# Econometria *for dummies*

Sergio Polini

24 giugno 2010

# Indice


**1 Introduzione** **1**
- 1.1 Articolazione . . . 2
- 1.2 Notazione . . . 2

**I Dati *cross-section*** **5**

**2 La regressione lineare** **7**
- 2.1 Aspettativa condizionata . . . 7
- 2.2 L'errore della regressione . . . 8
- 2.3 Varianza condizionata . . . 8
- 2.4 La regressione lineare . . . 9
  - 2.4.1 La regressione lineare come proiezione ortogonale . . . 10
  - 2.4.2 Il problema dell'identificazione . . . 13
  - 2.4.3 Il coefficiente di determinazione . . . 14
  - 2.4.4 Il modello lineare normale . . . 15
- 2.5 Applicazione a campioni di ampiezza finita . . . 17
  - 2.5.1 Valore atteso e varianza dello stimatore OLS . . . 17
  - 2.5.2 Il teorema di Gauss-Markov . . . 19
  - 2.5.3 I residui . . . 19
  - 2.5.4 Stima della varianza dell'errore . . . 20
  - 2.5.5 Multicollinearità . . . 20
- 2.6 Necessità di un approccio asintotico . . . 20

**3 L'ipotesi di esogeneità** **23**
- 3.1 L'importanza dell'ipotesi . . . 23
- 3.2 La stima dei parametri . . . 25
  - 3.2.1 Consistenza . . . 26
  - 3.2.2 Normalità asintotica . . . 26
  - 3.2.3 Stima della varianza . . . 28
- 3.3 Test di ipotesi e intervalli di confidenza . . . 31
  - 3.3.1 Test $z$ . . . 31
  - 3.3.2 Intervalli di confidenza . . . 33
  - 3.3.3 Test di Wald . . . 34
  - 3.3.4 Test $F$ . . . 36
- 3.4 Il problema delle variabili omesse . . . 38

3.5 Il problema degli errori di misura . . . . . 40

**4 Le variabili strumentali** **43**
4.1 Una sola variabile strumentale . . . . . 43
4.2 Più variabili strumentali . . . . . 45

**5 Variabile risposta qualitativa** **49**
5.1 Logit e probit . . . . . 49

**II Serie storiche** **51**

**6 La regressione spuria** **53**
6.1 Matrimoni religiosi e mortalità . . . . . 53
6.2 Processi stocastici . . . . . 54
6.2.1 Con memoria . . . . . 54
6.2.2 Senza memoria . . . . . 55
6.3 Definizioni . . . . . 56
6.3.1 Persistenza . . . . . 57
6.3.2 Stazionarietà ed ergodicità . . . . . 57
6.3.3 *White noise* e *Random walk* . . . . . 59
6.3.4 Cointegrazione . . . . . 61

**7 I processi ARMA** **63**
7.1 L: l'operatore ritardo . . . . . 63
7.2 MA: processi a media mobile . . . . . 64
7.2.1 Medie mobili finite . . . . . 64
7.2.2 Medie mobili infinite . . . . . 65
7.3 AR: processi autoregressivi . . . . . 66
7.3.1 Processi AR(1) . . . . . 67
7.3.2 Processi AR(p) . . . . . 68
7.4 ARMA: una generalizzazione . . . . . 71
7.5 Inferenza . . . . . 71
7.5.1 Consistenza e normalità asintotica . . . . . 72
7.5.2 Test di radice unitaria . . . . . 74
7.5.3 Test di stazionarietà . . . . . 74
7.5.4 La scomposizione di Beveridge-Nelson . . . . . 75

**8 I processi VAR** **77**
8.1 Macroeconomia e realtà . . . . . 77
8.2 Condizioni di stazionarietà . . . . . 78
8.3 Inferenza . . . . . 80

**9 Cointegrazione** **81**
9.1 Definizioni . . . . . 81
9.2 Modelli a correzione d'errore . . . . . 82
9.3 Il teorema di rappresentazione di Granger . . . . . 83

**III Appendici** **85**

**A Complementi di algebra lineare** **87**
A.1 Matrici inverse e inverse generalizzate . . . . . 87
A.2 Matrici di proiezione . . . . . 89
A.3 Immagine di una matrice . . . . . 91
A.4 Proiezione ortogonale sull'immagine di una matrice . . . . . 94

**B Equazioni alle differenze** **95**
B.1 Equazioni alle differenze del primo ordine . . . . . 95
B.2 Equazioni alle differenze di ordine $p$ . . . . . 96

**C Richiami di probabilità e di statistica** **103**
C.1 Variabili aleatorie multidimensionali . . . . . 103
C.2 Aspettativa condizionata . . . . . 104
C.2.1 Legge dell'aspettativa totale (LTE) . . . . . 104
C.2.2 Legge della varianza totale (LTV) . . . . . 105
C.3 La funzione caratteristica di una variabile aleatoria . . . . . 105
C.4 Successioni di variabili aleatorie . . . . . 106
C.4.1 Convergenza in distribuzione e in probabilità . . . . . 106
C.4.2 La legge dei grandi numeri . . . . . 107
C.4.3 Il teorema del limite centrale . . . . . 108

# Capitolo 1

# Introduzione

Questi appunti si basano prevalentemente sulle lezioni e le dispense del corso di econometria del prof. Massimo Franchi (Università di Roma La Sapienza, Facoltà di Scienze Statistiche, a.a. 2009-2010, http://w3.uniroma1.it/mfranchi/) e sui testi da lui indicati:

– Jeffrey M. Wooldridge (2002), *Econometric Analysis of Cross Section and Panel Data*;
– James D. Hamilton (1994), *Time Series Analysis*.

Mi sono poi avvalso di altri testi trovati "navigando nella Rete". In realtà, ho iniziato dando un'occhiata a gretl (http://gretl.sourceforge.net/), un software *open source* per l'analisi econometrica, e al suo notevole manuale utente (Cottrell e Lucchetti 2010). Da qui agli *Appunti di analisi delle serie storiche* del prof. Riccardo Lucchetti (Univestità Politecnica delle Marche) il passo è stato breve.

Gli *Appunti* mi sono stati utili perché si propongono espressamente come una «introduzione divulgativa» (Lucchetti 2008, p. 69) e l'obiettivo appare perfettamente raggiunto; in particolare, concetti tutt'altro che banali come *persistenza* e, soprattutto, *ergodicità* vengono introdotti con parole semplici che ne spiegano il "senso", anche se non vengono definiti formalmente.

Un'affermazione a pag. 5, tuttavia, ha scatenato ulteriori curiosità: «In linea generale, si può dire che l'inferenza è possibile solo se il processo stocastico che si sta studiando è stazionario ed ergodico». La ricerca di un'esposizione un po' più formale, ma non... al livello di Hamilton, mi ha condotto al *draft graduate textbook* del prof. Bruce E. Hansen (2010), dell'Università del Wisconsin.

Il suo *Econometrics* contiene proprio quello che cercavo: una definizione accessibile di ergodicità e del teorema ergodico, accompagnata dalla dimostrazione della loro necessità per l'inferenza. In realtà la parte sulle serie storiche appare appena abbozzata ed è dichiaratamente incompleta, ma i capitoli sulla regressione si sono rivelati una piacevole sorpresa.

Vi è un riepilogo della regressione classica che mi è risultato molto utile dopo aver seguito il corso di Modelli statistici della prof.ssa Cecilia Vitiello. Quel corso, infatti, era espressamente dedicato agli studi sperimentali e al modello lineare normale con ipotesi di omoschedasticità.[1] Hansen rivisita la regressione preparando il terreno all'abbandono

[1] I miei appunti tratti da quel corso sono in http://web.mclink.it/MC1166/ModelliStatistici/ModStat.html.

di quell'ipotesi fin da pag. 15 e poi, quando giunge all'approccio asintotico, dimostra sia la normalità asintotica dello stimatore OLS nel caso generale dell'eteroschedasticità, sia la consistenza della matrice di White (che per Wooldridge è "solo" il problema 4.4).

## 1.1 Articolazione

Dopo letture così illuminanti, mi è sembrato utile mettere insieme note prima sparse e pensare perfino ad una tendenziale organicità. Ho quindi pomposamente articolato gli appunti in due parti, dati *cross section* e serie storiche, con l'intento di aggiungere in futuro una parte sui dati *panel*, nonché capitoli su altri aspetti non trattati durante il corso.

Il capitolo 2 riepiloga gli aspetti fondamentali della regressione lineare seguendo l'impostazione di Hansen, il capitolo 3 tratta dell'ipotesi di esogeneità integrando Wooldridge con Hansen. I capitoli 4 e 5, dedicati alle variabili strumentali e al caso di variabile risposta qualitativa, sono basati su Wooldridge ma sono ancora solo abbozzi.

Il capitolo 6 introduce le serie storiche muovendo dal problema posto da Yule (1926) e cerca soprattutto di definire alcuni concetti chiave: *persistenza stazionarietà*, *ergodicità*, *integrazione* e *cointegrazione*.

Il capitolo 7 è dedicato ai processi MA, AR e ARMA. Le condizioni di stazionarietà e le relative dimostrazioni, apprese dal corso, sono diventate condizioni e dimostrazioni di stazionarietà ed ergodicità grazie a Hansen e Hamilton.

Il capitolo 8 si apre con l'introduzione dei processi VAR da parte di Sims (1980): un interessante spezzone di storia dell'analisi econometrica illustrato negli *Appunti* del prof. Lucchetti ed anche, con maggiore dettaglio, in altre dispense trovate in Rete (Carlucci e Girardi sd). Seguono le condizioni di stazionarietà e la relativa dimostrazione come apprese nel corso ma estese anche qui all'ergodicità. Il capitolo si conclude con accenni ai test di radice unitaria e di stazionarietà ed alla scomposizione di Beveridge-Nelson, tratti anch'essi dagli *Appunti* del prof. Lucchetti.

Il capitolo 9 è dedicato alla cointegrazione, ai modelli a correzione d'errore e al teorema di rappresentazione di Granger. È piuttosto sintetico perché la lettura di Engle e Granger (1987) e di Johansen (1991) mi ha fatto pensare che, per capire meglio, occorre estendere la casistica dei processi stocastici (introducendo trend lineari, intercette ecc.) rispetto a quanto trattato nel corso.

In sostanza, è solo un *work in progress* e, soprattutto, riflette quanto ho creduto di poter capire (il titolo, *Econometria for dummies*, è autoreferenziale).

## 1.2 Notazione

In matematica si usa scrivere le variabili con lettere minuscole in corsivo ($x^2 = 4$, $x = \pm 2$), i vettori e le matrici con lettere, rispettivamente, minuscole e maiuscole in neretto ($\mathbf{Ax} = \mathbf{b}$, $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$). In probabilità si usa scrivere le variabili aleatorie con lettere maiuscole ($Z \sim N(0,1)$), le loro realizzazioni con lettere minuscole.

