La Regressione Lineare

1. Cos'è l'Analisi della Regressione Multipla? L'analisi della regressione multipla è una tecnica statistica che può essere impiegata per analizzare la relazione tra una **variabile dipendente** e diverse **variabili indipendenti** (**predittori**). Obiettivo di tale analisi è utilizzare le variabili indipendenti, i valori delle quali sono noti, per prevedere il valore della variabile dipendente oggetto di interesse del ricercatore. Per far questo, la procedura di regressione attribuisce un peso ad ogni predittore, così da assicurare al modello la miglior capacità previsiva sull'insieme delle variabili indipendenti presenti nella combinazione lineare. Ogni peso denota il contributo relativo alla previsione dato dalla specifica variabile indipendente e insieme facilita l'interpretazione e la comprensione dell'influenza che ogni variabile ha sul risultato della previsione, anche se l'esistenza di correlazione tra le variabili indipendenti complica il processo interpretativo. L'insieme delle variabili indipendenti ponderate costituisce l'**equazione di regressione**, definita come la combinazione lineare delle variabili indipendenti che meglio prevede la variabile dipendente.

L'analisi della regressione multipla è una tecnica che si fonda su una relazione di dipendenza. Pertanto, per poterla utilizzare, è necessario individuare, tra le variabili oggetto di studio, qual è quella dipendente e quali sono quelle indipendenti. L'analisi della regressione è inoltre uno strumento statistico che dovrebbe essere utilizzato solo quando tutte le variabili (criterio e predittori) sono di tipo metrico. Comunque, sotto certe condizioni, è possibile includere anche dati non metrici sia per l'insieme delle variabili indipendenti (trasformando variabili ordinali o nominali in variabili simulate (dummy), sia per la variabile dipendente (utilizzando una misura binaria e ricorrendo alla regressione logistica). In sintesi, per poter ricorrere all'analisi della regressione multipla: (1) i dati devono essere metrici o altrimenti devono essere opportunamente trasformati, e (2) prima di ricavare l'equazione di regressione, il ricercatore deve decidere quale variabile è dipendente e quali, tra le restanti variabili, sono quelle indipendenti.

2. Un esempio di Regressione Semplice e Multipla L'obiettivo dell'analisi della regressione è prevedere i valori assunti da una variabile dipendente a partire dalla conoscenza di quelli osservati su una o più variabili indipendenti. Quando il problema coinvolge una sola variabile indipendente, la tecnica statistica viene definita **regressione semplice**. Quando invece il problema coinvolge due o più variabili indipendenti, è detta **regressione multipla**.

Per illustrare i concetti basilari di queste tecniche, è utile fare un esempio riguardante un campione di otto famiglie e relativo all'uso di carte di credito. Lo scopo consiste nell'individuazione e nella valutazione dei fattori che influenzano il numero di carte di credito utilizzate. Sono stati identificati tre potenziali fattori (ampiezza della famiglia, reddito della famiglia e numero di automobili possedute) e sono stati raccolti tutti i dati necessari su ciascuna delle otto famiglie (si veda la Tabella 1). Nella terminologia dell'analisi della regressione, la variabile dipendente (Y) è il numero di carte di credito usate, mentre le tre variabili indipendenti (V_1, V_2 e V_3) sono rispettivamente l'ampiezza della famiglia, il reddito familiare e il numero di automobili possedute.

TABELLA 1 Risultati dell'indagine sull'utilizzazione di carte di credito.

Numero identificativo della famiglia	*Numero di Carte di Credito utilizzate* Y	*Ampiezza della famiglia* V_1	*Reddito della famiglia (×000 $)* V_2	*Numero di automobili possedute* V_3
1	4	2	14	1
2	6	2	16	2
3	6	4	14	2
4	7	4	17	1
5	8	5	18	3
6	7	5	21	2
7	8	6	17	1
8	10	6	25	2

La discussione di questo esempio è divisa in tre parti, che consentono di mostrare le modalità con cui l'analisi della regressione stima la relazione tra le variabili indipendenti e la variabile dipendente. Gli argomenti affrontati riguardano:
(1) la previsione senza variabili indipendenti, con l'uso della sola media;
(2) la previsione con l'uso di una sola variabile indipendente – regressione semplice;
(3) la previsione con l'uso di diverse variabili indipendenti – regressione multipla.

3. Previsione senza variabili indipendenti

Prima di procedere alla stima della prima equazione di regressione, cominciamo calcolando il parametro di base rispetto al quale verrà confrontata la capacità predittiva dei vari modelli di regressione. Tale parametro dovrebbe rappresentare la miglior previsione ottenuta senza l'uso di alcuna variabile indipendente. Benché sia possibile ricorrere a molteplici soluzioni (ad esempio, utilizzare un valore specificato in precedenza, oppure una misura di tendenza centrale, come la media, la mediana o la moda), il parametro previsivo di base usato nella regressione è la semplice media della variabile dipendente, che ha diverse proprietà desiderabili. Nell'esempio, la media aritmetica del numero di carte di credito usate è sette. La previsione senza variabili indipendenti può quindi essere enunciata come segue: "il numero previsto di carte di credito usate da una famiglia è sette". Inoltre, possiamo scrivere la stessa enunciazione con una equazione di regressione:

Numero previsto di carte di credito = Numero medio di carte di credito

oppure

$$\hat{Y} = \bar{y}$$

Il ricercatore, però, vuole rispondere anche alla domanda seguente: "Quanto è precisa la previsione effettuata?" Poiché la media non fornisce una previsione perfetta di ciascun valore assunto dalla variabile dipendente, bisogna individuare un metodo per la valutazione dell'accuratezza previsiva che possa essere usato sia per la previsione senza variabili indipendenti sia per tutti i possibili modelli di regressione con variabili indipendenti. Il modo consueto per valutare l'accuratezza di una previsione è dato dall'esame degli errori prodotti nella previsione della variabile dipendente. Ad esempio, se si sfrutta la previsione ottenuta con la media, secondo cui ciascuna famiglia utilizza sette carte di credito, considerando la famiglia 1 (si veda la Tabella 2) si sovrastima il numero di carte di credito da essa usato di tre. Pertanto, in tal caso, l'errore è +3. Se si procedesse allo stesso modo per tutte le famiglie, si osserverebbero alcune stime troppo elevate, altre troppo basse, altre ancora esatte. Benché ci si possa aspettare di ottenere una misura dell'accuratezza della stima semplicemente sommando gli errori, in realtà questa operazione non è di alcuna utilità, dal momento che la somma degli errori rispetto alla media aritmetica è sempre uguale a zero. Di conseguenza, la semplice somma degli errori non cambierebbe mai, indipendentemente dal livello di precisione nella previsione della variabile dipendente ottenuta con la media. Per superare questo problema, gli errori vengono prima elevati al quadrato e quindi sommati. Il totale è definito **devianza** (somma degli scarti al quadrato - nella terminologia inglese, *sum of squared*

errors, ***SSE***), quantità che fornisce una misura dell'accuratezza della previsione, variabile in funzione dell'errore di previsione. L'obiettivo è quindi ottenere la più piccola possibile somma dei quadrati degli errori, poiché ciò significa rendere più accurata la previsione.

Come termine di confronto è stata scelta la media aritmetica poiché essa produce sempre la più piccola devianza rispetto ad ogni altra misura di tendenza centrale, includendo tra queste la mediana, la moda, qualsiasi altro valore medio Pertanto nell'esempio, utilizzando la media a fini previsivi senza considerare alcuna variabile indipendente, si ottiene la migliore previsione del numero di carte di credito, con una somma dei quadrati degli errori pari a 22 (Tabella 2). La previsione data dalla media verrà utilizzata come base di confronto proprio perché essa rappresenta la miglior previsione possibile senza l'uso di variabili indipendenti.

TABELLA 2 Previsione con l'uso della media della variabile dipendente

Numero identificativo della famiglia	*Numero di Carte di Credito utilizzate*	*Previsione*[a]	*Errore di previsione*[b]	*Errore di previsione al quadrato*
1	4	7	–3	9
2	6	7	–1	1
3	6	7	–1	1
4	7	7	0	0
5	8	7	+1	1
6	7	7	0	0
7	8	7	+1	1
8	10	7	+3	9
	-----		-----	-----
Totale	56		0	22

a Numero medio di carte di credito = 56 ÷ 8 = 7
b Con il termine *errore* ci si riferisce alla differenza tra il valore vero della variabile dipendente e il valore stimato.

4. Previsione con una sola variabile indipendente (la Regressione Semplice)

I ricercatori sono sempre interessati al miglioramento delle loro previsioni. In precedenza si è detto che, quando non si utilizzano variabili indipendenti, la miglior previsione è data dalla media. Ma nell'esempio sulle otto famiglie sono state raccolte informazioni anche su misure che potrebbero agire come variabili indipendenti. Proviamo allora a valutare se la conoscenza di una di queste variabili può aiutare il processo previsivo attraverso quella che è chiamata *regressione semplice*.

La regressione semplice è una procedura per la previsione dei dati (così come quella che adotta la media) che utilizza la regola già vista – la minimizzazione della somma dei quadrati degli errori di previsione. Già sappiamo che, non volendo impiegare l'ampiezza della famiglia come variabile indipendente, possiamo descrivere al meglio il numero di carte di credito utilizzate attraverso il valore medio, sette. Il ricorso alla regressione semplice, però, risponde all'obiettivo di migliorare la previsione fornita dalla media attraverso l'uso di una variabile indipendente.

5. Il ruolo del coefficiente di correlazione

Servendosi delle maggiori informazioni ottenute dall'indagine, è possibile migliorare la stima del numero di carte di credito utilizzate attraverso la riduzione degli errori di previsione. Per farlo, tali errori devono essere associati (correlati) con una delle potenziali variabili indipendenti (V_1, V_2 o V_3). Il concetto di associazione, rappresentato dal **coefficiente di correlazione** (***r***), è fondamentale per l'analisi della regressione poiché descrive la relazione tra due variabili, che si dicono correlate se le variazioni dell'una sono associate alle variazioni dell'altra. In questo caso, quando una variabile cambia, sappiamo in che

modo si modifica l'altra. Se V_i fosse correlata con l'uso delle carte di credito, potremmo sfruttare questa relazione per prevedere, come descritto di seguito, il numero di carte di credito utilizzate:

Numero di carte di credito previsto	=	Variazione nel numero di carte di credito usate, associata all'unità di variazione in V_i	×	Valore di V_i

oppure

$$\hat{Y} = b_1 V_i$$

Un'illustrazione della procedura adottata su dati ipotetici con una sola variabile indipendente X_1 è messa in luce nella Tabella 3. Se osserviamo che quando X_1 cresce di una unità la variabile dipendente si incrementa (in media) di due, possiamo fare previsioni per tutti i valori della variabile indipendente. Ad esempio, quando X_1 assume il valore 4, possiamo prevedere che la variabile dipendente assumerà il valore 8 (si veda la parte A della Tabella 3). Perciò, il valore previsto sarà sempre due volte il valore di X_1 ($2X_1$). Spesso accade che la previsione venga migliorata aggiungendo un valore costante. Nella parte A della Tabella 3 possiamo vedere come la previsione condotta utilizzando semplicemente il valore $2X_1$ produce sempre un errore pari a due. Di conseguenza, cambiando il modello con l'aggiunta della costante due otteniamo una previsione sempre perfetta (si veda la parte B della Tabella 3). Vedremo in seguito che nella stima di un'equazione di regressione è spesso opportuno includere una costante, definita **intercetta**.

6. La specificazione dell'equazione di Regressione Semplice

Nell'esempio possiamo selezionare la "miglior" variabile indipendente in base al valore assunto dal coefficiente di correlazione, poiché al crescere del suo valore aumenta la forza del legame tra le variabili e di conseguenza l'accuratezza della previsione. La Tabella 4 mostra la matrice di correlazione tra la variabile dipendente (Y) e le variabili indipendenti (V_1, V_2 e V_3). Scorrendo la prima colonna, si vede che V_1, la variabile ampiezza della famiglia, presenta il più alto coefficiente di correlazione con la variabile dipendente ed è perciò la prima candidata per la nostra regressione semplice. La matrice di correlazione contiene inoltre i coefficienti di correlazione tra le variabili indipendenti, cosa che risulta di estrema importanza dell'analisi della regressione multipla (con due o più variabili indipendenti).

TABELLA 3 Miglioramento dell'accuratezza della previsione con l'aggiunta dell'intercetta in un'equazione di regressione

Valore di X_1	*Variabile dipendente*	*Previsione*	*Errore di previsione*
PARTE A: PREVISIONE SENZA INTERCETTA Equazione di regressione: $Y = 2X_1$			
1	4	2	2
2	6	4	2

3	8	6	2
4	10	8	2
5	12	10	2

PARTE B: PREVISIONE CON INTERCETTA
Equazione di regressione: $Y = 2.0 + 2X_1$

1	4	2	0
2	6	6	0
3	8	8	0
4	10	10	0
5	12	12	0

TABELLA 4 Matrice di correlazione dello studio sulle carte di credito

Variabile	Y	V_1	V_2	V_3
Y Numero di carte di credito usate	1.000			
V_1 Ampiezza della famiglia	0.866	1.000		
V_2 Reddito della famiglia	0.829	0.673	1.000	
V_3 Numero di automobili	0.342	0.192	0.301	1.000

Possiamo ora stimare il primo modello di regressione semplice per il campione di otto famiglie e vedere in che modo si adatta ai nostri dati. Tale modello di regressione può essere descritto nel modo seguente:

Numero di carte di credito previsto = Intercetta + Variazione nel numero di carte di credito usate per unità di variazione dell'ampiezza della famiglia × Valore di V_i

oppure

$$\hat{Y} = b_0 + b_1 V_i$$

Nell'equazione di regressione b_0 rappresenta l'intercetta, mentre il termine b_1, definito **coefficiente di regressione**, rappresenta la variazione stimata della variabile dipendente per una unità di variazione di quella indipendente. L'errore di previsione, dato dalla differenza tra il valore vero e il valore stimato per la variabile dipendente, è definito **residuo** (e). L'analisi della regressione consente anche di valutare statisticamente la significatività dell'intercetta e del(i) coefficiente(i) di regressione, in modo da controllare se essi divergono significativamente da zero (e quindi se il loro effetto sulla variabile dipendente è o non è nullo).

Usando una procedura matematica nota come **metodo dei minimi quadrati**, si possono stimare i valori di b_0 e b_1 in modo tale che la somma dei quadrati degli errori di previsione sia minima. Nell'esempio, i valori appropriati sono: per la costante (b_0) 2.87 e per il coefficiente di regressione (b_1) relativo all'ampiezza della famiglia 0.97. L'equazione indica che, per ogni membro che si aggiunge al nucleo familiare, il possesso di carte di credito aumenta in media di 0.97. La costante 2.87 è interpretabile solo in funzione dell'intervallo di valori che può assumere la variabile indipendente. In questo caso, è impossibile ammettere una famiglia costituita da zero elementi e pertanto il valore dell'intercetta, da solo, non ha significato pratico. Ciò comunque non invalida la sua utilizzazione, poiché esso migliora la previsione dell'uso di carte di credito per ogni possibile valore dell'ampiezza familiare (che nell'esempio va da 1 a 5). Nei casi in cui le variabili indipendenti possono assumere valore zero, l'intercetta ha una precisa interpretazione. In alcune situazioni particolari, quando è noto che la relazione tra le variabili passa attraverso l'origine degli assi, il termine corrispondente all'intercetta può essere eliminato (in tal caso si parla di "regressione attraverso l'origine"). In questi casi l'interpretazione dei residui e dei coefficienti di regressione cambia leggermente. Nella Tabella 5 sono

mostrati l'equazione di regressione semplice e i valori risultanti della previsione e dei residui per ognuna delle otto famiglie.

TABELLA 5 Risultati della regressione semplice usando come variabile indipendente l'ampiezza della famiglia

Equazione di regressione: $Y = b_0 + b_1 V_1$
Equazione di previsione: $Y = 2.87 + 0.97 V_1$

Numero identificativo della famiglia	*Numero di Carte di Credito usate*	*Ampiezza della famiglia* (V_1)	*Previsione con la regressione semplice*	*Errore di previsione*	*Errore di previsione al quadrato*
1	4	2	4.81	–0.81	0.66
2	6	2	4.81	1.19	1.42
3	6	4	6.75	–0.75	0.56
4	7	4	6.75	0.25	0.06
5	8	5	7.72	0.28	0.08
6	7	5	7.72	–0.72	0.52
7	8	6	8.69	–0.69	0.48
8	10	6	8.69	1.31	1.72
	-----				-----
Totale	56				5.50

Poiché abbiamo sfruttato lo stesso criterio di stima (minimizzazione della somma dei quadrati degli errori, ovvero metodo **minimi quadrati**), confrontando la previsione data dal modello di regressione semplice con quella ottenuta senza variabili indipendenti possiamo determinare se la conoscenza dell'ampiezza del nucleo familiare ci ha aiutato a migliorare la stima del possesso di carte di credito. La somma dei quadrati degli errori commessi usando la media (il parametro di base) era 22; ricorrendo alla nuova procedura con una sola variabile indipendente, la somma dei quadrati degli errori diminuisce fino a 5.50 (si veda la Tabella 5). È chiaro allora che l'approccio della regressione semplice, con l'uso del metodo dei minimi quadrati e di una variabile indipendente, è decisamente migliore, a fini previsivi, rispetto a quello che utilizza la sola media.

7. La costruzione di un intervallo di confidenza per la previsione

Poiché non è stata ottenuta una previsione perfetta della variabile dipendente, è importante stimare l'intervallo dei valori che potremmo aspettarci di osservare, piuttosto che limitarci soltanto al calcolo di una stima puntuale. Comunque, essa rappresenta la miglior stima della variabile dipendente e si dimostra essere la previsione media ottenibile per un qualsiasi dato valore della variabile indipendente. A partire da questa stima, possiamo poi calcolare l'intervallo dei valori previsti basandoci su una misura dell'errore di previsione atteso. Nota come **errore standard della stima** (nella terminologia inglese, *standard error of the estimate*, ***SEE***), tale misura può essere definita semplicemente come la deviazione standard degli errori di previsione. È possibile costruire un intervallo di confidenza per una variabile intorno al suo valore medio sommando a questo (con segno positivo e negativo) la deviazione standard moltiplicata per una certa quantità. Ad esempio, aggiungendo alla media ±1.96 volte la deviazione standard, si ottiene un intervallo per grandi campioni che contiene al 95% di probabilità il valore vero della variabile.

Possiamo utilizzare un metodo simile per le previsioni ricavate da un modello di regressione. Usando la stima

puntuale, possiamo aggiungere ad essa (con segno più e meno) l'errore standard della stima moltiplicato per una data quantità (in funzione del livello di confidenza desiderato e della dimensione campionaria), così da ottenere i limiti superiore e inferiore dell'intervallo per la previsione ottenuta con una qualsiasi variabile indipendente. L'errore standard della stima (*SSR*) è calcolato come mostrato di seguito

$$SSR = \sqrt{\frac{\text{Somma dei quadrati degli errori}}{\text{Numerosità campionaria - 2}}} \ .$$

La costante con cui moltiplicare l'*SSR* per derivare l'intervallo di confidenza è determinata dal livello di significativà (α) e dalla numerosità campionaria (*N*), che restituiscono un valore *t*. L'intervallo di confidenza è quindi ottenuto calcolando i due estremi nel modo seguente: per l'estremo inferiore, sottraendo al valore previsto la quantità (*SSR* × *t*); per l'estremo superiore, aggiungendo al valore previsto la stessa quantità (*SSR* × *t*). Nel nostro modello di regressione semplice, *SSR* = 0.957 (radice quadrata di 5.50 diviso per 6). L'intervallo di confidenza per la previsione è costruito selezionando la costante moltiplicativa dell'errore standard in una tavola della distribuzione *t* di Student, in base al livello di confidenza voluto e alla numerosità campionaria. Nell'esempio, per un livello di confidenza del 95% e per 6 gradi di libertà (numerosità campionaria meno il numero dei coefficienti nel modello, 8 – 2 = 6), il valore *t* è pari a 2.447. La quantità da sommare (con segno positivo e negativo) al valore previsto è allora (0.957 × 2.477) 2.34. Se sostituiamo l'ampiezza media della famiglia (4.25) nell'equazione di regressione, il valore previsto è 6.99 (diverso dalla media, che è pari a 7, solo per effetto dell'arrotondamento). L'intervallo atteso per il numero di carte di credito usate va da 4.65 (6.99 – 2.34) a 9.33 (6.99 + 2.34).

