ANALISI STATISTICA DELLE VENDITE E METODI PER LA PREVISIONE

Prof. Domenico SUMMO

1. Premessa

Un imprenditore, nell'esplicare la propria attività economica, non fa altro che prevedere quali potranno essere i bisogni e le preferenze dei consumatori in futuro. La previsione, in tale contesto, è semplicemente un'ipotesi riguardante il futuro, fatta in base ad informazioni passate. Conoscere per prevedere è diventato da lungo tempo una proposizione così ovvia da sfiorare la banalità[1].

La previsione è efficace quando è possibile conoscere le modalità di manifestazione di un evento con un anticipo di tempo tale da intraprendere azioni orientate a sfruttare i vantaggi, se l'evento è favorevole, o a limitare i danni in caso contrario. Le azioni attuate in conseguenza delle previsioni possono tendere a modificare in qualche modo l'evento previsto; ad esempio basti pensare alle politiche di marketing (politiche di prodotto, di prezzo, pubblicitarie e di distribuzione) rivolte ad attenuare gli effetti della stagionalità della domanda.

La previsione delle vendite ricopre un ruolo centrale nella gestione di una azienda e costituisce senz'altro il problema più complesso e delicato. Essa compare come un input importante in tutte le fasi del processo decisionale di pianificazione aziendale.

La previsione, per poter essere veramente efficace, non può essere un esercizio discontinuo ma deve esplicarsi continuamente nel tempo; essa si forma e si modifica momento per momento e richiede la conoscenza continua di atti e di fatti che formano il tessuto economico dell'azienda[2].

---

[1] Giusti F., Vitali O., *Statistica economica*, op. cit..

[2] Caprara G., *Previsioni e programmazione in un impresa industriale*, Giuffrè, Milano, 1965.

2. L'organizzazione dei dati

Per una più puntuale analisi del fenomeno è opportuno organizzare il data set secondo alcune variabili descrittive. Da ogni atto di vendita si possono ricavare una molteplicità di dati: la tipologia del bene ceduto, la quantità ceduta, il valore del bene, la modalità di distribuzione, le condizioni di vendita, il luogo di destinazione, ecc. In tale ottica, le analisi condotte sulle previsioni di vendita dei beni e servizi di un'azienda consentono di ricavare tutte quelle informazioni necessarie all'organizzazione delle risorse disponibili e delle diverse funzioni aziendali; si tratta, allora, di analisi che per l'azienda ricoprono un ruolo centrale in quanto interessano tutti i diversi settori dell'azienda: dall'area degli acquisti, alla produzione, al personale, alla commercializzazione e distribuzione e, non per ultimo, all'area economico-finanziaria.

Le registrazioni delle vendite si possono organizzare secondo una matrice in cui le righe sono riferite alle singole vendite, mentre le colonne riportano tutte o alcune delle informazioni innanzi dette. L'analisi statistica delle vendite riguarderà, in tal caso, la classificazione ed elaborazione dei dati riguardanti i caratteri ricavati da tale matrice.

I dati opportunamente classificati possono rappresentarsi in tabelle doppie o multiple a seconda dei caratteri esaminati. Per ogni carattere considerato possono essere costruiti diversi rapporti statistici, a condizione che ci sia un nesso logico tra i caratteri a confronto. Ad esempio, indicando con i = 1, 2,..., s le tipologie di beni e servizi venduti e con j = 1, 2,..., n le aree territoriali, i dati possono essere organizzati in una tabella a doppia entrata del tipo seguente.

Vendite	**Area territoriale**						***Totale***
	1	2	...	j	...		
1	Y_{11}	Y_{12}	...	Y_{1j}	...	Y_{1n}	$Y_{1.}$
2	Y_{21}	Y_{22}	...	Y_{2j}	...	Y_{2n}	$Y_{2.}$
...							
i	Y_{i1}	Y_{i2}	...	Y_{ij}	...	Y_{in}	$Y_{i.}$
...							
s	Y_{s1}	Y_{s2}	...	Y_{sj}	...	Y_{sn}	$Y_{s.}$
Totale	$Y_{.1}$	$Y_{.2}$	...	$Y_{.j}$	...	$Y_{.n}$	$Y_{..}$

Pertanto, prendendo in considerazione il totale delle vendite dell'area territoriale j-esima o del prodotto i-esimo, si possono calcolare per riga e per colonna i seguenti quozienti:

1) rapporti di coesistenza tra le vendite dei singoli prodotti:

(per riga) $$\frac{Y_{ij}}{Y_{i1}}, \frac{Y_{ij}}{Y_{i2}}, \ldots, \frac{Y_{ij}}{Y_{in}}$$

(per colonna) $$\frac{Y_{ij}}{Y_{1j}}, \frac{Y_{ij}}{Y_{2j}}, \ldots, \frac{Y_{ij}}{Y_{sj}}$$

2) rapporti di composizione:

$\frac{Y_{ij}}{Y_{.j}}$ rispetto al totale delle colonne, mentre $\frac{Y_{ij}}{Y_{i.}}$ per il totale delle righe.

Considerando ancora la variabile tempo, si possono calcolare i *saggi di variazione* delle vendite tra il tempo t ed il tempo t-1, sia riferiti ai singoli prodotti che alle aree:

$$R = \frac{Y_{ij}(t)}{Y_{ij}(t-1)}$$

Quando i confronti avvengono tra aree territoriali diverse, è necessario tener conto delle diverse situazioni socio-economiche. Ad esempio per un bene di largo consumo le vendite, a parità di altre condizioni, dipendono dalla consistenza numerica della popolazione presente. Per eseguire delle comparazioni spaziali è quindi necessario depurare il valore complessivo delle vendite dall'incidenza della diversa numerosità delle popolazioni, calcolando prima i rapporti pro-capite.

3. L'orizzonte temporale delle vendite e metodi di previsione.

Le analisi sulle vendite vengono generalmente condotte facendo riferimento ad un determinato intervallo di tempo.

Non esiste una delimitazione temporale in merito alla classificazione delle vendite, tuttavia è importante distinguere le previsioni a breve, a medio e lungo termine; questa distinzione implica differenze che dipendono non soltanto dalla durata dell'intervallo della previsione, ma anche dal tipo di analisi che viene effettuata e dagli obiettivi che ci si prefigge di raggiungere. Sia i metodi di previsione a breve che quelli a medio termine sono normalmente adottati utilizzati a livello aziendale, invece, i metodi di previsione a lungo termine riguardano in prevalenza ambiti di ricerca di carattere macroeconomico.

Le previsioni a breve termine riguardano un periodo di tempo fino ad un anno e consentono di anticipare la dimensione della domanda di beni, delle materie prime, della forza lavoro, del fabbisogno di liquidità.

Le previsioni a medio termine abbracciano un periodo che varia da 2 a 5 anni e riguardano essenzialmente l'area della produzione aziendale e quella finanziaria, per prevedere o controllare la redditività degli investimenti e le diverse strategie e attività di marketing da mettere in pratica.

Le previsioni a lungo termine si riferiscono ad un periodo che varia fra i cinque e i dieci anni, e riguardano soprattutto l'andamento futuro della domanda globale di mercato di un determinato bene o servizio (domanda primaria), al cui interno si pone la previsione del volume delle vendite a livello di azienda (domanda secondaria). Connessa a questa previsione è la programmazione circa l'espansione della capacità produttiva e quindi l'installazione di nuovi impianti. Altri obiettivi connessi alla previsione delle vendite riguardano le fonti di finanziamento necessarie per sostenere i programmi di investimento, i fabbisogni di personale e di materie prime, unitamente alle fonti dove reperire queste ultime. La componente *trend* che caratterizza le serie

storiche dei dati di vendita rappresenta la finalità delle previsioni di lungo termine; tuttavia, non si tratta di far riferimento ad una estrapolazione acritica dei dati passati, ma, piuttosto, di individuare la fase in cui presumibilmente si trova il ciclo di vita del comparto o del prodotto[3]. La previsione a lungo termine, infine, non può, né deve proporsi di formulare una predizione, essa può inquadrare invece un ragionevole numero di alternative sulla base di una approfondita conoscenza del passato e delle tendenze che si colgono nel presente[4].

Da un punto di vista operativo i numerosi metodi di previsione delle vendite si differenziano per presupposti e contenuti metodologici, non si adattano tutti allo stesso modo al tipo di problema che di volta in volta ci si trova a dover risolvere e possono essere implementati per analisi sia a carattere temporale sia territoriale. Fra i diversi metodi di previsione una distinzione sostanziale è quella fra metodi statistici e metodi congetturali; questi ultimi, fondati essenzialmente su elementi come il giudizio personale e l'esperienza, presentano il limite della soggettività. I metodi di previsione a carattere prettamente statistico si basano generalmente su modelli che assumono come valide per il futuro quelle relazioni che si sono già verificate sia a livello territoriale (per aree) che temporale.

Nell'ambito delle diverse metodologie statistiche per la previsione, si distinguono: *metodi endogeni e metodi esogeni*.

I metodi endogeni utilizzano gli stessi dati del fenomeno organizzati su base temporale con scopo di prolungarli nel tempo utilizzando le tecniche statistiche di analisi tipiche delle serie temporali. Dopo aver scomposto la serie storica nelle sue componenti (trend, andamento congiunturale e stagionale), i principali metodi endogeni che possono essere impiegati sono:

- il metodo della media semplice e mobile,
- il metodo dell'estrapolazione del trend,
- il metodo del livellamento esponenziale,
- il metodo di Box e Jenkins.

I metodi esogeni di previsione, invece, considerano la relazione che esiste tra la variabile risposta (le vendite) e le possibili variabili indipendenti o esplicative. Si tratta di specificare il modello che lega queste variabili che rappresentano le determinanti delle vendite e ipotizzare che nel futuro queste cause possano continuare ad influire sul fenomeno oggetto di analisi: nel nostro caso le vendite. In questa classe di metodi rientrano senz'altro i modelli lineari e non lineari di regressione semplice e multipla. Altri metodi esogeni che possono essere impiegati, sono: i sondaggi, le tavole input-output di Leontief ed il calcolo iterativo dei parametri di una interpolante lineare.

[3] Valdani E.-Busacca B., *Previsioni delle vendite e ciclo di vita del prodotto*, ETAS, Torino, 1987.

[4] Marbach G., *Le previsioni di lungo periodo: analisi esplorative*, Franco Angeli Editore, Milano, 1980.

4. Il metodo della media semplice e mobile

Il metodo della media aritmetica semplice è applicabile a quelle serie storiche stazionarie, cioè in assenza di trend e di stagionalità. In tali condizioni, ogni nuovo dato conduce ad una migliore approssimazione della previsione successiva.

Se dunque al tempo s sono disponibili s osservazioni y_1, y_2,..., y_t, ..., y_s relative alle vendite, si può calcolare la media aritmetica semplice, data da:

$$ {}_s\bar{y} = \frac{\sum_{t=1}^{s} y_t}{s}, $$

che rappresenta il valore medio delle vendite alla fine del periodo s.

Dalle considerazioni precedenti risulta dunque che la previsione per il periodo s+1 è data dalla media stessa, cioè:

$$ \hat{y}_{s+1} = {}_s\bar{y} \quad . $$

Quando al tempo s+1 si rende effettivamente disponibile un nuovo dato v_{S+1}, la media può essere aggiornata, aggiungendo al numeratore il nuovo valore ed aumentando il denominatore di una unità:

$$ {}_{s+1}\bar{y} = \frac{\sum_{t=1}^{s+1} y_t}{s+1}, $$

la quale diventa la previsione per il tempo s+2, effettuata utilizzando il valore medio delle vendite alla fine del periodo s+1, e così via.

L'espressione precedente può essere scritta in modo applicativamente più semplice:

$$ {}_{s+1}\bar{y} = \frac{s\,{}_s\bar{y} + y_{s+1}}{s+1}, $$

che consente di calcolare la media al tempo s+1, e dunque la previsione per il tempo s+2, in funzione della media al tempo s e del nuovo dato rilevato al tempo s+1. Pertanto, per effettuare la previsione in un certo periodo, sono sufficienti tre dati: la media calcolata per il periodo precedente, il numero t dei termini che compongono la serie per cui è calcolata tale media, ed il valore dell'ultimo dato osservato.

L'impiego della media semplice, se da un lato è intuitivo, dall'altro presenta alcuni difetti, nel senso che attribuisce un peso sempre uguale a tutte le vendite precedenti; è noto, invece, che le vendite più lontane nel tempo hanno un'importanza minore rispetto alle più recenti. Inoltre il metodo non tiene conto delle componenti sistematiche (trend, stagionalità)[5].

Il metodo delle medie mobili è meno soggetto a tali problemi: esso, infatti, consente di modificare i dati storici della serie man mano che si rende disponibile una

[5] Valdani E.-Busacca B., *Previsioni delle vendite e ciclo di vita del prodotto*, op. cit..

nuova osservazione, escludendo dal calcolo l'osservazione più vecchia ed includendo la più recente, mantenendo costante il numero dei termini da cui scaturisce la media.

Per il calcolo della media mobile, occorre stabilire il numero k di termini da considerare. In caso che k sia dispari, la media mobile si dice *centrata*; nel caso di k pari, invece, la media non è centrata e risulta necessario effettuare una successiva media mobile a due termini. La media mobile iniziale a k termini si calcola come nel caso della previsione a media semplice:

$$_k\overline{y} = \frac{\sum_{t=1}^{k} y_t}{k},$$

ed anche in questo caso tale valore rappresenta la previsione per il periodo successivo: $\hat{v}_{k+1} = {}_k\overline{v}$.

La differenza sta nel fatto che, all'acquisizione del dato relativo al periodo n+1, il calcolo della media relativa a quell'istante, la quale costituisce perciò la previsione per il periodo n+2, diventa:

$$_{k+1}\overline{y} = \frac{\sum_{t=1}^{k+1} y_t}{k+1},$$

e così via.

Rispetto al metodo della media aritmetica semplice, in questo caso vi è lo svantaggio di dovere necessariamente tenere in archivio gli ultimi k dati, anziché solo l'ultima media calcolata, ma presenta il vantaggio di tenere conto, in una certa misura, sia del trend che della componente stagionale.

5. Il Metodo dell'estrapolazione del trend

Il metodo consiste nella valutazione delle tendenze evolutive statisticamente significative dei dati storici relativi alle vendite, in modo da estrapolarne l'andamento ad un periodo successivo.

Si tratta dunque di un metodo piuttosto semplice, basato sulle usuali tecniche di analisi delle serie storiche, mediante le quali i dati originari vengono depurati della componente stagionale, nonché della ciclica e della accidentale. Ciò richiede l'assunzione che il futuro andamento delle vendite possa essere dedotto da quelle che sono state le manifestazioni passate dello stesso fenomeno, ritenendo costanti le condizioni che hanno dato origine alla passata evoluzione.

Questo metodo di previsione, quindi, consente risultati accettabili se rimangono costanti per un certo periodo di tempo i fattori che contribuiscono alla crescita o alla diminuzione del fenomeno da prevedere.

Per poter formulare ipotesi sulla più probabile evoluzione futura delle vendite, si applicano alcune funzioni matematiche in grado di rappresentare significativi modelli tipici della dinamica delle stesse per un certo periodo di tempo. È buona norma procedere ad una preliminare rappresentazione grafica dei dati storici, in modo da

evidenziare un certo tipo di andamento e selezionare a priori una famiglia di funzioni adeguate, scegliendo poi tra queste quella che meglio di ogni altra compenetra tra loro significatività statistica e semplicità di calcolo e di interpretazione.

Individuata la funzione matematica che meglio spiega l'andamento delle vendite, si tratta di applicare ai dati il procedimento dei minimi quadrati per il calcolo dei parametri.

Non sempre l'applicazione di questa metodologia conduce a risultati univocamente determinati, nel senso che a volte più funzioni danno luogo a indici statistici ugualmente accettabili; in tal caso si può usare il metodo dell'analogia, o della previsione per confronto, condensando in una unica valutazione i risultati ottenuti.

Poste, allora, le osservazioni ai tempi 1, 2, ..., t, ..., s:

$$y_1,\ y_2,\ \ldots,\ y_t,\ \ldots,\ y_s\ ;$$

se la funzione prescelta è $\hat{y}_t = a + bt$, il sistema per il calcolo dei parametri risulta:

$$\begin{cases} sa + b\sum_{t=1}^{s} t = \sum_{t=1}^{s} y_t \\ a\sum_{t=1}^{s} t + b\sum_{t=1}^{s} t^2 = \sum_{t=1}^{s} y_t t \end{cases} ;$$

una volta stimati tali parametri, la previsione all'istante t=s+k sarà:

$$y_{s+k} = a + b(s+k)\ .$$

Tuttavia poiché in questi casi le previsioni possono avere carattere iterativo, sorge il problema del modo con cui modificare il valore dei parametri dell'espressione lineare impiegata per effettuare le previsioni dopo che diventa finalmente disponibile una nuova osservazione della serie relativa all'istante s+1. Pertanto assume primaria importanza poter disporre di un procedimento che consenta di stimare i nuovi parametri utilizzando i calcoli effettuati in precedenza[6].

