Statistica per le decisioni aziendali

Seconda edizione

con MyLab

L'attività didattica e di apprendimento del corso è proposta all'interno di un ambiente digitale per lo studio, che ha l'obiettivo di completare il libro offrendo risorse didattiche fruibili in modo autonomo o per assegnazione del docente.

Il codice studente monouso presente sulla copertina di questo libro consente l'accesso per 18 mesi a MyLab, una piattaforma digitale interattiva specificamente pensata per accompagnare e verificare i progressi durante lo studio.

MyLab offre la possibilità di accedere al manuale online: l'edizione digitale del testo arricchita da funzionalità che permettono di personalizzarne la fruizione, attivare la lettura audio digitalizzata, inserire segnalibri anche su tablet e smartphone.

Le attività formative e valutative sono dettagliate nella pagina di catalogo dedicata al libro, consultabile all'indirizzo linkpearson.it/CBEF0435 o tramite QR code.

Per accedere alla piattaforma **MyLab**, registrati come **studente** universitario sul sito **linkpearson.it/CE8622E3** attivando il codice studente monouso che trovi sulla copertina del libro.
Il **docente** può richiedere l'accesso alla piattaforma MyLab al consulente universitario di zona: **pearson.it/consulenti-universitari**.

€34,00

9788891931917

Gli autori e le autrici

Luigi Biggeri è Professore Emerito di Statistica economica presso l'Università degli Studi di Firenze, già Presidente dell'Istituto Nazionale di Statistica.

Matilde Bini è Professore Ordinario di Statistica economica presso l'Università Europea di Roma, è stata titolare di insegnamenti di Statistica in diverse Università Italiane.

Alessandra Coli è Ricercatrice di Statistica economica presso l'Università di Pisa, e dal 2020 ha un incarico temporaneo come *Statistical officer* presso Eurostat, Lussemburgo.

Laura Grassini è Professore Ordinario di Statistica economica presso l'Università degli Studi di Firenze.

Mauro Maltagliati è Professore Associato di Statistica economica presso l'Università degli Studi di Firenze.

Pearson

9788891931917A

Statistica per le decisioni aziendali
Seconda edizione
Luigi Biggeri
Matilde Bini
Alessandra Coli
Laura Grassini
Mauro Maltagliati

Statistica per le decisioni aziendali

Seconda edizione

Luigi Biggeri, Matilde Bini, Alessandra Coli, Laura Grassini, Mauro Maltagliati

MyLab Codice studente

Pearson

Statistica per le decisioni aziendali

Seconda edizione

Luigi Biggeri, Matilde Bini, Alessandra Coli, Laura Grassini, Mauro Maltagliati

Pearson

Realizzazione editoriale: Carmelo Giarratana
Grafica di copertina: Simone Tartaglia
Immagine di copertina: VladyslaV Travel photo/Shutterstock
Stampa: NEOGRAFIA A.S. – Martin-Priekopa, Slovacchia

9788891931917

2ª edizione: maggio 2023

Ristampa
00 01 02 03 04

Anno
23 24 25 26 27

Sommario

Prefazione

In un ambiente economico-aziendale sempre più globalizzato, dinamico e in continua trasformazione, il manager si trova a prendere decisioni in condizioni di incertezza, spesso in tempi brevi, avendo da analizzare una quantità notevole di informazioni quantitative (oltre che qualitative). La statistica ha perciò assunto un ruolo sempre più importante a supporto del processo decisionale del manager aziendale e, come vedremo, "pervade" qualsiasi azione e decisione aziendale.

Per affrontare un problema reale occorre partire dal contesto aziendale e dalla conoscenza della programmazione e gestione strategica ed operativa dell'azienda per definire e individuare, caso per caso a seconda del problema, i dati necessari per analizzarlo, e scegliere un appropriato metodo statistico per effettuare le analisi ed interpretare, infine, i risultati ottenuti.

Il manager e, soprattutto, l'esperto di statistica in azienda devono conoscere bene le due componenti necessarie per effettuare le analisi: le informazioni statistiche, incluse quelle derivanti dai Big Data (fonti, caratteristiche e qualità dei dati disponibili e/o di quelli da rilevare con apposite indagini) e i metodi statistici (sia descrittivi che inferenziali) che potrebbero essere adeguatamente utilizzati.

La Statistica aziendale – nell'ambito della quale questo volume didattico si colloca – si occupa proprio delle informazioni e dei metodi statistici per l'analisi dei fenomeni inerenti alla gestione dell'impresa a supporto delle decisioni manageriali, fornendo la cosiddetta "cassetta degli attrezzi" per affrontare adeguatamente i problemi e prendere le decisioni più opportune.

Nella preparazione del volume abbiamo, pertanto, cercato di attenerci a questa impostazione, privilegiando, rispetto agli aspetti analitici matematico-statistici, quelli concettuali e logici che ne giustificano l'applicazione ai casi concreti, e focalizzando l'attenzione su alcuni problemi che molto spesso i manager aziendali devono affrontare.

Il contenuto del volume, che è illustrato con dettaglio nel Capitolo 1, può essere distinto in due parti. La prima, costituita dai Capitoli 2 e 3, affronta temi di carattere generale inerenti all'informazione statistica disponibile e/o da produrre e all'interpretazione e comparazione dei dati, problematiche che sostanzialmente sempre si devono affrontare quando si impiega la statistica in azienda. La seconda parte, comprendente i restanti cinque capitoli, è dedicata alla descrizione ed applicazione di appositi metodi statistici ad alcune tra le problematiche più frequentemente oggetto di analisi in azienda. Da notare che la presentazione degli argomenti segue l'ordine logico della programmazione e della gestione delle attività e dei processi nei vari ambiti aziendali quale risulta dalla analisi di un processo produttivo, così come viene fatto nei testi di economia e gestione delle imprese.

CAPITOLO 6

Misura delle relazioni tra variabili per le decisioni aziendali

Matilde Bini

COSA CI SERVE SAPERE

È utile ripassare i seguenti argomenti:

covarianza, indici di correlazione e di associazione, tabelle di contingenza, distribuzioni: normale, **normale standardizzata**, *t*-Student, Chi-quadrato, *F*-Fisher. Teorie della stima puntuale e intervallare e del test delle ipotesi. Valore atteso. *Normal plot* e *q-q plot*. Scomposizione della varianza. Significato del *p-level*.

6.1 Relazioni tra variabili aziendali: considerazioni generali e contenuto del capitolo

Nella gestione operativa dell'impresa capita spesso, prima di prendere determinate decisioni, di volere verificare l'esistenza e l'intensità delle relazioni tra le variabili di interesse, al fine di tenerne conto nell'organizzazione e nell'impiego dei fattori produttivi e nell'organizzazione della fase di commercializzazione e di vendita. Per esempio, può risultare di interesse misurare e interpretare le relazioni tra le assenze dal lavoro e le qualifiche professionali del lavoratore o la sua anzianità in azienda; tra gli incidenti sul lavoro e l'orario di lavoro giornaliero e/o settimanale o l'età del lavoratore; tra i costi dei vari input e le quantità prodotte; tra il prezzo di un prodotto e le quantità e caratteristiche delle sue componenti; tra le vendite e il prezzo dei prodotti; tra le spese per la promozione e le vendite, e così via.

In tutti questi casi l'impiego di adeguati metodi di analisi statistica, che rientrano nell'ambito delle misure di correlazione e dei modelli di regressione, è certamente di aiuto per indirizzare le analisi e aiutare a definire le scelte del manager aziendale che in questo modo ha maggiori informazioni sulle caratteristiche delle "leve" da azionare.

Per presentare alcune esemplificazioni, ci riferiamo qui a due tipologie di analisi frequentemente svolte anche con l'ausilio dei metodi statistici: l'analisi dei costi di produzione e l'analisi delle vendite. Ma gli esempi che si potrebbero considerare in campo aziendale sono davvero tanti.

La prima tipologia assume un ruolo prioritario, sia per i riflessi direttamente esercitati sulla strategia competitiva, sia per l'incidenza assunta dai costi di produzione nel conto economico aziendale (Sciarelli, 1997).

A parte tutte le problematiche tipicamente aziendali, per il cui approfondimento rinviamo al testo citato, è evidente che il costo dei fattori impiegati per ottenere la produzione dei beni e/o servizi è il valore che dovrà essere recuperato con i ricavi ottenuti dalla loro vendita.

L'ammontare dei costi elementari dei fattori e del costo di produzione del prodotto naturalmente variano al variare dell'ammontare della produzione (anche in relazione alle cosiddette economie di scala). I costi e in particolare quelli variabili sono per definizione collegati al volume di attività (di produzione e di vendita) e l'analisi di questo legame è indispensabile per molteplici obbiettivi di analisi aziendale; per esempio, per il dimensionamento di un impianto (break-even point analysis), per verificare la validità del mix quali-quantitativo delle materie prime impiegate, per minimizzare i costi unitari di produzione, per determinare i prezzi di vendita, per valutare il volume minimo di produzione e di vendita al fine di recuperare integralmente i costi fissi e i costi variabili e, infine, nell'ambito della cosiddetta "Value Analysis"[1].

È pertanto importante derivare, almeno in prima istanza, la natura e il comportamento dei costi variabili, e di conseguenza anche dei costi totali, dall'analisi statistica dei dati disponibili in modo da evidenziare e misurare la relazione esistente tra tali costi e il volume della produzione.

Di altrettanto interesse è lo studio della relazione tra le vendite e il prezzo (o più prezzi) del prodotto (o più prodotti), e possibilmente con i prezzi di prodotti "concorrenti", per verificare se e quanto i prezzi influenzano le vendite dell'azienda e se conviene esercitare la leva prezzo per aumentare le vendite e/o i profitti. Come pure interessante è studiare la relazione tra l'ammontare delle vendite e le spese pubblicitarie e, in genere, per la promozione del prodotto. Anche in questo caso, l'analisi statistica dei dati disponibili consente di evidenziare e misurare le relazioni esistenti e di fornire utili indicazioni al manager sull'importanza delle leve di cui dispone in termini di politica dei prezzi e di politica di promozione.

L'obbiettivo di questo capitolo è proprio quello di fornire gli strumenti statistici per lo studio di tali relazioni, strumenti che ovviamente possono essere utilizzati anche per l'analisi delle relazioni tra gli altri fenomeni e variabili aziendali indicate prima.

Iniziamo, nel Paragrafo 6.2, con l'introduzione dei metodi statistici per l'analisi della relazione tra due variabili o bivariata, come tecnicamente viene definita. In questo caso in genere interessa:

a) in primo luogo, valutare il grado (intensità) e la forma della relazione attraverso l'impiego dell'analisi statistica della correlazione;

[1] La Value Analysis è un approccio di analisi per migliorare il valore di un prodotto o di un processo attraverso la comprensione delle sue componenti e dei relativi costi. Essa cerca di migliorare le componenti sia riducendo i loro costi (cioè individuando le aree di possibile risparmio), sia aumentando il valore delle loro funzioni (Miles, 1989).

b) in secondo luogo, se possibile, sintetizzare con una funzione analitica il legame tra le due variabili, attraverso l'utilizzo di un modello di regressione, in modo da conoscere in quale misura la variabile presa in considerazione (nei due esempi prima fatti il costo di produzione e l'ammontare delle vendite) dipende da un'altra variabile a cui è collegata (nel primo esempio potrebbe essere il volume di produzione; nel secondo il prezzo del prodotto). Analizzando il legame tra due sole variabili si parla di correlazione semplice e di regressione semplice.

È comunque evidente che spesso al manager interessa anche, e potremmo dire di più, l'analisi simultanea della relazione tra la variabile di interesse e due o più variabili che la influenzano, in modo da decidere quali tra le leve di cui dispone gli conviene azionare (per esempio è per lui importante conoscere come l'ammontare delle vendite è influenzato contemporaneamente dal prezzo del prodotto, dai prezzi dei prodotti concorrenti e dalle spese promozionali).

In questo caso, essendo coinvolte due o più variabili, si impiegano le misure di correlazione multipla e i modelli di regressione multipla (che sono presentati nel Paragrafo 6.3).

Come si vedrà, presentiamo essenzialmente modelli statistici lineari che sono ben noti e sono i più frequentemente utilizzati nelle analisi aziendali. Ciò non toglie che alcuni dei problemi in esame possano essere affrontati anche con l'impiego di altri e più sofisticati modelli per lo studio delle relazioni che si sono sviluppati negli ultimi decenni, per l'approfondimento dei quali rinviamo al testo di Chatterjee e Hadi, 2006, uno tra i vari testi sull'argomento proposti in letteratura.

La presentazione dell'analisi di correlazione e dei modelli di regressione punterà soprattutto a mettere in evidenza quali sono le modalità di impiego nei casi concreti, chiarendo, quindi, quali sono le ipotesi che stanno alla base per poterli utilizzare, la logica delle operazioni che si compiono con la loro applicazione, l'interpretazione dei risultati che si ottengono e, infine, poiché in genere si utilizzano dati provenienti da rilevazioni campionarie, l'attendibilità delle stime dei parametri (indici) e dell'adeguatezza complessiva del modello utilizzato.

6.2 Analisi e impiego della correlazione semplice e del modello di regressione lineare semplice

6.2.1 Analisi e misura della correlazione semplice

Come è noto, l'analisi della correlazione tra due variabili quantitative si conduce inizialmente attraverso la predisposizione del diagramma di dispersione (*scatterplot* o *scattergram*): la disposizione dei punti corrispondenti alle n coppie di valori delle due variabili, dette Y e X, fornisce una prima indicazione sull'eventuale legame.

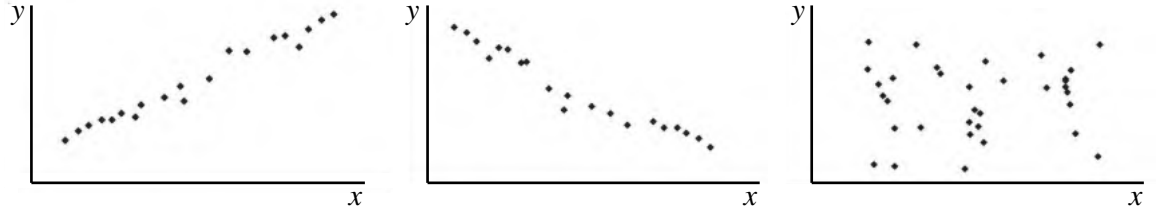

Figura 6.1 Aspetti grafici della correlazione.

A titolo esemplificativo, le tre rappresentazioni riportate nella Figura 6.1 mettono in evidenza, rispettivamente, una correlazione positiva, una correlazione negativa e una mancanza di correlazione.

Successivamente, in genere, per avere una misura dell'intensità della relazione si calcola l'**indice** parametrico di **correlazione di Pearson** ρ_{xy}.

L'indice viene utilizzato quando le variabili sono continue, cioè misurate su una scala a intervalli o razionale, sono distribuite normalmente e la loro relazione deve essere di tipo lineare

$$corr(x,y) = \rho_{xy} = \frac{\text{cov}(x,y)}{ss_{xx} \cdot ss_{yy}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{ss_{xx}^2 \cdot ss_{yy}^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i)-\bar{x}\,\bar{y}}{ss_{xx} \cdot ss_{yy}}$$

I valori che il coefficiente può assumere sono compresi tra –1 e +1; i valori estremi esprimono una relazione perfetta tra le variabili, mentre in assenza di una relazione il valore del coefficiente è 0. Il segno positivo o negativo corrisponde a quello della covarianza. Una correlazione positiva misura una relazione bidirezionale positiva tre le due variabili nel senso che la variazione unitaria di una variabile comporta una variazione nella stessa direzione (aumentativa o diminutiva) dell'altra variabile. Al contrario, quando il segno del coefficiente è negativo, le variazioni delle due variabili sono opposte. L'intensità di questo legame (positivo o negativo), e quindi in che proporzione avviene la variazione di una variabile rispetto alla variazione unitaria dell'altra, è rappresentato dal valore del coefficiente.

Come è noto dai testi di Statistica metodologica, cui si rinvia per i dettagli, quando l'indice di correlazione viene calcolato su dati provenienti da un campione, piuttosto che da un'intera popolazione obiettivo, occorre fare inferenza sull'indice procedendo con un **test di ipotesi** sulla **significatività** del suo valore.

La statistica test per la verifica dell'ipotesi nulla $H_0: \rho = 0$ (contro l'ipotesi $H_1: \rho \neq 0$) è:

$t = \dfrac{r-0}{\sqrt{(1-r^2)/(n-2)}} = r\sqrt{\dfrac{n-2}{1-r^2}}$ dove r è il valore dell'indice calcolato sui dati campionari e

$\sqrt{(1-r^2)/(n-2)}$ è l'errore standard di r. Se le variabili X e Y sono casuali e

distribuite normalmente[2], allora sotto l'ipotesi che H_0 sia vera, la statistica test t è distribuita come una variabile t di Student con $n - 2$ gradi di libertà.

Esempio 6.1

Studio della relazione tra costi e volume di produzione

Come anticipato nel primo paragrafo, affrontiamo qui a titolo esemplificativo l'analisi, con metodi statistici, dei costi di produzione in relazione al volume delle vendite, che in termini aziendali è detta *break-even point analysis*.

Rinviando ai testi di economia aziendale per una spiegazione dettagliata di questo tipo di analisi (si veda per esempio Sciarelli, 1997), ricordiamo che in genere essa si basa sulla costruzione di un grafico come quello riportato di seguito (Figura 6.2). Il grafico mette in evidenza la relazione esistente tra i costi (distinti in costanti e variabili) e l'ammontare della produzione, nonché l'ammontare dei ricavi in relazione alle unità prodotte e vendute, dati i prezzi di vendita. Diagrammi di questo tipo aiutano il manager ad analizzare l'andamento dei ricavi dalle vendite rispetto ai costi di produzione, evidenziando l'ammontare di utile o di perdita corrispondenti a un certo volume di vendita, in un determinato intervallo di tempo. Il punto di efficienza si ha per un volume di vendite di unità, dove i ricavi sono uguali ai costi; sotto tale punto le vendite di una quantità inferiore dei prodotti dà luogo a una perdita e viceversa sopra il punto l'azienda ricava un utile.

Nella costruzione del diagramma si presuppone che i costi siano sotto controllo e che non si verifichino sensibili fluttuazioni dei prezzi dei materiali, della manodopera e dei prodotti finiti, e che i costi variabili siano direttamente proporzionali, in modo deterministico, in relazione al volume di produzione.

Tuttavia queste ipotesi sono difficili da giustificare, se non nel breve periodo e a priori nulla si può dire sulla relazione tra costi variabili e ammontare della produzione, tanto che alcuni aziendalisti ipotizzano differenti tipi di relazioni come evidenziato dalla Figura 6.2.

È evidente quindi che le ipotesi che si possono fare a priori sulle relazioni tra i vari costi e il volume della produzione devono essere verificate empiricamente attraverso una misura della correlazione esistente tra le variabili, soprattutto attraverso la misura della dipendenza dei costi di produzione dal volume delle vendite (con la stima dei parametri di un modello di regressione).

Supponiamo che un'azienda del settore alimentare abbia diversi stabilimenti (oppure molti reparti all'interno di uno stabilimento) di dimensioni diverse secondo il numero di addetti, che producono lo stesso tipo di prodotto con volumi di produzione e di vendite differenti, e che sono localizzati in tutto il territorio nazionale, per esempio nei capoluoghi di regioni italiane. L'azienda ha rilevato, attraverso un campione casuale, i dati sui volumi di produzione e costi totali riportati nella Tabella 6.1.

[2] Occorre prestare attenzione alla presenza di valori anomali nei dati che potrebbero inficiare la natura delle distribuzioni e conseguentemente la validità del test.

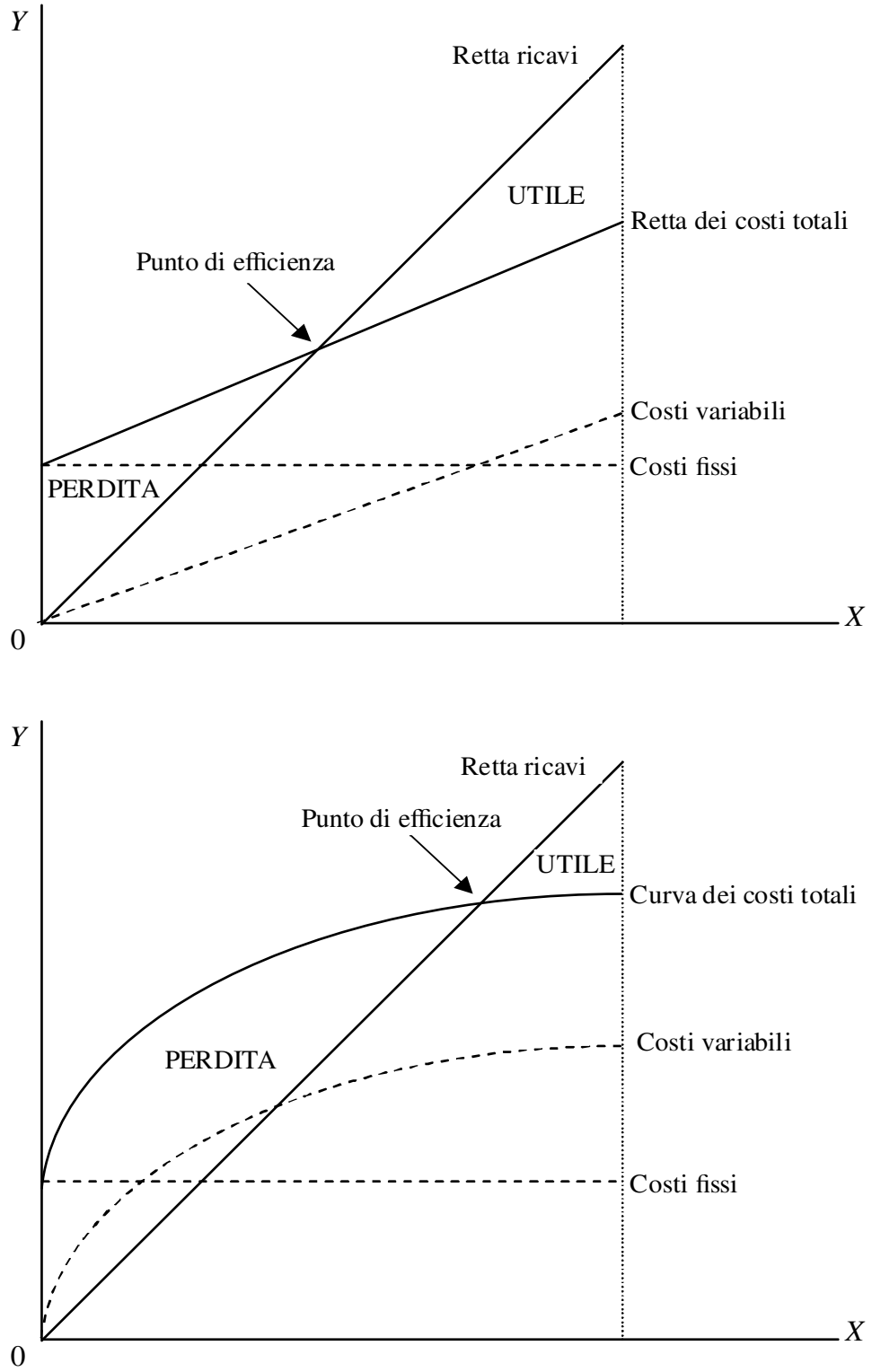

Figura 6.2 Diagramma di efficienza.

Tabella 6.1 Dati della produzione totale annua e dei costi totali annui.

