OBIETTIVI

- ✓ *Introdurre il modello di regressione lineare semplice come mezzo per compiere delle previsioni su una variabile mediante un'altra*
- ✓ *Verificare la capacità di adattamento ai dati del modello di regressione lineare semplice*
- ✓ *Studiare le trappole che si possono incontrare nell'uso del modello di regressione lineare semplice*
- ✓ *Introdurre la correlazione come misura dell'associazione tra due variabili*

Introduzione

Nei capitoli precedenti abbiamo preso in considerazione una sola variabile quantitativa: ne abbiamo studiato le misure di sintesi (Capitolo3) ed è stata oggetto di diversi metodi di inferenza statistica volti a determinare delle stime e trarre delle conclusioni su di essa (Capitoli 5-8). Oggetto di questo capitolo è, invece, lo studio della relazione tra due variabili, e a tale scopo introdurremo due tecniche di analisi: la regressione e la correlazione.

La **regressione** ha come scopo principale la previsione: si mira, vale a dire, alla costruzione di un modello attraverso cui prevedere i valori di una **variabile dipendente** o **risposta** a partire dai valori di almeno una **variabile indipendente** o **esplicativa**. In questo capitolo studieremo la regressione lineare semplice in cui si utilizza una *sola* variabile quantitativa indipendente X per prevedere una variabile quantitativa dipendente, Y. Nel Capitolo 10, introdurremo il modello di regressione lineare *multipla*, che invece impiega *diverse* variabili esplicative ($X_1, X_2 \ldots, X_p$) per prevedere una variabile quantitativa Y.[1]

La **correlazione** ha, invece, come scopo lo studio dell'associazione tra variabili quantitative. Nel paragrafo 9.11, ad esempio, studieremo la correlazione tra il marco tedesco e lo yen giapponese per un periodo di 10 anni. L'attenzione in questo caso è focalizzata non tanto sulla possibilità di prevedere una variabile mediante un'altra, quanto sullo studio delle relazioni che possono sussistere tra due variabili quantitative.

[1] *Tra i modelli di regressione in cui la variabile dipendente è una variabile qualitativa ricordiamo la regressione logistica (riferimento bibliografico 4)*

◆ APPLICAZIONE *Previsione delle vendite di un negozio di abbigliamento*

Nel corso degli ultimi 25 anni, una catena di negozi di abbigliamento femminile ha accresciuto la sua quota di mercato grazie all'apertura di nuove filiali. La decisione relativa alla dimensione di una nuova filiale non è mai stata oggetto di un approccio sistematico, ma quest'anno il manager responsabile del progetto di apertura di nuove filiali intende basare le sue decisioni su uno studio che gli consenta di prevedere le vendite annuali dei nuovi negozi. ◆

9.1 I MODELLI DI REGRESSIONE

Nel Capitolo 2 abbiamo visto come una variabile possa essere descritta facendo ricorso a diverse rappresentazioni grafiche. In maniera analoga nella regressione, per studiare la relazione tra due variabili, si fa uso di un grafico detto **diagramma di dispersione**, che si

ottiene riportando sull'asse delle ascisse i valori della variabile indipendente X e sull'asse delle ordinate i valori della variabile dipendente Y. Tra due variabili possono sussistere diversi tipi di relazione, esprimibili mediante funzioni matematiche dalle più semplici alle più complesse. La **relazione lineare**, illustrata dalla Figura 9.1, è senz'altro la forma più semplice di relazione tra variabili.

Il modello di regressione lineare è definito come segue:

Il modello di regressione lineare semplice

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (9.1)$$

dove

β_0 = l'intercetta per la popolazione

β_1 = l'inclinazione per la popolazione

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

L'**inclinazione** β_1 indica come varia Y in corrispondenza di una variazione unitaria di X. L'**intercetta** β_0 corrisponde al valore medio di Y quando X è uguale a 0. L'ultima componente del modello, ε_i, rappresenta l'elemento di errore casuale contenuto in Y in corrispondenza di ciascuna osservazione i.

La scelta del modello matematico appropriato per rappresentare la relazione che lega Y a X è suggerita dal modo in cui si distribuiscono i valori delle due variabili nel diagramma di dispersione. Prendiamo in considerazione i riquadri A-F della Figura 9.2. Nel Riquadro A, i valori di Y crescono linearmente al crescere dei valori di X, situazione simile a quella rappresentata nella Figura 9.3, che evidenzia l'esistenza di una relazione diretta tra la dimensione del negozio (in piedi al quadrato) e l'ammontare delle vendite nelle filiali della catena di abbigliamento femminile considerate nell'Applicazione.

Nel Riquadro B, si riporta un esempio di relazione inversa tra X e Y: all'aumentare di X, Y diminuisce. Una relazione analoga sussiste ad esempio tra il prezzo e l'ammontare delle vendite di un prodotto.

Il Riquadro C si riferisce a un dataset in cui non sussiste alcuna relazione tra X e Y: in corrispondenza di ciascun valore di X si possono osservare tanto valori elevati quanto valori bassi di Y.

Il Riquadro D fornisce un esempio di relazione *polinomiale diretta* tra X e Y: i valori di Y crescono al crescere dei valori di X, ma con un tasso di crescita che si riduce da un certo punto in poi (una volta superati alcuni valori della X). Una relazione di questo tipo sussiste, ad esempio, tra l'età e i costi di manutenzione di un macchinario. All'aumentare dell'età di un macchinario, i costi di manutenzione in principio aumentano rapidamente, per poi stabilizzarsi dopo un certo numero di anni.

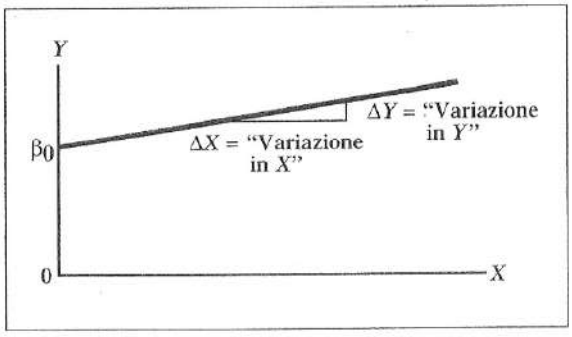
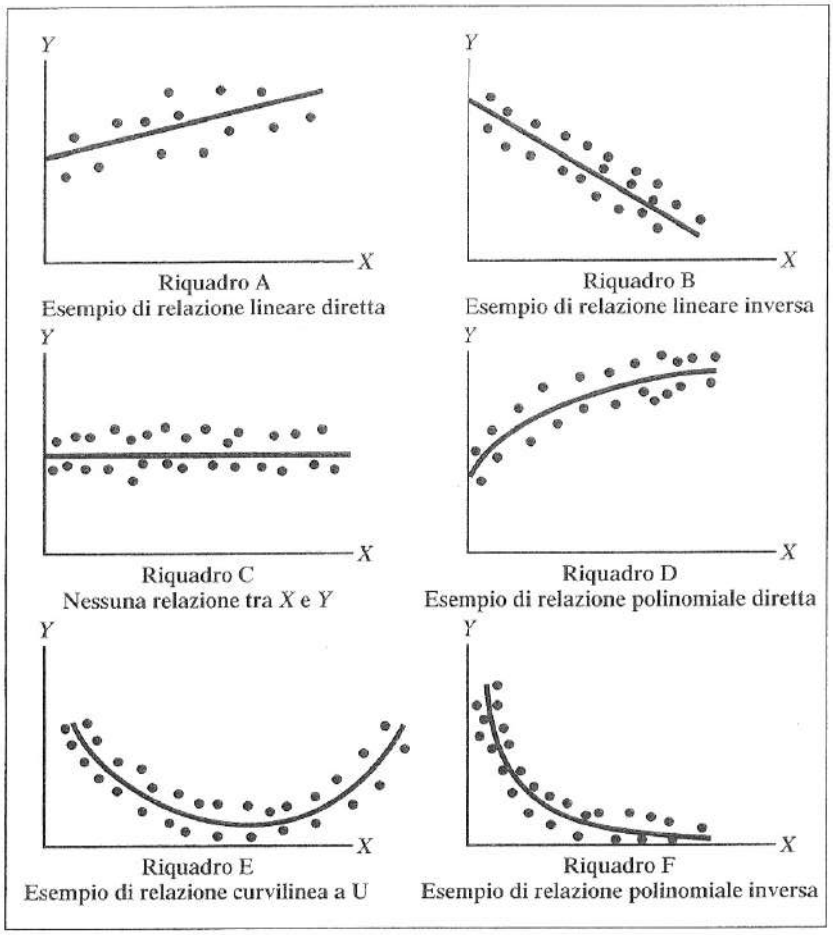


FIGURA 9.1
Una relazione lineare positiva

FIGURA 9.2
Esempi di relazioni tra variabili evidenziate dai diagrammi di dispersione



Nel Riquadro E si osserva una relazione parabolica o a forma di U tra X e Y: all'aumentare dei valori di X, Y diminuisce sino a un certo punto a partire dal quale comincia a crescere. Una relazione di questo genere sussiste, ad esempio, tra il numero di errori all'ora, che si compiono quando si svolge un certo lavoro, e il numero delle ore di lavoro. Il numero di errori si riduce mano a mano che si acquista dimestichezza con il lavoro, ma poi comincia a crescere come conseguenza della stanchezza o della noia.

Infine, nel Riquadro F si osserva una relazione esponenziale o polinomiale inversa tra X e Y: i valori di Y decrescono al crescere dei valori di X, ma con un tasso di decrescita che si riduce da un certo punto in poi (una volta superati alcuni valori della X). Una relazione di questo tipo sussiste tra il prezzo di vendita di un'automobile e la sua età: il prezzo di un'automobile si riduce drasticamente dopo il primo anno dall'acquisto, ma diminuisce meno rapidamente negli anni successivi.

In questo paragrafo abbiamo introdotto brevemente vari modelli per lo studio della relazione tra due variabili, individuati a mezzo del diagramma di dispersione. Sebbene i diagrammi di dispersione siano dei validi strumenti di analisi, tecniche statistiche più sofisticate consentono di pervenire alla scelta del modello di regressione più appropriato. Nei paragrafi successivi ci concentreremo sul modello di regressione lineare.

9.2 L'EQUAZIONE DELLA RETTA DI REGRESSIONE

Torniamo all'Applicazione con cui abbiamo aperto il capitolo. Supponete che il manager intenda esaminare la relazione tra la dimensione e l'ammontare annuo delle vendite di un negozio. La Tabella 9.1 riporta i valori delle due variabili (dimensione e ammontare delle vendite) per un campione di 14 negozi.

Il diagramma di dispersione relativo al dataset della Tabella 9.1, riportato nella Figura 9.3, mostra l'esistenza di una relazione lineare crescente tra la dimensione dei negozi (X) e l'ammontare annuo delle vendite (Y): all'aumentare della dimensione, le vendite aumentano come su una retta. Ora, se riteniamo che la relazione tra X e Y possa essere rappresentata a mezzo di una retta, non resta che stabilire come individuare la retta che meglio si adatta ai dati.

DATASET SITE

Tabella 9.1 *Dimensione del negozio (in piedi al quadrato) e vendite annue (in migliaia di dollari) per un campione di 14 filiali di una catena di negozi di abbigliamento*

Negozio	Dimensione (Piedi al quadrato)	Vendite Annue ($ 000)	Negozio	Dimensione (Piedi al quadrato)	Vendite Annue ($ 000)
1	1726	3681	8	1102	2694
2	1642	3895	9	3151	5468
3	2816	6653	10	1516	2898
4	5555	9543	11	5161	10 674
5	1292	3418	12	4567	7585
6	2208	5563	13	5841	11 760
7	1313	3660	14	3008	4085

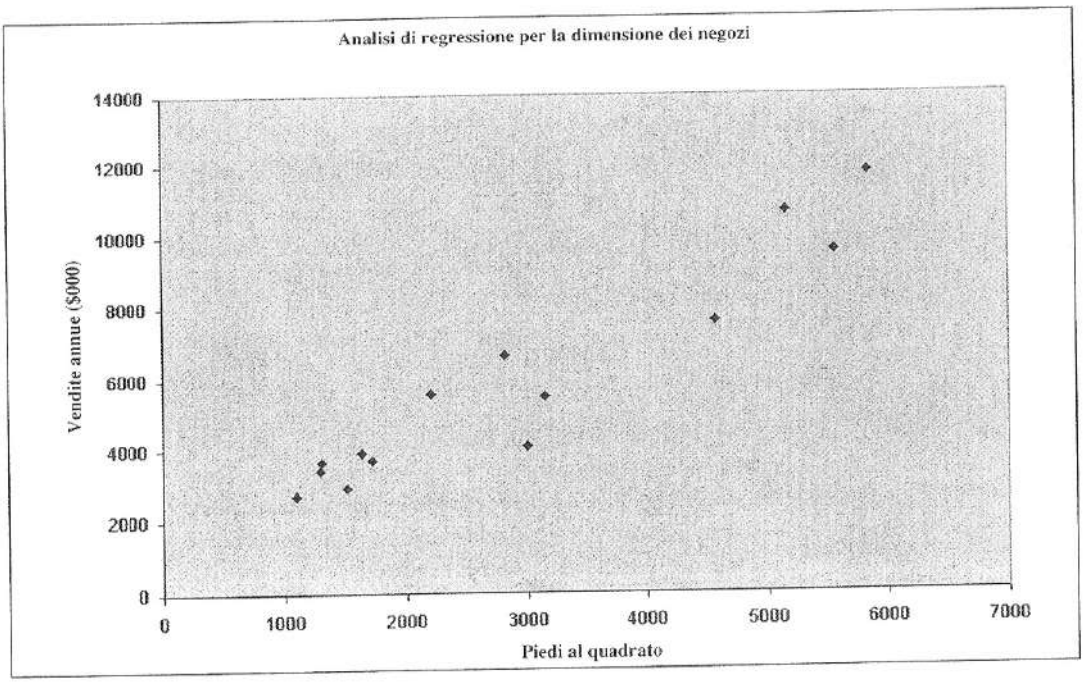


FIGURA 9.3 Diagramma di dispersione ottenuto con Microsoft Excel per il dataset SITE

Il metodo dei minimi quadrati

Nel paragrafo precedente abbiamo proposto un modello statistico (il modello di regressione semplice) per studiare la relazione tra due variabili quantitative (la dimensione dei negozi e l'ammontare annuo delle vendite) alla luce dei soli dati campionari. Si dimostra che sotto determinate ipotesi (paragrafo 9.4) l'intercetta campionaria b_0 e l'inclinazione campionaria b_1 si possono usare come stimatori dei parametri della popolazione β_0 e β_1: si ottiene in questa maniera la forma campionaria del modello di regressione lineare semplice.

