6, prendiamo in considerazione il caso di più di un modello con C_p minore o uguale a $(p + 1)$.

Esempio 13.6 *Scelta tra modelli di regressione alternativi*

Prendete in considerazione il seguente output ottenuto ricorrendo all'approccio *Best-Subsets*. Quale modello scegliereste come migliore?

VARIABLES	r^2	ADJ. r^2	C_p	
1	12.1	11.9	113.9	X_4
1	9.3	9.0	130.4	X_1
1	8.3	8.0	136.2	X_3
2	21.4	21.0	62.1	X_3X_4
2	19.1	18.6	75.6	X_1X_3

VARIABLES	r^2	ADJ. r^2	C_p	
2	18.1	17.7	81.0	$X_1 X_4$
3	28.5	28.0	22.6	$X_1 X_3 X_4$
3	26.8	26.3	32.4	$X_3 X_4 X_5$
3	24.0	23.4	49.0	$X_2 X_3 X_4$
4	30.8	30.1	11.3	$X_1 X_2 X_3 X_4$
4	30.4	29.7	14.0	$X_1 X_3 X_4 X_6$
4	29.6	28.9	18.3	$X_1 X_3 X_4 X_5$
5	31.7	30.8	8.2	$X_1 X_2 X_3 X_4 X_5$
5	31.5	30.6	9.6	$X_1 X_2 X_3 X_4 X_6$
5	31.3	30.4	10.7	$X_1 X_3 X_4 X_5 X_6$
6	32.3	31.3	6.8	$X_1 X_2 X_3 X_4 X_5 X_6$
6	31.9	30.9	9.0	$X_1 X_2 X_3 X_4 X_5 X_7$
6	31.7	30.6	10.4	$X_1 X_2 X_3 X_4 X_6 X_7$
7	32.4	31.2	8.0	$X_1 X_2 X_3 X_4 X_5 X_6 X_7$

Output parziale di una regressione *best-subsets*.

SOLUZIONE

Cominciamo con l'individuare i modelli per i quali la statistica C_p assume un valore minore o uguale a $(p + 1)$. Due modelli soddisfano tale criterio: il modello con sei variabili esplicative ($X_1 X_2 X_3 X_4 X_5 X_6$) per il quale la statistica C_p assume valore 6.8, che è minore di $p + 1 = 6 + 1 = 7$, e il modello completo con variabili esplicative ($X_1 X_2 X_3 X_4 X_5 X_6 X_7$) per il quale C_p assume valore 8.0. Un modo per effettuare una scelta tra modelli che soddisfano le medesime condizioni consiste nello stabilire se tali modelli hanno delle variabili in comune e quindi nel valutare il contributo delle altre variabili. Nel caso in considerazione i due modelli differiscono per la sola variabile X_7 e quindi non resta che stabilire se tale variabile dia un contributo significativo al modello in cui siano già state incluse le variabili $X_1 X_2 X_3 X_4 X_5 X_6$ dovrebbe essere inclusa nel modello solo se il suo contributo è statisticamente significativo.

Nel riquadro 13.2 sintetizziamo i passi principali da seguire nella costruzione del modello.



Riquadro 13.2 La costruzione del modello

- ✓ **1.** Scegliete un insieme di potenziali variabili esplicative.
- ✓ **2.** Stimate un modello contenente tutte le variabili individuate e calcolate il *variance inflationary factor (VIF)* per ciascuna variabile esplicative.
- ✓ **3.** Stabilite se vi sono variabili con $VIF > 5$.
- ✓ **4.** Tre sono i possibili risultati:
 - (a)** Nessuna delle variabili esplicative ha $VIF > 5$. In questo caso passate al punto 5.
 - (b)** Una delle variabili ha esplicative $VIF > 5$. In questo caso eliminate questa variabile e passate al punto 5.

(continua)

Riquadro 13.2 La costruzione del modello (seguito)

- (c) Più di una variabile esplicativa ha $VIF > 5$. In questo caso, eliminate le variabili indipendenti che hanno VIF più alto e tornate al punto 2.
- ✓ 5. Conducete un'analisi di tipo *best subsets* o *all subsets* sulle variabili restanti per individuare i modelli migliori (in termini di C_p).
 - ✓ 6. Elencate i modelli con $C_p \leq (p + 1)$.
 - ✓ 7. Tra i modelli individuati al punto 6 scegliete il modello migliore (come discusso nell'Esempio 13.6)
 - ✓ 8. Conducete un'analisi completa del modello individuato, inclusa l'analisi dei residui.
 - ✓ 9. Sulla base dell'analisi dei residui, aggiungete se necessario una componente non lineare e analizzate di nuovo il modello.
 - ✓ 10. Usate il modello selezionato per effettuare delle previsioni.

Di seguito riportiamo la tabella riassuntiva dei passaggi principali della costruzione del modello.

Esercizi del paragrafo 13.9

- 13.27 Supponete che per la costruzione di un modello di regressione si prendono in considerazione sei variabili indipendenti. In base a un campione di 40 osservazioni si ottengono i seguenti risultati:

$$n = 40 \quad p = 2 \quad T = 6 + 1 = 7 \quad R_p^2 = 0.274 \quad R_T^2 = 0.653$$

- (a) Calcolate il valore della statistica C_p per questo modello con due variabili indipendenti.
- (b) Sulla base dei risultati al punto (a) il modello considerato soddisfa i criteri per poter essere considerato il modello migliore?

