CAPITOLO PRIMO

PREVISIONI A BREVE TERMINE: VENDITE E DOMANDA DI MERCATO

Sommario: 1. Introduzione alle previsioni in azienda. - 2. Metodi di previsione a lungo termine. - 3. Metodi di previsione a breve termine. - 4. Metodo esogeno: il modello di regressione. - 5. Metodo di Box e Jenkins.

1. INTRODUZIONE ALLE PREVISIONI IN AZIENDA

Le aziende generalmente per svolgere le operazioni di gestione devono effettuare due tipologie di **previsioni**:

— la previsione delle vendite di un'azienda;
— la previsione della domanda globale di mercato.

In particolare, la **previsione delle vendite** rappresenta una operazione di notevole importanza, in quanto da essa derivano tutte le informazioni necessarie all'azienda per poter ottimizzare le risorse disponibili alla diverse funzioni aziendali: da quella degli acquisti a quella di produzione, alla funzione personale, alla funzione commerciale ed infine alla gestione finanziaria.

Ovviamente, affinchè le stime delle vendite siano quanto più attendibili e affidabili è necessario che ciascuna azienda conosca al meglio il contesto economico-ambientale in cui opera, quindi, è necessaria un'altrettanto affidabile **analisi previsionale della domanda di mercato**.

Per poter procedere alla realizzazione di tali analisi, le aziende si servono prevalentemente di tecniche e strumenti della statistica. Di seguito saranno illustrati i metodi maggiormente diffusi nella pratica aziendale.

Tali metodi di previsione possono essere distinti in diverse tipologie anche se una prima distinzione può essere fatta tra **metodi congetturali** e **metodi statistici**.

I primi si fondano sull'esperienza e sul giudizio personale dello statistico mentre i secondi si fondano su modelli che ipotizzano valide per il futuro relazioni realizzatesi in passato.

A seconda dell'orizzonte temporale cui si riferiscono si distinguono **previsioni a breve** e **previsioni a medio o a lungo termine**.

2. METODI DI PREVISIONE A LUNGO TERMINE

Tra i metodi di previsione utilizzati per il lungo periodo un cenno merita il **metodo Delphi**. Si tratta di una particolare forma di panel consistente in una serie di questionari scritti; esso equivale a una discussione anonima tra esperti, i quali esprimono il loro parere, indicando il loro accordo e/o il loro dissenso rispetto ad affermazioni loro presentate in merito a un tema stabilito. I pareri in tal modo raccolti sono, quindi, sintetizzati in forma aggregata e anonima e sono nuovamente inviati ai partecipanti per una nuova consultazione, in modo che ciascuno dei partecipanti possa confrontare la propria opinione con quelle espresse dal gruppo; in tale fase sono inviati ai partecipanti anche i nuovi questionari elaborati in base alle risposte ottenute. Successivamente, si realizza una nuova aggregazione dei risultati, tale da condurre a un affinamento della previsione e si sviluppa un nuovo questionario. Questa fase, se necessario, è nuovamente ripetuta e, infine, i risultati sono distribuiti ai partecipanti.

Il metodo non solo trova applicazione nell'ambito di problematiche connesse alla previsione aziendale ma è altresì finalizzato ad ottenere indicazioni prospettiche (ricerche di scenari relativi a periodi futuri) su temi di grande importanza come ad esempio l'innovazione tecnologica, le energie alternative, le frontiere della biogenetica ecc.

Il **metodo degli scenari** è un altro metodo utilizzato per le previsioni a lungo termine. Attraverso tale metodo si esaminano le potenziali **conseguenze** di situazioni future e si individuano le **condizioni** per il loro verificarsi. Si tratta di uno strumento strategico valido per tutte le aziende, dalle più grandi alle più piccole.

3. METODI DI PREVISIONE A BREVE TERMINE

Relativamente alle **previsioni a breve termine** concernenti, cioè, la domanda di mercato, si distingue tra *metodi endogeni* e *metodi esogeni*.

I **metodi endogeni** usano dati in **serie storica** (come la domanda di mercato riferita ai mesi di un dato anno oppure a dati anni, o il livello delle vendite in analoghi archi temporali); la previsione si basa sullo studio dell'andamento temporale della serie stessa estendendo i dati a intervalli di tempo al di fuori di quello di osservazione. Tra i metodi endogeni meritano un cenno a parte l'*estrapolazione* e il *metodo di Box e Jenkins* che consiste

in un approccio stocastico all'analisi delle serie storiche e di cui ci occuperemo di seguito, nel corso di questo capitolo.

L'**estrapolazione** è il processo di determinazione di una successione di valori teorici di frequenze o intensità, ottenuti in corrispondenza di modalità di un carattere quantitativo in una distribuzione di frequenza, o modalità di tempo in una serie storica, esterne all'intervallo di osservazione.

Il procedimento si attua sia *analiticamente* sia *graficamente*.

I **metodi esogeni**, invece, individuano, innanzi tutto, i fattori o le cause che agiscono su una data variabile (come la domanda di mercato o il livello delle vendite) che sono assunti come variabili indipendenti o variabili esplicative, e specificano la relazione tra questi ultimi e la prima attraverso l'esplicitazione di un'**equazione** (o un **sistema di equazioni** se le variabili esplicative sono più di una) esprimente la relazione esistente tra le variabili, allo scopo di prevedere i valori della variabile al variare della variabile (o delle variabili) da cui dipende. Essa implica l'esistenza di una antecedenza logica tra variabili. Il termine **regressione** esprime il concetto di dipendenza funzionale tra due o più variabili. La previsione dei valori della variabile si basa sull'ipotesi secondo cui la relazione individuata permane nel tempo, quindi è valida non solo rispetto alle osservazioni empiriche ma anche per il futuro.

Altri metodi di previsione a breve termine sono costituiti dai *sondaggi d'opinione* e dalle *tavole input-output.*

Un **sondaggio d'opinione** è un'indagine volta a conoscere le tendenze dell'opinione pubblica su un determinato argomento; è uno strumento importante nelle previsioni a breve termine sia in campo economico sia in campo politico.

Lo strumento di previsione più importante è la **tavola input-output** (**tavola *I - O***), o **tavola delle interdipendenze settoriali**, o **tavola intersettoriale dell'economia**. Essa intende i sistemi economici nazionali come divisi in settori, misura e prevede gli scambi tra un settore e l'altro, ossia i **flussi intermedi**, sulla base di coefficienti tecnici, esprimendo le relazioni con un sistema di equazioni.