8. La valutazione dell'accuratezza della previsione

Se la somma dei quadrati degli errori (*SSR*) rappresenta una misura dell'errore di previsione, dovremmo essere in grado di ricavare anche una misura del successo della nostra previsione, che possiamo chiamare **devianza di regressione** (somma dei quadrati degli scarti di regressione – nella terminologia inglese *sum of squares regression*, ***SSR***). Aggregate, queste due misure dovrebbero uguagliare la **devianza totale** (somma totale degli scarti dalla media al quadrato – nella terminologia inglese *total sum of squares*, ***TSS***), lo stesso valore ottenuto nel caso della previsione con la sola media. Quando il ricercatore utilizza variabili indipendenti, la somma totale degli scarti dalla media al quadrato può essere scomposta nella (1) somma dei quadrati degli scarti tra i valori previsti con la(e) variabile(i) indipendente(i) e il valore medio, che è detta devianza di regressione, e nella (2) somma dei quadrati degli errori di previsione, detta devianza di dispersione o devianza d'errore:

$\sum_{i=1}^{n}(y_i - \bar{y})^2$	=	$\sum_{i=1}^{n}(y_i - \hat{y})^2$	+	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
TSS	=	*SSR*	+	*SSM*
Devianza totale	=	Devianza di dispersione	+	Devianza di regressione

dove

$\bar{y}$	=	media di tutte le osservazioni
y_i	=	valore osservato sull'individuo *i*
$\hat{y}$	=	valore previsto per l'osservazione *i*

Tale scomposizione della devianza totale è utilizzabile per valutare in che modo l'equazione di regressione descrive il possesso di carte di credito delle famiglie. Il numero medio di carte di credito possedute dalle famiglie campione è la miglior stima del numero di carte di credito posseduto da una famiglia qualsiasi. Sappiamo che questa non è una stima estremamente accurata, ma è la migliore previsione disponibile quando non si utilizzano altre variabili. In precedenza,

la bontà della previsione ottenuta usando la media era stata valutata calcolando la somma dei quadrati degli errori di previsione (devianza = 22). Ora che abbiamo adattato un modello di regressione semplice utilizzando l'ampiezza del nucleo familiare, ci chiediamo se questo spiega meglio la variabilità dei dati rispetto a quanto abbia fatto la media. Sappiamo che in qualche modo esso è migliore, poiché la somma dei quadrati degli errori è ora 5.50. Possiamo quindi valutare la bontà dell'adattamento del modello ai dati esaminando l'entità di questo miglioramento.

Somma dei quadrati degli errori (previsione con la media)	DEV_{Totale}	o *SST*	22.0
– Somma dei quadrati degli errori (regressione semplice)	DEV_{Errore}	o *SSR*	–5.5
Somma dei quadrati spiegati dal modello (regressione semplice)	$DEV_{\text{Regressione}}$	o *SSM*	16.5

Pertanto, una devianza di regressione pari a 16.5 può essere giustificata con il passaggio dal modello senza variabili indipendenti al modello di regressione semplice con la variabile indipendente ampiezza della famiglia. Tale cambiamento rappresenta un miglioramento del 75% (16.5 ÷ 22 = 0.75) rispetto al modello senza variabili indipendenti. Un modo alternativo per esprimere il livello di accuratezza della previsione è dato dal **coefficiente di determinazione lineare** (***R*²**), calcolato rapportando la devianza di regressione alla devianza totale, come mostrato nella seguente equazione:

$$\text{Coefficiente di determinazione lineare}\,(R^2) = \frac{\text{Devianza di regressione}}{\text{Devianza totale}}$$

Se il modello di regressione con la variabile ampiezza del nucleo familiare prevedesse perfettamente il numero di carte di credito posseduto da ciascuna famiglia, si avrebbe $R^2 = 1.0$. Se invece l'uso dell'ampiezza del nucleo familiare non consentisse una migliore previsione rispetto a quella ottenuta con la media, si avrebbe $R^2 = 0$. Quando l'equazione di regressione contiene più di una variabile indipendente, il valore di R^2 rappresenta l'effetto combinato dell'intera equazione sulla previsione. Tale valore è semplicemente il quadrato della correlazione tra i valori veri e quelli previsti.

Quando il coefficiente di correlazione (r) viene utilizzato per valutare la relazione tra le variabili dipendente e indipendente, il suo segno ($+r$, $-r$) denota la pendenza della retta di regressione. In ogni caso, la "forza" della relazione lineare tra le variabili è rappresentata da R^2, che naturalmente è sempre positivo. Nel nostro esempio, $R^2 = 0.75$, indicando che il 75% della variabilità della variabile dipendente è spiegato dalla indipendente. Si osservi che, quando si parla di variabilità della variabile dipendente, ci si riferisce sempre alla devianza totale, misura di variabilità che l'analisi di regressione cerca di spiegare attraverso una o più variabili indipendenti.

9. Previsione con diverse variabili indipendenti (la Regressione Multipla)

È stato precedenza dimostrato come la regressione semplice aiuti a migliorare la previsione. Avvalendosi dei dati sull'ampiezza del nucleo familiare, il numero di carte di credito usate da ciascuna famiglia è stato stimato in modo più accurato di quanto era stato fatto attraverso la sola media aritmetica. Questo risultato induce a pensare che si potrebbe affinare ulteriormente la previsione usando dati addizionali ottenuti sulle famiglie. Ci si può chiedere allora: la nostra previsione migliorerebbe sfruttando non soltanto le informazioni sull'ampiezza del nucleo familiare ma anche dati relativi ad un'altra variabile, come il reddito o il numero di automobili possedute dalla famiglia?

10. L'effetto della multicollinearità

L'inserimento di una ulteriore variabile indipendente può migliorare la previsione della variabile dipendente in relazione non soltanto alla loro correlazione ma anche alla correlazione esistente tra la variabile indipendente addizionale e le variabili indipendenti già inserite nell'equazione di regressione. Per **collinearità** si intende

l'associazione, misurata dalla correlazione, tra due variabili indipendenti. Con **multicollinearità** ci si riferisce invece alla correlazione fra tre o più variabili indipendenti (evidenziata quando una di queste è posta in una relazione di dipendenza lineare con le altre). Benché ci sia una precisa distinzione tra i due termini statistici, è pratica comune utilizzarli in modo intercambiabile.

L'effetto della multicollinearità consiste nel ridurre la capacità previsiva di ogni singola variabile indipendente in modo proporzionale alla forza della sua associazione con le altre variabili indipendenti. Al crescere della collinearità, decresce la varianza spiegata da ogni singola variabile indipendente mentre aumenta la frazione di variabilità spiegata, collettivamente, da tutte le variabili. Poiché la capacità previsiva dovuta all'effetto di tutte le variabili indipendenti può essere contata soltanto una volta, quando si inseriscono altre variabili indipendenti con forte collinearità, la capacità di previsione totale del modello aumenta molto più lentamente. Per massimizzare la bontà della previsione ottenuta con un dato numero di variabili indipendenti, il ricercatore dovrebbe utilizzare quelle con bassa multicollinearità, che però abbiano un'alta correlazione con la variabile dipendente.

11. L'equazione di Regressione Multipla

Per migliorare ulteriormente la previsione sul possesso di carte di credito, proviamo a utilizzare i dati addizionali osservati sulle otto famiglie. La seconda variabile indipendente da includere nel modello di regressione è il reddito della famiglia (V_2), che dopo la variabile ampiezza del nucleo familiare ha la più alta correlazione con la variabile dipendente. Benché V_2 abbia un certo grado di correlazione con V_1, già inserita nell'equazione, rimane la miglior variabile da includere nel modello, perché V_3 ha una correlazione con la variabile dipendente più bassa. Pertanto, il precedente modello di regressione semplice può essere migliorato con l'uso di due variabili indipendenti, così come mostrato di seguito:

$$\text{Numero previsto di carte di credito usate} = b_0 + b_1V_1 + b_2V_2 + e$$

dove

b_0	=	numero costante di carte di credito indipendente dall'ampiezza e dal reddito familiare
b_1	=	variazione nel numero di carte di credito usate associata ad una unità di variazione nell'ampiezza della famiglia
b_2	=	variazione nel numero di carte di credito usate associata ad una unità di variazione nel reddito della famiglia
V_1	=	ampiezza della famiglia
V_2	=	reddito della famiglia

Stimando questo modello di regressione multipla utilizzando il metodo dei minimi quadrati si ottiene una costante pari a 0.482 e coefficienti di regressione per V_1 e V_2 rispettivamente pari a 0.63 e 0.216. Di nuovo, possiamo calcolare i residui del modello prevedendo Y e sottraendo la stima dal valore vero. Quindi, eleviamo al quadrato gli errori di previsione risultanti, come mostrato nella Tabella 6. La somma dei quadrati degli errori commessi utilizzando il modello di regressione multipla con l'ampiezza e il reddito della famiglia è pari a 3.04. Tale valore può essere comparato con quello ottenuto con il modello di regressione semplice, pari a 5.50 (si veda la Tabella 5), che utilizzava come variabile indipendente soltanto l'ampiezza del nucleo familiare.

Quando anche il reddito della famiglia è utilizzato nell'analisi della regressione, R^2 aumenta fino a 0.86.

$$R^2_{\text{(ampiezza della famiglia + reddito della famiglia)}} = \frac{22.0\text{-}3.04}{22.0} = \frac{18.96}{22.0} = 0.86$$

Ciò significa che l'inserimento del reddito familiare nel modello di regressione migliora la previsione dell'11% (0.86 – 0.75), per merito dell'aumentata capacità previsiva dovuta all'apporto della variabile reddito della famiglia.

TABELLA 6 Risultati della regressione multipla usando come variabili indipendenti l'ampiezza del nucleo familiare e il reddito della famiglia

Equazione di regressione: $Y = b_0 + b_1V_1 + b_2V_2$
Equazione di previsione: $Y = 0.482 + 0.63V_1 + 0.216V_2$

Numero identificativo della famiglia	*Numero di Carte di Credito usate*	*Ampiezza della famiglia* (V_1)	*Reddito della famiglia* (V_2)	*Previsione con la regressione multipla*	*Errore di previsione*	*Errore di previsione al quadrato*
1	4	2	14	4.76	–0.76	0.58
2	6	2	16	5.20	0.80	0.64
3	6	4	14	6.03	–0.03	0.00
4	7	4	17	6.68	0.32	0.10
5	8	5	18	7.53	0.47	0.22
6	7	5	21	8.18	–1.18	1.39
7	8	6	17	7.95	0.05	0.00
8	10	6	25	9.67	0.33	0.11

Totale						3.04

12. L'inserimento di una terza variabile indipendente

Abbiamo visto che passando dal modello di regressione semplice a quello multiplo si ottiene un miglioramento nell'accuratezza della previsione, ma abbiamo anche notato che l'aggiunta di variabili indipendenti diventa, ad un certo punto, meno vantaggiosa e in alcune circostanze addirittura controproducente. Nell'indagine sull'uso di carte di credito è possibile inserire nel modello di regressione multipla un'ulteriore variabile, il numero di automobili possedute (V_3). Se ora specifichiamo l'equazione di regressione includendo tutte e tre le variabili indipendenti, possiamo osservare un certo miglioramento nella bontà della previsione, non simile, però, a quello osservato in precedenza. Il valore di R^2 cresce a 0.87, aumentando soltanto di 0.01 rispetto al modello di regressione antecedente. In più, come abbiamo visto nel paragrafo appena concluso, il coefficiente di regressione di V_3 non è statisticamente significativo. Di conseguenza, in questo caso, per fare previsioni è preferibile impiegare il modello di regressione multipla con due variabili indipendenti (ampiezza e reddito della famiglia) senza utilizzare la terza variabile indipendente (numero di automobili possedute).

13. Sommario

L'analisi della regressione è una tecnica statistica semplice e potente che studia relazioni di dipendenza tra variabili e che può fornire al ricercatore previsioni e interpretazioni del fenomeno in analisi. L'esempio illustrato nei paragrafi precedenti ha spiegato i concetti basilari e le procedure che sottostanno all'analisi della regressione, nel tentativo di sviluppare una migliore comprensione del significato e delle problematiche di questa tecnica, nelle sue forme principali. Tutti questi temi verranno approfonditi in dettaglio nel seguito della trattazione, che fornirà inoltre gli elementi indispensabili per affrontare il processo decisionale che consente di adottare l'analisi della regressione negli appropriati problemi di ricerca.

14. Un processo decisionale per l'Analisi della Regressione Multipla

Abbiamo analizzato in precedenza esempi di regressione semplice e multipla. Si è mostrato, in tali esempi, come diversi fattori influenzino la possibilità di trovare il modello di regressione migliore. Tuttavia, fino a questo punto le problematiche esaminate sono state descritte considerandone gli aspetti essenziali, senza tener conto del loro effetto complessivo sull'analisi della regressione

multipla. Nei paragrafi seguenti si illustrerà un processo a sei fasi per la costruzione di un modello di regressione, che verrà utilizzato come una sorta di cornice per la discussione dei fattori che influenzano la creazione, la stima, l'interpretazione e la validazione dell'analisi della regressione. Il processo comincia con la specificazione degli obiettivi dell'analisi di regressione, che determinano la scelta della variabile dipendente e delle variabili indipendenti. Quindi procede con la definizione del modello, che dipende da diversi fattori, tra i quali, ad esempio, la dimensione campionaria e la necessità di utilizzare trasformazioni delle variabili. Una volta formulato il modello, tutte le assunzioni che sottostanno all'analisi della regressione vengono testate su ciascuna variabile. Se le assunzioni sono soddisfatte, il modello viene stimato. Dopo aver ottenuto i risultati, si prosegue con strumenti diagnostici per controllare che il modello soddisfi le assunzioni della regressione e che nessuna osservazione abbia un'influenza particolarmente elevata sui risultati. Il passo successivo consiste nell'interpretazione dell'equazione di regressione e si effettua esaminando il ruolo giocato da ciascuna variabile indipendente nella previsione della variabile dipendente. Infine, i risultati vengono validati per consentire la loro generalizzazione alla popolazione. Le Figure 1 e 6 rappresentano rispettivamente le fasi 1-3 e 4-6 attraverso una rappresentazione grafica del processo di costruzione di un modello di regressione multipla, mentre i paragrafi seguenti descrivono ciascuna fase nel dettaglio.

15. FASE 1: Obiettivi della Regressione Multipla

L'analisi della regressione multipla, che rappresenta una tra le possibili forme dei modelli lineari generalizzati, è una tecnica statistica utilizzata per esaminare la relazione tra una singola variabile dipendente e un insieme di variabili indipendenti. Il necessario punto di partenza dell'analisi, così come di tutte le altre tecniche statistiche multivariate, è il quesito della ricerca. Benché le sue caratteristiche di flessibilità e adattabilità ne permettano l'adozione per lo studio di pressoché tutte le relazioni di dipendenza, per selezionare i casi in cui la regressione multipla è la tecnica di analisi più opportuna, il ricercatore deve considerare tre questioni primarie: (1) l'adeguatezza del problema indagato dalla ricerca, (2) la specificazione di una relazione statistica, (3) la selezione della variabile dipendente e di quelle indipendenti.

16. Problemi di ricerca appropriati per la Regressione Multipla

La regressione multipla è una delle tecniche multivariate più utilizzate. Per la sua ampia applicabilità, essa è stata usata per molti scopi. La maggior parte dei problemi di ricerca affrontati con la regressione multipla cade in due ampie classi: previsione e spiegazione di un fenomeno. Ma i due problemi di ricerca non sono mutuamente esclusivi e l'uso di questa tecnica può essere indirizzato verso una delle due problematiche o verso entrambe.

17. La previsione con la Regressione Multipla

Uno degli obiettivi fondamentali della regressione multipla è prevedere la variabile dipendente attraverso un insieme di variabili indipendenti. Per farlo, la regressione multipla deve rendere massimizza la capacità previsiva delle variabili indipendenti, considerate globalmente così come rappresentate nell'equazione. La loro combinazione lineare è costruita in modo tale da essere il miglior predittore della variabile dipendente. La regressione multipla fornisce uno strumento oggettivo per valutare la capacità previsiva di un insieme di variabili indipendenti. Nelle applicazioni della tecnica focalizzate su questo obiettivo, il ricercatore è interessato primariamente all'ottenimento della previsione migliore: la regressione multipla permette allora di sfruttare molte opzioni, sia nella forma, sia nella specificazione delle variabili indipendenti, che consentono di modificare l'equazione in modo tale da accrescere la sua capacità previsiva. Spesso, però, questa viene potenziata a scapito dell'interpretazione. Un esempio tipico è dato da una variante della regressione, l'analisi delle serie temporali, in cui l'unico scopo è la previsione, mentre l'interpretazione dei risultati è utile soltanto come strumento per migliorare l'accuratezza previsiva del modello. In altre situazioni, la capacità previsiva è cruciale per assicurare la validità dell'insieme delle variabili indipendenti, permettendo così la conseguente interpretazione dell'equazione di regressione. Oltre a misure della capacità previsiva di un modello, esistono test statistici per il controllo della sua significatività. In ogni caso, sia o non sia la previsione l'obiettivo principale, l'analisi della regressione deve raggiungere livelli accettabili di accuratezza previsiva per giustificare il suo utilizzo. Il ricercatore deve comunque preoccuparsi di tenere in

considerazione entrambe le significatività, quella statistica così come quella che garantisce l'utilità pratica dei risultati (si veda la discussione della Fase 4).

La regressione multipla può perseguire anche un secondo obiettivo, che consiste nel confrontare due o più insiemi di variabili indipendenti per valutare la capacità previsiva di ogni equazione di regressione. Esempio di un approccio confermativo alla ricerca di modelli, questa utilizzazione della regressione multipla riguarda la comparazione dei risultati ottenuti con due o più modelli alternativi o in competizione. Lo scopo principale di questo tipo di analisi è la valutazione della capacità previsiva relativa dei modelli, benché in qualche caso la previsione di quello selezionato debba dimostrare significatività statistica e pratica.

18. La spiegazione con la Regressione Multipla

La regressione multipla fornisce inoltre uno strumento di valutazione oggettiva del grado e delle caratteristiche della relazione tra la variabile dipendente e le variabili indipendenti, attraverso la costruzione dell'equazione di regressione. Oltre alla capacità previsiva che le variabili indipendenti possiedono collettivamente, esse possono essere valutate per il loro contributo individuale alla previsione. L'interpretazione dell'equazione può essere fondata su uno qualsiasi dei tre seguenti aspetti: l'importanza delle variabili indipendenti, le tipologie di relazioni osservate, le interazioni tra le variabili indipendenti.