In econometria è necessario esprimere sia modelli matematici che la loro interpretazione probabilistica. Si adotta quindi spesso una sorta di compromesso:

a) le lettere minuscole in corsivo indicano sempre scalari, siano essi variabili aleatorie oppure le loro realizzazioni, essendo normalmente chiaro dal contesto a cosa ci si riferisce; in particolare:
   - la variabile risposta compare senza indici quando ci si riferisce al modello della popolazione, con un indice $i = 1, 2, \ldots, n$ quando ci si riferisce alla $i$-esima unità del campione estratto (dati *cross-section*), oppure con un indice $t = 1, 2, \ldots, T$ quando ci si riferisce all'osservazione effettuata al tempo $t$ (serie storiche);
   - le variabili esplicative, quando indicate con una stessa lettera, vengono distinte mediante un indice $j = 1, 2, \ldots, k$; se $x_j$ è una variabile esplicativa, la sua realizzazione rilevata sull'$i$-esima unità si indica con $x_{ij}$;

b) le lettere minuscole in neretto indicano vettori; in particolare, se sono presenti $k$ variabili esplicative $x_j$, $j = 1, \ldots, k$, queste vengono collettivamente indicate con $\mathbf{x}$;

c) le lettere maiuscole in neretto indicano matrici; in particolare, le osservazioni delle realizzazioni di $k$ variabili esplicative $x_j$ su $n$ unità vengono collettivamente indicate con $\mathbf{X}$, una matrice di $n$ righe e $k$ colonne; le righe della matrice vengono indicate con $\mathbf{x}_i$ e intese come vettori colonna $k \times 1$ (si tratta delle $i$-esime realizzazioni di $k$ variabili aleatorie; in questo caso, quindi, $\mathbf{x}$ è un vettore di variabili aleatorie, $\mathbf{x}_i$ un vettore di loro realizzazioni);

d) le lettere greche indicano i parametri incogniti di un modello econometrico; se in neretto indicano vettori di parametri. Gli stimatori dei parametri vengono indicati ponendo un accento circonflesso "ˆ", detto comunemente *hat* (cappello), sul relativo simbolo oppure con la corrispondente lettera dell'alfabeto latino; ad esempio si possono usare sia $\hat{\beta}$ che $b$ per lo stimatore del parametro $\beta$.

In queste note, infine, uso parentesi quadre per vettori e matrici, ma parentesi tonde per indicare su una sola riga vettori colonna:

$$(x_1, \ldots, x_n) \equiv \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}'$$

# Parte I

# Dati *cross-section*

# Capitolo 2

# La regressione lineare

In econometria si usa spesso il metodo dei *minimi quadrati* (OLS, *Ordinary Least Squares*), noto anche come *regressione*, con il quale si cerca di stimare l'aspettativa condizionata di una variabile (detta *variabile risposta* o *variabile dipendente*) dato un insieme di altri variabili (dette *variabili esplicative*, o *regressori* o *covariate*). In questo capitolo si analizzano le proprietà della regressione, in particolare della regressione lineare, si richiamano gli aspetti fondamentali dell'applicazione della regressione a campioni di ampiezza finita, si conclude mostrando la necessità di un approccio asintotico nelle analisi econometriche.[1]

## 2.1 Aspettativa condizionata

Siano $y$ una variabile risposta e $\mathbf{x} = x_1, x_2, \ldots, x_k$ un vettore di variabili esplicative, tutte con momento secondo finito:

- $\mathbb{E}[y^2] < \infty$;
- $\mathbb{E}[x_j^2] < \infty$ per ogni $j = 1, \ldots, k$;

Tale ipotesi assicura che tutte le variabili abbiano media e varianza finite. In particolare, è necessario che $\mathbb{E}[|y|] < \infty$ perché possa esistere la sua aspettativa condizionata, definita come segue (v. anche l'appendice C per la definizione e le proprietà dell'aspettativa condizionata):

$$\mathbb{E}[y \mid \mathbf{x}] = \int_{-\infty}^{+\infty} y \, f(y \mid \mathbf{x}) \, dy$$

L'aspettativa condizionata di $y$ varia al variare di $\mathbf{x}$ ed è quindi una funzione $\mathbb{R}^k \to \mathbb{R}$. Viene anche detta *funzione di regressione*, in quanto lo scopo della regressione è appunto quello di stimare l'aspettativa condizionata di $y$ dato un valore di $\mathbf{x}$.

Ad esempio, se un modello è del tipo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u = \mathbf{x}'\boldsymbol{\beta} + u$$

dove $u$ viene detto *errore* (termine su cui si ritornerà), l'aspettativa condizionata di $y$ è:

$$\mathbb{E}[y \mid \mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = \mathbf{x}'\boldsymbol{\beta}$$

Obiettivo della regressione è trovare stime $\mathbf{b}$ per i parametri incogniti $\boldsymbol{\beta}$.

[1] Il capitolo si ispira largamente a Hansen (2010, capp. 2-4).

## 2.2 L'errore della regressione

L'errore $u$ è la differenza tra la variabile $y$ e la sua aspettativa condizionata:

$$u = y - \mathbb{E}[y \mid \mathbf{x}]$$

e gode delle seguenti proprietà:

1) $\mathbb{E}[u \mid \mathbf{x}] = 0$, infatti, per la linearità dell'aspettativa condizionata:

$$\mathbb{E}[u \mid \mathbf{x}] = \mathbb{E}[(y - \mathbb{E}[y \mid \mathbf{x}]) \mid \mathbf{x}] = \mathbb{E}[y \mid \mathbf{x}] - \mathbb{E}[y \mid \mathbf{x}] = 0$$

2) $\mathbb{E}[u] = 0$, infatti, per la legge dell'aspettativa totale:

$$\mathbb{E}[u] = \mathbb{E}\Big[\mathbb{E}[u \mid \mathbf{x}]\Big] = \mathbb{E}[0] = 0$$

3) $\mathbb{E}[f(\mathbf{x})u] = 0$ per qualsiasi funzione $f(\mathbf{x})$ a valori finiti; infatti, per la legge dell'aspettativa totale $\mathbb{E}[f(\mathbf{x})u] = \mathbb{E}[\mathbb{E}[f(\mathbf{x})u \mid \mathbf{x}]]$ e per per la linearità dell'aspettativa condizionata $\mathbb{E}[f(\mathbf{x})u \mid \mathbf{x}] = f(\mathbf{x})\mathbb{E}[u \mid \mathbf{x}]$, quindi:

$$\mathbb{E}[f(\mathbf{x})u] = \mathbb{E}\Big[\mathbb{E}[f(\mathbf{x})u \mid \mathbf{x}]\Big] = \mathbb{E}\Big[f(\mathbf{x})\mathbb{E}[u \mid \mathbf{x}]\Big] = 0$$

analogamente per una funzione a valori vettoriali $\mathbf{f}(\mathbf{x})$;

4) $\mathbb{E}[\mathbf{x}u] = \mathbf{0}$, caso particolare della precedente.

Va notato $\mathbb{E}[u \mid \mathbf{x}] = 0$ non comporta che $\mathbf{x}$ e $u$ siano indipendenti. Ad esempio, se si avesse $y = xv$, con $x$ e $v$ indipendenti e $\mathbb{E}[v] = 1$, si avrebbe anche $\mathbb{E}[y \mid x] = x$ e si potrebbe scrivere $y = x + u$ con $u = x(v - 1)$; in questo caso $u$ sarebbe chiaramente dipendente da $x$, ma si avrebbe comunque $\mathbb{E}[u \mid x] = 0$.

Da $\mathbb{E}[u] = 0$ e $\mathbb{E}[\mathbf{x}u] = 0$ segue invece che $\mathbf{x}$ e $u$ sono incorrelati:

$$\mathrm{Cov}(\mathbf{x}, u) = \mathbb{E}[\mathbf{x}u] - \mathbb{E}[\mathbf{x}]\mathbb{E}[u] = \mathbf{0}$$

## 2.3 Varianza condizionata

L'aspettativa condizionata fornisce una buona approssimazione della distribuzione condizionata di $y$, ma va considerata anche la dispersione di tale distribuzione, comunemente misurata dalla *varianza condizionata*:[2]

$$\begin{aligned}\mathbb{V}[y \mid \mathbf{x}] &= \mathbb{E}[y^2 \mid \mathbf{x}] - \mathbb{E}[y \mid \mathbf{x}]^2 \\ &= \mathbb{E}\Big[(y - \mathbb{E}[y \mid \mathbf{x}])^2 \mid \mathbf{x}\Big] \\ &= \mathbb{E}[u^2 \mid \mathbf{x}]\end{aligned}$$

[2] Si ha:

$$\begin{aligned}\mathbb{E}\big[(y - \mathbb{E}[y \mid \mathbf{x}])^2 \mid \mathbf{x}\big] &= \mathbb{E}\big[(y^2 + \mathbb{E}[y \mid \mathbf{x}]^2 - 2y\mathbb{E}[y \mid \mathbf{x}]) \mid \mathbf{x}\big] \\ &= \mathbb{E}[y^2 \mid \mathbf{x}] + \mathbb{E}\big[\mathbb{E}[y \mid \mathbf{x}]^2 \mid \mathbf{x}\big] - 2\mathbb{E}\big[y\mathbb{E}[y \mid \mathbf{x}] \mid \mathbf{x}\big] \\ &= \mathbb{E}[y^2 \mid \mathbf{x}] + \mathbb{E}[y \mid \mathbf{x}]^2 - 2\mathbb{E}\big[y\mathbb{E}[y \mid \mathbf{x}] \mid \mathbf{x}\big]\end{aligned}$$

poiché $\mathbb{E}[y \mid \mathbf{x}]$ è una funzione di $\mathbf{x}$, $\mathbb{E}\big[y\mathbb{E}[y \mid \mathbf{x}] \mid \mathbf{x}\big] = \mathbb{E}[y \mid \mathbf{x}]\mathbb{E}[y \mid \mathbf{x}]$:

$$= \mathbb{E}[y^2 \mid \mathbf{x}] + \mathbb{E}[y \mid \mathbf{x}]^2 - 2\mathbb{E}[y \mid \mathbf{x}]^2 = \mathbb{E}[y^2 \mid \mathbf{x}] - \mathbb{E}[y \mid \mathbf{x}]^2$$

La varianza condizionata è una funzione delle variabili esplicative $\mathbf{x}$, ma si considera spesso un caso particolare in cui ciò non avviene. Si distingue quindi tra due diverse situazioni:

a) *eteroschedasticità*: si tratta della situazione tipica e più frequente nella pratica; come appena visto:
$$\mathbb{V}[y \mid \mathbf{x}] = \mathbb{E}[u^2 \mid \mathbf{x}] = \sigma^2(\mathbf{x})$$
ovvero la varianza condizionata è funzione di $\mathbf{x}$ (qui $\sigma^2$ denota una funzione);

b) *omoschedasticità*: la varianza condizionata *non* dipende da $\mathbf{x}$:
$$\mathbb{V}[y \mid \mathbf{x}] = \mathbb{E}[u^2 \mid \mathbf{x}] = \mathbb{E}[u^2] = \sigma^2$$
(qui $\sigma^2$ è un numero).

L'ipotesi di omoschedasticità semplifica molto alcuni aspetti della teoria, ma non si deve dimenticare che si tratta solo di una comoda eccezione utile sul piano astratto.