- 13.28 Intendete costruire un modello di regressione per prevedere il prezzo di vendita delle case di una determinata città sulla base del valore accertato, del periodo di tempo in cui la casa è stata venduta e sulla base di una variabile dummy che indica se la casa è nuova o meno (0 = no, 1 = sì). Si estrae un campione di 30 case per studiare la relazione tra il prezzo di vendita e il valore accertato. I risultati sono riportati nella tabella seguente.



DATASET
HOUSE1

OSSERVAZIONE	VALORE ACCERTATO	PREZZO DI VENDITA	TEMPO	NUOVA	OSSERVAZIONE	VALORE ACCERTATO	PREZZO DI VENDITA	TEMPO	NUOVA
	(\$ 000)	(\$ 000)				(\$ 000)	(\$ 000)		
1	78.17	94.10	10	1	16	84.36	106.70	12	0
2	80.24	101.90	10	1	17	72.94	81.50	5	0
3	74.03	88.65	11	0	18	76.50	94.50	14	1
4	86.31	115.50	2	0	19	66.28	69.00	1	0
5	75.22	87.50	5	0	20	79.74	96.90	3	1
6	65.54	72.00	4	0	21	72.78	86.50	14	0
7	72.43	91.50	17	0	22	77.90	97.90	12	1
8	85.61	113.90	13	0	23	74.31	83.00	11	0
9	60.80	69.34	6	0	24	79.85	97.30	12	1
10	81.88	96.90	5	1	25	84.78	100.80	2	1
11	79.11	96.00	7	0	26	81.61	97.90	6	1
12	59.93	61.90	4	0	27	74.92	90.50	12	0
13	75.27	93.00	11	0	28	79.98	97.00	4	1
14	85.88	109.50	10	1	29	77.96	92.00	9	0
15	76.64	93.75	17	0	30	79.07	95.90	12	1

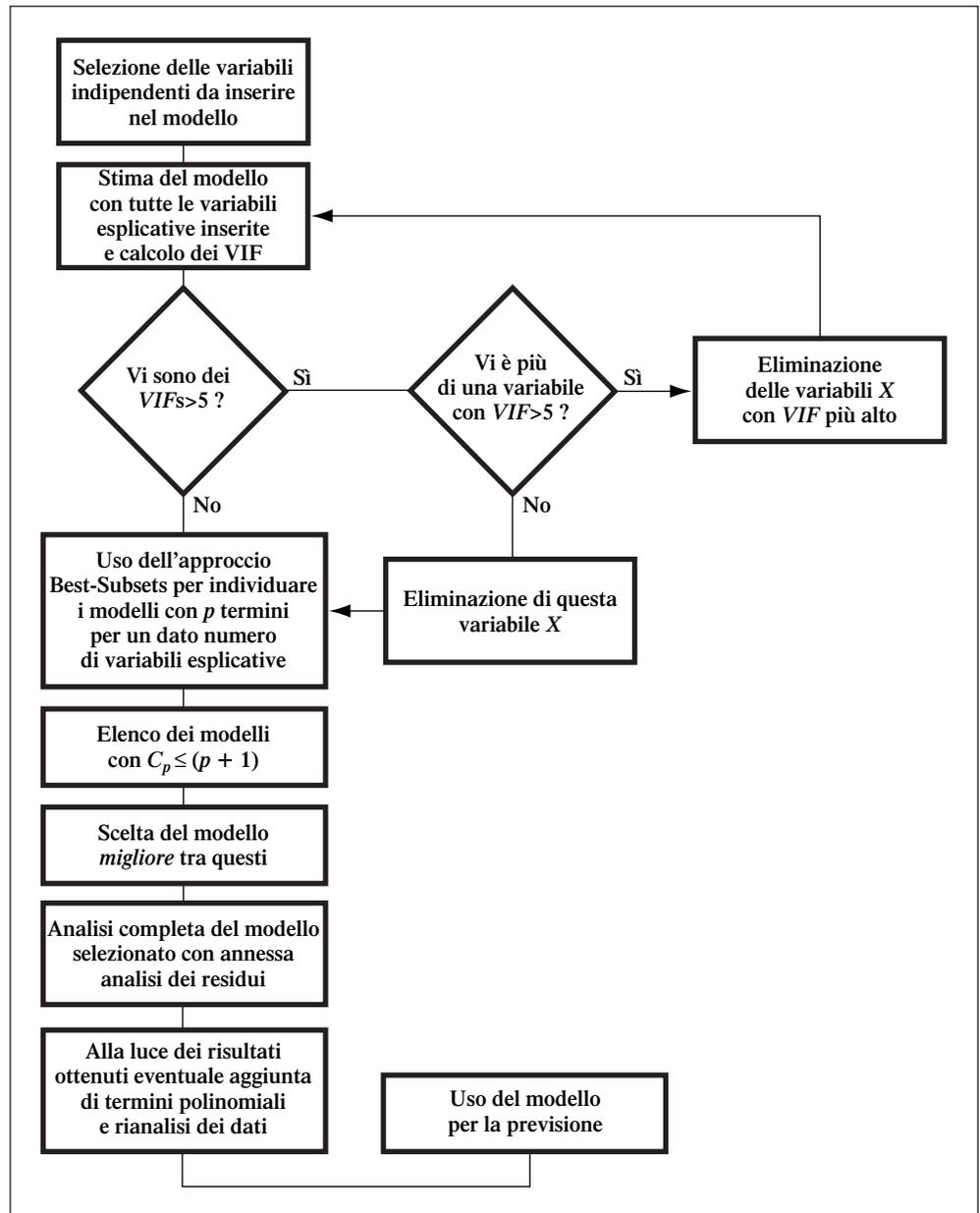


Tabella riassuntiva del capitolo 13.