Per impostare tale procedura, operiamo innanzitutto una traslazione sull'asse dei tempi e spostiamo l'origine all'istante (s+1)/2. Ciò equivale a porre

$$z_t = t - \frac{s+1}{2}$$

ad esempio:

$$z_1 = 1 - \frac{s+1}{2} = -\frac{s-1}{2} \qquad\qquad z_2 = 2 - \frac{s+1}{2} = -\frac{s-3}{2}$$

Pertanto i nuovi tempi saranno:

$$-\frac{s-1}{2}, -\frac{s-3}{2}, \ldots\ldots, \frac{s-3}{2}, \frac{s-1}{2}$$

[6] Giusti F., Vitali O., *Statistica economica*, op. cit..

e la funzione rappresentatrice sarà dalla seguente:

$$\hat{y}_{z_t} = {}_s a + {}_s b z_t$$

I suoi parametri saranno ricavati dal sistema:

$$\begin{cases} s \cdot {}_s a + {}_s b \sum z_t = \sum y_t \\ {}_s a \sum z_t + {}_s b \sum z_t^2 = \sum y_t \cdot z_t \end{cases},$$

per t che varia da –(s-1)/2 a (s-1)/2; è semplice verificare che la sommatoria delle z_t è uguale a 0, per cui il calcolo dei parametri è immediato:

$${}_s a = \frac{\sum y_t}{s}; \qquad {}_s b = \frac{\sum y_t z_t}{\sum z_t^2};$$

Si dimostra[7] che:

$$\sum_{t=-(s-1)/2}^{(s-1)/2} z_t^2 = \frac{s\left(s^2-1\right)}{12};$$

pertanto:

$${}_s \hat{b} = \frac{12\sum y_t z_t}{s\left(s^2-1\right)} = \frac{12\left(-\frac{s-1}{2}y_1 - \frac{s-3}{2}y_2 + \ldots + \frac{s-3}{2}y_{t-1} + \frac{s-1}{2}y_t\right)}{s\left(s^2-1\right)} =$$

[7] La somma dei primi s numeri naturali elevati al quadrato è fornita da: $\sum_{t=1}^{s} t^2 = \frac{s(s+1)(2s+1)}{6}$ di conseguenza, ponendo z = 1-(s+1)/2, si ottiene che la sommatoria di z^2 al tempo t diventa: $\sum_{t=-\frac{s-1}{2}}^{\frac{s-1}{2}} z_t^2 = 2\cdot\sum_{t=0}^{\frac{s-1}{2}} z_t^2 =$

$$= 2\frac{\left(\frac{s-1}{2}\right)\left(\frac{s-1}{2}+1\right)\left(2\frac{s-1}{2}+1\right)}{6} = \frac{\left[\left(\frac{s-1}{2}\right)^2 + \left(\frac{s-1}{2}\right)\right]s}{3} = \frac{\left(\frac{s^2-2s+1}{4} + \frac{s-1}{2}\right)s}{3} =$$

$$= \frac{1}{3}\cdot\left(\frac{s^2-1}{4}\right)s = \frac{s\left(s^2-1\right)}{12}$$

Cfr. Girone G.-Salvemini T., *Lezioni di statistica, vol.2*, Cacucci, Bari, 2000.

$$= \frac{12}{s\left(s^2-1\right)}\left[-\frac{s-1}{2}y_1 - \frac{s-3}{2}y_2 + ... + \frac{s-3}{2}y_{t-1} + \frac{s-1}{2}y_t\right] .$$

La previsione a *k* istanti futuri è data dalla seguente espressione:

$$\hat{y}_{\frac{s-1}{2}+k} = {}_s a_t + {}_s b\left(\frac{s-1}{2}+k\right)$$

Supponiamo che venga ora acquisito il nuovo dato y_{S+1} della serie oggetto di studio; i nuovi parametri della funzione rappresentatrice risultano, aggiungendo il nuovo dato e sottraendo il primo:

$$_{s+1}a = \frac{\sum y_s + y_{s+1} - y_1}{s} = {}_s a + \frac{y_{s+1} - y_1}{s};$$

$$_{s+1}a = {}_s a + \frac{12}{s(s^2-1)}\left(\frac{s-1}{2}y_{s+1} + \frac{s+1}{2}y_1 - s\cdot {}_s a\right) .$$

Ad esempio, se consideriamo una serie formata da sette osservazioni annuali relative alle vendite di un determinato bene (Tab. 5.1), si osserva che sulla base del metodo dell'aggiornamento dei parametri l'ottava osservazione si calcola nel modo seguente, dopo aver preventivamente calcolato i due coefficienti di regressione:

$${}_s a = \frac{\sum y_t}{s} = \frac{737,7}{7} = 105,39; \qquad {}_s b = \frac{\sum y_t z_t}{\sum z_t^2} = \frac{56,3}{28} = 2,01;$$

$per\, t \in \left[-\frac{s-1}{2}, \frac{s-1}{2}\right]$ si ricava:

$$\hat{y}_{z_t} = 105,39 + 2,01 \cdot z_t .$$

La previsione a *k* istanti futuri è data

$$\hat{y}_{z_t+k} = \hat{y}_{\frac{s-1}{2}+k} = {}_s a_1 + {}_s a\left(\frac{s-1}{2}+k\right)$$

nel nostro caso, per s=7 e per k=1 sarà:

$$\hat{y}_4 = {}_s a + {}_s b \cdot 4 = 105,39+2,01\cdot 4 = 113,42$$

Tab. 5.1- Esempio di aggiornamento dei parametri di una interpolante lineare.

t	Anni	Vendite	y_t = n..i. 1980=100	$z_t = t - \frac{s+1}{2}$			*pesi modificati dopo aver acquisito il nuovo dato (1992)*
1	1996	302.181	100,0	$-\frac{s-1}{2} = -3$	9	-300,0	$-\left(\frac{s-1}{2}+1\right) = -4$
2	1997	307.388	101,7	$-\frac{s-3}{2} = -2$	4	-203,4	$-\left(\frac{s-3}{2}+1\right) = -3$
3	1998	311.721	103,6	$-\frac{s-5}{2} = -1$	1	-103,6	$-\left(\frac{s-5}{2}+1\right) = -2$
4	1999	314.357	104,0	$-\frac{s-7}{2} = 0$	0	-	$-\left(\frac{s-7}{2}+1\right) = -1$
5	2000	321.167	106,3	$\frac{s-5}{2} = 1$	1	106,3	$\frac{s-5}{2} - 1 = 0$
6	2001	330.317	109,3	$\frac{s-3}{2} = 2$	4	218,6	$\frac{s-3}{2} - 1 = 1$
7	2002	340.728	112,8	$\frac{s-1}{2} = 3$	9	338,4	$\frac{s-1}{2} - 1 = 2$
	Totali		*737,7*		*28*	*56,3*	
8	2003		113,6				$\frac{s-1}{2} = 3$

Fonte: elaborazione su dati dell'indagine empirica.

N.B.: Si fa notare che i parametri, rispetto ai valori originari, sono:

$$_{s+1}b = {}_sb = 2{,}01 \quad ,$$

mentre $_{s+1}a$ sarà dato da

$$_{s+1}a = {}_sa - {}_sb\left(\frac{s+1}{2}\right) = 105{,}39 - 2{,}01\left(\frac{7+1}{2}\right) = 97{,}34;$$

pertanto la previsione all'anno s+1 con tali parametri è ancora:

$$\hat{y}_{s+1} = 97{,}34 + 2{,}01 \cdot 8 = 113{,}42 \ .$$

Ritorniamo, adesso, ai parametri $_s$a e $_s$b, ma nell'osservare in particolare il secondo che, sulla base dei sette dati originari che compongono la serie delle vendite, si ha:

$$_s b = \frac{\sum y_t z_t}{\sum z_t^2} =$$

$$= \frac{12}{s(s^2-1)}\left[-\frac{s-1}{2}y_1 - \frac{s-3}{2}y_2 - \frac{s-5}{2}y_3 - \frac{s-7}{2}y_4 + \frac{s-5}{2}y_5 + \frac{s-3}{2}y_6 + \frac{s-1}{2}y_7\right].$$

Supponiamo che venga acquisito il nuovo dato relativo al 2003, che, trasformato in termini relativi (1996=100), risulta pari a 113,6. Volendo rideterminare i parametri e far concorrere solo gli ultimi sette dati, dobbiamo applicare il concetto di media mobile ai vecchi parametri; in altri termini, dobbiamo trascurare il primo dato e considerare l'ultimo dato acquisito. Si osservi che ciò comporta una modifica dei pesi originari, così come si legge nell'ultima colonna della tabella. Pertanto si avrà, nell'esempio considerato:

$$_{s+1}a = {}_s a + \frac{y_{s+1} - y_1}{s} = 105{,}39 + \frac{113{,}6 - 100}{7} = 105{,}39 + 1{,}94 = 107{,}3$$

$$_{s+1}b = {}_s b + \frac{12}{s(s^2-1)}\left(\frac{s-1}{2}y_{s+1} + \frac{s+1}{2}y_1 - s \cdot {}_s a\right) =$$

$$= 2{,}01 + \frac{12}{336}\left(4 \cdot 100 + 3 \cdot 113{,}6 - 737{,}7\right) = 2{,}1216$$

La previsione delle vendite al nuovo tempo 9, ossia al 2004, sarà quindi (sempre posto il 1996=100):

$$\hat{y}_4 = a + b \cdot 4 = 107{,}3 + 2{,}1216 \cdot 4 = 115{,}79$$

6. Metodo del livellamento esponenziale

Il metodo del livellamento esponenziale (*esponential smooting*) è utilizzato di frequente in ambito aziendale per la perequazione dei dati e la previsione a breve termine. Questo metodo parte dalla considerazione che il dato più recente fornisca il contributo più importante alla previsione delle vendite (y). Questa impostazione conduce a determinare un tasso di decremento di importanza dei termini della serie cronologica oggetto di studio, denominato costante di livellamento, di solito indicato con la lettera α, la quale assume un valore compreso tra 0 e 1.

Data una serie storica, ad esempio le vendite, si vuole prevedere il fenomeno ad un tempo $s + k$, in cui $k \geq 1$ è *l'orizzonte di previsione*.

Posto ad esempio una serie con trend costante, o tutto al più alterato da fattori accidentali ($y_t = a + \varepsilon$) è logico che il calcolo della previsione al tempo ***s+1*** coincide con la media semplice dei valori osservati fino all'istante s con peso uguale.

Pertanto indicando con:

$$y_1, y_2, \ldots\ldots y_t, \ldots\ldots y_s$$

la serie storica delle vendite ai tempi da *1, 2,... t,.....s* la previsione al tempo *s+1* sarà:

$$\hat{y}_{s+1} = \frac{1}{s}\sum_{t=1}^{s} y_t \ .$$

Se invece il livello della serie muta in modo stocastico (ossia se la serie è costante solo in alcuni istanti di tempo) è più realistico calcolare $\hat{v}_{n-1}$ attraverso una media *ponderata* delle osservazioni precedenti dando peso maggiore a quelle più recenti.

Posto che il dato più recente fornisca un contributo più importante alla previsione, il metodo del livellamento esponenziale viene così denominato perché alla successione delle vendite v_t viene sostituita una successione " perequata" che si configura come una media ponderata di tutte le osservazioni disponibili in cui i pesi sono costituiti da una *successione esponenziale* con intensità decrescente all'aumentare della distanza dal tempo t in cui viene effettuata la previsione.

Pertanto posto un determinato tasso di decremento di importanza dei termini della serie cronologica oggetto di studio, denominato *costante di livellamento*, di solito indicato con la lettera α, la quale assume un valore compreso tra 0 e 1 e con $y_1, y_2, \ldots\ldots y_t$ i valori assunti da una serie storica delle vendite in corrispondenza, rispettivamente, dei tempi *1, 2, ..., t*; la previsione al periodo *t+1* si ottiene mediante la formula:

$$y^*_{t+1} = \alpha y_t + (1-\alpha) y^*_t \qquad [1]$$

In cui, come è stato detto, α è il coefficiente di livellamento, mentre y^*_t è il valore livellato al tempo t.

Si consideri che il valore livellato al tempo *t* è:

$$y^*_t = \alpha y_{t-1} + (1-\alpha) y^*_{t-1} \qquad [2]$$

per cui, sostituendo si ha:

$$y^*_{t+1} = \alpha y_t + (1-\alpha)\left[\alpha y_{t-1} + (1-\alpha) y^*_{t-1}\right] =$$

$$= \alpha y_t + \alpha(1-\alpha) y_{t-1} + (1-\alpha)^2 y^*_{t-1} \ . \qquad [3]$$

Per sostituzioni successive si ottiene la formula generale:

$$y_{t+1}^{*} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \ldots + \alpha(1-\alpha)^{t-1} y_1 + (1-\alpha)^t y_1^{*} \quad [4]$$

Essendo $0 \le (1-\alpha) \le 1$, il peso attribuito ai termini della serie va decrescendo man mano che ci si allontana nel tempo.

Dal punto di vista operativo il metodo richiede la determinazione della costante di livellamento la stima del dato relativo alla prima previsione y_1^{*} per iniziare il calcolo della [4] .Questo dato può essere il primo valore della serie oppure un valore medio. Tuttavia occorre precisare che la scelta arbitraria ha una bassa incidenza sui calcoli essendo tale valore moltiplicato per il coefficiente $(1-\alpha)^t$, valore molto piccolo; pertanto l'errore risulta trascurabile.

Punto di forza di questo metodo è che per effettuare la previsione sono sufficienti tre soli dati: α, y_1^{*}, e y_S.

Per quanto riguarda il valore di α, ci si affida all'esperienza del ricercatore, oppure può essere ottenuto minimizzando l'errore quadratico medio. All'aumentare di α, aumentando il peso attribuito ai termini più recenti, aumenta anche il potere del modello di adattarsi a variazioni dei valori osservati, per cui esso rappresenta la reattività del modello.

Tab. 6.1 - Esempio di previsione delle vendite con il metodo del livellamento esponenziale.

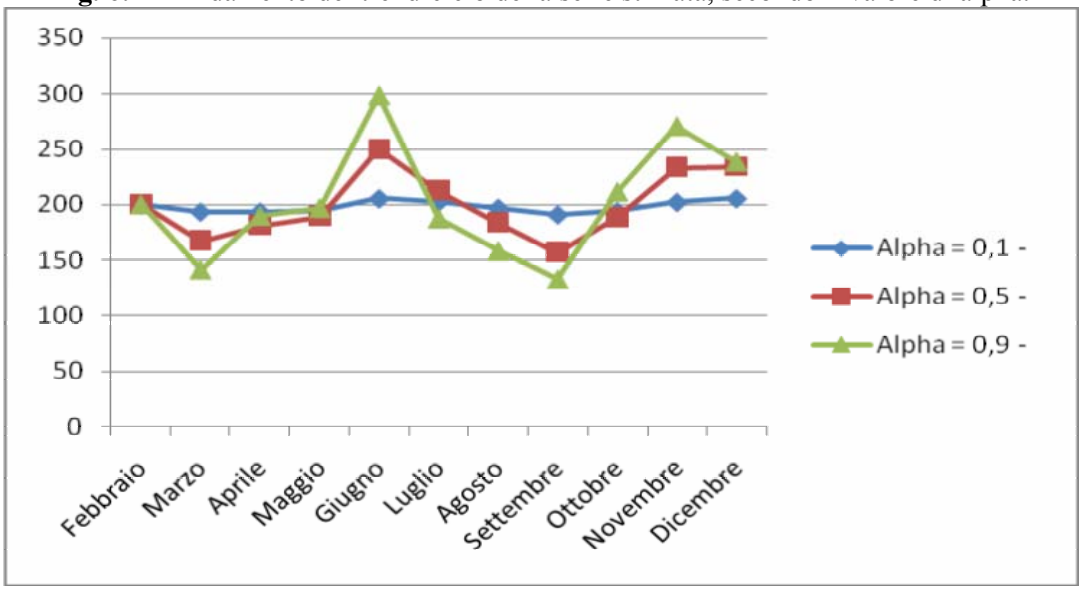
			Valori calcolati con coefficiente:		
Mese	Periodo	Valori osservati	$\alpha = 0,1$	$\alpha = 0,5$	$\alpha = 0,9$
Gennaio	1	200,0	-	-	-
Febbraio	2	135,0	200,0	200,0	200,0
Marzo	3	195,0	193,5	167,5	141,5
Aprile	4	197,5	193,6	181,3	189,7
Maggio	5	310,0	194,0	189,4	196,7
Giugno	6	175,0	205,6	249,7	298,7
Luglio	7	155,0	202,6	212,3	187,4
Agosto	8	130,0	197,0	183,7	158,2
Settembre	9	220,0	191,0	156,8	132,8
Ottobre	10	277,5	193,9	188,4	211,3
Novembre	11	235,0	202,3	233,0	270,9
Dicembre	12	-	205,6	234,0	238,6

Risulta, per α=0,1:

$$y_2^{*} = 0,1 \cdot 135 + (1 - 0,1) \cdot 200 = 193,5$$
$$y_3^{*} = 0,1 \cdot 195 + (1 - 0,1) \cdot 193,5 = 193,6$$
$$y_4^{*} = 0,1 \cdot 235 + (1 - 0,1) \cdot 202,3 = 205,6$$

Similarmente si ottengono le previsioni per α=0,5 e α=0,9.

Fig. 6.1 – Andamento del trend-ciclo della serie stimata, secondo il valore di alpha.



La discrezionalità del metodo deriva dalla scelta arbitraria del parametro α, che ha effetti sul grado del livellamento della serie stimata. Considerato che un adattamento perfetto dell'osservato presuppone che le misure siano esatte e quindi gli errori siano nulli, si osserva che quanto minore è il valore prescelto per α, tanto minore sarà infatti l'appiattimento della serie stimata di trend-ciclo, come risulta evidente dalla figura 6.1.