Capoluoghi in cui sono localizzati gli stabilimenti	Produzioni annue in tonnellate	Costi totali annui in milioni di euro
Ancona	3557,7	29,61
Aosta	3296,4	27,83
Bari	3437,1	28,39
Bologna	3336,6	29,17
Cagliari	3618,0	30,17
Campobasso	3376,8	30,28
Catanzaro	3195,9	28,06
Firenze	4060,2	33,28
Genova	3859,2	29,28
L'Aquila	3658,2	29,51
Milano Nord	3678,3	31,28
Milano Sud	3825,0	31,06
Napoli	3396,9	29,83
Palermo	3497,4	28,39
Perugia	3296,4	28,17
Potenza	3638,1	29,28
Roma Nord	3879,3	31,06
Roma Sud	4502,4	35,27
Torino	3396,9	27,28
Trento	3457,2	28,72
Trieste	3206,0	27,56
Venezia	3356,7	28,94

L'obiettivo dell'analisi è individuare e misurare la relazione tra i costi e volume della produzione, che sono le due variabili in gioco. La verifica dell'eventuale relazione tra le due variabili e la successiva determinazione della forma analitica di tale relazione serviranno ovviamente per costruire il diagramma di efficienza.
La prima cosa da fare per la verifica dell'esistenza della relazione è la costruzione del diagramma di dispersione, riportato nella Figura 6.3.

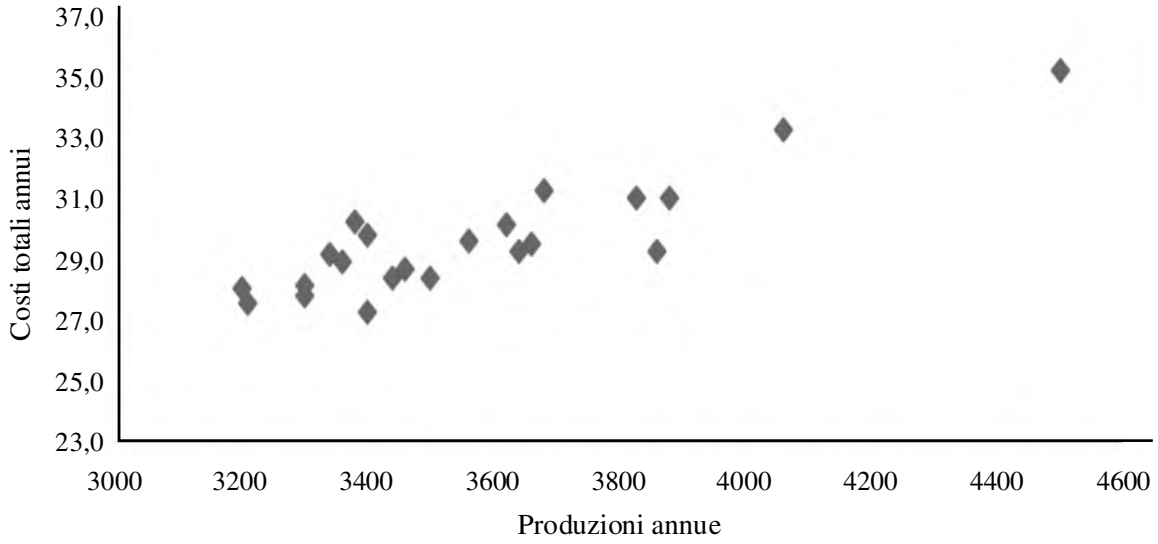

Figura 6.3 Diagramma dei costi totali e dei volumi di produzione.

Come si vede dalla figura, all'aumentare del livello di produzione, aumenta il valore del costo totale: esiste pertanto una relazione positiva tra le due variabili.

Tuttavia al manager interessa conoscere anche l'intensità di questo legame, che si ottiene calcolando l'indice di correlazione di Pearson.

Riportiamo le operazioni necessarie per misurare la covarianza e l'indice di correlazione:

Si calcolino innanzitutto le medie aritmetiche $M(X) = \bar{x}$ e $M(Y) = \bar{y}$:

$$\bar{x}\,\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{78526{,}68}{22} = 3569{,}39 \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{652{,}39}{22} = 29{,}65$$

Tali valori sono utilizzati nel calcolo della covarianza e dell'indice di correlazione:

$$ss_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 500{,}52$$

$$ss_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\,\bar{y} = \frac{1}{22}\cdot 2339644{,}8 - (3569{,}39 \cdot 29{,}65) = 500{,}52$$

$$corr(x,y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\,\bar{y}}{ss_{xx}\cdot ss_{yy}} = \frac{500{,}52}{\sqrt{92531{,}01 \cdot 3{,}42}} = \frac{500{,}52}{304{,}19 \cdot 1{,}85} = \frac{500{,}52}{562{,}14} = 0{,}89$$

Come si rileva dal risultato ottenuto, per l'azienda di prodotti alimentari il valore dell'indice di correlazione tra le due variabili è molto alto. Si può concludere, pertanto, che la relazione tra il volume dei costi, indicato con y, e il volume della produzione (x) è quasi del 90%.

Essendo i dati ottenuti attraverso un campione, si esegue un test a due code dell'ipotesi nulla di assenza di correlazione a un livello di significatività pari, per esempio, ad $\alpha = 0{,}05$. Il valore della statistica t è:

$$t = 0{,}89\sqrt{\frac{22-2}{1-(0{,}89)^2}} = 8{,}74$$

Per una distribuzione t di Student con 20 gradi di libertà, la probabilità di osservare un valore maggiore di $|8{,}74|$ è $p = 0{,}000000028$. Si rifiuta quindi l'ipotesi nulla: sulla base dei dati campionari osservati, c'è evidenza che la correlazione reale nella popolazione sia diversa da 0. ❑

BOX 6.1

L'analisi della relazione tra due (o più) variabili può essere affrontata anche quando i dati sono tutti di natura categorica oppure mista, cioè quantitativi e qualitativi insieme. In ambito aziendale, si potrebbero, per esempio, riscontrare problemi legati a esigenze di ottimizzazione della programmazione delle carriere che vedono implicati i livelli salariali con i livelli di titolo di studio del personale, oppure la diversa localizzazione degli stabilimenti con il livello di produttività, ma anche relazioni tra voci diverse del bilancio, infine indagini di marketing che prevedono molte variabili qualitative nominali e ordinali. Nel seguente prospetto sono riassunti tutti gli indici statistici di correlazione e associazione che possono essere utilizzati a seconda della combinazione tra la diversa natura delle variabili.

Dati	Continuo	Ordinale	Nominale
Continuo	ρ di Pearson		
Ordinale	ρ di Spearman	ρ di Spearman τ di Kendall γ di Goodman e Kruskall	
Nominale	Rapporto di correlazione η^2		χ^2 di connessione di Pearson ϕ^2 di contingenza quadratica V di Cramer

Una trattazione di questi indici con relativi esempi è rimandata al paragrafo sugli indici di correlazione e associazione nella piattaforma MyLab.

6.2.2 Modello di regressione lineare semplice

Caratteristiche del modello

Gli indici descrittivi di correlazione che abbiamo presentato misurano una relazione lineare biunivoca tra due variabili. Ma in azienda, come si è accennato nel caso dell'analisi dei costi descritto nel precedente esempio, nasce quasi sempre

l'esigenza di studiare e misurare la correlazione intesa come una relazione di dipendenza di una variabile (o più variabili) rispetto a un'altra variabile (o più variabili).

Focalizzando per il momento l'attenzione sulla relazione tra due variabili (saranno proprio i costi totali e volume della produzione di cui all'esempio già citato), la relazione di dipendenza lineare di Y da X può essere espressa attraverso l'impiego di un **modello di regressione lineare semplice**.

La relazione di dipendenza lineare può essere una relazione funzionale del tipo: $Y = \beta_0 + \beta_1 x_i + \epsilon_i$; la "valutazione" dei valori di β_0 e β_1 aiuterà a quantificare l'entità della dipendenza di Y da X, a visualizzare la relazione tra le due variabili, e a "prevedere" Y in funzione di X (interpolazione/estrapolazione). I concetti che stanno alla base dell'analisi di regressione lineare sono infatti importanti sia per lo studio della dipendenza fra caratteri che per la costruzione di modelli di previsione, basati sulla relazione e dipendenza tra variabili.

La forma generale per un **modello di regressione probabilistico** è la seguente

$$Y = \textit{componente deterministica} \text{ (o } \textit{sistematica}\text{)} + \textit{errore accidentale}$$

Come vedremo, la componente di errore gioca un ruolo fondamentale nell'analisi inferenziale del modello e cioè nella determinazione di stime di intervallo o nella conduzione di test delle ipotesi sulla parte deterministica del modello.

In questo capitolo consideriamo il modello più semplice che è lineare in x e la cui componente deterministica individua una retta.

L'espressione generale del modello ipotizzato, che si definisce di **regressione lineare semplice di I ordine**, è pertanto la seguente:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

Y: variabile dipendente o variabile risposta

x: variabile indipendente o predittore (o variabile esplicativa)

ϵ: componente di errore accidentale o disturbo

$\mu = \beta_0 + \beta_1 x$: componente deterministica o retta di regressione

β_0: intercetta della retta di regressione

β_1: pendenza della retta di regressione.

dove β_0 è moltiplicato per un valore costante 1, β_1 è moltiplicato per x.

Si può osservare inoltre, che β_0 è il valore di μ quando x è uguale a zero; β_1 indica di quanto aumenta μ se x aumenta di una unità. β_1 è la derivata prima della componente deterministica rispetto a x.

Da quanto detto segue che:

- β_0 è espresso nella stessa unità di misura della variabile risposta;
- β_1 ha unità di misura che dipende da y e da x (se x è espressa in kg e y in euro, β_1 è espresso in euro/kg).

Un'altra grandezza interessante è l'*elasticità* di μ rispetto a x[3] che, per il modello sopra indicato, è pari a $\beta_1 / (\mu / x)$. Se tale grandezza è, per esempio, 0,3 significa che al variare di x dell'1%, y varia dello 0,3%.

Il termine lineare, usato nella definizione del modello, è dovuto al fatto che è lineare nei parametri; è semplice perché contiene una sola variabile indipendente; è di primo ordine perché la variabile indipendente ha esponente unitario (cioè il modello è lineare anche in x).

Si può notare che, nell'espressione del modello sopra definito, abbiamo scritto con le lettere maiuscole il simbolo della variabile risposta. Il motivo è che Y è una variabile casuale in quanto dipendente da ϵ, la componente stocastica del modello. La variabile x mantiene invece la natura di variabile deterministica (variabile matematica). Nel **modello probabilistico** la componente deterministica rappresenta la linea delle medie e cioè $E(Y \mid x) = \beta_0 + \beta_1 x + \epsilon$. Come si vede, $E(Y \mid x)$ è detto anche valore atteso condizionato della v.c. Y rispetto a un determinato valore di x; per esempio, nell'espressione sotto, μ^* è il valore atteso della variabile risposta nel punto x^* e cioè *condizionatamente a* $x = x^*$.

$$E(Y \mid x = x^*) = \mu^* = \beta_0 + \beta_1 x^*$$

I parametri β_0 e β_1 sono generalmente incogniti e quindi devono essere stimati con procedure statistiche, sulla base di osservazioni raccolte su x e y. Nel caso di dati osservazionali, si rilevano le due variabili su ogni singola unità.

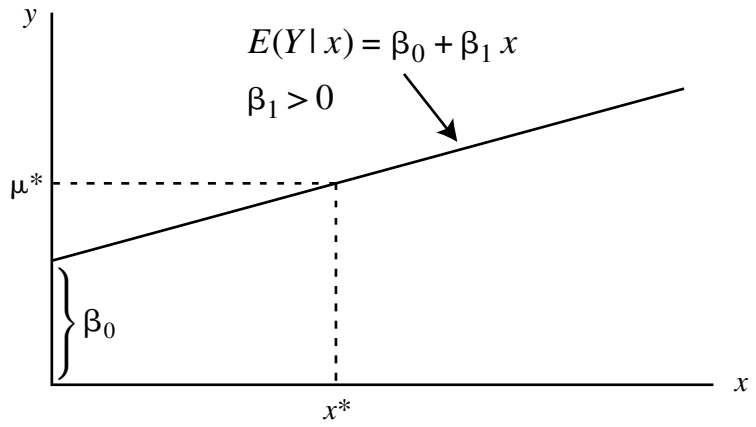

Figura 6.4 Retta di regressione.

[3] Data una funzione matematica $y = f(x)$ l'elasticità di y rispetto a x indicata è $\partial f(x)/(y/x)$.

Stima e impiego del modello

Le fasi attraverso le quali viene spiegato il modello di regressione lineare semplice del I ordine sono le seguenti.

Fase 1 – Specificazione della componente deterministica del modello che mette in relazione la media condizionata $E(y \mid x) = \mu$ con la variabile indipendente.

Fase 2 – Utilizzo dei dati campionari per stimare i parametri incogniti della componente deterministica.

Fase 3 – Specificazione delle caratteristiche distributive della componente di disturbo e stima della varianza di detta componente.

Fase 4 – Valutazione della validità del modello mediante analisi descrittiva e inferenziale.

Fase 5 – Impiego del modello.

Fase 6 – Interpretazione dei risultati della stima del modello.

Le illustriamo brevemente di seguito.

Specificazione della componente deterministica

A questo punto definiamo il modello con riferimento a un campione di n coppie di osservazioni sulle due variabili. Nelle formule verrà aggiunto il pedice i per indicare la generica unità statistica i. In base a quanto illustrato nel paragrafo precedente, il modello probabilistico è così definito:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{per } i = 1, \ldots, n$$

da cui deriviamo la formula della retta di regressione (retta delle medie); dall'espressione del valore atteso condizionato sopra definito e considerando che in media gli errori sono nulli ($E(\epsilon_i) = 0$), si deduce che $E(Y_i \mid x_i) = \mu_i = \beta_0 + \beta_1 x_i + \epsilon_i$ per ogni i.

La retta dei Minimi Quadrati Ordinari e la stima dei parametri

La retta dei Minimi Quadrati Ordinari (MQO) è quella che, fra tutte le rette possibili, minimizza la somma dei quadrati degli scarti fra valori osservati e i valori interpolati della variabile risposta, da cui ricava che soluzione del sistema di equazioni che ammette un'unica soluzione

$$b_1 = \frac{ss_{xy}}{ss_{xx}} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

dove

$$ss_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad (\textit{devianza di } x)$$

$$ss_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \quad (\textit{codevianza di } xy)$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sum_{i=1}^{n} y_i h_i \qquad h_i = \frac{(x-\bar{x})}{ss_{xx}}$$

con n = *dimensione campionaria*.

Vediamo qui come sono stati stimati i parametri della componente deterministica del modello.

Supponiamo di voler trovare la funzione lineare in x (la formula di una retta) che produce la migliore stima di y, intendendo come migliore quella che minimizza (sui dati disponibili) la somma dei quadrati degli scarti fra i valori osservati di y e quelli stimati attraverso la funzione lineare suddetta.

Questa strategia di stima è chiamata metodo dei Minimi Quadrati Ordinari. L'aggettivo "ordinari" distingue il metodo da altri procedimenti più sofisticati (per esempio, MQ ponderati, MQ generalizzati ecc.).

La Figura 6.5 riporta, oltre ai punti, anche la retta stimata mediante il metodo MQO. La retta dei MQO è denominata in vari modi: retta di regressione, retta interpolata, retta stimata; alcuni termini anglosassoni spesso usati e che si trovano in letteratura sono: least squares line, regression line, least squares prediction equation, fitted line.

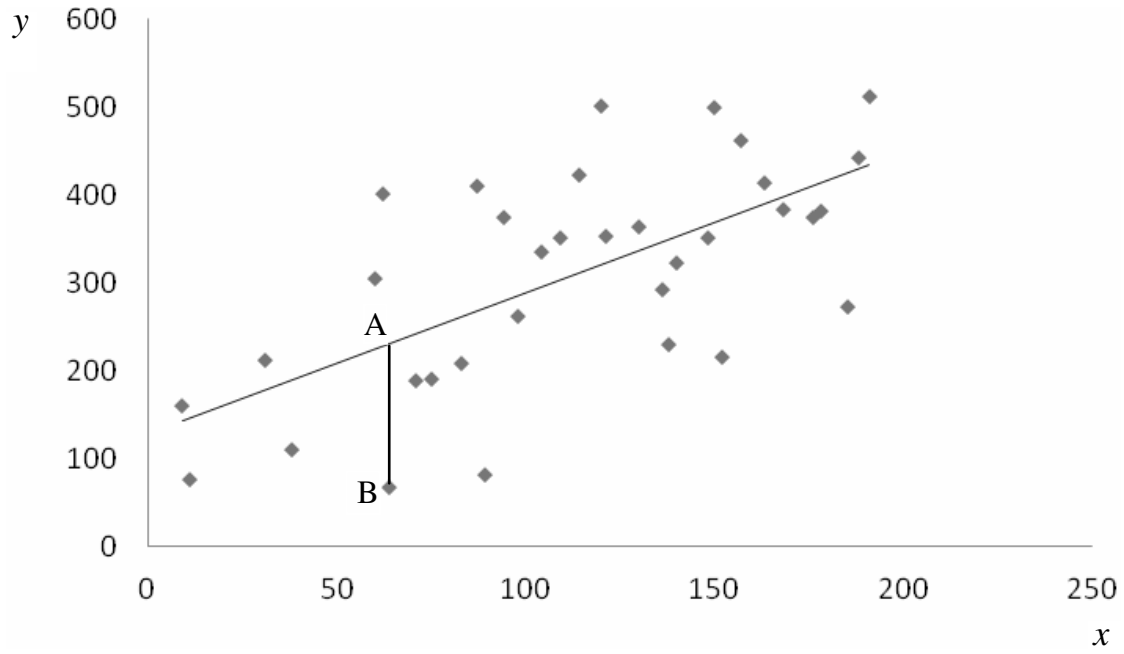

Figura 6.5 Diagramma a dispersione delle unità e un esempio di retta di regressione stimata con il metodo MQO.

Il metodo dei MQO si basa quindi sulla minimizzazione di una funzione che sintetizza le "distanze" fra valori osservati della variabile risposta y_i e valori stimati (o interpolati o predicted) $\hat{y}_i$, dove

$$\hat{y}_i = b_0 + b_1 x_i$$

e b_0 e b_1 sono i valori stimati (o stime) dei parametri incogniti β_0 e β_1.

Le distanze fra valori osservati e valori interpolati sono rappresentate nella Figura 6.5 dai segmenti verticali di tipo AB. Tali distanze, denominate anche scarti e indicate con d_i, vengono calcolate come:

$$d_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

La somma dei quadrati degli scarti che è minima in corrispondenza dei valori stimati dai MQO è:

$$sse = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_i) \right]^2$$

Sottolineiamo, pertanto, che *sse* (*sum of squared errors*) indica il minimo valore che può assumere la somma dei quadrati dagli scarti. Tale minimizzazione è realizzata, per l'appunto, dal metodo dei MQO.

La bontà di adattamento del modello

Si può dare un giudizio descrittivo sull'adattamento del modello, mediante l'indice di determinazione R^2 che è uguale al rapporto fra devianza di regressione e devianza totale della variabile dipendente.

$$R^2 = \frac{Devianza\ \ di\ \ regressione}{Devianza\ \ totale} = 1 - \frac{Devianza\ \ residua}{Devianza\ \ totale} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Questo indicatore è capace di riassumere l'adattamento globale e la capacità esplicativa complessiva del modello in relazione ai dati campionari. I valori che può assumere questo indicatore sono compresi tra 0 e 1 ($0 \leq R^2 \leq 1$). Quando $R^2 = 0$, il modello non spiega per niente la variabile risposta (la devianza di regressione è nulla e la variabilità di y non dipende dalla relazione con x).

La retta di regressione è parallela all'asse x (caso di indipendenza interpolativa). Quando invece $R^2 = 1$, il modello spiega perfettamente la variabile risposta. I punti sono allineati sulla retta di regressione.

Nei casi reali non si otterrà nessuna delle due situazioni limite: il significato di R^2 consiste, allora, nel misurare la percentuale di variabilità totale "spiegata" mediante la retta di regressione.

Caratteristiche distributive del disturbo e la significatività delle stime

Nel paragrafo precedente abbiamo introdotto il modello probabilistico come composto da una parte deterministica e una parte stocastica. In questo paragrafo viene completata la specificazione del modello con le ipotesi sulla componente di disturbo che riportiamo qui di seguito.

Specificazione del modello di regressione lineare semplice di I ordine

a) $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{per } \forall\, i = 1, ..., n$

b) $E(\epsilon_i) = 0 \quad \rightarrow \quad E(Y_i \mid x_i) = \mu_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{per } \forall\, i = 1, ..., n$

c) $V(\epsilon_i) = \sigma^2 \quad \rightarrow \quad V(Y_i) = \sigma^2 \quad \text{per } \forall\, i = 1, ..., n$

d) La distribuzione di ϵ_i è normale con varianza omoschedastica:

$$\epsilon_i \sim N(0, \sigma^2) \quad \rightarrow \quad Y_i \sim N((\beta_0 + \beta_1 x_i), \sigma^2))$$

e) Gli errori ϵ_i sono indipendenti, vale a dire che il disturbo associato all'osservazione i-esima non ha nessun effetto su quello dell'osservazione j-esima:

$$Cov(\epsilon_i, \epsilon_j) = 0 \quad \forall\, i \neq j \;\; \text{con } i = 1, ..., n$$

Come si può osservare, è presente un altro parametro distributivo: la varianza σ^2 della componente stocastica; tale varianza è assunta costante per ogni i (e quindi, è costante al variare di x) e per questo viene detta omoschedastica (in caso contrario viene detta eteroschedastica).

Le implicazioni delle prime quattro ipotesi possono essere facilmente apprezzate dalla Figura 6.6. Le distribuzioni degli errori sono centrate sulla media ovvero su un punto che si trova sulla retta (e infatti è stata denominata anche retta delle medie). Le distribuzioni sono tutte uguali, al variare della x, perché $\epsilon_i \sim N(0, \sigma^2)$ per ogni i. L'ipotesi di normalità della componente di disturbo è utile per poter effettuare analisi inferenziale come: stima di intervallo e test delle ipotesi sui parametri della componente deterministica.

Se, nelle formule delle stime qui di seguito riportate, al posto dei valori osservati della variabile dipendente y_i poniamo il simbolo della variabile casuale Y_i otteniamo l'espressione degli stimatori dei parametri della componente deterministica del modello, che indichiamo con B_0 e B_1.