Costruite un appropriato modello di regressione per prevedere il prezzo di vendita. Conducete un'estesa analisi dei residui del modello selezionato e motivate le vostre conclusioni.

- 13.29** Il file AUTO96 contiene dati relativi a 89 modelli di automobile. Tra le variabili esplicative considerate vi sono il peso, la larghezza, la lunghezza di ciascuna automobile e se l'automobile è a trazione anteriore o posteriore. Costruite un appropriato modello di regressione per prevedere il prezzo di vendita. Conducete un'estesa analisi dei residui del modello selezionato e motivate le vostre risposte.



LE TRAPPOLE DELL'ANALISI DI REGRESSIONE

Nel capitolo precedente abbiamo messo in evidenza le difficoltà di cui si deve tener conto nell'impiego del modello di regressione. Ora che abbiamo preso in considerazione modelli di regressione differenti siamo in grado di porre in evidenza le precauzioni ulteriori a cui si deve ricorrere nella costruzione di un modello di regressione lineare.

Riquadro 13.3 Le ulteriori trappole del modello di regressione

- ✓ **1.** Il valore del coefficiente di regressione corrispondente ad una data variabile esplicativa deve essere interpretato supponendo costanti i valori dei coefficienti relativi a tutte le altre variabili.
- ✓ **2.** Si devono valutare i grafici dei residui relativi a ciascuna delle variabili esplicative.
- ✓ **3.** Si devono valutare i termini di interazione per stabilire se l'inclinazione della variabile dipendente rispetto a ciascuna variabile esplicativa è la stessa per un dato valore della variabile dummy.
- ✓ **4.** Si deve calcolare il *VIF* per ciascuna variabile esplicativa prima di stabilire quale variabile includere nel modello.
- ✓ **5.** I possibili modelli di regressione vanno valutati usando l'approccio *best-subsets* alla selezione del modello.



RIEPILOGO DEL CAPITOLO

In questo capitolo, come illustrato dalla tabella riassuntiva del capitolo, abbiamo introdotto il modello di regressione multipla con due variabili esplicative e abbiamo illustrato i metodi per verificare la significatività del modello completo e la significatività del contributo di ciascuna variabile esplicativa. Abbiamo inoltre definito i coefficienti di determinazione parziali e abbiamo studiato il modello di regressione quadratica, le variabili dummy, la multicollinearità e i metodi di costruzione dei modelli di regressione.



PAROLE CHIAVE

approccio Best-Subsets 42
 coefficiente di determinazione parziale 21
 coefficiente di determinazione 6
 coefficiente netto di regressione 4

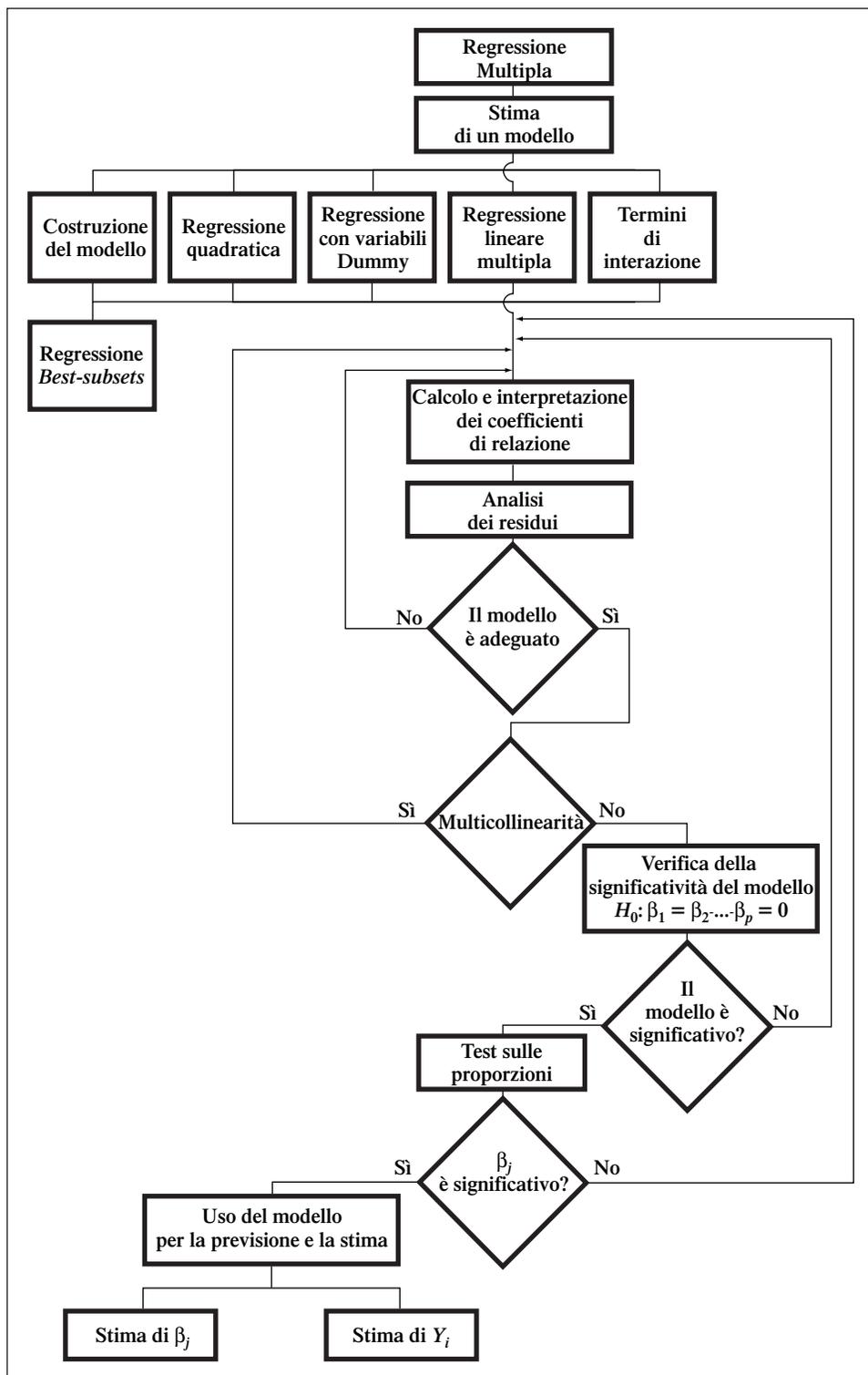
corretto r^2 7
 criterio del test F parziale 17
 modello di regressione quadratica 23
 multicollinearità 38
 regressione multipla 2