L'interpretazione più diretta dell'equazione di regressione ha attinenza con la determinazione dell'importanza relativa di ogni variabile indipendente nella previsione della dipendente. In tutte le applicazioni, la selezione delle variabili da inserire nel modello dovrebbe essere basata sulla relazione che dal punto di vista teorico dovrebbe esistere tra ciascuna variabile indipendente e la variabile dipendente. L'analisi della regressione consente poi di valutare oggettivamente l'importanza e la direzione (positiva o negativa) di queste relazioni. La caratteristica della regressione multipla che la differenzia dalla sua forma univariata, sta nella possibilità di valutare simultaneamente le relazioni che legano la variabile dipendente ad ogni variabile indipendente. Facendo questa valutazione simultanea, viene determinata anche l'importanza relativa di ciascuna variabile indipendente.

Oltre a mettere sotto controllo l'importanza di ogni variabile, la regressione multipla fornisce al ricercatore anche uno strumento per valutare la natura della relazione tra la variabile dipendente e le variabili indipendenti. La relazione assunta dal modello è di tipo lineare, basata sulla correlazione esistente tra le variabili indipendenti e quella dipendente. Si possono però utilizzare trasformazioni o variabili addizionali per controllare l'eventuale esistenza di altri tipi di relazione, in particolare curvilinee. Questa flessibilità assicura che il ricercatore possa esaminare la vera natura del legame tra le variabili, al di là della relazione lineare assunta dal modello.

Infine, la regressione multipla permette la comprensione delle relazioni esistenti tra le variabili indipendenti in funzione del loro effetto sulla previsione. Queste interazioni sono importanti per due ragioni. Primo, l'esistenza di correlazione tra le variabili indipendenti può renderne alcune ridondanti a fini previsivi, e pertanto non necessarie per produrre la previsione ottimale. Una tale eventualità non rifletterebbe la loro relazione individuale con la variabile dipendente, ma indicherebbe che nel contesto multivariato esse non sono necessarie se, per spiegare la varianza del fenomeno, viene impiegato un altro insieme di variabili indipendenti. Il ricercatore deve guardarsi perciò dal determinare l'importanza delle variabili indipendenti soltanto in base alla risultante equazione di regressione, poiché i legami esistenti tra le variabili indipendenti possono "mascherare" relazioni non necessarie per scopi previsivi, che tuttavia costituiscono importanti risultanze. Le interazioni tra le variabili possono avere influenza non soltanto sulla loro capacità previsiva, ma anche sulla stima dei loro effetti. Questa osservazione è facilmente constatabile ad esempio quando l'effetto di una variabile indipendente è condizionato ad un'altra variabile indipendente. La regressione multipla offre la possibilità di condurre analisi diagnostiche per determinare se tali effetti esistono, basandosi su fondamenti logici, empirici o teorici. Indicazioni di forte interazione (multicollinearità) tra le variabili indipendenti possono suggerire l'uso di scale aggregate.

19. Specificazione di una relazione statistica

La regressione multipla è appropriata quando il ricercatore è interessato all'analisi di una relazione di tipo statistico, non di tipo deterministico. Ad esempio, si esamini la relazione seguente:

Costo totale = Costo variabile + Costo fisso

Se il costo variabile e il costo fisso sono rispettivamente, per unità, 2 dollari ($) e 500$, se produciamo 100 unità il costo totale sarà esattamente pari a 700$, ed ogni deviazione da questi 700$ è dovuta a un nostro errore nella valutazione dei costi, dato che la loro relazione è fissata. Una relazione di questo tipo è definita *relazione funzionale*, poiché non ci si attende alcun errore di previsione.

Ma nell'esempio precedente, con dati che rappresentavano un aspetto del comportamento umano, avevamo assunto che la nostra descrizione dell'uso di carte di credito era solo approssimata e non costituiva una previsione perfetta del comportamento reale. Era stata definita perciò come una **relazione statistica**, poiché nella relazione esaminata era sempre presente una componente casuale. In una relazione statistica, per ogni valore di una variabile indipendente si osservano generalmente più valori della variabile dipendente. Si assume che quest'ultima sia una variabile aleatoria, e pertanto possiamo solo sperare di stimare il suo valore medio in corrispondenza di un valore assunto da una data variabile indipendente. Ad ulteriore chiarimento, nell'esempio sulla regressione semplice avevamo trovato due famiglie con due membri, due con quattro membri, e così via, che possedevano un diverso numero di carte di credito. Le due famiglie con quattro membri avevano in media 6.5 carte di credito, e la nostra previsione era di 6.75. Questa valutazione non era accurata quanto avremmo voluto, ma era comunque migliore di quella ottenuta usando la media su tutto il collettivo, pari a 7 carte di credito. In una relazione di tal genere si assume che l'errore di previsione commesso è interpretabile come il risultato del comportamento casuale dei possessori di carte di credito.

In breve, una relazione funzionale calcola un valore esatto, mentre una relazione statistica stima un valore medio. Per tutta la trattazione, parleremo sempre di relazioni statistiche. Entrambe le tipologie di relazioni sono mostrate in Figura 2.

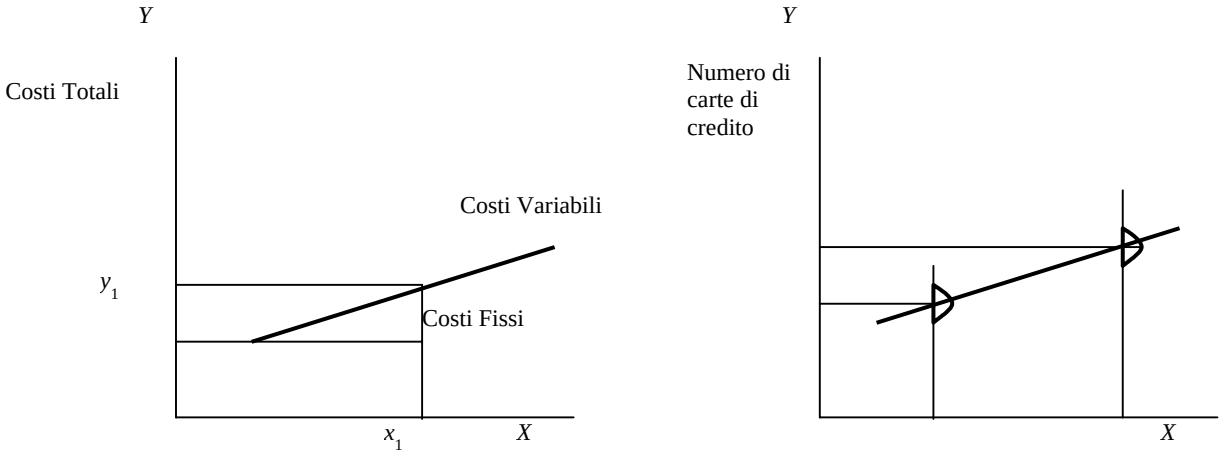


FIGURA 2 Confronto tra relazione funzionale e relazione statistica

20 Alcuni risultati teorici. Immaginiamo che fra variabili y e alcune variabili indipendenti $x_1,\ldots,x_k$ sussista la relazione lineare

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki} + \varepsilon_i$$

con ε_i che è una variabile casuale con media nulla e varianza costante, qualsiasi i.
per $i = 1,2,\ldots,n$.

Se sono verificate le seguenti condizioni

1) $y_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki} + \varepsilon_i$ è il modello "vero", ovvero non vi sono variabili omesse né variabili ridondanti, e la relazione è effettivamente linare;
2) tutte le variabili indipendenti sono deterministiche;
3) gli errori casuali sono indipendenti;
4) la varianza dell'errore casuale è costante;
5) nessuna variabile indipendente è combinazione lineare delle altre

Nota che le assunzioni sul termine d'errore equivalgono a dire che gli ε_i sono indipendenti e

identicamente distribuiti.
possiamo utilizzare il metodo dei minimi quadrati ordinari per stimare i coefficienti β_0 ,β_1,..,β_k.

Le stime $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ ottenute dall'impiego del metodo dei minimi quadrati hanno le seguenti proprietà:

a) sono funzioni lineari di y_1, ,y_n ;
b) sono corrette;
c) hanno varianza minima (per il teorema di Gauss-Markov).

La varianza dell'errore σ^2_ε viene stimata da s^2 che è dato da

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - y_i^*)^2}{n-k-1}$$

dove y_i^* è il valore riprodotto dal modello

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_k x_{ki}$$

Si dimostra che il valore atteso di s^2

$$E(s^2) = \sigma^2_\varepsilon$$

Si dimostra inoltre che la varianza delle stime campionarie ottenute con il metodo dei minimi quadrati è data da:

$$Var(\hat{\beta}_h) = \frac{s^2}{X'X}$$

dove β_h è il generico coefficiente della variabili h-ma e $X'X$ è la matrice della varianze-covarianze delle k variabili indipendenti.

Se assumiamo che gli errori casuali oltre ad essere i.i.d. seguano anche una distribuzione normale, ne segue necessariamente che anche le stime dei coefficienti seguono una distribuzione normale

$$\hat{\beta}_h = N(\beta_h, Var(\beta_h)).$$

Per sottoporre a verifica l'ipotesi nulla
$H_0 : \beta_h = c$

si calcola la statistica test

$$t = \frac{\hat{\beta}_h - c}{Var(\hat{\beta}_h)}$$

che segue la distribuzione di una *t* di Student con *n-k-1* gradi di libertà.

Sempre sotto l'ipotesi di normalità degli errori, per sottoporre a verifica l'ipotesi nulla che tutti i coefficienti del modello di regressione siano uguali a zero

$$H_0 = \beta_1 = \beta_2 = ... = \beta_k = 0,$$

si calcola la statistica test F di Fisher-Snedecor, con *k* e *n-k-1* gradi di libertà, che ha la seguente formula

$$F_{k,n-k-1} = \frac{SSR / k}{SSM / n-k-1}$$

dove

SSR è la somma degli errori stimati al quadrato e *SSM* è la devianza di *y* spiegata dal modello di regressione:

$$SSM = \sum_{i=1}^{n} (y_i^* - \bar{y})^2 .$$

21. Selezione delle variabili dipendente e indipendenti

La "riuscita" finale di ogni tecnica multivariata, e quindi della regressione multipla, comincia con la selezione delle variabili da utilizzare nell'analisi. Poiché la regressione multipla è una tecnica che studia relazioni di dipendenza, il ricercatore deve specificare qual è la variabile dipendente e quali sono quelle indipendenti. La selezione di entrambi i tipi di variabili dovrebbe essere basata principalmente su argomentazioni concettuali e teoriche. È il ricercatore che deve prendere le decisioni fondamentali per la scelta delle variabili, anche se sono disponibili molte opzioni e programmi che possono aiutarlo nella stima del modello. Se il ricercatore non seleziona le variabili ma invece (1) le inserisce nel modello indiscriminatamente oppure (2) le seleziona soltanto in base a considerazioni empiriche, vengono violati diversi principi fondamentali per lo sviluppo del modello.

La selezione di una variabile indipendente è spesso guidata dall'oggetto della ricerca. Ma in tutti i casi, il ricercatore dev'essere consapevole dell'esistenza dell'**errore di misura**, specialmente per la variabile dipendente. L'errore di misura è connesso alla capacità di individuare una variabile che rappresenti una misura accurata e coerente del concetto in studio. Se la variabile utilizzata come dipendente è affetta da un importante errore di misura, allora anche la miglior variabile indipendente non è in grado di raggiungere livelli accettabili di accuratezza previsiva. L'errore di misura può essere generato da varie fonti. Quando è problematico, può essere affrontato con l'uso di scale aggregate. Il ricercatore deve comunque cercare di ottenere sempre la miglior misura, dipendente e indipendente, ricorrendo a considerazioni concettuali ed empiriche.

Nella selezione delle variabili indipendenti, la questione maggiormente problematica è **l'errore di specificazione**, connesso all'inclusione di variabili irrilevanti o all'omissione di variabili rilevanti dall'insieme delle variabili indipendenti. Anche se l'inclusione di variabili irrilevanti non distorce i risultati relativi alle altre variabili, ha però un certo impatto su di loro. In primo luogo, riduce la parsimonia del modello, aspetto che può rivelarsi critico per l'interpretazione dei risultati. In secondo luogo, l'effetto delle variabili addizionali può mascherare o sostituirsi all'effetto di variabili più utili, in particolar modo se viene utilizzata una forma di stima sequenziale del modello. Infine, le variabili addizionali possono rendere meno precisi i test sulla significatività statistica delle variabili indipendenti e insieme ridurre la significatività statistica e pratica dell'analisi.

Dati i problemi associati all'aggiunta di variabili irrilevanti, ci si può chiedere se il ricercatore debba preoccuparsi dell'esclusione di variabili rilevanti. La risposta è certamente sì, poiché l'esclusione di tali variabili può produrre serie distorsioni sui risultati e influenzare negativamente la loro interpretazione. Nella situazione più semplice, le variabili omesse non sono correlate con quelle incluse nel modello, e il solo effetto dell'omissione consiste nella riduzione

dell'accuratezza previsiva dell'analisi. Ma quando una correlazione tra le variabili incluse ed escluse esiste, l'effetto di quelle inserite nel modello è tanto più distorto quanto maggiore è la correlazione con le variabili omesse. Maggiore la correlazione, maggiore la distorsione. Gli effetti stimati per le variabili incluse nel modello di regressione rappresentano ora non soltanto i loro effetti reali ma anche quelli condivisi con le variabili non inserite. Ciò può portare seri problemi per l'interpretazione del modello e la valutazione della significatività statistica e pratica dei risultati.

Il ricercatore deve porre grande attenzione alla selezione delle variabili, per evitare entrambi i tipi di errore di specificazione. Forse è più grave l'omissione di variabili rilevanti, poiché il loro effetto non può essere valutato senza il loro inserimento nell'equazione di regressione. In ogni caso, la necessità di un adeguato supporto teorico e pratico per la soluzione del problema della selezione delle variabili in un modello di regressione multipla si evidenzia comunque.

L'errore di misura influenza anche le variabili indipendenti, riducendo la loro capacità previsiva quanto più esso aumenta. La regressione multipla non dispone di strumenti diretti per la correzione degli errori di misura per le variabili dipendente e indipendenti, anche se l'intensità dell'errore è nota. Se il ricercatore sospetta che un errore di misura può avere effetti problematici, è opportuno che consideri il ricorso alle equazioni strutturali come strumento per tener conto dell'errore di misura nella stima degli effetti delle variabili indipendenti.

22. FASE 2: il disegno della ricerca nell'analisi della Regressione Multipla

Nel disegno dell'analisi della regressione multipla, il ricercatore deve tener conto di problematiche quali: l'ampiezza campionaria, la natura delle variabili indipendenti, la possibilità di creare nuove variabili per rappresentare relazioni particolari tra la variabile dipendente e le variabili indipendenti. Se si preoccupa si esaminare tali problematiche, i criteri per la significatività statistica e pratica del modello saranno certamente soddisfatti. La capacità della regressione multipla di analizzare molti quesiti di ricerca è fortemente influenzata dal disegno della ricerca stessa, la cui problematica va indubbiamente affrontata.

23. La dimensione del campione

Nel disegno della ricerca, l'ampiezza campionaria da utilizzare nella regressione multipla è forse l'elemento più importante sotto il controllo diretto del ricercatore. I suoi effetti si osservano immediatamente nella potenza statistica dei test d'ipotesi e nella generalizzazione dei risultati. Entrambe le tematiche verranno affrontate nei prossimi paragrafi.

24. Potenza statistica e dimensione campionaria

La dimensione del campione ha un impatto diretto sull'opportunità di utilizzare la regressione multipla e sulla sua potenza statistica. Piccoli campioni, solitamente caratterizzati da numerosità inferiori alle 20 osservazioni, sono appropriati soltanto per l'analisi della regressione semplice. Ma anche in questa circostanza, soltanto relazioni lineari molto forti possono essere individuate con un certo grado di fiducia. Parimenti, grandi campioni di 1000 osservazioni e più rendono i test statistici troppo sensibili, agendo in modo tale da far risultare significatività statistiche anche in assenza di relazioni lineari particolarmente forti. Con campioni molto grandi il ricercatore deve assicurarsi perciò che il criterio della significatività pratica sia soddisfatto quanto quello della significativa statistica.

TABELLA 7 Minimo R^2 statisticamente significativo con una potenza di 0.80 per diversi numeri di variabili indipendenti inseriti nel modello e varie dimensioni campionarie

	Livello di significatività (α) = 0.01 *N° di variabili indipendenti*				*Livello di significatività (α) = 0.05* *N° di variabili indipendenti*			
Dimensione campionaria	*2*	*5*	*10*	*20*	*2*	*5*	*10*	*20*
20	45	56	71	NA	39	48	64	NA
50	23	29	36	49	19	23	29	42
100	13	16	20	26	10	12	15	21

250	5	7	8	11		4	5	6	8
500	3	3	4	6		3	4	5	9
1000	1	2	2	3		1	1	2	2

NA = non applicabile

Nella regressione multipla, la **potenza** si riferisce alla probabilità di individuare un dato valore di R^2 o un coefficiente di regressione significativi (in termini statistici) ad uno specificato livello di significatività e per una data numerosità campionaria. La dimensione del campione ha un impatto diretto e considerevole sulla potenza. La Tabella 7 illustra il gioco congiunto della numerosità campionaria, del livello di significatività (α) prescelto e del numero di variabili indipendenti incluse nel modello, per l'ottenimento di un R^2 significativo. All'interno della tavola si possono leggere i valori minimi di R^2 che risultano statisticamente significativi al livello di confidenza α e con probabilità (potenza) 0.80, quando si dispone di un campione con dimensione pari a quella specificata. Per esempio, se il ricercatore impiega cinque variabili indipendenti, specifica un livello di significativà 0.05 ed è soddisfatto quando ottiene un R^2 significativo l'80% delle volte (ciò corrisponde ad una probabilità di 0.80), un campione di 50 rispondenti produrrà valori di R^2 pari al 23% e oltre. Se la numerosità campionaria aumenta fino a 100 rispondenti, si individueranno valori di R^2 significativi pari o al di sotto del 12%. Ma se sono disponibili soltanto 50 rispondenti, e il ricercatore vuole un livello di significativà di 0.01, l'analisi individuerà valori di R^2 significativi soltanto al di sopra del 29%. Il ricercatore dovrebbe sempre tenere in considerazione l'influenza della dimensione campionaria sui test di significatività, prima di raccogliere i dati. Se ci si attendono relazioni deboli, il ricercatore può esprimere il suo giudizio in merito alla numerosità campionaria necessaria a cogliere tali relazioni, se esse esistono. Ad esempio, la Tabella 7 dimostra che, considerando un modello di regressione contenente fino a 10 variabili indipendenti e un livello di significatività di 0.05, un campione di 100 osservazioni individua R^2 significativi anche per valori del coefficiente piuttosto bassi (dal 10% al 15%). Comunque, nelle stesse circostanze, se l'ampiezza campionaria arriva soltanto a 50, il valore di R^2 minimo che può risultare statisticamente significativo raddoppia. Il ricercatore deve essere sempre consapevole della potenza che ci si attende in ogni analisi della regressione multipla, e deve comprendere quali elementi del disegno della ricerca possono essere cambiati per soddisfare le condizioni che rendono l'analisi accettabile.