La prima tavola input-output fu elaborata nella prima metà del secolo scorso dall'economista russo Wassily Leontief su dati della contabilità nazionale statunitense e per la cui costruzione si è aggiudicato il premio Nobel per l'economia nel 1973.

Leontief realizzò l'idea del fisiocratico François Quesnay che nel 1758, nel *Tableau économique*, per primo indagò le relazioni economiche esisten-

ti tra classi sociali e settori produttivi, considerando, quindi, l'economia come un insieme di settori. Nel *Tableau* due sono i settori produttivi:

— quello primario (agricoltura e industria estrattiva);
— quello secondario (manifatture e commercio).

Una tavola *I-O* è un quadro contabile che evidenzia un elevato numero di relazioni nell'economia, essa costituisce la base di un modello utilizzato a fini interpretativi e previsivi e consente, inoltre, mediante specifici procedimenti statistico-matematici, di stimare le ripercussioni sul livello di produzione e sui fabbisogni delle singole branche provocate da modificazioni della domanda finale (consumi, investimenti, esportazioni); ciò permette di effettuare previsioni e di supportare decisioni di politica economica o di programmazione.

Per comprendere il modo di funzionamento di una **tavola input - output** focalizziamo l'attenzione su un particolare paese. Si immagini che le attività produttive nel paese abbiano luogo in un certo numero di settori. Concettualmente lo schema definito da Leontief presuppone che ogni settore produca un solo bene, ma nella costruzione concreta di tavole input - output è inevitabile l'aggregazione, che di norma è effettuata definendo raggruppamenti omogenei dal punto di vista tecnico produttivo. Pertanto, i settori possono essere industrie specializzate o diversi livelli di aggregazione delle stesse, o imprese. L'**unità elementare** di riferimento delle tavole è la **branca**; a differenza dei conti istituzionali di un paese, l'analisi delle operazioni effettuata per branca riguarda **relazioni economico - tecniche** e non di comportamento, per cui risulta possibile solo fino al livello del risultato di gestione.

I dati necessari alla costruzione di una tavola input - output di un'economia sono rappresentati dai flussi di prodotti da ciascun settore a sé stesso e agli altri dell'economia. Questi flussi sono quantificati per un dato intervallo di tempo, generalmente un anno.

Cosa si evince leggendo una tavola input-output nel senso delle righe e nel senso delle colonne?

Si distingue tra:

— **lettura della tavola nel senso delle colonne** che consente di analizzare, per ciascuna branca, il **processo di formazione delle risorse** (ossia produzione e importazioni) e la **struttura dei costi di produzione**, per cui i totali di ogni colonna rappresentano gli acquisti di ciascuna branca;

— **lettura della tavola nel senso delle righe** che consente di analizzare la produzione delle branche secondo la **destinazione**, ossia come il risultato dell'attività produttiva si ripartisca tra le branche per gli impieghi intermedi e gli utilizzatori finali, per cui i totali di ogni riga rappresentano le vendite effettuate dalla branca.

Una tavola input-output consta di tre sezioni:

— una **tavola degli impieghi intermedi** che è la **matrice input-output**, costituita da una tabella a doppia entrata in cui sono indicate per riga e per colonna le medesime branche;
— una **tavola degli impieghi finali**, a destra della prima, in cui sono riportate per colonna, per branca di origine, gli usi finali delle risorse quali consumi, investimenti, variazioni delle scorte e, se si tratta di un sistema economico aperto, esportazioni. Nell'ultima colonna della tavola può figurare il **totale degli impieghi**;
— una **tavola degli impieghi primari e delle risorse**, in cui sono registrati i costi relativi ai fattori primari, e che è rappresentata da righe intestate alle componenti del valore aggiunto.

Un'ultima riga della tavola è destinata al **totale delle risorse**.

Per costruire queste tavole si parte dal raggruppamento delle imprese per *settori omogenei* (industria, agricoltura, servizi ecc.) che saranno riportati nella tabella a doppia entrata:

— sulla *colonna* come *settore acquirente*;
— sulla *riga* come *settore venditore*;

ciò perché ogni settore economico allo stesso tempo acquista beni o servizi (*input*) dagli altri settori e vende loro la propria produzione (*output*).

La tavola intersettoriale indica, inoltre, i c.d. **reimpieghi**, ossia i beni o servizi utilizzati dallo stesso settore che le ha prodotte (valori indicati nelle caselle corrispondenti a righe e colonne intestate allo stesso settore).

Sia dato un sistema economico aperto agli scambi con l'estero costituito da n branche, una tavola input-autput in tale sistema è del tipo di quella di seguito riportata, in cui:

— $a_{11}, a_{12}, \ldots, a_{nn}$ sono gli impieghi intermedi;
— $x_1, x_2, \ldots, x_n$ sono gli impieghi intermedi totali;
— $Z_{11}, Z_{12}, \ldots, Z_{n4}$ sono gli impieghi finali;
— $X_1, X_2, \ldots, X_n$ sono le risorse totali;
— ${}_PZ, {}_IZ, \ldots, {}_TZ$ sono gli impieghi finali soddisfatti, rispettivamente, dalla produzione interna (P) dalle importazioni (I) e gli impieghi finali totali (T);

— ${}_{P.}X, {}_{I.}X, \ldots, {}_{T.}X$ rappresentano il totale delle risorse per branche di origine;
— $y_{11}, y_{12}, \ldots, y_{4n}$ rappresentano il valore aggiunto e le sue componenti, ossia i flussi di redditi primari corrisposti ai fattori della produzione (lavoro, capitale e imprese) a titolo di retribuzione dei servizi resi;
— $Y_{1.}, Y_{2.}, \ldots, Y_{4.}$ rappresentano i totali delle risorse primarie;
— $X_{.1}, X_{.2}, \ldots, X_{.n}$ rappresentano i totali delle risorse per branche di destinazione.