regressione stepwise 40
 statistica C_p 43
 termine di interazione 33
 variabili dummy 31
 variance inflationary factor (*VIF*) 38

Verifica della comprensione

- 13.30** In che cosa l'interpretazione dei coefficienti di regressione nel caso del modello di regressione multipla differisce dall'interpretazione che se ne dà nel caso del modello lineare semplice?
- 13.31** In che cosa la verifica della significatività dell'intero di modello di regressione multipla differisce dalla verifica della significatività del contributo di ciascuna variabile esplicativa.
- 13.32** In che cosa i coefficienti di determinazione parziali differiscono dal coefficiente di determinazione?
- 13.33** Come e perché le variabili dummy possono essere impiegate?

Tabella riassuntiva del capitolo.



- 13.34** In che modo possiamo stabilire se l'inclinazione della variabile dipendente rispetto a una variabile esplicativa è uguale per tutti i valori della variabile dummy?
- 13.35** In che modo possiamo stabilire se le variabili dipendenti sono non correlate?
- 13.36** In quale caso potremmo decidere di includere una variabile dummy in un modello di regressione?
- 13.37** Che assunzione concernente l'inclinazione della variabile indipendente Y rispetto alla variabile esplicativa X deve essere fatta quando si include una variabile dummy nel modello di regressione?
- 13.38** Quale differenza intercorre tra l'approccio *stepwise* alla costruzione del modello e l'approccio *best-subsets*?
- 13.39** In che modo possiamo effettuare una scelta tra i modelli selezionati in base al valore della statistica C_p nell'ambito dell'approccio *best-subsets* alla costruzione del modello?

Esercizi di riepilogo del capitolo



DATASET
HOUSE2

- **13.40** Intendete sviluppare un modello di regressione multipla per prevedere il valore delle case monofamiliari di una data città sulla base della superficie riscaldata e dell'età. A tale scopo si estrae un campione di 15 case e i dati relativi al valore (in migliaia di dollari), la superficie riscaldata (in migliaia di piedi al quadrato) e l'età delle case (in anni) sono riportati nella tabella seguente:

CASA	SUPERFICIE RISCALDATA (IN MIGLIAIA DI PIEDI)			CASA	SUPERFICIE RISCALDATA (IN MIGLIAIA DI PIEDI)		
	VALORE ACCERTATO (\$ 000)	AL QUADRATO)	ETÀ (IN ANNI)		VALORE ACCERTATO (\$ 000)	AL QUADRATO)	ETÀ (IN ANNI)
1	84.4	2.00	3.42	9	78.5	1.59	1.75
2	77.4	1.71	11.50	10	79.2	1.50	2.75
3	75.7	1.45	8.33	11	86.7	1.90	0.00
4	85.9	1.76	0.00	12	79.3	1.39	0.00
5	79.1	1.93	7.42	13	74.5	1.54	12.58
6	70.4	1.20	32.00	14	83.8	1.89	2.75
7	75.8	1.55	16.00	15	76.8	1.59	7.17
8	85.9	1.93	2.00				

- Formulate l'espressione del modello di regressione multipla.
- Fornite un'interpretazione delle inclinazioni.
- Prevedete il valore accertato di una casa con superficie riscaldata pari a 1750 piedi al quadrato e 10 anni di età.
- Conducete un'analisi dei residui e valutate l'adeguatezza del modello.
- Per un livello di significatività pari a 0.05 stabilite se vi sia una relazione significativa tra il valore delle case e le due variabili esplicative (area riscaldata e età).
- Con riferimento al punto (e) calcolate il valore del p -value e interpretatene il significato.
- Interpretate il significato del coefficiente di determinazione $r_{Y,12}^2$.
- Calcolate l' r^2 corretto.
- Per un livello di significatività pari a 0.05 valutate il contributo di ciascuna variabile esplicativa al modello di regressione. Sulla base dei risultati ottenuti indicate quale modello andrebbe usato.
- Con riferimento al punto (e) calcolate il valore del p -value e interpretatene il significato.
- Costruite un intervallo di confidenza del 95% per l'inclinazione del valore delle case rispetto alla superficie riscaldata. In che modo l'interpretazione dell'inclinazione differisce da quella data nell'Esercizio 12.66?
- Calcolate i coefficienti di determinazione parziali ($r_{Y1,2}^2$ e $r_{Y2,1}^2$), e spiegate il significato.