Tale suddetta arbitrarietà può essere limitata scegliendo il parametro stesso secondo un criterio di ottimalità; in sostanza, per misurare la bontà del livello di adattamento del modello nel periodo osservato e la relativa capacità di estrapolazione previsiva sulla base dei diversi valori di α, possono essere presi in considerazione gli errori medi e assoluti, derivati dall'implementazione dei diversi modelli di livellamento esponenziale.

7. Le previsione con modelli stocastici: il procedimento di BOX E JENKINS

L'impostazione moderna delle serie cronologiche nasce dalla considerazione che la descrizione del loro comportamento mediante semplici funzioni del tempo emerge dall'applicazione di procedure statistiche neutrali, cioè identiche per tutte le serie storiche, senza tener conto della natura del meccanismo che determina il prodursi dei dati della serie, delle strutture probabilistiche ad esse eventualmente soggiacenti.

Per far sì che sia la serie osservata ad orientare verso un possibile modello generatore dei dati che la costituiscono, è necessario rifarsi al concetto di "processo stocastico". Ciò comporta che i valori rilevati siano interpretabili come una delle tante realizzazioni finite di uno stesso processo stocastico. Si tratta, in sostanza, di capire che tipo di meccanismo ha agito sull'andamento della variabile considerata (le vendite) meccanismo che si può ripetere nel futuro. L'analisi stocastica delle serie temporali si

fonda sul presupposto che le manifestazioni di un fenomeno come le vendite siano prodotte da una struttura probabilistica che non conosciamo, ma della quale possiamo stimare i parametri ricorrendo ai metodi dell'inferenza induttiva mediante lo studio delle relazioni temporali che riusciamo a riconoscere fra i dati rilevati.

L'impiego dei modelli stocastici nella previsione delle vendite si adatta bene, ad esempio, quando i dati di base sono variazioni trimestrali delle vendite del prodotto k espresse in quantità fisiche in un dato intervallo temporale.

L'assenza di stagionalità è una condizione che rende più spedita la procedura da seguire e consente un corretto impiego del modello stocastico.

Alla base del metodo di Box e Jenkins vi è la nozione di funzione di autocorrelazione totale e parziale e, inoltre, una particolare classe di processi stocastici cosiddetti a parametro discreto caratterizzati da particolari requisiti. Nella ricerca del modello che descrive la serie che riproduca il meccanismo che la genera è importante tener presente come individuare il processo generatore dei dati della serie, come verificarne la scelta e come utilizzare il risultato per la previsione. La sua applicazione richiede, allora, come presupposto che tra i dati della serie vi sia autocorrelazione, ciò sta a significare che i coefficienti di correlazione totale e parziale devono essere calcolati sui dati della serie e sugli stessi dati slittati k volte. Dall'andamento di queste due funzioni si cerca di risalire mediante un procedimento inferenziale alla stima dei parametri del modello rappresentativo del processo.

In sintesi, le fasi principali per arrivare a specificare e stimare un modello utilizzabile a fini previsionali sono le seguenti:

- analisi preliminare dei dati di base:
- identificazione del modello,
- stima dei parametri,
- verifica della qualità statistica delle stime,previsione.

7.1. Definizione del processo stocastico[8]

Un processo stocastico, o processo aleatorio X_S è costituito da una famiglia di variabili casuali Y_1, Y_2,...,Y_t,…,Y_s descritte ed ordinate da un parametro t che indica il tempo e che intenderemo discreto.

Fissato un valore t (ad esempio t=2) ed una determinata prova sperimentale (cioè osservando il valore che per la v.c. $Y_t = Y_2$) si ottiene un numero reale interpretabile come realizzazione finita della v.c. considerata. Per le serie cronologiche, pertanto, gli s valori osservati y_1, y_2..., y_S, ecc. ai tempi 1, 2..., s, ecc. costituiscono nel loro insieme una parte finita di una famiglia, di una singola realizzazione di un processo stocastico soggiacente.

Utilizzando tali valori "campionari", si tratta di ricavare con procedure statistiche stime attendibili delle grandezze che caratterizzano il processo generatore.

Ovviamente ogni processo stocastico ha sue caratteristiche e proprietà e un processo stocastico può essere conosciuto solo se sono note le funzioni di densità per ogni k-pla di variabili componenti. E' facile rendersi conto dei complessi problemi che si presentano per studiarne le caratteristiche e le proprietà; ciò induce a semplificare e a

[8] Giusti F., Vitali O. (1983), *Statistica Economica*, Cacucci Editore, Bari.

prendere in considerazione soltanto particolari processi stocastici per i quali è più facile applicare la metodologia inferenziale.

Per ogni v.c. componente il processo stocastico, possono definirsi il valore medio e la varianza e considerando a due a due le v.c. si può calcolare l'autocovarianza e pertanto si avrà:

Inoltre per facilitare i confronti tra diversi processi stocastici, la funzione di autocovarianza viene standardizzata dividendola per la funzione varianza e si ottiene la funzione autocorrelazione.

Un particolarissimo processo stocastico è quello denominato "White noise" (rumore bianco); tale processo si indica con il simbolo W_t ed è costituito da v.c. W_1, W_2..., incorrelate, ha valore medio nullo e varianza σ_w costante ed indipendente dal tempo t. Se le v.c. sono normali, si ha il processo stocastico "White noise" gaussiano.

Nell'analisi delle serie cronologiche assumono notevole importanza i processi stocastici stazionari. Esistono tuttavia almeno due definizioni di stazionarietà. La prima, denominata stazionarietà in senso forte, è molto restrittiva in quanto implica l'invarianza delle varie strutture probabilistiche che caratterizzano il processo ed è scarsamente applicabile nei casi concreti.

Pertanto è preferibile far riferimento ad una definizione meno rigida, definendo come processo stocastico in senso debole, o semplicemente "processo stazionario", quei processi per i quali:

1) Il valore medio è costante al variare del tempo:
$$E(Y_t) = \mu_t ;$$
questa è la condizione di invarianza in media, che implica l'assenza di trend nella stessa media.
2) La varianza è finita e costante al variare del tempo:
$$E(Y_t - \mu)^2 = T^2$$
questa è la condizione di omoscedasticità.
3) L'autocovarianza tra X_t e $X_{t+\tau}$ dipende soltanto dallo sfasamento temporale (lag): τ
$$\gamma(\tau) = E(Y_t - \mu)(Y_{t-\tau} - \mu)$$
ed esprime pertanto la connessione fra le v.c. componenti al variare della loro distanza.

Per facilitare confronti tra diversi processi stocastici, la funzione di autocovarianza viene standardizzata dividendo per la varianza. Si ottiene così la funzione di autocorrelazione $\rho(\tau)$, simmetrica, nel senso che $\rho(\tau) = \rho(-\tau)$, la quale esprime il coefficiente di autocorrelazione binario fra Y_t e $Y_{t+\tau}$.

$$\rho(\tau) = \frac{E(Y_t - \mu)(Y_{t-\tau} - \mu)}{E(Y_t - \mu)^2}$$

che può scriversi anche nella forma:

$$\rho(\tau) = \left[\frac{(Y_t - \mu)}{\sigma}\right] \cdot \left[\frac{(Y_{t-\tau} - \mu)}{\sigma}\right] = \frac{\gamma(\tau)}{\gamma(0)} \quad .$$

Il grafico della funzione di autocorrelazione al variare di τ è detto correlogramma. Pertanto i coefficienti di autocorrelazione, $\rho(\tau)$, misurano la concordanza o la discordanza tra i valori della serie storica e quelli differiti di τ unità di tempo; consentendo di analizzare la sua struttura interna, ossia i legami tra i termini della stessa. In particolare $\rho(0)=1$ (coefficiente di correlazione tra la serie e sé stessa); gli altri coefficienti sono compresi tra -1 e 1.

Infine, un processo stocastico si definisce "invertibile" se per ogni valore del parametro t sussiste la relazione:

$$Y_t = f(Y_{t-1},\ Y_{t-2}....) + W_t$$

Un processo invertibile può essere espresso cioè da una funzione della storia trascorsa del fenomeno e da un rumore bianco.

Volendo, allora, analizzare una serie storica di un fenomeno economico, giova ricordare che essa rappresenta una realizzazione finita di un processo sconosciuto di cui occorre ricercare le caratteristiche, impiegando procedure di inferenza statistica.

Il problema è indubbiamente complicato dal fatto che i dati y_1, y_2..., y_n della serie costituiscono una successione di n campioni di ampiezza unitaria su altrettante v.c. distinte.

Tuttavia se il processo è stazionario e se sono verificate, rispetto ai vari parametri, alcune condizioni di ergodicità[9], è possibile derivare dai dati stime consistenti dei parametri stessi e delle funzioni che caratterizzano il processo.

Date tali ipotesi, si hanno le seguenti stime:

per la media: $$\hat{\mu} = \frac{\sum_{t=1}^{s} y_t}{s} \quad ;$$

per la varianza: $$S^2 = \hat{\gamma}(0) = \frac{\sum_{t=1}^{s} (y_t - \mu)^2}{s-1} \quad ;$$

per l'autocovarianza: $$\hat{\gamma}(\tau) = \frac{\sum_{t=1}^{s-\tau} (y_t - \mu)(y_{t-\tau} - \mu)}{s} \quad ;$$

[9]La condizione che assicura l'ergodicità di un processo rispetto al valore medio è che la sua funzione di correlazione tenda a zero; un processo si dice ergodico se la media relativa a un gran numero di manifestazioni in un dato istante di tempo tende alla media temporale.

per l'autocorrelazione: $$\hat{\rho}(\tau) = \frac{\sum_{t=1}^{s-\tau} (y_t - \mu)(y_{t-\tau} - \mu)}{\sum_{t}^{s} (y_t - \mu)^2} \quad ;$$

per l'autocorrelazione parziale: $$\hat{\phi}_{\tau\tau} = \frac{|Q_\tau|}{|P_\tau|}$$

ottenuta calcolando i determinanti delle matrici si ottiene:

$$|Q_T| = \begin{vmatrix} 1 & \hat{\rho}(1) & \hat{\rho}(2) & \ldots. & \hat{\rho}(1) \\ \hat{\rho}(1) & 1 & \hat{\rho}(1) & \ldots. & \hat{\rho}(2) \\ \hat{\rho}(2) & \hat{\rho}(1) & 1 & \ldots. & \hat{\rho}(3) \\ \ldots. & \ldots. & \hat{\rho}(1) & \ldots. & \ldots. \\ \hat{\rho}(\tau-1) & \ldots. & \hat{\rho}(\tau-3) & \ldots. & \hat{\rho}(\tau) \end{vmatrix}$$

$$|P_T| = \begin{vmatrix} 1 & \hat{\rho}(1) & \hat{\rho}(2) & \ldots. & \hat{\rho}(\tau-1) \\ \hat{\rho}(1) & 1 & \hat{\rho}(1) & \ldots. & \hat{\rho}(\tau-2) \\ \hat{\rho}(2) & \hat{\rho}(1) & 1 & \ldots. & \hat{\rho}(\tau-3) \\ \ldots. & \ldots. & \hat{\rho}(1) & \ldots. & \ldots. \\ \hat{\rho}(\tau-1) & \ldots. & \hat{\rho}(\tau-3) & \ldots. & 1 \end{vmatrix}$$

7.2. Modelli AR, MA, ARMA e ARIMA.

Il processo stocastico autoregressivo di ordine "p", che si indica con il simbolo AR(p), è un processo che genera la serie storica:

$$Y_t = \eta_1 Y_{t-1} + \eta_2 Y_{t-2} + ... + \eta_p Y_{t-p} + W_t$$

cioè il valore osservato al tempo t è dato dalla combinazione lineare di p termini immediatamente precedenti e da una componente aleatoria, per cui il dato attuale è il risultato di tante manifestazioni passate.

Il processo stocastico a media mobile di ordine "q", che si indica con il simbolo MA(q), è, invece, un processo che genera la serie storica:

$$Y_t = W_t + \varphi_1 W_{t-1} + \varphi_2 W_{t-2} + \varphi_q W_{t-q}$$

cioè il valore osservato al tempo t è dato dalla combinazione lineare dei valore di una componente casuale allo stesso tempo t e nei q tempi immediatamente precedenti, per cui il dato attuale si configura come il risultato di una successione di impulsi aleatori.

Le serie storiche di molti fenomeni economici possono essere efficacemente descritte da modelli riconducibili ai processi AR (autoregressivi) o MA (media mobile) oppure da modelli misti costituiti congiuntamente da una componente autoregressiva e da una componente media mobile.

Emergono pertanto i processi ARMA (processi stocastici autoregressivi di ordine p e media mobile di ordine q.

In altri termini da un punto di vista interpretativo si può ragionevolmente pensare che il valore di x_t di una serie storica osservato al tempo t=1, 2, ... s sia funzione lineare dei valori osservati in tempi precedenti (la struttura AR) e di avvenimenti casuali verificatisi nel tempo t e in tempi precedenti (la struttura MA).

Nell'ambito dei processi ARMA, notevole rilievo assumono quelli stazionari e invertibili, nonché la maggior parte delle serie storiche di natura economica e di tipo evolutivo crescenti o decrescenti. E' necessario in via preliminare effettuare differenze e trasformazioni dei valori osservati y_t, in modo che la serie risultante "Y_t" possa ritenersi generata da un processo stocastico stazionario. Le trasformazioni più frequenti sono date da:

$Y_t = \Delta y_t = y_t - y_{t-1}$ differenze prime;

$Y_t = \Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$ differenze seconde;

$Y_t = \Delta \log y_t = \log y_t - \log y_{t-1}$ differenze prime dei logaritmi.

Nel caso di dati mensili:

$Y_t = \Delta_{12} y_t = y_t - y_{t-12}$ differenze con sfasamento di 12 mesi;

$Y_t = \Delta_{12} \log y_t = \log y_t - \log y_{t-12}$ differenze prime dei logaritmi con sfasamento di dodici mesi.

I modelli stocastici relativi ai valori delle differenze vengono chiamati integrati ed indicati con la sigla ARIMA, cioè processo stocastico autoregressivo di ordine ***p*** integrato ***d*** volte e media mobile di ordine ***q.***

Per la scelta del tipo di differenze e della trasformazione più opportuna di una serie storica, ci si avvale solitamente del grafico dei valori y_t; in base ad esso, si determina l'operatore ritenuto adeguato, e si effettua la rappresentazione grafica dei valori y_t, che dovrebbero apparire stazionari. In caso negativo si considerano differenze o trasformazioni ulteriori.

Il modello - che deve essere individuato ed i cui parametri devono essere stimati - è lo strumento impiegato per descrivere o interpretare un fenomeno governato da leggi probabilistiche, noto come processo stocastico.

E' consuetudine indicare i modelli di Box-Jenkins come modelli della classe ARMA, perché la loro struttura è formata da componenti autoregressive, ossia osservazioni sulla stessa serie sfasata temporalmente, e da componenti a media mobile, anch'esse ritardate.

Il procedimento di Box-Jenkins inizia con un esame preliminare dei dati della serie per studiare le caratteristiche (mediante grafici, indicatori, calcolo di costanti, ecc.) e per individuare le trasformazioni più convenienti capaci di ricondurre la serie oggetto di analisi ad una serie modificata esprimibile con un modello ARMA (un modello si dice invertibile quando nella sua parte a media mobile deve dare un minor peso alle osservazioni meno recenti).

L'eliminazione del trend può effettuarsi o interpolando tra punti una funzione matematica atta a rappresentarlo, o determinando le differenze di ordine sufficientemente elevato per assicurare la stazionarietà.

Se la varianza non è costante al variare del tempo (eteroscedasticità), prima di eliminare il trend conviene operare una trasformazione logaritmica.

Quando la serie è mensile, trimestrale, ecc. e quando si ha ragione di ritenere che la stagionalità giochi un ruolo rilevante, occorre eliminarne gli effetti, ciò che in sede operativa viene realizzato prendendo in considerazione le differenze tra i valori della serie sfasati di 12 mesi, di 3 mesi, ecc. Con tale artificio l'influenza della stagionalità viene quanto meno contenuta.

Per l'accertamento della stazionarietà in media e varianza della serie temporale – oltre ad esaminare graficamente le oscillazioni circa l'andamento della stessa serie intorno al relativo valore medio - è necessario calcolare i coefficienti di autocorrelazione globale e parziale; prima di arrivare al calcolo degli stessi coefficienti è necessario calcolare la media, la varianza e l'autocovarianza che ne costituiscono i presupposti indispensabili.

La stazionarietà in media si manifesta mediante una funzione di autocorrelazione globale che scende verso lo zero con molta rapidità. Se la serie è stazionaria il coefficiente $\boldsymbol{\eta}$ deve inoltre risultare in valore assoluto minore dell'unità.

Consideriamo una serie storica formata da n osservazioni, una regola pratica suggerisce di calcolare i coefficienti di correlazione totale e parziale in numero non superiore ad un quarto; se, inoltre, i coefficienti sono significativamente diversi da zero occorre per ognuno di essi calcolare il test $\boldsymbol{t}$ e confrontarlo con i valori soglia.

7.3. Identificazione di un modello della classe ARMA

Per identificare il modello si possono confrontare i coefficienti di autocorrelazione calcolati con modelli teorici di correlogrammi relativi a processi stazionari AR che sono di andamento noto (riportati in appendice).

Scelto il modello si tratta di stabilire il grado dello stesso, questa decisione è in funzione della autocorrelazione parziale; prima di tutto, è necessario effettuare un controllo sul coefficiente autoregressivo η tramite il *test t* ponendo come sempre l'ipotesi che sia $\eta = 0$.