Branche di origine	**Branche di destinazione**	**Totale impieghi intermedi**	**Impieghi finali**				
	$S_1 \quad S_2 \quad \ldots \quad S_n$		**Consumi**	**Investimenti**	**Esportazioni**	**Variazioni delle scorte**	**Totale risorse**
S_1	$a_{11} \quad a_{12} \quad \ldots \quad a_{1n}$	$x_{1.}$	Z_{11}	Z_{12}	Z_{13}	Z_{14}	$X_{1.}$
S_2	$a_{21} \quad a_{22} \quad \ldots \quad a_{2n}$	$x_{2.}$	Z_{21}	Z_{22}	Z_{23}	Z_{24}	$X_{2.}$
$\ldots$	$\ldots \quad \ldots \quad \ldots$	...	$\ldots$	$\ldots$	$\ldots$	$\ldots$	...
S_n		$x_{n.}$	Z_{n1}	Z_{n2}	Z_{n3}	Z_{n4}	$X_{n.}$
	$a_{n1} \quad a_{n2} \quad \ldots \quad a_{nn}$						
	Risorse primarie		${}_{P.}Z$				${}_{P.}X$
			${}_{I.}Z$				${}_{I.}X$
			${}_{T.}Z$				${}_{T.}X$
Salari e stipendi	$y_{11} \quad y_{12} \quad \ldots \quad y_{1n}$	$Y_{1.}$					
Valore aggiunto	$y_{21} \quad y_{22} \quad \ldots \quad y_{2n}$	$Y_{2.}$					
Imposte indirette nette	$y_{31} \quad y_{32} \quad \ldots \quad y_{3n}$	$Y_{3.}$					
Importazioni	$y_{41} \quad y_{42} \quad \ldots \quad y_{4n}$	$Y_{4.}$					
Totale	$X_{.1} \quad X_{.2} \quad \ldots \quad X_{.n}$						

4. METODO ESOGENO: IL MODELLO DI REGRESSIONE

Nell'analisi statistica la **regressione** è volta alla ricerca di un modello atto a descrivere la relazione esistente tra una **variabile dipendente**, e una o più **variabili indipendenti** o **esplicative**.

La scelta dell'una o dell'altra variabile come indipendente non è arbitraria ma legata alla natura del fenomeno: si sceglie come indipendente la variabile che sia *logicamente antecedente* rispetto all'altra. In un modello di regressione, le variabili esplicative (dette anche **regressori**) spiegano, prevedono, simulano, controllano la variabile dipendente.

Il termine *regressione* fu coniato da Galton che, nel misurare la relazione tra statura dei padri e quella dei figli, osservò una regressione dei valori delle altezze dei figli verso la media.

Per effettuare una regressione si fa riferimento a modelli teorici di vario tipo: lineare, parabolico, esponenziale, logaritmico etc.

I **modelli di regressione** si basano, appunto, sull'esistenza di una relazione di antecedenza logica tra il fenomeno osservato e una o più variabili esplicative, essi sono costituiti da una sola o più equazioni lineari, o non, nei parametri.

Bisogna fare una ulteriore distinzione tra **modelli di regressione semplice** e **multipla**, i primi descrivono come una data variabile (indipendente o esplicativa) *spieghi* un'altra variabile (dipendente), mentre i secondi sono utilizzati quando le variabili indipendenti sono più di una.

Il **modello di regressione** di cui ci occuperemo in questo testo è un modello di regressione multipla che intende *spiegare* il legame funzionale esistente tra m variabili indipendenti $X_1, X_2, ..., X_m$ e una variabile dipendente Y. Tale modello è indubbiamente più realistico del modello di regressione semplice, in quanto è estremamente raro che un fenomeno economico o sociale sia spiegato da una sola variabile.

Tuttavia, questo modo di procedere presenta degli inconvenienti: non è possibile affermare che tra le due variabili X e Y esiste una perfetta relazione matematica del tipo:

$$Y = f\left(X_1, X_2, ..., X_m\right)$$

in quanto, innanzitutto, non si dispone di tutte le informazioni relative alle variabili X_i e al fenomeno Y ma solo di un campione di osservazioni su di essi estratto da una data popolazione, che darà luogo a coppie di ordinate le

quali solo in termini probabilistici esprimono la relazione esistente tra le corrispondenti variabili della popolazione; inoltre, a causa di fenomeni imprevedibili, errori di misurazione, scarti accidentali, non esiste un legame di tipo deterministico tra le variabili. Costituisce, quindi, una semplificazione affermare che il fenomeno *Y* è *spiegato* dal fenomeno *X*, in quanto nella realtà esistono interrelazioni tra variabili che non possono essere compendiate in alcuna scrittura. Per questi motivi il modello di riferimento sarà del tipo:

$$Y = f\left(X_1, X_2, ..., X_m\right) + \varepsilon$$

La variabile *Y* è una **variabile casuale** (v.c.) risultante dalla somma di una **componente deterministica** $f\left(X_1, X_2, ..., X_m\right)$ e di una **componente stocastica** ε, dove ε è una v.c. che ha funzioni compensative per le discrepanze esistenti tra il modello e la realtà.

La v.c. ε è definita **errore** o **scarto** tra le costruzioni teoriche e la realtà osservata, infatti, si ha:

$$\varepsilon = Y - f\left(X_1, X_2, ..., X_m\right)$$

Dopo aver provveduto ad individuare le variabili che *spiegano* il fenomeno in oggetto si passa alla fase di **specificazione del modello** che consiste nella sua rappresentazione formale; si assume che la relazione tra le variabili sia lineare nei parametri, dunque, il modello di regressione sarà del tipo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m + \varepsilon$$

Tale modello presuppone che la variabile *Y* sia data da una combinazione lineare di *m* variabili esplicative $X_1, X_2, ..., X_m$ e da un termine di disturbo.

Lo stadio successivo relativo alla costruzione di un modello riguarda la **stima dei parametri**; è necessario, a questo punto, trovare una relazione lineare tra le variabili esplicative e la variabile *Y* in modo da minimizzare il valore della componente stocastica. Il modello di regressione avrà espressione:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_m X_m$$

dove $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_m$ sono i coefficienti di regressione stimati attraverso il modello.

Essendo un modello di regressione una rappresentazione semplificata della realtà, è indispensabile sapere quanta parte della variabilità di *Y* è spiegata sotto le ipotesi specificate; una importante misura della bontà di una modello di regressione ai dati osservati si basa sugli **errori** che, non essendo osservabili a priori, possono essere calcolati come residui tra le singole osservazioni e i valori teorici. Siano *N* le osservazioni, per la *i*-esima osservazione, $i = 1,2,...,N$, sia Y_i il valore osservato e $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_m X_{mi}$ il corrispondente valore teorico calcolato con la regressione, l'espressione analitica dell'errore è la seguente:

$$e_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_m X_{mi}\right)$$

Nell'eseguire una stima puntuale dei parametri è necessario trovare una relazione tra la variabile *Y* e le variabili esplicative in modo da **minimizzare** il valore dell'errore.