L'equazione campionaria del modello di regressione lineare

La previsione di Y in base al modello di regressione lineare è data dalla somma tra l'intercetta campionaria e il prodotto tra il valore di X e l'inclinazione campionaria

$$\hat{Y}_i = b_0 + b_1 X_i \tag{9.2}$$

dove

$\hat{Y}_i$ = previsione di Y per l'osservazione i

X_i = valore di X per l'osservazione i

La previsione di Y richiede, allora, il calcolo dei due coefficienti di regressione b_0 e b_1. Una volta determinati, possiamo tracciare la retta di regressione nel diagramma di dispersione e valutare visivamente la capacità esplicativa del modello, osservando se la retta stimata si avvicina o meno ai dati osservati.

La regressione mira a individuare la retta che meglio si adatta ai dati, laddove tale capacità di adattamento può essere valutata in base a criteri diversi. Il criterio più semplice consiste nel valutare le differenze tra i valori osservati (Y_i) e i valori previsti in base alla retta stimata ($\hat{Y}_i$) e quindi cercare quella retta che minimizza tali differenze. Tuttavia, siccome le differenze considerate possono essere negative per alcune osservazioni e positive per altre, si tratterà di *minimizzare* la somma dei loro quadrati:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

dove

Y_i = il vero valore di Y per l'osservazione di i

$\hat{Y}_i$ = il valore previsto di Y per l'osservazione di i

Dal momento che in base al modello proposto $\hat{Y}_i = b_0 + b_1 X_i$, si tratta di minimizzare la seguente espressione:

$$\sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

rispetto alle due incognite b_0 e b_1.

La tecnica matematica che consiste nel determinare i valori di b_0 e b_1 che rendono minima l'espressione precedente prende il nome di **metodo dei minimi quadrati**.

Nella Figura 9.4 si riporta l'output di Excel relativo al dataset della Tabella 9.1. I valori stimati dei **coefficienti di regressione** sono rispettivamente $b_0 = 901.247$ e $b_1 = 1.686$ e quindi la retta stimata ha la seguente espressione:

$$\hat{Y}_i = 901.247 + 1.686\, X_i$$

	A	B	C	D	E	F	G
1	Analisi di regressione per la dimensione dei negozi						
2							
3	*Statistica della regressione*						
4	R multiplo	0,953824159					
5	R al quadrato	0,909780526					
6	R al quadrato corretto	0,902262236					
7	Errore standard	936,8500077					
8	Osservazioni	14					
9							
10	ANALISI VARIANZA						
11		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*	
12	Regressione	1	106208119,7	106208119,7	121,0089774	1,26653E-07	
13	Residuo	12	10532255,24	877687,9369			
14	Totale	13	116740374,9				
15							
16		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
17	Intercetta	901,2465701	513,0227603	1,756737985	0,10442398	-216,5339829	2019,027123
18	Dimensione	1,68613497	0,153279311	11,00040805	1,26653E-07	1,352168946	2,020101895

SQR, SQE, SQT, b_0, b_1

FIGURA 9.4 Analisi di regressione per la dimensione dei negozi

Un valore dell'inclinazione b_1 uguale a +1.686 significa che se X aumenta di un'unità Y aumenta in media di 1.686 unità. In altri termini, in base al modello proposto, si prevede che se la dimensione del negozio aumenta di 1 unità, l'ammontare delle vendite aumenta di 1686 dollari. L'inclinazione della retta rappresenta, allora, la proporzione delle vendite che varia in funzione della dimensione del negozio.

L'intercetta rappresenta il valore medio di Y quando X è uguale a 0. Nel caso considerato, dal momento che la dimensione del negozio non può essere uguale a 0, il valore di b_0 si deve intendere come la proporzione delle vendite che varia in funzione di fattori diversi dalla dimensione del negozio.

Nella Figura 9.5 si riporta il diagramma di dispersione relativo ai dati della Tabella 9.1 e la corrispondente retta di regressione stimata con il metodo dei minimi quadrati.

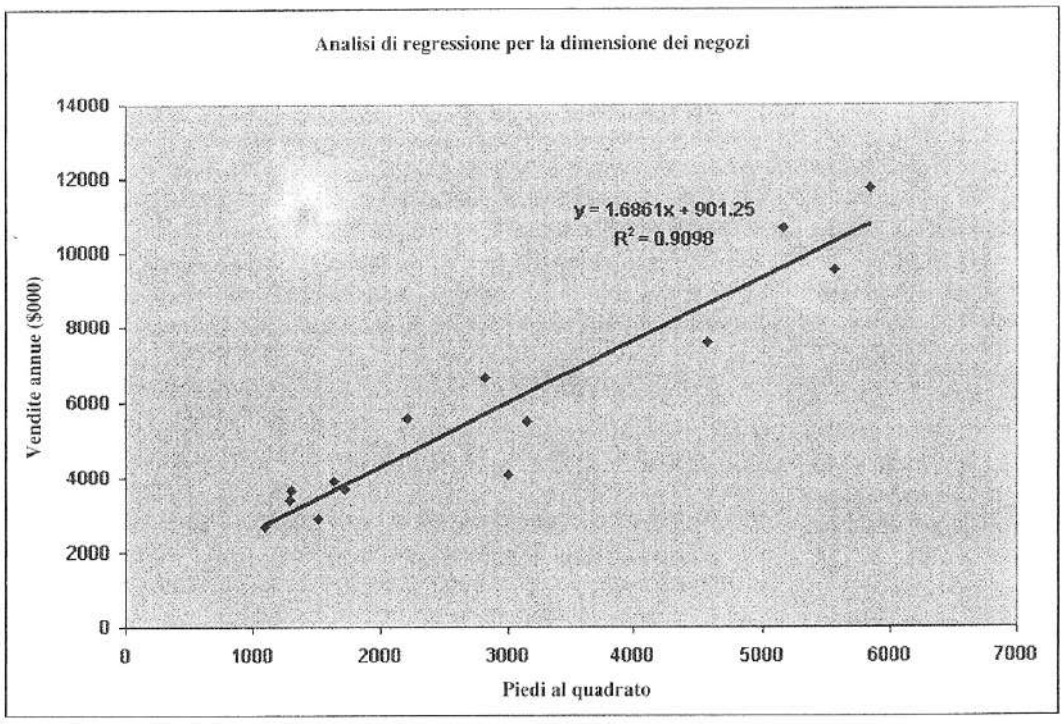


FIGURA 9.5 Diagramma di dispersione e retta di regressione ottenuti con Excel per il dataset SITE

Esempio 9.1 *Interpretazione dell'intercetta b_0 e dell'inclinazione b_1*

Un economista intende usare il tasso di crescita annuo della produttività negli Stati Uniti (X) per prevedere la variazione percentuale dell'indice Standard & Poor di 500 azioni (Y). Sulla base di dati annui per periodo di 50 anni, l'economista ottiene la seguente stima della retta di regressione di Y su X:

$$\hat{Y}_i = -5.0 + 7X_i$$

Come si possono interpretare l'intercetta b_0 e l'inclinazione b_1?

SOLUZIONE

Alla luce del valore assunto dall'intercetta, $b_0 = -5$, ci aspettiamo, in base al modello proposto, che l'indice Standard e Poor diminuisca del 5% se il tasso di produttività è uguale a 0. L'inclinazione $b_1 = 7$ ci dice che in corrispondenza di una variazione della produttività dell'1% dovremmo osservare una variazione dell'indice del 7%.

Esempio 9.2 *Previsione delle vendite annue di un negozio a partire dalla sua dimensione*

Prevedete in base al modello di regressione stimato (tabulato Excel della Figura 9.4), l'ammontare delle vendite per un negozio di 4000 piedi al quadrato.

SOLUZIONE

Si tratta di sostituire il valore 4000 al valore di X_i nella retta di regressione stimata:

$$\hat{Y}_i = 901.247 + 1.686X_i$$

$$\hat{Y}_i = 901.247 + 1.686(4000) = 7645.786 \text{ o } \$\ 7645.786$$

Pertanto si prevede un ammontare delle vendite pari a $ 7645.786 per un negozio di 4000 piedi al quadrato.

La previsione nel modello di regressione: l'interpolazione a confronto con l'estrapolazione

Nel fare previsioni alla luce di un modello di regressione dobbiamo sempre tenere presente quale è l'intervallo dei valori assunti dalla variabile indipendente e questo perché possiamo prevedere Y solo in corrispondenza dei valori di X che cadono in questo intervallo. Quando facciamo previsioni in corrispondenza a valori di X che cadono in questo intervallo, diciamo che stiamo facendo delle *interpolazioni*, mentre quando cerchiamo di prevedere il valore di Y corrispondente a valori di X che non cadono in questo intervallo, stiamo cercando di fare delle *estrapolazioni*. Per chiarire meglio la differenza tra le due operazioni, consideriamo i dati nella Tabella 9.1, e supponiamo di utilizzare la superficie del negozio per prevedere l'ammontare delle vendite; i valori osservati con riferimento alla variabile esplicativa vanno da 1102 a 5841 piedi al quadrato. Di conseguenza, potremo fare previsioni sulle vendite di negozi con una superficie compresa in questo intervallo. Qualunque previsione sulle vendite di negozi con superficie al di fuori di questo intervallo può essere fatta soltanto sotto l'ipotesi che la relazione stimata tra le due variabili rimanga la stessa anche al di fuori di questo intervallo (e questa è un'ipotesi che si basa su nostre supposizioni e non su dati osservati).

Esercizi del paragrafo 9.2

• **9.1** Alla luce della seguente stima della retta di regressione di Y su X:

$$\hat{Y}_i = 2 + 5X_i$$

(a) fornite un'interpretazione dell'intercetta b_0;
(b) fornite un'interpretazione dell'inclinazione b_1;
(c) prevedete il valore di Y per $X = 3$.
(d) dite per quali dei seguenti valori di X usereste la retta stimata per prevedere Y, sapendo che i valori assunti da X sono compresi tra 2 e 25.
(1) 3? (4) 24?
(2) −3? (5) 26?
(3) 0?

Nota: fate ricorso a Microsoft Excel per la risoluzione degli esercizi seguenti.



DATASET PETFOOD

• **9.2** Il manager di una catena di supermercati intende stabilire in quale maniera la vendita di cibo per animali è influenzata dallo spazio sugli scaffali destinato al prodotto. La seguente tabella riporta i valori dello spazio sugli scaffali (in piedi) e dell'ammontare delle vendite settimanali di cibo per animali, in 12 supermercati della medesima grandezza.

SUPERMARKET	SPAZIO SUGLI SCAFFALI, X (PIEDI)	VENDITE SETTIMANALI, Y (MIGLIAIA DI DOLLARI)	SUPERMARKET	SPAZIO SUGLI SCAFFALI, X (PIEDI)	VENDITE SETTIMANALI, Y (MIGLIAIA DI DOLLARI)
1	5	1.6	7	15	2.3
2	5	2.2	8	15	2.7
3	5	1.4	9	15	2.8
4	10	1.9	10	20	2.6
5	10	2.4	11	20	2.9
6	10	2.6	12	20	3.1

(a) Disegnate il diagramma di dispersione per i dati della tabella.
(b) Nell'ipotesi che tra le due variabili sussista una relazione lineare, stimate con il metodo dei minimi quadrati i coefficienti di regressione b_0 e b_1.
(c) Fornite un'interpretazione di b_1.
(d) Prevedete l'ammontare delle vendite settimanali di cibo per animali se lo spazio destinato ai prodotti è uguale a 8 piedi.
(e) Supponete che l'ammontare delle vendite del negozio 12 sia 2.6 e rispondete di nuovo ai punti (a)-(d).



DATASET MOVIE

9.3 Una società che si occupa della vendita di videocassette di film per uso domestico (home video) intende stimare il numero di cassette che potrà vendere in base al successo di botteghino riscosso dal film. La seguente tabella riporta i dati relativi all'ammontare delle vendite dei biglietti (in milioni di dollari) e il numero corrispondente di videocassette vendute (in migliaia) per un campione di 30 film.
(a) Create il diagramma di dispersione per i dati della tabella.
(b) Stimate con il metodo dei minimi quadrati i coefficienti di regressione b_0 e b_1.
(c) Fornite un'interpretazione di b_0 e b_1 con riferimento al problema considerato.
(d) Prevedete il numero di videocassette che potrebbero essere vendute per un film per il quale sono stati venduti $ 20 milioni di biglietti.