Si passa infine alla stima della funzione di autocovarianza e di autocorrelazione, i cui andamenti sono in grado di suggerire gli ordini p e q dei processi autoregressivi a media mobile sottostanti. E' questa certamente la fase più complessa e impegnativa del procedimento, in cui si riscontra la maggiore dose di soggettività, specialmente se il processo soggiacente alla serie trasformata non è solamente autoregressivo o solamente media mobile.

E' necessario orientarsi verso modelli che non contengono parametri in numero elevato.

L'idoneità del modello adottato per descrivere le caratteristiche della serie, intesa come realizzazione finita del processo stocastico da esso configurato, viene identificata effettuando varie analisi dei residui per controllare la loro casualità e normalità. Si procede, pertanto, ad esami grafici, allo studio delle caratteristiche distributive dei residui e specialmente alla stima della loro autocorrelazione, costruendo altresì bande di confidenza da utilizzarsi quali criteri di adeguatezza del modello. Se infatti uno o più valori ricadono all'esterno delle bande, il modello dovrà essere certamente perfezionato.

Se a seguito dei controlli effettuati il modello viene accettato, esso può essere utilmente impiegato per varie finalità di descrizione, di previsione, di simulazione e di analisi, ecc. Se viene rifiutato, viene iterata la successione delle fasi, che vanno dalla identificazione alla verifica, tenendo conto degli elementi emersi che hanno condotto al suo rigetto.

I punti deboli sono costituiti essenzialmente dalla insicurezza di garantire la validità delle numerose ipotesi di base e dall'elevato grado di soggettività insito in talune fasi del procedimento.

7.4. Un'applicazione empirica con il procedimento di Box e Jenkins

Nella presente esemplificazione empirica, si considera una serie mensile composta di quaranta periodi di osservazione, relativa alle vendite di un determinato prodotto di largo consumo; nel supporre che la suddetta serie sia stata prodotta da una struttura probabilistica che non conosciamo, si vuole stimare i relativi parametri ricorrendo ai metodi dell'inferenza induttiva mediante lo studio delle relazioni temporali che riusciamo a riconoscere fra i dati rilevati.

Per stimare un adeguato modello di previsione si è supposto in partenza di suddividere la serie stessa in due parti: la prima – composta dai periodi che vanno dal tempo t_1 al tempo t_{36} – per implementare un adeguato modello di previsione e stimarne i relativi parametri, la seconda – dal tempo t_{37} al tempo t_{40} – per validare il modello stimato ai fini previsivi (Tab. 7.4.1)[10].

[10] Brasini S., Freo M., Tassinari F., Tassinari G. (2002), Statistica aziendale e analisi di mercato, Il Mulino, Bologna.

Tab. 7.4.1 – Distribuzione delle vendite.

Tempo	**Vendite**	**Lag**					
		t-1	**t-2**	**t-3**	**t-4**	**….**	**t-9**
		Per la stima del modello					
T_1	95						
T_2	92	95					
T_3	88	92	95				
T_4	88	88	92	95			
T_5	93	88	88	92	95		
T_6	97	93	88	88	92	….	
T_7	104	97	93	88	88	….	
T_8	110	104	97	93	88	….	
T_9	110	110	104	97	93	….	
T_{10}	120	110	110	104	97	….	95
T_{11}	110	120	110	110	104	….	92
T_{12}	105	110	120	110	110	….	88
T_{13}	90	105	110	120	110	….	88
T_{14}	85	90	105	110	120	….	93
T_{15}	88	85	90	105	110	….	97
T_{16}	101	88	85	90	105	….	104
T_{17}	107	101	88	85	90	….	110
T_{18}	112	107	101	88	85	….	110
T_{19}	103	112	107	101	88	….	120
T_{20}	97	103	112	107	101	….	110
T_{21}	92	97	103	112	107	….	105
T_{22}	89	92	97	103	112	….	90
T_{23}	85	89	92	97	103	….	85
T_{24}	91	85	89	92	97	….	88
T_{25}	96	91	85	89	92	….	101
T_{26}	99	96	91	85	89	….	107
T_{27}	103	99	96	91	85	….	112
T_{28}	107	103	99	96	91	….	103
T_{29}	104	107	103	99	96	….	97
T_{30}	100	104	107	103	99	….	92
T_{31}	97	100	104	107	103	….	89
T_{32}	93	97	100	104	107	….	85
T_{33}	90	93	97	100	104	….	91
T_{34}	86	90	93	97	100	….	96
T_{35}	92	86	90	93	97	….	99
T_{36}	96	92	86	90	93	….	103
Media	*97,64*						
Varianza	*76,35*						
Std. Dev.	*8,74*						

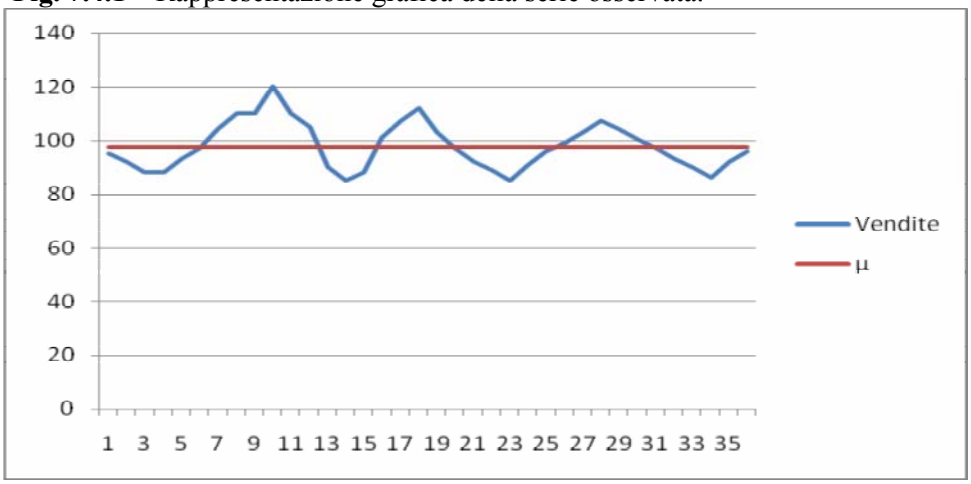
	Per la previsione
T_{37}	99
T_{38}	102
T_{39}	105
T_{40}	107

Fonte: elaborazione su dati dell'indagine empirica.

Dato che nella ricerca del modello che descrive la serie che riproduca il meccanismo che la genera è importante tener presente come individuare il processo generatore dei dati della serie, come verificarne la scelta e come utilizzare il risultato per la previsione, la sua applicazione richiede, allora, come presupposto che tra i dati della serie vi sia autocorrelazione. I coefficienti di correlazione totale e parziale devono essere calcolati sui dati della serie e sugli stessi dati slittati k volte.

Nel voler prevedere l'andamento delle vendite secondo il procedimento di Box-Jenkins è necessario partire con un esame preliminare dei dati della serie per studiarne le caratteristiche.

Fig. 7.4.1 – Rappresentazione grafica della serie osservata.



Fonte: elaborazione su dati dell'indagine empirica.

Per l'accertamento della stazionarietà in media e varianza della serie temporale considerata, esaminiamo la figura precedente (Fig. 7.4.1) dalla quale emerge che i dati oscillano attorno allo loro media ($\overline{Y} = 97{,}639$) e che l'ampiezza delle oscillazioni nel tempo tende a mantenersi stabile. L'analisi empirica sembrerebbe indicare che le oscillazioni sono di natura casuale; in ogni modo, ogni genere di valutazione deve trovare conferme obbiettive. A questo punto proprio per stabilire se davvero la serie è stazionaria in media è necessario calcolare i coefficienti di autocorrelazione e prima di questi la media, la varianza e l'autocovarianza che costituiscono i presupposti indispensabili. I calcoli da dover svolgere per ottenere il coefficiente di autocorrelazione globale di ordine 1, $\rho(1)$, sono indicati di seguito:

$$\hat{\mu} = \frac{1}{36}\sum_{t=1}^{36} Y_t = 97{,}639$$

La varianza è:

$$\hat{\gamma}_0 = s_Y^2 = \frac{1}{35}\sum_{1}^{36}(Y_t - \overline{Y})^2 = 76{,}352$$

L'autocovarianza di ordine 1 è:

$$\hat{\gamma}_1 = \frac{1}{35}\sum_{2}^{36}(Y_t - \overline{Y})(Y_{t-1} - \overline{Y}) = 68{,}464$$

Il coefficiente di autocorrelazione globale stimato di ordine 1 è pertanto:

$$\rho(1) = \frac{\hat{\gamma}_1}{\hat{\gamma}_0} = \frac{68{,}464}{76{,}352} = 0{,}8966$$

Il coefficiente di autorrelazione parziale di ordine 1 è, invece, dato da:

$$\phi_{1,1} = \frac{|Q_k|}{|P_k|} = \frac{4{,}14683}{4{,}6246} = 0{,}897$$

da cui nello specifico:

$$|Q_1| = \begin{vmatrix} 0{,}897 & 0{,}897 & 0{,}324 & -0{,}172 & -0{,}487 & -0{,}576 & -0{,}417 & -0{,}155 & 0{,}140 \\ 0{,}324 & 1 & 0{,}897 & 0{,}324 & -0{,}172 & -0{,}487 & -0{,}576 & -0{,}417 & -0{,}155 \\ -0{,}172 & 0{,}897 & 1 & 0{,}897 & 0{,}324 & -0{,}172 & -0{,}487 & -0{,}576 & -0{,}417 \\ -0{,}487 & 0{,}324 & 0{,}897 & 1 & 0{,}897 & 0{,}324 & -0{,}172 & -0{,}487 & -0{,}576 \\ -0{,}576 & -0{,}172 & 0{,}324 & 0{,}897 & 1 & 0{,}897 & 0{,}324 & -0{,}172 & -0{,}487 \\ -0{,}417 & -0{,}487 & -0{,}172 & 0{,}324 & 0{,}897 & 1 & 0{,}897 & 0{,}324 & -0{,}172 \\ -0{,}155 & -0{,}576 & -0{,}487 & -0{,}172 & 0{,}324 & 0{,}897 & 1 & 0{,}897 & 0{,}324 \\ 0{,}140 & -0{,}417 & -0{,}576 & -0{,}487 & -0{,}172 & 0{,}324 & 0{,}897 & 1 & 0{,}897 \\ 0{,}309 & -0{,}155 & -0{,}417 & -0{,}576 & -0{,}487 & -0{,}172 & 0{,}324 & 0{,}897 & 1 \end{vmatrix}$$

$$|P_9| = \begin{vmatrix} 1 & 0,897 & 0,324 & -0,172 & -0,487 & -0,576 & -0,417 & -0,155 & 0,140 \\ 0,897 & 1 & 0,897 & 0,324 & -0,172 & -0,487 & -0,576 & -0,417 & -0,155 \\ 0,324 & 0,897 & 1 & 0,897 & 0,324 & -0,172 & -0,487 & -0,576 & -0,417 \\ -0,172 & 0,324 & 0,897 & 1 & 0,897 & 0,324 & -0,172 & -0,487 & -0,576 \\ -0,487 & -0,172 & 0,324 & 0,897 & 1 & 0,897 & 0,324 & -0,172 & -0,487 \\ -0,576 & -0,487 & -0,172 & 0,324 & 0,897 & 1 & 0,897 & 0,324 & -0,172 \\ -0,417 & -0,576 & -0,487 & -0,172 & 0,324 & 0,897 & 1 & 0,897 & 0,324 \\ -0,155 & -0,417 & -0,576 & -0,487 & -0,172 & 0,324 & 0,897 & 1 & 0,897 \\ 0,140 & -0,154 & -0,417 & -0,576 & -0,487 & -0,172 & 0,3243 & 00,897 & 1 \end{vmatrix}$$

Poiché la serie storica in esame è formata da 36 osservazioni e dato che una regola pratica suggerisce di calcolare i coefficienti di correlazione totale e parziale in numero non superiore ad un quarto, sono stati calcolati 9 coefficienti di autocorrelazione totale e parziale. Per ciascun coefficiente calcolato è stato misurato il suo relativo livello di significatività; al fine di verificare quanti e quali siano risultati significativamente diversi da zero è stato, per ognuno di essi, calcolato il test ***t*,** che - confrontato con i valori soglia – ha permesso di mettere in risalto che solo i primi due coefficienti di autocorrelazione globale e solo il primo coefficiente di autocorrelazione parziale risultano essere significativamente diversi da zero al livello del 5 per cento (Tab. 7.4.2); esaminando ancora più nello specifico i coefficienti di autocorrelazione globale, si osserva ancora che i successivi valori si smorzano rapidamente verso lo zero.

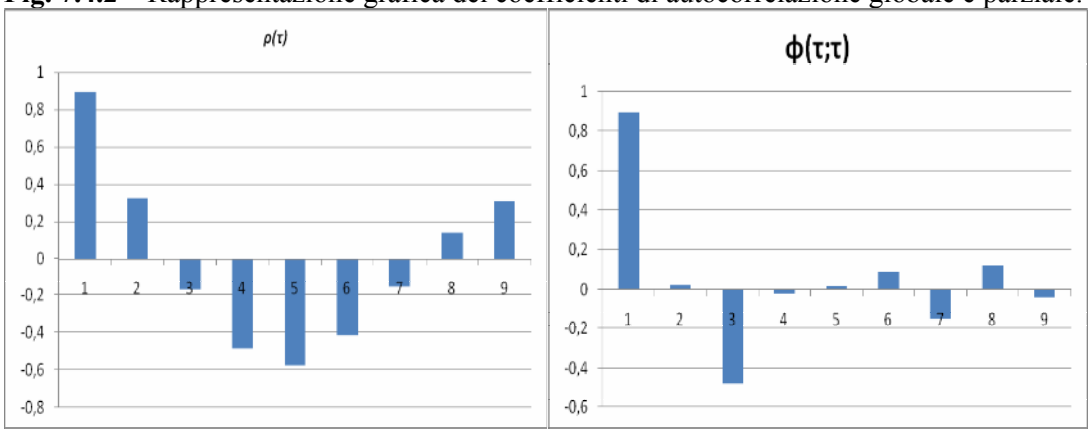
Nel caso specifico, la stazionarietà in media della serie storia è, allora, fornita oltre che dall'andamento della funzione di autocorrelazione globale che scende verso lo zero con molta rapidità, anche dal valore del primo coefficiente di autocorrelazione parziale che è in valore assoluto minore dell'unità ($\varphi_{1,1} = 0,896689$).

Tab. 7.4.2 - Coefficienti di autocorrelazione globale e parziali.

Lag	ρ_k	Test t	φ_{kk}	Test t	$t_{n-k-2,\alpha = 0,05}$
1	0,897	32,365	0,897	32,365	1,692
2	0,324	11,849	0,022	0,807	1,694
3	-0,172	-6,517	-0,477	-18,084	1,695
4	-0,487	-19,144	-0,017	-0,676	1,698
5	-0,576	-22,862	0,017	0,677	1,699
6	-0,417	-16,527	0,088	3,483	1,701
7	-0,155	-6,241	-0,151	-6,077	1,703
8	0,140	6,057	0,120	5,165	1,706
9	0,309	14,295	-0,039	-1,807	1,708

Fonte: elaborazione su dati dell'indagine empirica.

Fig. 7.4.2 – Rappresentazione grafica dei coefficienti di autocorrelazione globale e parziale.



Fonte: elaborazione su dati dell'indagine empirica.

Nel passare all'identificazione di un modello di previsione, si osserva che - prendendo in considerazione le successioni dei valori dei coefficienti di autocorrelazione globale, di autocorrelazione parziale ed i relativi correlogrammi teorici di processi lineari AR che sono di andamento noto – la scelta si indirizza verso un modello ARMA, ovvero un modello misto di tipo autoregressivo di ordine 2 (AR(2)) a media mobile di ordine 1 (MA(1)); la ragione è semplice. Si è voluto partire con un modello misto con una componente autoregressiva ed una a media mobile anche in considerazione del fatto che dopo il primo i coefficienti di autocorrelazione parziali decadono velocemente verso lo zero.

In sostanza, il modello stimato è del tipo:

$$Y_t = C + \eta_1 Y_{t-1} + \eta_2 Y_{t-2} + \varphi_1 W_{t-1} + a_t$$

Tab. 7.4.3 – Stima dei parametri ARMA (con p = 2 e q = 1).

Parametri	**Stima**	**Std. Error**	**Test t**	$t_{28;\alpha=0,025}$
η_1	1,495	0,095	15,695	2,048
η_2	-0,868	0,096	-9,013	
φ_1	-0,800	0,615	-1,302	
Costante	36,701	5,808	6,319	

Fonte: elaborazione su dati dell'indagine empirica.

Da cui si ha:

$$Y_t = 36,701 + 1,495 Y_{t-1} - 0,868 Y_{t-2} - 0,800 W_{t-1} + a_t$$

Prima di passare alla valutazione della capacità previsiva del modello stesso, è doveroso innanzitutto verificare il livello di significatività di ciascun coefficiente stimato sulla base del test t di Student; come si può osservare dalla precedente tabella

(Tab. 7.4.3), solo il coefficiente relativo alla componente a media mobile non risulta essere significativamente diverso da zero. Nello specifico, si ha:

$$t = \frac{\hat{\varphi} - 0}{s(\hat{\varphi})} = \frac{-0{,}800}{0{,}615} = -1{,}302 > t_{28;\alpha=0{,}025} = -2{,}048$$

Dal risultato ottenuto, si desume che i coefficienti di autocorrelazione parziale pur decadendo velocemente verso lo zero, la serie studiata presenta sostanzialmente un andamento stazionario in media, confermando ancor di più il fatto che le oscillazioni presenti sono sostanzialmente da attribuirsi al regolare andamento delle vendite stesse intorno al suo valore medio.