Una misura dell'adeguatezza del modello di regressione ampiamente usata è basata sul valore dell'errore per le *N* osservazioni e assume la denominazione di **indice di determinazione**, la sua espressione analitica è la seguente:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}$$

dove $\overline{Y}$ è il valore medio della variabile *Y*.

Essendo il rapporto di una parte della variabilità, cioè quella dovuta alla regressione, rispetto alla variabilità totale si verifica che:

$$0 \le R^2 \le 1$$

quanto più è prossimo a 1 tanto più si può affermare che il modello è adeguato.

Si dimostra, tuttavia, che, al crescere del numero di variabili esplicative si ha un valore più elevato di R^2. Pertanto, non è necessariamente vero che se l'indice di determinazione R^2 assume un valore tendente all'unità, il mo-

dello di regressione è adeguato; l'aggiunta di una variabile esplicativa accrescerà sempre R^2 a prescindere dal fatto che la variabile aggiunta contribuisca o meno al modello; è, dunque, possibile che i modelli che hanno valori elevati di R^2 siano poco adatti alla previsione o alla stima.

Un valore dell'indice più alto non fa dello stesso, quindi, un criterio per identificare un modello migliore; per ovviare a simili difficoltà di interpretazione si preferisce usare l'**indice di determinazione corretto** (*adjusted* R^2) la cui espressione analitica, tenendo conto che m è il numero di variabili esplicative, è la seguente:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} e_i^2 \Big/ \left(N - m\right)}{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2 \Big/ \left(N - 1\right)}$$

il quale, non necessariamente aumenta al crescere del numero di variabili esplicative inserite nel modello.

All'indice suddetto si ricorre soprattutto se si vogliono confrontare modelli di regressione che intendono spiegare la medesima variabile indipendente utilizzando un numero diverso di variabili esplicative.

L'**errore standard** di tale modello, noto anche con la denominazione di **errore quadratico medio dei residui**, ha la seguente espressione analitica:

$$s_e = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N - m}}$$

e misura la dispersione delle osservazioni attorno al modello di regressione.

Le ipotesi sottostanti un modello di regressione si configurano alla stregua di congetture la cui validità può essere confermata o messa in forse a seguito dell'evoluzione reale del fenomeno allo studio, così come si manifesta con le osservazioni empiriche; per questo occorre porsi in atteggiamento critico rispetto alle stesse. La **verifica statistica d'ipotesi** in un modello di regressione mira a valutare la **significatività dei coefficienti**, ossia a valutare se la variabile o le variabili esplicative siano statisticamente in grado di spiegare la variabile dipendente.

Una volta costruito un modello di regressione si può stabilire, attraverso l'uso di un test, se esiste o no una relazione significativa tra la variabile dipendente e l'insieme delle variabili esplicative.

Talvolta si vuole verificare se esiste una relazione lineare tra la variabile Y e le m variabili esplicative.

Pertanto, l'ipotesi da testare è:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0$$

Il rifiuto dell'ipotesi nulla implica che almeno una delle variabili contribuisce significativamente a *spiegare* la variabile dipendente. L'ipotesi nulla può essere verificata utilizzando, a un dato **livello di confidenza**, il **test *F* di Snedecor**. Se il valore assunto dal test, funzione delle osservazioni, è tale da condurre a un **rifiuto** dell'ipotesi nulla, si afferma che il modello è efficace nello spiegare il fenomeno.

Per valutare, invece, la significatività dei singoli coefficienti di regressione si usa il **test *t*** dato, per ciascun coefficiente, dal rapporto tra la stima del coefficiente e il suo errore standard.

Il **valore di significatività**, o ***p*-value**, ossia il livello di significatività associato a t calcolato empiricamente, misura la confidenza con cui l'ipotesi relativa alla nullità di ciascun coefficiente è rifiutata. Dal confronto tra il p-value con il livello di confidenza prescelto, si può stabilire se rifiutare oppure no l'ipotesi di nullità dei coefficienti.

Data la difficoltà di calcolo di tali indici e, poiché tale testo vuole evitare eccessive formulazioni matematiche, ci serviamo di un esempio in cui gli indici suddetti sono ottenuti attraverso un foglio di lavoro di Microsoft Excel e di una particolare tipologia di modello lineare generale fornita dal seguente modello a tre variabili:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad i = 1, 2, ..., N$$

in cui le osservazioni campionarie si presentano sotto forma di una successione di terne del tipo $\left(X_{1i}, X_{2i}, Y_i\right)$ e i loro punti-immagine danno luogo a un diagramma a scatter nello spazio tridimensionale. Dallo stesso diagramma si evincono le eventuali relazioni esistenti tra i caratteri.

I coefficienti incogniti β_0, β_1 e β_2 sono stimati ricorrendo al **metodo dei minimi quadrati** in virtù del quale la somma dei quadrati dei residui da minimizzare è:

$$G(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2$$

ossia deve essere:

$$G(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2 = \min$$

Uguagliando a zero le derivate parziali rispetto ai parametri si ottiene un sistema da cui, semplificando, si ottiene un sistema di **equazioni normali** nelle tre incognite β_0, β_1 e β_2, la cui soluzione, $\hat{\beta}_0, \hat{\beta}_1$ e $\hat{\beta}_2$, è rappresentata da formule troppo complesse per essere esposte in questo testo.

Caso

Un'azienda operante nel settore della ristorazione importa vino dalla Francia. La tabella seguente riporta, relativamente agli anni indicati, le importazioni, i prezzi all'importazione, le spese di trasporto, espresse in termini di variazioni relative rispetto al periodo precedente.