DATASET
UNIV&COL

- (m) Il perito immobiliare sostiene che il valore delle case non è influenzato dalla loro età. Alla luce dei risultati dei punti (a)-(l) ritenete corretta questa affermazione?

- 13.41** Il file UNIV&COL contiene dei dati su 80 college e università statunitensi. Tra le variabili incluse vi sono il costo annuale (in migliaia di dollari), il punteggio medio nel test SAT (*Scholastic Aptitude Test*) e le spese vitto e alloggio (in migliaia di dollari). Intendete costruire il modello di regressione per prevedere il costo annuale totale sulla base del punteggio al SAT e delle spese di vitto e alloggio.
- (a) Formulate l'espressione del modello di regressione multipla.
 - (b) Fornite un'interpretazione delle inclinazioni.
 - (c) Prevedete il costo totale per un'università che ha conseguito un punteggio SAT pari a 1100 e in cui l'ammontare delle spese di vitto e alloggio è pari a \$5000.
 - (d) Conducete un'analisi dei residui e valutate l'adeguatezza del modello.
 - (e) Per un livello di significatività pari a 0.05 stabilite se vi sia una relazione significativa tra il costo totale e le due variabili esplicative (punteggio SAT e spese di vitto e alloggio).
 - (f) Con riferimento al punto (e) calcolate il valore del p -value e interpretatene il significato.
 - (g) Interpretate il significato del coefficiente di determinazione $r_{Y,12}^2$.
 - (h) Calcolate l' r^2 corretto.
 - (i) Per un livello di significatività pari a 0.05 valutate il contributo di ciascuna variabile esplicative. Sulla base dei risultati ottenuti indicate quale modello andrebbe usato.
 - (j) Con riferimento al punto (e) calcolate il valore del p -value e interpretatene il significato.
 - (k) Costruite un intervallo di confidenza del 95% per l'inclinazione del costo totale rispetto al punteggio SAT.
 - (l) Calcolate i coefficienti di determinazione parziali ($r_{Y1,2}^2$ e $r_{Y2,1}^2$), e spiegate il significato.
 - (m) Spiegate perché l'inclinazione del costo totale rispetto alle spese di vitto e alloggio sembra differire in maniera consistente da 1.0.
 - (n) Quali fattori non inclusi nel modello potrebbero spiegare la forte relazione positiva tra i costi e il punteggio SAT?



DATASET
AUTO96

- **13.42** Il file AUTO96 contiene dati relativi a 89 modelli di automobile rilevati a partire dal 1996. Tra le variabili esplicative considerate vi sono la distanza percorsa per litro di benzina (in pollici) e il peso dell'automobile (in libbre). Intendete sviluppare un modello di regressione per prevedere la distanza percorsa sulla base della lunghezza e del peso delle automobili.
- (a) Formulate l'espressione del modello di regressione multipla.
 - (b) Fornite un'interpretazione delle inclinazioni.
 - (c) Prevedete la distanza percorsa da un'automobile lunga 195 pollici e con il peso pari a 3000 libbre.
 - (d) Conducete un'analisi dei residui e valutate l'adeguatezza del modello.
 - (e) Per un livello di significatività pari a 0.05 stabilite se vi sia una relazione significativa tra la distanza percorsa e le due variabili esplicative (lunghezza e peso).
 - (f) Con riferimento al punto (e) calcolate il valore del p -value e interpretatene il significato.
 - (g) Interpretate il significato del coefficiente di determinazione $r_{Y,12}^2$.
 - (h) Calcolate l' r^2 corretto.
 - (i) Per un livello di significatività pari a 0.05 valutate il contributo di ciascuna variabile esplicative. Sulla base dei risultati ottenuti indicate quale modello andrebbe usato.
 - (j) Con riferimento al punto (e) calcolate il valore del p -value e interpretatene il significato.
 - (k) Costruite un intervallo di confidenza del 95% la distanza percorsa rispetto al peso.
 - (l) Calcolate i coefficienti di determinazione parziali ($r_{Y1,2}^2$ e $r_{Y2,1}^2$), e spiegate il significato.



DATASET
UNIV&COL

- 13.43** Il file UNIV&COL contiene dei dati su 80 college e università statunitensi. Tra le variabili incluse vi sono il tipo di calendario accademico (1 = semestralità, 0 = altro), il costo annuale (in migliaia di dollari), il punteggio medio nel test SAT (*Scholastic Aptitude Test*), le spese

vitto e alloggio (in migliaia di dollari), il tipo di istituzione (pubblica o privata), una variabile che indica se il punteggio al TOEFL è in media almeno uguale a 550 e l'indebitamento medio degli studenti al momento del diploma.

Costruite il modello di regressione più appropriato per prevedere l'ammontare dell'indebitamento al momento del diploma. Conducete un'estesa analisi dei residui del modello selezionato e motivate le vostre risposte.