Movie	Box Office Gross ($ milioni)	Home Video Units Sold (000)	Movie	Box Office Gross ($ milioni)	Home Video Units Sold (000)
1	1.10	57.18	16	9.36	190.80
2	1.13	26.17	17	9.89	121.57
3	1.18	92.79	18	12.66	183.30
4	1.25	61.60	19	15.35	204.72
5	1.44	46.50	20	17.55	112.47
6	1.53	85.06	21	17.91	162.95
7	1.53	103.52	22	18.25	109.20
8	1.69	30.88	23	23.13	280.79
9	1.74	49.29	24	27.62	229.51
10	1.77	24.14	25	37.09	277.68
11	2.42	115.31	26	40.73	226.73
12	5.34	87.04	27	45.55	365.14
13	5.70	128.45	28	46.62	218.64
14	6.43	126.64	29	54.70	286.31
15	8.59	107.28	30	58.51	254.58



DATASET RENT

9.4 Un agente immobiliare intende prevedere gli affitti mensili degli appartamenti sulla base della loro dimensione. Nella tabella seguente si riportano i valori degli affitti e della dimensione di 25 appartamenti di una zona residenziale.

Apartment	Monthly Rent ($)	Size (square feet)	Apartment	Monthly Rent ($)	Size (square feet)
1	950	850	14	1800	1369
2	1600	1450	15	1400	1175
3	1200	1085	16	1450	1225
4	1500	1232	17	1100	1245
5	950	718	18	1700	1259
6	1700	1485	19	1200	1150
7	1650	1136	20	1150	896
8	935	726	21	1600	1361
9	875	700	22	1650	1040
10	1150	956	23	1200	755
11	1400	1100	24	800	1000
12	1650	1285	25	1750	1200
13	2300	1985			

(a) Create il diagramma di dispersione per i dati della tabella.
(b) Stimate con il metodo dei minimi quadrati i coefficienti di regressione b_0 e b_1.
(c) Fornite un'interpretazione di b_0 e b_1 con riferimento al problema considerato.
(d) Prevedete l'ammontare dell'affitto mensile per un appartamento di 1000 piedi al quadrato.
(e) Perché non si può utilizzare la retta stimata per prevedere l'affitto mensile di appartamenti di 500 piedi al quadrato?

9.3 Le misure di variabilità

In questo paragrafo introduciamo alcune misure di variabilità che consentono di valutare le capacità previsive del modello statistico proposto. La somma totale dei quadrati (SQT)

misura la variabilità dei valori Y_i attorno alla media $\hat{Y}_i$ Nella regressione, la **variabilità totale** oppure **somma totale dei quadrati** si scompone nella **variabilità spiegata** oppure **somma dei quadrati della regressione** (*SQR*), che rappresenta la parte di variabilità della *Y* attribuibile alla relazione che sussiste tra *X* e *Y*, e **nella variabilità non spiegata** oppure **somma dei quadrati degli errori** (*SQE*), che invece è la parte di variabilità non imputabile alla relazione tra *X* e *Y*. La Figura 9.6 illustra tali misure di variabilità.

La somma dei quadrati della regressione (*SQR*) è data dalla differenza tra $\hat{Y}_i$ (i valori di *Y* previsti in base al modello di regressione proposto) e $\overline{Y}$ (la media dei valori di *Y*). La somma dei quadrati degli errori (*SQE*) è data dalla differenza tra Y_i e $\hat{Y}_i$. Di seguito sono riportate le definizioni delle misure di variabilità introdotte.

Le misure di variabilità nella regressione

Somma totale dei quadrati = somma dei quadrati della regressione + somma dei quadrati degli errori

$$SQT = SQR + SQE \tag{9.3}$$

La somma totale dei quadrati (*SQT*)

La somma totale dei quadrati (*SQT*) è data dalla somma dei quadrati delle differenze tra i valori osservati di *Y* e la loro media.

$$SQT = \text{variabilità totale} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{9.4}$$

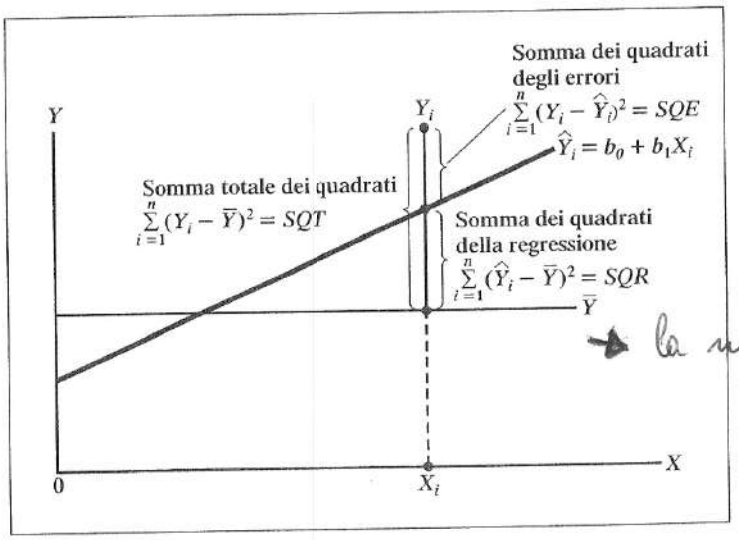
La somma dei quadrati della regressione (*SQR*)

La somma dei quadrati della regressione (*SQR*) è data dalla somma dei quadrati delle differenze tra i valori previsti di *Y* e la media di *Y*.

$$SQR = \text{variabilità spiegata} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 \tag{9.5}$$

$$= SQT - SQE$$

FIGURA 9.6
Le misure di variabilità nel modello di regressione



→ la media dei valori Y

La somma dei quadrati degli errori (*SQE*)

La somma dei quadrati degli errori (*SQE*) è data dalla somma dei quadrati delle differenze tra i valori osservati e i valori previsti di Y

$$SQE = \text{variabilità non spiegata} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{9.6}$$

In base alle Figure 9.4 e 9.5, osserviamo

$$SQR = 106{,}208{,}120; \quad SQE = 10{,}532{,}255; \quad \text{e} \quad SQT = 116{,}740{,}375$$

In base alle Figure 9.4 4 9.5, osserviamo che:

$$SQR = 106\,208\,120; \quad SQE = 10\,532\,255; \quad SQT = 116\,740\,375$$

Inoltre, in base all'equazione (9.3):

$$SQT = SQR + SQE$$

$$116\,740\,375 = 106\,208\,120 + 10\,532\,255$$

La somma dei quadrati delle differenze dalla media, 116 740 375, si suddivide nella somma dei quadrati spiegata dalla regressione 106 208 120 e nella somma dei quadrati residua, 10 532 255.

COMMENTO: Notazione scientifica

In alcune versioni di Excel, le cifre molto piccole o molto grandi sono espresse non in formato numerico, ma facendo ricorso alla "notazione scientifica". Ad esempio potremmo trovare la *SQR* dell'esempio precedente (*SQR* = 106208120) formattata come 1.06E+08. I numeri che seguono la lettera E corrispondono al numero di cifre decimali per le quali si deve spostare la virgola verso sinistra (se negativi) o verso destra (se positivi) per ottenere l'abituale formato numerico. 3.7431E–02 è il numero che si ottiene da 3.7431 spostando la virgola di 2 posti verso sinistra 0.037431 e mentre 3.7431E+02 è il numero che si ottiene da 3.7431 spostando la virgola di 2 posti verso destra, 374.31. Pertanto 1.06E+08 non è altro che 106000000. Osservate come nella notazione scientifica si faccia ricorso a poche cifre significative, cosa che comporta un'approssimazione dei numeri considerati.

Il coefficiente di determinazione

Le somme dei quadrati precedentemente introdotte (*SQT*, *SQR* e *SQE*) forniscono, se considerate da sole, informazioni limitate sulla bontà del modello statistico proposto. Tuttavia il rapporto tra la somma dei quadrati della regressione e la somma totale dei quadrati si configura come una misura utile per valutare il modello di regressione.
Tale misura prende il nome di **coefficiente di determinazione** ed è di seguito definita.

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.

$$r^2 = \frac{SQR}{SQT} \tag{9.7}$$

Il coefficiente di determinazione misura la parte di variabilità di Y spiegata dalla variabile indipendente X nel modello di regressione. Tornando all'esempio riguardante la

catena di negozi di abbigliamento, dal momento che SQR = 106208120, SQE = 10532255 e SQT = 116740375, il coefficiente di determinazione è dato da:

$$r^2 = \frac{106\,208\,120}{116\,740\,375} = 0.91$$

la variabilità delle vendite annue risulta spiegata per il 91% dalla dimensione dei negozi. In questo caso osserviamo che esiste una relazione lineare forte tra le due variabili considerate, perché solo il 9% della variabilità della variabile dipendente si deve ascrivere a fattori diversi dalla variabile indipendente prescelta.

L'errore standard della stima

Sebbene il metodo dei minimi quadrati consenta di individuare la retta che rende minima la differenza tra i valori osservati e i valori previsti, la retta di regressione non conduce mai a previsioni scevre da errori, se non nel raro caso in cui tutti i dati siano disposti su di essa. Pertanto, si rende necessaria una statistica campionaria che misuri la variabilità degli scostamenti dei valori osservati da quelli previsti. Tale misura prende il nome di **errore standard della stima**, indicato con il simbolo S_{YX}.

L'errore standard della stima

$$S_{YX} = \sqrt{\frac{SQE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} \qquad (9.8)$$

dove

Y_i = il valore di Y in corrispondenza X_i

$\hat{Y}_i$ = il valore previsto di Y in corrispondenza di X_i

SQE = somma dei quadrati degli errori

In base all'equazione (9.8), con SQE = 10532255, avremo:

$$S_{YX} = \sqrt{\frac{10\,532\,255}{14-2}} = 936.85$$

L'errore standard della stima uguale a \$ 936.85 è indicato come Errore Standard nel tabulato di Excel della Figura 9.4. L'errore standard ha il medesimo significato dello scarto quadratico medio: come lo scarto quadratico misura la variabilità delle osservazioni attorno alla media, così l'errore standard misura la variabilità delle osservazioni attorno alla retta di regressione. Inoltre, come vedremo nei paragrafi 9.7 e 9.8, l'errore standard consente di stabilire se tra le due variabili considerate sussiste una relazione significativa dal punto di vista statistico e rende possibili inferenze sul valore previsto di Y.

Esercizi del paragrafo 9.3

9.5 Come si interpreta un coefficiente di determinazione uguale a 0.80?

• **9.6** Se SQR= 36 e SQE=4, calcolate la SQT e il coefficiente di determinazione r^2. Spiegate il significato di r^2.



DATASET PETFOOD

• **9.7** In base all'output di Excel ottenuto per risolvere l'esercizio 9.2:

(a) calcolate il coefficiente di determinazione r^2 e spiegatene il significato;

(b) calcolate l'errore standard della stima;

(c) ritenete che il modello di regressione sia un utile strumento di previsione delle vendite di cibo per animali?

DATASET MOVIE

9.8 In base all'output di Excel ottenuto per risolvere l'esercizio 9.3:
(a) calcolate il coefficiente di determinazione r^2 e spiegatene il significato;
(b) calcolate l'errore standard della stima;
(c) ritenete che il modello di regressione sia un utile strumento di previsione delle vendite di videocassette?

DATASET RENT

9.9 In base all'output di Excel ottenuto per risolvere l'esercizio 9.4:
(a) calcolate il coefficiente di determinazione r^2 e spiegatene il significato;
(b) calcolate l'errore standard della stima;
(c) ritenete che il modello di regressione sia un utile strumento di previsione degli affitti mensili?

9.4 LE ASSUNZIONI DEL MODELLO

Quando abbiamo introdotto la teoria della verifica di ipotesi e dell'analisi della varianza, abbiamo più volte sottolineato come una corretta applicazione delle procedure statistiche dipenda in genere dal soddisfacimento delle ipotesi su cui esse si fondano. Le assunzioni alla base del modello di regressione sono analoghe a quelle su cui si fonda l'analisi della varianza, perché i due modelli di analisi ricadono nello stesso insieme: la classe dei *modelli lineari* (riferimento bibliografico 6).

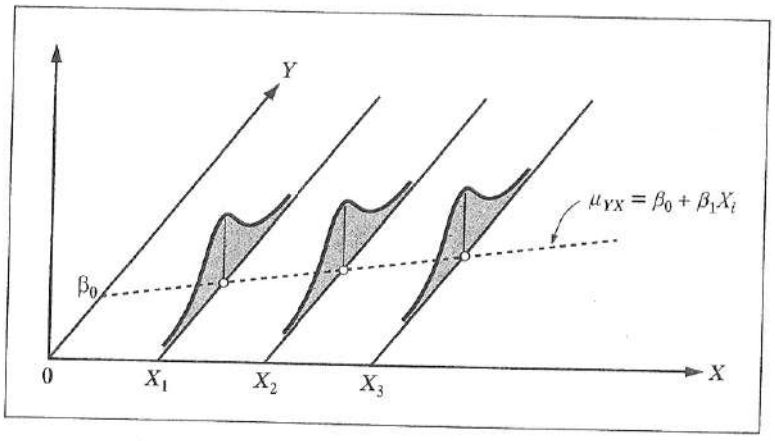
Nel Riquadro 9.1 si riportano le **ipotesi del modello di regressione**.

Riquadro 9.1 Le ipotesi del modello di regressione

✓ **1.** Distribuzione normale degli errori.
✓ **2.** Omoschedasticità.
✓ **3.** Indipendenza degli errori.