Tabella 1

Anni	Importazioni	Prezzi	Pubblicità
1989	1,4750	0,0210	0,0000
1990	-0,0232	0,0029	-0,0020
1991	0,0155	-0,0146	0,1203
1992	0,0163	0,0139	-0,0235
1993	0,0070	-0,0078	-0,0240
1994	-0,5224	0,0010	0,0398
1995	0,8333	0,0000	-0,0100
1996	0,0977	-0,0177	-0,0745
1997	-0,0797	0,0050	-0,1948
1998	0,4398	-0,0269	-0,1531
1999	-0,3992	-0,0041	0,0335
2000	-0,4278	0,0010	0,1481
2001	1,5909	-0,0134	0,2285
2002	-0,1404	0,0333	0,0500
2003	0,0204	0,0111	0,0524
2004	-0,0360	0,0000	-0,1629
2005	-0,0871	0,0379	-0,0346
2006	-0,0909	0,0115	0,0414
2007	-0,1250	0,0275	-0,1398

Per analizzare le importazioni di vino nel tempo si è tenuto conto, quindi, di due fattori che ne possono influenzare la quantità: il prezzo all'importazione e le spese pubblicitarie.
Per evidenziare l'entità di tali influenze consideriamo un foglio elettronico di Excel e otteniamo la **matrice di correlazione** da cui si evincono i **coefficienti di correlazione di Bravais e Pearson** tra le variabili. A questo punto, sempre in riferimento ai dati riportati in tabella 1, dopo aver impostato un foglio Excel, selezionare il menu **Strumenti**, quindi, **Analisi dati**. Si tratta di un indice che misura la relazione lineare esistente tra due caratteri statistici. È espresso dal rapporto tra la covarianza tra le due variabili ***X* e *Y* ed il prodotto dei rispettivi scarti quadratici medi:**

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

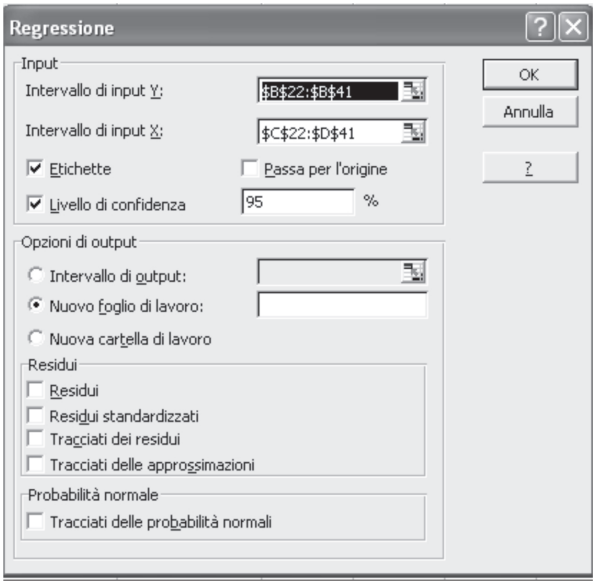
In generale, più il valore assoluto di $r_{XY} \to 1$ tanto più forte è il legame lineare tra *X* e *Y*.
Nella finestra di dialogo **Strumenti di analisi** selezionare **Correlazione**.

	A	B	C	D
1		Importazioni	Prezzi	Pubblicità
2	Importazioni	1		
3	Prezzi	-0,1508	1	
4	Pubblicità	0,2052	-0,0715	1

Siano:

— *Y* le importazioni;
— X_1 i prezzi;
— X_2 le spese pubblicitarie;

dalla matrice si evince che la correlazione tra le importazioni e i prezzi è, evidentemente, negativa e pari a $r_{YX_1} = -0{,}1508$, mentre la correlazione tra le importazioni e la pubblicità è lievemente positiva e pari a $r_{YX_2} = 0{,}2052$. Le due variabili esplicative sono pressoché incorrelate tra loro, infatti $r_{X_1X_2} = -0{,}0715$.
Excel dispone anche di uno strumento di analisi in grado di eseguire un'analisi lineare della regressione utilizzando il metodo dei minimi quadrati.
Lo strumento in questione è **Regressione**.
A questo punto, sempre in riferimento ai dati riportati in tabella 1, considerando il medesimo foglio Excel visto, selezionare il menu **Strumenti**, quindi, **Analisi dati**. Nella finestra di dialogo **Strumenti di analisi** selezionare **Regressione**.
La finestra di dialogo va completata nel modo illustrato.



L'output di riepilogo dello strumento Regressione è il seguente:

	A	B	C	D	E	F	G	H	I
1	OUTPUT RIEPILOGO								
2									
3	*Statistica della regressione*								
4	R multiplo	0,246413249							
5	R al quadrato	0,060719489							
6	R al quadrato corretto	-0,05669057							
7	Errore standard	0,590196062							
8	Osservazioni	19							
9									
10	ANALISI VARIANZA								
11		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*			
12	Regressione	2	0,36028679	0,180143	0,517157449	0,605846342			
13	Residuo	16	5,573340036	0,348334					
14	Totale	18	5,933626826						
15									
16		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*	*Inferiore 95,0%*	*Superiore 95,0%*
17	Intercetta	0,160174538	0,139815774	1,145611	0,268802681	-0,136221596	0,456570672	-0,136221596	0,456570672
18	Prezzi	-4,53892129	8,060579762	-0,5631	0,581171978	-21,62658319	12,5487406	-21,62658319	12,5487406
19	Pubblicità	1,031024626	1,281620021	0,80447	0,43291725	-1,685887835	3,747937087	-1,685887835	3,747937087
20									
21									
22									
23									

La lettura del foglio è immediata, basta solo dire che $\hat{\beta}_0 = 0{,}16, \hat{\beta}_1 = -4{,}54$ e $\hat{\beta}_2 = 1{,}03$ per cui il modello sarà del tipo:

$$\text{IMPORTAZIONI} = 0{,}16 - 4{,}54 \times \text{PREZZI} + 1{,}03 \times \text{PUBBLICITA'} + e$$

da cui si evince facilmente la relazione inversa esistente tra importazioni e prezzi all'importazione.

A) Previsione

Consideriamo ora uno degli obiettivi fondamentali dell'analisi: la **previsione** di valori della variabile dipendente in corrispondenza di valori esterni al campo di osservazione della o delle variabili esplicative.

Dopo aver stimato, con il metodo dei minimi quadrati, i parametri del modello di regressione è possibile utilizzarli per eseguire previsioni dei valori della *Y* in corrispondenza di dati valori delle variabili esplicative.

Così come si hanno stime puntuali e intervallari dei parametri del modello, è possibile eseguire **previsioni puntuali** e **previsioni intervallari** di valori della variabile *Y* in corrispondenza di dati valori della variabile *X*.