Poiché per la componente a media mobile l'ipotesi nulla non è stata rigettata, si può, a questo punto, ipotizzare di costruire un modello più semplice unicamente di tipo autoregressivo; pertanto, dall'eliminazione della componente a media mobile è stato implementato un secondo modello unicamente di tipo autoregressivo di ordine 2, quale:

$$Y_t = C + \eta_1 Y_{t-1} + \eta_2 Y_{t-2} + a_t$$

Tab. 7.4.4 – Stima dei parametri del modello AR(2).

Parametri	**Stima**	**Std. Error**	**Test t**	$f_{29;\alpha=0{,}025}$
η_1	1,489	0,094	15,761	
η_2	-0,867	0,095	-9,076	2,045
Costante	37,180	5,800	6,410	

Fonte: elaborazione su dati dell'indagine empirica.

Da cui:

$$Y_t = 37{,}180 + 1{,}489 Y_{t-1} - 0{,}867 Y_{t-2} + a_t$$

Dopo aver determinato il modello di regressione il passo conclusivo è quello di andare a misurare l'errore medio assoluto del modello (EMA)

$$EMA = \frac{100}{n} \sum_{i=1}^{34} \left| \frac{\hat{a}_i}{\hat{x}_i} \right| = \frac{100}{34} \times 0{,}026 = 0{,}075.$$

A tal proposito, si riscontra un elevato grado di accostamento dei dati al modello stimato.

A questo punto è doveroso valutare se i risultati ottenuti possono essere accettabili anche ai fini previsivi (Tab. 7.4.5).

Tab. 7.4.5 – Distribuzione della serie secondo i valori stimati da modello.

Tempo	*Valori osservati*	*Valori stimati*	*Errore*
T_1	95	-	-
T_2	92	-	-
T_3	88	91,849	-3,849
T_4	88	90,844	-2,844
T_5	93	91,927	1,073
T_6	97	98,780	-1,780
T_7	104	102,463	1,537
T_8	110	106,123	3,878
T_9	110	109,959	0,041
T_{10}	120	106,295	13,705
T_{11}	110	112,414	-2,414
T_{12}	105	104,062	0,938
T_{13}	90	92,860	-2,860
T_{14}	85	85,458	-0,458
T_{15}	88	82,405	5,595
T_{16}	101	93,683	7,317
T_{17}	107	107,878	-0,878
T_{18}	112	111,929	0,071
T_{19}	103	109,192	-6,192
T_{20}	97	98,364	-1,364
T_{21}	92	89,733	2,267
T_{22}	89	90,599	-1,599
T_{23}	85	90,579	-5,579
T_{24}	91	89,335	1,665
T_{25}	96	96,169	-0,169
T_{26}	99	103,323	-4,323
T_{27}	103	102,441	0,559
T_{28}	107	103,685	3,315
T_{29}	104	106,123	-2,123
T_{30}	100	101,326	-1,326
T_{31}	97	95,319	1,681
T_{32}	93	94,433	-1,433
T_{33}	90	92,790	-2,790
T_{34}	86	91,349	-5,349
T_{35}	92	89,706	2,294
T_{36}	96	91,553	4,447
T_{37}	***99***	***95,727***	***3,273***
T_{38}	***102***	***100,342***	***1,658***
T_{39}	***105***	***103,593***	***1,407***
T_{40}	***107***	***104,434***	***2,566***

Fonte: elaborazione su dati dell'indagine.

Avendo già in partenza escluso dal modello gli ultimi quattro dati della serie relativi al periodo compreso tra il tempo t_{37} e t_{40}, si possono stimare i dati relativi a questi ultimi quattro periodi e confrontarli con quelli effettivi posseduti. Sostituendo pertanto a $z_{(t-1)}$ il valore stimato al tempo t_{36} ed a $z_{(t-2)}$ il valore stimato al tempo t_{35}, si ottiene il valore stimato al tempo t_{37} che confrontato con quello effettivo permette di misurare l'errore stesso da modello. Procedendo in maniera analoga anche per i periodi successivi fino a t_{40}, si ottengono gli altri valori stimati e un errore medio assoluto molto basso, uguale a 0,065.

8. La previsione delle vendite mediante un modello di regressione multipla

Le vendite di una azienda dipendono da variabili endogene, come ad esempio i costi di pubblicità, e da variabili esogene, relative prevalentemente a indicatori di tipo socio-economico[11].

Se si dispone, ad esempio, di n variabili esplicative osservate in alcune aree territoriali (come la popolazione residente, il reddito disponibile, lo stato delle concorrenza, ecc.), le vendite possono essere rappresentate da un modello di regressione multipla. Per poter valutare l'andamento delle vendite sulla base della dinamica di talune variabili ritenute significative, si deve naturalmente assumere l'invarianza delle leggi che legano tra loro tali variabili.

Si osservi che, essendo tale analisi basata sulla valutazione del legame esistente tra le vendite e le variabili cosiddette esplicative, è possibile pervenire ad una valutazione previsiva per ognuna delle possibili tendenze delle variabili stesse. In altre parole, è possibile ipotizzare diverse dinamiche delle variabili indipendenti e stimare, per ognuna di queste ipotesi, il probabile andamento delle vendite. Da queste semplici considerazioni si evince l'importanza della scelta delle variabili da prendere in considerazione per la costruzione del modello.

L'analisi mediante un modello di regressione può essere schematizzata nelle seguenti fasi:

1) Scelta delle variabili indipendenti. Se si tratta di una sola variabile, si parla di regressione semplice, altrimenti di regressione multipla. Si tratta, intuitivamente, di una fase molto delicata ed importante, poiché dalla selezione di uno o più tra i fenomeni in esame dipendono in maniera diretta i risultati ottenuti. Da un lato si tende a prendere in considerazione molte variabili per ottenere risultati che tengano conto del maggior numero possibile di fenomeni, dall'altro è necessario salvaguardare innegabili necessità di semplicità di calcolo e di interpretazione.
2) Scelta del tipo di funzione. È infatti necessario selezionare tra i tanti possibili un solo tipo di relazione da adottare come legame tra la variabile dipendente (vendite) e le variabili indipendenti (esplicative) che si è deciso di adoperare per la

[11] Mazzali A., *Lezioni di Statistica aziendale*, op. cit..

valutazione, cercando di contemperare le esigenze opposte di un maggior grado di rappresentatività ottenibile dal modello con la maggiore semplicità possibile.

3) Calcolo dei parametri. Una volta scelta la funzione di regressione che meglio si adatta allo scopo, occorre procedere al calcolo dei parametri che compaiono nel modello, preferibilmente con il metodo dei minimi quadrati.

La scelta delle variabili è un passo fondamentale in quanto dovranno essere presi in considerazione solo fenomeni effettivamente dotati di un potere esplicativo dell'andamento delle vendite, ed inoltre tali fenomeni dovranno essere indipendenti fra loro (altrimenti si finisce per attribuire ad un fenomeno un'influenza che in realtà compete al fenomeno ad esso correlato).

La scelta delle variabili indipendenti è dunque la fase più delicata dell'intero processo di previsione e deriva molto spesso più dall'esperienza dell'analista che da considerazioni di altro tipo. Tali decisioni, infatti, dipendono fortemente dalla tipologia dei beni e/o servizi per i quali si vuole effettuare la previsione, per cui occorre prestare la massima attenzione alle caratteristiche di rappresentatività delle variabili stesse[12].

Il modello di regressione multipla consente di esprimere una generica osservazione come somma di due componenti: una deterministica, che esprime il comportamento del valore medio della variabile dipendente (Y), e un'altra di tipo accidentale, costituita da elementi che sfuggono alla osservazione e che prende il nome di errore casuale o disturbo. Da cui consegue:

$$y_i=\mu_i+\varepsilon_i \quad .$$

L'analisi della regressione multipla consiste nel determinare una funzione che esprima nella maniera migliore il legame esistente in media tra le variabili indipendenti (X_1, X_2,..... X_n) e la variabile dipendente Y.

Assumendo che il legame sia di tipo lineare, si ha:

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_n x_{in}+\varepsilon_i \quad ,$$

che è l'equazione dell'*iperpiano di regressione*[13]. Nel caso di due sole variabili indipendenti, si ha il *piano di regressione*:

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\varepsilon_i \quad .$$

Disponendo di n osservazioni delle variabili esplicative, i problemi che si pongono sono:

1) stima dei parametri incogniti β_0, β_1, ... , β_n e della varianza dell'errore ε_i;
2) definizione della bontà dell'accostamento del modello ai dati empirici;

[12] Naturalmente la bontà dei risultati va valutata sulla base di indici statistici, per cui si effettuano alcuni test di verifica: precisamente, si può ricorrere al test t di Student per la valutazione della significatività statistica delle variabili prescelte, al test F di Snedecor-Fisher per stabilire la validità esplicativa del modello, e al test di Durbin-Watson di autocorrelazione dei residui della regressione, in base al quale, in caso di mancata casualità dei residui, si conclude per l'esistenza di una variabile significativa che non è stata presa in considerazione ai fini della definizione del modello ma che provoca *perturbazioni correlate* nelle serie di dati. Cfr., ad es., Delvecchio F., *Statistica per la ricerca sociale*, Cacucci, 1995.

[13] Girone G. - Salvemini T., *Lezioni di statistica, vol. 2*, Cacucci, Bari, 1991.

3) utilizzo del modello per effettuare previsioni spaziali e temporali.

Per quanto riguarda la stima dei parametri incogniti β_0, β_1.... β_n, se le variabili sono grandezze note e non variabili casuali (e quindi gli errori ε_i hanno media nulla, varianza costante σ^2 e sono incorrelati, per cui $Cov(\varepsilon_i,\varepsilon_j)=0$ per $i \neq j$), il metodo dei minimi quadrati può essere utilizzato per stimare i parametri incogniti. Si assume dunque che le vendite del bene i-esimo dipendano dalle n variabili X_1, X_2.... X_n, in funzione dei parametri β_0, β_1, ..., β_n, e da una componente erratica ε_i.

Come è noto, il modello di regressione con il metodo dei minimi quadrati si basa sulla minimizzazione dei quadrati degli scarti tra i valori teorici e quelli empirici:

$$\sum \left(y_i^* - y_i\right)^2 = \text{minimo}$$

che si risolve eguagliando a zero le derivate parziali rispetto ai parametri incogniti. In simboli, ponendo:

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_n x_{in} \quad ,$$

si ha:

$$\Sigma(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_n x_{in} - y_i) = 0$$

che, derivando, conduce al seguente sistema di equazioni:

$$\begin{cases} n\beta_0 + \beta_1 \sum x_{i1} + \beta_2 \sum x_{i2} + ... + \beta_n \sum x_{in} = \sum y_i \\ \beta_0 \sum x_{i1} + \beta_1 \sum x_{i1}^{\ 2} + \beta_2 \sum x_{i1} x_{i2} + ... + \beta_n \sum x_{i1} x_{in} = \sum x_{i1} y_1 \\ \beta_0 \sum x_{i2} + \beta_1 \sum x_{i1} x_{i2} + \beta_2 \sum x_{i2}^{\ 2} + ... + \beta_n \sum x_{i2} x_{in} = \sum x_{i2} y_2 \\ ..... \\ \beta_0 \sum x_{in} + \beta_1 \sum x_{in} x_{i1} + \beta_2 \sum x_{in} x_{12} + ... + \beta_n \sum x_{in}^{\ 2} = \sum x_{in} y_2 \end{cases}$$

Il sistema può essere scritto anche in forma matriciale:

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}.$$

Poiché nella realtà i coefficienti β_i non sono generalmente noti, si procede ad una loro stima, a cui si perviene sostituendo ai β_i, nel precedente sistema di equazioni, i corrispondenti valori empirici b_i. Si ottiene così un vettore **b** (di ordine n+1) di stime dei coefficienti di regressione presenti nel modello, per i quali, con le tecniche di inferenza statistica, è possibile calcolare la variabilità e gli intervalli di confidenza, nonché il grado di accostamento della funzione trovata al fenomeno reale. L'espressione in forma matriciale del modello diventa dunque:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{E}$$

dove **E** è il vettore delle stime degli errori aleatori ε_i.

Per valutare dunque l'attitudine del modello a rappresentare la realtà rilevata, tramite la varianza dell'errore, si tratta di stimare la dispersione dei valori osservati

intorno al piano di regressione[14]. La devianza totale risulta dalla somma della devianza di regressione e della devianza dell'errore:

$$Dev(Y) = Dev(R) + Dev(E)$$

Il primo addendo esprime la misura in cui il modello riesce a spiegare il comportamento del fenomeno: quanto più Dev(R) si avvicina a Dev(Y), tanto maggiore è la validità esplicativa del modello e minore è la Dev(E).

$$Dev(Y) = \sum_i (Y_i - \overline{Y})^2 \qquad \text{(Dev. totale)}$$

$$Dev(R) = \sum_i (Y_i^* - \overline{Y})^2 \qquad \text{(Dev. di regressione)}$$

$$Dev(E) = \sum_i (Y_i - Y_i^*)^2 \qquad \text{(Dev. residua)}$$

Come misura del grado di accostamento, si definisce l'indice di determinazione:

$$R^2 = \frac{Dev(R)}{Dev(Y)}.$$

Ovviamente è $0 \leq R^2 \leq 1$, e la validità dell'accostamento è tanto migliore quanto più R^2 si avvicina a 1.

Ottenuto il modello, questo può essere utilizzato per la previsione delle vendite in un'altra regione.

8.1. Ulteriori riflessioni sul modello di regressione multipla

Se da una parte l'utilizzo della regressione multipla risulta particolarmente efficace ed al tempo stesso agevole, non è possibile un suo utilizzo indiscriminato senza una valutazione delle condizioni circa la sua applicazione nelle diverse situazioni concrete,oltre a interpretare correttamente i risultati al fine della previsione:

La scelta stessa in primis delle variabili da inserire nel modello è una fase particolarmente delicata per la messa a punto di un modello che spieghi la maggior parte della variabilità. A tal fine è possibile ricorrere ad esperienze precedenti, anche a supposizioni teoriche ma soprattutto occorre far riferimento al "contesto di studio" ovvero avere un conoscenza approfondita del contesto economico, sociale in cui ci si muove. In questa fase è importante avere una idea precisa in merito ai segni algebrici, che dovranno assumere i coefficienti della equazione di regressione; in altri termini, bisogna sapere sin dall' inizio se le singole variabili indipendenti influiscono sulle vendite in termini positivi o negativi. E' importante valutare la qualità dei dati statistici di base; ad esempio, è necessario verificare l'esistenza di eventuali errori di misura, soprattutto, per quanto riguarda la variabile dipendente. Se questa ad, esempio, è affetta

[14] Sadocchi S., *Manuale di analisi statistica multivariata*, Franco Angeli, Milano, 1980.

da un errore di misura importante, allora anche la migliore variabile indipendente non è in grado di raggiungere livelli accettabili di accuratezza previsiva.

Una delle ipotesi alla base della regressione multipla è la linearità della relazione che lega la variabile dipendente alle variabili indipendenti. Talvolta tale ipotesi poco si adatta ai dati originari per cui non si riesce a rappresentare gli effetti curvilinei. In questi casi si ricorre ad espressioni del modello polinomiali oppure a variabili di comodo (dummy) che possano integrare le informazioni a nostra disposizione; queste variabili assumono due soli valori (in genere 0 e 1) a significare rispettivamente assenza e presenza di una data caratteristica qualitativa.

Nel procedere alla stima dei coefficienti dell'equazione di regressione specificata e poi alla valutazione della capacità previsiva delle variabili indipendenti, devono essere calcolati opportuni test statistici. In questa fase le operazioni fondamentali sono la scelta del metodo per la stima dei coefficienti di regressione, la valutazione della significatività statistica dell'intera equazione e dei singoli coefficienti e l'individuazione di osservazioni che possono influenzare in misura eccessiva i risultati dell'analisi (i cosiddetti *outliers*).

Il metodo di stima utilizzato più frequentemente è quello dei minimi quadrati sia per le sue proprietà statistiche e poi per la possibilità di applicarlo indipendentemente dalla forma della distribuzione della variabile dipendente ; questo metodo in sostanza si configura come caso particolare del metodo della massima verosimiglianza di R. Fisher. Questo metodo consiste nell'assumere come stima di un parametro incognito la corrispondente statistica campionaria che ha la massima probabilità di verificarsi[15].

Stimata l'equazione è necessario controllare le assunzioni e le condizioni che sono alla base del modello di regressione. Si tratta di ipotesi che influenzano tutta la procedura interessando sia le singole variabili sia la relazione nel suo insieme che è necessario rispettare.

Questi test poggiano essenzialmente sui residui della regressione che abbiamo indicato con e_i, che misurano lo scarto fra valori empirici e valori teorici (Y_i - Y*). Questi costituiscono gli errori di previsione.

Occorre valutare se i residui sono di entità trascurabile e se la loro somma algebrica sia nulla e, di conseguenza, media nulla; il che lascerebbe supporre la loro accidentalità. Se ciò non verifica i residui possono indicare che la relazione che è stata prevista fra le variabili in realtà non c'è, oppure possono trarre origine da caratteristiche dei dati che l'equazione utilizzata non riesce a cogliere .Inoltre i residui devono risultare omoschedastici (nel senso che tutti gli errori devono avere la stessa varianza), indipendenti e distribuiti normalmente.