In base alla prima ipotesi, la **normalità**, si richiede che gli errori abbiamo, per ogni valore di X, una distribuzione normale (Figura 9.7). Analogamente al test t e al test F dell'ANOVA, il modello di regressione risulta robusto rispetto a scostamenti dall'ipotesi di normalità: le inferenze sulla retta di regressione e sui coefficienti non risultano seriamente influenzate da una distribuzione degli errori solo *approssimativamente* normale.

FIGURA 9.7
Le ipotesi del modello di regressione



La seconda ipotesi, l'**omoschedasticità**, richiede che la variabilità degli errori sia costante per ciascun valore di X. Gli errori devono, vale a dire, variare di un medesimo ammontare sia in corrispondenza di valori elevati, che in corrispondenza di valori piccoli di X. L'omoschedasticità degli errori è cruciale ai fini dell'applicazione del metodo dei

minimi quadrati. Se tale ipotesi non è soddisfatta, si devono trasformare opportunamente i dati oppure ricorrere a metodi di stima diversi (i minimi quadrati ponderati, riferimento bibliografico 6).

In base alla terza ipotesi, l'**indipendenza**, gli errori devono essere indipendenti per ciascun valore di X. Questa ipotesi diventa importante quando i dati sono frutto di osservazioni nel corso del tempo. In tal caso, infatti, può accadere che gli errori di un certo periodo di tempo siano correlati con gli errori del periodo precedente.

9.5 L'ANALISI DEI RESIDUI

In questo paragrafo introduciamo l'**analisi dei residui**, come approccio grafico, che consente sia di valutare l'adattamento ai dati del modello di regressione stimato, sia di individuare eventuali violazioni delle ipotesi sottostanti al modello.

Valutazione dell'adattamento del modello di regressione stimato

Si definisce **residuo** o errore stimato, e_i, la differenza tra il valore osservato (Y_i) e il valore previsto ($\hat{Y}_i$), della variabile dipendente per un dato valore X_i della variabile indipendente.

Il residuo

Il residuo è uguale alla differenza tra valore osservato e il valore previsto di Y:

$$e_i = Y_i - \hat{Y}_i \tag{9.9}$$

La capacità di adattamento ai dati della retta di regressione può essere valutata rappresentando in un grafico i residui (sull'asse delle ordinate) rispetto ai corrispondenti valori di X_i (sull'asse delle ascisse). Se il modello è appropriato, nel grafico dei residui non si riconosce nessun andamento particolare. Se, invece, il modello non è adeguato il grafico rivelerà l'esistenza di una relazione tra i valori X_i e i residui e_i. La Figura 9.8 illustra una situazione di questo genere.

Pur rivelando la presenza di un trend in Y all'aumentare di X, i dati del Riquadro A sembrano suggerire l'esistenza di una relazione tra le due variabile non lineare ma polinomiale, perché il tasso di crescita di Y si riduce all'aumentare dei valori di X. Pertanto, in questo caso il modello di regressione lineare non sembra appropriato. Il Riquadro B conferma l'esistenza di una relazione non lineare: si può infatti osservare l'andamento curvilineo del

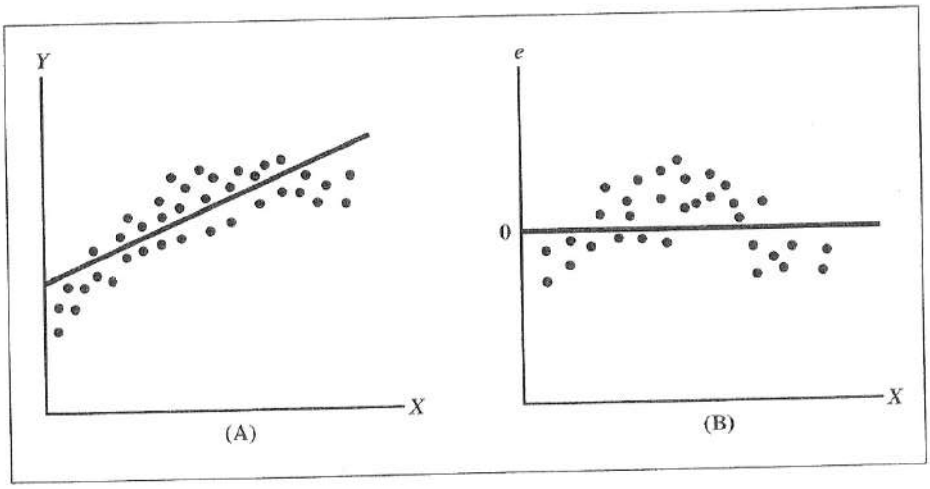


FIGURA 9.8
Valutazione dell'adeguatezza del modello di regressione lineare semplice

grafico dei residui e_i rispetto a X_i. Con il grafico dei residui si rimuove il trend lineare di Y rispetto a X e si evidenzia quindi lo scarso adattamento ai dati del modello di regressione (*lack of fit*). Possiamo quindi concludere che un modello polinomiale in questo caso è più appropriato di un modello lineare semplice.

Torniamo all'esempio relativo alla catena di negozi di abbigliamento. Nella Figura 9.9 si riporta l'output di Excel relativo all'analisi dei residui, contenente le previsioni della variabile risposta (l'ammontare delle vendite annue) e i corrispondenti residui.

Nella Figura 9.10, si riporta il grafico dei residui rispetto ai valori della variabile indipendente (la dimensione dei negozi). Il grafico non rivela alcuna relazione particolare tra i residui e X_i: i residui sembrano distribuirsi in maniera uguale al di sopra e al di sotto dello 0. Possiamo concludere che, in questo caso, il modello di regressione è adeguato.

21			
22	OUTPUT RESIDUI		
23			
24	*Osservazione*	*Vendite previste*	*Residui*
25	1	3811,515529	-130,516
26	2	3669,880191	225,1198
27	3	5649,402647	1003,597
28	4	10267,72633	-724,726
29	5	3879,732952	338,267
30	6	4624,232585	938,7674
31	7	3115,141786	544,8582
32	8	2759,367307	-65,3673
33	9	6214,257862	-746,258
34	10	3457,427185	-559,427
35	11	9603,389152	1070,611
36	12	8601,82498	-1016,82
37	13	10749,96093	1010,039
38	14	5973,140561	-1888,14

FIGURA 9.9
L'output di Excel relativo all'analisi dei residui per il problema di scelta della dimensione delle nuove filiali

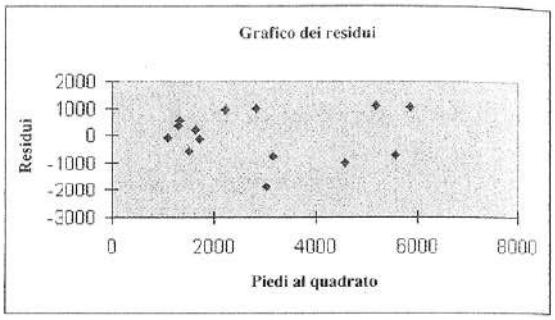


FIGURA 9.10
Grafico dei residui rispetto alla dimensione dei negozi ottenuto con Excel

Valutazione delle ipotesi

◆ ***Omoschedasticità*** Il grafico dei residui rispetto a X_i consente di stabilire anche se l'ipotesi di omoschedasticità è soddisfatta. Nel grafico della Figura 9.11, la variabilità dei residui varia a seconda dei valori assunti da X, segno di una violazione dell'ipotesi di omoschedasticità: i residui sembrano disporsi a ventaglio, a indicare un aumento della variabilità all'aumentare dei valori di X. Nella Figura 9.10, invece, non si osservano delle differenze significative nella variabilità dei residui in corrispondenza di valori diversi di X: l'ipotesi di omoschedasticità sembra soddisfatta.

◆ ***Normalità*** L'analisi dei residui consente di verificare l'ipotesi di normalità degli errori. Si tratta di costruire la distribuzione di frequenza dei residui e di darne una rappresentazione grafica a mezzo dell'istogramma.

Considerando ancora l'esempio relativo alla catena di negozi di abbigliamento, nella Tabella 9.2 si riporta la distribuzione di frequenza dei residui, mentre il corrispondente istogramma è rappresentato nella Figura 9.12.

È chiaramente, difficile, valutare l'ipotesi di normalità della distribuzione degli errori alla luce di un campione di sole 14 osservazioni e questo indipendentemente dallo strumento a cui si ricorre (l'istogramma, il diagramma ramo-foglia, il diagramma scatola e baffi o il *normality plot*). Dalla Figura 9.12, possiamo solo osservare che la distribuzione dei dati, sebbene non sembri normale, non è particolarmente asimmetrica. Pertanto la robustezza della regressione rispetto agli

scostamenti dall'ipotesi di normalità e la dimensione limitata del campione fanno sì che il riconoscimento di una distribuzione degli errori solo approssimativamente normale non risulti preoccupante.

FIGURA 9.11
Violazione dell'ipotesi di omoschedasticità

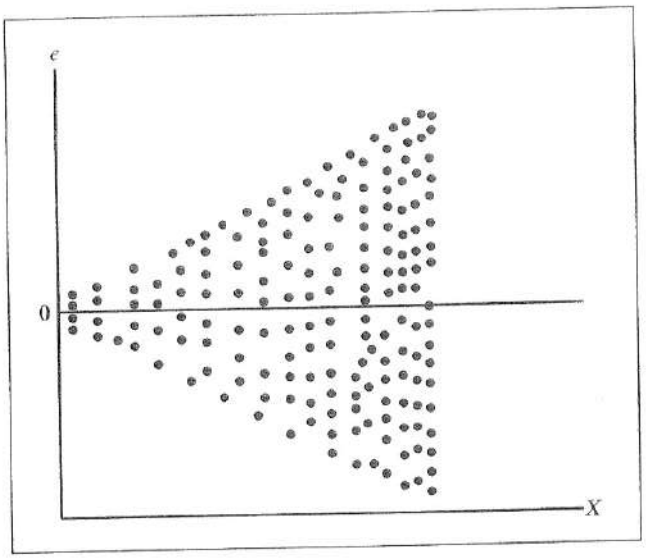
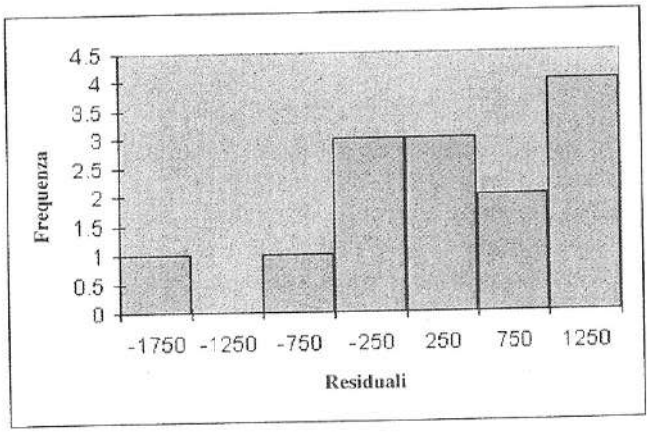


Tabella 9.2 *Distribuzione di frequenza dei residui per i dati relativi alla scelta della dimensione dei negozi*

RESIDUI	FREQUENZA
da −2250 a −1750	1
da −1750 a −1250	0
da −1250 a −750	1
da −750 a −250	3
da −250 a +250	3
da +250 a +750	2
da +750 a +1250	4
Totale	14

FIGURA 9.12
Istogramma dei residui ottenuto con Excel, per i dati relativi alla scelta della localizzazione dei negozi

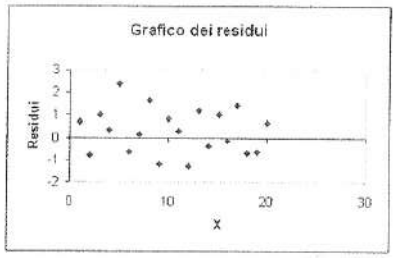


◆ ***Indipendenza*** L'ipotesi di indipendenza degli errori può essere verificata rappresentando i residui nell'ordine con cui i dati sono stati raccolti: nei dati raccolti nel corso del tempo è spesso presente un'*autocorrelazione* tra osservazioni successive. Il grafico dei residui rispetto al tempo consente allora di evidenziare la presenza di una relazione di questo genere. L'autocorrelazione dei residui viene misurata a mezzo della statistica di Durbin-Watson di cui ci occuperemo nel paragrafo 9.6.

Esercizi del paragrafo 9.5

• **9.10** Di seguito si riporta la tabella contenente i valori di X e dei residui ottenuti dalla stima di un modello di regressione, e il relativo grafico dei residui.

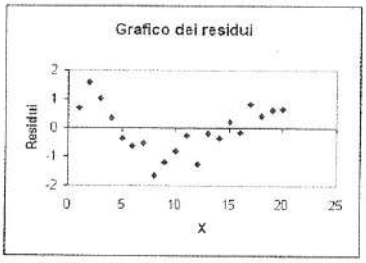
	X	Residui
2	1	0,7
3	2	-0,78
4	3	1,03
5	4	0,33
6	5	2,39
7	6	-0,67
8	7	0,16
9	8	1,65
10	9	-1,19
11	10	0,84
12	11	0,29
13	12	-1,28
14	13	1,21
15	14	-0,37
16	15	1,02
17	16	-0,16
18	17	1,42
19	18	-0,71
20	19	-0,63
21	20	0,67



È riconoscibile una struttura nei residui? Commentate.

9.11 Di seguito si riporta la tabella contenente i valori di X e dei residui ottenuti dalla stima di un modello di regressione, e il relativo grafico dei residui.

	X	Residui
2	1	0,7
3	2	1,58
4	3	1,03
5	4	0,33
6	5	-0,39
7	6	-0,67
8	7	-0,56
9	8	-1,65
10	9	-1,19
11	10	-0,84
12	11	-0,29
13	12	-1,28
14	13	-0,21
15	14	-0,37
16	15	0,22
17	16	-0,16
18	17	0,82
19	18	0,41
20	19	0,63
21	20	0,67
22		



È riconoscibile una struttura nei residui? Commentate.