- 13.44** Il data set HOME contiene dei dati relativi alle case monofamiliari vendute nella città di Cincinnati nell'arco di un anno. Le variabili comprese sono:

Price: il prezzo di vendita in dollari

Location: indice di posizione della casa da 1 a 5, 1 indica la posizione migliore e 5 la peggiore

Condition: indice di stato della casa da 1 a 5, 1 indica pessimo stato e 5 ottimo stato

Bedrooms: numero di stanze da letto

Bathrooms: numero di bagni

Other Rooms: numero di locali esclusi i bagni e le camere da letto

Fonte: Dati reali raccolti dagli autori.

Costruite un modello di regressione assumendo il prezzo come variabile dipendente e le restanti cinque variabili come variabili esplicative.

- Formulate l'espressione del modello di regressione multipla.
- Fornite un'interpretazione delle inclinazioni.
- Per un livello di significatività pari a 0.05 valutate il contributo di ciascuna variabile esplicativa.
- Con riferimento al punto (c) calcolate il valore del p -value e interpretatene il significato.
- Prevedere il prezzo di una casa con Bedrooms = 3, Bathrooms = 2.5, Other Rooms = 4, Location = 4 e Condition = 4.
- Interpretate il significato del coefficiente di determinazione $r_{Y.12345}^2$.
- Calcolate l' r^2 corretto.
- Conducete un'analisi dei residui e stabilite se il modello è adeguato.
- Eliminate le variabili che non contribuiscono in maniera significativa al modello in base all'approccio *best-subsets*. Quale modello ritenete migliore? Giustificate la vostra risposta.



DATASET
OMNI

- 13.45** Prendete in considerazione i dati relativi alle vendite della barretta energetica OMNI riportati nella Tabella 13.1. Nella Figura 13.1 si riportano i risultati del modello di regressione multipla in cui si assumono come variabili esplicative il prezzo (X_1) e la spesa in attività promozionali (X_2). Costruite un ulteriore modello di regressione in cui si aggiungono come variabili esplicative i due termini quadratici $X_3 = X_1^2$ e $X_4 = X_2^2$.
- Formulate l'espressione del modello di regressione multipla.
 - Calcolate e interpretate il significato del coefficiente di determinazione $r_{Y.1234}^2$.
 - Calcolate l' r^2 corretto.
 - Per un livello di significatività pari a 0.05 stabilite se i due termini addizionali contribuiscono in maniera significativa al modello di regressione.
 - Confrontate il modello ottenuto con quello della Figura 13.1. Quale ritenete più appropriato?

IL CASO

UNA SOCIETÀ DI TRASLOCHI

Il proprietario di una società di traslochi ricorre da tempo ad un certo metodo statistico per stimare il numero di ore necessarie per effettuare un trasloco. Tale metodo si è rivelato utile nel passato, ma ora il proprietario vorrebbe disporre di un me-

todo di stima più accurato. A tale scopo ha provveduto alla raccolta di dati riguardanti 36 traslochi con un tempo di viaggio trascurabile rispetto al numero di ore di lavoro. I dati rilevati sono riportati nella tabella seguente.

ORE DI OSSERVAZIONE LAVORO LOCALI			SUPERFICIE TRASLOCATA (IN PIEDI AL QUADRATO)	ORE DI OSSERVAZIONE LAVORO LOCALI			SUPERFICIE TRASLOCATA (IN PIEDI AL QUADRATO)
1	24.00	3.5	545	19	25.00	3.0	557
2	13.50	2.0	400	20	45.00	5.5	1028
3	26.25	2.5	562	21	29.00	4.5	793
4	25.00	3.0	540	22	21.00	3.0	523
5	9.00	1.0	220	23	22.00	3.5	564
6	20.00	3.0	344	24	16.50	2.5	312
7	22.00	3.5	569	25	37.00	4.0	757
8	11.25	2.0	340	26	32.00	3.5	600
9	50.00	5.0	900	27	34.00	4.0	796
10	12.00	1.5	285	28	25.00	3.5	577
11	38.75	5.0	865	29	31.00	3.0	500
12	40.00	4.5	831	30	24.00	4.0	695
13	19.50	3.0	344	31	40.00	5.5	1054
14	18.00	2.5	360	32	27.00	3.0	486
15	28.00	4.0	750	33	18.00	3.0	442
16	27.00	3.5	650	34	62.50	5.5	1249
17	21.00	3.0	415	35	53.75	5.0	995
18	15.00	2.5	275	36	79.50	5.5	1397



Sviluppate un modello per prevedere le ore di lavoro necessarie per un trasloco sulla base del numero di locali e del numero di piedi al quadrato da

traslocare. Redigete una breve sintesi delle conclusioni a cui pervenite e allegate in appendice i risultati statistici ottenuti e le relative spiegazioni.

BIBLIOGRAFIA

- Hocking, R.R., "Developments in Linear Regression Methodology: 1959–1982," *Technometrics* 25 (1983): 219–250.
- Hosmer, D.W., e S. Lemeshow, *Applied Logistic Regression* (New York: Wiley, 1989).
- Marquardt, D.W., "You Should Standardize the Predictor Variables in Your Regression Models," discussion of "A Critique of Some Ridge Regression Methods," by G. Smith and F. Campbell, *Journal of the American Statistical Association* 75 (1980): 87–91.
- Microsoft Excel 2000* (Redmond, WA: Microsoft Corp., 1999).
- Neter, J., M. Kutner, C. Nachtsheim, e W. Wasserman, *Applied Linear Statistical Models, 4th ed.* (Homewood, IL: Irwin, 1996).
- Snee, R.D., "Some Aspects of Nonorthogonal Data Analysis, Part I. Developing Prediction Equations," *Journal of Quality Technology* 5 (1973): 67–79.