Peraltro, anche assumendo eteroschedasticità è possibile definire $\sigma^2$ come varianza dell'errore:
$$\mathbb{E}\Big[(y - \mathbb{E}[y \mid \mathbf{x}])^2\Big] = \mathbb{E}[u^2] = \sigma^2$$
intendendola come valore atteso della varianza condizionata:
$$\sigma^2 = \mathbb{E}[u^2] = \mathbb{E}\Big[\mathbb{E}[u^2 \mid \mathbf{x}]\Big] = \mathbb{E}[\sigma^2(\mathbf{x})]$$

## 2.4 La regressione lineare

In generale, l'aspettativa condizionata può assumere qualsiasi forma funzionale. Si usa comunque spesso la forma *lineare nei parametri*:
$$\mathbb{E}[y \mid \mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$
dove il primo parametro, $\beta_0$, viene detto *intercetta*. Si dice lineare *nei parametri* perché i parametri $\beta_j$ compaiono tutti con esponente 1, ma nulla vieta che qualche $x_j$ sia una qualsiasi funzione di qualche altro; ad esempio, l'equazione precedente potrebbe essere in realtà:
$$\mathbb{E}[y \mid \mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \cdots + \beta_k x_1^k$$
con $x_j = x_1^j$.

Quando si scrive l'aspettativa condizionata come funzione di un vettore, $\mathbb{E}[y \mid \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, si intende $\mathbf{x}$ come un vettore di $k+1$ elementi il primo dei quali sia 1:
$$\mathbb{E}[y \mid \mathbf{x}] = \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_k \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$
Altre volte, in modo del tutto equivalente (forse preferibile), si intende $x_1 = 1$ e si scrive:
$$\mathbb{E}[y \mid \mathbf{x}] = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \begin{bmatrix} 1 & x_2 & \dots & x_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$
intendendo $\mathbf{x}$ come vettore di $k$ elementi.

### 2.4.1 La regressione lineare come proiezione ortogonale

La forma lineare dell'aspettativa condizionata $\mathbb{E}[y \mid \mathbf{x}]$ è semplice, ma probabilmente poco accurata sul piano empirico, niente più che un'approssimazione. Per migliorare la qualità dell'approssimazione si cerca di minimizzare l'errore quadratico medio ($MSE$, *Mean Squared Error*):

$$S(\boldsymbol{\beta}) = \mathbb{E}[u^2] = \mathbb{E}[(y - \mathbf{x}'\boldsymbol{\beta})^2]$$

che può essere riscritta così:

$$S(\boldsymbol{\beta}) = \mathbb{E}[y^2] - 2\boldsymbol{\beta}'\mathbb{E}[\mathbf{x}y] + \boldsymbol{\beta}'\mathbb{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta}$$

La condizione del primo ordine per la minimizzazione è:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbb{E}[\mathbf{x}y] + 2\mathbb{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta} = 0$$

da cui:

$$\mathbb{E}[\mathbf{x}y] = \mathbb{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta}$$

Se ora si assume che $\mathbb{E}[\mathbf{x}\mathbf{x}']$ sia una matrice a rango pieno, quindi invertibile, si ottiene:

$$\boldsymbol{\beta} = \mathbb{E}[\mathbf{x}\mathbf{x}']^{-1}\mathbb{E}[\mathbf{x}y]$$

Il parametro $\boldsymbol{\beta}$ così definito viene detto *coefficiente di regressione*, o anche *coefficiente di proiezione lineare*. Analogamente, l'errore $u = y - \mathbf{x}'\boldsymbol{\beta}$ viene detto *errore di proiezione*.

Il motivo per cui si parla di proiezione risulta più chiaro se si passa alla stima di $\boldsymbol{\beta}$. Una volta definito un modello quale $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$, si osservano i valori di $y$ e di $\mathbf{x}$ su $n$ unità e si ottengono $n$ osservazioni del tipo:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i \qquad\qquad y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

In forma matriciale:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

dove:

- $\mathbf{y}$ è un vettore $n \times 1$ contenente le $n$ osservazioni della variabile risposta;
- $\mathbf{X}$ è una matrice $n \times k$ contenente in ciascuna riga le $k$ osservazioni delle variabili esplicative sull'unità $i$-esima; la prima colonna è costituita da tutti 1;
- $\mathbf{x}_i$ è il vettore colonna della $i$-esima riga della matrice $\mathbf{X}$;
- $\boldsymbol{\beta}$ è un vettore $k \times 1$ contenente i parametri (i coefficienti di regressione o di proiezione);
- $\mathbf{u}$ è un vettore $n \times 1$.

L'errore quadratico medio da minimizzare diventa:

$$S_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$$

dove le differenze $y_i - \mathbf{x}_i\boldsymbol{\beta}$ vengono dette *residui* e spesso indicate con $e_i$.

Essendo $n$ dato, si tratta di minimizzare la somma dei quadrati dei residui $RSS_n(\boldsymbol{\beta}) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$ (*Residual Sum of Squares*) e si ha:

$$RSS_n(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

$$\frac{\partial\, RSS_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0 \quad \Rightarrow \quad \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Se $\mathbf{X}'\mathbf{X}$ risulta, oltre che simmetrica, anche invertibile, si ottiene $\mathbf{b}$ come stima di $\boldsymbol{\beta}$ da:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^n \mathbf{x}_i y_i\right) = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i y_i\right)$$

In sostanza, si stimano i momenti di popolazione $\mathbb{E}[\mathbf{x}\mathbf{x}']$ e $\mathbb{E}[\mathbf{x}y]$ con le rispettive medie campionarie.

È questo il *metodo dei minimi quadrati*, detto anche OLS (*Ordinary Least Squares*). Lo stimatore così ottenuto viene quindi detto *stimatore OLS*.

L'aspettativa condizionata $\mathbb{E}[y \mid \mathbf{x}]$ viene stimata da

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

La matrice $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ risulta simmetrica (in quanto prodotto di matrici con le loro trasposte) e idempotente, in quanto:

$$\begin{aligned}\mathbf{H}^2 &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{X}\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}\end{aligned}$$

È quindi una matrice di proiezione ortogonale di rango $k$ che proietta $\mathbf{y}$ sullo spazio generato dalle colonne di $\mathbf{X}$ (cfr. l'appendice A).

I *residui* $\mathbf{e}$, a loro volta, sono dati da:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

dove $\mathbf{I} - \mathbf{H}$ è una matrice di rango $n - k$, anch'essa simmetrica e idempotente, che proietta $\mathbf{y}$ in uno spazio che è il complemento ortogonale di quello generato dalle colonne di $\mathbf{X}$. La lunghezza del vettore $\mathbf{e}$ misura quindi la distanza tra $\mathbf{y}$ e la sua proiezione ortogonale $\hat{\mathbf{y}}$ (v. figura 2.1).

Analogamente a quanto si ha per l'errore, anche i residui hanno media nulla e sono incorrelati con le variabili esplicative. Infatti:

$$\begin{aligned}\mathbb{E}[\mathbf{e} \mid \mathbf{X}] &= \mathbb{E}[\mathbf{y} \mid \mathbf{X}] - \mathbb{E}[\mathbf{X}\mathbf{b} \mid \mathbf{X}] = \mathbb{E}[\mathbf{y} \mid \mathbf{X}] - \mathbb{E}[(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) \mid \mathbf{X}] \\ &= \mathbb{E}[\mathbf{y} \mid \mathbf{X}] - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{y} \mid \mathbf{X}] \\ &= \mathbf{X}\mathbf{b} - \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})]\mathbf{b} = \mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b} = \mathbf{0}\end{aligned}$$

Da ciò seguono $\mathbb{E}[\mathbf{e}] = 0$ e $\mathbb{E}[\mathbf{X}\mathbf{e}] = 0$, quindi anche:

$$\mathrm{Cov}(\mathbf{X}, \mathbf{e}) = \mathbb{E}[\mathbf{X}\mathbf{e}] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{e}] = 0$$

che è un altro modo di esprimere il fatto che, da un punto di vista geometrico, il vettore $\mathbf{e}$ è ortogonale al piano generato dalle colonne di $\mathbf{X}$.

**Esempio 2.1.** Sia $y$ una variabile che si ritiene spiegata da una sola variabile esplicativa. Siano $\mathbf{x} = (1, 2, 3)$ e $\mathbf{y} = (2.9, 5.2, 6.9)$ i valori osservati su tre unità. Si ha:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad \begin{bmatrix} 2.9 \\ 5.2 \\ 6.9 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{u}$$

La stima di $\boldsymbol{\beta}$ porta a:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left( \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2.9 \\ 5.2 \\ 6.9 \end{bmatrix} = \begin{bmatrix} b_1 = 1 \\ b_2 = 2 \end{bmatrix}$$

Oppure, con R:

```
> x <- c(1, 2, 3)
> y <- c(2.9, 5.2, 6.9)
> reg <- lm(y ~ x)
> coef(reg)
(Intercept)           x
          1           2
```

Ne seguono le stime $\hat{\mathbf{y}}$ dell'aspettativa condizionata, dette *valori teorici* o *valori predetti*, e i residui (che hanno media 0):

$$\begin{aligned} \hat{y}_1 &= b_1 + b_2 x_{12} = 1 + 2 \cdot 1 = 3 & \qquad e_1 &= y_1 - \hat{y}_1 = 2.9 - 3 = -0.1 \\ \hat{y}_2 &= b_1 + b_2 x_{22} = 1 + 2 \cdot 2 = 5 & \qquad e_2 &= y_2 - \hat{y}_2 = 5.2 - 5 = \phantom{-}0.2 \\ \hat{y}_3 &= b_1 + b_2 x_{32} = 1 + 2 \cdot 3 = 7 & \qquad e_3 &= y_3 - \hat{y}_3 = 6.9 - 7 = -0.1 \end{aligned}$$

Con R:

```
> predict(reg)
1 2 3
3 5 7
> residuals(reg)
   1    2    3
-0.1  0.2 -0.1
```

La matrice $\mathbf{H}$ è:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{bmatrix} 5/6 & 1/3 & -1/6 \\ 1/3 & 1/3 & 1/3 \\ -1/6 & 1/3 & 5/6 \end{bmatrix}$$

Lo spazio su cui $\mathbf{y}$ viene proiettato è l'immagine della matrice $\mathbf{H}$, ovvero lo spazio generato dalle sue colonne linearmente indipendenti. Dato che $\mathbf{H}$ risulta dal prodotto di matrici di rango 2 e delle loro trasposte, ha anch'essa rango 2. Essendo peraltro simmetrica, è possibile e conveniente diagonalizzarla, pervenendo a $\mathbf{H} = \mathbf{M}\boldsymbol{\Lambda}\mathbf{M}^{-1}$:

$$\begin{bmatrix} 5/6 & 1/3 & -1/6 \\ 1/3 & 1/3 & 1/3 \\ -1/6 & 1/3 & 5/6 \end{bmatrix} = \begin{bmatrix} 2 & -1 & 1 \\ 1 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & -1 & 1 \\ 1 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix}^{-1}$$

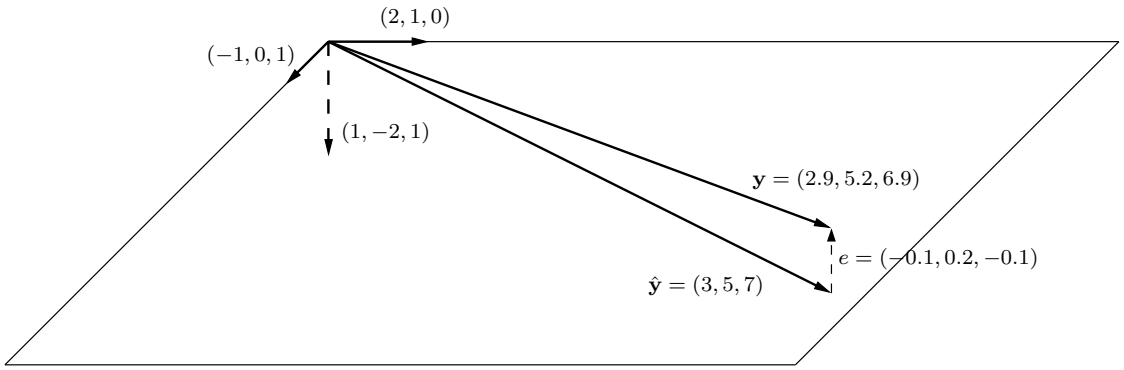


*Figura 2.1.* La regressione lineare come proiezione ortogonale.