Il ricercatore può anche determinare la dimensione campionaria necessaria per individuare gli effetti delle variabili indipendenti singolarmente prese, dato l'effetto atteso della relazione (correlazione), il livello di significatività α e la potenza desiderata. I calcoli relativi sono troppo complessi perché possano essere presentati in questa sede, ma i lettori interessati possono fare riferimento ai testi sull'analisi della potenza o ai programmi informatici che, per un dato problema di ricerca, calcolano la numerosità campionaria e la potenza statistica.

25. Generalizzazione e dimensione campionaria

Oltre al ruolo giocato nella determinazione della potenza statistica, la dimensione del campione influenza anche la possibilità di estendere alla popolazione i risultati ottenuti sul campione, in funzione del rapporto tra le osservazioni e le variabili indipendenti. Una regola generale dice che tale rapporto non dovrebbe mai cadere al di sotto del 5 a 1, e in sostanza che dovremmo disporre di almeno cinque osservazioni per ogni variabile indipendente nell'equazione di regressione. Quando il rapporto è inferiore al 5 a 1, il ricercatore va incontro al rischio di "sovradattare" l'equazione ai dati disponibili, legando eccessivamente il risultato dell'analisi al campione utilizzato e facendo perdere alle conclusioni tratte il valore di generalità. Benché il rapporto minimo sia 5 a 1, il suo livello desiderabile cade tra le 15 e le 20 osservazioni per ogni variabile indipendente. Quando tale livello è raggiunto, se il campione è rappresentativo i risultati dovrebbero essere generalizzabili. Comunque, se viene impiegata una procedura stepwise, il livello raccomandato cresce fino a 50 a 1. Nei casi in cui il campione adottato non soddisfa questi criteri, il ricercatore deve preoccuparsi di validare i risultati per garantire la loro generalizzabilità.

26. Gli effetti fissi e aleatori dei predittori

Gli esempi di modelli di regressione discussi fino a questo punto sono stati costruiti assumendo che i livelli delle variabili indipendenti fossero fissati. Per chiarire questo concetto, si consideri il caso in cui si voglia conoscere l'impatto sulla preferenza di una bevanda di tre livelli di dolcezza: si preparerebbero allora tre diversi lotti e li si farebbe assaggiare ad un campione di consumatori. Si stimerebbe quindi il tasso di preferenza di

ogni bevanda usando il livello di dolcezza come variabile indipendente, fissando perciò il grado zuccherino e analizzando il suo effetto in base ai tre livelli determinati. Non si assumerebbe che i tre lotti di bevande con diversi gradi zuccherini siano un campione casuale estratto dall'insieme di tutti i possibili lotti con un qualsiasi livello di dolcezza. Ebbene, una variabile indipendente è aleatoria quando i suoi livelli sono scelti a caso. Se si utilizza una tale variabile indipendente, l'interesse non è semplicemente nei livelli esaminati, quelli che ha già assunto, ma è piuttosto nella popolazione più ampia dei possibili livelli che essa può assumere, dalla quale abbiamo estratto un campione.

La maggior parte dei modelli di regressione basati su dati campionari sono modelli ad effetti aleatori. Ne costituisce un esempio l'indagine campionaria condotta su una certa popolazione con l'obiettivo di valutare la relazione esistente tra l'età dei rispondenti e la frequenza del ricorso al medico. La variabile indipendente "età dei rispondenti" venne selezionata a caso dalla popolazione in modo da non ottenere risultati validi soltanto per gli individui del campione, ma per fare inferenza sulla relazione tra le variabili nell'intera popolazione in esame.

Per i modelli che utilizzano entrambi i tipi di variabili indipendenti, le procedure di stima rimangono le stesse, ad eccezione che per i termini d'errore poiché nei modelli con effetti aleatori una porzione dell'errore casuale viene dal campionamento delle variabili indipendenti. In ogni caso, le procedure statistiche basate sui modelli con effetti fissi sono piuttosto robuste, così che l'analisi condotta su modelli ad effetti casuali come se fossero modelli ad effetti fissi (e così si comporta la maggior parte dei pacchetti informatici) può essere appropriata con una ragionevole approssimazione.

27. Creazione di variabili addizionali

La relazione fondamentale indagata dalla regressione multipla, che si basa sul coefficiente di correlazione, è l'*associazione* lineare tra una variabile dipendente *metrica* e diverse variabili indipendenti. Un problema spesso affrontato dai ricercatori è l'inserimento nell'equazione di regressione di dati non metrici, come il genere o l'occupazione. Ma, come abbiamo più volte sottolineato, la regressione si limita all'analisi di dati metrici, e tra l'altro non consente di individuare direttamente relazioni non lineari, cosa che costituisce un notevole limite per il ricercatore, soprattutto quando si trova ad affrontare situazioni in cui la teoria suggerisce relazioni non lineari (come ad esempio forme a U) o quando i dati stessi rivelano relazioni di questo tipo.

In casi del genere, è necessario creare nuove variabili attraverso delle **trasformazioni**, poiché per incorporare nel modello variabili non metriche o per rappresentare relazioni diverse da quella lineare la regressione multipla deve completamente affidarsi alla creazione di nuove variabili. La trasformazione dei dati è uno strumento con il quale il ricercatore può modificare la variabile dipendente o le variabili indipendenti con una o due finalità: per migliorare o modificare la relazione tra le variabili, oppure per consentire l'inserimento di variabili non metriche nel modello. Tali trasformazioni possono essere fondate su ragioni sia "teoriche" (nei casi in cui sono rese necessarie per la natura delle variabili) sia "derivate dai dati" (nei casi in cui sono strettamente suggerite da un attento esame dei dati). In ogni caso, il ricercatore deve procedere più volte per tentativi ed errori, valutando in che modo il modello migliora in relazione alle trasformazioni dei dati impiegate. Questi temi verranno qui di seguito approfonditi perché consentono all'analisi della regressione di meglio rappresentare i dati e perché consentono di integrare l'informazione contenuta nelle variabili originarie attraverso la creazione di nuove variabili.

Le trasformazioni che descriveremo sono facilmente implementabili con comandi semplici, disponibili in tutti i pacchetti statistici più diffusi. Si osservi comunque che esistono metodi di trasformazione più sofisticati e complessi.

28. Inserimento di dati non metrici per mezzo di variabili dummy

Una situazione comune che i ricercatori si trovano ad affrontare è la presenza di variabili indipendenti non metriche. Fino a questo punto, però, la nostra illustrazione ha considerato soltanto variabili dipendenti e indipendenti misurate su scala metrica. Se tuttavia la variabile dipendente è dicotomica (0, 1), per condurre uno studio è necessario ricorrere all'analisi discriminante oppure a una forma particolare di regressione (la regressione logistica). Ma cosa possiamo fare quando le variabili indipendenti non sono metriche e si presentano con due o più modalità? In tal caso, si può ricorrere a variabili dicotomiche, note anche come **variabili dummy**, che possono essere utilizzate in sostituzione di variabili indipendenti. Ogni variabile dummy rappresenta una modalità di una variabile indipendente non metrica e ciascuna variabile indipendente non metrica, con k categorie, può essere rappresentata attraverso $k - 1$ variabili dummy.

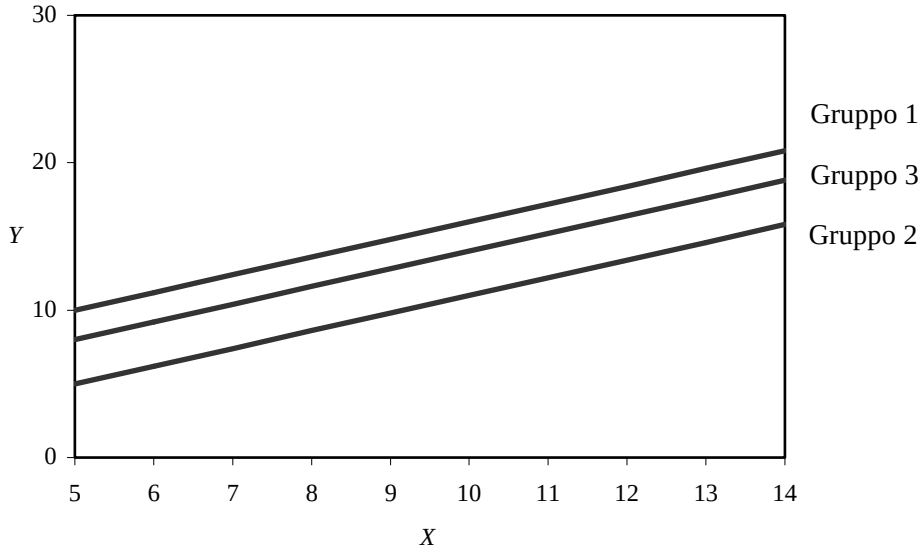
Esistono due forme per la rappresentazione di variabili dummy, ma la più comune è quella in cui ogni categoria è codificata con 1 o con 0. I coefficienti di regressione delle variabili dummy rappresentano le differenze tra le medie,

calcolate sulla variabile dipendente, di ciascun gruppo di rispondenti individuati dalle diverse variabili dummy e la **categoria di riferimento** (vale a dire il gruppo omesso, individuato da tutti gli zeri). Queste differenze tra i gruppi possono essere valutate direttamente poiché i coefficienti di regressione sono espressi nella stessa unità di misura della variabile dipendente. Tale metodo di codifica ottenuto ricorrendo alle variabili dummy può essere descritto anche in relazione alle diverse intercette che le equazioni di regressione dei diversi gruppi si trovano ad assumere (si veda la Figura 3). Nell'esempio illustrato, una variabile non metrica a tre categorie è stata rappresentata con due variabili dummy (D_1 e D_2), che individuano rispettivamente i gruppi 1 e 2, mentre il gruppo 3 rappresenta la **categoria di riferimento**. I coefficienti di regressione sono pari a 2.0 per D_1 e –3.0 per D_2. Tali coefficienti si traducono in tre rette parallele. Il gruppo di riferimento (in questo caso il numero 3) è definito dall'equazione di regressione con entrambe le variabili dummy uguali a zero. La retta del gruppo 1 è al di sopra di due unità rispetto a quella del gruppo di riferimento, mentre quella del gruppo 2 è al di sotto di tre unità rispetto alla retta del gruppo 3. Il parallelismo tra le rette di regressione dimostra che le variabili dummy non modificano la natura della relazione, ma attribuiscono ai diversi gruppi soltanto differenti intercette. Questa forma di codifica è particolarmente appropriata quando esiste un gruppo di riferimento logico, come accade ad esempio nei disegni sperimentali. Ogni volta che si ricorre alla codifica tramite variabili dummy, bisogna essere consapevoli dell'esistenza del gruppo di confronto e ricordare che i coefficienti di regressione rappresentano le differenze tra le medie dei gruppi rispetto a quella del gruppo di riferimento.

Un metodo alternativo di codifica delle variabili dummy è simile al precedente a parte il fatto che sul gruppo di confronto (quello con tutti gli zeri) le variabili dummy assumono il valore –1 invece che 0. In questo modo i coefficienti di regressione rappresentano le differenze tra le medie di ciascun gruppo e la media generale calcolata su tutti i gruppi e non più la differenza rispetto alla media del gruppo di riferimento. Entrambe le forme di codifica forniscono comunque gli stessi risultati, sia per la previsione, sia sul coefficiente di determinazione, sia sui coefficienti di regressione delle variabili continue. L'unica differenza consiste nell'interpretazione dei coefficienti delle variabili dummy.

Equazioni di regressione con variabili dummy (D_1 e D_2)	
Modello specificato	$Y = a + b_1X + b_2D_1 + b_3D_2$
Modello stimato	
Globale	$Y = 2 + 1.2X + 2D_1 - 3D_2$
Specifico per ogni gruppo	
Gruppo 1 ($D_1 = 1$, $D_2 = 0$)	$Y = 2 + 1.2X + 2(1)$
Gruppo 2 ($D_1 = 0$, $D_2 = 1$)	$Y = 2 + 1.2X \quad - 3(1)$
Gruppo 3 ($D_1 = 0$, $D_2 = 0$)	$Y = 2 + 1.2X$

FIGURA 3 Inserimento di variabili non metriche attraverso variabili dummy



29. FASE 3: Controllo delle assunzioni della Regressione Multipla

Abbiamo mostrato come sia possibile migliorare la previsione della variabile dipendente aggiungendo variabili indipendenti ed anche trasformandole per descrivere eventuali aspetti non lineari della relazione. Per farlo, però, sulla relazione che lega la variabile dipendente alle variabili indipendenti dobbiamo porre alcune assunzioni, che influenzano la procedura statistica (minimi quadrati) usata per la stima del modello di regressione multipla. Nei paragrafi seguenti verranno descritte le metodologie appropriate per la verifica di tali assunzioni e le necessarie azioni correttive da attuare quando si presentano delle violazioni.

30. Confronto tra le singole variabili e l'equazione di regressione

Le assunzioni che costituiscono le basi teoriche dell'analisi della regressione multipla interessano sia le variabili singolarmente prese (dipendente e indipendenti) sia la relazione considerata nel suo complesso. Esistono metodi che consentono di valutare la sussistenza delle assunzioni su ciascuna variabile. Ma nella regressione multipla, una volta che l'equazione è stata costruita, essa agisce sulla previsione della variabile dipendente nella sua interezza. Ciò comporta la necessità di valutare l'esistenza delle assunzioni non solo per le singole variabili ma anche per l'equazione stessa. Questo paragrafo esamina l'equazione e la sua relazione con la variabile dipendente, per controllare se soddisfa le assunzioni della regressione multipla. Tale analisi, in realtà, può essere condotta soltanto *dopo* la stima del modello di regressione (Fase quattro). Pertanto, i test sulle assunzioni devono essere previsti non soltanto nelle fasi iniziali della regressione, ma anche dopo che il modello è stato stimato.

Nel calcolo dei coefficienti di regressione e nella previsione della variabile dipendente, la questione centrale è la verifica del fatto che le assunzioni dell'analisi della regressione siano state rispettate. Ci si deve chiedere infatti se gli errori di previsione sono il risultato di una reale assenza di relazione tra le variabili o se sono piuttosto causati da alcune caratteristiche dei dati che il modello di regressione non coglie. Pertanto, le assunzioni da porre sotto controllo sono le seguenti:

- Linearità del fenomeno misurato
- Varianza costante dei termini di errore
- Indipendenza dei termini di errore
- Normalità distributiva dei termini di errore

La misura principale dell'errore di previsione dell'equazione di regressione è il **residuo**, cioè la differenza tra valori osservati e valori previsti per la variabile dipendente. La rappresentazione grafica dei residui, rispetto alle variabili indipendenti o rispetto alla previsione, costituisce un metodo fondamentale per individuare violazioni alle assunzioni relative alla relazione complessiva. Quando si ricorre a questo strumento, è raccomandato l'uso di una qualche forma di standardizzazione dei residui, che li renda direttamente confrontabili. (Nella loro forma originale, i valori previsti più alti hanno per natura residui più alti). I più utilizzati sono i **residui studentizzati**,

i cui valori sono analoghi ai valori di *t*. Questa corrispondenza rende piuttosto facile la verifica della significatività statistica dei residui particolarmente ampi. Nei prossimi paragrafi analizzeremo una serie di test statistici che possono completare l'esame visivo dei grafici dei residui.

Il diagramma più comune è quello che vede i residui (r_i) in ordinata e i valori previsti della variabile dipendente in ascissa (Y_i). In un modello di regressione semplice, il grafico dei residui può essere costruito ponendo sull'asse delle ascisse sia la variabile dipendente che le variabili indipendenti, poiché esse sono in relazione diretta. Nella regressione multipla, invece, soltanto i valori previsti della variabile dipendente rappresentano l'effetto complessivo dell'equazione di regressione. Perciò, a meno che l'analista non voglia concentrarsi su una sola variabile, in ascissa vengono utilizzati soltanto i valori previsti. Le violazioni alle assunzioni possono essere identificate attraverso particolari andamenti nei grafici dei residui.

31. La linearità del fenomeno

La **linearità** della relazione tra variabile dipendente e variabili indipendenti rappresenta il grado con cui le variazioni nella dipendente sono associate a variazioni nelle variabili indipendenti: è misurabile considerando di quanto varia in media la variabile dipendente per ogni unità di variazione delle variabili indipendenti, e valutando se le medie delle distribuzioni della variabile dipendente, condizionate ai valori assunti dalle variabili indipendenti, giacciono sulla retta (relazione semplice) o sull'iperpiano (relazione multivariata) di regressione. Ciascun coefficiente di regressione rappresenta una misura di tale concomitanza di variazioni, e rimane costante sull'insieme dei valori assunti dalla variabile indipendente. Poiché il concetto di correlazione è legato a relazioni di tipo lineare, essendo alla base dell'analisi della regressione ciò solleva una questione critica e rende necessario il controllo dell'assunzione di linearità della relazione. Fortunatamente, la linearità è facilmente controllabile attraverso il grafico dei residui. Un deciso andamento curvilineo nel grafico dei residui indica che un'azione correttiva sul modello ne migliorerebbe sia la capacità previsiva sia la validità della stima dei coefficienti. Benché esistano metodi che consentono di riportare i dati alla linearità attraverso delle trasformazioni, il ricercatore può voler includere le relazioni non lineari direttamente nel modello di regressione. In tal caso, trasformazioni dei dati come la creazione di termini polinomiali, descritta nella Fase 2, o metodi particolari come la regressione non lineare, possono cogliere gli effetti curvilinei delle variabili indipendenti o addirittura relazioni non lineari più complesse.

L'esame dei residui, nella regressione multipla, rivela gli effetti combinati di tutte le variabili indipendenti, ma non consente di valutare ciascuna variabile indipendente separatamente dalle altre. Per farlo, è possibile ricorrere ai cosiddetti **grafici di regressione parziale** (nella terminologia inglese, ***partial regression plot***), che mostrano la relazione di ogni singola variabile indipendente con la variabile dipendente. Essi si distinguono dal diagramma dei residui appena discusso per il fatto che la linea che attraversa il baricentro dei punti, orizzontale nei grafici precedenti (Figura 5), ha ora una pendenza verso l'alto o il basso in funzione del segno (positivo o negativo) del coefficiente di regressione della relativa variabile indipendente. L'esame dei residui intorno a questa linea è poi condotto esattamente come in precedenza.

Nei grafici di regressione parziale, *pattern* curvilinei nei residui rivelano una relazione non lineare tra la variabile dipendente e una specifica variabile indipendente. Si tratta del metodo più utile, quando il modello è multivariato, per controllare la sussistenza dell'assunzione di linearità su ciascuna variabile indipendente e, nel caso si osservino violazioni, per capire su quali variabili sia necessario applicare gli appropriati rimedi per superare tali violazioni. Inoltre, poiché esamina una variabile alla volta, il metodo facilita l'identificazione di outlier e di osservazioni che possono avere una particolare influenza sui risultati.

32. Varianza costante del termine d'errore

La presenza di varianze diseguali nei termini d'errore (**eterochedasticità**) è una delle più comuni violazioni alle assunzioni di base e si osserva quando le distribuzioni della variabile dipendente condizionate ai valori assunti dalle variabili indipendenti non hanno varianze identiche. Il suo controllo è effettuato con i grafici dei residui o con semplici test statistici. Se la varianza non è costante, nella rappresentazione grafica dei residui (studentizzati) rispetto ai valori previsti della variabile dipendente e dal confronto con il grafico dei residui casuali si osservano *pattern* particolari. Il più comune è forse quello a forma triangolare, che può essere orientato in entrambe le direzioni. Un *pattern* a forma di diamante si presenta di solito quando si impiegano dati percentuali, in cui generalmente si riscontra maggior variabilità nella zona centrale dell'intervallo di variazione piuttosto che nelle code. Molte volte si

incontrano diversi tipi di violazioni anziché una sola, come mostrato ad esempio nella Figura 5h, che rappresenta un caso di non linearità e insieme di eteroschedasticità.