DATASET
PETFOOD

• **9.12** Tornate all'esercizio 9.2 e conducete un'analisi dei residui. Alla luce dei risultati ottenuti:
(a) valutate la capacità di adattamento ai dati del modello;
(b) verificate se le ipotesi alla base del modello di regressione sono soddisfatte.

DATASET MOVIE

DATASET RENT

9.13 Tornate all'esercizio 9.3 e conducete un'analisi dei residui. Alla luce dei risultati ottenuti:
(a) valutate la capacità di adattamento ai dati del modello;
(b) verificate se le ipotesi alla base del modello di regressione sono soddisfatte:

9.14 Tornate all'esercizio 9.4 e conducete un'analisi dei residui. Alla luce dei risultati ottenuti:
(a) valutate la capacità di adattamento ai dati del modello;
(b) verificate se le ipotesi alla base del modello di regressione sono soddisfatte:

9.6 L'AUTOCORRELAZIONE E LA STATISTICA DI DURBIN-WATSON

Come abbiamo osservato nel paragrafo precedente, l'ipotesi di indipendenza degli errori spesso è violata quando i dati sono raccolti nel corso del tempo. In questo caso, infatti, i residui che si riferiscono a un certo periodo di tempo tendono a essere simili ai residui che si riferiscono a periodi di tempo adiacenti. In presenza di un andamento di tale genere si dice che tra i residui vi è **autocorrelazione** e un'elevata autocorrelazione dei residui può pregiudicare seriamente la validità del modello di regressione.

Il grafico dei residui come strumento per evidenziare l'autocorrelazione

Come abbiamo visto nel paragrafo 9.5, il grafico dei residui rispetto al tempo è lo strumento più semplice per individuare la presenza di autocorrelazione tra i residui. In presenza di un'autocorrelazione positiva, ad esempio, nel grafico si evidenzieranno gruppi di residui dello stesso segno, indice della presenza di un legame di dipendenza tra gli stessi.

Supponete che il manager di un negozio che si occupa di consegne a domicilio intenda prevedere l'ammontare delle entrate settimanali sulla base del numero dei clienti. La Tabella 9.3 riporta i dati relativi al numero dei clienti e all'ammontare delle entrate per un periodo di 15 settimane.

DATASET CUSTSALE

Tabella 9.3 *Numero di clienti e ammontare delle entrate per un periodo di 15 settimane*

SETTIMANA	CLIENTI	ENTRATE ($ 000)	SETTIMANA	CLIENTI	ENTRATE ($ 000)
1	794	9.33	9	880	12.07
2	799	8.26	10	905	12.55
3	837	7.48	11	886	11.92
4	855	9.08	12	843	10.27
5	845	9.83	13	904	11.80
6	844	10.09	14	950	12.15
7	863	11.01	15	841	9.64
8	875	11.49			

L'output di Excel relativo a tale insieme di dati è, invece, riportato nella Figura 9.13. R^2 è uguale a 0.657 a indicare che il 65.7% della variabilità delle entrate risulta spiegato in base al modello di regressione dal numero dei clienti; l'intercetta b_0 è uguale a −16.032 e l'inclinazione b_1 è uguale a 0.03076. Tuttavia per poter accettare tali risultati e eventualmente poter prevedere alla luce di essi l'ammontare delle entrate, dobbiamo procedere a un'attenta analisi dei residui per verificare l'ipotesi di indipendenza degli errori. Poiché i dati sono stati raccolti nel corso del tempo, è infatti possibile, alla luce di quanto osservato precedentemente, che i residui siano autocorrelati. Nella Figura 9.14, si riporta il grafico dei residui rispetto al tempo: i residui presentano un'oscillazione ciclica verso l'alto e verso

	A	B	C	D	E	F	G
1	Analisi di regressione per il negozio di consegna a domicilio						
2							
3	*Statistica della regressione*						
4	R multiplo	0,810829997					
5	R al quadrato	0,657445284					
6	R al quadrato corretto	0,631094922					
7	Errore standard	0,936036681					
8	Osservazioni	15					
9							
10	ANALISI VARIANZA						
11		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*	
12	Regressione	1	21,860433	21,86043264	24,95014171	0,000245105	
13	Residuo	13	11,390141	0,876164669			
14	Totale (b_0)	14	33,250573				
15							
16		*Coefficienti*	*rore standa.*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
17	Intercetta (b_1)	-16,0321936	5,3101671	-3,019150493	0,009868641	-27,50410993	-4,560277262
18	Customers	0,030760228	0,0061582	4,995011683	0,000245105	0,017456271	0,044064185

FIGURA 9.13 Output di Excel relativo ai dati della Tabella 9.3

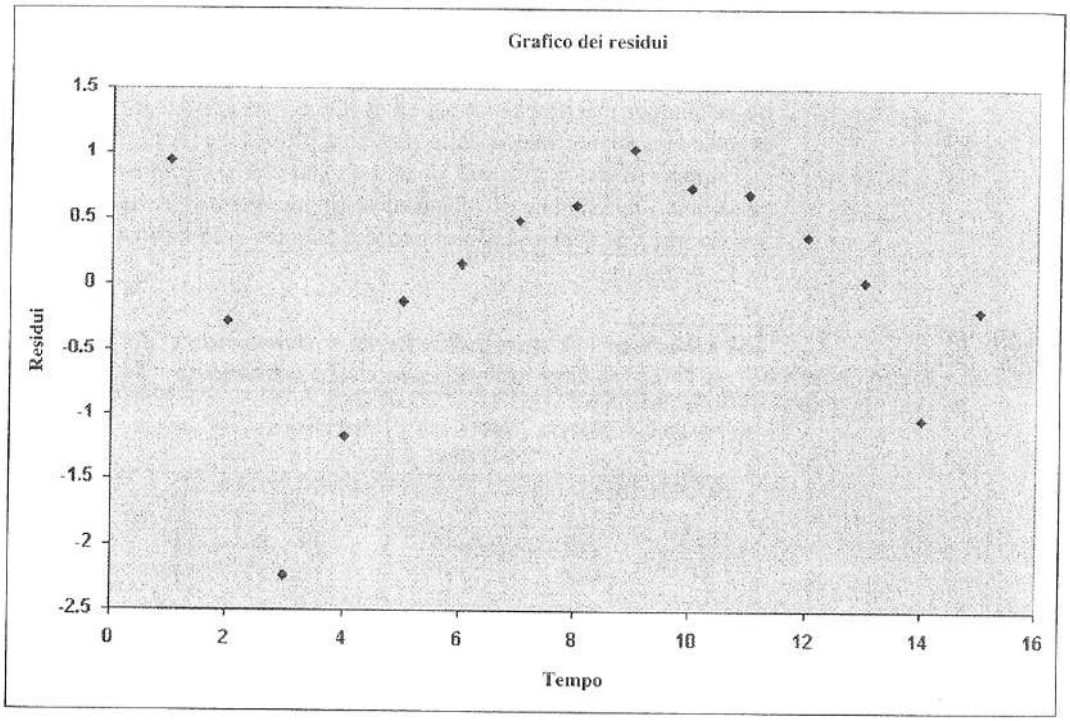


FIGURA 9.14 Grafico dei residui relativo ai dati della Tabella 9.3

il basso, che fa pensare alla presenza di una certa autocorrelazione tra di essi e quindi alla violazione dell'ipotesi di indipendenza degli errori.

La statistica di Durbin-Watson

L'autocorrelazione dei residui può essere individuata e misurata facendo ricorso a una particolare statistica campionaria, la **statistica di Durbin-Watson**, che misura la correlazione tra ciascun residuo e quello che lo precede.

La statistica di Durbin-Watson

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad (9.10)$$

dove e_i = residuo relativo al periodo di tempo i

Il numeratore della statistica D, $\sum_{i=2}^{n}(e_i - e_{i-1})^2$, è dato dalla somma dei quadrati delle differenze tra ciascun residuo e quello precedente, mentre il denominatore, $\sum_{i=1}^{n} e_i^2$, coincide con la somma dei quadrati di tutti i residui. In presenza di un'autocorrelazione positiva dei residui, D assume un valore vicino a 0, mentre in assenza di autocorrelazione il valore di D è vicino a 2 (se i residui presentano un'autocorrelazione negativa, D assume un valore maggiore di 2 e può anche raggiungere il suo valore massimo, 4). Nella Figura 9.15 si riporta l'output relativo al calcolo della statistica di Durbin-Watson, ottenuto con l'aggiunta PHStat. Il valore della statistica D calcolata con riferimento ai dati riportati nella Tabella 9.3 è uguale a:

$$D = \frac{10.058}{11.39} = 0.883$$

FIGURA 9.15
Output dell'aggiunta PHStat con la statistica di Durbin-Watson relativa ai dati della Tabella 9.3

	A	B	C
1	Analisi di regressione per il negozio di consegna a domicilio		
2			
3	Sum of Squared Difference of Residuals	10,05752	
4	Sum of Squared Residuals	11,39014	
5	Durbin-Watson Statistic	0,883003	
6			

Si tratta allora di stabilire per quali valori di D possiamo concludere che vi è autocorrelazione tra i residui e che, pertanto, l'ipotesi di indipendenza degli errori è violata. Tali valori dipendono da n, il numero delle osservazioni, e da p, il numero di variabili indipendenti nel modello (nel modello lineare semplice $p = 1$). Nella Tabella 9.4, si riporta la tavola dei valori critici della statistica di Durbin-Watson (Tavola E.7).

Tabella 9.4 *I valori critici della statistica di Durbin-Watson*

$\alpha = 0.05$

	$p = 1$		$p = 2$		$p = 3$		$p = 4$		$p = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06

Fonte: Tavola E.7.

Nella Tabella 9.4 si riportano, per ciascuna combinazione di α (il livello di significatività), n (la dimensione del campione) e p (il numero delle variabili indipendenti), due valori della statistica D. Il primo d_L è il valore critico inferiore di D, cioè il più piccolo dei valori di D corrispondenti a una situazione di assenza di autocorrelazione dei residui: se D è minore di d_L, concludiamo che vi è prova della presenza di un'autocorrelazione positiva tra i residui. Il secondo valore d_U è, invece, il valore critico superiore di D: se D è maggiore di d_U, concludiamo che non vi è prova della presenza di un'autocorrelazione positiva tra i residui. Se D è compreso tra d_L e d_U non possiamo giungere a nessuna conclusione.

Pertanto per i dati della Tabella 9.3, siccome vi è una sola variabile esplicativa ($p = 1$) e si considerano 15 osservazioni ($n = 15$), avremo in base alla Tabella 9.4, $d_L = 1.08$ e $d_U = 1.36$. Siccome $D = 0.883 < 1.08$, possiamo concludere che vi è autocorrelazione tra i residui. L'analisi di regressione condotta risulta inappropriata e si rende necessario ricorrere a approcci diversi all'analisi della relazione tra variabili (riferimento bibliografico 6).

Esercizi del paragrafo 9.6

• **9.15** La tabella seguente riporta i valori dei residui per dati raccolti nel corso di 10 periodi di tempo:

Periodo di tempo	Residui	Periodo di tempo	Residui
1	−5	6	+1
2	−4	7	+2
3	−3	8	+3
4	−2	9	+4
5	−1	10	+5

(a) Rappresentate graficamente i residui rispetto al tempo. Commentate.
(b) Calcolate la statistica di Durbin-Watson.
(c) Alla luce delle risposte ai punti (a) e (b) a quali conclusioni si può pervenire in merito all'autocorrelazione dei residui?



Dataset PETFOOD

• **9.16** Riprendete l'esercizio 9.2.
(a) È necessario in questo caso calcolare la statistica di Durbin-Watson?
(b) In che caso sarebbe necessario calcolare la statistica di Durbin-Watson, prima di procedere alla stima con il metodo dei minimi quadrati del modello di regressione?



Dataset ELECUSE

9.17 Il proprietario di una casa monofamiliare, in una zona residenziale nel Nord-Est degli Stati Uniti, intende fare ricorso a un modello statistico per prevedere il consumo complessivo di elettricità sulla base della temperatura atmosferica (misurata in gradi Fahrenheit, °F). Nella tabella seguente si riportano i valori dei chilowatt consumati e della temperatura relativi a un periodo di 24 mesi.
(a) Disegnate il diagramma di dispersione per i dati della tabella.
(b) Nell'ipotesi che tra le due variabili sussista una relazione lineare, stimate con il metodo dei minimi quadrati i coefficienti di regressione b_0 e b_1.
(c) Fornite un'interpretazione di b_1.
(d) Prevedete il consumo medio in chilowatt corrispondente a una temperatura media di 50 gradi Fahrenheit.
(e) Calcolate il coefficiente di determinazione r_2 e interpretatene il significato.
(f) Calcolate l'errore standard della stima.
(g) Rappresentate graficamente i residui rispetto alla temperatura atmosferica media.
(h) Rappresentate graficamente i residui rispetto al tempo.
(i) Calcolate la statistica di Durbin-Watson. Per $\alpha = 0.05$, si può ritenere che vi sia autocorrelazione tra i residui?

(i) Calcolate la statistica di Durbin-Watson. Per $\alpha = 0.05$, si può ritenere che vi sia autocorrelazione tra i residui?
(j) Alla luce delle risposte ai punti (g)-(i), si deve mettere in discussione la validità del modello?
(k) Supponete che l'ammontare delle vendite del giorno 21 sia 1.75. Rispondete ai punti (a)-(j) e commentate le differenze rispetto ai risultati precedentemente ottenuti.