Si ottengono così tre autovettori (le colonne di $\mathbf{M}$), i primi due dei quali, essendo non nulli i relativi autovalori, costituiscono una base dell'immagine. Si nota anche che il terzo autovettore (una base del kernel) è ortogonale ai primi due, che generano il piano cui appartiene il vettore $\hat{\mathbf{y}}$:

$$\begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix} = 5 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + 7 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

I residui appartengono invece allo spazio immagine della matrice $\mathbf{I} - \mathbf{H}$; diagonalizzando:

$$\begin{bmatrix} 1/6 & -1/3 & 1/6 \\ -1/3 & 2/3 & -1/3 \\ 1/6 & -1/3 & 1/6 \end{bmatrix} = \begin{bmatrix} 1 & 2 & -1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}^{-1}$$

si ritrovano gli stessi autovettori, ma ora c'è un solo autovalore non nullo e il relativo autovettore, che costituisce una base dell'immagine, è ortogonale agli altri due. Si vede così che il vettore dei residui, $(-0.1, 0.2, -0.1) = -\frac{1}{10}(1, -2, 1)$, appartiene ad uno spazio ad una dimensione *ortogonale* a quello cui appartiene il vettore delle stime (v.figura 2.1).

### 2.4.2 Il problema dell'identificazione

Si dice che il vettore $\boldsymbol{\beta}$ è *identificato* quando è univocamente determinato. Il problema dell'identificazione, nel caso della regressione lineare, si riduce al rango della matrice $k \times k$ $\mathbb{E}[\mathbf{x}\mathbf{x}']$: se la matrice è a rango pieno, l'equazione

$$\mathbb{E}[\mathbf{x}y] = \mathbb{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta}$$

ha un'unica soluzione, si possono cioè trovare valori univoci per i $k$ parametri $\beta_j$.

In caso contrario, l'equazione ha infinite soluzioni. Si può trovare una soluzione usando la pseudoinversa di Moore-Penrose (v. appendice A):

$$\boldsymbol{\beta} = \mathbb{E}[\mathbf{x}\mathbf{x}']^{+}\mathbb{E}[\mathbf{x}y]$$

ma risulta così identificata solo l'aspettativa condizionata $\mathbb{E}[y \mid \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, non anche i singoli elementi di $\boldsymbol{\beta}$.

### 2.4.3 Il coefficiente di determinazione

I dati osservati nel vettore $\mathbf{y}$ presentano una variabilità che si tenta di spiegare con la sua proiezione $\hat{\mathbf{x}}$ sul piano generato dalle colonne della matrice $\mathbf{X}$. In tale contesto, una misura tipica della variabilità è costituita dalla somma dei quadrati degli scarti tra i singoli valori di $\mathbf{y}$ e la loro media aritmetica $\bar{y}$, che viene detta $TSS$ (*Total Sum of Squares*). Analogamente, viene detta $ESS$ (*Explained Sum of Squares*) la somma degli scarti degli $\hat{y}_i$ dalla media $\bar{y}$. Si verifica facilmente che:

$$TSS = ESS + RSS$$

ovvero:

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

Si dice anche che la *devianza totale* è uguale alla somma della *devianza spiegata* e della *devianza residua*.

Si usa calcolare la bontà dell'adattamento della funzione di regressione ai dati mediante il rapporto tra devianza spiegata e devianza totale, detto *coefficiente di determinazione multipla* e indicato con $R^2$:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \qquad 0 \leq R^2 \leq 1$$

Si considera l'adattamento tanto migliore quanto più $R^2$ si avvicina a 1.

In realtà $R^2$ aumenta con l'aumentare del numero delle variabili esplicative. Per tenere conto di ciò, Henri Theil propose un $R^2$ *corretto*:

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

dove $n-k$ sono i gradi di libertà della devianza residua ($n$ e $k$ sono le dimensioni della matrice $mX$) e $n-1$ quelli della devianza totale.

Si deve inoltre tenere presente che non esiste alcuna "legge" che stabilisca un'associazione tra il valore dei coefficienti di determinazione e la "bontà" di una regressione, e che anche in caso di valori "piccoli" è possibile una stima accurata dei coefficienti di regressione se l'ampiezza del campione è grande.

**Esempio 2.2.** Usando la semplice regressione dell'esempio precedente:

$$\begin{aligned}
TSS &= (2.9-5)^2 + (5.2-5)^2 + (6.9-5)^2 = 4.41 + 0.04 + 3.61 = 8.06 \\
ESS &= (3-5)^2 + (5-5)^2 + (7-5)^2 = 4 + 0 + 4 = 8 \\
RSS &= (2.9-3)^2 + (5.2-5)^2 + (6.9-7)^2 = 0.01 + 0.04 + 0.01 = 0.06 \\
R^2 &= 8/8.06 = 0.9926 \\
\overline{R}^2 &= 1 - \frac{0.06/(3-2)}{TSS/(3-1)} = 0.9851
\end{aligned}$$

La figura 2.2 mostra l'output del comando `summary()` di R, con i coefficienti $R^2$ e $\overline{R}^2$ insieme ad altri risultati che verranno commentati nella sezione successiva.

```
> summary(reg)

Call:
lm(formula = y ~ x)

Residuals:
   1    2    3
-0.1  0.2 -0.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0000     0.3742   2.673    0.228
x             2.0000     0.1732  11.547    0.055 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2449 on 1 degrees of freedom
Multiple R-squared: 0.9926,Adjusted R-squared: 0.9851
F-statistic: 133.3 on 1 and 1 DF,  p-value: 0.055
```

*Figura 2.2.* Output del comando `summary()` di R per una semplice regressione di $\mathbf{y} = (2.9, 5.2, 6.9)$ su $\mathbf{x} = (1, 2, 3)$.

### 2.4.4 Il modello lineare normale

Negli studi sperimentali molto spesso l'errore viene indicato con $\varepsilon$ e si assume che sia distribuito normalmente; poiché $\mathbf{x}'\boldsymbol{\beta}$ è il prodotto di un vettore di dati osservati e di parametri, ne segue che anche $y$ è una variabile aleatoria normale, in quanto trasformazione lineare di una variabile aleatoria normale.

L'ipotesi di normalità comporta anche che, se $y$ e $\varepsilon$ sono incorrelati, sono anche indipendenti; da ciò segue naturalmente l'ipotesi di omoschedasticità:

$$\begin{cases} y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \quad \Rightarrow \quad \begin{cases} \mathbb{E}[y \mid \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} \\ \mathbb{V}[y \mid \mathbf{x}] = \mathbb{E}[\varepsilon^2 \mid \mathbf{x}] = \mathbb{E}[\varepsilon^2] = \sigma^2 \end{cases} \quad \Rightarrow \quad y \sim N(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$$

In econometria l'ipotesi di normalità non appare utile, in quanto i dati economici ben difficilmente presentano distribuzioni normali. Si può comunque notare che, data la funzione di densità:

$$f(\mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

la funzione di log-verosimiglianza è:

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}$$

Si vede che $\ell(\boldsymbol{\beta}, \sigma^2)$, per qualsiasi valore di $\sigma^2$, è massimizzata dai valori di $\boldsymbol{\beta}$ che minimizzano il numeratore dell'ultimo termine, che a sua volta altro non è che la quantità $RSS_n(\boldsymbol{\beta})$, minimizzata da $\mathbf{b}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Il metodo di massima verosimiglianza porta dunque ad uno stimatore uguale a quello OLS.

L'ipotesi di normalità consente di definire test per la verifica di ipotesi sia sulla stima dei singoli coefficienti di regressione, sia sull'intera funzione di regressione.

Quanto alla stima di un singolo coefficiente, da $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ e da $\mathbb{V}[y \mid \mathbf{x}] = \sigma^2$ segue:

$$\mathbb{E}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$
$$\mathrm{Cov}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Indicando con $a_{ii}$ l'$i$-esimo elemento della diagonale principale della matrice $(\mathbf{X}'\mathbf{X})^{-1}$:

$$\mathbb{E}[b_i] = \beta_i,\ \mathbb{V}[b_i] = \sigma^2 a_{ii} \qquad \text{ovvero:} \qquad b_i \sim N(\beta_1, \sigma^2 a_{ii})$$

Sotto ipotesi nulla $\beta_i = 0$, si può definire la variabile normale standard $\dfrac{b_i - 0}{\sqrt{\sigma^2 a_{ii}}}$. Poiché $\sigma^2$ non è nota, si può sostituire con una stima data dalla devianza residua divisa per i suoi gradi di libertà, ottenendo così la statistica test:

$$t^* = \frac{b_i}{\sqrt{\dfrac{RSS}{n-k}a_{ii}}} \sim t_{n-k}$$

che è distribuita come una $t$ di Student. Il denominatore $\sqrt{\dfrac{RSS}{n-k}a_{ii}}$ viene detto *errore standard* (*standard error*).

Quanto all'intero modello ci si avvale del teorema di Cochran, che può essere formulato come segue:

**Teorema 2.3** (Cochran)**.** *Se $n$ osservazioni $y_i$ provengono dalla stessa distribuzione normale con media $\mu$ e varianza $\sigma^2$, se la devianza totale $TSS$ è scomposta nella somma di una devianza spiegata $ESS$ con $k-1$ gradi di libertà e di una devianza residua $RSS$ con $n-k$ gradi di libertà, allora $ESS/\sigma^2$ e $RSS/\sigma^2$ si distribuiscono come $\chi^2$ indipendenti con gradi di libertà, rispettivamente, $k-1$ e $n-k$:*

$$\frac{ESS}{\sigma_2} \sim \chi^2_{k-1} \qquad\qquad \frac{RSS}{\sigma_2} \sim \chi^2_{n-k}$$

L'ipotesi nulla consiste nel supporre pari a zero tutti i coefficienti tranne l'intercetta (in due dimensioni, retta di regressione orizzontale). Ciò vuol dire ipotizzare che tutti gli $y_i$ siano uguali all'intercetta e che abbiano pertanto la stessa media, oltre che la stessa varianza. Si può quindi costruire la statistica test:

$$F^* = \frac{\dfrac{ESS}{\sigma^2}/(k-1)}{\dfrac{RSS}{\sigma^2}/(n-k)} = \frac{\dfrac{ESS}{k-1}}{\dfrac{RSS}{n-k}} \sim F_{k-1,n-k}$$

che si distribuisce come una $F$ di Snedecor.

I software statistici propongono sempre i risultati dei test $t$ e $F$ (v. esempio 2.2). Tuttavia, se non si assume normalità i test devono essere diversamente fondati.