Ogni pacchetto statistico prevede test per saggiare l'ipotesi di omoschedasticità. SPSS, per esempio, mette a disposizione il test di Levene per l'uguaglianza delle varianze di una coppia di variabili, test particolarmente raccomandato per la sua robustezza rispetto all'allontanamento dalla normalità distributiva delle variabili in esame, problema che si presenta frequentemente nell'analisi della regressione.

Nel caso sia presente eteroschedasticità, si può ricorrere a due soluzioni. Se la violazione può essere attribuita a una sola variabile indipendente, si può impiegare la procedura dei minimi quadrati pesati. In ogni caso, in modo più semplice e diretto, si possono utilizzare trasformazioni dei dati che stabilizzano la varianza e che ci permettono di inserire le variabili trasformate direttamente nel modello di regressione.

33. Indipendenza dei termini d'errore Nell'analisi della regressione si assume che i valori previsti sono tra loro indipendenti. Ma ciò significa che ogni valore previsto non è legato a nessun'altra previsione, e cioè che tutti questi valori non risultano ordinati in base ad alcun criterio. È possibile identificare meglio la situazione descritta rappresentando graficamente i residui rispetto a una variabile sequenziale. Se sono indipendenti, il *pattern* rivelato dal grafico deve risultare simile a quello casuale. In caso contrario, visualizzando particolari andamenti, si identificano violazioni all'assunzione. Un altro *pattern* frequente ricorre quando le condizioni fondamentali del modello cambiano, ma tale cambiamento non viene tenuto in considezione. A titolo esemplificativo, si consideri un modello di regressione per le vendite mensili di costumi da bagno, rilevate per 12 mesi e con due stagioni invernali contro una sola stagione estiva, che non preveda alcun indicatore stagionale. Il *pattern* dei residui di un tale modello mostrerà valori negativi per i mesi invernali e valori positivi per i mesi estivi. Questo tipo di violazione delle assunzioni può essere superata utilizzando trasformazioni dei dati, come ad esempio le differenze prime di una serie storia, oppure includendo variabili indicatrici, oppure ancora ricorrendo a modelli di regressione formulati in modo particolare.

34. Normalità distributiva dei termini d'errore La violazione dell'assunzione che si incontra più frequentemente è forse quella relativa alla normalità distributiva della variabile dipendente o delle variabili indipendenti, o addirittura di entrambi i tipi di variabili. Lo strumento diagnostico più semplice per il controllo di tale assunzione, da adottarsi per l'insieme delle variabili indipendenti, è l'istogramma dei residui, che costituisce una sorta di controllo visivo sull'approssimarsi della loro distribuzione alla curva gaussiana. Benché risulti attraente per la sua semplicità, l'uso di questo metodo è particolarmente difficoltoso quando si dispone di piccoli campioni, sui quali le distribuzioni non sono individuabili nella loro forma reale. Un metodo migliore si ottiene allora ricorrendo al **grafico di probabilità normale** (nella terminologia inglese, ***normal probability plot***), che si differenzia dal diagramma dei residui per il fatto che in questo caso vengono confrontati con la distribuzione normale. Essa è rappresentata con una retta bisettrice del primo quadrante, e costituisce il termine di confronto per i residui raffigurati sul piano cartesiano. Se la distribuzione è normale, la linea dei residui si approssima alla diagonale. La stessa procedura può essere usata per confrontare con la gaussiana sia la variabile dipendente, sia ciascuna variabile indipendente.

35. Sommario L'analisi dei residui, sia tramite grafici sia tramite test statistici, costituisce uno strumento analitico semplice ma potente per esaminare la corrispondenza del modello alle assunzioni della regressione. Tuttavia, però, troppo spesso queste analisi non sono condotte e le violazioni alle assunzioni vengono lasciate agire. Quando ciò avviene, gli utilizzatori dei risultati non sono a conoscenza delle potenziali imprecisioni in essi contenute, imprecisioni che possono riguardare vari elementi dell'analisi, dai test di significatività sui coefficienti di regressione (che possono rivelare significatività statistica quando non è presente o viceversa), alle previsioni della variabile dipendente (che possono risultare distorte). Si raccomanda quindi di applicare sempre questi metodi, qualsiasi insieme di dati e qualsiasi modello di regressione sia oggetto di studio. L'uso di rimedi appropriati, nel caso si osservino violazioni alle assunzioni, migliorerà la fiducia nell'interpretazione e nella previsione fornita dalla regressione multipla.

36. FASE 4: Stima del modello di Regressione e valutazione della bontà dell'adattamento ai dati Dopo aver specificato gli obiettivi dell'analisi della regressione, selezionato le variabili dipendente e indipendenti, affrontato le problematiche del

disegno della ricerca, e valutato la rispondenza delle variabili alle assunzioni dell'analisi, il ricercatore a questo punto è pronto per la stima del modello e la conseguente valutazione della capacità previsiva delle variabili indipendenti. In questa fase, egli deve portare a termine tre compiti fondamentali: (1) scegliere un metodo per la stima del modello di regressione, (2) valutare la significatività statistica del modello nella previsione della variabile indipendente, e (3) determinare se qualche osservazione esercita una influenza eccessiva sui risultati dell'analisi.

37. Approccio generale alla selezione delle variabili Nella maggior parte dei casi in cui si utilizza la regressione multipla, il ricercatore dispone di un certo numero di potenziali variabili indipendenti, dal quale selezionare quelle da inserire nel modello. A volte capita che l'insieme delle variabili indipendenti sia ben specificato, e in tal caso il modello di regressione ha una funzione prevalentemente confermativa. In altre situazioni, invece, il ricercatore può voler effettuare una selezione nell'insieme delle variabili indipendenti. Per aiutarlo a trovare il modello di regressione "migliore", esistono allora diverse tecniche (metodi di ricerca sequenziale e processi combinatori), che verranno illustrate nei prossimi paragrafi.

39. Specificazione confermativa Il metodo più semplice, ma forse più richiesto, per la specificazione del modello di regressione è quello che utilizza una prospettiva di tipo confermativo, e ad esso si ricorre nei casi in cui il ricercatore vuole specificare autonomamente l'insieme delle variabili indipendenti da includere nel modello. A confronto con i metodi che verranno discussi di seguito, il ricercatore ha qui il controllo totale sulla selezione delle variabili. Benché la specificazione confermativa sia concettualmente semplice, il suo uso implica che il ricercatore sia certo che l'insieme delle variabili selezionate raggiunga la massima capacità previsiva mantenendo le fondamentali caratteristiche di parsimonia.

40. Uno sguardo generale sui metodi di selezione del modello Che si scelga il metodo confermativo, di ricerca sequenziale o combinatorio, l'aspetto più importante per la buona riuscita dell'analisi è la conoscenza del contesto dello studio di cui è dotato il ricercatore, al quale spetta la scelta del ricorso ad una prospettiva oggettiva per l'inclusione delle variabili oppure del ricorso ai segni attesi e alla grandezza dei coefficienti. Senza questa conoscenza, i risultati della regressione possono avere un'alta accuratezza previsiva ma essere privi di rilevanza empirica o teorica. Il ricercatore non dovrebbe mai affidarsi completamente a nessuno di questi approcci, ma dovrebbe invece usarli dopo averli attentamente considerati e dovrebbe quindi accettare i risultati soltanto dopo un esame scrupoloso.

41. Test sull'equazione di regressione per la valutazione delle assunzioni Dopo aver scelto le variabili da inserire nell'equazione e dopo aver stimato i coefficienti di regressione, il ricercatore deve ora controllare se il modello soddisfa le assunzioni sottostanti l'analisi della regressione multipla. Come già illustrato nella Fase 3, le variabili considerate devono singolarmente rispettare le assunzioni di linearità, varianza costante, indipendenza e normalità. Ma anche l'equazione di regressione deve soddisfare le stesse assunzioni. I test diagnostici discussi nella Fase 3 possono allora essere utilizzati per valutare l'effetto collettivo dell'equazione attraverso l'esame dei residui. Se si osservano violazioni importanti, il ricercatore deve mettere in atto misure correttive e quindi stimare nuovamente il modello di regressione.

42. Esame della significatività statistica del modello scelto Se estraessimo ripetuti campioni di otto famiglie e chiedessimo loro il numero di componenti e di carte di credito utilizzate, rare volte otterremmo esattamente gli stessi valori per $Y = b_0 + b_1 X_1$ da tutti di campioni. Ci aspetteremmo infatti che il caso producesse delle differenze sulle stime ricavate dai diversi campioni. Poiché di solito, come nell'esempio, si utilizza un solo campione e su di esso si

costruisce il modello predittivo, si pone la necessità di testare l'ipotesi che tale modello rappresenti la popolazione di tutte le famiglie che usufruiscono di carte di credito e non soltanto quel campione di otto famiglie. I test che consentono di sottoporre a verifica questa ipotesi possono avere una o due forme principali: una è data dal test sulla varianza spiegata (coefficiente di determinazione) e una dal test sui coefficienti.

43. Significatività del modello complessivo: il coefficiente di determinazione

Per porre sotto controllo l'ipotesi che l'ammontare della variabilità spiegata dal modello di regressione sia maggiore della variabilità spiegata dalla media viene utilizzata la statistica *F*.

Il test *F* per il modello di regressione multipla con due variabili indipendenti è (18.96 ÷ 2)/(3.04 ÷ 5) = 15.59. Anche questo test è statisticamente significativo, e indica perciò che la variabile indipendente aggiunta al modello di regressione ha apportato un contributo importante per il miglioramento della sua capacità previsiva.

Sappiamo però che R^2 è influenzato dal gioco combinato del numero di variabili indipendenti inserite nell'equazione e della numerosità campionaria. A questo proposito, sono state suggerite diverse regole empiriche, dall'uso di un numero di osservazioni variabile da 10 a 15 per ogni variabile indipendente fino ad un valore minimo assoluto di 4 osservazioni per ogni variabile indipendente. Come ci si approssima o si va al di là di questi limiti, diviene necessario correggere R^2 per controllare la sua eccessiva crescita dovuta a una sorta di "sovradattamento" ai dati. Tutti i programmi informatici che effettuano analisi della regressione calcolano perciò, insieme al coefficiente di determinazione, anche il **coefficiente di determinazione lineare corretto** (**R^2 corretto**). Interpretato alla stessa stregua del primo, il valore di R^2 corretto diminuisce quando si riduce il numero osservazioni per variabile indipendente. Tale coefficiente è particolarmente utile per confrontare equazioni di regressione costruite con un diverso numero di variabili indipendenti o calcolate su campioni di diversa ampiezza, poiché tiene conto del numero specifico di variabili considerate nel modello e della dimensione del campione sul quale è stato stimato. Nell'esempio sulle carte di credito, il valore di R^2 per il modello di regressione semplice è 0.751, mentre il valore del coefficiente corretto è 0.709. Aggiungendo la seconda variabile indipendente, R^2 cresce fino a 0.861, mentre R^2 corretto aumenta soltanto fino a 0.806. In entrambe le equazioni, quest'ultimo coefficiente riflette l'andamento del rapporto tra il numero dei coefficienti da stimare e la numerosità del campione e compensa il "sovradattamento" del modello ai dati.

44. Test di significatività sui coefficienti di regressione

Nell'analisi della regressione, quando lo studio è basato su dati ricavati da un campione estratto dalla popolazione piuttosto che da un censimento, è opportuno e necessario condurre test statistici sulla significatività dei coefficienti stimati. Infatti, anche se il ricercatore utilizza dati campionari per la stima di un'equazione di regressione, non è interessato a stimare un modello che valga per quello specifico campione, ma piuttosto che possa essere generalizzato all'intera popolazione. Qualsiasi campione fosse estratto dalla popolazione fornirebbe valori diversi per i coefficienti dell'equazione. Inoltre, con campioni di piccola numerosità, i coefficienti stimati varierebbero ampiamente da campione a campione. Ma al cresce della loro dimensione, i campioni diventano maggiormente rappresentativi della popolazione e la variabilità delle stime prodotte diviene sempre più bassa. Questo è vero fino a quando l'analisi è condotta direttamente sull'intera popolazione. In tal caso, non c'è necessità di effettuare test di significatività poiché il "campione" è esattamente eguale alla popolazione e quindi perfettamente rappresentativo. La variabilità attesa nella stima dei coefficienti dell'equazione (sia nella costante che nei coefficienti di regressione) è definita **errore standard** dei coefficienti.

I test di significatività sui coefficienti di regressione offrono uno strumento statistico fondato sulla probabilità che consente di valutare se i coefficienti stimati su un gran numero di campioni di una certa dimensione si mantengono diversi da zero. Se la numerosità del campione è bassa, l'errore campionario può essere troppo grande per permettere di dire con un certo grado di confidenza (definito *livello di significatività*) che il coefficiente in esame è diverso da zero. Ma al crescere della dimensione del campione cresce anche la precisione del test, poiché la variabilità nei coefficienti diviene più bassa. In ogni caso, grandi campioni non garantiscono che i coefficienti siano diversi da zero, ma piuttosto che il test sia più preciso.

Un esempio di variabilità campionaria per un coefficiente di regressione Per illustrare questo aspetto dell'analisi sono stati estratti 20 campioni casuali di quattro diverse dimensioni (10, 25, 50 e

100 rispondenti) da un grande insieme di dati. Su ciascuno di essi è stata condotta un'analisi della regressione semplice e i coefficienti di regressione stimati sono stati riportati nella Tabella 8. Come possiamo vedere, la variabilità delle stime dei coefficienti è maggiore sui campioni di 10 rispondenti, dove i valori stimati vanno da 2.20 a 6.06. Al crescere della dimensione campionaria fino a 25 e 50 rispondenti, l'errore di campionamento diminuisce considerevolmente. Infine, il campione di 100 rispondenti mostra un intervallo di variazione delle stime che è quasi dimezzato rispetto a quello dei campioni di 10 rispondenti (2.10 contro 3.86). Da tutto questo si comprende come il test statistico che consente di determinare se il coefficiente è realmente diverso da zero è sempre più preciso al crescere della numerosità campionaria.

TABELLA 7 Variabilità campionaria dei coefficienti di regressione stimati

	Dimensione campionaria			
Campione	*10*	*25*	*50*	*100*
1	2.58	2.52	2.97	3.60
2	2.45	2.81	2.91	3.70
3	2.20	3.73	3.58	3.88
4	6.06	5.64	5.00	4.20
5	2.59	4.00	4.08	3.16
6	5.06	3.08	3.89	3.68
7	4.68	2.66	3.07	2.80
8	6.00	4.12	3.65	4.58
9	3.91	4.05	4.62	3.34
10	3.04	3.04	3.68	3.32
11	3.74	3.45	4.04	3.48
12	5.20	4.19	4.43	3.23
13	5.82	4.68	5.20	3.68
14	2.23	3.77	3.99	4.30
15	5.17	4.88	4.76	4.90
16	3.69	3.09	4.02	3.75
17	3.17	3.14	2.91	3.17
18	2.63	3.55	3.72	3.44
19	3.49	5.02	5.85	4.31
20	4.57	3.61	5.12	4.21
Minimo	2.20	2.52	2.91	2.80
Massimo	6.06	5.64	5.85	4.90
Intervallo di variazione	3.86	3.12	2.94	2.10
Deviazione standard	1.28	0.85	0.83	0.54

Test di significatività nell'esempio di regressione semplice Quando abbiamo affrontato il modello di regressione semplice nell'esempio relativo all'uso di carte di credito, abbiamo visto come l'equazione di regressione per il numero di carte di credito fosse $Y = b_0 + b_1 X_1 = 2.87 + 0.971(\text{ampiezza della famiglia})$. Vorremmo ora testare le due ipotesi sui coefficienti di regressione di questo modello (2.87 e 0.971).

Ipotesi 1. Il valore dell'intercetta (termine costante) pari a 2.87 è dovuto all'errore campionario e il vero termine costante appropriato per la popolazione è zero.

Ponendo sotto controllo questa ipotesi, vorremmo semplicemente valutare se il termine costante ha un effetto diverso da zero e se quindi dev'essere inserito nel modello. Se non risulta significativamente diverso da zero, dovremmo decidere di non impiegare la costante a scopi previsivi. Lo strumento appropriato per il controllo di questa ipotesi è il test *t*, che è disponibile su tutti i programmi per l'analisi computerizzata della regressione. Per un qualsiasi coefficiente di un'equazione di regressione il valore di *t* è pari al coefficiente stesso diviso per il suo errore standard. Per esempio, per un coefficiente che vale 2.5 e che ha un errore standard di 0.5, *t* sarà pari a 5.0. Per determinare se il coefficiente è significativamente diverso da zero, il valore di *t* calcolato sui dati campionari viene confrontato con il valore tabulato, scelto in funzione della numerosità del campione e del livello di significatività prescelto. Se il valore calcolato è maggiore di quello tabulato, possiamo ritenere con un certo grado di confidenza (al livello di significatività stabilito) che il coefficiente ha un effetto statisticamente significativo all'interno dell'equazione di regressione.

Lo stesso test dev'essere effettuato anche sull'intercetta. Non è invece necessario se i dati usati per sviluppare il modello non includono osservazioni con valori tutti nulli sulle variabili indipendenti, poiché in tal caso il termine costante non è "contenuto" nei dati ed agisce soltanto sul posizionamento del modello.

Ipotesi 2. Il coefficiente 0.971 indica che una variazione unitaria nell'ampiezza della famiglia è associata a una variazione pari a 0.971 nel numero medio di carte di credito possedute dalla famiglia e che tale coefficiente è significativamente diverso da zero.

Se il valore del coefficiente fosse generato dal solo effetto dell'errore campionario, concluderemmo che l'ampiezza della famiglia non ha influenza sul numero di carte di credito possedute. Si noti che il test non valuta se il coefficiente è uguale a un dato valore, ma piuttosto se dev'essere utilizzato nell'equazione oppure no. Anche in questo caso il test appropriato è il *t* di Student. È bene ricordare che il test statistico sul coefficiente di regressione serve a valutare se, nell'insieme dei possibili campioni estraibili dalla popolazione, tale coefficiente è diverso da zero. Nell'esempio sulla regressione semplice, l'errore standard dell'ampiezza della famiglia è 0.229. Al valore della statistica *t* calcolata sul campione, che è 4.25 (0.971 ÷ 0.229), è associata una probabilità di 0.005. Ciò significa che, con un alto grado di fiducia (99.5%), possiamo ritenere che il coefficiente debba essere mantenuto nell'equazione di regressione.

In alcuni casi è necessario confrontare modelli in cui l'uno è un sottoinsieme dell'altro (si tratta dei cosidetti modelli **annidati**). Un esempio è il seguente.
Poniamo che il modello base (nodello 1) sia

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

a cui si contrappone un modello alternativo, privo della variabile x_3 (modello2).

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

La contrapposizione tra modello 1 e modello 2 può essere espressa sotto forma di ipotesi da sottoporre a test nella seguente forma:

$$H_0 : \beta_3 = 0.$$

Il test statistico da utilizzare per sottoporre a verifica questa ipotesi nulla è il seguente:

$$F_{k1-k2;\, n-k1} = (SSR_2 - SSR_1) / (k_1-k_2) / SSR_1 / n-k_1$$

in cui
SSR_1 è la somma dei residui al quadrato dal modello 1
SSR_2 é la somma dei residui al quadrato dal modello 2
k_1 è il numero di variabili indipendenti del modello1
k_2 è il numero di variabili indipendenti del modello 2

che segue una distribuzione di Fisher Snedecor con (k_1-k_2) *e* $(n-k_1)$ gradi di libertà.