9.7 Inferenza sull'inclinazione della retta di regressione

Nei paragrafi precedenti, il modello di regressione è stato impiegato con uno scopo principalmente descrittivo. In questa ottica, abbiamo introdotto il metodo dei minimi quadrati per stimare i coefficienti della retta di regressione e, quindi, per poter prevedere Y sulla base dei valori di X. Abbiamo anche definito l'errore standard della stima e il coefficiente di determinazione.

Infine abbiamo visto come attraverso un'analisi dei residui, si possa verificare se le ipotesi alla base del modello di regressione sono soddisfatte e quindi se il modello stesso è appropriato. Passiamo ora a considerare le inferenze che si possono condurre sull'esistenza di una relazione lineare tra le variabili di una popolazione alla luce dei risultati campionari.

Il test *t* sull'inclinazione della retta

Possiamo stabilire se tra le variabili X e Y sussiste una relazione lineare significativa sottoponendo a verifica l'ipotesi che b_1 (l'inclinazione della popolazione) sia uguale a 0. Nel caso di rifiuto di tale ipotesi, possiamo concludere che vi è prova dell'esistenza di una relazione lineare tra le variabili considerate. L'ipotesi nulla e l'ipotesi alternativa sono, allora, rispettivamente:

H_0: $\beta_1 = 0$ (non vi è una relazione lineare)
H_1: $\beta_1 \neq 0$ (vi è una relazione lineare)

La statistica test è di seguito definita.

Il test *t* per la verifica di ipotesi sull'inclinazione β_1

La statistica t è data dalla differenza tra l'inclinazione campionaria e l'inclinazione ipotizzata della popolazione, il tutto diviso per l'errore standard dell'inclinazione.

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

dove (9.11)

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SQX}}$$

$$SQX = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

La statistica t ha una distribuzione t di Student con $n - 2$ gradi di libertà.

Torniamo all'esempio relativo alla catena di negozi di abbigliamento e verifichiamo se sussiste una relazione lineare significativa tra la dimensione del negozio e l'ammontare

annuo delle vendite, per un livello di significatività uguale a 0.05. Dall'output di Excel della Figura 9.4, ricaviamo[2]:

$$b_1 = +1.686 \qquad n = 14 \qquad S_{b_1} = 0.1533$$

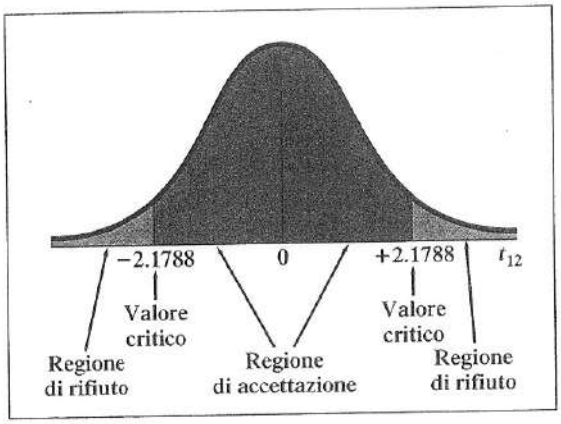
Pertanto il valore della statistica t è dato da:

$$t = \frac{b_1}{S_{b_1}} = \frac{1.686}{0.1533} = 11.00$$

Il valore della statistica t è riportato nell'output di Excel nella colonna **Stat** t. Poiché $t = 11 > t_{12} = 2.1788$, rifiutiamo H_0. In base all'approccio del p-value rifiutiamo H_0 perché il p-value è approssimativamente uguale a 0. (Nell'output di Excel il valore del p-value è riportato nella colonna *Significatività* con la notazione scientifica 1.27E – 07, che corrisponde a 0.000000127 e pertanto è inferiore a $\alpha = 0.05$). Possiamo, quindi, concludere che esiste una relazione lineare significativa tra l'ammontare medio annuo delle vendite e la dimensione del negozio. (Figura 9.16).

[2] *Dettagli ulteriori sul calcolo della statistica t sono dati nel paragrafo 9.10*

FIGURA 9.16
Verifica di ipotesi sull'inclinazione della retta di regressione, con $\alpha = 0.05$ e 12 gradi di libertà



Il test *F* per l'inclinazione

La significatività dell'inclinazione della retta di regressione può essere sottoposta a verifica anche ricorrendo al test F (Tabella 9.5). Nel paragrafo 7.2 abbiamo visto che il test F ha per oggetto il rapporto tra due varianze. Nel verificare la significatività dell'inclinazione, si impiega come misura dell'errore casuale la varianza dei residui (data dalla somma dei quadrati degli errori divisa per il numero dei gradi di libertà). Pertanto il test F è dato dal rapporto tra la varianza dovuta alla regressione (la somma dei quadrati della regressione divisa per il numero delle variabili indipendenti) e la varianza dei residui (equazione 9.12).

Il test *F* per la verifica di ipotesi sull'inclinazione β_1

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

$$F = \frac{MQR}{MQE} \tag{9.12}$$

dove

$$MQR = \frac{SQR}{p}$$

$$MQE = \frac{SQE}{n - p - 1}$$

Se α è il livello di significatività scelto, si ottiene la seguente regola decisionale:

> Rifiuta H_0
> se $F > F_U$ il valore critico che si trova nella coda
> di sinistra della distribuzione F con $n - p - 1$ gradi di libertà;
> altrimenti accetta H_0.

I risultati dell'analisi possono essere riportati in maniera sintetica mediante la tabella dell'analisi della varianza (ANOVA), contenuta nella Tabella 9.5.

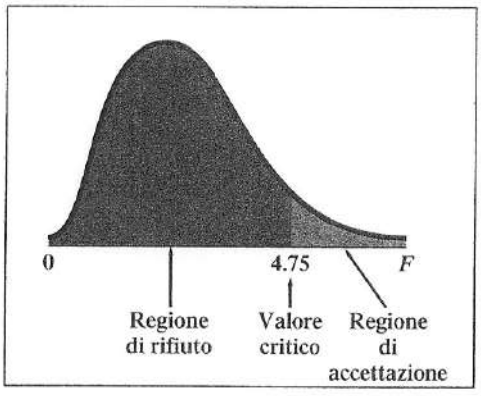
Tabella 9.5 *Tabella dell'ANOVA per la verifica della significatività del coefficiente di regressione*

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		

L'output di Excel relativo alla regressione contiene la tabella completa dell'ANOVA, come possiamo vedere dalla Figura 9.4. La statistica F in questo caso è uguale a 121.01 e il p-value è inferiore a 0.001.

Se $\alpha = 0.05$, il valore critico della statistica F con 1 e 12 gradi di libertà si ricava dalla Tavola E.5 ed è uguale a 4.75, come illustrato nella Figura 9.17. Rifiutiamo l'ipotesi H_0 perché $F = 121.01 > 4.75$ o perché il p-value $= 0.000000127 < 0.05$ e quindi possiamo concludere che sussiste una relazione significativa tra la dimensione del negozio e l'ammontare annuo delle vendite.

Figura 9.17
Verifica della significatività dell'inclinazione, con $\alpha = 0.05$ e 1 e 12 gradi di libertà



L'intervallo di confidenza per l'inclinazione β_1

L'esistenza di una relazione lineare tra le variabili considerate può essere valutata anche costruendo un intervallo di confidenza per β_1 e verificando se il valore ipotizzato ($\beta_1 = 0$) vi è compreso.

L'intervallo di confidenza per l'inclinazione

L'intervallo di confidenza per β_1 si ottiene addizionando e sottraendo all'inclinazione campionaria b_1 il prodotto tra il valore critico della statistica t e l'errore standard dell'inclinazione.

$$b_1 \pm t_{n-2} S_{b_1} \tag{9.13}$$

Dall'output di Excel della Figura 9.4, ricaviamo:

$$b_1 = +1.686 \qquad n = 14 \qquad S_{b_1} = 0.1533$$

Pertanto

$$\begin{aligned} b_1 \pm t_{n-2} S_{b_1} &= +1.686 \pm (2.1788)(0.1533) \\ &= +1.686 \pm 0.334 \\ +1.352 \leq \beta_1 &\leq +2.02 \end{aligned}$$

L'intervallo di confidenza per l'inclinazione per un livello di confidenza del 95% è compreso tra gli estremi +1.352 e +2.02. Poiché 0 non è compreso nell'intervallo, concludiamo che l'ammontare delle vendite è legato da una relazione lineare significativa alla dimensione del negozio.

Esercizi del paragrafo 9.7

• **9.19** Intendete verificare la significatività dell'inclinazione della retta di regressione. A tale scopo estraete un campione di ampiezza $n = 18$ e ottenete i seguenti risultati:

$$b_1 = +4.5 \qquad S_{b_1} = 1.5$$

(a) Calcolate il valore della statistica t.
(b) Determinate i valori critici della statistica per $\alpha = 0.05$.
(c) Alla luce dei punti (a) e (b), a quale decisione statistica dovreste pervenire?
(d) Costruite un intervallo di confidenza di livello 0.95 per l'inclinazione β_1.



DATASET PETFOOD

• **9.20** Riprendete l'output di Excel ottenuto per risolvere l'esercizio 9.2.
(a) Per $\alpha = 0.05$, si può ritenere che tra l'ammontare delle vendite di cibo per animali e lo spazio destinato sugli scaffali al prodotto sussista una relazione lineare significativa?
(b) Costruite un intervallo di confidenza di livello 95% per l'inclinazione β_1.



DATASET MOVIE

9.21 Riprendete l'output di Excel ottenuto per risolvere l'esercizio 9.3.
(a) Per $\alpha = 0.05$, si può ritenere che tra il successo di botteghino e le vendite di videocassette sussista una relazione lineare significativa?
(b) Costruite un intervallo di confidenza di livello 95% per l'inclinazione β_1.



DATASET RENT

9.22 Riprendete l'output di Excel ottenuto per risolvere l'esercizio 9.4.
(a) Per $\alpha = 0.05$, si può ritenere che tra la dimensione degli appartamenti e l'ammontare degli affitti sussista una relazione lineare significativa?
(b) Costruite un intervallo di confidenza di livello 95% per l'inclinazione β_1.

9.23 La volatilità delle azioni viene spesso misurata facendo ricorso all'indice beta. Tale indice si ottiene stimando un modello di regressione in cui la variabile dipendente è la variazione settimanale dell'azione e la variabile indipendente è la variazione settimanale di un indice di mercato. In genere l'indice di mercato utilizzato è l'indice S&P 500. Pertanto se si volesse calcolare l'indice beta per le azioni dell'IBM, dovremmo stimare il seguente modello di regressione:

$$\begin{aligned} &(\text{variazione \% settimanale di IBM}) = \\ &\beta_0 + \beta_1(\text{variazione \% settimanale di S\&P 500}) + \epsilon \end{aligned}$$

Beta coincide con lo stimatore dei minimi quadrati, b_1. Se l'indice beta è uguale a 1, l'azione tende a presentare la medesima variabilità del mercato; se beta è uguale a 1.5, l'azione tende a presentare una variabilità pari a una volta e mezzo quella del mercato; infine se beta è uguale 0.6, l'azione ha una variabilità uguale al 60% di quella del mercato. Nella tavole seguente si riportano i valori dell'indice beta per le azioni di alcune società.

SOCIETÀ	BETA
Ford Motor Company	0.92
Houston Industries	0.43
IBM	1.09
LSI Logic	1.80

Fonte: Standard & Poor'50 Guide, Edition 1999 (McGraw-Hill Inc. 1999)

(a) Spiegate il significato dell'indice beta per ciascuna società.
(b) L'investitore può essere guidato nelle sue scelte dal valore dell'indice beta?

9.8 LA STIMA DELLA PREVISIONE

In questo paragrafo, introduciamo i metodi che consentono di fare inferenza sulla media di Y e su un suo singolo valore Y_I.

L'intervallo di confidenza per la risposta media

Nell'esempio relativo alla catena di negozi di abbigliamento, in base al modello di regressione abbiamo ottenuto la previsione delle vendite (7645.786) per un negozio di 4000 piedi al quadrato: tale valore è una *stima puntuale* della media della popolazione delle vendite. In questo paragrafo, introduciamo **l'intervallo di confidenza per la media della variabile risposta (la risposta media).**

L'intervallo di confidenza per $\mu_{Y|X_i}$ la media di Y

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{h_i} \tag{9.14}$$

dove

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$\hat{Y}_i$ = previsione del valore medio Y: $\hat{Y}_i = b_0 + b_1 X_i$

S_{YX} = errore standard della stima

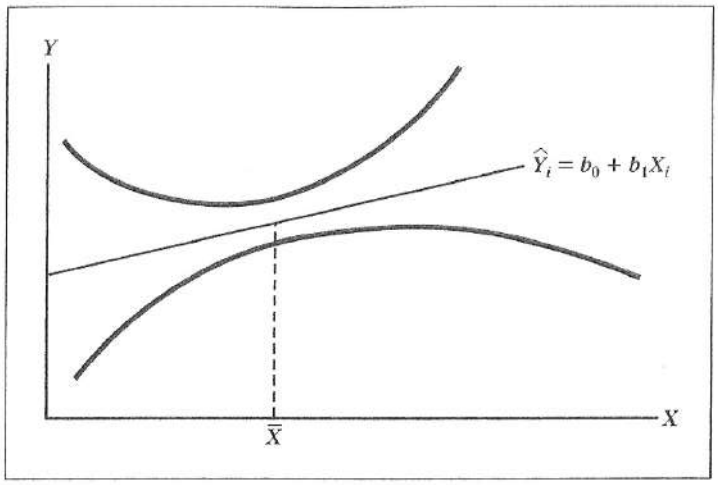
n = ampiezza del campione

X_i = i–esimo del valore di X

In base all'equazione (9.14), l'ampiezza dell'intervallo risulta dipendere da diversi fattori. Per un dato livello di confidenza, un aumento della variabilità attorno alla retta di regressione, misurata dall'errore standard della stima, dà luogo a un aumento della dimensione dell'intervallo, un aumento dell'ampiezza del campione ne comporta invece una riduzione. Osserviamo, inoltre, che l'ampiezza dell'intervallo dipende anche dai valori di X.