## 2.5 Applicazione a campioni di ampiezza finita

La regressione lineare è tradizionalmente applicata a campioni di ampiezza finita. Con ciò si intende che, data una popolazione, si assume che si possano estrarre da essa più campioni di ampiezza $n$ e si considerano le possibilità di inferenza sui parametri della popolazione per $n$ dato. In econometria, per i motivi che si vedranno, si preferisce un approccio asintotico: si cerca di inferire i parametri della popolazione sulla base di un campione di ampiezza $n \to \infty$. È comunque opportuno approfondire alcuni aspetti dell'approccio tradizionale.

Gli assunti di partenza sono:

1) *indipendenza e identica distribuzione*: possibilità di estrarre campioni casuali contenenti $n$ osservazioni $y_i, \mathbf{x}_i$;
2) *linearità*: esistenza di una relazione lineare del tipo

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

3) *attesa condizionata nulla dell'errore*:

$$\mathbb{E}[u_i \mid \mathbf{x}_i] = 0$$

4) *momenti secondi finiti per tutte le variabili*:

$$\mathbb{E}[y_i^2] < \infty \qquad \forall j = 2, \ldots, k : \ \mathbb{E}[x_{ij}^2] < \infty$$

5) *invertibilità* della matrice $\mathbb{E}[\mathbf{x}\mathbf{x}']$:

$$\operatorname{rk}(\mathbb{E}[\mathbf{x}\mathbf{x}']) = k$$

### 2.5.1 Valore atteso e varianza dello stimatore OLS

L'attesa condizionata di $\mathbf{y}$, il vettore degli $n$ valori della variabile risposta, rispetto ai valori delle $k$ variabili esplicative, è:

$$\mathbb{E}[\mathbf{y} \mid \mathbf{X}] = \begin{bmatrix} \vdots \\ \mathbb{E}[y_i \mid \mathbf{X}] \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \mathbb{E}[y_i \mid \mathbf{x}_i] \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \mathbf{x}_i\boldsymbol{\beta} \\ \vdots \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}$$

Ne segue che lo stimatore $\mathbf{b}$ di $\boldsymbol{\beta}$ è uno stimatore *corretto*:

$$\mathbb{E}[\mathbf{b} \mid \mathbf{X}] = \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \mid \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{y} \mid \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

La correttezza implica che, ripetendo la regressione su più campioni, la media degli stimatori tende al valore vero del parametro.

Applicando la legge dell'aspettativa totale si ha anche:

$$\mathbb{E}[\mathbf{b}] = \mathbb{E}\Big[\mathbb{E}[\mathbf{b} \mid \mathbf{X}]\Big] = \boldsymbol{\beta}$$

Si tratta di un risultato che rafforza il precedente, in quanto afferma che $\mathbf{b}$ è uno stimatore corretto quale che sia la matrice $\mathbf{X}$.

Quanto alla varianza, in generale per un vettore $n \times 1$ di variabili aleatorie $\mathbf{z}$ si ha:

$$\mathbb{V}[\mathbf{z}] = \mathbb{E}\Big[(\mathbf{z} - \mathbb{E}[\mathbf{z}])(\mathbf{z} - \mathbb{E}[\mathbf{z}])'\Big] = \mathbb{E}[\mathbf{z}\mathbf{z}'] - \mathbb{E}[\mathbf{z}]\mathbb{E}[\mathbf{z}]'$$

che è una matrice $n \times n$. La varianza condizionata rispetto ad una matrice $\mathbf{X}$ è invece:

$$\mathbb{V}[\mathbf{z} \mid \mathbf{X}] = \mathbb{E}\Big[(\mathbf{z} - \mathbb{E}[\mathbf{z} \mid \mathbf{X}])(\mathbf{z} - \mathbb{E}[\mathbf{z} \mid \mathbf{X}])' \mid \mathbf{X}\Big]$$

Poiché $\mathbb{E}[\mathbf{u} \mid \mathbf{X}] = \mathbf{0}$, la matrice di varianza e covarianza condizionate del vettore $\mathbf{u}$ è una matrice diagonale $n \times n$:

$$\mathbf{D} = \mathbb{E}[\mathbf{u}\mathbf{u}' \mid \mathbf{X}]$$

Si tratta di una matrice diagonale in quanto gli elementi della diagonale prinicipale sono:

$$\mathbb{E}[u_i^2 \mid \mathbf{X}] = \mathbb{E}[u_i^2 \mid \mathbf{x}_i] = \sigma_i^2$$

mentre gli altri sono, per l'ipotesi di indipendenza:

$$\mathbb{E}[u_i u_j \mid \mathbf{X}] = \mathbb{E}[u_i \mid \mathbf{x}_i]\mathbb{E}[u_j \mid \mathbf{x}_j] = 0$$

Se si assume omoschedasticità, $\mathbf{D} = \sigma^2 \mathbf{I}_n$.

Poiché $\mathbb{V}[\mathbf{y} \mid \mathbf{X}] = \mathbb{E}[\mathbf{u}\mathbf{u}' \mid \mathbf{X}]$ (cfr. sez. 2.3), la matrice $\mathbf{D}$ è anche la matrice di varianza e covarianza di $\mathbf{y}$.

Se una variabile aleatoria $\mathbf{v}$ è data dal prodotto di un'altra v.a. $\mathbf{z}$ per una matrice $\mathbf{A}$, allora $\mathbb{V}[\mathbf{v}] = \mathbf{A}\mathbb{V}[\mathbf{z}]\mathbf{A}'$. Nel caso dello stimatore $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, si ha:

$$\mathbb{V}[\mathbf{b} \mid \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

È utile notare che:[3]

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \qquad \mathbf{X}'\mathbf{D}\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \sigma_i^2$$

in particolare, $\mathbf{X}'\mathbf{D}\mathbf{X}$ è una versione ponderata di $\mathbf{X}'\mathbf{X}$. Se poi si assume omoschedasticità, $\mathbf{X}'\mathbf{D}\mathbf{X}$ diventa $\mathbf{X}'\mathbf{X}\sigma^2$.

---

[3] Se $\mathbf{X} = \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix}$, allora $\mathbf{X}'\mathbf{X} = \begin{bmatrix} a & c & e \\ b & d & f \end{bmatrix} \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} = \begin{bmatrix} a^2 + c^2 + e^2 & ab + cd + ef \\ ab + cd + ef & b^2 + d^2 + f^2 \end{bmatrix}$, che è la somma di:

$$\mathbf{x}_1\mathbf{x}_1' = \begin{bmatrix} a \\ b \end{bmatrix} \begin{bmatrix} a & b \end{bmatrix} = \begin{bmatrix} a^2 & ab \\ ab & b^2 \end{bmatrix}, \quad \mathbf{x}_2\mathbf{x}_2' = \begin{bmatrix} c \\ d \end{bmatrix} \begin{bmatrix} c & d \end{bmatrix} = \begin{bmatrix} c^2 & cd \\ cd & d^2 \end{bmatrix}, \quad \mathbf{x}_3\mathbf{x}_3' = \begin{bmatrix} e \\ f \end{bmatrix} \begin{bmatrix} e & f \end{bmatrix} = \begin{bmatrix} e^2 & ef \\ ef & f^2 \end{bmatrix}$$

Inoltre,

$$\mathbf{X}'\mathbf{D}\mathbf{X} = \begin{bmatrix} a & c & e \\ b & d & f \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} = \begin{bmatrix} a\sigma_1^2 & c\sigma_2^2 & e\sigma_3^2 \\ b\sigma_1^2 & b\sigma_2^2 & b\sigma_3^2 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix}$$
$$= \begin{bmatrix} a^2\sigma_1^2 + c^2\sigma_2^2 + e^2\sigma_3^2 & ab\sigma_1^2 + cd\sigma_2^2 + ef\sigma_3^2 \\ ab\sigma_1^2 + cd\sigma_2^2 + ef\sigma_3^2 & b^2\sigma_1^2 + d^2\sigma_2^2 + f^2\sigma_3^2 \end{bmatrix}$$

che è la somma di:

$$\mathbf{x}_1\mathbf{x}_1'\sigma_1^2 = \begin{bmatrix} a^2\sigma_1^2 & ab\sigma_1^2 \\ ab\sigma_1^2 & b^2\sigma_1^2 \end{bmatrix}, \quad \mathbf{x}_2\mathbf{x}_2'\sigma_2^2 = \begin{bmatrix} c^2\sigma_2^2 & cd\sigma_2^2 \\ cd\sigma_2^2 & d^2\sigma_2^2 \end{bmatrix}, \quad \mathbf{x}_3\mathbf{x}_3'\sigma_3^2 = \begin{bmatrix} e^2\sigma_3^2 & ef\sigma_3^2 \\ ef\sigma_3^2 & f^2\sigma_3^2 \end{bmatrix}$$

### 2.5.2 Il teorema di Gauss-Markov

**Teorema 2.4** (Gauss-Markov)**.** *In un modello di regressione lineare con ipotesi di omoschedasticità, lo stimatore lineare corretto di minima varianza è lo stimatore OLS*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})'\mathbf{X}'\mathbf{y}$$

*In un modello di regressione lineare con eteroschedasticità, lo stimatore lineare corretto di minima varianza è:*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{y}$$

La prima parte del teorema afferma sì che lo stimatore OLS è efficiente (minima varianza) in caso di omoschedasticità, ma lascia aperta la possibilità che risultino ancora migliori stimatori non lineari oppure distorti.

La seconda parte definisce uno stimatore lineare efficiente per il caso generale, che viene detto *stimatore GLS* (*Generalized Least Squares*); si tratta tuttavia di uno stimatore non direttamente praticabile, in quanto la matrice $\mathbf{D}$ non è nota. Si usa quindi un un approccio detto *FGLS*, *Feasible GLS*, in cui le varianze $\sigma_i^2$ vengono sostituite con loro stime.

### 2.5.3 I residui

A rigore, il vettore dei residui $\mathbf{e}$ non è uno stimatore del vettore degli errori $\mathbf{u}$, ma una sua trasformata:

$$\begin{aligned}\mathbf{e} &= (\mathbf{I}-\mathbf{H})\mathbf{y} = (\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}-\mathbf{H})\mathbf{u} = [\mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})]\boldsymbol{\beta} + (\mathbf{I}-\mathbf{H})\mathbf{u} \\ &= (\mathbf{X}-\mathbf{X})\boldsymbol{\beta} + (\mathbf{I}-\mathbf{H})\mathbf{u} \\ &= (\mathbf{I}-\mathbf{H})\mathbf{u}\end{aligned}$$

dove $\mathbf{H}$ è la matrice di proiezione ortogonale definita nella sezione 2.4.1.