45. Sommario I test di significatività sui coefficienti di regressione rappresentano per il ricercatore uno strumento di valutazione empirica del loro "vero" effetto. Anche se non costituiscono test di validazione, consentono comunque di determinare se l'impatto delle variabili dipendenti, sintetizzato dai coefficienti di regressione, è generalizzabile ad altri campioni della popolazione in studio. Si osservi che i valori stimati per i coefficienti di regressione sono strettamente legati al campione utilizzato, del quale rappresentano le migliori stime, e che possono variare notevolmente da campione a campione, evidenziando la necessità di validare ogni analisi della regressione su differenti campioni. È poi certo che, nel mettere in atto una procedura di validazione, il ricercatore debba aspettarsi che i coefficienti varino. Egli comunque deve cercare di dimostrare che la relazione vale anche su altri campioni e che quindi i risultati possono essere generalizzati ad ogni campione estratto dalla popolazione.

46. FASE 5: Interpretazione dell'equazione di Regressione Il successivo compito del ricercatore è quello di interpretare l'equazione di regressione stimata, valutando i coefficienti di regressione in base al loro contributo alla spiegazione della variabile dipendente. Nel caso in cui sia stato impiegato un metodo di ricerca sequenziale o un approccio combinatorio, il ricercatore deve valutare non soltanto il modello di regressione stimato, ma anche le variabili indipendenti potenziali che sono state omesse dall'equazione. In queste tecniche, infatti, il risultato finale può essere influenzato in modo sostanziale dalla multicollinearità. Pertanto, oltre a valutare i coefficienti stimati, il ricercatore deve anche valutare l'impatto potenziale delle variabili omesse, per assicurare al modello il significato operativo insieme alla significatività statistica.

47. Uso dei coefficienti di Regressione I coefficienti di regressione stimati sono usati per calcolare i valori previsti per ciascuna osservazione e per esprimere la variazione attesa nella variabile dipendente per ogni unità di variazione nella(e) variabile(i) indipendente(i). Oltre a fare previsioni, molte volte attraverso la regressione si cerca di dare una spiegazione ad un fenomeno – valutando l'impatto di ciascuna variabile indipendente sulla previsione della variabile dipendente. Nell'esempio sulla regressione multipla già discusso si voleva conoscere quale variabile – ampiezza o reddito familiare – aveva la più rilevante influenza sulla previsione del numero di carte usate dalla famiglia.

Sfortunatamente, i coefficienti di regressione (b_0, b_1 e b_2) non danno in molti casi questa informazione. Un esempio piuttosto semplice può chiarire tale affermazione. Si supponga di voler prevedere quanto spendono in media i teenager per l'acquisto di CD (Y) usando le due seguenti variabili indipendenti: il reddito dei genitori in migliaia di dollari (X_1) e la disponibilità mensile di denaro dei teenager misurata in dollari (X_2). Si supponga inoltre di aver stimato, attraverso il metodo dei minimi quadrati, il seguente modello:

$$Y = -0.01 + X_1 + 0.001X_2$$

Si potrebbe ritenere allora che X_1 sia la variabile più importante, perché il suo coefficiente è 1000 volte più grande del coefficiente di X_2. In realtà, tale supposizione non è affatto vera. Un incremento di 10$ nel reddito dei genitori produce un cambiamento di 0.01$ (1 × 10$ ÷ 1000$) negli acquisti

medi di CD (si divide 10$ per 1000 perché il valore di X_1 è misurato in migliaia di dollari). Ma anche una variazione di 10$ nella disponibilità mensile dei teenager produce una variazione di 0.01$ (0.001 × 10$) nella spesa media per i CD (poiché la disponibilità dei giovani è stata misurata in dollari). Dunque, un cambiamento di 10$ nel reddito dei genitori ha lo stesso effetto di un cambiamento di pari intensità nella disponibilità di denaro dei teenager. Entrambe le variazioni sono ugualmente importanti, ma i coefficienti di regressione non rilevano direttamente questo dato di fatto. Il problema può essere risolto utilizzando un coefficiente modificato detto *coefficiente beta*.

48. Standardizzazione dei coefficienti di Regressione: i coefficienti beta

Se tutte le variabili indipendenti fossero state **standardizzate** prima di procedere alla stima dell'equazione di regressione, si sarebbero ottenuti coefficienti di regressione diversi. I coefficienti stimati sui dati standardizzati sono detti **coefficienti beta**. Il vantaggio dato dal loro calcolo consiste nell'eliminazione del problema della differenza nell'unità di misura delle diverse variabili indipendenti, poiché essi riflettono l'effetto prodotto sulla variabile dipendente da una variazione in unità standard osservata sulla relativa variabile indipendente. Potendo in tal modo leggere il modello secondo una unità di misura comune, è possibile determinare quale variabile ha la maggior influenza nell'equazione di regressione.

Quando si utilizzano i coefficienti beta è necessario considerare tre aspetti. In primo luogo, essi dovrebbero essere usati come guida per la determinazione dell'importanza relativa delle variabili indipendenti soltanto quando la collinearità è minima. In secondo luogo, ciascun coefficiente beta può essere interpretato soltanto nell'intero contesto dell'equazione di regressione, anche tenendo conto di tutte le altre variabili. Ad esempio, il valore beta per l'ampiezza del nucleo familiare manifesta la sua importanza soltanto in relazione al reddito della famiglia, non in senso assoluto. Se si aggiunge all'equazione una terza variabile, è molto probabile che il coefficiente beta per l'ampiezza del nucleo familiare cambi, a causa di una possibile relazione tra la dimensione della famiglia e la nuova variabile. In terzo luogo, i coefficienti beta sono influenzati dai livelli delle variabili (ad esempio, per l'ampiezza del nucleo familiare, 5, 6 e 7). Se avessimo osservato famiglie con 8, 9 e 10 componenti, i valori di tali coefficienti sarebbero stati molto probabilmente diversi. In breve, i coefficienti beta dovrebbero essere usati soltanto come guida all'individuazione dell'importanza relativa delle variabili indipendenti inserite nell'equazione, e limitatamente all'intervallo dei valori osservati sui dati campionari.

49. Valutazione della multicollinearità

Una questione chiave nell'interpretazione dell'equazione di regressione è data dalla correlazione tra le variabili indipendenti. È questo un problema di dati, non di specificazione del modello. La situazione ideale per un ricercatore è quella in cui egli dispone di variabili indipendenti altamente correlate con la variabile dipendente ma poco correlate tra loro. In molte circostanze, però, e in particolare quando i dati coinvolgono risposte di un campione di consumatori, esiste certamente un qualche grado di multicollinearità. In altre situazioni, invece, è lo stesso ricercatore a creare un'alta multicollinearità, ad esempio quando utilizza variabili dummy per codificare variabili non metriche o termini polinomiali per rappresentare componenti non lineari. Compito del ricercatore è allora (1) valutare il grado di multicollinearità e (2) determinare il suo impatto sui risultati e, se necessario, approntare gli opportuni rimedi. Nei paragrafi seguenti si descrivono gli effetti della multicollinearità, alcune utili procedure diagnostiche e le possibili soluzioni al problema.

50. L'effetto della multicollinearità

L'effetto della multicollinearità può interessare sia la capacità di *spiegazione* del modello sia la sua *stima*. L'effetto definito in termini di spiegazione concerne principalmente la capacità della procedura di regressione e del ricercatore di rappresentare e capire l'influenza che ciascuna variabile indipendente ha sull'equazione di regressione. Quando esiste multicollinearità (anche a livelli relativamente bassi, come a 0.30 o a valori simili), il processo di separazione degli effetti individuali diviene molto più complesso. Per prima cosa, la multicollinearità fa diminuire il valore del coefficiente di determinazione e rende sempre più difficoltoso il suo aumento, anche se si aggiungono al modello nuove variabili con un loro proprio potere esplicativo. Come seconda osservazione, importante come la precedente,

bisogna dire che la sua presenza rende problematica la determinazione dei contributi individuali delle variabili indipendenti, perché i loro effetti vengono "mescolati" o confusi. La multicollinearità si esprime infatti in una vasta porzione di variabilità condivisa da tutte le variabili e in bassi livelli di variabilità spiegata da ogni singola variabile. Soltanto in base a questi ultimi, però, possono essere determinati gli effetti individuali delle variabili indipendenti. A titolo di esempio, si assuma che una variabile indipendente (X_1) abbia una correlazione di 0.60 con la variabile dipendente, e che una seconda variabile indipendente (X_2) abbia invece una correlazione di 0.50. In base a questi valori, X_1 dovrebbe spiegare il 36% (ottenuto elevando al quadrato il coefficiente di correlazione, pari a 0.60) della varianza della variabile dipendente, mentre X_2 dovrebbe spiegarne il 25% (0.50 al quadrato). Se le due variabili indipendenti non sono tra loro correlate, esse non "sovrappongono" (o equivalentemente non "condividono") la loro capacità previsiva. La loro spiegazione totale sarebbe pari alla somma delle spiegazioni individuali, e cioè pari al 61%. Ma al crescere della collinearità, cresce la "condivisione" della capacità previsiva mentre diminuisce la capacità previsiva collettiva delle variabili indipendenti.

Considerando diversi livelli di collinearità, la Figura 9 rappresenta le porzioni di varianza comune e specifica per l'esempio delle due variabili indipendenti. Se la collinearità tra tali variabili è zero, allora esse prevedono rispettivamente il 36% e il 25% della varianza della variabile dipendente, per una previsione complessiva del 61%. Ma come la collinearità cresce, la varianza spiegata totale diminuisce. Inoltre, l'ammontare di varianza spiegata singolarmente da ciascuna variabile indipendente viene ridotta a livelli tali da rendere piuttosto problematica la stima del loro effetto individuale.

Oltre a influenzare la capacità esplicativa del modello, la multicollinearità può avere effetti sostanziali sulla stima dei coefficienti di regressione e sui loro test di significatività. Si osservi innanzitutto che nel caso estremo in cui due o più variabili sono perfettamente correlate, caso che viene definito **singolarità**, è addirittura impossibile procedere alla stima di uno qualsiasi dei coefficienti. In una situazione come questa, prima di procedere alla stima dei coefficienti è necessario rimuovere la singolarità. Ma se anche la multicollinearità non è massima e si presenta con un alto grado può portare a coefficienti di regressione non correttamente stimati o addirittura stimati con segno opposto. L'esempio seguente illustra questa situazione (si veda la Tabella 9). Esaminando la matrice di correlazione e le regressioni semplici, appare chiaro che la relazione tra Y e V_1 è positiva, mentre la relazione tra Y e V_2 è negativa. L'equazione di regressione multipla, però, non mantiene tutti i segni delle relazioni rivelate dai modelli di regressione semplice. All'osservatore che esaminasse soltanto i coefficienti dell'equazione multivariata, sembrerebbe che entrambe le relazioni (Y e V_1, Y e V_2) siano negative, mentre noi sappiamo che ciò non vale per Y e V_1. Intuitivamente il segno del coefficiente di regressione di V_1 è sbagliato, ma è dovuto alla forte correlazione negativa tra V_1 e V_2 che si manifesta nel segno opposto attribuito a V_1. Anche se questi effetti sulle procedure di stima si presentano principalmente ad alti livelli di multicollinearità (al di sopra di 0.80), la possibilità di ottenere risultati contrari a quanto atteso o errati rende indispensabile l'attento esame di ogni equazione di regressione, per il controllo dell'eventuale esistenza di multicollinearità.

51. L'identificazione della multicollinearità

Si è visto in precedenza che l'effetto della multicollinearità può essere sostanziale. In una qualunque analisi della regressione la sua valutazione dovrebbe essere compiuta in due fasi: (1) identificazione dell'ordine di grandezza con cui si presenta e (2) valutazione del grado di influenza sulla stima dei coefficienti. Se è poi necessario adottare azioni correttive, sono disponibili diverse opzioni. In questo paragrafo si illustreranno le procedure di identificazione e di valutazione, per esaminare quindi alcuni dei possibili rimedi.

Lo strumento più semplice e ovvio per l'individuazione della collinearità è l'esame della matrice di correlazione delle variabili indipendenti. La presenza di coefficienti di correlazione con valori molto alti (generalmente uguali o superiori a 0.90) è il primo indicatore di una collinearità importante. L'assenza di valori di correlazione alti, però, non assicura l'assenza di collinearità. Essa infatti può essere dovuta all'effetto combinato di due o più variabili indipendenti.

Due delle misure più comuni per valutare la presenza di collinearità tra coppie e tra insieme più grandi di variabili sono (1) la **tolleranza** e (2) la sua inversa – il **fattore di accrescimento della varianza** (nella terminologia inglese, *variance inflaction factor*, ***VIF***). Tali misure ci dicono qual è il grado con cui ciascuna variabile indipendente è spiegata dalle altre variabili indipendenti. Più semplicemente, ogni variabile indipendente viene espressa in funzione delle rimanenti variabili indipendenti diventando in tal modo variabile dipendente. La tolleranza rappresenta allora l'ammontare di variabilità della prescelta variabile indipendente che rimane non spiegata dalle altre variabili

indipendenti. Pertanto, quando la tolleranza assume valori molto piccoli (e di conseguenza il *VIF* assume valori elevati, poiché *VIF* = 1 / tolleranza) esiste un'alta collinearità. Una soglia (nella terminologia inglese, *cutoff*) comunemente usata per stabilire un livello accettabile di collinearità è il valore di tolleranza 0.10, che corrisponde a un *VIF* al di sopra del 10. È opportuno però che il ricercatore determini di volta in volta il grado di collinearità accettabile, perché molte soglie raccomandate o di default ammettono livelli di collinearità consistenti. Ad esempio, la soglia sopra suggerita per il valore della tolleranza, pari a 0.10, corrisponde a una correlazione multipla di 0.95. Inoltre, un valore di correlazione multipla pari a 0.9, misurato tra una variabile indipendente e tutte le altre variabili indipendenti (in modo simile alla regola applicata nella matrice di correlazione tra coppie di variabili), darebbe luogo a un valore di tolleranza inferiore a 0.19. Perciò, ogni variabile con valori di tolleranza inferiori a 0.19 (o con valori del *VIF* superiori a 5.3) avrebbe una correlazione più alta di 0.90.

Quando si ricorre alle procedure informatizzate per l'analisi della regressione, si consiglia di specificare sempre i valori soglia della tolleranza, poiché i valori per l'esclusione delle variabili con alta collinearità previsti di default ammettono generalmente livelli di collinearità estremamente elevati. Per esempio, in SPSS il valore della tolleranza di default per l'esclusione di una variabile è 0.0001, il che significa che, fino a quando la percentuale di varianza spiegata dalle altre variabili indipendenti non supera il 99.99%, la variabile in questione non può essere esclusa dall'equazione di regressione. Si osservi infine che è possibile valutare l'effetto reale di una elevata collinearità sulla stima dei coefficienti, ma ciò va al di la degli scopi di questo testo.

TABELLA 8 Stime di regressione in presenza di multicollinearità

A. DATI

	Dimensione campionaria		
	Dipendente	*Indipendenti*	
Rispondenti	Y	V_1	V_1
1	5	6	13
2	3	8	13
3	9	8	11
4	9	10	11
5	13	10	9
6	11	12	9
7	17	12	7
8	15	14	7

B. MATRICE DI CORRELAZIONE

	Y	V_1	V_2
Y	1.000		
V_1	0.823	1.000	
V_2	– 0.977	– 0.977	1.000

C. STIME DI REGRESSIONE

Regressione semplice (V_1):	$Y = – 4.75 + 1.5V_1$
Regressione semplice (V_2):	$Y = 29.75 + –1.95V_2$
Regressione multipla (V_1 e V_2):	$Y = 44.75 + – 0.75V_1 + – 2.7V_2$

Anche disponendo di diagnosi sulla collinearità ottenute usando il *VIF* o la tolleranza, non è detto che si conosca quali variabili non sono tra loro correlate. Una procedura sviluppata da Belsley et al. consente di identificare le variabili non correlate anche quando esiste correlazione tra alcune delle variabili utilizzate. Tale procedura costituisce per il ricercatore uno strumento di grande potenza diagnostica per la valutazione del livello e dell'impatto della collinearità.

52. Provvedimenti per la multicollinearità

I rimedi da adottare nel caso si osservi multicollinearità vanno dalla trasformazione dell'equazione di regressione fino all'uso di procedure di stima particolari. Una volta determinato il grado di collinearità, il ricercatore può scegliere tra

alcune diverse opzioni:

- Omettere una o più variabili indipendenti altamente correlate e identificare altre variabili indipendenti per aiutare la previsione. Utilizzando questa procedura, però, il ricercatore deve prestare particolare attenzione, onde evitare di creare un errore di specificazione cancellando una o più variabili indipendenti.
- Usare il modello con le variabili indipendenti altamente correlate soltanto a fini previsivi (in sostanza, senza interpretare i coefficienti di regressione).
- Usare i coefficienti di correlazione semplici calcolati tra ciascuna variabile indipendente e la variabile dipendente per comprendere la relazione tra le coppie di variabili dipendente-indipendente.
- Usare un metodo di analisi più sofisticato, come la regressione Bayesiana (oppure un suo caso particolare – la *ridge regression*) o la regressione sulle componenti principali, per ottenere un modello che rifletta più chiaramente gli effetti individuali delle variabili indipendenti.

Ciascuna di queste opzioni richiede che il ricercatore esprima un giudizio sulle variabili incluse nell'equazione di regressione, giudizio che dovrebbe sempre essere guidato dai fondamenti teorici dello studio.

53. FASE 6: Validazione dei risultati

Dopo aver identificato il miglior modello di regressione, il passo finale consiste nell'accertarsi che sia rappresentativo della relazione tra le variabili nella popolazione (generalizzazione) e che sia appropriato per le circostanze in cui dovrà essere utilizzato (trasferibilità). La linea guida migliore per valutare questi aspetti è il controllo della conformità del modello di regressione stimato ad un modello teorico esistente o ad un insieme di risultati già validati sullo stesso argomento di studio. Precedenti risultati o teorie, però, non sempre sono disponibili. Pertanto, descriveremo anche gli approcci empirici al problema della validazione del modello.

54. Campioni aggiuntivi o suddivisione del campione

L'approccio empirico più appropriato per la validazione del modello consiste nel testare statisticamente l'equazione di regressione su un campione estratto dalla popolazione generale. Un nuovo campione può assicurare la rappresentatività ed essere utilizzato in vari modi. In primo luogo, è possibile valutare la capacità previsiva del modello prevedendo i valori del nuovo campione, e quindi calcolare misure della bontà dell'adattamento ai dati. In secondo luogo, è possibile utilizzarlo per stimare un nuovo modello, da confrontare con l'equazione originale in base a caratteristiche quali, ad esempio: la significatività delle variabili impiegate; il segno, le dimensioni e l'importanza relativa delle variabili; l'accuratezza della previsione. In entrambi i casi, il ricercatore giudica la validità del modello originale mettendolo a confronto con modelli di regressione stimati sul nuovo campione.

Molte volte la possibilità di ottenere nuovi dati è limitata o preclusa da fattori come i costi, i tempi, la disponibilità di rispondenti. In questi casi il ricercatore può dividere il campione in due parti: un primo sottocampione sul quale effettuare la stima per la costruzione del modello di regressione, un secondo sottocampione da utilizzare per la validazione dell'equazione. Per effettuare questa suddivisione dei dati si può ricorrere a varie procedure, casuali o sistematiche, che estraggono due campioni indipendenti dall'insieme di dati originario. Tutti i pacchetti statistici più diffusi dispongono di specifiche opzioni che consentono di effettuare la stima e la validazione del modello su sottocampioni distinti.