L'ampiezza dell'intervallo è piccola in prossimità del valor medio $\overline{X}$ e aumenta mano a mano che ci si allontana da questo. Tale effetto è dovuto alla presenza della radice quadrata del rapporto nell'equazione (9.14) ed è illustrato dalla Figura 9.18.

FIGURA 9.18
Intervallo di confidenza per μ_{YX}



Esempio 9.3 *Costruzione di intervallo di confidenza del 95% per la risposta media μ_{YX}*

Per il problema introdotto nell'Applicazione, relativo alla catena di negozi di abbigliamento, abbiamo ottenuto la seguente stima della retta di regressione $\hat{Y}_i = 901.247 + 1.686\, X_i$. Costruite un intervallo di confidenza del 95% per la media delle vendite di tutti i negozi di 4000 piedi al quadrato.

SOLUZIONE

Dalla stima della retta di regressione:

$$\hat{Y}_i = 901.247 + 1.686\, X_i$$

e per $X_i = 4{,}000$, otteniamo

$$\hat{Y}_i = 901.247 + 1.686(4000) = 7645.786$$

Inoltre

$$\overline{X} = 2921.2857; \qquad S_{YX} = 936.85; \qquad \sum_{i=1}^{n} (X_i - \overline{X})^2 = 37\,357\,090.86$$

e dalla Tavola E.3, $t_{12} = 2.1788$. Pertanto,

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{h_i}$$

dove

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

di modo che

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

e

$$7645.786 \pm (2.1788)(936.85)\sqrt{\frac{1}{14} + \frac{(4{,}000 - 2921.2857)^2}{37\,357\,090.86}}$$

$$= 7\,645\,786 \pm 653.756$$

quindi

$$6992.031 \leq \mu_{YX} \leq 8299.542$$

Pertanto possiamo concludere che la media delle vendite settimanali per i negozi di 4000 piedi al quadrato è compresa tra 6992.031 e 8299.542 migliaia di dollari. Tale risultato poteva essere facilmente ottenuto con l'aggiunta PHStat.

L'intervallo di confidenza per la previsione

L'intervallo di confidenza per la previsione ha una forma simile a quella dell'intervallo di confidenza per la risposta media, sebbene in questo caso si stimi un singolo valore e non un parametro. L'**intervallo di confidenza per la previsione di un singolo valore Y_i (di una singola risposta)** è dato dall'equazione 9.15.

Esempio 9.4 *Costruzione di intervallo di confidenza del 95% per la previsione*

Per il problema introdotto nell'Applicazione, relativo alla catena di negozi di abbigliamento, abbiamo ottenuto la seguente stima della retta di regressione $\hat{Y}_i = 901.247 + 1.686\,X_i$. Costruite un intervallo di confidenza del 95% per la previsione delle vendite di un negozio di 4000 piedi al quadrato.

L'intervallo di confidenza per la previsione di una singola risposta Y_i

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{1 + h_i} \tag{9.15}$$

dove

h_i, $\hat{Y}_i$, S_{YX}, n e X_i sono definiti come nell'equazione 9.15

Soluzione

Dalla stima della retta di regressione:

$$\hat{Y}_i = 901.247 + 1.686\,X_i$$

e poi $X_i = 4000$, otteniamo

$$\hat{Y}_i = 901.247 + 1.686(4000) = 7645.786$$

Inoltre

$$\bar{X} = 2921.2857; \qquad S_{YX} = 936.85; \qquad \sum_{i=1}^{n} (X_i - \bar{X})^2 = 37\,357\,090.86$$

e dalla tavola E.3, $t_{12} = 2.1788$. Pertanto,

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{1 + h_i}$$

dove:

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

di modo che

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

e

$$7645.786 \pm (2.1788)(936.85)\sqrt{1 + \frac{1}{14} + \frac{(4{,}000 - 2921.2857)^2}{37\,357\,090.86}}$$

$$= 7645.786 \pm 2143.357$$

quindi

$$5502.43 \leq Y_I \leq 9789.143$$

Pertanto possiamo concludere che l'ammontare delle vendite settimanali per un negozio di 4000 piedi al quadrato è compreso tra 5502.43 e 9789.143 migliaia di dollari.

L'esempio 9.4 illustra il calcolo dell'intervallo di confidenza per la previsione di una singola risposta con riferimento ai dati relativi alla catena di negozi di abbigliamento, anche se tale intervallo può essere ottenuto facilmente con l'aggiunta PHStat.

Confrontando i risultati degli esempi 9.3 e 9.4, osserviamo che l'ampiezza dell'intervallo per la previsione di una singola risposta è maggiore dell'ampiezza dell'intervallo della risposta media e questo perché nella previsione di un valore singolo vi è senz'altro una variabilità maggiore di quella che accompagna la previsione di un valore medio.

Esercizi del paragrafo 9.8

9.24 Per un campione di 20 osservazioni, si ottiene con il metodo dei minimi quadrati la seguente retta di regressione: $\hat{Y}_i = 5 + 3X_i$ inoltre, $S_{YX} = 1.0$, $\overline{X} = 2$, e $\sum_{i=1}^{n}(X_i - \overline{X})^2 = 20$.

(a) Costruite un intervallo di confidenza del 95% per la risposta media corrispondente a $X = 2$.

(b) Costruite un intervallo di confidenza del 95% per il valore della singola risposta corrispondente a $X = 2$.



DATASET PETFOOD

• **9.25** Riprendete l'output di Excel ottenuto per risolvere l'esercizio 9.2.

(a) Costruite un intervallo di confidenza del 95% per la media delle vendite settimanali di tutti i negozi in cui i cibi per animali occupano uno spazio di 8 piedi.

(b) Costruite un intervallo di confidenza del 95% per le vendite settimanali di un negozio in cui i cibi per animali occupano uno spazio di 8 piedi.

(c) Spiegate le differenze tra i risultati dei due punti.



DATASET MOVIE

9.26 Riprendete l'output di Excel ottenuto per risolvere l'esercizio 9.3.

(a) Costruite un intervallo di confidenza del 95% per le vendite medie delle videocassette di tutti film per i quali sono stati acquistati biglietti per 10 milioni di dollari.

(b) Costruite un intervallo di confidenza del 95% per le vendite delle videocassette di un film di cui sono stati acquistati biglietti per 10 milioni di dollari.

(c) Spiegate le differenze tra i risultati dei due punti.

DATASET
RENT

9.27 Riprendete l'output di Excel ottenuto per risolvere l'esercizio 9.4.
(a) Costruite un intervallo di confidenza del 95% per l'affitto medio mensile di tutti gli appartamenti con una dimensione di 1000 piedi al quadrato.
(b) Costruite un intervallo di confidenza del 95% per l'affitto mensile di un appartamento con una dimensione di 1000 piedi al quadrato.
(c) Spiegate le differenze tra i risultati dei due punti.

9.9 LE TRAPPOLE DELL'ANALISI DI REGRESSIONE

Il modello di regressione si presenta come una tecnica statistica a cui si ricorre ampiamente nelle applicazioni economiche. Tuttavia, spesso viene impiegato in maniera non corretta. Il riquadro seguente illustra le difficoltà di cui si deve tener conto nell'impiego del modello di regressione.

Riquadro 9.2 Le difficoltà del modello di regressione

✓ **1.** Scarsa conoscenza delle assunzioni alla base del modello.
✓ **2.** Scarsa conoscenza del modo in cui valutare tali assunzioni.
✓ **3.** Scarsa conoscenza dei modelli alternativi a quello di regressione lineare semplice.
✓ **4.** Uso del modello di regressione senza una conoscenza adeguata della teoria sottostante.

L'ampia diffusione di fogli elettronici e pacchetti statistici ha facilitato l'uso delle tecniche di regressione per la previsione. Tuttavia, l'uso sempre maggiore di tali tecniche non si è accompagnato a una conoscenza adeguata delle stesse.

Di seguito, discuteremo un esempio classico nella letteratura statistica con cui si illustra la necessità di andare oltre i primi risultati della stima della retta di regressione (la stima dei coefficienti e il calcolo di r^2) ottenibili facilmente a mezzo dei moderni pacchetti statistici, e di condurre delle analisi più approfondite, ricorrendo a tecniche quali l'analisi dei residui. Nella Tabella 9.6 si riportano i valori di X e di Y per quattro dataset costruiti ad hoc.

Anscombe (riferimento bibliografico 1) mostra che per tutti e quattro i dataset si ottengono i seguenti risultati:

$$SQR = \text{variabilità spiegata} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = 27.51$$

$$SQE = \text{variabilità non spiegata} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 13.763$$

$$SQT = \text{variabilità totale} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 41.273$$

Pertanto i quattro dataset sono equivalenti con riferimento queste statistiche della retta di regressione. L'analisi non si può quindi arrestare a questo punto, pena la perdita di alcune informazioni importanti. Prendiamo, allora, in considerazione i diagrammi di dispersione dei quattro dataset (Figura 9.19) e i relativi grafici dei residui (Figura 9.20).

I grafici evidenziano le differenze tra i quattro dataset. L'unico dataset per cui il modello di regressione lineare semplice sembra appropriato è il dataset A: i punti del diagramma di dispersione sembrano disporsi lungo una retta e i residui non presentano nessun andamento riconoscibile e nessun outlier. Al contrario, il diagramma di dispersione del dataset B sembra

Tabella 9.6 *Quattro insiemi di dati artificiali*

DATASET A		DATASET B		DATASET C		DATASET D	
X_i	Y_i	X_i	Y_i	X_i	Y_i	X_i	Y_i
10	8.04	10	9.14	10	7.46	8	6.58
14	9.96	14	8.10	14	8.84	8	5.76
5	5.68	5	4.74	5	5.73	8	7.71
8	6.95	8	8.14	8	6.77	8	8.84
9	8.81	9	8.77	9	7.11	8	8.47
12	10.84	12	9.13	12	8.15	8	7.04
4	4.26	4	3.10	4	5.39	8	5.25
7	4.82	7	7.26	7	6.42	19	12.50
11	8.33	11	9.26	11	7.81	8	5.56
13	7.58	13	8.74	13	12.74	8	7.91
6	7.24	6	6.13	6	6.08	8	6.89

Fonte: "Graphs in Statistical Analysis," F.J. Anscombe, *The American Statistician* 27 (1973); 17–21.
Copyright © 1973 The American Statistical Association.