Da ciò segue che, come per l'errore, l'aspettativa condizionata dei residui è zero:

$$\mathbb{E}[\mathbf{e} \mid \mathbf{X}] = \mathbb{E}[(\mathbf{I}-\mathbf{H})\mathbf{u} \mid \mathbf{X}] = (\mathbf{I}-\mathbf{H})\mathbb{E}[\mathbf{u} \mid \mathbf{X}] = 0$$

Quanto alla varianza:

$$\mathbb{V}[\mathbf{e} \mid \mathbf{X}] = \mathbb{V}[(\mathbf{I}-\mathbf{H})\mathbf{u} \mid \mathbf{X}] = (\mathbf{I}-\mathbf{H})\mathbb{V}[\mathbf{u} \mid \mathbf{X}](\mathbf{I}-\mathbf{H}) = (\mathbf{I}-\mathbf{H})\mathbf{D}(\mathbf{I}-\mathbf{H})$$

L'espressione si semplifica nel caso di omoschedasticità; ricordando che la matrice $\mathbf{I}-\mathbf{H}$ è simmetrica e idempotente:

$$\mathbf{D} = \sigma^2\mathbf{I} \quad \Rightarrow \quad \mathbb{V}[\mathbf{e} \mid \mathbf{X}] = (\mathbf{I}-\mathbf{H})\sigma^2$$

In particolare, per la $i$-esima osservazione si ha:

$$\mathbf{D} = \sigma^2\mathbf{I} \quad \Rightarrow \quad \mathbb{V}[e_i \mid \mathbf{X}] = (1-h_{ii})\sigma^2$$

dove $1-h_{ii}$ è l'$i$-esimo elemento della diagonale principale della matrice $\mathbf{I}-\mathbf{H}$. Si vede così che, anche nell'ipotesi che l'errore sia omoschedastico, i residui sono eteroschedastici e non indipendenti: $\mathbb{V}[e_ie_j \mid \mathbf{X}] = (1-h_{ij})\sigma^2$.

Ciò nonostante i residui, come l'errore, sono incorrelati con le variabili esplicative in quanto sono una proiezione di $\mathbf{y}$ su uno spazio ortogonale a quello generato dalle colonne di $\mathbf{X}$; ciò consente di usare i residui per una stima della varianza dell'errore.

### 2.5.4 Stima della varianza dell'errore

La varianza dell'errore, $\sigma^2\mathbb{E}[u^2]$ (sez. 2.3), misura la variabilità di $\mathbf{y}$ non spiegata dalla regressione. Il suo stimatore col metodo dei momenti è:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

In forma matriciale, usando la simmetria e l'idempotenza della matrice $\mathbf{I}-\mathbf{H}$ e le proprietà dell'operatore traccia:

$$\hat{\sigma}^2 = \frac{1}{n}\mathbf{e}'\mathbf{e} = \frac{1}{n}\mathbf{u}'(\mathbf{I}-\mathbf{H})\mathbf{u} = \frac{1}{n}\operatorname{tr}\Big(\mathbf{u}'(\mathbf{I}-\mathbf{H})\mathbf{u}\Big) = \frac{1}{n}\operatorname{tr}\Big((\mathbf{I}-\mathbf{H})\mathbf{u}\mathbf{u}'\Big)$$

da cui:

$$\mathbb{E}[\hat{\sigma}^2 \mid \mathbf{X}] = \frac{1}{n}\operatorname{tr}\Big(\mathbb{E}[(\mathbf{I}-\mathbf{H})\mathbf{u}\mathbf{u}' \mid \mathbf{X}]\Big) = \frac{1}{n}\operatorname{tr}\Big((\mathbf{I}-\mathbf{H})\mathbb{E}[\mathbf{u}\mathbf{u}' \mid \mathbf{X}]\Big) = \frac{1}{n}\operatorname{tr}\Big((\mathbf{I}-\mathbf{H})\mathbf{D}\Big)$$

Le matrici idempotenti hanno traccia uguale al rango (cfr. appendice A). In caso di omoschedasticità, quindi, l'espressione si semplifica:

$$\mathbf{D} = \sigma^2\mathbf{I} \quad \Rightarrow \quad \mathbb{E}[\hat{\sigma}^2 \mid \mathbf{X}] = \frac{1}{n}\operatorname{tr}\Big((\mathbf{I}-\mathbf{H})\sigma^2\Big) = \left(\frac{n-k}{n}\right)\sigma^2$$

e si vede così che lo stimatore è distorto. Si può ottenere uno stimatore corretto dividendo per $n-k$:

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n} e_i^2$$

### 2.5.5 Multicollinearità

Si ha *multicollinearità stretta* quando il rango della matrice $\mathbf{X}'\mathbf{X}$ è minore di $k$; in questo caso, $\mathbf{b}$ non è definito (sez. 2.4.2).

Più frequente il caso della (*quasi*) *multicollinearità*, che si verifica quando la matrice $\mathbf{X}'\mathbf{X}$ è quasi singolare. Si tratta di una definizione vaga (che vuol dire "quasi"?), da cui segue comunque, nella pratica, che i calcoli numerici possono produrre risultati errati, ma, soprattutto, che le stime dei singoli coefficienti diventano imprecise.

Accade infatti che, essendo i regressori tra loro correlati, diventa difficile distinguere i loro effetti sulla variabile risposta, quindi stimare i relativi coefficienti di regressione. Lo *standard error* dei singoli stimatori risulta ampio, conseguentemente ampi i relativi intervalli di confidenza, anche se gli stimatori rimangono corretti.

Si può comunque notare che, come nel caso della varianza campionaria, si possono ottenere risultati migliori aumentanto la dimensione del campione.

## 2.6 Necessità di un approccio asintotico

L'approccio dei campioni finiti risulta poco utile nell'analisi econometrica, in quanto succede raramente di poter estrarre più campioni da una stessa popolazione (la popolazione, infatti, cambia nel tempo).

Negli studi sperimentali il ricercatore ha il pieno controllo dell'esperimento: sceglie alcuni fattori di cui vuole indagare l'effetto su alcune unità; a tale scopo sceglie diversi trattamenti, corrispondenti a diversi livelli di quei fattori (le variabili esplicative), e li somministra alle unità sperimentali in modo casuale; osserva quindi i valori di una variabile risposta per verificare se essi possono essere intesi come effetti delle variabili esplicative, oppure se la variabilità osservata nella risposta è imputabile solo a fattori accidentali. La somministrazione dei trattamenti è a tal punto sotto il controllo del ricercatore, che le variabili esplicative vengono spesso intese come variabili deterministiche, non aleatorie.

Tutto ciò in econometria è impossibile. Si possono solo osservare i valori di alcune variabili assunte come esplicative (valori osservati, non scelti dal ricercatore), senza alcuna garanzia di aver considerato tutte le variabili che potrebbero avere effetto sulla variabile risposta. Non è possibile, inoltre, ripetere lo studio a piacimento; ad esempio, per studiare l'effetto delle spese promozionali sulle vendite non si può provare prima con un ammontare, poi con un altro, poi con un altro ancora; per studiare l'effetto del livello di istruzione sul salario non si possono far studiare fino a livelli diversi gruppi di ragazzi scelti a caso e poi, dopo qualche anno, rilevare i loro salari. Ne segue che anche le variabili esplicative vanno intese come variabili aleatorie e che non ha molto senso la ricerca di stimatori corretti; si preferisce quindi effettuare ricerche su grandi campioni contando su proprietà quali la consistenza e la normalità asintotica degli stimatori.

Risultano ancora meno utili ipotesi di distribuzione normale, in quanto i fenomeni economici sono tipicamente non-normali. Si può notare, al riguardo, che negli studi sperimentali l'errore viene tradizionalmente inteso come errore sperimentale, come effetto di una variabilità del tutto accidentale presente sia nel fenomeno studiato che nelle misurazioni effettuate; in tale contesto è ragionevole assumere sia che l'errore $\boldsymbol{\varepsilon}$ presenti una distribuzione normale, sia che non risulti correlato con le variabili esplicative

In econometria, invece, l'errore viene indicato preferibilmente con $u$, per *unobserved*, in quanto contiene anche variabili che possono avere effetto sulla variabile risposta ma non sono state osservate; può trattarsi di variabili per le quali non sono disponibili dati attendibili, o anche di variabili non direttamente misurabili (ad esempio, l'abilità individuale come fattore del livello del salario).

Risulta necessario, pertanto, assumere inizialmente un *modello della popolazione* che appaia ragionevolmente completo dal punto di vista della teoria economica. Nel caso si tratti di un modello lineare nei parametri, l'approccio più semplice consiste nell'applicare la regressione lineare assumendo che le variabili non osservate non siano correlate con quelle osservate; rimane così possibile mantenere la definizione di errore come differenza tra $y$ e la sua aspettativa condizionata, quindi le proprietà:

$$\mathbb{E}[u \mid \mathbf{x}] = 0 \qquad \mathbb{E}[u] = 0 \qquad \mathbb{E}[\mathbf{f}(\mathbf{x})u] = \mathbf{0} \qquad \mathbb{E}[\mathbf{x}u] = \mathbf{0}$$

L'assunzione di tali proprietà dell'errore viene detta *ipotesi di esogeneità* e ad essa è dedicato il prossimo capitolo.

# Capitolo 3

# L'ipotesi di esogeneità

In economia una variabile viene detta endogena se è determinata nell'ambito di un modello, ad esempio se è variabile dipendente in equazioni in cui compaiono altre variabili, dette esogene, i cui valori sono assunti come dati. In econometria, invece, una variabile esplicativa viene detta *endogena* se è correlata con la variabile non osservabile $u$, *esogena* in caso contrario.

Nella regressione lineare con ipotesi di esogeneità si muove da un modello della popolazione del tipo $y = \mathbf{x}'\boldsymbol{\beta} + u$ assunto come "vero" e si assume, inoltre, l'*ipoesi di esogeneità* $\mathbb{E}[u \mid \mathbf{x}] = 0]$. Come si è visto (sez. 2.2), da ciò seguono $\mathbb{E}[u] = 0$, $\mathbb{E}[\mathbf{x}u] = \mathbf{0}$ e $\mathrm{Cov}(\mathbf{x}, u) = \mathbf{0}$.

Obiettivo dell'analisi è la stima degli *effetti parziali* delle variabili esplicative sull'aspettativa condizionata di $y$:

$$\frac{\partial}{\partial x_j}\mathbb{E}[y \mid \mathbf{x}] = \frac{\partial}{\partial x_j}(\beta_1 + \beta_2 x_2 \cdots + \beta_k x_k) = \beta_j$$

In questo capitolo si illustra l'importanza dell'ipotesi di esogeneità e si mostra come, grazie ad essa, sia possibile ottenere stimatori consistenti e asintoticamente normali degli effetti parziali. Si discutono poi alcuni test di ipotesi e si conclude mostrando i rimedi più semplici alle frequenti situazioni di endogeneità.[1]

## 3.1 L'importanza dell'ipotesi

Per apprezzare l'importanza dell'ipotesi di esogeneità, si può ipotizzare che il modello "vero" (assunto come tale) della popolazione sia:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u \qquad \mathbb{E}[u \mid x_2, x_3] = 0$$

Va notato che il modello non afferma che $y$ dipende solo da $x_2, x_3$, ma piuttosto che, anche se $u$ contiene altre variabili che hanno effetto su $y$, queste non sono correlate alle due considerate.

Il modello consente di definire gli effetti parziali delle variabili $x_2, x_3$ sull'aspettativa condizionata di $y$; ad esempio:

$$\frac{\partial}{\partial x_2}\mathbb{E}[y \mid x_2, x_3] = \frac{\partial}{\partial x_2}(\beta_1 + \beta_2 x_2 + \beta_3 x_3) = \beta_2$$

[1] Questo capitolo e il successivo seguono liberamente la traccia di Wooldridge (2002, capp. 4-5), con elementi tratti da Hansen (2010, capp. 5-6).