Formule degli stimatori del MQO del modello di regressione lineare semplice di I ordine

$$B_1 = \frac{SS_{xy}}{ss_{xx}} \qquad B_0 = \overline{Y} - B_1 \overline{x}$$

dove

$$ss_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad (\textit{devianza di } x)$$

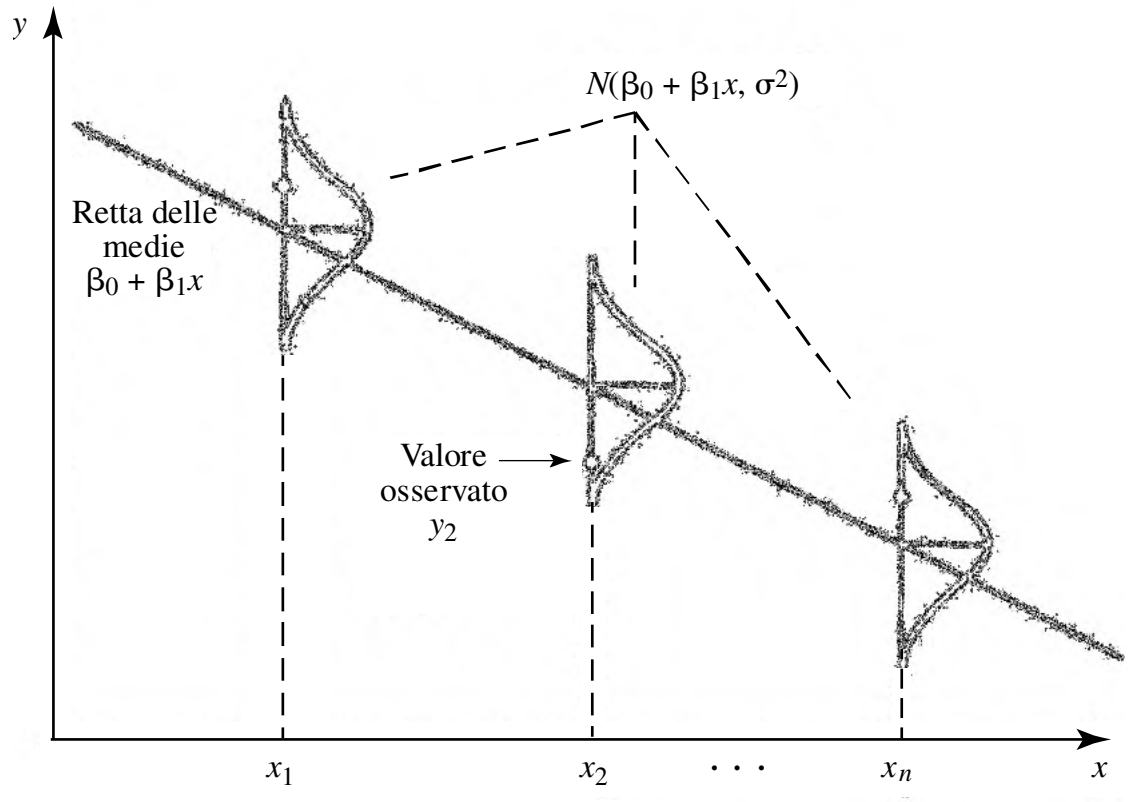

Figura 6.6 Distribuzione del disturbo e retta di regressione.

$$SS_{xy} = \sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x}) \quad (\textit{codevianza di } x,y)$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} Y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sum_{i=1}^{n} Y_i h_i \qquad h_i = \frac{(x - \bar{x})}{ss_{xx}}$$

con n = *dimensione campionaria.*

Definiamo inoltre la variabile casuale:

$$\hat{Y}_i = B_0 + B_1 x_i$$

Questa variabile casuale può essere considerata:

1. uno stimatore della media $\mu = \beta_0 + \beta_1 x$;
2. un predittore del valore della variabile casuale $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$; in questo caso si parla di predittore e non di stimatore perché esso ha lo scopo di "stimare" la determinazione di una variabile casuale e non il valore di un parametro (costante).

Nel caso 1 abbiamo che $\hat{Y}_i$ è stimatore corretto di μ_i. L'errore di stima pari a

$$W_i = \hat{Y}_i - \mu_i = \hat{Y}_i - (\beta_0 + \beta_1 x_i) = (B_0 - \beta_0) + (B_1 - \beta_1) x_i$$

La varianza di e_i (che è uguale alla varianza di $\hat{Y}_i$) è

$$V(W_i) = E(\hat{Y}_i - \mu_i)^2 = V(\hat{Y}_i) = \sigma^2 \left(\frac{1}{n} + h_i^2 \right)$$

Pertanto, quanto più è grande $V(\hat{Y}_i)$ tanto meno è precisa la statistica $\hat{Y}_i$ nella stima di μ_i.

Nel caso 2 viene introdotto l'errore di previsione E_i che è

$$E_i = \hat{Y}_i - Y_i = B_0 + B_1 x_i - (\beta_0 + \beta_1 x_i + \epsilon_i) = (B_0 - \beta_0) + (B_1 - \beta_1) x_i + \epsilon_i$$

La varianza di E_i è:

$$V(E_i) = E\left[(\hat{Y}_i - \mu_i) - (Y_i - \mu) \right]^2 = \sigma^2 + V(W_i) = \sigma^2 \left(1 + \frac{1}{n} + h_i^2 \right)$$

Pertanto, quanto più grande è questa varianza tanto peggiore sarà $\hat{Y}_i$ quale predittore di Y_i. Si noti che la varianza dell'errore di previsione è più grande di quella dell'errore di stima. In quest'ultimo è presente solo la variabilità campionaria nella stima di μ_i; nell'errore di previsione è presente anche la componente stocastica del modello.

Da notare, inoltre che la variabile casuale Y_i e il suo predittore $\hat{Y}_i$ hanno lo stesso valore atteso e cioè $E(Y_i) = E(\hat{Y}_i) = \mu_i$.

La stima corretta della varianza σ^2 viene ricavata attraverso i residui dei MQO e precisamente

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - 2}$$

La somma dei quadrati dei residui viene divisa per la grandezza [numero di osservazioni – numero dei parametri – 1)]. La grandezza

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

è lo stimatore corretto di σ^2.

Sfruttando l'ipotesi di normalità della componente stocastica è possibile determinare intervalli di confidenza per i parametri e per μ_i e intervallo di previsione per Y_i. È possibile anche condurre test delle ipotesi.

Una nota metodologica sugli stimatori è rimandata al paragrafo sugli stimatori nella piattaforma MyLab.

La significatività delle stime

Un particolare test delle ipotesi risponde alla seguente domanda: è significativo o meno l'effetto della variabile esplicativa sulla variabile dipendente? Il test che possiamo formulare per questo quesito è il seguente:

$H_0 : \beta_1 = 0$ contro l'ipotesi alternativa $H_1 : \beta_1 \neq 0$ oppure, se si hanno delle idee sul valore del parametri, si può usare un test unilaterale.

Questo non è altro che un test delle ipotesi sulla media in quanto B_1 ha come media proprio il valore del parametro β_1 (si veda l'appendice metodologica riportata nella piattaforma MyLab). Quindi la statistica test è

$$\frac{B_1 - \beta_1}{\sqrt{\hat{V}(B_1)}} \sim t_{n-k-1}$$

che sotto H_0 assume la forma

$$\frac{B_1 - 0}{\sqrt{\hat{V}(B_1)}} \sim t_{n-k-1}.$$

Per gli intervalli di confidenza dei parametri si veda l'appendice metodologica (nella piattaforma MyLab) nella quale si possono trovare anche le formule per la stima di intervallo di μ_i e per l'intervallo di previsione per la variabile casuale Y_i in corrispondenza a un prefissato valore x_i.

Se si vuole verificare l'ipotesi $H_0 : \beta_j = 0$ contro l'ipotesi alternativa $H_1 : \beta_j \neq 0$ (test bidirezionale) con un livello di significatività α, e quindi rifiutare tale ipotesi, occorrerà individuare la regione di rifiuto data dai valori della statistica t che sono maggiori del valore teorico $t_{\alpha/2,n-k-1}$ (dove k è il numero dei parametri nel modello, escluso quello dell'intercetta), dedotto dalla tavola delle probabilità della distribuzione t di Student.

Data l'ipotesi nulla, la regione di rifiuto di $H_0 : \beta_j = 0$ è l'area sotto la curva della distribuzione t Student con $\alpha / 2, n - k - 1$ gradi di libertà, delineata dai valori $|t| > t_{\alpha/2,n-k-1}$. La Figura 6.7 mostra la regione di rifiuto per un test bidirezionale.

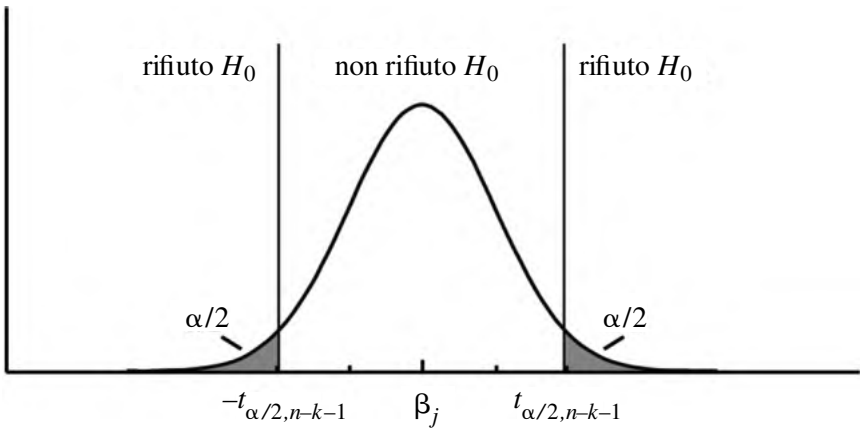

Figura 6.7 Funzione di densità di probabilità di β sotto l'ipotesi $H_0 : \beta_j = 0$.

Si può anche valutare che un coefficiente possa essere uguale a un valore numerico e verificare l'ipotesi che $H_0 : \beta_j = \beta_j^*$. In tal caso le statistiche test per verificare l'ipotesi nulla per ciascun coefficiente β_j saranno

$$\frac{\left|\beta_j - \beta_j^*\right|}{S\sqrt{h_j}} \sim t_{n-k-1}.$$

Il livello di significatività del test α è la probabilità di effettuare un errore del I tipo se l'ipotesi nulla è vera, $\alpha = \Pr(\text{rifiutare } H_0 | H_0)$, cioè la probabilità di rifiutare H_0 quando H_0 è vera, quindi la probabilità di rifiutare *erroneamente* H_0. Nel caso del test bidirezionale, il *p-value* è la probabilità che una variabile casuale t Student con $g = n-k-1$ gradi di libertà generi una realizzazione maggiore in valore assoluto di quella ottenuta $p\text{-}value = P(|t_{n-k-1}| > |t_{calc}|)$. Si noti che $P(|t_{n-k-1}| > |t_{calc}|) = \alpha = \Pr(\text{rifiutare } H_0 | H_0)$, quindi gli intervalli $[-\infty, -t_{\alpha/2, n-k-1})$ e $(t_{\alpha/2, n-k-1}, \infty]$ sono *la regione di rifiuto di* H_0 e di conseguenza *la regione di non rifiuto di* H_0 è $[-t_{\alpha/2, n-k-1}, t_{\alpha/2, n-k-1}]$.

6.2.3 Analisi dei residui

L'ultima fase dell'analisi consiste nello studio della componente stocastica del modello ϵ_i come ulteriore verifica della bontà del modello adattato e per controllare se valgono alcune assunzioni fondamentali alla base del modello lineare: le ipotesi di normalità e di omoschedasticità.

Solo se questi controlli convalidano le assunzioni fatte si può considerare valida l'operazione di stima mediante i MQO e quindi anche il risultato del test delle ipotesi.

Ma come si fa se la variabile non è osservabile? Forse è meglio ricorrere all'analisi dei residui anche attraverso opportuni grafici.

Si analizzano pertanto i residui di regressione $r_i = y_i - \hat{y}_i$. Bisogna considerare che il residuo i-esimo è la determinazione di una variabile casuale con media nulla ma con varianza che dipende da i (e cioè dai valori della variabile esplicativa). Il procedimento corretto sarebbe quello di studiare non i residui grezzi (che hanno varianze diverse e quindi non sono omoschedastici anche se rimangono incorrelati fra loro) bensì quelli standardizzati che sono ricavati tenendo presente la seguente formula

$$r_{i,std} = \frac{r_i}{s(r_i)}$$

dove $s(r_i)$ è la devazione standard stimata del residuo i.

Essi sarebbero in vero delle variabili casuali t-Student tuttavia, per una numerosità campionaria n abbastanza grande, si può applicare la distribuzione normale.

Verifica empirica dell'ipotesi di normalità

I residui dovrebbero presentare una distribuzione empirica piuttosto simile alla normale. Per confrontare la distribuzione dei residui con la distribuzione normale può essere utilizzato un grafico detto q-qnormal plot (il grafico viene costruito con i quantili della distribuzione empirica dei residui e i quantili della distribuzione normale standard).

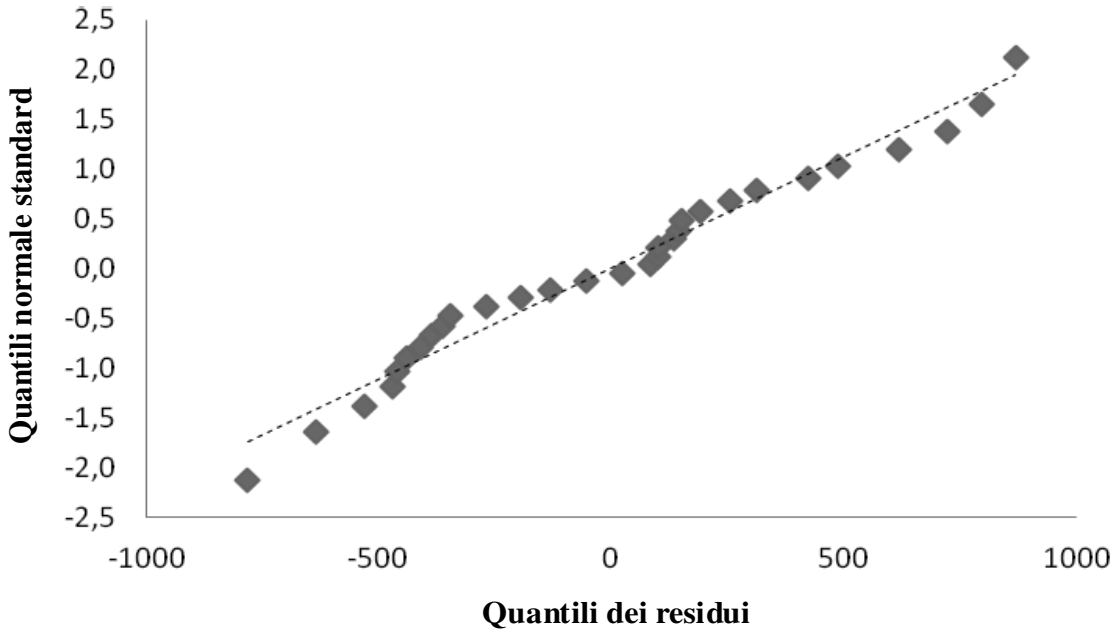

Figura 6.8 Un esempio di q-q normal plot.

Un'altra rappresentazione utile per verificare l'ipotesi di normalità è quella rappresentata nella Figura 6.9. I residui dovrebbero disporsi in maniera casuale intorno all'asse delle ascisse come rappresentato nella figura. Sotto l'ipotesi di normalità, il 95% di essi dovrebbe essere compreso tra $\pm 1{,}96 \cdot \hat{\sigma}$.

Verifica empirica dell'ipotesi di omoschedasticità

La presenza di strutture particolari nel grafico dei residui può indicare errori di specificazione nel modello. Se per esempio nel grafico che può essere costruito rispetto a una **variabile esplicativa** o ai **valori stimati della funzione di regressione**, l'ordine di grandezza in cui si dispongono i residui si modifica come rappresentato nella Figura 6.10, questo può indicare la presenza di eteroschedasticità (quindi assenza di omoschedasticità).

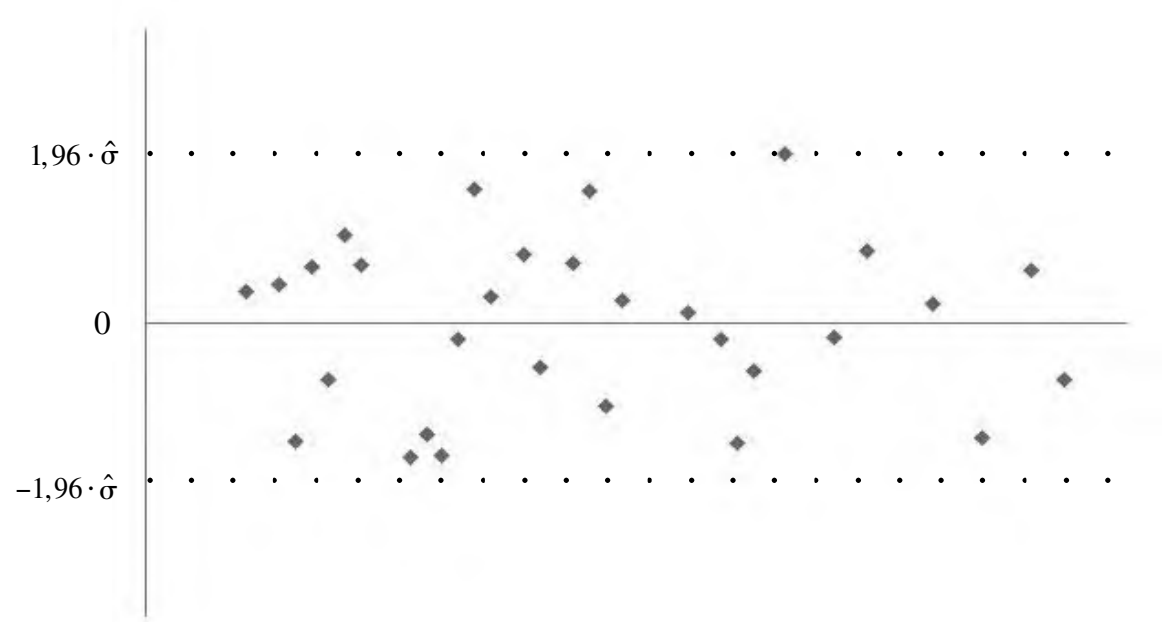

Figura 6.9 Grafico dei residui.

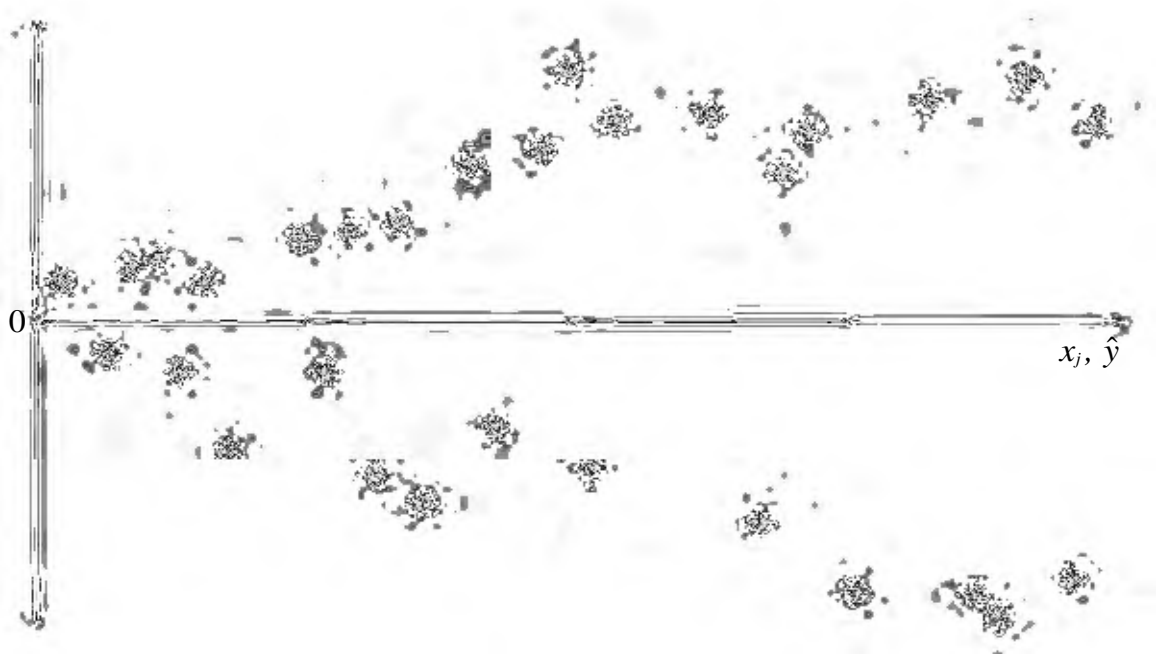

Figura 6.10 Grafico dei residui in presenza di eteroschedasticità.

Altri casi particolari

Ricordiamo altri casi particolari sui grafici dei residui. Se per esempio si presenta una forma come quella rappresentata nella Figura 6.11, allora siamo in presenza di autocorrelazione positiva degli errori[4].

[4] Se le ipotesi di omoschedasticità e incorrelazione non sono soddisfatte esistono stimatori più efficienti dello stimatore dei minimi quadrati.

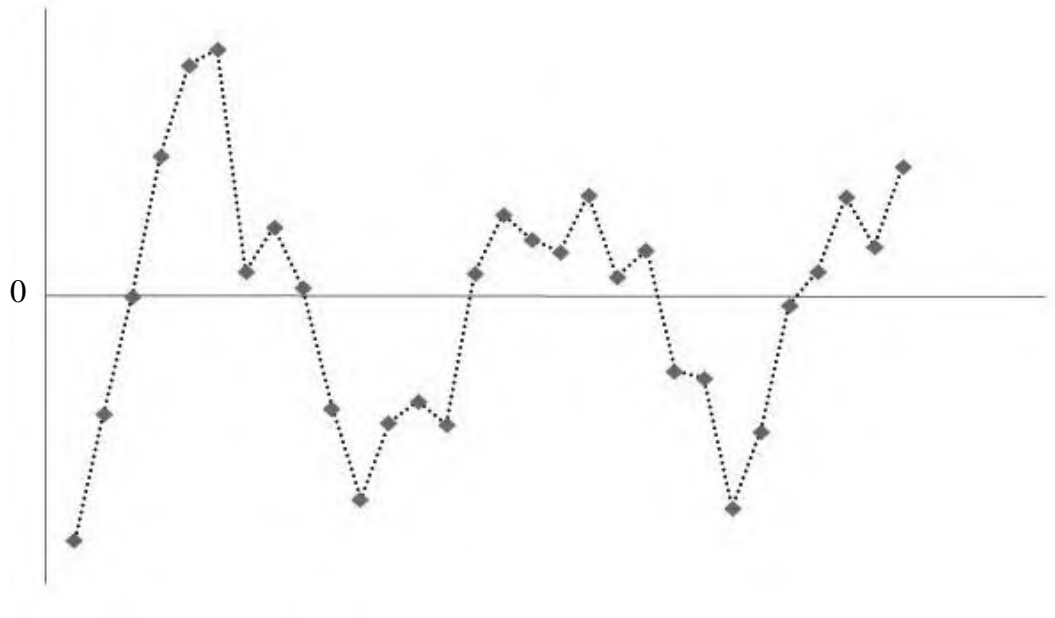

Figura 6.11 Residui in presenza di autocorrelazione.

La Figura 6.12 rappresenta il caso in cui il modello di regressione non ha intercetta, e pertanto i residui hanno media non nulla[5]. Diversamente, se nel modello è stata omessa una variabile esplicativa, i residui presentano una struttura come quella rappresentata nella Figura 6.13.