Nel considerare più da vicino le ipotesi richiamate iniziando dalla linearità della relazione tra variabile dipendente e variabili indipendenti. La linearità rappresenta il grado con cui le variazione nella variabile dipendente sono associate a variazioni nelle variabili indipendenti ed è misurabile in due modi

a) considerando di quanto varia in media la variabile dipendente per ogni unità di variazione delle variabili indipendenti;
b) osservando se il valore medio della variabile dipendente giace sulla retta di regressione nel caso della regressione lineare semplice o sull'iperpiano di regressione se si tratta di una relazione multivariata.

[15] Leti G. (1983), *Statistica descrittiva*, Bologna, Il Mulino.

L'eteroscedasticità - contrario della omoscedasticità - è la presenza di varianze diseguali nei residui ed è una delle più frequenti violazioni alle assunzioni di base. Il suo controllo è effettuato con i grafici dei residui o con test statistici. In presenza di eteroscedasticità sono possibili due soluzioni:

- si può ricorrere in modo semplice e diretto a trasformazioni dei dati che riducono la variabilità e tendono a stabilizzare la varianza ; ad esempio si segnala la trasformazione logaritmica.
- se la violazione attribuibile ad una sola variabile indipendente, per stimare i coefficienti si può impiegare la procedura dei minimi quadrati ponderati, che equivale a trasformare le variabili dividendole per le rispettive varianze in modo da ottenere un modello con i residui omoscedastici.

Indipendenza dei residui significa che i valori teorici sono tra loro indipendenti, nel senso che non risultano ordinati in base ad alcun criterio. Per identificare ciò si può effettuare una rappresentazione grafica dei residui e osservare l'andamento degli stessi ; se sono indipendenti deve verificarsi una alternanza tra valori positivi e valori negativi. Se ciò non si verifica è necessario ricorrere a trasformazioni dei dati : ad esempio operando sulle differenze prime dei dati osservati.

La normalità della distribuzione dei residui viene verificata ricorrendo alla distribuzione grafica controllando la loro approssimazione alla curva gaussiana.

I test diagnostici cui si ricorre per le operazioni di controllo sono il test di Snedecor sulla varianza spiegata dalla relazione lineare e il test t di Student relativo ai singoli coefficienti.

Per controllare l'ipotesi che la variabilità spiegata dal modello di regressione sia maggiore della variabilità spiegata dalla media (che rappresenta la prima forma di previsione) o in altri termini che R^2 sia maggiore di zero si fa ricorso al test F di Snedecor. Il test si calcola nel modo seguente

$$F = \frac{\frac{dev(rgr)}{df(rgr)}}{\frac{dev(res)}{df(res)}}$$

in cui:

df(rgr): indica i gradi di libertà (cioè il numero delle osservazioni non vincolate) della devianza di regressione pari al numero dei coefficienti stimati (inclusa l'intercetta) meno uno;

df(res): indica i gradi di libertà della devianza residua, pari alla dimensione campionaria (n) meno il numero dei coefficienti stimati (inclusa l'intercetta).

Per quanto riguarda la F di Snedecor è bene mettere in evidenza due aspetti importanti:

a) la devianza residua divisa per il numero dei gradi di libertà al denominatore del rapporto è semplicemente la varianza degli errori di previsione

b) dal punto di vista intuitivo se il rapporto tra la devianza spiegata dal modello e quella residua è alto significa che il modello è efficace nello spiegare la relazione tra la variabile dipendente (le vendite) e le variabili indipendenti.

In merito al coefficiente di determinazione R^2 occorre precisare che esso è influenzato dal gioco combinato del numero di variabili indipendenti inserite nel modello e dal numero di osservazioni. L'esperienza consiglia di inserire nel modello un numero di osservazioni compreso fra 10 e 15 per ogni variabile indipendente. Oltre tali limiti occorre correggere R^2 per controllare la crescita dovuta a una sorta di sovra adattamento ai dati . La formula è la seguente

$$\overline{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-p} = 1 - \frac{dev(res)}{dev(rgr)} \cdot \frac{n-1}{n-p}$$

in cui p è il numero dei coefficienti stimati inclusa l'intercetta.

E' altrettanto importante valutare la significatività dei coefficienti di regressione stimati. La variabilità nella stima dei coefficienti dell'equazione (sia della costante che dei coefficienti di regressione) viene definita errore standard dei coefficienti e viene indicata con $\hat{\sigma}_\beta$.

Per valutare la significatività dei coefficienti di regressione si ricorre al test t di Student dato dal rapporto tra il coefficiente di regressione stimato e il suo errore standard

$$t = \frac{b}{\hat{\sigma}_\beta}$$

in cui $\hat{\sigma}_B$ è dato da:

$$\hat{\sigma}_\beta = \frac{\sum (Y_i - \overline{Y}_i^*)^2 / (n-2)}{\sum (X_i - \overline{X}_i)^2}$$

da cui, per valori significativamente maggiori dei relativi valori teorici, si tenderà a rifiutare l'ipotesi nulla H_0: $\beta_i = 0$; in tal caso, allora, esiste una significativa relazione di dipendenza tra le vendite ed il prezzo.

Altro problema è il valore degli *outliers*. Non sempre si tratta di valori "cattivi" e quindi devono essere trattati con la dovuta cautela (Esempio dei dati di bilancio e relativi indici) Vedere metodi di stima per contenere l'impatto degli outliers sui risultati dell'analisi.

L'interpretazione del modello di regressione e la generalizzazione dei risultati viene effettuata valutando i coefficienti di regressione in base al contributo fornito nella spiegazione della variabile dipendente. Questi coefficienti ci consentono di effettuare le previsioni delle vendite per ogni unità di variazione delle variabili dipendenti. Per raggiungere tale scopo bisogna utilizzare i coefficienti b, che vengono stimati non su dati originari ma su dati standardizzati in modo da eliminare l'unità di misura e leggere

tutti i coefficienti secondo una unità di misura comune. Per la standardizzazione si effettuano per tutte le variabili (dipendente e indipendenti) gli scarti dalla media e si divide per lo scarto quadratico medio. In questo modo tutte le variabili hanno media nulla e varianza unitaria.

Altro problema fondamentale può essere la presenza di multicollinearità cioè la correlazione tra le variabili indipendenti. La situazione ideale è che vi sia la massima correlazione tra la variabile dipendente (nel nostro caso le vendite) e le variabili indipendenti ma poco correlate tra loro. Si pone pertanto il problema di: valutare il grado di multicollinearità e determinare l'impatto sui risultati e, se necessario trovare i rimedi opportuni.

Per quanto riguarda il primo aspetto si tratta di esaminare la matrice di correlazione delle variabili indipendenti. L'eventuale presenza di coefficienti di correlazione superiore a 0,90 è indice collinearità. Ma occorre anche considerare che coefficienti di correlazione non elevati non assicurano che non ci sia multicollinearità: essa può dipendere anche dall'effetto combinato delle variabili indipendenti. Per valutare la multicollinearità si ricorre al fattore di accrescimento della varianza; detto coefficiente è denominato ***FIV*** *(Variance Inflaction Factor)****;*** per una generica variabile X_i si ha:

$$FIV = \frac{1}{1 - R_i^2}$$

in cui R_i^2 è il coefficiente di determinazione lineare multipla nella regressione della variabile esplicativa X_i su tutte le altre variabili indipendenti (con tre variabili indipendenti R_i^2 è il coefficiente di X_1 su X_2 e X_3). Quando le variabili sono incorrelate l'indice FIV è uguale ad 1; se sono altamente correlate potrebbe risultare superiore a 10. La soglia prudenziale oltre la quale ci si trova in presenza di multicollinearità è il valore 5.

Infine nel caso in cui sono stati utilizzati dati campionari si pone il problema di ricondurre i risultati ottenuti alla popolazione. In questi casi si confrontano le stime ottenute con un modello teorico o con precedenti esperienze.

In ultimo, per confrontare più modelli di regressione è necessario esaminare il valore di R^2 che aumenta al crescere delle variabili inserite nel modello; di conseguenza, per confrontare modelli con un numero di variabili diverso è necessario utilizzare R^2 corretto che tiene conto sia delle possibili differenze nelle numerosità campionarie sia nel numero delle variabili indipendenti.

8.2. Il modello di regressione: un'esemplificazione empirica

Nel presente paragrafo si vuole presentare un'esemplificazione empirica di un modello di regressione, ripercorrendo quanto già esplicitato dal punto di vista prettamente teorico.

Supponiamo, allora, di voler misurare la previsione delle vendite di un determinato prodotto di largo consumo in riferimento ad alcune determinate variabili esplicative, quali:

- il prezzo unitario (X_1),

- il reddito medio mensile delle famiglie (X_2),
- la spese per pubblicità (X_3),
- la qualità percepita del prodotto (X_4),
- l'immagine dell'azienda (X_5).

I dati di base sono stati riportati nella tabella seguente (Tab. 8.2.1); nel caso specifico, si è ipotizzato che, di fatto, l'ammontare delle vendite (Y) possano essere dipese nel periodo di osservazione da variabili indipendenti, come: il prezzo unitario, il reddito medio delle famiglie, le spese sostenute dall'azienda in pubblicità, la qualità percepita del prodotto e l'immagine dell'azienda. Prima di procedere le variabili – dipendente ed esplicative – sono state standardizzate al fine di eliminare l'influenza derivante dalle diverse scale di misura.

Tab. 8.2.1 – Andamento delle vendite in riferimento al periodo di osservazione.

Tempo	**Variabile dipendente**		**Variabili esplicative**									
	Y = *Vendite*		*X_1 =* *Prezzo*		*X_2 =* *Reddito*		*X_3 =* *Pubblicità*		*X_4 =* *Qualità prodotto*		*X_5 =* *Immagine*	
	V. assoluti	V. stand	V. assoluti	V. stand	V. assoluti	V. stand	V. assoluti	V. stand	V. assoluti	V. stand	V. assoluti	V. stand
T_1	40	-2,49	14	-1,67	2,00	-2,45	1,00	-1,41	4,00	0,31	5	-1,53
T_2	60	-1,97	16	-1,46	2,50	-2,23	2,00	-0,94	2,00	-1,30	8	-2,63
T_3	60	-1,97	16	-1,46	3,00	-2,02	2,00	-0,94	3,00	-0,49	8	-2,08
T_4	70	-1,71	17	-1,36	3,50	-1,80	1,00	-1,41	2,00	-1,30	8	-2,63
T_5	80	-1,46	18	-1,26	4,00	-1,59	3,00	-0,46	6,00	1,92	4	-0,44
T_6	70	-1,71	17	-1,36	4,50	-1,37	2,00	-0,94	2,00	-1,30	10	-2,63
T_7	80	-1,46	17	-1,36	3,20	-1,93	1,00	-1,41	5,00	1,12	8	-0,98
T_8	110	-0,69	25	-0,54	3,00	-2,02	2,00	-0,94	4,00	0,31	7	-1,53
T_9	110	-0,69	23	-0,74	4,20	-1,50	3,00	-0,46	2,00	-1,30	8	-2,63
T_{10}	120	-0,43	22	-0,85	4,10	-1,54	4,00	0,01	4,00	0,31	9	-1,53
T_{11}	114	-0,58	25	-0,54	2,70	-2,15	1,00	-1,41	3,00	-0,49	6	-2,08
T_{12}	116	-0,53	26	-0,44	3,90	-1,63	5,00	0,49	4,00	0,31	8	-1,53
T_{13}	157	0,52	27	-0,33	3,00	-2,02	2,00	-0,94	5,00	1,12	7	-0,98
T_{14}	138	0,03	28	-0,23	4,10	-1,54	5,00	0,49	3,00	-0,49	7	-2,08
T_{15}	132	-0,12	20	-1,05	5,10	-1,11	4,00	0,01	3,00	-0,49	7	-2,08
T_{16}	122	-0,38	29	-0,13	2,10	-2,41	2,00	-0,94	5,00	1,12	5	-0,98
T_{17}	145	0,21	27	-0,33	3,70	-1,72	4,00	0,01	5,00	1,12	6	-0,98
T_{18}	155	0,47	28	-0,23	3,90	-1,63	6,00	0,96	5,00	1,12	6	-0,98
T_{19}	144	0,19	30	-0,03	2,90	-2,06	1,00	-1,41	3,00	-0,49	7	-2,08
T_{20}	135	-0,04	4	-2,69	3,40	-1,85	3,00	-0,46	3,00	-0,49	7	-2,08
T_{21}	140	0,08	32	0,18	4,50	-1,37	5,00	0,49	3,00	-0,49	6	-2,08
T_{22}	156	0,50	37	0,69	3,70	-1,72	5,00	0,49	2,00	-1,30	5	-2,63
T_{23}	134	-0,07	35	0,49	3,60	-1,76	6,00	0,96	5,00	1,12	8	-0,98

Continua

Segue

T_{24}	130	-0,17	34	0,38	3,80	-1,67	8,00	1,91	5,00	1,12	7	-0,98
T_{25}	125	-0,30	35	0,49	4,50	-1,37	1,00	-1,41	2,00	-1,30	4	-2,63
t_{26}	150	0,34	34	0,38	3,80	-1,67	2,00	-0,94	3,00	-0,49	5	-2,08
T_{27}	140	0,08	32	0,18	2,90	-2,06	3,00	-0,46	5,00	1,12	7	-0,98
T_{28}	133	-0,10	33	0,28	3,30	-1,89	4,00	0,01	5,00	1,12	8	-0,98
T_{29}	165	0,73	37	0,69	3,50	-1,80	6,00	0,96	4,00	0,31	10	-1,53
T_{30}	150	0,34	31	0,08	2,80	-2,11	3,00	-0,46	2,00	-1,30	8	-2,63
T_{31}	156	0,50	38	0,80	2,70	-2,15	7,00	1,43	3,00	-0,49	4	-2,08
T_{32}	160	0,60	37	0,69	3,80	-1,67	8,00	1,91	5,00	1,12	9	-0,98
T_{33}	159	0,57	36	0,59	4,10	-1,54	5,00	0,49	5,00	1,12	4	-0,98
T_{34}	145	0,21	35	0,49	4,00	-1,59	7,00	1,43	4,00	0,31	4	-1,53
T_{35}	156	0,50	37	0,69	3,90	-1,63	3,00	-0,46	2,00	-1,30	9	-2,63
T_{36}	170	0,86	38	0,80	4,20	-1,50	5,00	0,49	3,00	-0,49	4	-2,08

Fonte: elaborazione su dati dell'indagine empirica.

Prima di passare alla costruzione di un modello di regressione è fondamentale condurre un'analisi esplorativa sui dati; da un lato è importante conoscere le caratteristiche descrittive di tutte le variabili prese in esame e dall'altro - proprio attraverso un'analisi del grado di associazione esistente tra i diversi caratteri - è importante anche stabilire in che misura le diverse variabili selezionate influenzino la domanda del bene osservato.

Tab. 8.2.2 – Analisi esplorativa.

Analisi esplorativa	**Variabile dipendente**		*Variabili esplicative*									
	Y = Vendite		*X_1 = Prezzo*		*X_2 = Reddito*		*X_3 = Pubblicità*		*X_4 = Qualità prodotto*		*X_5 = Immagine*	
	V. assoluti	V. stand	V. assoluti	V. stand	V. assoluti	V. stand	V. assoluti	V. stand	V. assoluti	V. stand	V. assoluti	V. stand
Media	136,705	1,00	30,25	1,00	3,63	1,00	3,98	1,00	3,61	1,00	6,79	1,00
Dev. Sta	38,9105	0,00	9,74	0,00	0,68	0,00	2,11	0,00	1,24	0,00	1,82	0,00
Variabilità	1514,03	1	94,93	1,01	0,46	0,08	4,44	1	1,55	1	3,33	0,46
Min.	40	-2,48	4	-2,69	2	-2,45	1	-1,41	1	-2,10	4	-3,17
Max.	198	1,57	46	1,61	5,1	-1,11	8	1,91	6	1,92	10	-0,43
Asimm.	-0,6807	-0,68	-0,52	-0,52	-0,47	-0,47	0,28	0,28	-0,12	-0,12	-0,01	-0,12
1° quart.	119	-0,45	24,5	-0,59	3,15	-1,95	2	-0,94	3	-0,49	5	-2,08
Mediana	142	0,136	32	0,18	3,8	-1,67	4	0,01	4	0,31	7	-1,53
3° quartile	159,25	0,57	37	0,69	4,1	-1,54	5,25	0,60	5	1,11	8	-0,98

Fonte: elaborazione su dati dell'indagine empirica.

Nel tentativo di stabilire se e in che misura le variabili esplicative selezionate influenzino le vendite occorre calcolare la matrice di correlazione, che riassume per riga e colonna il grado di correlazione esistente (attraverso il coefficiente di correlazione di Bravais fra le variabili in gioco). Dalla seguente tabella (Tab. 8.2.3), si evidenzia:

- *che i segni algebrici risultano coerenti con le attese:* le vendite sono correlate positivamente con il prezzo ($r_{y,x1}$) ed in misura minore con la pubblicità ($r_{y,x3)}$) ed il reddito medio mensile delle famiglie ($r_{y,x2}$). Analizzando le variabili esplicative si osserva – sia pure in misura debole – un'associazione negativa tra la qualità del prodotto ed il reddito ($r_{x4,x2}$);
- *che le variabili qualità ed immagine dell'azienda risultano essere sostanzialmente incorrelate sia con le vendite che con il prezzo del prodotto, mentre risulta essere lievemente correlata con la pubblicità; dai risultati è evidente una strettissima correlazione, invece, tra le due variabili qualità ed immagine dell'azienda.* Sulla base dei risultati ottenuti e considerato l'ortogonalità tra quest'ultime e le prime variabili esplicative prese in esame, ipotizzare di trovarsi in presenza di multicollinearità e di ridondanza dei dati.