Se però si usasse il modello

$$y = \beta_1 + \beta_2 x_2 + v \qquad v = \beta_3 x_3 + u$$

e se $x_2$ e $x_3$ fossero correlate, si avrebbe in realtà, per qualche $c$,

$$\mathbb{E}[y \mid x_2] = \beta_1 + \beta_2 x_2 + \beta_3 (c x_2) \qquad \text{quindi} \qquad \frac{\partial}{\partial x_2} \mathbb{E}[y \mid x_2] = \beta_2 + \beta_3 c$$

dove $\beta_2$ sarebbe l'effetto diretto, $\beta_3 c$ quello indiretto, di $x_2$; $\beta_2$ non potrebbe quindi essere considerato l'effetto parziale di $x_2$. In altri termini, non sarebbe possibile concludere: se $x_2$ aumenta di una unità, allora $\mathbb{E}[y \mid x_2]$ aumenta di $\beta_2$.

Da altro punto di vista, non si potrebbe più definire l'errore $v$ come differenza tra $y$ e la sua aspettativa condizionata, infatti:

$$y - \mathbb{E}[y \mid x_2] = \beta_1 + \beta_2 x_2 + v - \beta_1 - \beta_2 x_2 - \beta_3 (c x_2) = v - \beta_3 (c x_2) \neq v$$

Ne seguirebbe:

$$\mathbb{E}[v] = \mathbb{E}[\beta_3 x_3 + u] = \mathbb{E}[\beta_3 c x_2 + u] = \beta_3 c \mathbb{E}[x_2] \neq 0$$

In pratica, si cerca di costruire modelli in cui compaiano, oltre alle variabili esplicative di cui interessa studiare l'effetto parziale, anche altre variabili esplicative *di controllo*, il cui scopo è fare in modo che il termine $u$ possa sì contenere variabili non osservate, ma solo variabili non correlate con quelle di interesse.

L'ipotesi di esogeneità risulta particolarmente importante anche perché solo se risulta assumibile si può pervenire a stime consistenti degli effetti parziali, come si vedrà nella sezione successiva.

**Osservazione 3.1.** Si è visto nel capitolo 2 che il vettore dei residui è ortogonale al sottospazio generato dalle colonne della matrice $\mathbf{X}$ e che si ha:

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{u}$$

In grandi campioni la matrice $\mathbf{H}$ tende a diventare poco rilevante, in quanto le sue somme di riga e di colonna sono sempre 1 e la somma degli elementi della diagonale principale è sempre pari a $k$. Per $n \to \infty$, quindi, $\mathbf{e} \xrightarrow{p} \mathbf{u}$. Tuttavia questo avviene *sempre* e, pertanto, non consente di verificare l'ipotesi di esogeneità; né c'è altro modo. Vi sono comunque situazioni in cui l'ipotesi appare manifestamente infondata:

a) *variabili omesse*: il modello non comprende tutte le variabili di controllo perché non si disponde dei dati necessari; si può ovviare usando *variabili proxy* (sez. 3.4);

b) *errore di misura*: alcune variabili possono essere rilevate sono in modo imperfetto (ad esempio, perché il loro valore dipende dall'accuratezza e dall'attendibilità delle unità di rilevazione; il problema è discusso nella sez. 3.5);

c) *simultaneità*: una o più variabili esplicativi sono in parte funzioni della variabile risposta (ad esempio, se $y$ è il numero di omicidi in una città e $x_j$ è l'organico delle forze di polizia, $x_j$ è determinata in parte da $y$).

## 3.2 La stima dei parametri

Il modello della popolazione viene espresso più sinteticamente nella forma:

$$y = \mathbf{x}'\boldsymbol{\beta} + u$$

in cui $\mathbf{x}$ indica il vettore $k \times 1$ delle variabili esplicative.

Assumendo di estrarre un campione dalla popolazione, si avranno $n$ osservazioni del tipo:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

in cui $\mathbf{x}_i$ indica il vettore colonna della $i$-esima riga della matrice $\mathbf{X}$, contenente tante righe quante sono le osservazioni e tante colonne quante sono le variabili esplicative.

L'analisi di regressione lineare si basa su un modello di popolazione che soddisfi i seguenti assunti.

**Assunti 3.2.**

1) *Indipendenza e identica distribuzione*: possibilità di estrarre campioni casuali contenenti le variabili iid $y_i, \mathbf{x}_i$, $i = 1, \dots, n$.
2) *Linearità*: $y = \mathbf{x}'\boldsymbol{\beta} + u$.
3) *Esogeneità*: $\mathbb{E}[u \mid \mathbf{x}] = 0$.
4) *Momenti quarti finiti per* $\mathbf{x}$ *e* $u$.
5) *Rango pieno* (invertibilità) della matrice $\mathbb{E}[\mathbf{x}\mathbf{x}']$.

Premoltiplicando il modello della popolazione per $\mathbf{x}$ e calcolando i valori attesi:

$$\mathbb{E}[\mathbf{x}y] = \mathbb{E}[\mathbf{x}\mathbf{x}'\boldsymbol{\beta} + \mathbf{x}u] = \boldsymbol{\beta}\mathbb{E}[\mathbf{x}\mathbf{x}'] + \mathbb{E}[\mathbf{x}u] = \boldsymbol{\beta}\mathbb{E}[\mathbf{x}\mathbf{x}']$$

in quanto $u$ e $\mathbf{x}$ sono incorrelate (per l'ipotesi di esogeneità). Si ottiene così:

$$\boldsymbol{\beta} = \mathbb{E}[\mathbf{x}\mathbf{x}']^{-1}\mathbb{E}[\mathbf{x}y]$$

Per stimare $\mathbb{E}[\mathbf{x}\mathbf{x}']$ e $\mathbb{E}[\mathbf{x}y]$ si può ricorrere al metodo dei momenti, sostituendoli con le rispettive medie campionarie:

$$\begin{aligned}\mathbf{b} &= \left(n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\mathbf{x}_i y_i\right) = n\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} n^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i y_i\right)\\ &= \left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i y_i\right)\end{aligned}$$

Oppure, in forma matriciale,[2]

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

dove $\mathbf{X}$ è la matrice con righe $\mathbf{x}_i'$, $i = 1, \dots, n$, e $\mathbf{y}$ è il vettore colonna $[\ y_1\ \dots\ y_n\ ]'$. Poiché $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, si può anche scrivere:

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\\ &= \boldsymbol{\beta} + \left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i u_i\right)\end{aligned}$$

[2] Cfr. cap. 2, nota 3 a pag. 18.

Lo stimatore $\mathbf{b}$ non è altro che lo *stimatore OLS*. Tuttavia, mentre nel caso di campioni di ampiezza finita risulta uno stimatore corretto, nell'approccio asintotico rilevano consistenza e normalità asintotica.

### 3.2.1 Consistenza

**Teorema 3.3.** *Se valgono gli assunti 3.2, lo stimatore OLS di $\boldsymbol{\beta}$:*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \boldsymbol{\beta} + \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{x}_i u_i\right)$$

*è consistente:*

$$\mathbf{b} \xrightarrow{p} \boldsymbol{\beta}$$

*Dimostrazione.* L'espressione di $\mathbf{b}$ in termini di medie campionarie rende evidente che $\mathbf{b}$ dipende anche da $n$; si può quindi considerare la successione

$$\mathbf{b}_n = \boldsymbol{\beta} + \left(\frac{\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'}{n}\right)^{-1}\left(\frac{n^{-1}\sum_{i=1}^{n} \mathbf{x}_i u_i}{n}\right)$$

$\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'$ è in termine $n$-esimo di una successione di variabili aleatorie assunte iid, il cui valore atteso è $\mathbb{E}[\mathbf{x}\mathbf{x}']$, assunto finito; quindi, per la legge dei grandi numeri:

$$\frac{\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'}{n} \xrightarrow{p} \mathbb{E}[\mathbf{x}\mathbf{x}']$$

Analogamente, e per l'ipotesi di esogeneità:

$$\frac{\sum_{i=1}^{n} \mathbf{x}_i u_i}{n} \xrightarrow{p} \mathbb{E}[\mathbf{x}u] = \mathbf{0}$$

Si assume inoltre il rango pieno di $\mathbb{E}[\mathbf{x}\mathbf{x}']$, quindi l'esistenza di $\mathbb{E}[\mathbf{x}\mathbf{x}']^{-1}$. Per il lemma di Slutsky (v. appendice C), essendo l'inversa una funzione continua,

$$\frac{\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'}{n} \xrightarrow{p} \mathbb{E}[\mathbf{x}\mathbf{x}'] \quad \Rightarrow \quad \left(\frac{\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'}{n}\right)^{-1} \xrightarrow{p} \mathbb{E}[\mathbf{x}\mathbf{x}']^{-1} < \infty$$

e si ha:

$$\mathbf{b}_n \xrightarrow{p} \boldsymbol{\beta} + \mathbb{E}[\mathbf{x}\mathbf{x}']^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}$$ □

### 3.2.2 Normalità asintotica

**Teorema 3.4.** *Se valgono gli assunti 3.2, lo stimatore OLS di $\boldsymbol{\beta}$:*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \boldsymbol{\beta} + \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{x}_i u_i\right)$$

*è asintoticamente normale:*

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$$

*Dimostrazione.* La successione $\mathbf{b}_n$ può essere riscritta come segue, portando a sinistra $\boldsymbol{\beta}$ e moltiplicando entrambi i membri per $\sqrt{n}$:

$$\sqrt{n}(\mathbf{b}_n - \boldsymbol{\beta}) = \left( \frac{\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'}{n} \right)^{-1} \left( \frac{\sum_{i=1}^n \mathbf{x}_i u_i}{\sqrt{n}} \right)$$

Si è appena visto che $\left( \frac{\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'}{n} \right)^{-1} \xrightarrow{p} \mathbb{E}[\mathbf{x}\mathbf{x}']^{-1} < \infty$.

Quanto a $\frac{\sum_{i=1}^n \mathbf{x}_i u_i}{\sqrt{n}}$, dall'ipotesi di esogeneità segue che $\mathbb{E}[\mathbf{x}u] = 0$. Inoltre:

$$\mathbb{V}[\mathbf{x}u] = \mathbb{E}[u^2 \mathbf{x}\mathbf{x}'] - \mathbb{E}[\mathbf{x}]\mathbb{E}[u] = \mathbb{E}[u^2 \mathbf{x}\mathbf{x}']$$

Per la disuguaglianza di Cauchy-Schwarz e per l'assunto dei momenti quarti finiti:

$$\mathbb{E}\left[ \left| u^2 \mathbf{x}\mathbf{x}' \right| \right] \leq \sqrt{\mu_4(u)\mu_4(\mathbf{x})} < \infty$$

Si può quindi applicare il teorema del limite centrale alla successione $\sum_{i=1}^n \mathbf{x}_i u_i$:

$$\frac{\sum_{i=1}^n \mathbf{x}_i u_i}{\sqrt{n}} \xrightarrow{d} N(\mathbf{0}, \mathbb{V}[\mathbf{x}u])$$

$\sqrt{n}(\mathbf{b}_n - \boldsymbol{\beta})$ risulta così una trasformazione lineare (una moltiplicazione per una quantità che tende a $\mathbb{E}[\mathbf{x}\mathbf{x}']^{-1}$, che è una matrice simmetrica) di una successione di v.a. asintoticamente normali e si ha:

$$\sqrt{n}(\mathbf{b}_n - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbb{E}[\mathbf{x}\mathbf{x}']^{-1}\mathbb{E}[u^2 \mathbf{x}\mathbf{x}']\mathbb{E}[\mathbf{x}\mathbf{x}']^{-1}) \qquad \square$$

Ponendo

$$\mathbf{A} = \mathbb{E}[\mathbf{x}\mathbf{x}'] \qquad \mathbf{B} = \mathbb{V}[\mathbf{x}u] = \mathbb{E}[u^2 \mathbf{x}\mathbf{x}']$$

il teorema consente di dire che, per grandi campioni, $\mathbf{b}$ si distribuisce approssimativamente come una normale con media $\boldsymbol{\beta}$ e varianza $\frac{\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}}{n}$:[3]

$$\mathbf{b} \overset{a}{\sim} N\left( \boldsymbol{\beta}, \frac{\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}}{n} \right)$$

Rimane comunque da stimare la varianza asintotica di $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta})$, quindi quella approssimata di $\mathbf{b}$.