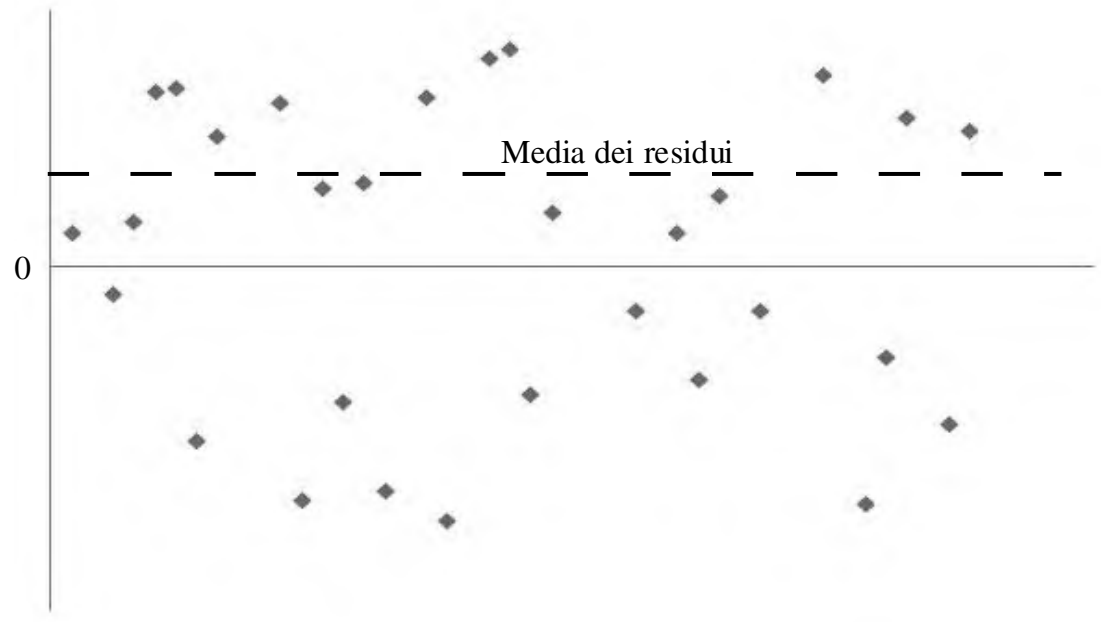

Figura 6.12 Residui nel caso di una stima del modello senza intercetta.

[5] Ciò perché nella stima dei parametri non si impone il vincolo che $\Sigma \hat{\epsilon}_i = 0$.

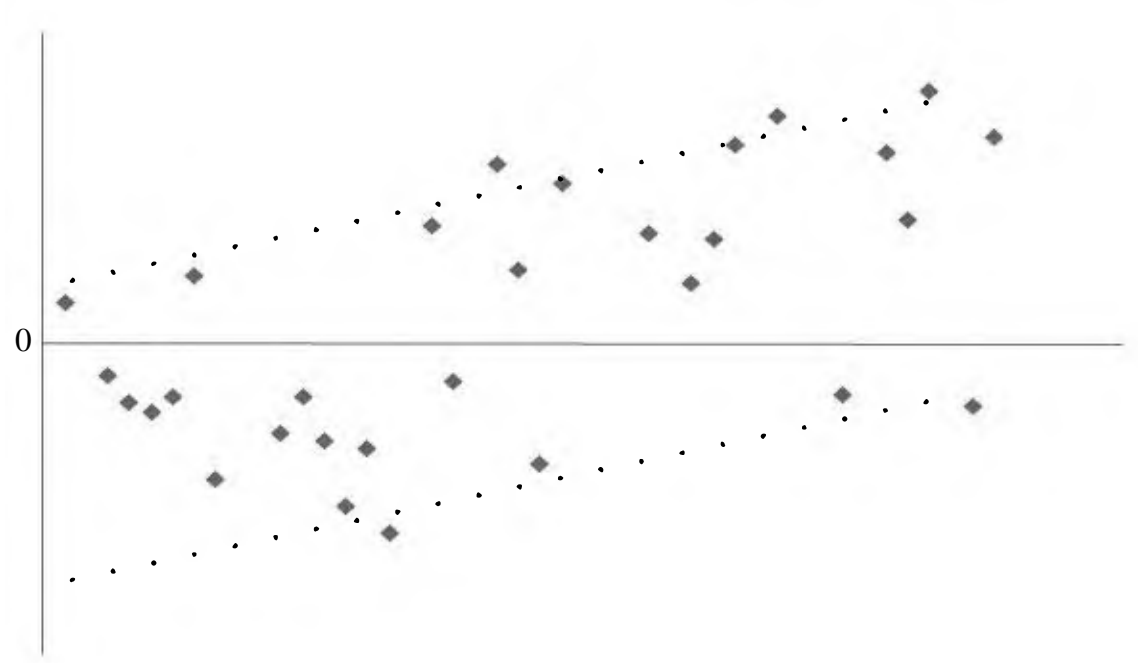

Figura 6.13 Residui nel caso di una stima del modello con una variabile esplicativa omessa.

Un cambiamento strutturale nella relazione fra la variabile dipendente e le variabili esplicative invece, in cui le osservazioni sono divise in due gruppi generati da due modelli con diversi valori dei parametri, dà luogo a gruppi di residui (Figura 6.14).

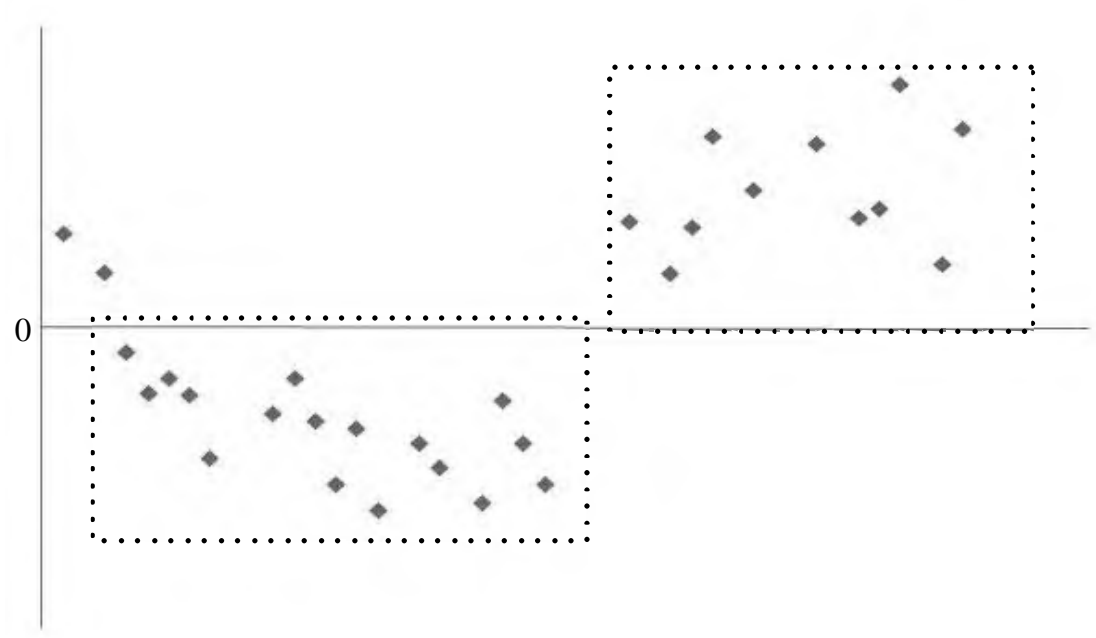

Figura 6.14 Residui in presenza di cambiamenti strutturali.

6.2.4 Casi di relazioni non lineari e presenza di dati anomali

Il coefficiente di correlazione non risulta più una misura di intensità (e segno) del legame tra variabili sia quando la relazione fra le variabili è di tipo non lineare, sia in presenza di valori anomali che sono molto diversi dal resto dei valori osservati della variabile.

Nella Figura 6.15 si riportano alcune tipiche situazioni in cui la correlazione pur potendosi calcolare come misura, non esprime correttamente il legame lineare tra le due variabili.

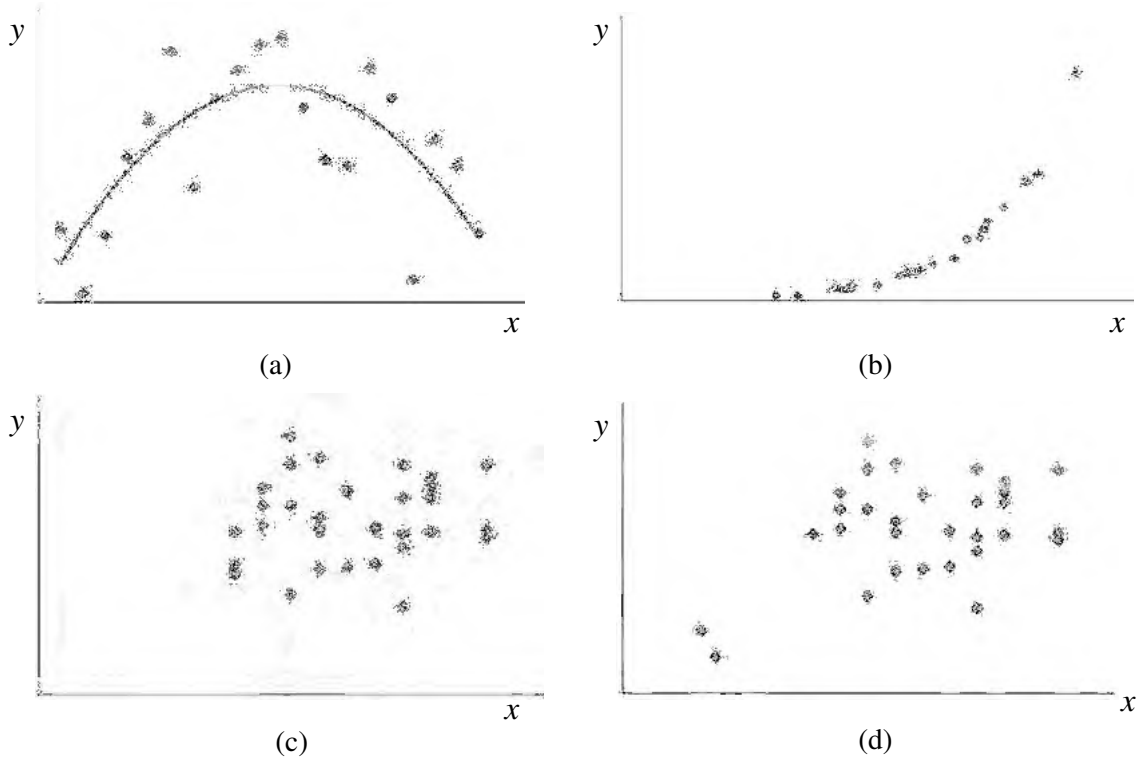

Figura 6.15 Relazioni non lineari tra variabili e presenza di valori anomali.

In particolare, nei quattro diagrammi (a, b, c, d) a dispersione sopra esposti sono rappresentate situazioni come (a) una relazione quadratica, (b) una relazione esponenziale, (c) una relazione quasi nulla, (d) una relazione quasi nulla con la presenza anche di valori anomali; le due unità sono molti distanti dal corpo dei dati e riducono ulteriormente il valore del coefficiente di correlazione. Tali valori potrebbero addirittura appartenere a un campione diverso (o popolazione diversa da quella da cui è stato estratto il campione dei dati).

Quando le relazioni non sono lineari occorre utilizzare una forma funzionale appropriata che esprima la vera relazione esistente tra le variabili. Le funzioni possono essere quadratiche oppure polinomiali, sebbene queste ultime siano nella

pratica di scarsa utilità, per la difficoltà di interpretare il significato economico dei coefficienti. La funzione di una relazione quadratica, invece, può essere di utilità pratica, anche per la costruzione del diagramma del punto di efficienza. Infatti, come precedentemente ricordato, i costi variabili di produzione non necessariamente sono crescenti in modo lineare, ma piuttosto variano in modo più o meno che proporzionalmente (si veda la Figura 6.2), soprattutto se l'arco temporale considerato è un lungo periodo di approvvigionamento e di lavoro (in tal caso sono presenti fluttuazioni dei prezzi dei materiali, della manodopera e dei prodotti finiti).

L'espressione del modello di regressione quadratica include anche un termine al quadrato:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \quad \text{per } i = 1, \ldots, n$$

dove β_0 è l'intercetta del modello, β_1 è il coefficiente dell'effetto lineare di x su y, β_2 è il coefficiente dell'effetto quadrato di x su y.

Se β_1 è positivo, a fronte di un incremento unitario di x la funzione cresce di β_1, ma questo incremento varia di $2 \cdot \beta_2$ al variare unitario di x; se β_1 è minore di $2 \cdot \beta_2 x$, il modello declina. Se β_2 è negativo, allora l'incremento di y decresce di $2 \cdot \beta_2$ per un incremento unitario di x il decremento di y aumenta di $2 \cdot \beta_2$. Il modello viene stimato con il metodo dei Minimi Quadrati Ordinari. Un esempio di analisi di regressione quadratica è presentato nella piattaforma MyLab.

Un altro tipo di relazione non lineare molto utilizzato negli studi di fenomeni aziendali, economici e sociali (frequentissimo negli studi biologici ed epidemiologici) è il modello di regressione logistica.

$$P(y \mid x) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}}$$

Tale funzione modella la relazione tra una variabile dipendente y dicotomica (esito dicotomico come pezzo difettoso: sì/no), e un insieme di variabili $x_1, \ldots, x_k$ che possono essere dicotomiche (genere, assenza/presenza di un additivo) o categoriche (tipo di materiale, classe salariale, titolo di studio) o continue (età, livello di temperatura, voci di bilancio).

La funzione logistica verrà trattata successivamente nel Capitolo 8, e alcuni esempi sono presentati nella piattaforma MyLab.

Nella Figura 6.16, si rappresenta il caso in cui la relazione sottostante il modello di regressione non sia lineare, e pertanto anche i residui presentano un andamento non lineare.

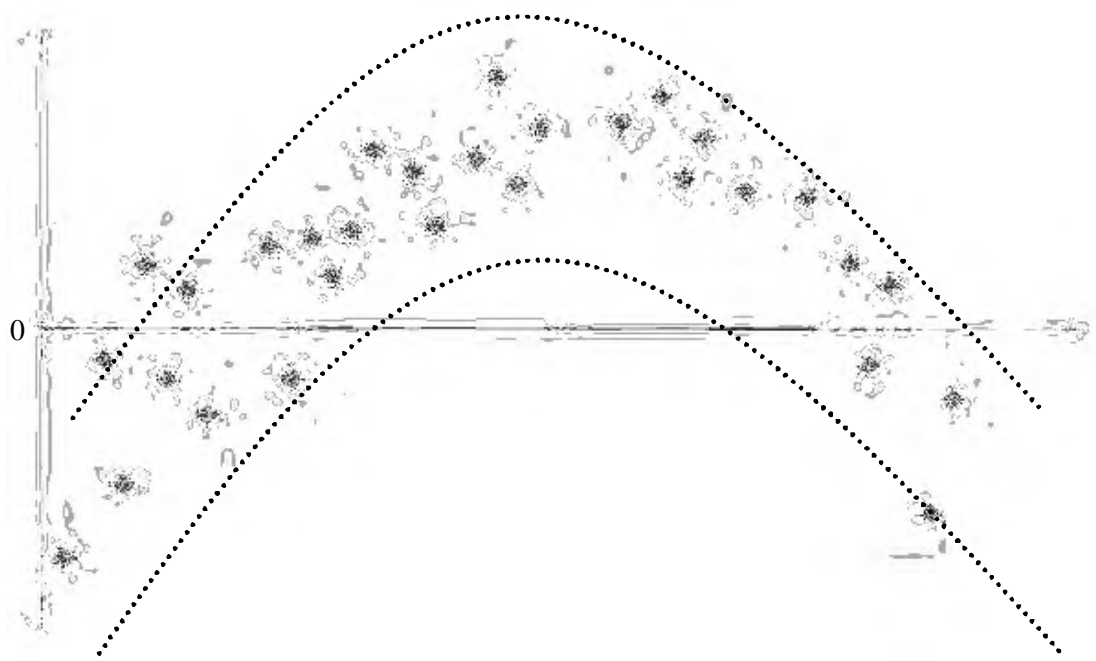

Figura 6.16 Residui in presenza di una relazione non lineare.

Valori anomali (outliers)

La presenza di valori anomali, detti outliers, nei dati (valori distanti dal corpo dei dati), genera la presenza di residui corrispondenti ai valori anomali molto distanti dagli altri residui (Figura 6.17).

In presenza di valori anomali nel corpo dei dati l'analisi statistica di regressione può procedere in vari modi: stimare il modello di regressione lineare (semplice

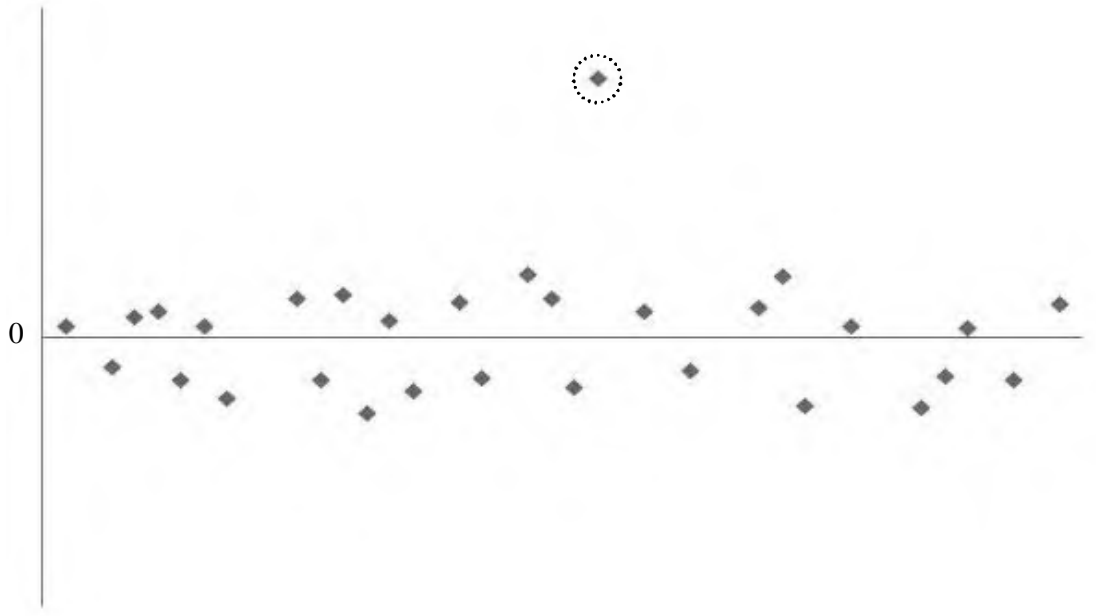

Figura 6.17 Residui in presenza di valori anomali.

o multiplo o non lineare) dopo avere preventivamente rimosso i valori anomali. Le tecniche di rimozione degli outliers non sono esenti da errori a causa di esclusioni errate ed errati mantenimenti conducendo a un framework altrettanto improprio quanto quello iniziale, che sconsiglia l'applicazione della teoria classica sulla normalità. Oppure, è possibile ricorrere a metodi di stima robusti (o resistenti) così qualificati in letteratura per la loro proprietà di produrre stime non facilmente condizionabili da dati contaminati. Questi metodi concordano nell'identificare come outliers le unità che evidenziano i residui più elevati. Tra i vari approcci robusti proposti, una menzione particolare spetta al metodo dei minimi quadrati mediani (Least Median of Square LMS – Rousseeuw, 1984) per via della sua intuitività e facilità d'uso. Gli stimatori robusti (LMS, MAD, trimmed mean ecc.) presentano però l'inconveniente di sottopesare o tralasciare alcune osservazioni; inoltre, possono completamente fallire se le osservazioni non provengono da un'unica popolazione, ma da più popolazioni distinte.

Un approccio alternativo al problema è quello di avvalersi delle cosiddette analisi diagnostiche[6], che prevedono il calcolo di statistiche in grado di individuare le anomalie e tra queste quelle influenti. Queste possono essere così esaminate, e successivamente eliminate o corrette, in modo da consentire il riadattamento del modello mediante le tecniche classiche.

Esempio 6.2

Dopo avere effettuato un'analisi di correlazione, il cui risultato ha prodotto un coefficiente molto alto (0,89), l'azienda del settore alimentare, di cui all'esempio precedente, si pone l'obiettivo di individuare una relazione funzionale tra le due variabili costi totali e volume di produzione, stimando un modello di regressione lineare semplice ovviamente utilizzando i dati della Tabella 6.1. I risultati della stima del modello sono riportati nella Tabella 6.2.

Tabella 6.2 Risultati della stima del modello lineare semplice.

	Coefficienti	Errore standard	Stat *t*	Valore di significatività	Intervallo di confidenza al 95%	
Intercetta	10,35	2,22	4,67	0,15·10–3	5,73	14,97
Produzioni annue in centinaia di tonnellate*	0,54	0,06	8,75	0,03·10–6	0,41	0,67

* I dati originari presentati nella Tabella 6.1 sono stati riproporzionati rispetto all'ordine di grandezza della variabile costi, pertanto le produzioni annue sono espresse in centinaia di tonnellate.

[6] Una tra le più recenti metodologie, che consentono di individuare outlier e strutture insospettabili nei dati e quindi utili per la costruzione di modelli robusti, è il metodo iterativo Forward Search (FS); (Atkinson A. C., Riani M., 2000; Atkinson A. C., Riani M. and Cerioli A., 2004).

Il primo risultato da commentare sono le stime dei parametri e la loro significatività. Il valore dell'intercetta $b_0 = 10{,}35$ esprime il valore dei costi quando la produzione è nulla. In altri termini, l'azienda sostiene dei costi fissi anche se non produce, situazione plausibile che rispecchia quanto detto nella presentazione del break-even point.

Il coefficiente di inclinazione della retta esprime invece la variazione dei costi in funzione di una variazione unitaria della produzione e quindi l'influenza dei costi variabili. Più specificatamente, l'aumento di 100 tonnellate di produzione determina in media un aumento di 0,54 milioni di euro di costi, all'interno del campo di variazione dei valori della variabile indipendente. Questi risultati sono validi all'interno del campo di variazione della variabile volume di produzione. Nella Figura 6.18 riportiamo lo scatterplot delle unità campionarie e della retta di regressione stimata.

Queste stime sono risultate altamente significative come si può osservare dai valori della significatività del test, riportati nella Tabella 6.2, inoltre l'intervallo di confidenza al 95% non contiene lo 0.

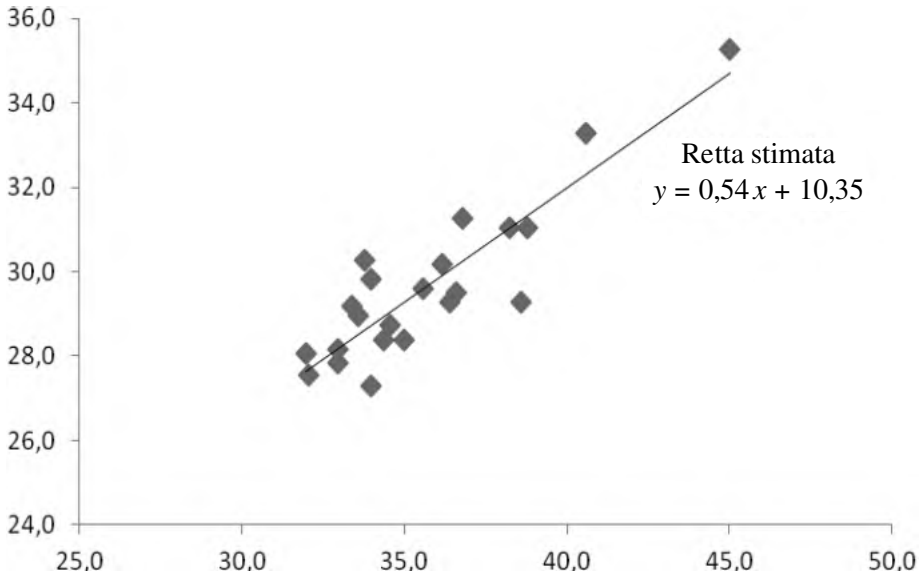

Figura 6.18 Unità campionarie e retta di regressione stimata.

L'azienda osserva il valore dell'indice di determinazione R^2 che è uguale al rapporto fra devianza di regressione e devianza totale della variabile dipendente, come indicatore descrittivo del grado di adattamento del modello ai dati campionari. In questo caso è $R^2 = 0{,}79$ e quindi la capacità del modello di rappresentare le relazione tra le due variabili è abbastanza buono, ma non conferma l'ipotesi che spesso viene adottata di diretta proporzionalità dei costi variabili al variare del volume di produzione. Comunque, il manager grazie a questo indicatore è in grado anche di valutare se la forma funzionale di primo ordine rispecchia le condizioni con costi variabili lineari sottostanti il modello di break-even point.