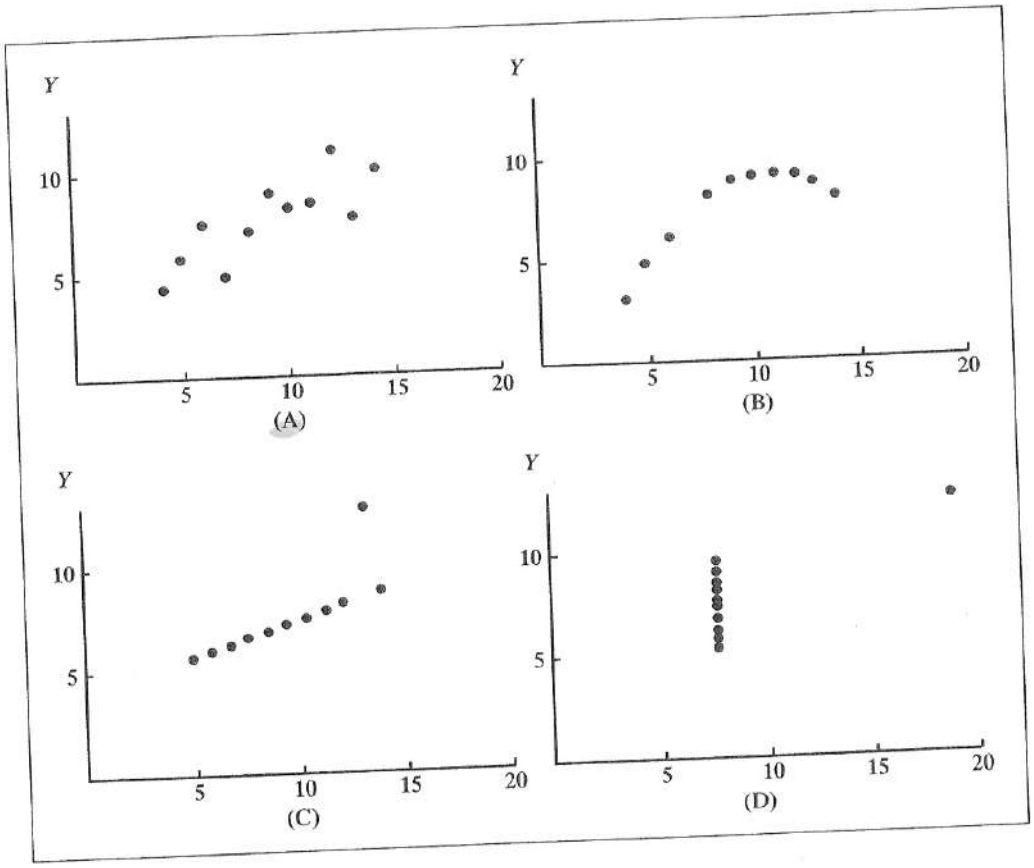


FIGURA 9.19 Diagrammi di dispersione per i quattro dataset

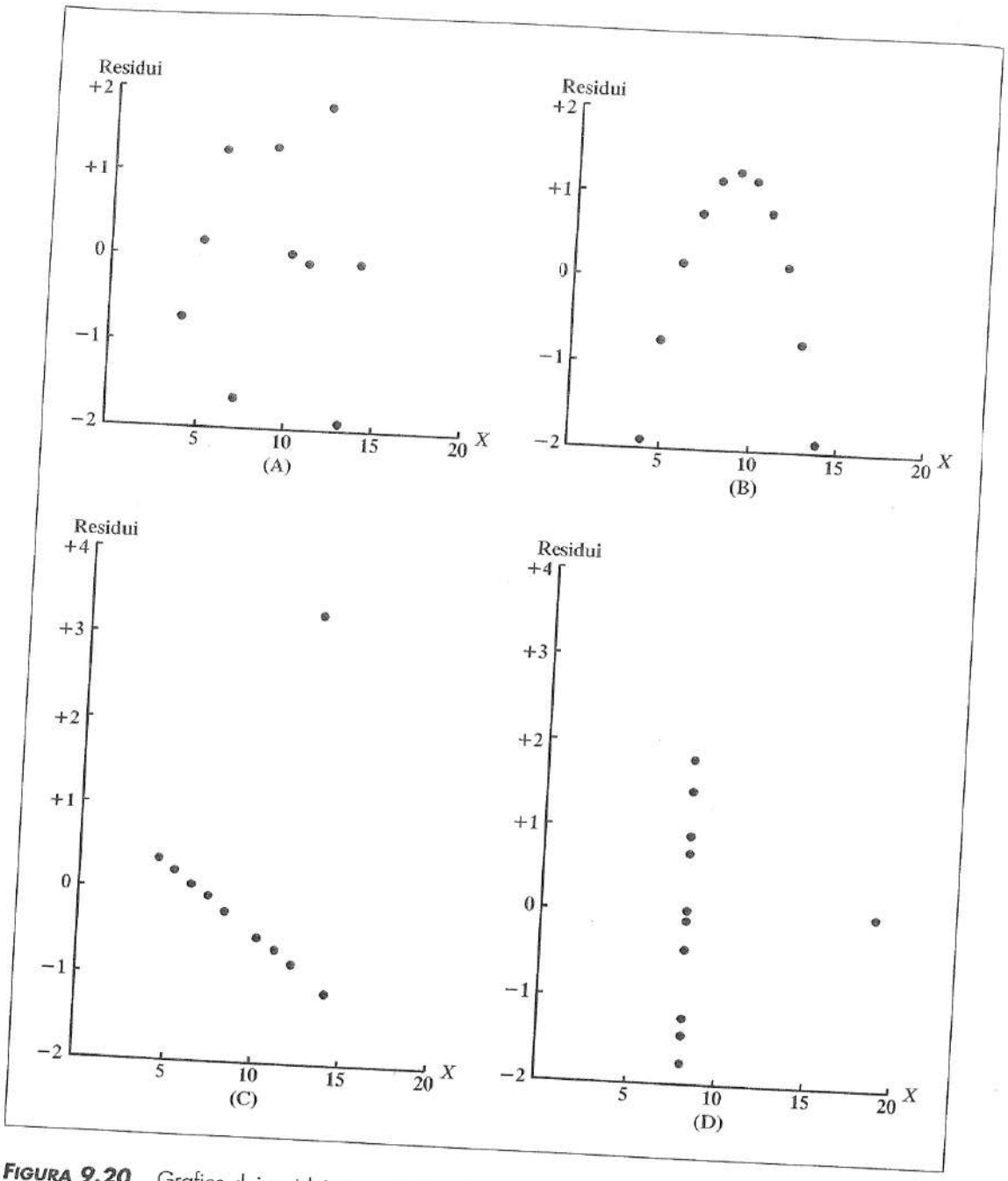


FIGURA 9.20 Grafico dei residui per i quattro dataset

Fonte: Anscombe, 1973, *The American Statistician*, Copyright © 1973 by The American Statistical Association.

suggerire l'adozione di un modello di regressione non lineare, conclusione che peraltro il grafico dei residui rafforza. Il diagramma di dispersione del dataset C evidenzia la presenza di un outlier, che si dovrebbe eliminare prima di procedere alla stima del modello. Analogamente in base al diagramma del dataset D, sembra che la stima della retta di regressione sia estremamente influenzata da un'osservazione (l'ottava osservazione per la quale $X_8 = 19$ e $Y_8 = 12.50$), situazione di cui si deve tener conto nella valutazione dei risultati dell'analisi di regressione.

Quest'esempio sottolinea l'importanza dell'analisi dei residui, che si deve sempre accompagnare alla stima del modello. Nel Riquadro 9.3 si suggerisce una strategia da seguire per evitare le trappole dell'analisi di regressione evidenziate nel Riquadro 9.2.

Riquadro 9.3 Una strategia per evitare le trappole della regressione

✓ **1.** Cominciate l'analisi sempre con un'attenta osservazione del diagramma di dispersione, per cogliere l'eventuale relazione tra X e Y.

✓ **2.** Verificate se le ipotesi alla base del modello di regressione sono soddisfatte dopo la stima del modello e prima di passare a impiegarne i risultati.

✓ **3.** Rappresentate graficamente i residui rispetto alla variabile dipendente per stabilire se il modello si adatta ai dati e se l'ipotesi di omoschedasticità è rispettata.

✓ **4.** Usate l'istogramma, il diagramma ramo-foglia o il diagramma scatola e baffi dei residui per verificare in quale misura l'ipotesi di normalità degli errori è rispettata.

✓ **5.** Se i dati sono raccolti in ordine sequenziale, rappresentate graficamente i residui nell'ordine con cui i dati sono stati raccolti e calcolate la statistica di Dubin-Watson.

✓ **6.** Se alla luce dei punti 3-5 ritenete che le ipotesi alla base del modello di regressione lineare siano violate, ricorrete ad altri metodi di stima del modello o ad altri modelli.

✓ **7.** Se alla luce dei punti 3-5 ritenete che le ipotesi alla base del modello di regressione lineare non siano violate, potete procedere ad alcune inferenze sul modello. Sottoponete a verifica la significatività dei coefficienti e costruite gli intervalli di confidenza per la risposta media e per la previsione.

9.10 I CALCOLI DELLA REGRESSIONE LINEARE SEMPLICE

In questo paragrafo illustriamo i calcoli che consentono di ottenere le statistiche del modello di regressione prese in considerazione.

Il calcolo dell'intercetta b_0 e dell'inclinazione b_1

Il metodo dei minimi quadrati per la stima dei coefficienti della retta di regressione comporta la risoluzione del sistema dato dalle seguenti equazioni (9.16a) e (9.16b).

Equazioni da risolvere per applicare il metodo dei minimi quadrati

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i \tag{9.16a}$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \tag{9.16b}$$

Dalla risoluzione del sistema si ottengono i seguenti risultati:

Formula per il calcolo dell'inclinazione b_1

$$b_1 = \frac{SQXY}{SQX} \tag{9.17}$$

$\bar{X})$ dove

$$SQXY = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SQX = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$$

e

Formula per il calcolo dell'intercetta b_0

$$b_0 = \bar{Y} - b_1\bar{X} \tag{9.18}$$

dove

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \quad \text{e} \quad \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Alla luce delle equazioni (9.16) e (9.18), si rende necessario il calcolo di cinque quantità per determinare b_0 e b_1: n, l'ampiezza del campione, $\sum_{i=1}^{n} X_i$, la somma dei valori della X, $\sum_{i=1}^{n} Y_i$, $\sum_{i=1}^{n} X_i Y_i$, la somma dei quadrati dei valori della X e la somma dei prodotti dei valori della X e della Y.

Nella Tabella 9.7 si riportano i calcoli di queste quantità (con l'aggiunta di $\sum_{i=1}^{n} Y_i^2$, la somma dei quadrati dei valori della Y, necessaria per il calcolo della SQT) con riferimento ai dati sulla dimensione e sull'ammontare delle vendite per la catena di negozi di abbigliamento.

Calcoliamo b_0 e b_1 in base alle equazioni (9.16) e (9.18):

$$b_1 = \frac{SQXY}{SQX}$$

$$SQXY = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$= 301\,298\,822 - \frac{(81\,577)(40\,898)}{14}$$

$$= 301\,298\,822 - 238\,309\,724.71$$

$$= 62\,989\,097.29$$

Tabella 9.7 *Calcoli per il problema della dimensione dei negozi*

Negozio	Piedi al quadrato X	Vendite Y	X^2	Y^2	XY
1	1,726	3,681	2,979,076	13,549,761	6,353,406
2	1,642	3,895	2,696,164	15,171,025	6,395,590
3	2,816	6,653	7,929,856	44,262,409	18,734,848
4	5,555	9,543	30,858,025	91,068,849	53,011,365
5	1,292	3,418	1,669,264	11,682,724	4,416,056
6	2,208	5,563	4,875,264	30,946,969	12,283,104
7	1,313	3,660	1,723,969	13,395,600	4,805,580
8	1,102	2,694	1,214,404	7,257,636	2,968,788
9	3,151	5,468	9,928,801	29,899,024	17,229,668
10	1,516	2,898	2,298,256	8,398,404	4,393,368
11	5,161	10,674	26,635,921	113,934,276	55,088,514
12	4,567	7,585	20,857,489	57,532,225	34,640,695
13	5,841	11,760	34,117,281	138,297,600	68,690,160
14	3,008	4,085	9,048,064	16,687,225	12,287,680
Totale	40,898	81,577	156,831,834	592,083,727	301,298,822

$$SQX = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}X_i^2 - \frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}$$

$$= 156\,831\,834 - \frac{(40\,898)^2}{14}$$

$$= 156\,831\,834 - 119\,474\,743.1$$

$$= 37\,357\,090.86$$

di modo che

$$b_1 = \frac{62\,989\,097.29}{37\,357\,090.86}$$

$$= 1.68613$$

e

$$b_0 = \bar{Y} - b_1\bar{X}$$

e

$$\bar{Y} = \frac{\sum_{i=1}^{n}Y_i}{n} = \frac{81\,577}{14} = 5826.929$$

$$\bar{X} = \frac{\sum_{i=1}^{n}X_i}{n} = \frac{40\,898}{14} = 2921.2857$$

di modo che

$$b_0 = 5{,}826.929 - (1.68613)(2{,}921.2857)$$

$$= 901.2$$

Il calcolo delle misure di variabilità

Di seguito riportiamo le formule per il calcolo di *SQT*, *SQR* e *SQE*, in base alle equazioni (9.4), (9.5) e (9.6).

Formula per il calcolo della somma totale dei quadrati (*SQT*)

$$SQT = \text{variabilità totale} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \tag{9.19}$$

Formula per il calcolo della somma dei quadrati della regressione (SQR)

$$SQR = \text{variabilità spiegata dalla regressione} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = b_0\sum_{i=1}^{n} Y_i + b_1\sum_{i=1}^{n} X_iY_i - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \tag{9.20}$$

Formula per il calcolo della somma dei quadrati degli errori (SQE)

$$SQE = \text{variabilità residua} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} Y_i^2 - b_0\sum_{i=1}^{n} Y_i - b_1\sum_{i=1}^{n} X_iY_i \tag{9.21}$$

In base alla Tabella 9.7, otteniamo i seguenti risultati:

$$SQT = \text{variabilità totale} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

$$= 592\,083\,727 - \frac{(81\,577)^2}{14}$$

$$= 592\,083\,727 - 475\,343\,352$$

$$= 116\,740\,375$$

$$SQR = \text{variabilità spiegata}$$

$$= \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

$$= b_0\sum_{i=1}^{n} Y_i + b_1\sum_{i=1}^{n} X_iY_i - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

$$= (901.2)(81\,577) + (1.68613)(301\,298\,822) - \frac{(81\,577)^2}{14}$$

$$= 106\,208\,120$$

$$
\begin{aligned}
SQE &= \text{variabilità residua} \\
&= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i Y_i \\
&= 592\,083\,727 - (901.2)(81\,577) - (1.68613)(301\,298\,822) \\
&= 10\,532\,255
\end{aligned}
$$

Il calcolo dell'errore standard dell'inclinazione

In questo paragrafo illustriamo i calcoli che consentono di ottenere l'errore standard dell'inclinazione, che abbiamo impiegato nella verifica di ipotesi sull'esistenza di una relazione lineare tra X e Y.

$$
\begin{aligned}
S_{b_1} &= \frac{S_{YX}}{\sqrt{SQX}} \\
SQX &= \sum_{i=1}^{n} (X_i - \overline{X})^2 \\
&= \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} \\
&= 156\,831\,834 - \frac{(40\,898)^2}{14} \\
&= 37\,357\,090.86 \\
S_{b_1} &= \frac{936.85}{\sqrt{37\,357\,090.86}} \\
&= 0.1533
\end{aligned}
$$

L'esempio 9.5 presenta un'ulteriore illustrazione dei calcoli necessari per la stima del modello di regressione.

Esempio 9.5 *Il calcolo di b_0, b_1, SQT, SQR, SQE e r^2*

Prendete in considerazione il dataset A della Tabella 9.6:

X_i	Y_i
10	8.04
14	9.96
5	5.68
8	6.95
9	8.81
12	10.84
4	4.26
7	4.82
11	8.33
13	7.58
6	7.24

Calcolate b_0, b_1, SQT, SQR, SQE e r^2.