L'USO DI MICROSOFT EXCEL NEI MODELLI DI REGRESSIONE MULTIPLA

In questo paragrafo illustriamo come utilizzare Excel per analizzare un modello di regressione lineare multipla. In particolar modo ricorreremo all'opzione **Regressione** del menu **Strumenti di analisi** e all'aggiunta PHStat. Innanzitutto è opportuno sottolineare che *l'insieme X di tutte le variabili esplicative* deve essere inserito in colonne *consecutive*, perché lo strumento **Regressione** consente di specificare come input della X solo una serie consecutiva di valori inseriti nel foglio di lavoro.

Ritorniamo ai dati relativi alla barretta energetica OMNI, contenuti nel file **OMNI.xls**. Dopo aver aperto **OMNI.xls**, fate clic sull'etichetta del foglio **Data** (dati) e verificate che i dati relativi all'ammontare delle vendite, al prezzo e alla spesa in attività promozionali siano inseriti rispettivamente nelle colonne A, B e C. Selezionate **PHStat | Regression | Multiple Regression** (regressione | regressione multipla). Nella finestra di dialogo che si apre, digitate **A1:A35** nel riquadro **Y Variable Cell Range** (intervallo dei valori della Y) e **B1:C35** nel riquadro **X Variable Cell Range** (intervallo dei valori della X). Selezionate il comando **First cells in both ranges contain label** (la prima cella di entrambi gli intervalli di valori contiene l'etichetta) e digitate **95** nel riquadro **Confidence Lvl. for regression coefficients** (livello di confidenza per i coefficienti di regressione). Selezionate i comandi: **Statistics Table** (tabella delle statistiche), **Anova and Coefficients Table** (Anova e tabella dei coefficienti), **Residual Table** (tabella dei residui) e **Residual Plot** (grafico dei residui).

Digitate **Analisi di regressione per le vendite della barretta Omni Power** nel riquadro **Output Title** (Titolo dell'output). Selezionate i comandi **Coefficients of Partial Determination** (coefficienti di determinazione parziali), **Variance Inflationary Factor (VIF)** e **Confidence and Prediction Interval** (intervalli di confidenza e previsione). Digitate **95** nel riquadro **Confidence Level for int. estimate** (livello di confidenza degli intervalli). Fate clic su **OK**. Per ottenere gli intervalli di confidenza e la previsione corrispondenti a determinati valori delle variabili esplicative fate clic sul comando **Intervals** (intervalli). Per i dati relativi alla barretta Omni, per ottenere la stima delle vendite per un prezzo pari a **79** cents e un ammontare delle spese promozionali pari a \$ 400 inserite **79** nella cella B5 e **400** nella cella B6.

L'uso di Excel per i modelli di regressione quadratica

Per usare Excel nell'analisi dei modelli di regressione in cui una variabile esplicativa figura al quadrato ci limiteremo a usare la semplice formula $=A2^2$ per creare il quadrato della variabile esplicativa e l'analisi potrà essere condotta analogamente a quanto descritto nel paragrafo precedente. Torniamo al problema illustrato nel paragrafo 13.6 relativo all'elasticità delle vendite rispetto al prezzo. Aprite il file **DISPRAZ.XLS** e fate clic sull'etichetta del foglio **Data** (dati) per verificare che i dati relativi alle vendite e al prezzo siano inseriti rispettivamente nelle colonne A e B. Posizionate il cursore nella cella **B1** e selezionate i comandi **Inserisci | Colonne**. Digitate il nome della variabile **Prezzo²** nella cella **B1**. Nella cella **B2** inserite la formula $=A2^2$ e copiatela sino alla riga 16. Selezionate **PHStat | Regression | Multiple Regression** (regressione | regressione multipla). Nella finestra di dialogo che si apre, digitate **C1:C16** nel riquadro **Y Variable Cell Range** (intervallo dei valori della Y) e **A1:B16** nel riquadro **X Variable Cell Range** (intervallo dei valori della X). Selezionate il comando **First cells in both ranges contain label** (la prima cella di entrambi gli intervalli di valori contiene l'etichetta) e digitate **95** nel riquadro **Confidence Lvl. for regression coefficients** (livello di confidenza per i coefficienti di regressione). Selezionate i comandi: **Statistics Table** (tabella delle statistiche), **Anova and Coefficients Table** (Anova e tabella dei coefficienti), **Residual Table** (tabella dei residui) e **Residual Plot** (grafico dei residui). Digitate **Analisi di regressione per l'elasticità del prezzo** nel riquadro **Output Title** (titolo dell'output) e fate clic sul comando **OK**.

L'uso di Excel per ottenere delle variabili dummy

Prendiamo in considerazione l'esempio relativo alla previsione del valore accertato delle case trattato nel Paragrafo 13.7. Per sviluppare un modello di regressione che contenga una variabile dummy a indicare la presenza o l'assenza di un caminetto, aprite il file **HOUSE3.XLS** e fate clic sull'etichetta del foglio **Data** (dati) per verificare che i dati relativi al valore delle case, alla superficie di riscaldamento e alla presenza o meno di un caminetto siano inseriti rispettivamente nelle colonne A, B e C.