Tab. 8.2.3 – Matrice di correlazione

	Vendite (Y)	*Prezzo (X_1)*	*Reddito (X_2)*	*Pubblicità (X_3)*	*Qualità (X_4)*	*Immagine (X_5)*
Vendite	1,000					
Prezzo	0,846	1,000				
Reddito	0,348	0,265	1,000			
Pubblicità	0,597	0,586	0,374	1,000		
Qualità	0,048	0,034	-0,112	0,325	1,000	
Immagine	0,048	0,035	-0,113	0,325	0,910	1,000

Fonte: elaborazione su dati dell'indagine empirica.

Esaminando ancora la matrice di correlazione, si osserva che la variabile maggiormente correlata con le vendite è il prezzo (0,846), seguono in misura minore la pubblicità (0,597) e il reddito (0,348). Per nulla correlate risultano essere la qualità e l'immagine.

Per la determinazione dei coefficienti di regressione, si può ricorrere al metodo di adattamento dei minimi quadrati che garantisce – nel caso di un modello lineare – le stime non distorte e più efficienti, anche in assenza di ipotesi sulla distribuzione degli errori. Prima di prendere in considerazione le stime dei coefficienti è necessario valutare se il modello è adeguato a descrivere – ed eventualmente ad interpretare e prevedere – il fenomeno in esame: le vendite del prodotto. La verifica si basa, allora, sulla coerenza delle proprietà statistiche degli errori, con riferimento al principio che - se non ci sono fattori rilevanti esclusi - il modello selezionato può essere considerato soddisfacente; considerato che gli errori non sono, però, osservabili, vanno calcolati successivamente alla stima del modello come residui per ogni osservazione, sottraendo da ciascun valore osservato di Y quello teorico Y^*.

Nella presente applicazione empirica si è posto anche il problema di determinare il migliore modello di regressione capace di spiegare il fenomeno in esame; in sostanza si è posto come problema quello di scegliere come modello di regressione, proprio quel

modello che meglio riesca ad adattarsi alla spezzata di regressione. Per raggiungere tale scopo, la metodologia seguita è stata quella a ritroso (*backword*): inizialmente le variabili esplicative sono state inserite tutte nel modello, poi – passo dopo passo e ad una ad una - le variabili esplicative sono state eliminate, secondo la capacità di spiegare la variabilità del fenomeno (si parte dall'eliminazione di quelle variabili con la minore capacità). Nel caso empirico in esame, i passi eseguiti sono stati quattro; l'ultimo modello determinato presenta solo il prezzo come variabile esplicativa (Tab. 8.2.4). A questo punto, si è posto il problema di determinare il modello di regressione che risulta meglio adattato alla spezzata di regressione (ipotesi H_0); per raggiungere tale scopo è stato considerato il seguente rapporto F_1, che si distribuisce secondo una F di Snedecor, con $g_1 = s - 2$ e $g_2 = n - s$ gradi di libertà.

$$F_1 = \frac{\dfrac{\sum_i \left(\bar{y}_i - \hat{y}_i \right)}{s-2}}{\dfrac{Dev\,(e)}{n-s}}$$

Fissato il livello di α, si accetta l'ipotesi nulla se il valore di F_1 risulta essere inferiore al valore soglia; in tal modo, il rapporto F_1 può essere considerato come un test di attendibilità del grado di adattamento della retta di regressione alla spezzata di regressione; cioè, F_1 può essere considerato come un test per la misura della relazione di linearità tra la variabile risposta e quelle esplicative inserite nel modello.

Tab. 8.2.4 – Modelli di regressione.

Modello	Variabili esplicative	R	R^2	R^2 corretto	Errore standard	Attendibilità		
						F_1	gdl_1	gdl_2
1	Immagine Prezzo Reddito Pubblicità	0,785	0,616	0,566	0,577	12,408	4	31
2	Prezzo Reddito Pubblicità	0,784	0,615	0,579	0,568	0,016	1	31
3	Prezzo Pubblicità	0,779	0,606	0,582	0,566	0,771	1	32
4	Prezzo	0,757	0,574	0,561	0,580	2,715	1	33

Fonte: elaborazione su dati empirici.

Esaminando i quattro modelli così ottenuti, è stato scelto il secondo proprio per il suo maggiore grado di attendibilità rispetto agli altri, presentando un valore di $F_1 = 0{,}016 < f_{1,31,\alpha=0,05} = 4{,}16$; nonostante anche il terzo modello presenti un valore di F_1 inferiore al valore soglia si è preferito considerare il secondo modello proprio per la sua maggiore capacità di spiegare la variabilità del fenomeno ($R^2 = 0{,}615$). Esaminando ancora più in dettaglio il fenomeno in esame, è evidente che - anche se la funzione di

regressione scelta presenta un buon grado di adattamento e consente di asserire che esiste una relazione di linearità tra la funzione e la spezzata di regressione – non è detto che proprio tale funzione spieghi bene il legame di dipendenza della variabile vendite con le variabili esplicative inserite nel modello stesso (prezzo, reddito, pubblicità).

Tab. 8.2.5 – Analisi della variabilità dei modelli di regressione.

Modello	Devianza	SQ	gdl	MQ	F_2
1	Regressione	16,517	4	4,129	12,408
	Residua	10,316	31	,333	
	Totale	26,834	35		
2	Regressione	16,512	3	5,504	17,063
	Residua	10,322	32	,323	
	Totale	26,834	35		
3	Regressione	16,263	2	8,132	25,386
	Residua	10,571	33	,320	
	Totale	26,834	35		
4	Regressione	15,393	1	15,393	45,748
	Residua	11,440	34	,336	
	Totale	26,834	35		

Fonte: elaborazione su dati empirici.

Dopo quanto detto, risulta, allora, necessario andare a verificare che all'interno del modello scelto la variabile vendite risulti effettivamente dipendente dalle variabili prezzo, reddito e pubblicità; posto allora come ipotesi nulla H_0 la condizione che i coefficienti β siano tutti uguali a zero e considerato il seguente rapporto F_2,

$$F_2 = \frac{\frac{Dev(R)}{1}}{\frac{Dev(e)}{n-s}}$$

che si distribuisce come una F di Snedecor con $g_1 = 1$ e $g_2 = n - s$ gradi di libertà e fissato il livello di $\alpha = 0,05$, si rifiuterà l'ipotesi nulla se risulta $F_2 < f_{g1,g2,\alpha}$. Dell'esame dei risultati ottenuti dal modello selezionato (Tab. 8.2.5), si osserva che il valore $F_2 = 17,063$ risulta essere maggiore dello stesso valore soglia ($f_{3,32,\ \alpha=0,05} = 2,90$), pertanto si rifiuta l'ipotesi nulla supponendo verosimilmente che tale modello lineare di regressione sia idoneo a rappresentare il fenomeno in esame.

Dopo aver ampiamente esaminato il modello lineare di regressione selezionato, è doveroso effettuare una digressione in riferimento all'indice di determinazione R^2 che, variabile da zero a uno, misura la quota di variabilità della variabile risposta spiegata dalla relazione stimata:

$$R^2 = \frac{dev(regr)}{dev(tot)} = 1 - \frac{dev(res)}{dev(tot)}.$$

Facendo riferimento al modello scelto, si evince che il 61,6 % della variabilità del fenomeno – vendita del prodotto – è spiegata dal modello e, quindi, di conseguenza dalle variabili esplicative considerate (prezzo, reddito e pubblicità). Considerato che nella presente applicazione la tecnica impiegata è stata quella a ritroso - dove le variabili meno correlate e significative nell'esprimere la variabilità del fenomeno sono state ad una ad una escluse dal modello – l'analisi dei valori assunti dall'indice di determinazione permettono di osservare proprio una diminuzione degli stessi, passando dal primo modello al quarto (in sostanza, tale coefficiente aumenta all'aumentare del numero delle variabili esplicative inserite nel modello). A titolo esemplificativo e per meglio spiegare la composizione della variabilità del fenomeno, si osserva che il 61,6 % della variabilità è spiegata dal modello di regressione (R^2), mentre il 38,4 % della restante variabilità è legata alla casualità del fenomeno stesso.

$$dev\ (regr\ ) = \sum_i (Y_i^* - \overline{Y})^2 \quad \text{devianza di regressione} = 16{,}512$$

$$dev\ (res\ ) = \sum_i (Y_i - Y_i^*)^2 \quad \text{devianza residua} = 10{,}322$$

$$dev\ (tot\ ) = \sum_i (Y_i - \overline{Y})^2 \quad \text{devianza totale} = 26{,}834$$

In tale modo, il modello di regressione così ottenuto presenta la seguente formulazione matematica:

$$y = -0{,}643x_1 + 0{,}295x_2 + 0{,}157x_3 + 0{,}448$$

dove:

y = vendite,
x_1 = prezzo,
x_2 = reddito,
x_3 = pubblicità.

I parametri b_0, b_1, b_2 e b_3 sono stati calcolati con il metodo dei minimi quadrati.

Tab. 8.2.6 – Coefficienti di regressione.

Modelli		Coefficienti		Stat. t		Intervallo di confidenza 95%		Collinearità	
		b	*Std. Error*	*T*	*Sig.*	*Inferiore*	*Superiore*	*Tolleranza*	*VIF*
Modelli di regressione									
1	Constante	0,505	0,758	0,667	0,510	-1,041	2,051	-	-
	Prezzo	- 0,645	0,135	4,791	0,000	0,370	0,919	0,687	1,455
	Reddito	0,308	0,355	0,866	0,393	- 0,417	1,033	0,806	1,240
	Pubblicità	0,151	0,135	1,117	0,272	- 0,124	0,425	0,536	1,867
	Immagine	0,020	0,160	0,128	0,899	- 0,305	0,346	0,828	1,208
2	Constante	0,448	0,602	0,744	0,462	- 0,778	1,675	-	-
	Prezzo	- 0,643	0,132	4,878	0,000	0,375	0,912	0,694	1,441
	Reddito	0,295	0,336	0,878	0,386	- 0,390	0,980	0,873	1,145
	Pubblicità	0,157	0,123	1,279	0,210	- 0,093	0,407	0,625	1,600
3	Constante	- 0,073	0,099	- 0,738	0,466	- 0,275	0,129	-	-
	Prezzo	- 0,639	0,131	4,869	0,000	0,372	0,906	0,695	1,439
	Pubblicità	0,191	0,116	1,648	0,109	- 0,045	0,427	0,695	1,439
4	Constante	- 0,068	0,102	- 0,664	0,511	- 0,274	0,139	-	-
	Prezzo	- 0,759	0,112	6,764	0,000	0,531	0,987	1,000	1,000
Variabili escluse dall'analisi									
	Qualità	46,186	-	1,049	0,303	-	-	0,000	15,608

Fonte: elaborazione su dati empirici.

L'equazione ottenuta consente di effettuare una prima previsione. I coefficienti stimati devono essere letti come una misura della variazione delle vendite, in relazione alla variazione relativa delle variabili indipendenti inserite nel modello stesso; ad esempio, all'aumentare dell'1 % dei prezzi, le vendite diminuiscono dello 0,64 %, mentre all'aumentare dell'1 % del reddito o delle spese pubblicitarie, le vendite aumentano rispettivamente dello 0,29 % e dello 0,16 %. L'intercetta, invece, misura il livello della variabile dipendente quando le variabili indipendenti assumono valori nulli e, quindi, pur in assenza di variazioni nei prezzi e di condizioni esterne, le vendite ammonterebbero comunque al 44,8 % del totale (come si evincerebbe dal modello osservato).

A questo punto, stimati i parametri, è necessario andare a fare inferenza sugli stessi parametri della retta di regressione e costruire intervalli di confidenza che con probabilità 1 - α contengano tali parametri; allora, in base al test t di Student, i coefficienti di regressione delle variabili indipendenti inserite nel modello risultano essere significativamente diverse da zero ad un livello di significatività del 5%. Risultati ottenuti facendo ricorso, allora, al test t di Student dato da:

$$t = \frac{b}{\hat{\sigma}_B}$$

Rispetto a quanto appena affermato, diversamente risulta per quanto attiene l'analisi dell'unica variabile esclusa dall'analisi stessa – la variabile qualità di prodotto

– dove il risultato della statistica t di Student - uguale a 1,049 – è inferiore allo stesso valore soglia di riferimento (t_{31}= 1,69); nel caso specifico, allora, la variabile "qualità di prodotto" risulta essere stata esclusa dai modelli di regressione determinati anche perché non fornisce alcun apporto informativo aggiuntivo nello spiegare le vendite mediante la relazione lineare che caratterizza il modello di regressione. Esaminando ancora la variabile, qualità di prodotto, si mette in evidenza che la seconda ragione che ha portato alla sua esclusione è dovuto essenzialmente al fatto di essere altamente correlata con la variabile immagine (collinearità). Tale situazione di collinearità non solo non apporta nuove informazioni ma diventa anche difficile determinare l'effetto prodotto da ciascuna delle due variabili indipendenti fortemente correlate tra loro rispetto a quella dipendente.

Se già l'eventuale presenza di coefficienti di correlazione superiore a 0,90 è indice di collinearità, per valutare, in generale, una eventuale situazione di multicollinearità si ricorre al fattore di accrescimento della varianza, cioè:

$$Variance\ Inflation\ Factor = FIV = \frac{1}{1 - R_i^2} = \frac{1}{1 - 0,94} = 15,608 \ .$$

Nel caso di una sola variabile tale quantità risulta pari ad uno. Nel secondo caso subisce un incremento notevole, del 90%. Per valori del FIV superiori a 10 (alcuni autori per prudenza suggeriscono di abbassare il valore a 5) la correlazione risulta eccessiva, pertanto è ridondante l'apporto di una delle due variabili, che risultano essere fortemente correlate tra loro. Analizzando il valore del FIV relativo alla variabile qualità si osserva che esso è pari a 15,608, valore decisamente superiore al valore soglia indicato.

8.3. Previsione delle vendite con dati in *cross-section*

Con questo ulteriore esempio si vogliono chiarire alcuni ulteriori aspetti metodologici rimasti nascosti nell'esempio precedente; si tratta di stimare il volume delle vendite di una azienda sulla base della percezione che la clientela ha della p*erformance* dei suoi prodotti. Si tratta di un aspetto nuovo che si collega ad un altro tema di analisi di mercato circa il comportamento di acquisto dei consumatori.

Dall'archivio dei clienti di una azienda è stato selezionato un campione di 100 unità le quali sono state intervistate su singoli quesiti ai quali gli intervistati hanno risposto assegnando un punteggio riportato su scale di misura definite nell'intervallo da zero a 10[16].

I dati di base non sono delle serie temporali come nell'esempio precedente ma dati in *cross-section*. La variabile da stimare è la quantità che i singoli clienti prevedono di acquistare, supposta in relazione con le variabili che riflettono la *performance* dell'azienda.

Le variabili esplicative considerate sono le seguenti:

X_1 = Velocità di consegna del prodotto

X_2 = Prezzo

[16] Metodo di previsione delle vendite sulla base della precesione che la clientela ha della performance dei suoi prodotti (vedasi J.F. Hair, R.E. Anderson, R.L. Tatham, W.C. Black, Multivariate Data Aanlysis, Englewood Cliffs, NJ Prentice Hall, 1998, 8° editing, pp.195-214).

X_3 = Variazione del prezzo
X_4 = Immagine del produttore
X_5 = Servizio complessivo
X_6 = Immagine della forza di vendita
X_7 = Qualità del prodotto

Per quanto riguarda l'ampiezza campionaria si è seguita la regola empirica di avere dalle 10 alle 15 unità per ogni variabile indipendente considerata; in tal modo, si è cercato di garantire l'ottenimento di stime soddisfacenti.

Sette variabili esplicative di partenza sono tante. Considerandole tutte indistintamente, un qualunque modello implementato, rappresentativo di un qualunque fenomeno osservato, risulterebbe abbastanza complesso e di difficile applicazione in un contesto di replicabilità; pertanto conviene adottare una procedura *stepwise regression* (procedura a gradini o per passi successivi). Questa procedura risponde al criterio della parsimonia, nel senso che conviene inserire nel modello di regressione un varabile per volta con l'intento di rendere massima la capacità previsiva dello stesso. Questa metodica richiede la selezione di una variabile per volta partendo da quella che contribuisce in misura maggiore a spiegare la variabilità; cioè quella che è maggiormente correlata con la variabile dipendente (nel nostro caso le vendite).

Successivamente si inseriscono una per volta le altre variabili fino a quando non sono esaurite o non si decide di arrestare la procedura in base ad un criterio di arresto prefissato. Il criterio da rispettare è che l'immissione di una nuova variabile deve ridurre al massimo la variabilità residua (devianza di dispersione data dalla somma dei quadrati degli errori tra dati empirici e dati teorici). La riduzione della devianza residua deve risultare statisticamente significativa e può essere controlla attraverso un test *F* di Snedecor.