Con l'analisi empirica sui residui, mediante l'osservazione dei grafici del q-q normal plot e dei residui standardizzati rispetto alla variabile risposta stimata (Figure 6.19 e 6.20), il manager verifica la conformità del modello adottato e delle analisi svolte alle ipotesi inferenziali sottostanti al modello teorico.

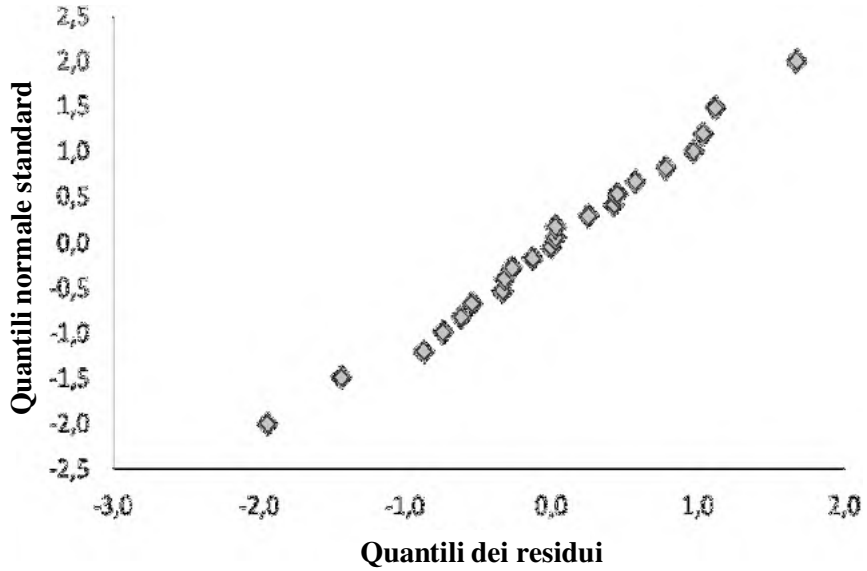

Figura 6.19 Q-qnormal plot dei residui.

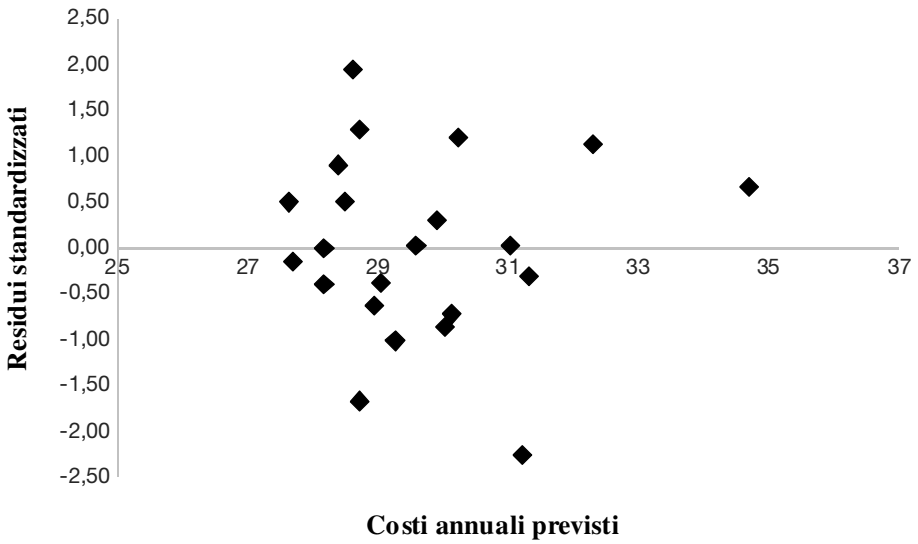

Figura 6.20 Grafico dei residui standardizzati.

Il risultato che si ottiene è che sia l'ipotesi di normalità che di omoschedasticità sono rispettate, anche se sono presenti due unità sospette di essere anomale. Su queste ultime il manager potrebbe indagare meglio e più approfonditamente per capire la loro natura (quali sono i motivi per cui sono diverse). ❑

6.3 Analisi e impiego della correlazione multipla e del modello di regressione lineare multipla

6.3.1 Analisi e misura della correlazione lineare multipla

Nella realtà aziendale come pure in contesti economici e sociali, data la natura complessa dei fenomeni, l'analisi della relazione tra variabili non si limita quasi mai allo studio di una singola coppia, ma piuttosto a due o più variabili insieme. Per esempio, pensare che il livello dei costi di produzione sia causato unicamente dal volume di produzione, è un'ipotesi piuttosto riduttiva. Così come il livello delle vendite non dipende da un solo fattore ma è contemporaneamente influenzato dal prezzo del prodotto, dai prezzi dei prodotti concorrenti e dalle spese promozionali. Al manager interessa l'analisi simultanea della relazione tra la variabile di interesse e due o più variabili che la influenzano, in modo che i risultati ottenuti siano uno strumento per prendere delle decisioni e intraprendere le azioni strategiche per modificare la variabile di interesse. Pertanto in queste circostanze, il manager di azienda dovrà implementare uno studio della correlazione e successivamente della regressione lineare tenendo conto di tutti i fattori che entrano in gioco nello studio di un determinato fenomeno. Il modello di regressione semplice non risponderà più a questa esigenza ma occorrerà invece utilizzare una sua estensione.

Analisi della correlazione multipla

Quando l'analisi della correlazione interessa più di due variabili, l'indice di correlazione può essere calcolato per tutte le possibili coppie di variabili, per esempio siano X, Y e Z, avremo la correlazione di x_1 vs y; x_2 vs y; x_1 vs x_2, e il risultato ottenuto viene riportato in una matrice detta matrice di correlazione; alternativamente sarà possibile utilizzare l'indice di correlazione multipla: il coefficiente di correlazione multipla, detto anche coefficiente di determinazione R^2, misura la proporzione di variabilità totale che viene spiegata dall'insieme di predittori.

Nella correlazione misurata, come visto nell'analisi della regressione semplice, l'informazione sul legame è specifico per ogni singola coppia di variabili, e rende possibile fare il confronto sul segno ed entità tra tutte le possibili coppie di variabili. Nel secondo caso, invece, il coefficiente di correlazione multipla è definito come rapporto tra devianza dovuta alla regressione e devianza totale, e l'informazione che fornisce è una misura della relazione lineare che intercorre tra più di due variabili considerate insieme.

Esempio 6.3

Nel caso dell'azienda alimentare, supponiamo che il manager si ponga il seguente quesito: se prendo in esame i vari tipi di costi variabili, qual è l'influenza che essi hanno sui costi totali? E quindi in sostanza la domanda è: come si determinano in questo caso le inclinazioni delle rette del ricavo o dei costi totali e il punto di partenza della retta dei costi, punto che rappresenta i costi fissi?

Mentre per i ricavi la risposta è immediata, perché fissato il prezzo di vendita la retta lineare è facilmente definita partendo da zero (se le vendite sono nulle si ha un ricavo nullo) e con un'inclinazione tale che per ogni unità di prodotto i ricavi si alzino di una certa proporzionalità, per i costi invece, la risposta non è immediata. Uno dei metodi che si possono seguire per determinare l'ordinata all'origine (costi fissi) e l'inclinazione della retta (inclinazione che non è altro che il costo variabile medio unitario che viene supposto proporzionale) è il metodo dell'analisi di regressione. Tale metodo può essere utilizzato anche per determinare lo standard di singole voci di costo variabili, semi variabili o fisse.

Il manager intende "raffinare" l'analisi sui costi totali relazionandoli non solo al volume della produzione ma, per esempio, anche ai costi relativi al personale (il costo del lavoro include salari, contributi sociali e pensioni). È interessato a capire quanto pesano le spese del capitale umano sul totale delle spese, scorporate da quest'ultimo, oltre al volume della produzione. Nella Tabella 6.3 sono riportati tutti i costi che hanno contributo alla determinazione del costo totale (ultima colonna). La nuova variabile che si aggiunge all'analisi è perciò la voce totale costi del personale.

Tabella 6.3 Volume di produzione e voci di costi variabili.

Unità	Produz. annue in centinaia di ton	Salari	Contrib. sociali	Pensioni	Totale costi pers. (A)	Imposte	Amm.ti	Interessi	Altre spese	Totale altri costi (B)	Costi totali (A) + (B)
Ancona	35,58	9,85	2,34	1,40	13,59	0,85	11,78	1,66	1,73	16,02	29,61
Aosta	32,96	7,11	2,20	1,40	10,71	0,89	13,44	1,66	1,13	17,12	27,83
Bari	34,37	7,70	2,40	1,40	11,50	0,94	12,43	1,66	1,86	16,89	28,39
Bologna	33,37	9,66	1,88	1,40	12,94	0,83	11,72	1,66	2,02	16,22	29,17
Cagliari	36,18	6,61	2,64	1,40	10,65	0,72	9,46	1,66	7,68	19,52	30,17
Campobasso	33,77	3,87	1,54	1,40	6,81	0,64	8,06	1,66	13,11	23,47	30,28
Catanzaro	31,96	4,87	1,94	1,40	8,21	0,67	8,72	1,66	8,80	19,85	28,06
Firenze	40,60	5,31	2,12	1,40	8,83	0,65	8,92	1,66	13,21	24,45	33,28
Genova	38,59	6,08	2,44	1,40	9,92	0,69	9,56	1,66	7,44	19,35	29,28
L'Aquila	36,58	8,03	3,22	1,40	12,65	0,69	11,10	1,66	3,42	16,86	29,51
Milano Nord	36,78	8,90	3,56	1,40	13,86	0,90	11,92	1,66	2,93	17,42	31,28
Milano Sud	38,25	5,51	2,21	1,40	9,12	0,69	9,70	1,66	9,89	21,94	31,06
Napoli	33,97	7,71	2,52	1,40	11,63	0,78	9,67	1,66	6,09	18,20	29,83
Palermo	34,97	8,14	3,78	1,40	13,32	0,73	9,01	1,66	3,66	15,07	28,39
Perugia	32,96	3,81	1,89	1,40	7,10	0,71	9,26	1,66	9,44	21,07	28,17

(*segue*)

Tabella 6.3 Volume di produzione e voci di costi variabili. (*continua*)

Unità	Produz. annue in centinaia di ton	Salari	Contrib. sociali	Pensioni	Totale costi pers. (A)	Imposte	Amm.ti	Interessi	Altre spese	Totale altri costi (B)	Costi totali (A) + (B)
Potenza	36,38	5,41	2,56	1,40	9,37	0,81	9,97	1,66	7,46	19,90	29,28
Roma Nord	38,79	10,54	3,59	1,40	15,53	0,90	11,11	1,66	1,86	15,53	31,06
Roma Sud	45,02	10,99	3,72	1,40	16,11	0,83	12,01	1,66	4,66	19,16	35,27
Torino	33,97	6,59	3,01	1,40	11,00	0,76	10,58	1,66	3,28	16,27	27,28
Trento	34,57	5,00	1,78	1,40	8,18	0,74	9,64	1,66	8,51	20,54	28,72
Trieste	32,06	4,36	1,94	1,40	7,70	0,82	9,27	1,66	8,11	19,86	27,56
Venezia	33,57	5,13	2,06	1,40	8,59	0,86	8,58	1,66	9,26	20,36	28,94

Per effettuare l'analisi suddetta viene innanzitutto calcolata la matrice di correlazione:

	Produzioni annue in centinaia di tonnellate	Totale costi personale	Totale altri costi
Produzioni annue in centinaia di tonnellate	1		
Totale costi personale	0,50 (*p-value* = 0,017)	1	
Totale altri costi	0,13 (*p-value* = 0,56)	–0,73 (*p-value* = 0,000)	1

Come risulta dalla tabella i costi totali hanno una relazione negativa forte (–0,73) con il volume della produzione, e una più bassa (0,50) e positiva con il costo del lavoro; mentre vi è un legame molto scarso e non significativo con il volume della produzione.

Come sappiamo la matrice di correlazione presenta dei risultati che riguardano le singole coppie di variabili, e in questo caso suggerisce di non stimare un modello di regressione tra produzione e totale altri costi. In effetti, se il manager stimasse tale modello otterrebbe un risultato non utilizzabile:

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Intervallo di confidenza al 95%*	
Intercetta	15,0571	6,50049	2,3163	0,031271318	1,497314	28,61688	1,497314
Produzioni annue in centinaia di ton	0,10673	0,18146	0,58818	0,562996706	–0,27179	0,485249	–0,27179

La variabile Produzioni annue, come già annunciato dal coefficiente di correlazione, non è significativa. Al manager tuttavia, interessa effettuare un'analisi più complessa sul totale degli altri costi in cui sono coinvolte contemporaneamente sia le produzioni annue che il costo del lavoro; inoltre, potrebbero cambiare i valori di stima e la significatività nelle produzioni annue. ❑

Analisi della regressione lineare multipla

La regressione multipla è usata per scopi predittivi, cioè sulla base delle correlazioni tra le variabili, si cerca di prevedere la risposta sulle unità statistiche per descrivere il comportamento di un determinato fenomeno, per individuare modelli causali, cioè forme funzionali che includendo una serie di variabili spieghino la risposta osservata su un campione di soggetti, interpretando pertanto il comportamento di un determinato fenomeno oggetto di studio.

Anziché stimare una retta che sia la più vicina possibile a tutti i punti di uno spazio bidimensionale, occorrerà stimare una funzione lineare che individui un piano, se le variabili sono tre compresa la dipendente, o un iperpiano in uno spazio k dimensionale se le variabili sono più di tre, che sia il più vicino possibile a tutti i punti dello spazio tridimensionale o k dimensionale.

Il modello di regressione lineare quando esprime la dipendenza di una variabile risposta rispetto a due (o più) variabili indipendenti è detto modello di regressione multipla.

Nella Figura 6.21 riportiamo una rappresentazione grafica di un modello di regressione lineare multipla generico nel caso di due variabili indipendenti (come è noto, non è possibile costruire una rappresentazione grafica superiore a tre dimensioni).

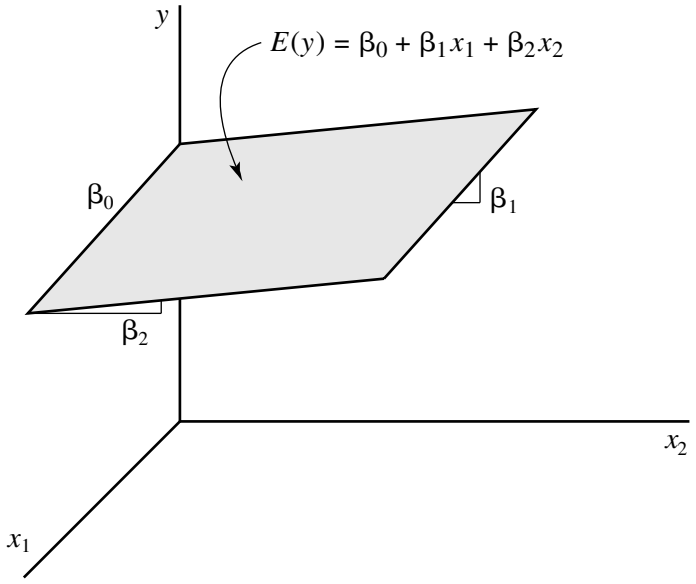

Figura 6.21 Grafico di un modello di regressione lineare multipla con due regressori.

In questo paragrafo sarà affrontata l'analisi delle regressione multipla nel caso di due covariate. Una nota metodologica sul modello a k variabili e sulla sua costruzione seguendo l'approccio stepwise è disponibile nella piataforma MyLab.

6.3.2 Modello di regressione lineare multipla

Vediamo le specificazioni di questo modello in generale. L'espressione della regressione è la stessa di come avevamo già visto $y = f(x_1, x_2) + \epsilon_i$, ma ovviamente con incluse la variabile x_1 e x_2, e con l'errore ϵ_i componente casuale non osservabile. L'errore rappresenta tutti quegli elementi di cui non è possibile valutare gli effetti e che hanno, nel complesso, un'azione di disturbo.

La forma funzionale che ci interessa è $f(x_1, x_2) = \beta_0 + \beta_1 h(x_1) + \beta_2 g(x_2)$ che è lineare nei parametri β_0, β_1, β_2. Alcuni esempi di funzioni lineari nei parametri sono:

1) $\beta_0 + \beta_1 x_1 + \beta_2 x_2$

2) $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$

3) $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ con $x_2 = x_1^2$

4) $\beta_0 + \beta_1 \log(x_1) + {}_2\log(x_2) = \log[A(x_1)^{\beta_1}(x_2)^{\beta_2}]$ e $\beta_0 = \log(A)$

Specificazione della componente deterministica

Specificazione del modello di regressione lineare multiplo di I ordine

Per un campione di n unità statistiche e sia i l'i-esimo elemento del campione, la specificazione del modello è

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

Le assunzioni sulla componente di disturbo sono riportate qui di seguito.

a) $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ per $i = 1, \ldots, n$

b) $E(\epsilon_i) = 0 \rightarrow E(Y_i \mid x_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \quad \forall i = 1, \ldots, n$

c) $V(\epsilon_i) = \sigma^2 I_n$ (dove I_n indica la matrice identità di ordine n) $\rightarrow V(Y_i) = \sigma^2 I_n$ per $\forall i = 1, \ldots, n$

d) La distribuzione di ϵ_i è normale con varianza omoschedastica:

$$\epsilon_i \sim NMV(0, \sigma^2 I_n) \rightarrow Y_i \sim NMV(\beta_0 + \beta_1 x_i + \beta_2 x_i, \sigma^2 I_n)$$

e) Gli errori ϵ_i sono indipendenti, vale a dire che il disturbo associato all'osservazione i-esima non ha alcun effetto su quello dell'osservazione j-esima: $Cov(\epsilon_i, \epsilon_j) = 0$ per $\forall i \neq j$ con $i = 1, \ldots, n$.

Le variabili x_{1i} e x_{2i} sono valori prefissati e y_i, essendo una combinazione lineare di una parte deterministica (variabili prefissate) e una componente stocastica costituita dall'errore, è essa stessa una variabile casuale (v.c.) che per effetto delle assunzioni fatte sulla componente di disturbo (errore), ha le seguenti caratteristiche:

a) il suo valore atteso è uguale alla componente deterministica $E(y_i)= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$;

b) la sua varianza è omoschedastica $V(y_i) = \sigma^2 I_n$; è la $\sigma^2 I_n$: variabilità dei valori di y intorno alla media $E(y_i)$;

c) i suoi valori sono tra loro incorrelati $Cov(y_i, y_j) = 0$ per $\forall\, i \neq j$;

d) la sua distribuzione è normale, in particolare $y_i \sim NMV(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}, \sigma^2 I_n)$.

Che cosa rappresentano β_1 e β_2? Sono i coefficienti che misurano gli effetti netti, vale a dire parziali delle variabili esplicative sulla variabile dipendente. In particolare, β_1 misura la variazione di $E(y_i)$ dovuta a una variazione unitaria di x_1 tenuta costante la variabile x_2; mentre β_2 misura la variazione di $E(y_i)$ dovuta a una variazione unitaria di x_2 tenuta costante la variabile x_1. Il termine noto β_0 misura $E(y)$ in corrispondenza di $x_1 = 0$ e $x_2 = 0$.

Il piano dei Minimi Quadrati Ordinari e la stima dei parametri

I coefficienti del modello di regressione multipla sono definiti seguendo il criterio del metodo dei Minimi Quadrati Ordinari utilizzato nella regressione semplice. Pertanto, le stime dei coefficienti β_0, β_1, β_2, definite come b_0, b_1 e b_2, sono funzioni dei valori osservati di y, x_1 e x_2, e tale per cui è minima la somma dei quadrati dei residui $e_i = y_i - \hat{y}_i$: min *sse* (*sum of squared errors*) dove $sse = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2}) \right]^2$.

Le condizioni del primo ordine sono tre e producono tre equazioni in tre incognite:

$$\frac{\delta sse}{\delta b_0} = 0, \qquad \frac{\delta sse}{\delta b_1} = 0, \qquad \frac{\delta sse}{\delta b_2} = 0$$

Dalla risoluzione del sistema di tre equazioni si ottengono così le tre espressioni di b_0, b_1 e b_2

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

$$b_1 = \frac{Cov(x_1, y)Var(x_2) - Cov(x_2, y)Cov(x_1, x_2)}{Var(x_1)Var(x_2) - \left[Cov(x_1, x_2)\right]^2}$$

$$b_2 = \frac{Cov(x_2, y)Var(x_1) - Cov(x_1, y)Cov(x_1, x_2)}{Var(x_1)Var(x_2) - \left[Cov(x_1, x_2)\right]^2}$$

Le espressioni degli stimatori dei parametri β_0, β_1, β_2 sono definite come

$$B_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$B_1 = \frac{Cov(X_1,Y)Var(X_2) - Cov(X_2,y)Cov(X_1,X_2)}{Var(X_1)Var(X_2) - Cov(X_1,X_2)^2}$$

$$B_2 = \frac{Cov(X_2,Y)Var(X_1) - Cov(X_1,Y)Cov(X_1,X_2)}{Var(X_1)Var(X_2) - \left[Cov(X_1,X_2)\right]^2}$$

L'espressione di β_0 è una semplice estensione dell'espressione dell'intercetta nella regressione semplice. Mentre le formule di β_1 e β_2 sono più complesse di quella relativa alla pendenza della retta di regressione lineare semplice.

La stima corretta della varianza σ^2 viene ricavata attraverso i residui dei MQO e precisamente

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-3}$$

La somma dei quadrati dei residui viene divisa per la grandezza [(numero di osservazioni – numero dei parametri – 1)]. Quindi la grandezza

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-3}$$

è lo stimatore corretto di σ^2, cioè $E(S^2) = \sigma^2$.

Gli stimatori B_0, B_1 e B_2 sono variabili casuali indipendenti da S^2. Queste espressioni sono importanti per formulare test delle ipotesi sui coefficienti del modello.

BOX 6.2 Proprietà degli stimatori dei Minimi Quadrati Ordinari

Gli stimatori B_0, B_1 e B_2 dei parametri β_0, β_1 e β_2, funzioni della v.c. Y e dell'errore ϵ_i, sono definiti come stimatori BLUE, cioè hanno la varianza minima nella classe degli stimatori Lineari e Corretti:

B: Best → varianza minima fra i Linear Unbiased Estimators (LUE) (teorema di Gauss-Markov)

L: Linear → funzioni lineari della componenti di errore

U: Unbiased (non distorto) → $E(B_0) = \beta_0 E(B_1) = \beta_1 E(B_2) = \beta_2$

E: Estimator (stimatore)

Nell'ipotesi che $\epsilon_i \sim NVM(0, \sigma^2 I_n)$, le distribuzioni degli stimatori B_0, B_1, B_2 e S_e^2 sono:

- $B_0 \sim N(\beta_0, \sigma^2 h_0)$
- $B_1 \sim N(\beta_1, \sigma^2 h_1)$
- $B_2 \sim N(\beta_2, \sigma^2 h_2)$
- $(n-3)S_e^2 / \sigma^2 \sim \chi^2_{n-3}$

dove h_0, h_1, h_2 sono funzioni[7] dei valori x_{1i} e x_{2i}; le varianze di B_0, B_1, B_2 sono funzioni di σ^2 e quindi devono essere stimate attraverso S^2.