Supposto, per semplicità di trattazione, un modello di regressione semplice del tipo:

$$Y = \beta_0 + \beta_1 X$$

con *X* rappresentante la variabile *Variazione relativa dei prezzi*, scriviamo che la **previsione puntuale** al tempo *T* del valore della variabile *Y* corrispondente al valore un anno avanti $X_T^a(1)$ della variabile *X* è fornita da:

$$Y_T^a = \hat{\beta}_0 + \hat{\beta}_1 X_T^a(1)$$

dove $\hat{\beta}_0$ e $\hat{\beta}_1$ sono le stime dei minimi quadrati dei coefficienti β_0 e β_1.

La **previsione intervallare** al tempo *T* del valore della variabile *Y* è fornita, invece, dal seguente intervallo:

$$\left[Y_T^a(1) - t_{\alpha/2;N-m} s_e \sqrt{1 + \frac{1}{N} + \frac{\left(X_T^a(1) - \overline{X}\right)^2}{(N-1)s_X^2}} ; Y_T^a(1) + t_{\alpha/2;N-m} s_e \sqrt{1 + \frac{1}{N} + \frac{\left(X_T^a(1) - \overline{X}\right)^2}{(N-1)s_X^2}} \right]$$

dove:

$t_{\alpha/2;N-m}$ è il quantile della funzione di *Student* in corrispondenza di $N - m$ gradi di libertà e per un livello di confidenza $(1-\alpha)$;

s_e è la stima dello scarto quadratico medio degli errori;

s_X^2 è la stima della varianza della variabile *Variazione relativa dei prezzi*.

Il processo di determinazione di uno o di una successione di valori teorici ottenuti in corrispondenza di valori esterni all'intervallo di osservazio-

ne presuppone che i valori campionari osservati delle variabili esplicative si siano succeduti in passato con una certa regolarità e che possa valere anche per specificati valori della variabile. In quanto basato solo sulla regolarità, in passato, del fenomeno che rappresenta, la previsione può essere poco attendibile non tenendo conto di cause perturbatrici che potrebbero verificarsi in futuro.

B) Multicollinearità nel modello di regressione

Una delle ipotesi alla base di un modello di regressione multipla è l'**ipotesi di non collinearità delle variabili esplicative** secondo cui le variabili esplicative sono linearmente indipendenti. Se tra le stesse non esiste dipendenza lineare si dice che esse sono **ortogonali**.

Tuttavia, in alcune situazioni c'è, tra variabili, dipendenza quasi lineare, ossia, almeno una di loro è una combinazione lineare delle altre, in tale caso si dice che si è in presenza di **multicollinearità**.

Per individuare la presenza di multicollinearità si elabora il **Fattore di Incremento della Varianza** (FIV) che, nella regressione della variabile esplicativa X_j, è espresso da:

$$FIV_j = \frac{1}{1 - R_j^2}$$

dove R_j^2 è l'indice di determinazione lineare nella regressione della variabile X_j sulle altre variabili. In presenza di due sole variabili esplicative, R_1^2 è il coefficiente di determinazione della regressione di X_1 su X_2.

Il FIV, per ciascun termine del modello, misura l'effetto della dipendenza tra variabili sulla varianza di quel termine.

Se le variabili sono tra loro incorrelate il FIV è uguale a 1, mentre se le stesse sono correlate tra loro il valore del FIV aumenta.

Il prospetto seguente illustra i diversi valori assunti dal *FIV* relativo alla *j* – esima variabile, in corrispondenza di valori crescenti di R_j^2.

R_i^2	0,2	0,6	0,8	0,9	0,92	0,95	0,98	0,99	0,999	0,9999
FIV_j	1,25	2,5	5	10	12,5	20	50	100	1.000	10.000

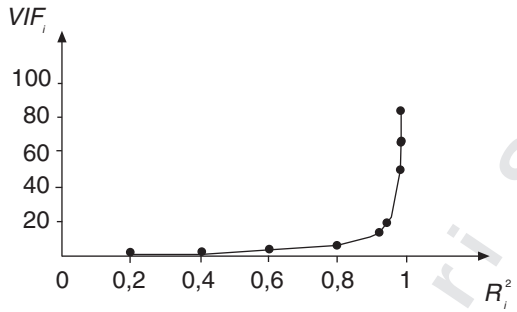


Fig. 1 – Come varia il FIV al variare di R_j^2

Al limite, quando la *j*–esima variabile dipende linearmente dalle rimanenti $R_j^2 = 1$ per cui FIV_j ha valore infinito.

Mardquardt (1970) sostiene che se FIV_j è maggiore di 10, vi è un'elevata correlazione tra X_j e le altre variabili esplicative. Altri studiosi, invece, suggeriscono di ricorrere a metodi di stima diversi dai minimi quadrati quando FIV_j è maggiore di 5.

5. METODO DI BOX E JENKINS

Tra i metodi endogeni di previsione, interessante è il metodo di **Box e Jenkins**. Per affrontare lo studio del metodo occorre premettere che esso si basa su un approccio moderno di analisi delle serie storiche. Infatti, l'analisi classica delle serie storiche non consente di studiare serie evolutive, ossia serie di dati che subiscono evoluzioni temporali di tipo non deterministico. L'**analisi moderna** delle serie storiche delle vendite, invece, ricerca un modello stocastico complessivo capace di generare le serie osservate.

Compito dell'analista è, innanzi tutto, quello di stimare i parametri del modello con riferimento ai legami esistenti tra i dati empirici, e successivamente di verificare la validità dello stesso, accertata la quale esso può essere utilizzato a fini di previsione. La procedura utilizzata in questo contesto è stata elaborata da Box e Jenkins.

Prima di procedere oltre è necessario, a questo punto, dare una nozione fondamentale nell'analisi moderna: quella di **processo stocastico**. Trattasi di una famiglia di v.c. *Z* dipendenti dal parametro tempo *t*, che varia in un insieme *T* di numeri reali, con $t \in T$.

A seconda della natura di T si hanno processi a **tempo discreto** o **continuo**. In questo contesto consideriamo successioni di v.c. continue con parametro t discreto.

Una **serie storica** osservata $\{Z_t, t \in T\}$ è una **parte finita di una realizzazione di un processo stocastico**, e solo una parte in quanto essa è un campione unico della famiglia di v.c. che caratterizzano il processo stocastico.

Nell'analisi delle serie storiche il problema fondamentale è risalire dalla serie storica al processo generatore.

Nell'approccio considerato si stabilisce un'**analogia tra serie storica osservata e campione casuale**, nel senso che così come ad una data popolazione statistica corrisponde un dato universo dei campioni, di cui il campione osservato consente di inferire sui parametri incogniti della popolazione, ad un processo stocastico corrisponde un insieme di serie storiche di cui quella osservata costituisce una realizzazione finita del processo incognito e costituisce la base per la costruzione di un modello statistico in grado di descrivere il meccanismo generatore del processo stocastico.