Copiate il contenuto della colonna C nella colonna D. Selezionate l'intervallo di celle **C2:C16** contenenti le risposte Yes (Sì) e No e quindi selezionate **Modifica | Sostituisci**. Nella finestra di dialogo che si apre inserite **Yes** nel riquadro **Trova** e **1** nel riquadro **Sostituisci con**. Fate clic su **Sostituisci tutto**. Selezionate di nuovo l'intervallo di celle **C2:C16** e quindi i comandi **Modifica | Sostituisci**. Nella finestra di dialogo che si apre inserite **No** nel riquadro **Trova** e **0**. Fate clic su **Sostituisci tutto**. I valori contenuti nelle colonne A, B e C possono ora essere usati per la costruzione di un modello di regressione con variabili dummy. A tale scopo selezionate **PHStat | Regression | Multiple Regression** (regressione | regressione multipla). Nella finestra di dialogo che si apre, digitate **A1:A16** nel riquadro **Y Variable Cell Range** (intervallo dei valori della Y) e **B1:C16** nel riquadro **X Variable Cell Range** (intervallo dei valori della X). Selezionate il comando **First cells in both ranges contain label** (la prima cella di entrambi gli intervalli di valori contiene l'etichetta) e digitate **95** nel riquadro **Confidence Lvl. for regression coefficients** (livello di confidenza per i coefficienti di regressione). Selezionate i comandi: **Statistics Table** (tabella delle statistiche), **Anova and Coefficients Table** (Anova e tabella dei coefficienti), **Residual Table** (tabella dei residui) e **Residual Plot** (grafico dei residui). Digitate **Analisi di regressione per la previsione dei caminetti** nel riquadro **Output Title** (titolo dell'output) e fate clic sul comando **OK**.

L'uso di Excel per la costruzione di un modello di regressione

Con l'aggiunta PHStat è possibile costruire il modello di regressione sia secondo l'approccio *Best-Subsets* che secondo l'approccio *Stepwise*.

Torniamo al problema relativo al numero di ore di standby di un gruppo di artisti grafici sindacalizzati illustrato nel Paragrafo 13.9. Aprite il file **STANDBY.XLS** e fate clic sull'etichetta del foglio **Data** (dati) per verificare che i dati relativi al numero di ore di standby, allo staff totale, al numero di ore di lavoro a distanza, alle ore Dubner e al numero totale di ore di lavoro siano inseriti rispettivamente nelle colonne da A a E. Selezionate **PHStat | Regression | Best Subsets** (regressione | Best-Subsets). Nella finestra di dialogo che si apre inserite **A1:A27** nel riquadro **Y Variable Cell Range** (intervallo dei valori della Y) e **B1:C16** nel riquadro **X Variable Cell Range** (intervallo dei valori della X). Selezionate il comando **First cells in both ranges contain label** (la prima cella di entrambi gli intervalli di valori contiene l'etichetta) e digitate **95** nel riquadro **Confidence Lvl. for regression coefficients** (livello di confidenza per i coefficienti di regressione). Inserite **Analisi Best-Subsets per il modello relativo alle ore di standby** nel riquadro **Output Title** (titolo dell'output) e fate clic sul comando **OK**.

L'aggiunta PHStat inserisce diversi fogli di lavori (in questo caso 16) tra cui il foglio **BESTWS** simile a quello riprodotto nella Figura 13.17 contenente la sintesi relativa all'analisi Best-Subsets. Questo foglio non può essere cambiato in maniera *dinamica*, per questo motivo qualora si volesse apportare una modifica ai dati, si rende necessario condurre una nuova analisi con PHStat analoga a quella descritta. (Nota La procedura *Best-Subsets* tollera sino a 7 variabili esplicative. Se il computer non ha un processore veloce oppure ha una limitata memoria di lavoro è possibile che tale procedura richieda diversi secondi se non minuti per essere completata. Quando si usano molte variabili esplicative è possibile che la procedura descritta si interrompa dando luogo a un messaggio di errore se il computer ha una limitata memoria di lavoro.)

Per illustrare l'uso dell'aggiunta PHStat per la costruzione del modello di regressione secondo l'approccio *Stepwise*, usiamo ancora i dati contenuti nel file **STANDBY.XLS**. Selezionate **PHStat | Regression | Stepwise Regression** (regressione | regressione Stepwise). Nella finestra di dialogo che si apre inserite **A1:A27** nel riquadro **Y Variable Cell Range** (intervallo dei valori della Y) e **B1:C16** nel riquadro **X Variable Cell Range** (intervallo dei valori della X). Selezionate il comando **First cells in both ranges contain label** (la prima cella di entrambi gli intervalli di valori contiene l'etichetta) e digitate **95** nel riquadro **Confidence Lvl. for regression coefficients** (livello di confidenza per i coefficienti di regressione). Nel riquadro **Stepwise criteria** (criteri stepwise) selezionate il comando **p values** e nel riquadro **Stepwise options** (opzioni stepwise) selezionare **General Stepwise** (stepwise generale). Nel riquadro **p value to enter** (*p*-value per l'inclusione del modello) inserite **0.05** e nel riquadro **p value to remove** (*p*-value per l'esclusione del modello) inserite **0.05**. Inserite **Analisi Stepwise per il modello relativo alle ore di standby** nel riquadro **Output Title** (titolo dell'output) e fate clic sul comando **OK**.