La bontà di adattamento del modello

Nell'analisi della regressione semplice abbiamo visto come l'indice di determinazione R^2 sia una misura della bontà di adattamento del modello ai dati. Quando si confrontano modelli di regressione lineare con un diverso numero di variabili esplicative, l'indice R^2 deve essere utilizzato con cautela. Accade, infatti, che l'inclusione di un'ulteriore variabile esplicativa nel modello incrementa il valore della devianza di regressione (o spiegata, Explained Sum of Squares), anche quando la variabile inclusa non è significativa per spiegare le variazioni di Y, mentre non altera il valore della variabilità totale TSS (Total Sum of Squares). Di conseguenza, R^2 dato che è il rapporto tra la devianza di regressione e la devianza totale, non diminuisce, anzi, generalmente cresce.

Per mitigare questo effetto si introduce l'indice di determinazione multipla corretto, detto $\bar{R}^2$ corretto, che tiene conto del numero delle variabili indipendenti presenti nel modello, includendo nell'espressione di R^2 i gradi di libertà rispettivamente della devianza di regressione e della devianza totale, come segue:

$$\bar{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

Il secondo termine dell'indice è il rapporto tra la stima non distorta della varianza degli errori con la stima non distorta della varianza della variabile dipendente.

La scelta di un modello tra diversi è legata al valore di $\bar{R}^2$ corrispondente: maggiore è l'indice, migliore sarà la bontà del modello. L'inclusione di un ulteriore variabile nel modello produce un aumento di $\bar{R}^2$ in quanto ne aumenta la capacità di adattamento, ma nel contempo la quantità $(n-k-1)$ diminuisce. Tuttavia, se quest'ultima, a fronte di un ingresso di una nuova variabile, viene compensata da una riduzione della devianza residua (Residual Sum of Squares), allora è preferibile il modello con un più elevato numero di regressori.

[7] Il termine h_j, con $j = 0, 1, 2$, è l'elemento posizionato sulla $(j+1)$-esima cella della diagonale principale della matrice ottenuta come inversa della matrice prodotto tra la matrice dei dati trasposta e la stessa matrice dei dati. Nella matrice dei dati è compresa una colonna di unità che individua la posizione del termine noto (la colonna $j = 0$).

La verifica dell'utilità del modello

Una volta stimato il modello, è importante verificare l'esistenza o meno di un legame lineare tra la variabile dipendente y e le variabili indipendenti $x_1 \ldots x_k$ che sono dentro al modello; in altri termini, si intende valutare la significatività dei parametri considerati *congiuntamente*, verificando l'ipotesi di indipendenza lineare

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

contro l'ipotesi alternativa

H_1 : almeno un'uguaglianza in H_0 è vera.

Se l'ipotesi nulla è verificata, cioè H_0 non è rifiutata, il test conduce all'individuazione ed esclusione di quelle variabili che non spiegano la dipendenza lineare di Y. Viceversa, se l'ipotesi nulla viene rifiutata, significa che almeno uno dei regressori contribuisce a spiegare, in termini di relazione lineare, la variabilità della variabile dipendente y. Occorrerà pertanto fare dei test di significatività separati sui coefficienti dei singoli regressori.

Analogamente a quanto già visto per la regressione semplice, la varianza totale di y può essere scomposta in due distinte varianze parziali, di regressione e residua. Ricordiamo che partendo dalla definizione della devianza totale some somma della devianza di regressione con la devianza residua, TSS (Total Sum of Squares) = ESS (Explained Sum of Squares) + RSS (Residual Sum of Squares):

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

dove $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$, possiamo ottenere le varianze dividendo le devianze per i rispettivi gradi di libertà:

$$\text{var}(Y) = \frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1} = s_Y^2 \quad \text{(TSS)}$$

$$\text{var}(Y)_{reg} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{k} = s_{reg}^2 \quad \text{(ESS)}$$

$$\text{var}(Y)_e = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-k-1} = s_e^2 \quad \text{(RSS).}$$

Se sono soddisfatte le condizioni di normalità distributiva della Y, di omoschedasticità e di indipendenza delle osservazioni, posta l'ipotesi di indipendenza lineare, le due variabili casuali campionarie s_{reg}^2 e s_e^2 sono entrambe stimatori corretti della varianza σ^2.

$$\frac{k s_{reg}^2}{\sigma^2} \sim \chi_k^2 \qquad (n-k-1)S_e^2 / \sigma^2 \sim \chi_{n-k-1}^2$$

La statistica test per la verifica di H_0 si basa sul rapporto tra le varianze parziali:

$$\text{Statistica test} = \frac{s^2_{reg}}{s^2_e} \sim F_{k,n-k-1}$$

Tale statistica, detta F, si distribuisce come una F di Fisher con k e $n-k-1$ gradi di libertà (un esempio di distribuzione di Fisher è rappresentato nella Figura 6.22).

Quando non vale l'ipotesi di indipendenza lineare il rapporto F tende a crescere; per esempio si consideri una probabilità fissata a priori $\alpha = P(F_{k,n-k-1} > F^*)$ di rifiutare H_0 quando è vera, dove F^* è il valore critico dedotto dalle tavole della distribuzione F di Fisher. Se il valore di F_{calc} calcolato con i dati è maggiore del valore F^* ($F_{calc} > F^*$), allora si rifiuta H_0 e si può ritenere che la variabilità spiegata dal modello sia significativamente più elevata della variabilità residua. In sintesi:

- si rifiuta H_0 se $p\text{-}value < \alpha$
- non si rifiuta H_0 se $p\text{-}value > \alpha$, dove $p\text{-}value = P(F_{k,n-k-1} > F_{calc})$

La statistica F può anche essere espressa in funzione dell'indice di determinazione R^2:

$$F(k, n-k-1) = \frac{ESS/k}{RSS/(n-k-1)} = \frac{\dfrac{ESS}{TSS}\Big/k}{\dfrac{RSS}{TSS}\Big/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

F è funzione monotona crescente di R^2. All'aumentare di R^2, il numeratore aumenta e il denominatore diminuisce, e quindi F cresce.

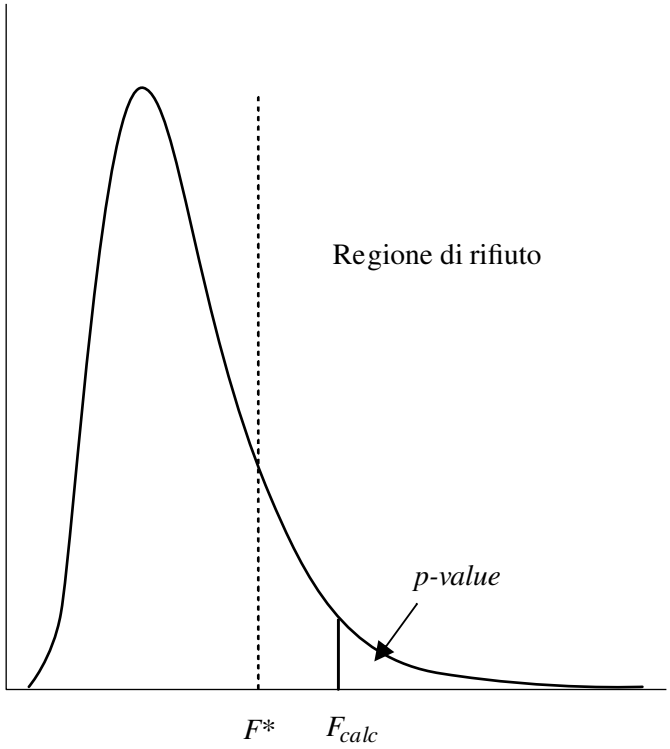

Figura 6.22 Distribuzione *F* di Fisher (un esempio).

Si ricorda che quando nel modello di regressione è compreso un solo regressore, il valore della statistica test t sulla significatività del parametro è uguale alla radice quadrata del valore della statistica test F sulla significatività complessiva del modello: $t = \sqrt{F}$.

Attenzione: le proprietà degli stimatori e i test delle ipotesi sono validi se sono verificate le proprietà della componente stocastica ϵ_i. È necessario pertanto procedere alla verifica delle assunzioni fatte riguardo all'errore.

La significatività delle stime

Se l'ipotesi nulla relativa alla significatività complessiva dei coefficienti viene rifiutata, l'analisi inferenziale passa al test sulla significatività dei singoli parametri.

Analogamente a quanto visto nella regressione lineare semplice, anche in questo caso in presenza di più variabili indipendenti, si ricorre alla procedura del test di ipotesi per rispondere alla domanda se ciascuna variabile esplicativa ha singolarmente un effetto significativo o meno sulla variabile dipendente. L'ipotesi per verificare la significatività di un singolo coefficiente è

$$H_0 : \beta_j = 0, \text{ con } j = 1, ..., k$$

se questa ipotesi non viene rifiutata, la variabile x_j non ha un potere esplicativo e può essere eliminata dal modello. Le statistiche test per verificare l'ipotesi nulla per ciascun coefficiente β_j sono

$$\frac{|\beta_j|}{S\sqrt{h_j}} \sim t_{n-k-1}$$

con $j = 1, ..., k$.

In sintesi, data l'ipotesi $H_0 : \beta_j = 0$ contro l'ipotesi alternativa $H_1 : \beta_j \neq 0$ (test bidirezionale) con un livello di significatività α, occorre individuare la regione di rifiuto ottenuta dai valori della statistica t che sono maggiori del valore teorico $t_{\alpha/2,n-k-1}$, dedotto dalla tavola delle probabilità della distribuzione t di Student.

Data l'ipotesi nulla, la regione di rifiuto di $H_0 : \beta_j = 0$ è l'area sotto la curva della distribuzione t Student con $\alpha/2, n-k-1$ gradi di libertà, delineata dai valori $|t| > t_{\alpha/2,n-k-1}$. La precedente Figura 6.7 mostra la regione di rifiuto per un test bidirezionale.

Nel caso del test bidirezionale, il *p-value* è la probabilità che una variabile casuale t Student con $g = n-k-1$ gradi di libertà generi una realizzazione maggiore in valore assoluto di quella ottenuta *p-value* = $P(|t_{n-k-1}| > |t_{calc}|)$. Si noti che $P(|t_{n-k-1}| > |t_{calc}|) = \alpha = \Pr(\text{rifiutare } H_0 | H_0)$, quindi gli intervalli $[-\infty, -t_{\alpha/2,n-k-1})$ e $(t_{\alpha/2,n-k-1}, \infty]$ sono *la regione di rifiuto di* H_0 e di conseguenza *la regione di non rifiuto di* H_0 è $[-t_{\alpha/2,n-k-1}, t_{\alpha/2,n-k-1}]$.

6.3.3 L'impiego del modello di regressione lineare in presenza di una variabile dicotomica

Nello studio della relazione tra variabili non sempre ci troviamo di fronte a fenomeni misurati solo su scala quantitativa. Una variabile Y (dipendente) potrebbe dipendere anche da una variabile di tipo qualitativo (categorico), e il modello di regressione lineare può essere ancora impiegato. Consideriamo il caso di una variabile particolare, detta dicotomica (o dummy), denominata ad esempio come X_d, che assume solo due modalità, o che si riferisce alla presenza o meno di una determinata caratteristica, rappresentate in entrambe i casi da due valori:

$X_{di} = 0$ se sull'unità statistica i-esima non è presente una determinata caratteristica del fenomeno oggetto di indagine

$X_{di} = 1$ se sull'unità statistica i-esima è presente una determinata caratteristica del fenomeno oggetto di indagine.

Un approfondimento metodologico sull'impiego del modello di regressione in presenza di una o più variabili qualitative con più di due modalità è rimandato alla piattaforma MyLab.

Supponiamo, ad esempio, che un manager debba valutare il rendimento misurato in termini di vendite complessive dei prodotti (in migliaia di euro) della sua azienda rispetto a 10 diverse campagne pubblicitarie svolte via radio (in migliaia di euro), e rispetto anche a modifiche della confezione dei prodotti. Per il manager, quindi, è importante non soltanto valutare l'effetto sulle vendite delle spese promozionali ma anche quello dei cambiamenti della confezione. I dati di questo studio sono riportati nella seguente tabella.

Campagna pubblicitaria	Vendite (in migliaia di euro) y_i	Spese (in migliaia di euro) X_i	Modifiche alla confezione X_{di}
1	74	51	0
2	70	68	1
3	93	97	1
4	67	55	0
5	99	95	0
6	73	74	1
7	33	20	1
8	91	91	1
9	80	74	0
10	86	80	1

Pertanto, per ciascuna campagna pubblicitaria si conosce anche se i prodotti sono stati venduti con la nuova confezione o meno.

La variabile qualitativa "modifiche sulla confezione" assume solo due valori, che hanno il seguente significato:

$X_d = 0$ se le confezioni dei prodotti non hanno subito modifiche

$X_d = 1$ se le confezioni dei prodotti hanno subito modifiche.

L'espressione del modello di regressione in cui il volume delle vendite dei prodotti dipende dalla spesa in pubblicità e dalla politica di diversificazione della confezione è il seguente:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{di} + \epsilon_i$$

dove

Y_i = valore accertato della i-esima campagna pubblicitaria

β_0 = intercetta

β_1 = l'effetto sul volume delle vendite dovuto al volume di spesa in pubblicità fermo restando l'effetto della presenza del cambiamento della confezione

β_2 = l'effetto addizionale della presenza del cambiamento della confezione sul volume delle vendite fermo restando l'effetto del volume di spesa in pubblicità

i = errore corrispondente alla i-esima campagna pubblicitaria.

Se $x_{di} = 0$, il modello diventa:

$$E(Y_i | = X_{1i}, X_{2i} = 1) = \beta_0 + \beta_1 X_i$$

invece, se $x_{di} = 1$:

$$E(Y_i | = X_i, X_{di} = 1) = (\beta_0 + \beta_2) + \beta_1 X_i$$

La presenza della modifica nella confezione ($x_{di} = 1$) contribuisce al valore dell'intercetta del modello.

Si riportano qui di seguito i risultati ottenuti dalla stima del modello, avendo utilizzato la funzione di regressione del tool di Excel.

Statistica della regressione	
R multiplo	0,98
R al quadrato	0,96
R al quadrato corretto	0,95
Errore standard	4,28
Osservazioni	10

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	2	3006,15	1503,07	82,04	0,00
Residuo	7	128,25	18,32		
Totale	9	3134,40			

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*
Intercetta	27,35	4,68	5,84	0,001
Spese (in migliaia di euro)	0,77	0,06	12,64	0,000
Modifiche alla confezione	-7,90	2,77	-2,85	0,025

L'espressione del modello di regressione stimato è $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \epsilon_i$.

Nel caso dell'esempio di cui sopra:

$$\hat{y}_i = 27{,}35 + 0{,}77 x_i - 7{,}90 x_{di}$$

Nel caso in cui i prodotti non abbiano subito modifiche nella confezione, quindi con $x_{di} = 0$, il modello stimato è:

$$\hat{y}_i = 27{,}35 + 0{,}77 x_i$$

Invece in presenza di modifiche alla confezione ($x_{di} = 1$), se $x_2 = 1$, il modello stimato diventa:

$$\hat{y}_i = 27{,}35 + 0{,}77 x_{1i} - 7{,}90$$

ovvero il valore 7,90 si sottrae a 27,35 con il seguente risultato:

$$\hat{y}_i = 19{,}45 + 0{,}77 x_i$$

L'interpretazione dei risultati è la seguente:

a) mantenendo costante l'effetto dovuto alla presenza o meno di modifiche nella confezione, ci si aspetta che il valore del volume delle vendite aumenti di 770 in corrispondenza di un aumento di 1000 Euro di spesa pubblicitaria;

b) mantenendo costante la spesa in pubblicità, prevediamo che la presenza di modifiche alla confezione dei prodotti diminuisca di 7900 euro il valore del volume delle vendite dei prodotti.

Osservando ancora i risultati nella tabella sopra riportata, vediamo che il valore della statistica *t* riferita alla dipendenza del volume delle vendite dalla pubblicità è 12,64, con un *p-value* (valore di significatività) pari all'incirca a 0,025; il valore della statistica *t* che si riferisce invece alla presenza o meno di cambiamenti nelle confezioni ha -2,85 e il corrispondente *p-value* è all'incirca 0,000. Pertanto, per livelli di significatività pari a 0,1, oppure 0,05, oppure 0,01, entrambe le variabili presentano un contributo significativo al modello di regressione. Inoltre, dal valore di R^2 vediamo che il 95% della variabilità totale nel valore delle vendite è spiegato dalla variabilità nelle spese pubblicitarie e nella presenza o meno di cambiamenti nella confezione dei prodotti.

Applicazione del modello di regressione in presenza di possibili interazioni tra le variabili indipendenti

Per completare l'analisi di regressione, è necessario verificare che il legame di dipendenza del volume delle vendite rispetto alla spesa in campagne pubblicitarie non dipenda dalla presenza o meno di cambiamenti nelle confezioni dei prodotti. A tal fine, introduciamo nel modello una nuova variabile indipendente definita come prodotto delle prime due variabili esplicative, x^*x_d, e verifichiamo se questa nuova variabile influenza significativamente il volume delle vendite. Questa nuova variabile è detta variabile di interazione.

Il nuovo modello che include la nuova variabile è il seguente:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{di} + \beta_3 X_i X_{di} + \epsilon_i$$

Quando $X_{di} = 0$ il modello diventa

$$E(Y_i | = X_i, X_{di} = 0) = \beta_0 + \beta_1 X_{di}$$

Se $X_{di} = 1$ il modello diventa

$$E(Y_i | = X_i, X_{di} = 1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i$$

La presenza della modifica nella confezione ($x_{di} = 1$) incide sia sul valore dell'intercetta sia sul parametro associato alla variabile spesa in campagne pubblicitarie.

I risultati della stima del nuovo modello sono riportati nelle tabella qui di seguito.

Statistica della regressione	
R multiplo	0,98
R al quadrato	0,97
R al quadrato corretto	0,95
Errore standard	4,15
Osservazioni	10

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	3	3030,84	1010,28	58,53	0,00
Residuo	6	103,56	17,26		
Totale	9	3134,4			

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*
Intercetta	35,85	8,43	4,25	0,01
Spese (in migliaia di euro)	0,64	0,12	5,40	0,00
Modifiche alla confezione	-19,26	9,88	-1,95	0,10
Variabile di interazione	0,16	0,14	1,20	0,28

Dai risultati ottenuti emerge che il valore della statistica *t* corrispondente all'interazione tra la spese in campagna pubblicitarie e le modifiche alla confezione è 1,20 con un valore del *p-value* = 0,28 che è maggiore di 0,05 (anche di 0,01 e di 0,1), perciò non rifiutiamo l'ipotesi nulla secondo cui il coefficiente della variabile

di interazione sia uguale a zero. Si conclude che l'interazione tra le due variabili non offre un contributo significativo al modello una volta che la spese in campagna pubblicitarie e modifiche alla confezione siano già state incluse come variabili esplicative.

6.3.4 Problemi da affrontare in presenza di correlazione tra le variabili indipendenti

L'esistenza di una correlazione fra le variabili indipendenti di un modello di regressione rappresenta una violazione alle assunzioni classiche del modello di regressione. Questa situazione viene identificata con il termine di multicollinearità.

Questa particolare situazione riduce la capacità previsiva di ogni singola variabile indipendente in modo proporzionale alla forza della sua relazione con le altre variabili indipendenti.

L'effetto della multicollinearità interessa non soltanto il modello nella sua capacità di spiegare le relazioni esistenti tra le variabili (non è chiara l'influenza di ciascuna variabile indipendente) ma anche la sua stima (gli effetti sono "mescolati" o confusi).

Quali sono le conseguenze della multicollinearità sulla stima del modello di regressione?

- Le stime dei Minimi Quadrati Ordinari dei parametri sono ancora corrette, ma le varianze e gli errori standard delle stime aumentano. Ciò comporta che il valore della statistica test *t* dei coefficienti stimati diminuisce (essendoci al denominatore valori degli errori standard maggiori), e la probabilità che l'ipotesi di non significatività dei coefficienti non venga rifiutata è più elevata.
- Eventuali cambiamenti nella specificazione del modello, mediante l'aggiunta o l'eliminazione di una o più variabili, modificano anche sensibilmente le stime dei parametri che sono molto sensibili a cambi di specificazione (aggiunta o eliminazione di una variabile).
- Le stime possono non essere significative anche in presenza di un coefficiente di Determinazione R^2 elevato.

La multicollinearità si distingue in due tipologie: la quasi multicollinearità, quando esiste una correlazione tra due o più regressori; tuttavia, la correlazione tra le variabili indipendenti è una condizione sufficiente ma non necessaria alla multicollinearità.

Il secondo tipo di multicollinearità si definisce perfetta quando un regressore è la combinazione lineare di un altro regressore (o più regressori). In tal caso l'effetto del regressore è già compreso nell'altro che lo esprime (o negli altri che lo esprimono), e il Metodo dei Minimi Quadrati non riesce a stimare i parametri perché la matrice dei dati è singolare.

Occorre pertanto verificare l'eventuale presenza di una multicollinerità e valutarla, mediante l'impiego di alcune misure.

a) L'indice di Determinazione lineare R^2_{m0} del modello di regressione in cui X_m è la m-esima variabile che dipende dagli altri $k-1$ regressori.

(R^2_{m0} è il quadrato del coefficiente che misura la correlazione fra la m-esima variabile esplicativa e tutte le altre $m-1$ variabili).

Se $R^2_{m0} > 0,9,$ allora siamo in presenza di multicollinearità.

b) L'Indice di Tolleranza: $Tolerance = 1 - R^2_{m0}$.

c) Il Fattore di Accrescimento della Varianza: $VIF = \dfrac{1}{1 - R^2_{m0}}$.

Quando $VIF > 5$ siamo in presenza di un'alta multicollinearità.

Tanto più elevato è il grado della multicollinearità, tanto maggiori sono i suoi effetti.

Esistono alcuni modi per individuare il caso di multicollinearità nei dati.

1. Si può innanzitutto ragionare sulla matrice di correlazione delle variabili. Se esistono dei valori della correlazione prossimi a $\pm 0,9,$ la multicollinearità è sospetta. Tuttavia, questo criterio si limita a individuare il problema solo per coppie di variabili.

2. Si possono stimare regressioni tra ogni covariata e le altre covariate. Se una di queste regressioni presenta un coefficiente di determinazione R^2 prossimo a 1, allora la covariata considerata come variabili dipendente sarà casua di multicollinearità nella regressione originale.

3. Si può utilizzare il metodo basato sul calcolo degli autovalori della matrice dei dati. Il $\det(X'X)$ è uguale al prodotto degli autovalori (λ_k). Se uno o più autovalori sono prossimi a zero, allora il valore del determinante sarà prossimo a zero. Si calcola la radice quadrata del rapporto tra gli autovalori massimo e minimo $\sqrt{\lambda_{\max}/\lambda_{\min}}$. Se la variabili sono tra loro indipendenti (quindi le colonne della matrice dei dati sono tra loro ortogonali), il rapporto tra i due autovalori è uguale a uno. Tale rapporto aumenta all'aumentare della collinearità tra le variabili. La multicollinearità viene individuata quando il rapporto è maggiore di 20.

4. Quando le statistiche test t e il test F danno risultati contraddittori, per esempio non vengono rifiutate le ipotesi sui singoli parametri mentre il test F conduce a un rifiuto dell'ipotesi congiunta, allora è possibile (è una condizione non necessaria) che sia presente il problema di multicollinearità tra due o più variabili.

Il problema della multicollinearità può essere risolto eliminando dal modello il regressore (o i regressori) che genera la multicollinearità, oppure apportando una modifica al regressore per esempio sostituendolo con la funzione lineare in cui compare l'altro regressore, o infine aumentando la dimensione campionaria in modo tale da ridurre l'errore standard delle stime.