A) Classificazione dei processi stocastici

Per condurre uno studio di tipo inferenziale su una serie storica si impongono delle restrizioni e limitazioni ai processi stocastici in modo da ottenerne una classe di più agevole analisi.

Processo stazionario

Un processo stocastico generatore di serie storiche si dice **stazionario in senso debole** o, semplicemente, **stazionario** se:

— ha valore medio costante al variare del tempo (**ipotesi di invarianza in media**);
— ha varianza finita e costante al variare del tempo (**ipotesi di omoschedasticità**);
— ha autocovarianza dipendente solo dal lag (ritardo).

Una tipologia di processo stocastico stazionario è il **processo White Noise** (***WN***), o **rumore bianco**, definito come una successione di v.c. di media nulla e varianza costante e incorrelate in tempi successivi.

Processo gaussiano

Un processo stocastico si dice **gaussiano** se sono normali le distribuzioni di **probabilità** delle v.c. costituenti il processo. Per cui se si tratta di un processo White Noise, allora si parla di processo *WN gaussiano*.

Processo ergodico

Un processo stocastico si dice **ergodico** quando l'autocovarianza tende a zero al crescere del lag.

Processo invertibile

Un processo stocastico Z_t si dice **invertibile** quando esiste una funzione lineare $f(\cdot)$ e un processo *WN*, ε_t, tale che può essere espresso in funzione della serie passata e da un White Noise, in termini formali:

$$Z_t = C + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + ... + \varepsilon_t$$

in cui α_1, α_2, ... sono coefficienti e *C* è una costante rappresentativa del livello iniziale della serie.

A questo punto si rende necessario introdurre una funzione utilizzata per confronti tra processi stocastici: la *funzione di autocorrelazione globale*.

La **funzione di autocorrelazione globale** (spesso indicata con l'acronimo **FAC**) esprime il coefficiente di correlazione lineare tra Z_t e Z_{t+k}, al variare di *k*. Essa si calcola a partire dalla funzione di autocovarianza, ossia dalla funzione che misura, per ogni lag *k,* la **covarianza** tra Z_t e Z_{t+k} supponendo che Z_t sia un processo stocastico stazionario, la sua formula è:

$$\gamma(k) = \mathrm{E}\ (Z_t - \mu)\ (Z_{t-k} - \mu)$$

essendo $\mu = E(Z_t)$ il valore medio del processo. La FAC si ottiene standardizzando la funzione di autocovarianza, ossia facendo il rapporto tra l'autocovarianza al tempo *t* e la varianza della serie che, per serie stazionarie, coincide con l'autocovarianza al lag 0; in simboli:

$$\rho(k) = Corr(Z_t, Z_{t+k}) = \frac{Cov(Z_t, Z_{t+k})}{\sqrt{Var(Z_t)\,Var(Z_{t+k})}}$$

che, siccome $Var(Z_t) = Var(Z_{t+k}) = \gamma(0)$, diviene:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} \qquad \text{per } k = 0, 1, 2, ...$$

ed esprime la correlazione interna alle osservazioni di una serie storica, detta **correlazione seriale**. Il grafico della funzione, al variare di *k* è detto

correlogramma. Il suo frequente utilizzo nell'analisi delle serie storiche è giustificato dal fatto che, trattandosi di un valore standardizzato consente confronti tra diversi processi stocastici.

La **funzione di autocorrelazione parziale** (spesso indicata con l'acronimo **FACP**) esprime, invece, la correlazione tra Z_t e Z_{t+k}, dopo aver eliminato gli effetti delle relazioni lineari intermedie. La funzione di autocorrelazione parziale di lag 2 misura la correlazione seriale tra Z_t e Z_{t+k} dopo aver eliminato la correlazione parziale di lag 1.

La stazionarietà in media della serie storica si realizza se la funzione di autocorrelazione decade verso 0 con una certa rapidità, invece, la non stazionarietà in media si realizza allorché la funzione di autocorrelazione globale decresce molto lentamente.

B) I processi ARMA

Per introdurre i processi elaborati da Box e Jenkins occorre premettere le seguenti definizioni di *processi autoregressivi* e di *processi media mobile.*

Un processo stocastico si dice **autoregressivo di ordine *p***, e si indica con *AR*(p) (acronimo di *Auto – Regressive*) se l'osservazione Z_t al tempo t è definita come una combinazione lineare di p termini immediatamente precedenti e di una componente casuale; in simboli:

$$Z_t = C + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + \varepsilon_t$$

Un processo stocastico si dice **media mobile di ordine *q***, e si indica con *MA*(q) (acronimo di *Moving Average* che, in inglese significa, appunto, media mobile), se l'osservazione Z_t al tempo t è una combinazione lineare di processi White Noise; in simboli:

$$Z_t = C + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$$

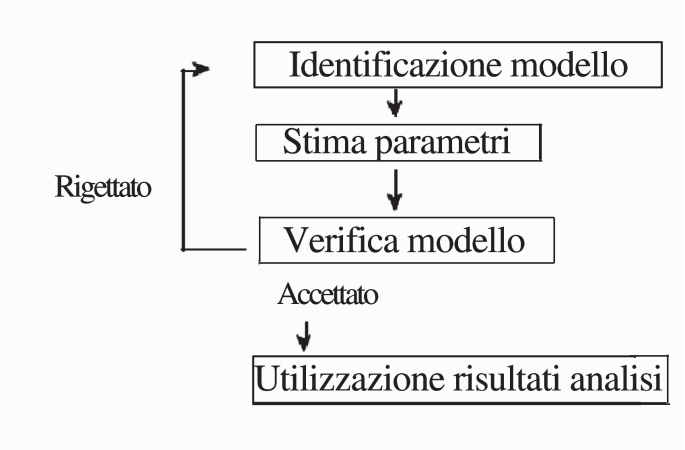
Un processo stocastico costituito congiuntamente da una parte autoregressiva e da una parte media mobile è definito **processo ARMA(*p*, *q*)**, **autoregressivo di ordine *p* e media mobile di ordine *q*** e la cui espressione analitica è la seguente:

$$Z_t = C + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$$

La condizione di stazionarietà di una serie storica non è quasi mai riscontrabile empiricamente; generalmente le serie storiche sono di tipo evo-

lutivo. Grazie al lavoro di Box e Jenkins del 1976, è possibile risalire a processi stazionari generalizzando la classe dei processi ARMA.