Che un nuovo campione venga estratto oppure no, è probabile che si osservino delle differenze tra il modello originario e quelli stimati per la validazione. Al ricercatore spetta allora il compito di mediare tra i diversi risultati, alla ricerca del miglior modello su tutti i campioni. La necessità di continui sforzi per la validazione e il perfezionamento del modello ci ricorda che nessuna equazione di regressione costituisce il modello finale e assoluto, a meno che non sia stimato sull'intera popolazione.

55. Confronto tra modelli di Regressione

Per confrontare diversi modelli di regressione, lo standard più comunemente usato è la bontà dell'adattamento globale ai dati. Si è

visto in precedenza che R^2 fornisce questo tipo di informazione, ma ha purtroppo uno svantaggio: al crescere delle variabili inserite nel modello, cresce anche il suo valore. Pertanto, per nessuna equazione di regressione si otterrà mai un valore di R^2 più elevato rispetto a quello del modello con tutte le variabili indipendenti; sarà possibile però ricavare un valore di R^2 pressoché equivalente con un numero inferiore di variabili indipendenti. Di conseguenza, per confrontare modelli con un numero di variabili indipendenti diverso è bene utilizzare R^2 corretto. Inoltre, questa misura è utile per il confronto tra modelli stimati su insieme di dati diversi, poiché tiene conto delle possibili differenze nelle numerosità campionarie.

56. Prevedere utilizzando il modello È sempre possibile fare previsioni utilizzando un modello, considerando nuovi valori delle variabili indipendenti, sostituendoli nell'equazione stimata e calcolando il valore previsto per la variabile dipendente. Procedendo in questo modo, si devono comunque tenere in considerazione alcuni aspetti che possono avere un serio impatto sulla qualità delle nuove previsioni:

1. Quando il modello viene utilizzato su un nuovo campione, si deve ricordare che ogni previsione non è influenzata soltanto dall'errore di campionamento del campione originario, ma anche da quello del nuovo campione. Pertanto, oltre alla stima puntuale, è sempre necessario calcolare gli intervalli di confidenza delle previsioni, in modo da disporre, per la variabile dipendente, dell'intervallo atteso dei suoi valori.
2. Si deve essere certi che le condizioni e le relazioni misurate al tempo dell'estrazione del campione originario non siano cambiate in modo determinante. Ad esempio, a proposito della ricerca sulle carte di credito, se la maggior parte delle compagnie cominciasse ad innalzare gli emolumenti delle carte, il possesso reale di carte di credito cambierebbe sostanzialmente; questa informazione, però, non sarebbe inclusa nel modello.
3. Infine, il modello non deve essere usato a fini previsivi con valori delle variabili indipendenti esterni all'intervallo di variazione osservato nel campione. Ad esempio, ancora nello studio sulle carte di credito, se la famiglia più numerosa del campione fosse composta da 6 membri, non sarebbe opportuno prevedere il numero di carte di credito possedute dalle famiglie con 10 membri. Non si può assumere infatti che le relazioni rimangano le stesse per valori delle variabili indipendenti molto superiori o inferiori a quelli osservati nel campione originario.

57 Analisi e previsione della domanda primaria

57.1. Previsioni aziendali e di settore (o di mercato).

Le previsioni delle vendite e del mercato sono la base per definire e programmare, nella migliore combinazione possibile, gli obiettivi e le risorse interne di tutte le funzioni aziendali: dalla produzione agli acquisti e al personale, dalla gestione finanziaria alla ricerca e sviluppo e alla commercializzazione dei prodotti.

Nella gestione di una azienda la stima delle vendite future integrata con le previsioni dello scenario economico-ambientale in cui l'azienda si troverà presumibilmente ad operare rappresenta pertanto una delle operazioni più importanti e delicate per le conseguenze che riverbera direttamente sulle politiche di marketing e sul processo di programmazione strategica. Due tipi di previsione quindi: quella riguardante le vendite dell'azienda stessa e quella aggregata delle vendite, vale a dire la previsione che riguarda il mercato in cui l'azienda agisce. Tale previsione detta anche della domanda primaria del prodotto o del gruppo di prodotti di interesse, risponde alla necessità per una azienda di stimare l'andamento delle vendite che presumibilmente si genereranno in una data zona o in un segmento di mercato in presenza di particolari condizioni dell'ambiente economico e in funzione delle attività sviluppate da tutte le aziende che operano nel settore. La previsione aggregata consente quindi ad una azienda di individuare i segmenti in cui rafforzare eventualmente la propria presenza o programmare il proprio ingresso.

Gli orizzonti temporali a cui le previsioni aziendali fanno riferimento sono essenzialmente il breve e il medio termine.

Di particolare importanza sono le previsioni a breve termine che vengono predisposte con grande

dettaglio all'inizio dell'esercizio ed aggiornate a brevi intervalli di tempo, di solito da uno a sei mesi[1]. Lo scopo è quello di stimare il presumibile livello della domanda dell'azienda, tenendo conto sia della sua capacità produttiva che delle condizioni della concorrenza, da utilizzare – conviene ribadirlo – come input per una programmazione efficiente della gestione del magazzino, del carico degli impianti e dei turni di lavoro, per stabilire le precedenze di lavorazione e le cadenze degli acquisti, per stimare le necessità di manodopera, ecc.

A loro volta, le previsioni a medio termine sono utilizzate per ripartire le risorse tra investimenti alternativi. Nella maggior parte dei casi assumono la forma di budget[2] assegnati ai singoli reparti aziendali per l'anno successivo, sulla base delle previsioni relative alla propria domanda e a quella della concorrenza; previsioni utili soprattutto per mettere a punto le politiche finanziarie, la rilevazione dei costi di produzione e il controllo budgetario, per indirizzare l'attività di ricerca e sviluppo, per confrontare l'andamento della quota di mercato con quella della concorrenza.

Prima di passare ad una rassegna schematica dei metodi di previsione conviene soffermare l'attenzione sui soggetti che in azienda o per l'azienda si occupano di previsioni.

I piani di previsione delle vendite sono elaborati sia da componenti interne all'azienda (management, ufficio studi, consulenti) sia da componenti decentrate quali i venditori e i clienti. Tra queste i venditori rappresentano la fonte informativa principale. Alla forza di vendita giungono ininterrottamente informazioni sulle probabili attività della clientela, sulle abitudini d'acquisto dei grossisti o dei consumatori finali rispetto alla marca o al prodotto, e sugli sviluppi dell'attività della concorrenza. I venditori sono quindi in grado di fornire una serie di dati raccolti direttamente sul mercato e stimare le vendite future sulle zone di loro competenza. Queste informazioni vengono successivamente rielaborate dal management in base ad elementi noti alla direzione aziendale. Ovviamente il contributo del management, della forza di vendita e dei clienti ha generalmente un contenuto soggettivo e congetturale in quanto si fonda sull'intuizione e sul giudizio dei singoli.

Sebbene i metodi di valutazione soggettiva siano strettamente legati alle valutazioni dei singoli, e criticabili quindi per la mancanza di rigore scientifico, mantengono una forte capacità previsionale poiché riflettono «esperienza» e giudizi non facilmente ottenibili con procedimenti statistici. A questo riguardo particolarmente rilevanti nel nostro paese sono le previsioni elaborate dall'Istituto di Studi e Analisi Economiche (ISEA).

Lo sviluppo dei sistemi informativi aziendali rende disponibili basi di dati che consentono di integrare le previsioni di tipo soggettivo con previsioni di tipo statistico. La necessità di interpretare, formalizzare, programmare le decisioni aziendali in condizioni di incertezza tramite adeguati strumenti operativi fa dell'analisi e della previsione delle serie temporali un ambito della statistica di notevole interesse metodologico e di grande rilevanza operativa. Ma in ogni caso va sottolineato che in azienda è ancora largamente seguito un criterio che impiega congiuntamente i due tipi di previsione: statistiche e congetturali.

I metodi e le tecniche statistiche di previsione concretamente disponibili sono assai numerosi, e sostenuti non di rado da un sofisticato apparato analitico. Tali tecniche, che si differenziano quanto ad impostazione, contenuto metodologico, requisiti in termini di disponibilità dei dati, contesto applicativo, possono essere suddivisi in due grandi famiglie, fondate su principi differenti.

I metodi cosiddetti "endogeni", utilizzano essenzialmente dati temporali riflettenti la dimensione assunta dal fenomeno analizzato (ammontare in valore o in quantità delle vendite di un dato prodotto per mese, trimestre o anno). In questo contesto si resta all'interno della serie storica, che si va a prolungare punto a punto. Si tratta di metodi utili soprattutto per previsioni a breve termine. Rientrano in questo gruppo prima di tutto il metodo di Box-Jenkins[3] per l'organicità e la compattezza dell'impostazione, nonché metodi tradizionali come l'estrapolazione grafica, le medie mobili, il livellamento esponenziale. Il metodo di Box e Jenkins (noto anche come analisi stocastica delle serie temporali) si fonda sul presupposto che le osservazioni di un fenomeno nel corso del tempo siano generate da una struttura probabilistica (ovviamente ignota) della quale è possibile stimare su base inferenziale i parametri mediante lo studio dei legami temporali riconoscibili nei dati osservati. L'approccio sviluppa un tipo di analisi capace di inglobare in maniera più efficiente di altri le informazioni contenute nelle osservazioni più recenti della serie temporale ed è questo uno dei motivi che lo rendono particolarmente adatto per le previsioni a breve termine.

[1] Sono generalmente considerate previsioni a brevissimo termine quelle decisioni necessarie per una efficace gestione del quotidiano.

[2] Il budget di vendita è una stima prudenziale del volume atteso delle vendite utilizzato soprattutto per prendere le decisioni riguardanti acquisti, produzione e *cash flow*.

[3] Sui metodi di analisi delle serie temporali e di previsione vi sono numerosi manuali. Una sintesi molto semplice si trova in R. Guarini – F. Tassinari, *Statistica economica. Problemi e metodi di analisi*, Bologna, Il Mulino, 1997, seconda edizione (capp. 2 e 3).

Una seconda classe di metodi di previsione è quello che raggruppa i cosiddetti metodi "esogeni". Si tratta di metodi che puntano prioritariamente alla identificazione delle cause (individuate tramite variabili indipendenti) che agiscono appunto sulle vendite dell'azienda o del settore assunte come variabile dipendente di interesse. Una volta individuate le cause occorre specificare successivamente la relazione (il modello) che lega le variabili indipendenti alla variabile dipendente e risolvere, quindi, stimandola statisticamente, l'equazione (o il sistema di equazioni) così ottenuto.
L'applicazione con finalità previsive di questi modelli, utili soprattutto per fare congetture sull'andamento futuro di un mercato, si basa sull'ipotesi che "le medesime cause producono gli stessi effetti". Come scrive con grande efficacia Edmond Malinvaud, le ipotesi sulla permanenza delle relazioni fra vendite (o spesa dal punto di vista degli acquirenti) e i fattori che le determinano sono formalmente rappresentabili tramite modelli solo alla condizione che le correlazioni che si osservano sui campioni empirici siano interpretabili come prova di una dipendenza estendibile ad un'ampia gamma di situazioni, con l'obiettivo di giungere se non a generalizzazioni vere e proprie almeno ad una buona approssimazione delle leggi che regolano il comportamento degli acquirenti (consumatori).
Fanno parte di questo gruppo le tecniche di analisi della regressione semplice e multipla il cui scopo principale è quello di fare previsioni sui valori che può assumere una variabile dipendente in base alla conoscenza di una o più variabili indipendenti.

57.2. La previsione con il modello di regressione. Da un punto di vista aziendale, un'importante obiettivo dell'analisi di regressione è la predizione. Cosa possiamo prevedere sulle vendite della nostra marca se possiamo predeterminare l'ammontare futuro delle spese per pubblicità? Quale prezzo di vendita al dettaglio possiamo imporre ai distributori per il nostro prodotto se vogliamo essere competitivi? A domande come queste si può dare risposta usando la funzione stimata con il modello di regressione per stimare la risposta che corrisponde ad un determinato valore della variabile indipendente (o ad una combinazione dei valori delle variabili indipendenti se facciamo ricorso ad un modello di regressione multipla)
Inizialmente ci concentreremo sulla predizione del valore medio, poi si passerà alla discussione del problema della predizione di un valore puntuale.

In accordo al modello di regressione semplice, la risposta media in corrispondenza al valore x^* della variabile x sia dato da

$b_0 + b_1 x^*$.

Il valore atteso è stimato da

$b_0' + b_1' x^* = y^*$

in cui $b_0' + b_1'$ sono le stime dei minimi quadrati ordinari dei parametri del modello di regressione semplice (per le condizioni di applicabilità del metodo dei minimi quadrati ordinari si veda Brasini-Tassinari-Tassinari, cap.3), che è il valore della linea di regressione stimata in corrispondenza
a $x = x^*$.
L'inferenza statistica concernente il valore atteso per $x = x^*$ si basa sulla distribuzione t di Student con n-2 gradi di libertà.

Un intervallo di confidenza all 100(1-a)% per la risposta media $b_0 + b_1 x^*$ è dato da

$$b_0' + b_1' x^* \pm t_{a/2} \, . \, s \left(1/n + ((x^* - x_m)^2 / S_{xx})^{1/2} \right.$$

in cui
s è lo scarto quadratico medio degli errori dal modello di regressione,
S_{xx} è la devianza della variabile indipendente
x_m è il valore medio artimetico di x sulle unità statistiche osservate

$t_{a/2}$ è il valore in ascissa della distribuzione *t* di Student che lascia a destra una massa di probabilità pari ad a/2.
La quantità

$$s\ (\ 1/n + ((x^*-x_m)^2/S_{xx})^{1/2}$$

rappresenta l'errore standard della stima della risposta media.
La formula per l'errrore standard della predizione mostra che, a parità di condizioni, il suo valore aumenta all'aumentare di $(x^*-x_m)^2$. Di conseguenza, l'intervallo di confidenza associato ad x^* è maggiore. In generale, la predizione è più precisa nei pressi della media campionaria rispetto a quella formulata per valori che sono lontani dalla media.

Per quanto riguarda la predizione del valore puntuale, la formula che consente di ottenere l'errore standard della previsione è la seguente:

$$s\ (1+\ 1/n + ((x^*-x_m)^2/S_{xx})^{1/2}$$

in cui i simboli hanno il significato già visto.
Consideriamo un esempio. Immaginiamo che si sia stimato un modello di regressione lineare che collega le spese pubblicitarie (*x)* con il valore delle vendite dell'impresa (*y*). Le stime dei coefficienti sono, rrispettivamente
$b_0 = 0{,}44$ e $b_1 = 2{,}62$.
Lo scarto quadratico medio dell'errore, *s,* vale 2,97, x_m è pari a 4,9 infine $S_{xx} = 40{,}9$ e $n= 10$..
Calcoliamo la previone di *y, y**, quando x^* è pari a 4,9, e l'intervallo di previsione al 95%.

La previsione puntuale $y^* = 0{,}44 + 2{,}62\ (4{,}5) = 12{,}23$.
La statistica *t* di Student con 8 gradi di libertà per a/2=0,025 è pari a 2,306, e quindi l'intervallo di confidenza è dato da

$$12{,}23 \pm 2{,}306 \times 2{,}97\ (1+1/10 +((4{,}5-4{,}9)^2/40{,}9)^{1/2}$$

Grande cautela deve inoltre essere utilizzata quando si estende la previsione al di fuori del campo di variazione della x utilizzato per la stima. L'intervallo di confidenza può diventare così ampio che le previsioni sono sostanzialmente inaffidabili. Inoltre, la natura della relazione tra le variabili può cambiare in modo drammatico, per valori distanti di x, ed i valori correnti non sono in grado di fornire informazioni per individuare tale cambiamento.

Quando si utilizza il modello di regressione con dati che provengono da serie temporali occorre porre particolare attenzione a due problemi, tra loro connessi,:

a) la mancata indipendenza dei residui e
b) la non stazionarietà in media della serie.

Pertanto, è buona norma, una volta calcolati i residui, procedere al calcolo dei coefficienti di autocorrelazione. L'impiego del metodo di stima dei minimi quadrati ordinari è adeguato soltanto se i residui risulteranno incorrelati. Infatti la stima con i minimi quadrati ordinari è lecita solo se non risulta falsa l'assunzione che gli errori siano casuali, e di conseguenze le variabili dipendenti, in corrispondenza di valori differenti delle variabili esplicative siano indipendenti. Ciò significa che il valore di *y* in corrispondenza di un certo valore di *x* è indipendente dal valore di *y* in corrispondenza di un altro valore di *x*. Tuttavia, se i dati provengono da una serie temporale, ciò in genere non è

vero. Le osservazioni prese in tempi successivi tendono ad essere in relazione le une alle altre. L'ammontare dei salari nel mese *t* è collegato al monte salari del mese precedente, il numero di occupati nel trimestre corrente è legato al numero di occupati del mese precedente, la quota di mercato di questa settimana è legata alla quota di mercato della settimana precedente, e così via. Le inferenze dai modelli di regressione costruiti con dati di serie temporali possono essere fuorvianti anche se vengono utilizzate le formule appropriate.
Per controllare che i residui dal modello non siano afflitti da correlazione seriale si possono impiegare diversi strumenti.
Il punto iniziale è il calcolo del coefficiente di autocorrelazione campionario che costituisce un elemento chiave per accertare la presenza di legami lineari tra le osservazioni sfasate di una serie temporale è pertanto allo sfasamento (*lag*) *k*, la cui formula è la seguente:

$$r_k = \frac{\sum_{t=1}^{n-k}(y_t - \mu_y)(y_{t+k} - \mu_y)}{n}$$

Per accertare la significatività dei coefficienti di autocorrelazione occorre procedere alla verifica dell'ipotesi
H_0 : r_k =0 contro l'ipotesi alternativa H_1 r_k diverso da zero. r_k è il valore vero del parametro nella popolazione.
Si procede con l'usuale test t di Student, stimando l'errore standard di r_k con la formula:

$$\sigma_{r_k} = (1/n + 2\sum_{i=1}^{k-1} r_i^2)^{1/2}$$

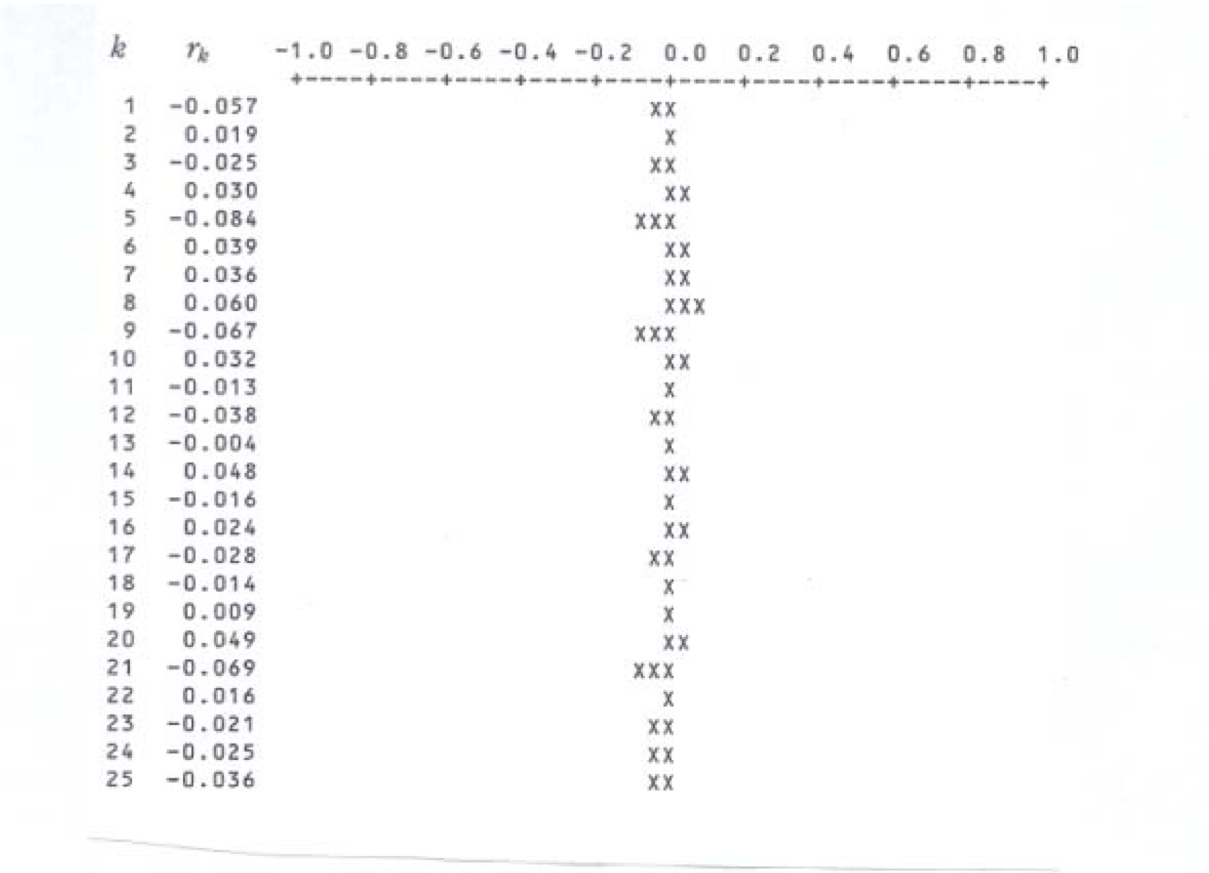
Un'ulteriore verifica dell'assenza di autocorrelazione nella serie può essere compiuta sottoponendo a verifica l'ipotesi nulla

$H_0 = r_1=r_2=\ldots=r_k=0$

nei confronti dell'ipotesi alternativa che almeno un coefficiente di autocorrelazione sia diverso da zero.