Esempio 6.4

Continuando con l'analisi dei dati presentati nell'esempio precedente, il manager dell'azienda alimentare, dopo aver valutato le relazioni tra le singole coppie di variabili, stima un modello di regressione multipla del Totale altri costi rispetto a Produzioni annue in centinaia di tonnellate e Totali costi personale. Il grafico delle tre variabili, qui di seguito riportato, mostra le unità nel piano tridimensionale e lascia intravedere la presenza di un piano stimato dalla funzione con il metodo dei Minimi Quadrati Ordinari.

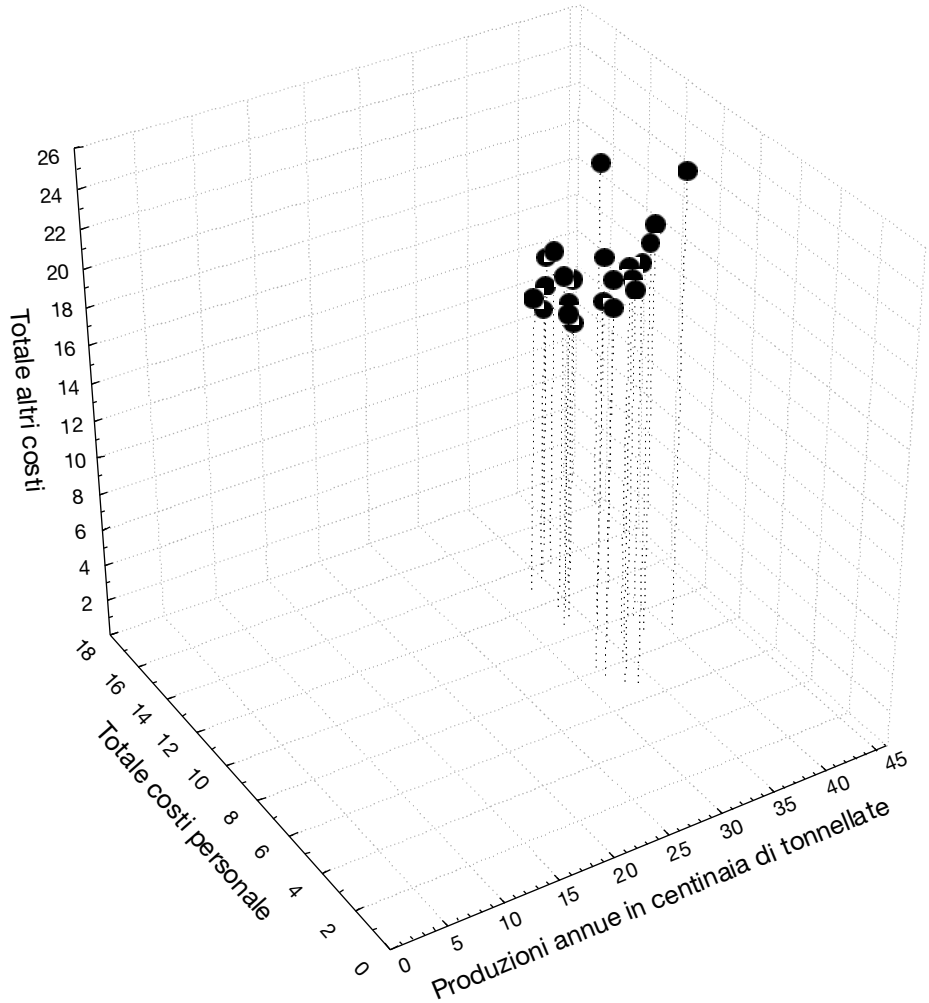

Figura 6.23 Diagramma a punti della matrice dei dati.

Nella Tabella 6.4 si riportano i risultati della stima del modello, mentre la Figura 6.24 presenta il piano del modello di regressione stimato.

Tabella 6.4 Stima e valori della significatività dei parametri.

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Intervallo di confidenza al 95%*	
Intercetta	10,21	2,30	4,43	0,000284641	5,39	15,03
Produzioni annue in centinaia di tonnellate	0,55	0,07	7,55	3,94·10–7	0,40	0,71
Totale costi personale	–1,03	0,09	12,06	2,381·10–10	–1,21	–0,85

Il totale costi personale risulta ancora negativo e significativo, e sorprendentemente anche le produzioni annue sono diventate significative. Questo sta a significare che il modello è capace di "catturare" una relazione grazie alla compresenza di un'altra covariata[8] internamente a una relazione lineare.

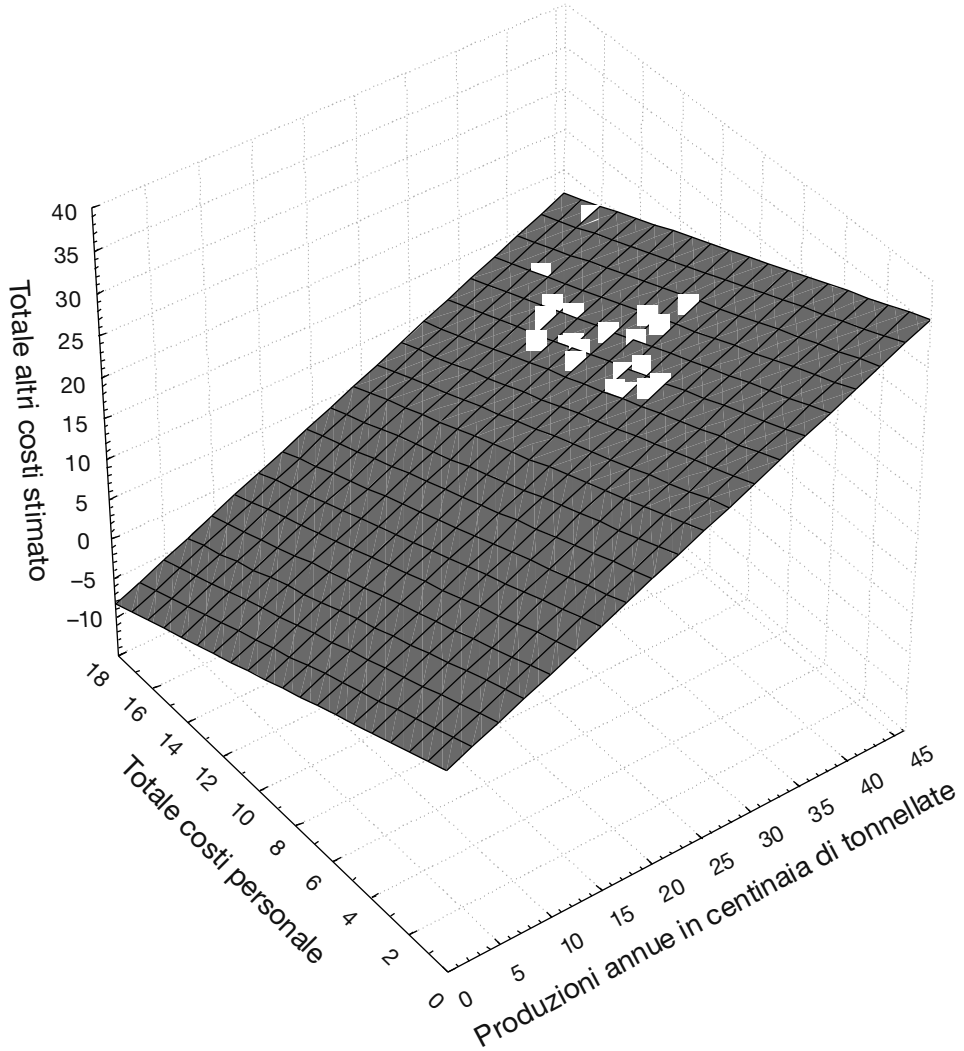

Figura 6.24 Piano del modello di regressione stimato.

[8] Ricordiamo che sebbene il coefficiente b_2 rappresenti un effetto puro, cioè al netto dell'effetto dell'altro coefficiente b_1, esso è definito $b_2 = \frac{Cov(x_2,y)Var(x_1) - Cov(x_1,y)Cov(x_1,x_2)}{Var(x_1)Var(x_2) - [Cov(x_1,x_2)]^2}$.

Le produzioni annue hanno un effetto contenuto sul Totale degli altri costi, attribuibile probabilmente alle voci Ammortamenti, Interessi e altre spese. Quando il costo del personale diminuisce, gli altri costi aumentano a fronte di un maggiore investimento tecnologico nel processo produttivo (aumentano gli ammortamenti, gli interessi passivi e altre spese varie).

Nel complesso il modello presenta un ottimo grado di adattamento ai dati campionari, presentando un valore di $\bar{R}^2$ (corretto) = 0,87. Anche il test F sui parametri restituisce un buon risultato, presentando un'elevata significatività, come riportato nella Tabella 6.5. ❑

Tabella 6.5 Analisi della varianza.

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	2	355926	177963	33	$1{,}83 \cdot 10^{-8}$
Residuo	31	165011	5323		
Totale	33	520937			

6.4 Caso di studio

CASO 6.1 **La dipendenza del fatturato dal prezzo in un'azienda alimentare**

Un aspetto importante nell'attività di marketing, riguarda la misura dell'efficacia di una politica di prezzi e, soprattutto, della promozione pubblicitaria. A tale scopo spesso si studia la dipendenza del fatturato dalla spesa in pubblicità e rispetto a una serie di fattori come per esempio il prezzo del prodotto e i prezzi dei prodotti concorrenti.

L'azienda di prodotti alimentari multinazionale intende introdurre sul mercato con un prezzo competitivo, un nuovo prodotto surgelato adatto per la ristorazione, e ritiene che vi possa essere un buon successo sul mercato. Prima di immettere definitivamente il surgelato sul mercato nazionale e in uno secondo momento anche all'estero, svolge un'analisi preventiva per decidere in maniera efficiente la politica dei prezzi e l'investimento in pubblicità. Un campione di 34 esercizi di ristorazione viene da essa selezionato mediante una ricerca di mercato. I ristoranti hanno all'incirca lo stesso volume di fatturato mensile. La divisione di marketing dell'azienda intende stabilire se esiste innanzitutto una relazione tra le spese di promozione e le vendite del prodotto. I dati dell'esempio sono riportati nella seguente Tabella 6.6, espressi in migliaia di euro.

La variabile dipendente y (detta anche variabile risposta) è il fatturato (per esempio mensile), cioè il volume delle vendite; la variabile indipendente x (detta anche variabile esplicativa o predittore di y) è la spesa pubblicitaria (per esempio mensile). La domanda che il manager si pone è la seguente: sussiste una relazione (elevata) fra le due variabili? In altre parole, conosciuta la spesa pubblicitaria si

può pensare di poter ricavare, mediante un'opportuna funzione matematica che lega y a x, l'esatto ammontare del fatturato che si otterrà?

Tabella 6.6 Dati sui prezzi del prodotto, sulla spesa di promozione e sul volume delle vendite.

Unità	*Prezzi del prodotto*	*Spesa per la promozione*	*Volume delle vendite*
1	139	64	67,5
2	142	11	76,1
3	155	89	82
4	146	38	109,6
5	151	9	160,2
6	169	71	188,2
7	156	75	191,6
8	179	83	208,8
9	144	31	211,4
10	183	152	215,9
11	190	138	229,5
12	178	98	261,8
13	199	185	273
14	167	136	292,7
15	186	60	305,6
16	247	140	322,4
17	201	104	335,4
18	191	148	350,7
19	196	109	351,9
20	197	121	353,2
21	180	130	363,6
22	200	94	374,6
23	251	176	375,4
24	212	178	382,5
25	238	168	384,2
26	243	62	401,1
27	256	87	411,3
28	247	163	414,1
29	173	114	422,6
30	249	188	442,1
31	245	157	463
32	253	150	500
33	255	120	501,5
34	189	191	512

La presenza di una relazione perfetta non sarà possibile per alcuni motivi, primo fra tutti per il fatto che il fatturato può dipendere anche da altri elementi oltre che dalla spesa pubblicitaria (per esempio il prezzo, l'andamento generale dell'economia ecc.). Tuttavia, anche se altre variabili vengono incluse nel modello, è poco probabile che si riesca a determinare esattamente il fatturato. Ci sarà sempre un andamento o variabilità non spiegabile dovuta a fenomeni di disturbo che non sono osservabili (chiamati errori accidentali o casuali).

Indichiamo con y il fatturato (in migliaia di euro) e con x le spese per la promozione pubblicitaria (in migliaia di euro). La domanda è, a questo punto, la seguente: come utilizzare al meglio l'informazione campionaria per stimare β_0 e β_1?

È senz'altro utile costruire un diagramma a punti dei dati in modo da verificare, in via descrittiva, l'opportunità di scegliere un modello lineare del primo ordine (Figura 6.25).

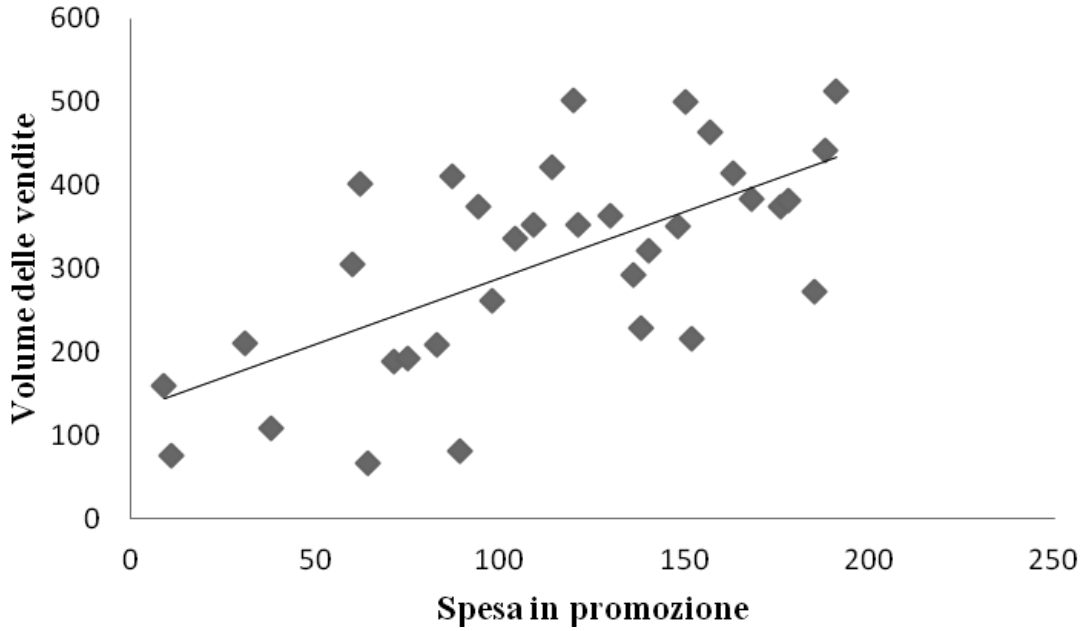

Figura 6.25 Diagramma a punti del volume delle vendite e retta stimata con il metodo dei Minimi Quadrati Ordinari.

Dall'osservazione del grafico sembra che la scelta del modello lineare sia appropriata. L'azienda stima il modello e ottiene i risultati riportato nella Tabella 6.7.

Tabella 6.7 Stime dei coefficienti, significatività e intervallo di confidenza al 95%.

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Intervallo di confidenza al 95%*	
Intercetta	129,57	40,80	3,18	0,003	46,46	212,69
Spesa promozione	1,60	0,33	4,84	$0{,}03 \cdot 10^{-3}$	0,92	2,27

È necessario ricordare che i valori ottenuti per i parametri sono stime ricavate da un'osservazione parziale (campione). Per pervenire a risultati di validità più generale occorrerà applicare l'analisi inferenziale. A tale scopo è necessario specificare meglio la natura della componente di disturbo del modello. Vale la pena soffermarci un momento sui valori stimati.

È importante interpretare il significato dell'intercetta e del coefficiente angolare della retta relativamente ai dati utilizzati per la stima del modello. In questo esempio, il valore $b_0 = 129{,}57$ (cioè 129570 euro) implica che il fatturato ammonta in media a 129570 euro quando la spesa in pubblicità è nulla. Quindi, mediamente l'azienda avrebbe un fatturato anche se non investisse in pubblicità. I valori stimati dei parametri devono essere interpretati soltanto all'interno del campo di variazione dei valori della variabile indipendente che, in questo caso, va da 9000 euro ($x = 9$) a 191000 euro ($x = 191$).

La pendenza della retta è uguale a 1,60, e sta a significare che l'aumento di 1 unità della variabile indipendente determina, in media, un aumento di 1,60 unità della variabile risposta. In altre parole: per ogni aumento di 1000 euro nella spesa pubblicitaria (si ricorda che x è espressa in migliaia di euro) si determina, in media, un aumento di fatturato pari a 1600 euro, all'interno del campo di variazione dei valori della variabile indipendente. Poiché non sono disponibili dati per spese pubblicitarie inferiori a 9000 euro o superiori a 191000 euro, il responsabile del reparto di marketing non ha modo di controllare se il modello rimane valido anche per valori esterni a detto intervallo.

Si può dare un giudizio descrittivo sull'adattamento del modello, mediante l'indice di determinazione R^2. Questo indicatore è capace di riassumere l'adattamento globale e la capacità esplicativa complessiva del modello in rapporto ai dati campionari. Nella Tabella 6.8 si riportano i calcoli delle due devianze.

$$R^2 = \frac{\text{Devianza di regressione}}{\text{Devianza totale}} = 1 - \frac{\text{Devianza residua}}{\text{Devianza totale}}$$

$$= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{300834{,}72}{520936{,}77} = 1 - 0{,}57 = 0{,}42$$

Un valore basso pari a 0,42 come nell'esempio, esprime uno scarso adattamento della retta ai dati. In effetti, osservando la retta nel grafico della Figura 6.25, i dati hanno una disposizione sparsa intorno alla retta: ciò significa che esiste un legame di dipendenza debole del volume delle vendite rispetto alla spesa in pubblicità.

Anche l'indice di correlazione tra spesa in pubblicità e volume delle vendite presenta, coerentemente, un valore piuttosto contenuto:

$$\rho_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \cdot \sigma_y} = 0{,}65$$

Tabella 6.8 Spesa promozione, volume delle vendite e calcolo della devianza totale e di regressione.

Unità	*Spesa promozione*	*Volume vendite*	$(y_i - \bar{y})^2$	$(y_i - \hat{y})^2$
1	64	67,5	58742,08	26975,04
2	11	76,1	54647,31	5045,93
3	89	82	51923,66	35966,88
4	38	109,6	40107,13	6502,15
5	9	160,2	22400,40	264,32
6	71	188,2	14803,02	2993,75
7	75	191,6	13987,24	3329,36
8	83	208,8	10214,67	2837,83
9	31	211,4	9695,88	1045,78
10	152	215,9	8829,92	24435,59
11	138	229,5	6458,96	14488,96
12	98	261,8	2310,50	586,44
13	185	273	1359,22	23073,05
14	136	292,7	294,73	2913,56
15	60	305,6	18,21	6439,19
16	140	322,4	157,06	940,21
17	104	335,4	651,90	1584,48
18	148	350,7	1667,28	229,02
19	109	351,9	1766,72	2335,19
20	121	353,2	1877,69	928,28
21	130	363,6	2887,17	702,28
22	94	374,6	4190,28	9019,10
23	176	375,4	4294,49	1234,19
24	178	382,5	5275,46	974,92
25	168	384,2	5525,30	183,88
26	62	401,1	8323,34	29774,15
27	87	411,3	10288,52	20404,22
28	163	414,1	10864,38	591,53
29	114	422,6	12708,58	12330,35
30	188	442,1	17485,40	154,08
31	157	463	23449,52	6855,74
32	150	500	36150,31	17154,14
33	120	501,5	36722,96	32531,19
34	191	512	40857,49	6009,94
Totale			520936,77	300834,72

Tuttavia, per capire se il risultato ottenuto con questo campione di dati abbia portata più generale, è necessario condurre un'analisi inferenziale.

Il manager ottiene dal test statistico (Tabella 6.7) che le stime sono molto significative; si noti inoltre che l'intervallo di confidenza al 95% non contiene il valore 0.

Anche in presenza di uno scarso livello di adattamento del modello ai dati, possiamo ottenere stime altamente significative.

Qui di seguito viene riportato il grafico q-qnormal plot per i dati dell'azienda alimentare.

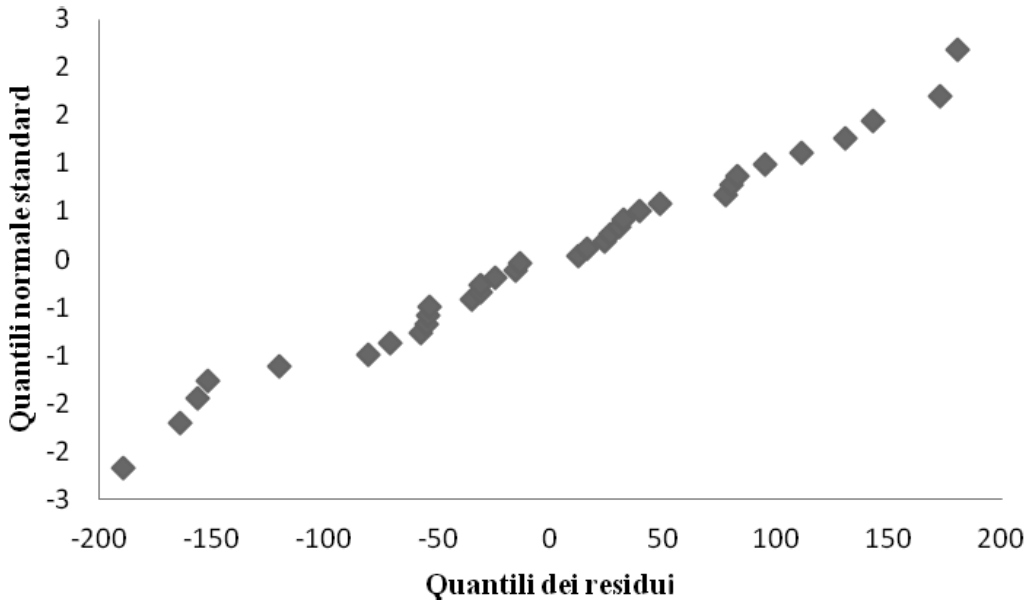

Figura 6.26 Q-qnormal plot dei residui.

Dall'analisi grafica il manager dell'azienda può concludere che l'ipotesi di normalità è rispettata quasi completamente. Con i risultati alla mano la nostra azienda multinazionale ottiene il grafico dei residui riportato nella Figura 6.27.

Il manager può concludere che l'ipotesi di omoschedasticità è rispettata.

Supponiamo adesso che la divisione marketing dell'azienda di prodotti alimentari intenda stabilire l'effetto che il prezzo e le promozioni in questi esercizi possono esercitare sulle vendite.

In questo caso occorre utilizzare un modello di regressione che relazioni il volume delle vendite y, con il prezzo del singolo prodotto (x_1) e la spesa mensile per le attività promozionali (x_2). I dati sono quelli della Tabella 6.6. Come primo passo dell'analisi, il manager produce i grafici a dispersione e la matrice di correlazione (Tabella 6.9).