La procedura elaborata da Box e Jenkins si occupa di costruire un modello tale da approssimare il processo generatore di una serie storica osservata. Essa consiste nell'iterare lo schema seguente:



Nella prima fase, si ha la **identificazione del modello** che consiste nella formalizzazione in termini matematico – statistici delle ipotesi teoriche, considerando che un modello ARMA può essere identificato, non solo per serie storiche stazionarie, ma anche per serie evolutive relativamente alle differenze successive dei dati originari. Il tipo di specificazione adottata dipende, generalmente, non solo dal particolare processo che si considera, ma anche dal materiale empirico a disposizione. In questo stadio bisogna risolvere problemi diversi quali la determinazione delle variabili influenti, la decisione circa la forma funzionale delle relazioni tra variabili, e infine, di fondamentale importanza, bisogna tener conto del carattere necessariamente aleatorio del modello e per questo considerare le deviazioni residuali tra relazioni teoriche e osservazioni empiriche.

In statistica, alcune **trasformazioni** sono in grado di rendere i dati osservati più coerenti con le ipotesi di un modello. In particolare, il modello ARMA da identificare dovrebbe essere applicato ad una serie storica stazionaria che rispetti l'ipotesi di normalità del rumore bianco.

Se una delle ipotesi alla base non si verifica, ovvero se:

— non è rispettata l'ipotesi di invarianza in media, è possibile operare con le differenze successive;
— se i dati non presentano omoschedasticità si effettua una trasformazione logaritmica degli stessi;
— se non è soddisfatta l'ipotesi di normalità del rumore bianco, allora si controlla che il processo stesso sia normale.

Nella seconda fase è trattato il problema della **quantificazione** delle relazioni, si procede alla stima dei parametri incogniti del modello. In questa fase si intende individuare una struttura del modello che sia il più possibile prossima alla incognita struttura vera, ossia alla reale rappresentazione del fenomeno in oggetto.

I parametri del modello sono stimati ricorrendo al **metodo di massima verosimiglianza**, meno frequente è il metodo dei minimi quadrati. Tale fase è macchinosa, perciò si ricorre a software specifici.

La **verifica** (terza fase) consiste in una sequenza di operazioni atte a valutare la validità del modello sulla base delle osservazioni disponibili sulle diverse variabili dello stesso.

La verifica della validità del modello stimato viene effettuata ricorrendo all'**analisi dei residui** tra valori osservati e valori stimati della serie storica.

Durante tale fase si procede alla verifica dell'ipotesi di mancanza di correlazione dei residui, ossia se essi sono realizzazioni di un processo White Noise.

I parametri stimati sono testati per accertarne la significatività statistica, e quelli non significativi sono cancellati dal modello.

Se il modello costruito non è accettato durante questa fase, si procede ad un'iterazione delle fasi: identificazione, stima, verifica. In caso di accettazione dello stesso si procede, invece, ad un suo utilizzo a **fini previsionali**.

Glossario

Covarianza: misura del legame lineare tra due variabili statistiche X ed Y. È data dalla media aritmetica del prodotto degli scarti di due variabili dalle loro rispettive medie, vale a dire:

$$Cov(X,Y) = \sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

Quando scarti positivi o negativi della variabile X tendono ad associarsi rispettivamente a scarti positivi o negativi della variabile Y, allora i loro prodotti saranno positivi e la covarianza risulterà positiva (concordanza). Quando, invece, scarti positivi della variabile X tendono ad associarsi a scarti negativi della variabile Y o viceversa, allora i loro prodotti saranno negativi e la covarianza risulterà negativa (discordanza). Essendo il numeratore del coefficiente di correlazione lineare la covarianza ne assume lo stesso segno ma a differenza del coefficiente di correlazione lineare, essa dipende dalle unità di misura di X e di Y.

Matrice: una matrice $n \times k$ è un insieme rettangolare di elementi disposti su n righe e k colonne. Quando la matrice contiene informazioni statistiche, la rappresentazione usuale della matrice dei dati contiene le rilevazioni di k variabili (le colonne) su n soggetti (le righe).
Una matrice si dice *quadrata* se è del tipo $n \times n$, ossia se il numero di righe è uguale al numero di colonne.

Panel: indagine caratterizzata da campioni, permanenti o continui, costituiti dalle stesse unità statistiche che vengono intervistate, salvo sostituzioni richieste da esigenze tecniche, in successivi periodi di tempo.

Probabilità: è un numero associato al verificarsi di un evento. Nonostante si tratta di un concetto primitivo, la probabilità ha ricevuto nel tempo definizioni diverse. Secondo la *definizione classica*, formalizzata da Laplace, la probabilità di un evento E è il rapporto fra il numero m dei casi favorevoli al verificarsi di un evento e il numero n dei casi possibili purché siano tutti ugualmente possibili; in simboli:

$$P(E) = \frac{m}{n}$$

Secondo la *definizione frequentista*, cui un contributo fondamentale è stato dato da *Von Mises*, la probabilità di un evento è il limite cui tende la frequenza relativa dell'evento quando il numero delle prove tende all'infinito. Per la *definizione soggettivista*, opera di de Finetti la probabilità di un evento è il grado di fiducia che un individuo coerente, sulla base delle sue conoscenze, attribuisce al verificarsi dell'evento in questione. A Kolmogorov si deve, infine, la *definizione assiomatica* di probabilità.

Variabile casuale: funzione *misurabile* e a valori reali definita sullo spazio campione. Si definisce una variabile casuale, indicata anche con v. c. quando si crea una corrispondenza tra l'insieme dei risultati di una prova e l'insieme dei numeri reali.
Le variabili casuali possono essere discrete o continue.

Varianza: indice di variabilità molto utilizzato e indicato con σ^2. Esso è definito come la media aritmetica del quadrato degli scarti dalla media aritmetica:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Se si dispone della distribuzione delle modalità x_i e delle rispettive frequenze n_i, $i = 1, 2, \ldots, k$, la sua espressione analitica è:

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \mu)^2 n_i}{\sum_{i=1}^{k} n_i}$$

Rappresenta il quadrato dello *scarto quadratico medio* ed è espressa nel quadrato della unità di misura del fenomeno X.