In sintesi il procedimento è il seguente:

1) si considera come prima variabile indipendente quella che presenta il più alto coefficiente di correlazione semplice con la variabile vendite; se *r* non è significativo la procedura si ferma e si conclude che il modello è del tipo $Y= b_0$;
2) si introduce come seconda variabile quella che presenta il più alto coefficiente di correlazione parziale rispetto alla Y. Se ad esempio, il coefficiente di correlazione parziale più elevato e significativo riguarda la variabile X_2 , si scrive l'equazione:

$$Y= b_0 + b_1X_1 + b_2X_2 ;$$

3) a questo punto si va a verificare la significatività di b_1 e b_2;
   a) se b_1 e b_2 sono entrambi significativi l'equazione è quella precedente e si vanno a considerare i coefficienti di correlazione parziale rispetto ad Y delle altre variabili. E' necessario calcolare i coefficienti di correlazione parziale perché è necessario eliminare l'influenza di X_1 e X_2 rispetto ad Y;
   b) se b_1 è significativo e b_2 non lo è allora si considera solo l'equazione:

$$Y= b_0 + b_1X_1$$

   c) se b_1 non è significativo e b_2 lo è, bisogna considerare le altre variabili indipendenti e considerare quella che ha il coefficiente di

correlazione parziale più elevato oltre alla X_2. Dopo aver scritto il modello di regressione occorre verificare la significatività dei relativi coefficienti di regressione parziale.

Il motivo per cui ogni volta si controlla la significatività di tutti i coefficienti di regressione parziale (compresi anche quelli che erano significativi al passo precedente) sta a dimostrare che l'aggiunta di una nuova variabile nel modello può rendere non significativo il coefficiente di regressione parziale di una variabile già scelta a causa di una eventuale alta correlazione fra esse; in questo caso la variabile che ha il coefficiente non significativo viene rimossa dal modello.

Naturalmente non esiste un modello perfetto che ci consente di scegliere le variabili. Il calcolo ci può dare solo utili indicazioni sulle variabili da usare soprattutto se siamo in una fase esplorativa. Molto importante risulta il buon senso e l'esperienza del ricercatore. Alcune variabili (anche se non risultano significativi i coefficienti di regressione parziale) devono essere mantenute per il loro significato logico; altre, invece, devono essere rimosse proprio perché non hanno nessun significato logico (anche i rispettivi coefficienti di regressione sono significativi)[17].

Tab. 8.3.1 - Matrice di correlazione

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y
X_1	1,00							
X_2	-0,35	1,00						
X_3	0,51	-0,49	1,00					
X_4	0,05	0,27	-0,12	1,00				
X_5	0,61	0,51	0,07	0,30	1,00			
X_6	0,08	0,19	-0,03	0,79	0,24	1,00		
X_7	-0,48	0,47	-0,45	0,20	-0,06	0,18	1,00	
Y	**0,68**	0,08	**0,56**	0,22	**0,70**	0,26	-0,19	1,00

Fonte: elaborazione su dati empirici.

Dal matrice di correlazione risulta che la variabile indipendente che risulta con un più elevato coefficiente di correlazione con la variabile dipendente è la X_5, cioè il "servizio complessivo". Il coefficiente di correlazione r è 0,70. In altri termini si è portati ad acquistare di più in quanto si giudica positivo il servizio reso. Costruito il modello di regressione, con il metodo dei minimi quadrati, si ha che il coefficiente di regressione stimato b_5 è pari 8,38 e l'intercetta è pari a 21,65; l'equazione risulta la seguente

$$Y = 21{,}65 + 8{,}38X_5$$

I valori degli indicatori necessari per controllare questo primo passo risultano:

- $R^2 = 0{,}49$,
- *Errore standard della stima* = 6.45,
- $t\,(X_5) = 9{,}72$,
- $\beta(X_5) = 0{,}70$

[17] Vedasi Delvecchio F. *Statistica per la ricerca sociale* , Cacucci , Bari, 1976.

E' necessario ricordare che per ogni variabile inserita nel modello è necessario esaminare:

- **L'indice di determinazione lineare** R^2, che, come già più volte ricordato, misura la frazione della variabilità totale di Y spiegata dalla relazione di Y con $X_{5.}$, sta a significare che la variabile "servizio complessivo" spiega circa la metà della variabilità degli acquisti che i 100 clienti intervistati ritengono di effettuare;
- **L'errore standard della stima**, che rappresenta una misura della variabilità dei valori osservati intorno alla retta di regressione, ossia una misura ulteriore dell'accuratezza della previsione. L'errore standard della stima si ottiene calcolando la radice quadrata della somma dei quadrati dei residui (che altro non sono che la devianza di dispersione) divisa per i suoi gradi di libertà. Nel nostro caso essendo sette le variabili e 14 le osservazioni i gradi di libertà sono 98.

$$Errore\ standard\ della\ stima = \sqrt{\frac{4071,69}{98}} = 6,45$$

- **Il valore della t di Student**, impiegato per il controllo dell'ipotesi che b_5 sia uguale a zero, ottenuto dal rapporto fra il coefficiente stimato ed il suo errore standard. Il suo valore critico ad un livello di significatività del 5% con 98 gradi di libertà risulta essere del 1,98;
- **Il coefficiente di regressione standardizzato** β che è pari a 0,70 e che consente di confrontare l'effetto che le singole variabili hanno su *Y.*

A questo punto restano da inserire le altre variabili allo scopo di migliorare il modello e diminuire la variabilità residua; in sostanza, andare a verificare se sono disponibili altre variabili da confrontare con X_5. Dalla matrice di correlazione sembrerebbe che la variabile da inserire con questa procedura s*tepwise* sia la variabile X_1 che ha un coefficiente di correlazione più elevato con Y. Ma non è così in quanto bisogna tener conto delle ***correlazioni parziali tra Y e le variabili indipendenti*** e scegliere quella variabile che presenta il più alto coefficiente di correlazione parziale.

La correlazione parziale è una misura della quota della variabilità di Y che non viene spiegata dalle variabili già inserite nel modello, ma dalle variabili che via via si inseriscono. A tal proposito bisogna precisare che quando si hanno tre o più variabili si può calcolare la correlazione parziale fra due avendo reso costante (e quindi ininfluente) la terza variabile ed eventualmente le altre. In questo caso si parla di correlazione parziale fra Y e X_3 al netto dell'influenza di X_5 e della correlazione fra Y ed X_1 sempre al netto dell'influenza di $X_{5.}$

Tab. 8.3.2 - Correlazioni parziali tra Y e le variabili indipendenti.

Correlazioni parziali	*Valori*
$r_{yx1.x5}$	0,44
$r_{yx2.x5}$	-0,45
$r_{yx3.x5}$	0,72
$r_{yx4.x5}$	0,02
$r_{yx6.x5}$	0,13
$r_{yx7.x5}$	-0,22

Fonte: elaborazione su dati empirici.

Da tabella risulta che il coefficiente di correlazione parziale più elevato al netto della variabile X_5 è il terzo pari a 0,72, (cioè la correlazione tra Y e X_3 eliminata la influenza di X_5 già inserita nel modello) mentre la correlazione parziale fra Y e X_1 è soltanto 0,44.

R^2 risulta essere pari a 0,518 (il quadrato di 0,72); come interpretare, allora, questo valore? Calcolando il quadrato di 0,72 si ottiene l'indice di determinazione lineare parziale R^2 pari a 0,518; quindi un ulteriore 51,8% della varianza non ancora spiegata, può essere colta inserendo nel modello la variabile X_3. Pertanto considerato che il 49,1 % della variabilità di Y è già spiegata dalla variabile X_5 e che un altro 51,8% del 50,9% della varianza totale residua (non spiegata con il primo modello) si spiega aggiungendo all'equazione di regressione la variabile X_3 si avrà un incremento del 26,4% della stessa variabilità spiegata. A questo punto possiamo passare allo studio del nuovo modello costruito con entrambe le variabili indipendenti X_3 e X_5, i cui risultati sono i seguenti:

$$\boldsymbol{Y = -3{,}49 + 3{,}43X_3 + 7{,}97X_5}$$

Con:

- *errore standard della stima* = 4,50,
- $\beta(X_3) = 0{,}52$ $\beta(X_5) = 0{,}67$
- $t(X_3) = 10{,}21$ $t(X_5) = 13{,}22$
- $R^2 = 0{,}76$
- $\overline{R}^2 = 0{,}75$

Dal confronto con il modello precedente si evince che R^2 è aumentato in modo significativo da 0,49 a 0,76. Dopo l'inserimento di b_3 il valore di b_5 si è modificato di poco, passando da 8,38 a 7,97. Questa è una indicazione ulteriore che le variabili X_5 e X_3 sono relativamente indipendenti (fra queste due variabili la correlazione è solo di 0,07). I valori del test *t* parziale indicano che sia X_5 che X_3 sono variabili esplicative di *Y* statisticamente significative. Il valore t per la variabile X_5 è ora 13,22, mentre valeva 9,72 al passo 1.

Il modello può proseguire inserendo la variabile X_6, variabile selezionata dopo aver nuovamente ricalcolato i coefficienti di correlazione parziale al netto di X_3 e di X_5. Completata la stima del modello, valutata la varianza di regressione ed effettuati i test di controllo possiamo scrivere l'equazione di regressione che include le variabili X_3, X_5 e X_6 (cioè la flessibilità del prezzo, il servizio complessivo e l'immagine della forza di vendita).

$$Y = -6{,}52 + 3{,}38\,X_3 + 7{,}62X_5 + 1{,}41X_6$$

La procedura a questo punto si ferma per il solo fato che i coefficienti delle altre variabili non superano il controllo del test *t* e non risultano significativamente diversi da zero.

Con l'equazione riportata sopra per ogni cliente può essere ricalcolata la quantità di prodotto che presumibilmente sarà acquistata.

Nel voler fare alcune considerazioni in merito al modello, si fa notare che dal loro confronto risulta che i due fattori che influenzano fortemente il probabile acquisto del prodotto sono: il servizio complessivo ed il prezzo. Incrementi in una di queste due variabili si ripercuotono in corrispondenti aumenti nelle quantità acquistate. In particolare un aumento di un punto nella percezione che il cliente ha del servizio erogato avrà come risultato un aumento nell'uso del prodotto di circa l'8%. Supponendo che un cliente assegni un punteggio pari a 4 al prodotto, il modello risulta essere il seguente con una previsione di acquisto di 43,12.

$$-6{,}52 + 3{,}38 *(4) + 7{,}62*(4) + 1{,}41* (4) = 43{,}12.$$

L'accuratezza previsiva del modello è buona, con una varianza spiegata del 77% ed un errore standard della stima della variabile dipendente è risultato pari al 4,4%.

Appendice:

Modelli teorici di correlogrammi relativi a processi stazionari AR che sono di andamento noto[18].

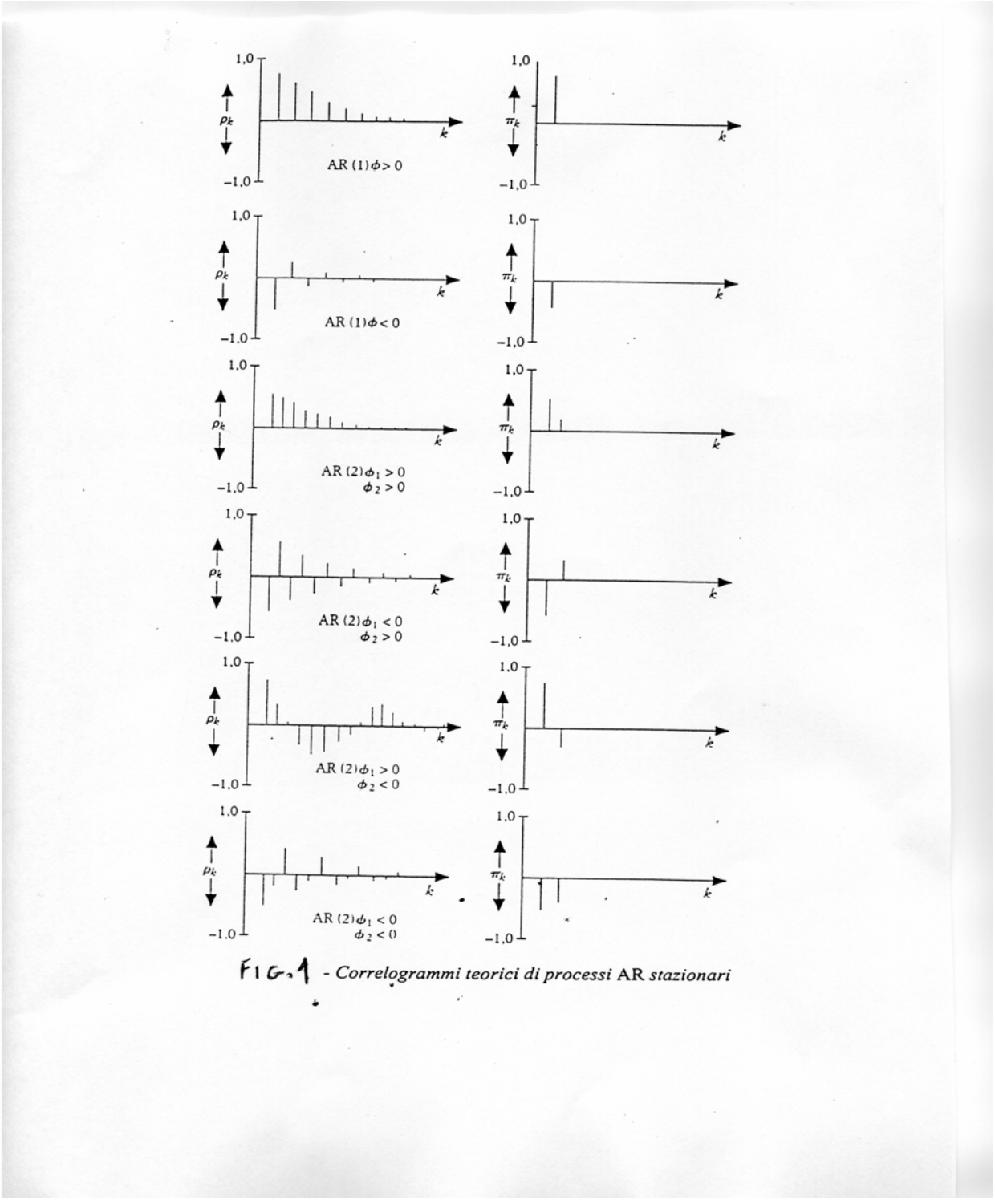


FIG. 1 - *Correlogrammi teorici di processi* AR *stazionari*

[18] Brasini S., Tassinari F. (2000), *Lezioni di Statistica Aziendale*, Società Editrice Esculapio, Bologna.

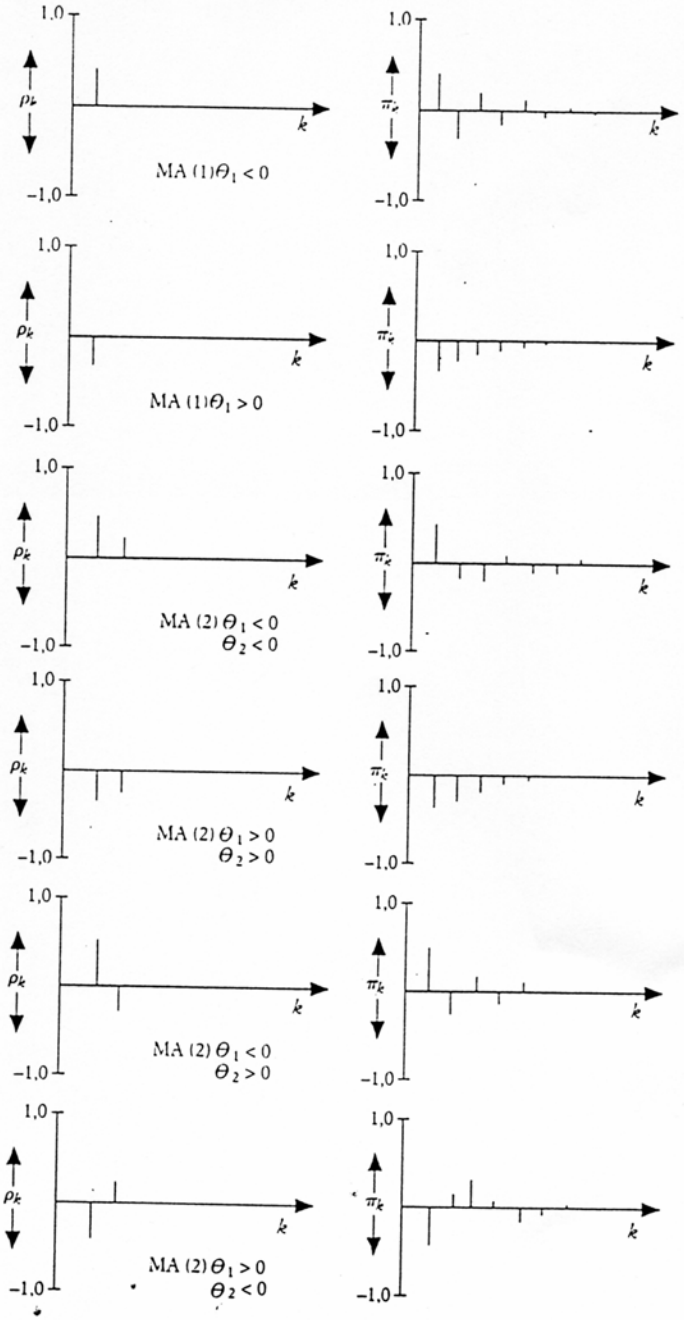


FIG. 2) *Correlogrammi teorici di processi* MA *stazionari*