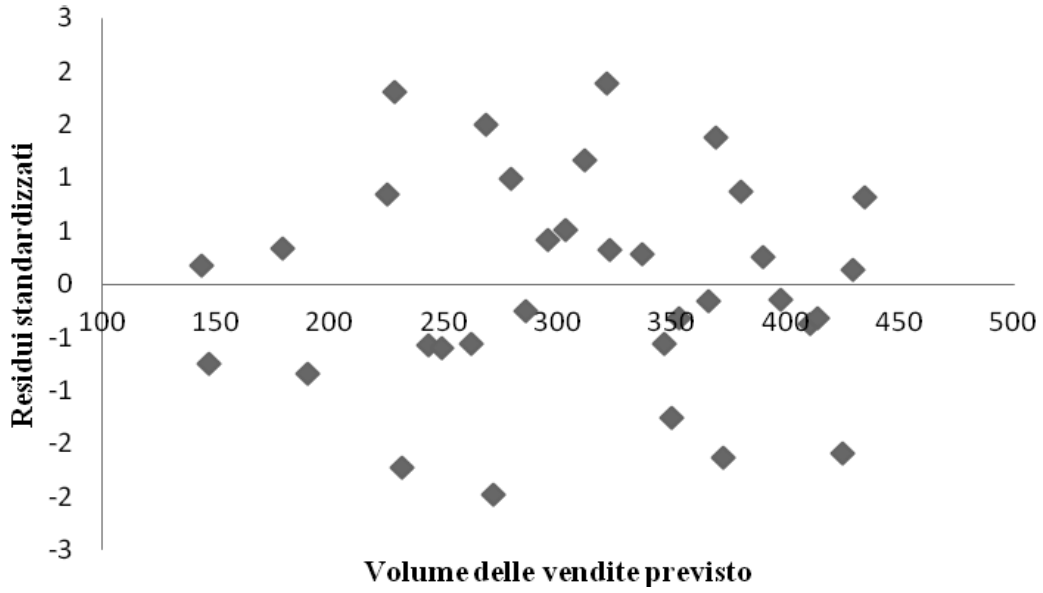

Figura 6.27 Grafico dei residui standardizzati.

Tabella 6.9 Matrice di correlazione dei dati.

	Prezzi del prodotto	*Spesa promozione*	*Volume vendite*
Prezzi del prodotto	1		
Spesa promozione	0,61 (*p-value* = 0,000)	1	
Volume vendite	0,80 (*p-value* = 0,000)	0,65 (*p-value* = 0,000)	1

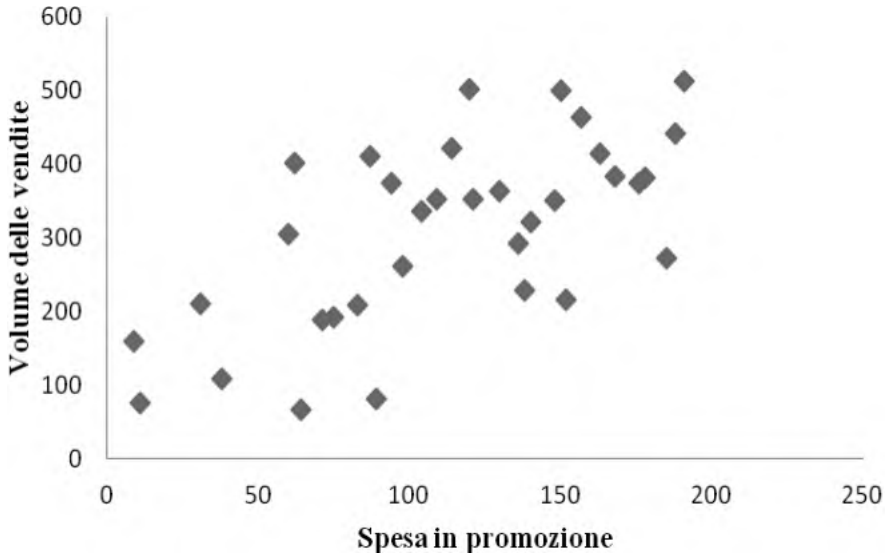

Figura 6.28 Diagramma a dispersione del volume delle vendite, rispetto alla spesa in promozione.

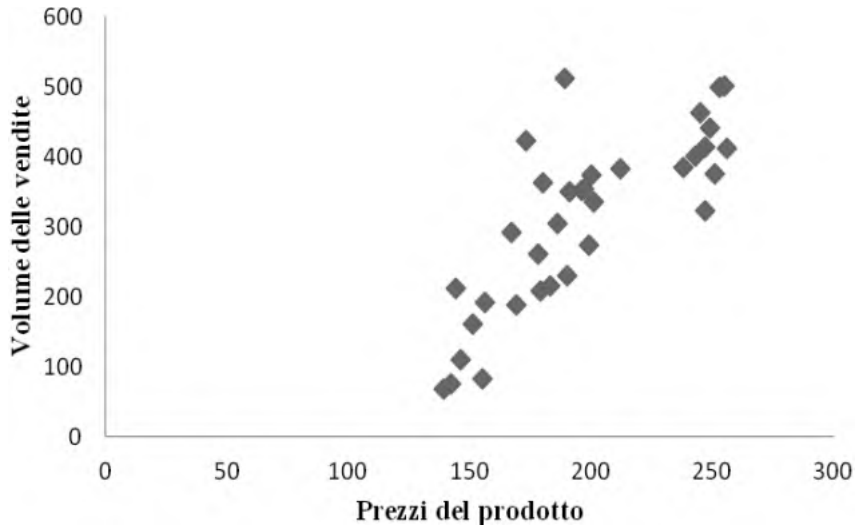

Figura 6.29 Diagramma a dispersione del volume delle vendite, rispetto ai prezzi del prodotto.

Le correlazioni sono tutte positive (già si conosceva quella tra le vendite e la spesa in promozione) e in particolare vi è un forte legame tra il volume delle vendite e i prezzi del prodotto, come si evince dal valore della correlazione e dalla posizione dei punti sul grafico (rispetto agli altri due grafici in cui le nubi dei punti sono disposte più orizzontalmente).

Come sappiamo l'indice di correlazione misura la relazione lineare tre due variabili. La costruzione di un grafico a dispersione rispecchia l'eventuale presenza di una relazione lineare di dipendenza di una variabile rispetto a un'altra. Pertanto occorre tenere presente che la dipendente viene convenzionalmente posizionata sull'asse delle ordinate. In presenza di due predittori, lo scambiarsi di uno con l'altro sugli assi genera due figure diverse con due differenti ipotetiche (in realtà sono entrambi due variabili indipendenti) funzioni di regressione diverse.

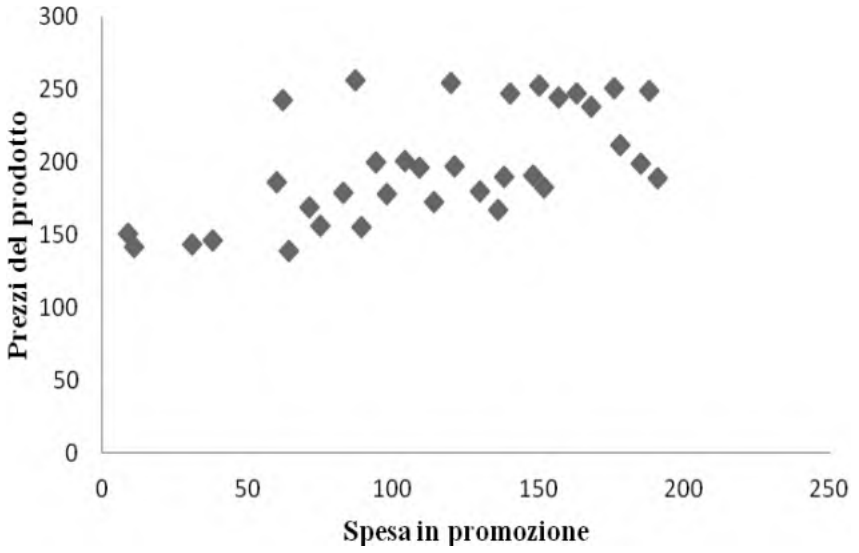

Figura 6.30 Diagramma a dispersione dei prezzi del prodotto rispetto alla spesa in promozione.

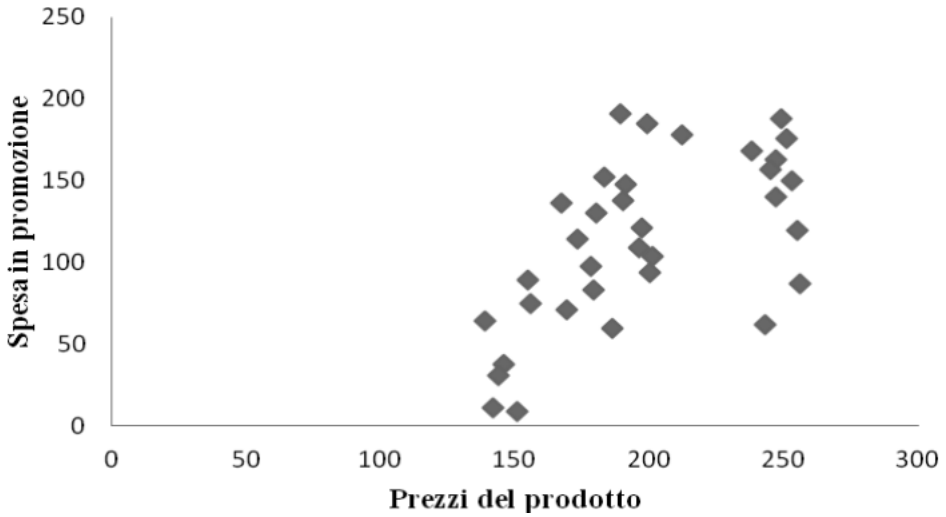

Figura 6.31 Diagramma a dispersione della spesa in promozione rispetto ai prezzi del prodotto.

Le ipotetiche funzioni di regressione semplice stimate relative alle Figure 6.30 e 6.31 sono rispettivamente $y = 0{,}4532x + 146{,}08$ e $y = 0{,}8153x - 47{,}89$. Ricordando la relazione tra il coefficiente di correlazione e il coefficiente di regressione lineare, abbiamo $0{,}61 = 0{,}45 \cdot \frac{50{,}40}{37{,}58} = 0{,}45 \cdot 1{,}34$ e $0{,}61 = 0{,}81 \cdot \frac{37{,}58}{50{,}40} = 0{,}81 \cdot 0{,}75$.

In questa analisi la componente deterministica $f(X)$ è composta dalle variabili Prezzo del prodotto e Spesa di promozione. Il manager produce un grafico a dispersione per vedere i punti nel piano cartesiano.

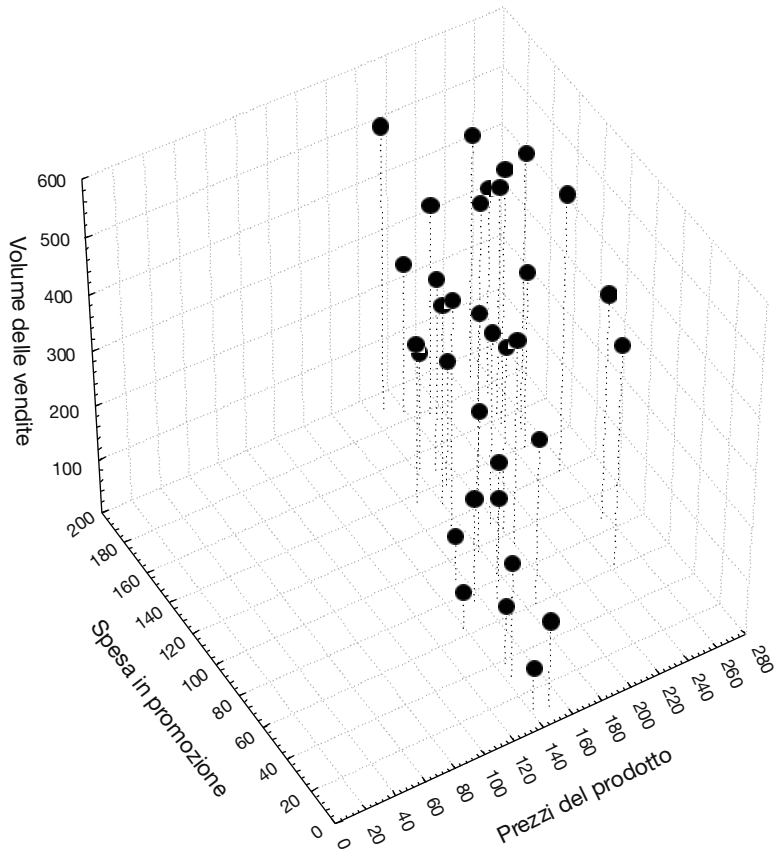

Figura 6.32 Diagramma a punti del volume delle vendite rispetto ai due regressori.

Da un'analisi preliminare grafica e di correlazione, in cui si sono manifestati il legame e il segno delle relazioni, il manager decide di procedere con un'analisi delle regressione multipla.

La divisione marketing dell'azienda multinazionale procede alla stima del modello con incluse le due variabili indipendenti (Tabella 6.10).

Tabella 6.10 Stime dei coefficienti, significatività e intervallo di confidenza al 95%.

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Intervallo di confidenza al 95%*	
Intercetta	–179,84	68,52	–2,62	0,01	–319,59	–40,10
Prezzi del prodotto	2,12	0,42	5,05	$0{,}02 \cdot 10^{-3}$	1,26	2,97
Spesa di promozione	0,64	0,31	2,04	0,05	0,00	1,27

I risultati possono essere rappresentati graficamente generando un piano di regressione.

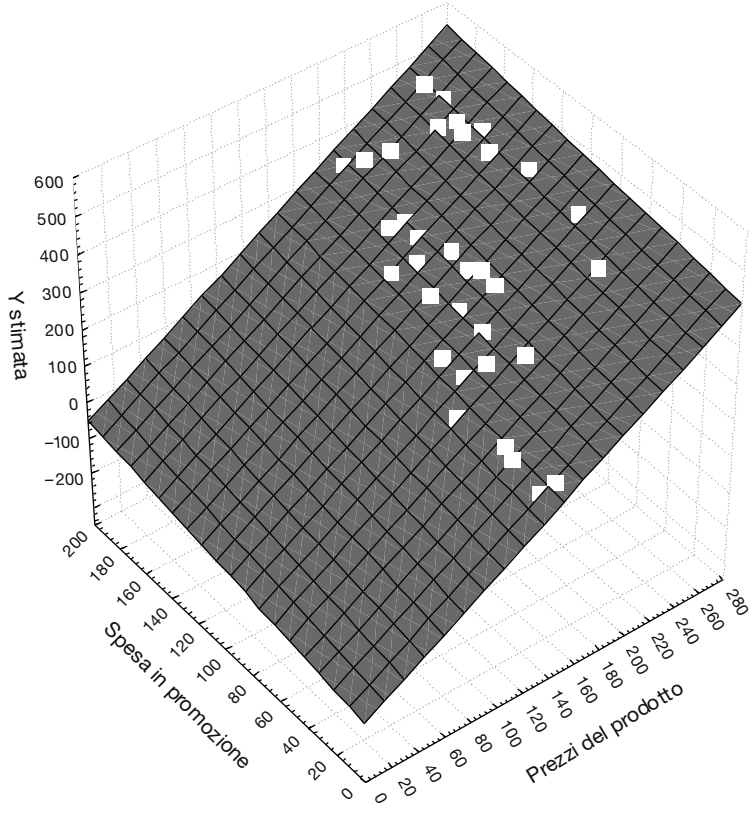

Figura 6.33 Piano del modello di regressione stimato.

Il manager, osservando le stime dei coefficienti, deduce che in assenza di prezzo e di spesa in promozione non c'è alcuna realizzazione delle vendite, anzi il valore negativo dell'intercetta spiega l'esistenza di una spesa. La variabile prezzo incide sul volume delle vendite (2,12) in misura superiore alla spesa in promozione, e in effetti il prezzo è un fattore piuttosto importante nella determinazione del fatturato delle vendite.

Si noti che dalla precedente analisi semplice del volume delle vendite rispetto alla promozione era emerso un valore positivo dell'intercetta, dunque, si può concludere che il valore negativo dell'intercetta di quest'ultima stima è attribuibile alla variabile prezzo. Del resto, anche la stima della funzione semplice delle vendite sul prezzo, $y = 2{,}637x - 210{,}32$ (si veda la Figura 6.29), ha generato un valore negativo dell'intercetta.

La capacità del modello di rappresentare questa relazione tra le variabili, sulla base del campione considerato, non risulta particolarmente elevata, dal momento che $\bar{R}^2$ (corretto) = 0,66; tuttavia, il manager vuole verificare statisticamente la validità del modello nel suo complesso, ricorrendo all'analisi della varianza basata sulla statistica *F*. Il risultato del test è riportato nella Tabella 6.11.

Tabella 6.11 Analisi della varianza.

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	2	355926	177963	33	$1{,}83 \cdot 10^{-8}$
Residuo	31	165011	5323		
Totale	33	520937			

Si rifiuta che tutti i parametri siano uguali a zero, pertanto esiste un effetto significativo dei regressori sul volume delle vendite.

Una volta verificata la validità complessiva del modello, il manager si concentra sulla significatività dei singoli parametri stimati, osservando i valori della significatività riportati nella Tabella 6.10. Dal test sulla significatività dei singoli parametri, per un livello di confidenza al 95% le stime dell'intercetta e della variabile prezzi del prodotto risultano significative, mentre la spesa in promozione che nella stima di un modello di regressione in cui compariva come unico predittore, aveva un *p-value* = $0{,}03 \cdot 10^{-3}$, adesso presenta un valore ridotto a 0,05 corrispondente al limite della significatività.

In generale, dato un modello di regressione lineare con una covariata, l'inclusione di un nuovo predittore modifica i valori delle stime e della significatività della covariata medesima, talvolta anche al punto di perdere di significatività. Il manager, pertanto deve fare le proprie valutazioni sulla decisione di includere o meno uno o più regressori nel modello, valutazioni non solo dettate dalla natura delle variabili e dalle relazioni economiche, ma anche valutazioni statistiche grazie all'ausilio dei test sulla significatività dei parametri.

Esercizi

6.1 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 manager 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 che segue si riportano i dati raccolti negli ultimi 24 mesi in relazione ai costi di distribuzione (in migliaia), alle vendite (in migliaia) e al numero di ordini ricevuti. Sulla base dei dati raccolti:

a) fornire un'interpretazione delle inclinazioni della variabile dipendente rispetto a ciascuna delle variabili esplicative;
b) verificare se sono rispettate le ipotesi normalità e omoschedasticità del modello di regressione;
c) fornire una previsione dei costi di distribuzione per un ammontare delle vendite pari a 400.000 euro e degli ordini pari a 4500.

Mese	Costi distribuzione	Vendite	Ordini
1	65,45	386	3015
2	84,16	446	3806
3	98,08	512	5309
4	76,19	401	4262
5	85,31	457	5000
6	80,94	458	4097
7	64,96	301	3213
8	83,27	484	4809
9	94,53	517	5237
10	86,89	503	4732
11	83,34	535	4413
12	66,58	353	2921
13	75,48	372	3977
14	79,8	328	4428
15	71,49	408	3806
16	91,88	501	4582
17	106,94	527	5582
18	72,24	444	2890
19	103	623	5079
20	105,74	596	5735
21	81,83	463	4269
22	66,21	389	3708
23	101,68	547	5387
24	79,3	415	4161

6.2 La seguente tabella contiene i dati relativi al processo di confezionamento di confetti per la gola, e precisamente le variabili: numero di pezzi prodotti (confezioni da 20 confetti), PROD; numero di ore di impiego del macchinario, ORE1; numero di ore di forza lavoro, ORE2.

PROD	ORE1	ORE2
3917	63	100
3026	46	86
3457	76	92
4157	94	108
4344	93	104
5000	98	105
3390	41	112
3564	64	95
4040	92	108
2525	48	79
2546	55	66
2924	43	117
2680	42	101
3015	50	85
3223	53	102
1998	41	64
3356	57	116
3753	61	92
3894	73	90
3908	61	120

Svolgere la seguente analisi:

a) costruire i diagrammi a dispersione per le coppie di variabili formate con il numero di pezzi prodotti, il numero di ore di impiego del macchinario e il numero di ore di forza lavoro;

b) stimare una funzione di regressione lineare che esprima il contributo dei macchinari e del lavoro umano alla produzione;

c) interpretare il significato dei coefficienti del modello;

d) commentare la significatività del modello;

e) commentare la significatività dei singoli coefficienti stimati;

f) spiegare il significato del coefficiente di determinazione R^2.

6.3 La rivista Motor Trend ha riportato nell'ultimo fascicolo alcuni dati relativi a un'indagine sul consumo di benzina relativi a 30 diversi modelli di macchina, con incluse le relative caratteristiche tecniche di costruzione. Si verifichi la relazione di dipendenza del consumo di benzina da due particolari specifiche quali la coppia massima e la compressione. Utilizzare i dati riportati nella seguente tabella.

Consumo di benzina	Coppia massima	Compressione
18,9	260	8
17	275	8,5
20	185	8,25
18,3	255	8
20,1	170	8,4
11,2	330	8,2
22,1	175	8
21,5	200	8,5
34,7	81	8,2
30,4	83	9
16,5	250	8,5
36,5	83	8,5
21,5	146	8,2
19,7	195	8
20,3	109	8,4
17,8	220	8
14,4	360	8,5
14,9	330	8,2
17,8	250	8,5
16,4	255	8,5
23,5	175	8
21,5	290	8,4
31,9	83	9
13,3	366	8
23,9	120	8,4
19,7	255	8,5
13,9	243	8
13,3	243	8
13,8	295	8,25
16,5	255	8,5

6.4 Un'impresa edile intende effettuare 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 estrae un campione di 15 case e i dati relativi al valore (in migliaia di euro), la superficie riscaldata (in migliaia di piedi al quadrato) e l'età delle case (in anni) sono riportati nella tabella qui di seguito:

Case	Valore accertato	Superficie riscaldata	Età
1	84,4	2	3,42
2	77,4	1,71	11,5
3	75,7	1,45	8,33
4	85,9	1,76	0
5	79,1	1,93	7,42
6	70,4	1,2	32
7	75,8	1,55	16
8	85,9	1,93	2
9	78,5	1,59	1,75
10	79,2	1,5	2,75
11	86,7	1,9	0
12	79,3	1,39	0
13	74,5	1,54	12,58
14	83,8	1,89	2,75
15	76,8	1,59	7,17

Svolgere i punti:

a) costruire un diagramma a dispersione per ogni coppia di variabili formata da valore, superficie riscaldata ed età della case;

b) calcolare le matrici di varianza-covarianza e di correlazione;

c) formulare l'espressione del modello di regressione multipla;

d) stimare il modello e fornire un'interpretazione dell'inclinazione del modello;

e) prevedere il valore accertato di una casa con superficie riscaldata pari a 1750 piedi al quadrato e 10 anni di età;

f) condurre un'analisi dei residui e sulla base di essa valutare l'adeguatezza del modello;

g) per un livello di significatività α pari a 0.05 stabilire se vi sia una relazione significativa tra il valore delle case e le due variabili esplicative;

h) con riferimento al punto g) calcolare il valore del *p-value* e interpretarne il significato;

i) interpretare il significato del coefficiente di determinazione;

j) per un livello di significatività α pari a 0.05 valutare il contributo di ciascuna variabile esplicativa al modello di regressione. Sulla base dei risultati esiste un modello migliore che andrebbe usato?

6.5 1. L'equazione della retta (deterministica) è $y = \beta_0 + \beta_1 x$. Se passa per il punto $y = 6$ e $x = 6$, determinare il valore dei parametri.

2. Individuare l'intercetta e la pendenza delle seguenti rette:

   (a) $y = 4 + x$ (b) $y = -4 + x$ (c) $y = 4 + 2x$

   (d) $y = -2x$ (e) $y = x$ (f) $y = 0{,}5 - 0{,}75x$

3. Disegnare due delle rette sopra introdotte.

4. Indicare quali valori possono assumere i parametri di una retta.

5. Secondo quanto definito, il modello probabilistico implica che ogni valore della variabile risposta si trovi esattamente sulla linea delle medie?

Le soluzioni degli esercizi sono disponibili sulla piattaforma MyLab.