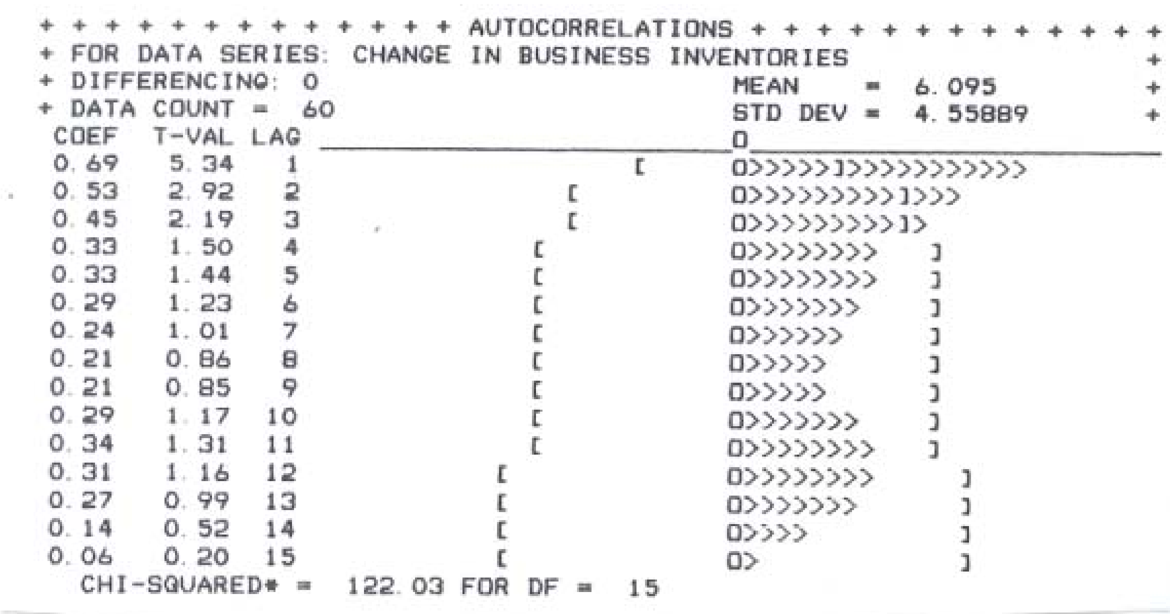
Nella **Figura 4** sono riprodotti i grafici di due serie di coefficienti di autocorrelazione, la a) che riproduce le autocorrelazione di una seri di residui indipendenti, mentre la b) mostra molti coefficienti di autocorrelazione significativi.

(a)



k	r_k
1	-0.057
2	0.019
3	-0.025
4	0.030
5	-0.084
6	0.039
7	0.036
8	0.060
9	-0.067
10	0.032
11	-0.013
12	-0.038
13	-0.004
14	0.048
15	-0.016
16	0.024
17	-0.028
18	-0.014
19	0.009
20	0.049
21	-0.069
22	0.016
23	-0.021
24	-0.025
25	-0.036

(b)



+ + + + + + + + + + + + + AUTOCORRELATIONS + + + + + + + + + + + + +
+ FOR DATA SERIES: CHANGE IN BUSINESS INVENTORIES +
+ DIFFERENCING: 0 MEAN = 6.095 +
+ DATA COUNT = 60 STD DEV = 4.55889 +

COEF	T-VAL	LAG
0.69	5.34	1
0.53	2.92	2
0.45	2.19	3
0.33	1.50	4
0.33	1.44	5
0.29	1.23	6
0.24	1.01	7
0.21	0.86	8
0.21	0.85	9
0.29	1.17	10
0.34	1.31	11
0.31	1.16	12
0.27	0.99	13
0.14	0.52	14
0.06	0.20	15

CHI-SQUARED* = 122.03 FOR DF = 15

Figura 4

La statistica test che si utilizza è nota come Q di Ljung e Box e ha la seguente formula:

$$Q = n(n+2)\sum_{k=1}^{K}(n-k)^{-1}r_k^2$$

La statistica test segue una distribuzione del chi-quadrato con k-p gradi di libertà.
Nello studio dei residui di un modello di regressione ha un impiego molto diffuso il test *d* di Durbin-Watson per verificare la presenza di autocorrelazione del primo ordine.
Tale test ha la seguente formula:

I valori critici sono stati tabulati da Durbin e Watson, e dipendono sia dal numero di osservazioni che dal numero di variabili indipendenti presenti nel modello di regressione. Il test varia tra 0 e +4, e prevede una zona di indeterminazione (compresa tra d_u e d_l) in cui il risultato della procedura di verifica è non conclusivo.
La lettura dei risultati (per testare l'ipotesi nulla r=0 contro l'alternativa r diverso da zero) va fatta secondo il seguente schema:

$(4-d_l)<d<4$	si rifiuta H_0 in favore di H_1, r<0,
$(4-d_u)<d<(4-d_l)$	indecisione,
$d_u<d<2$	si accetta $H_{0,}$
$d_l<d<d_u$	indecisione,
$0<d<d_l$	si rifiuta H_0 in favore di H_1, r>0.

Per quanto attiene al problema della non stazionarietà in media, si possono impiegare diversi strumenti per giungere alla sua identificazione.
In linea generale, già l'esame grafico permette di accertare se la serie presenti un trend, ovvero una tendenza pressocchè monotona crescente o decrescente, o se la media della serie riferita ad intervalli diversi è molto instabile (vedi la **Figura 5**).

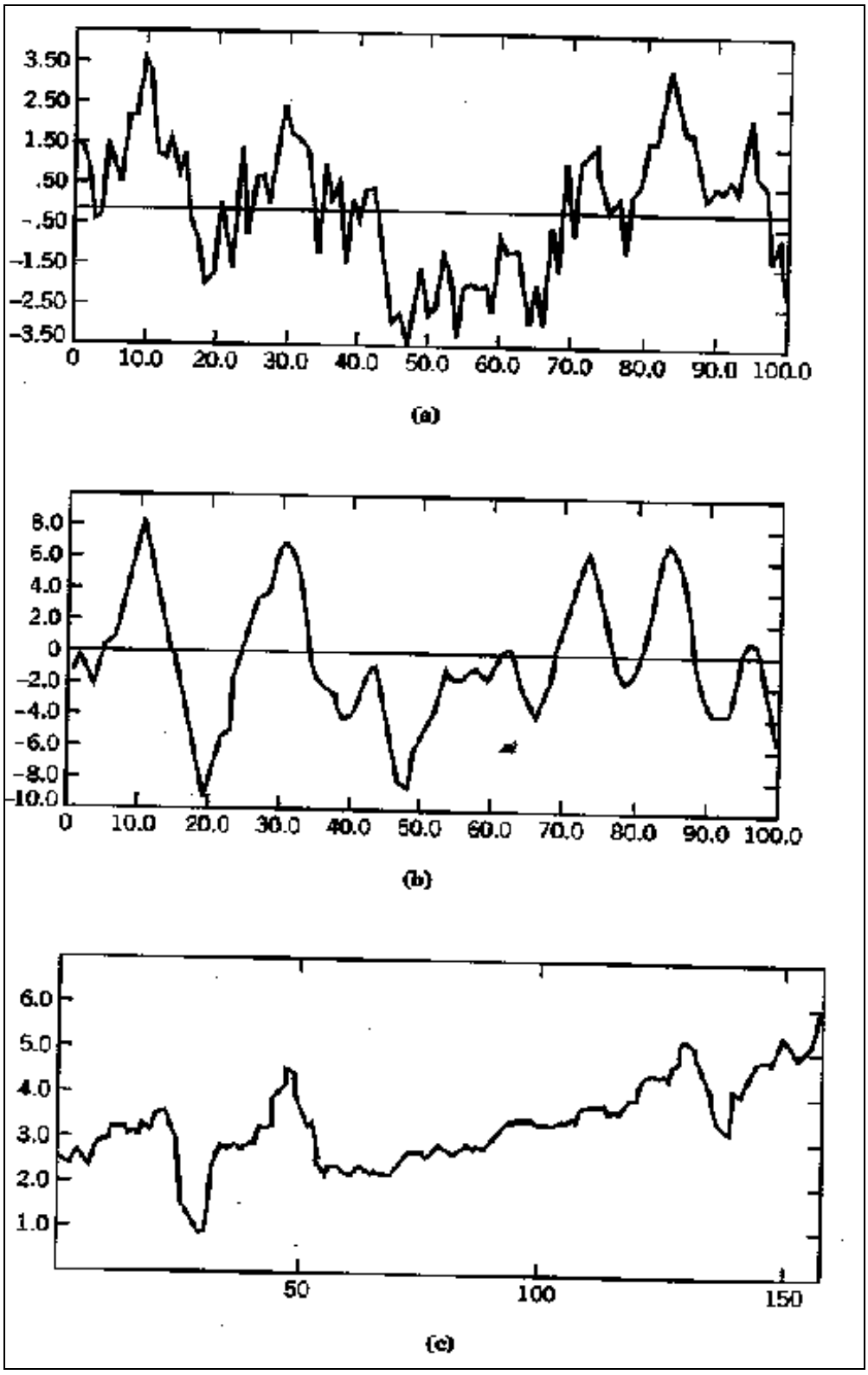


Figura 5

Un ulteriore strumento diagnostico è costituito dalla funzione di autocorrelazione dei valori originari della serie. Se questa è non stazionaria in media, si riscontreranno molti coefficienti

significativamente diversi da zero, e la forma del grafico sarà in generale lentamente decrescente, come mostrato nella **Figura 6**.

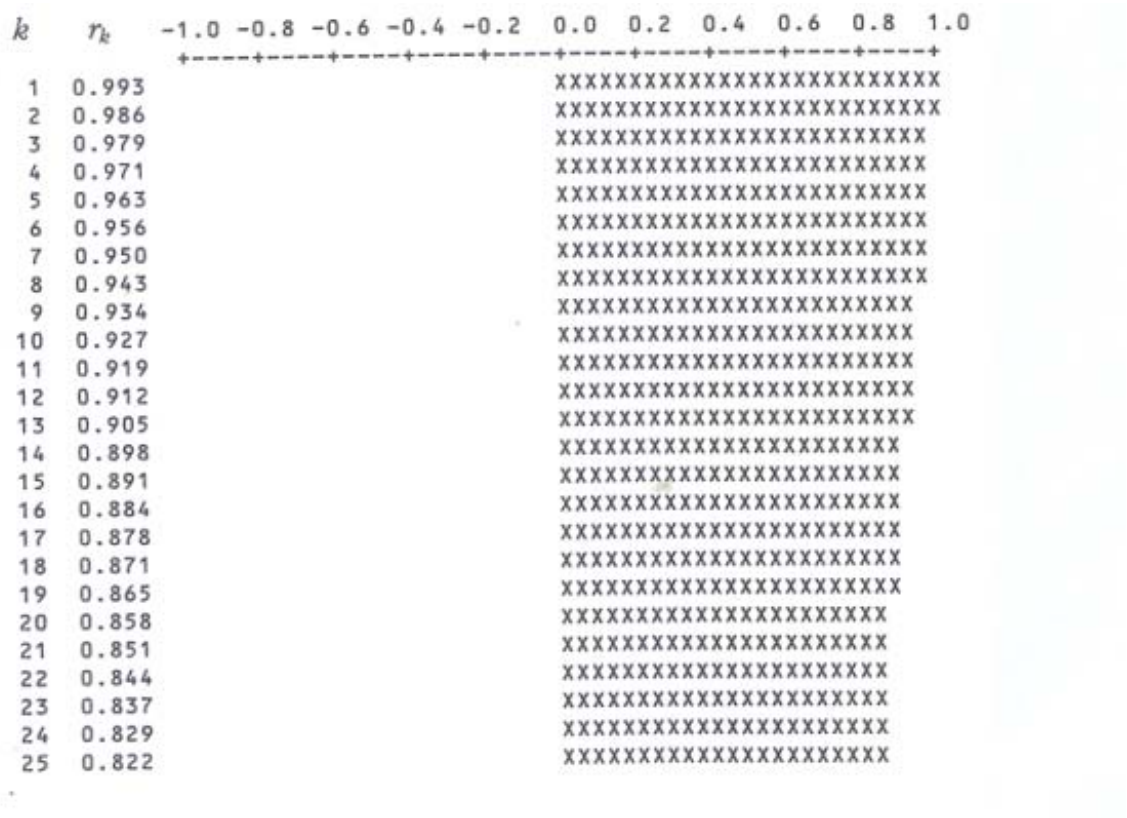


Figura 6

Infine, se la serie è non stazionaria, il modello di regressione:

$y_t = a + b\, y_{t-1+}\, e_t$

noto come modello autoregressivo del primo ordine presenta valori del coefficiente b non significativamente diversi da zero.
Si consideri il seguente esempio. La **Figura 7** mostra tra dollaro canadese e dollaro USA nel periodo compreso tra il 1986 e il 1995 (osservazioni settimanali). La serie è marcatamente non stazionaria in media, anche se non presenta una tendenza monotona.
La stima del modello autoregressivo del primo ordine fornisce i seguenti risultati:

(b)

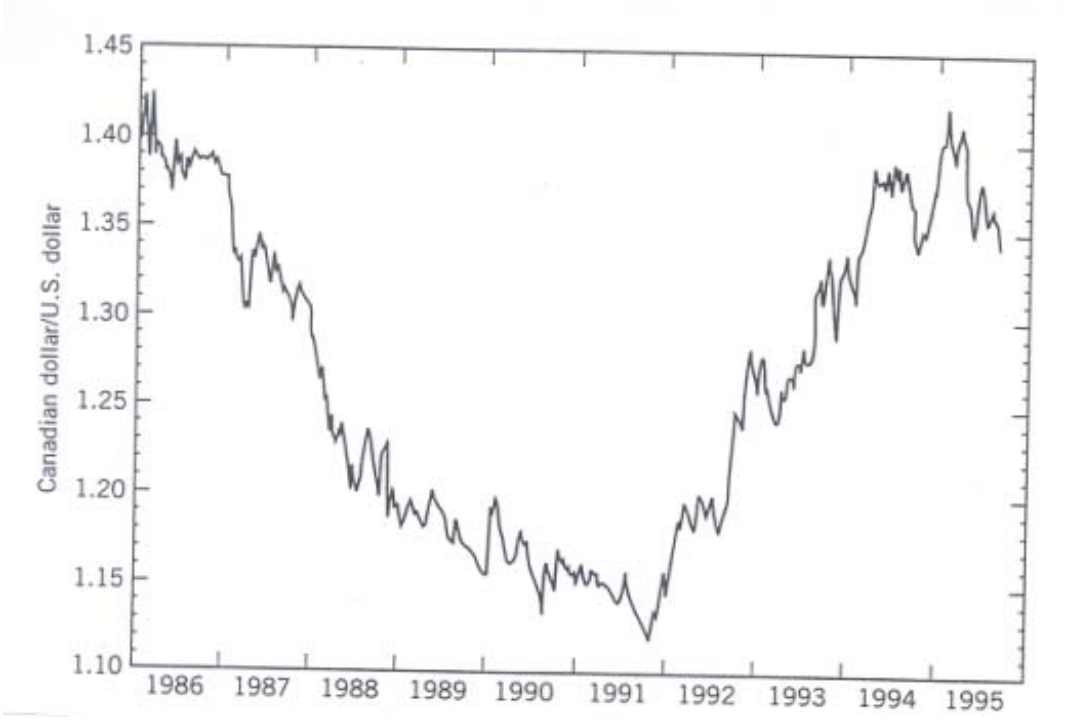
Variabile	Coefficiente	Errore standard	t di Student	significatività
Costante	0,007204	0,005171	1,39	0,164
y_{t-1}	0,994211	0,004078	234,79	0,000
R^2 = 99,2				
Analisi della	varianza			
Fonte	Gradi libertà	Somma quadrati	Errore medio	F

Regressione	1	4,0018	4,0018	59432,02
Residuo	501	0,0337	0,0001	
Totale	502	4,0355		

E' evidente che il coefficiente della variabile tasso di cambio ritardato di un periodo non è significativamente diverso da uno, mentre l'intercetta non è significativamente diversa da zero. In questi casi il modello che si adatta alla serie è il seguente:

$y_t = a + y_{t-1+} e_t$

che è noto anche come modello *random walk*, che è tipico per le serie non stazionarie in media. Se le serie coinvolte nel modello sono di questo tipo è necessario ricorre a metodi di stima dei coefficienti peculiari, noti in letteratura come modelli di cointegrazione, la cui trattazione non rientra negli scopi di questi appunti.



The regression equation is
CN/US Rt = 0.00720 + 0.994 CN/USLg1

Predictor	Coef	Stdev	t-ratio	p
Constant	0.007204	0.005171	1.39	0.164
CN/USLg1	0.994211	0.004078	243.79	0.000

s = 0.008206 R-sq = 99.2% R-sq(adj) = 99.2%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	4.0018	4.0018	59432.02	0.000
Error	501	0.0337	0.0001		
Total	502	4.0355			

Figura 7