[3] Il simbolo $\overset{a}{\sim}$ può stare per "asintoticamente distribuito come" (così Wooldridge 2002, p.38), oppure per "distribuito approssimativamente come". Nel primo significato, il simbolo è equivalente all'altro $\xrightarrow{d}$; inoltre, se $n \to \infty$ non si può dividere impunemente per $n$. Si usa quindi qui il secondo significato, intendendo che vale per $n$ grande, ma comunque finito.

### 3.2.3 Stima della varianza

Se si volesse assumere omoschedasticità, la varianza dell'errore sarebbe costante e non dipenderebbe da $\mathbf{x}$, né da $\mathbf{xx}'$. Si avrebbe quindi:

$$\mathbb{E}[u^2\mathbf{xx}'] = \sigma^2\mathbb{E}[\mathbf{xx}'] \qquad \sigma^2 = \mathbb{E}[u^2]$$

La varianza asintotica di $\sqrt{n}(\mathbf{b}-\boldsymbol{\beta})$ diventerebbe:

$$\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} = \mathbb{E}[\mathbf{xx}']^{-1}\sigma^2\mathbb{E}[\mathbf{xx}']\mathbb{E}[\mathbf{xx}']^{-1} = \sigma^2\mathbb{E}[\mathbf{xx}']^{-1}$$

Indicando con il simbolo $\mathbf{V}_{\mathbf{b}}^{o}$ la varianza approssimata di $\mathbf{b}$ in omoschedasticità,

$$\mathbf{V}_{\mathbf{b}}^{o} = \frac{\sigma^2\mathbb{E}[\mathbf{xx}']^{-1}}{n}$$

Si è già usato, per $\mathbb{E}[\mathbf{xx}']$, lo stimatore $(\mathbf{X}'\mathbf{X})/n$. Quanto a $\sigma^2$, si potrebbe usare come stimatore consistente la varianza campionaria dei residui, $\hat{\sigma}^2 = \dfrac{\mathbf{ee}'}{n}$.[4] Infatti:[5]

**Teorema 3.5.** *Se valgono gli assunti 3.2, la varianza campionaria dei residui:*

$$\hat{\sigma}^2 = \frac{\mathbf{ee}'}{n}$$

*è uno stimatore consistente di* $\sigma^2 = \mathbb{E}[u^2]$.

*Dimostrazione.* Muovendo da:

$$\begin{aligned} u_i &= y_i - \mathbf{x}_i'\boldsymbol{\beta} \\ e_i &= y_i - \mathbf{x}_i'\mathbf{b} = u_i + \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\mathbf{b} = u_i - \mathbf{x}_i'(\mathbf{b}-\boldsymbol{\beta}) \\ e_i^2 &= u_i^2 - 2u_i\mathbf{x}_i'(\mathbf{b}-\boldsymbol{\beta}) + (\mathbf{b}-\boldsymbol{\beta})'\mathbf{xx}'(\mathbf{b}-\boldsymbol{\beta}) \end{aligned}$$

si perviene a:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} e_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} u_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'u_i\right)(\mathbf{b}-\boldsymbol{\beta}) + (\mathbf{b}-\boldsymbol{\beta})'\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{b}-\boldsymbol{\beta}) \\ &\xrightarrow{p} \sigma^2 \end{aligned}$$

ricordando che per gli assunti e per la legge dei grandi numeri:

$$\frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'}{n} \xrightarrow{p} \mathbb{E}[\mathbf{xx}'] < \infty \qquad \frac{\sum_{i=1}^{n}\mathbf{x}_i u_i}{n} \xrightarrow{p} \mathbb{E}[\mathbf{x}u] = \mathbf{0} \qquad \mathbf{b} \xrightarrow{p} \boldsymbol{\beta} \qquad \square$$

[4] Si potrebbe usare anche una varianza campionaria corretta, dividendo per $n-1$ o $n-k$, in quanto la consistenza per $n \to \infty$ non ne risentirebbe.

[5] Le dimostrazioni dei teoremi 3.5 e 3.6 sono tratte da Hansen (2010, pp. 73-74, 76-77).

La varianza approssimata di $\mathbf{b}$ verrebbe così stimata da:

$$\hat{\mathbf{V}}_{\mathbf{b}}^{o} = \frac{\hat{\sigma}^2}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

Nel caso più generale (eteroschedasticità), occorre una diversa stima di $\mathbf{B} = \mathbb{E}[u^2\mathbf{x}\mathbf{x}']$. Il metodo dei momenti suggerisce lo stimatore $\frac{\sum_{i=1}^n u_i^2\mathbf{x}_i\mathbf{x}_i'}{n}$; dal momento che gli $u_i$ non sono osservabili, possono essere sostituiti dai residui $e_i$ e si dimostra che si ottiene così uno stimatore consistente $\hat{\mathbf{B}} = \frac{\sum_{i=1}^n e_i^2\mathbf{x}_i\mathbf{x}_i'}{n}$.

**Teorema 3.6.** *Se valgono gli assunti 3.2, lo stimatore:*

$$\hat{\mathbf{B}} = \frac{\sum_{i=1}^n e_i^2\mathbf{x}_i\mathbf{x}_i'}{n}$$

*è uno stimatore consistente di* $\mathbf{B} = \mathbb{E}[u^2\mathbf{x}\mathbf{x}']$.

*Dimostrazione.* Si può esprimere $\hat{\mathbf{B}}$ come segue (cfr. dimostrazione del teorema 3.5):

$$\begin{aligned}\hat{\mathbf{B}} &= \frac{1}{n}\sum_{i=1}^n e_i^2\mathbf{x}_i\mathbf{x}_i' \\ &= \frac{1}{2}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'u_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'(\mathbf{b}-\boldsymbol{\beta})'\mathbf{x}_i u_i\right) + \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\left((\mathbf{b}-\boldsymbol{\beta})'\mathbf{x}_i\right)^2\end{aligned}$$

Considerando separatamente i tre addendi, il generico elemento $hl$-esimo della matrice $\mathbf{x}_i\mathbf{x}_i'u_i^2$ è $x_{ih}x_{il}u_i^2$. Per la disuguaglianza di Cauchy-Schwarz e per l'assunto dei momenti quarti finiti:

$$\begin{aligned}\mathbb{E}\left[\left|x_{ih}x_{il}u_i^2\right|\right] &\leq \mathbb{E}[x_{ih}^2x_{il}^2]^{1/2}\mathbb{E}[u_i^4]^{1/2} \\ &\leq \mathbb{E}[x_{ih}^4]^{1/4}\mathbb{E}[x_{il}^4]^{1/4}\mathbb{E}[u_i^4]^{1/2} < \infty\end{aligned}$$

Quindi, per la legge dei grandi numeri:

$$\frac{1}{2}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'u_i^2 \xrightarrow{p} \mathbb{E}[u^2\mathbf{x}\mathbf{x}'] = \mathbf{B}$$

Applicando la disuguaglianza triangolare alla norma del secondo addendo, la disuguaglianza di Schwarz, poi l'uguaglianza $\|\mathbf{v}\mathbf{v}'\| = \|\mathbf{v}\|^2$, infine ancora la disuguaglianza di Schwarz:

$$\begin{aligned}\left\|\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'(\mathbf{b}-\boldsymbol{\beta})'\mathbf{x}_i u_i\right\| &\leq \frac{2}{n}\sum_{i=1}^n \left\|\mathbf{x}_i\mathbf{x}_i'(\mathbf{b}-\boldsymbol{\beta})'\mathbf{x}_i u_i\right\| \\ &\leq \frac{2}{n}\sum_{i=1}^n \left\|\mathbf{x}_i\mathbf{x}_i'\right\| \left|(\mathbf{b}-\boldsymbol{\beta})'\mathbf{x}_i\right| |u_i| \\ &\leq \left(\frac{2}{n}\sum_{i=1}^n \|\mathbf{x}_i\|^3 |u_i|\right)\|\mathbf{b}-\boldsymbol{\beta}\|\end{aligned}$$

Per la disuguaglianza di Hölder (di cui quella di Cauchy-Schwarz è caso particolare) e per l'assunto dei momenti quarti finiti:

$$\mathbb{E}\left[\|\mathbf{x}_i\|^3 |u_i|\right] \leq \mathbb{E}\left[\|\mathbf{x}_i\|^4\right]^{3/4} \mathbb{E}\left[u_i^4\right]^{1/4} < \infty$$

Per la legge dei grandi numeri:

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^3 |u_i| \xrightarrow{p} \mathbb{E}\left[\|\mathbf{x}_i\|^3 |u_i|\right] < \infty$$

Poiché $\mathbf{b} - \boldsymbol{\beta} \xrightarrow{p} \mathbf{0}$, il secondo addendo converge in probabilità a zero. Analogamente:

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' \left((\mathbf{b}-\boldsymbol{\beta})'\mathbf{x}_i\right)^2\right\| &\leq \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\mathbf{x}_i'\| \left((\mathbf{b}-\boldsymbol{\beta})'\mathbf{x}_i\right)^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^4 \|\mathbf{b}-\boldsymbol{\beta}\|\end{aligned}$$

quindi anche il terzo addendo converge in probabilità a zero. □

Dal teorema segue che, indicando con $\mathbf{V_b}$ la varianza approssimata di $\mathbf{b}$ nel caso generale (eteroschedasticità), una sua stima consistente è:

$$\hat{\mathbf{V}}_\mathbf{b} = \frac{\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}}{n} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} e_i^2 \mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}$$

Si tratta di una matrice detta HCCME, per *Heteroskedasticity-Consistent Covariance Matrix Estimator*, introdotta da H. White nel 1980. Le radici quadrate degli elementi della diagonale principale vengono detti *errori standard di White*.

Sono state proposte diverse varianti della matrice, considerando che quasi certamente la somma dei quadrati dei residui è minore di quella dei quadrati degli errori non osservati; in altri termini, poiché lo stimatore $\mathbf{b}$ minimizza $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \mathbf{x}_i'\mathbf{b})^2$, tale somma è quasi certamente minore di $\sum_{i=1}^n u_i^2 = \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$ (Cottrell e Lucchetti 2010, pp. 104-105). Le varianti principali sono (cfr. Zeileis 2004):

- HC o HC0: la matrice originale di White;
- HC1: la matrice di White moltiplicata per $(n-k)/n$, una correzione per i gradi di libertà;
- HC2: $e_i^2/(1-h_{ii})$ invece di $e_i^2$ (cfr. sez. 2.5.3);
- HC3: $e_i^2(1-h_{ii})^2$ invece di $e_i^2$;
- HC4: $e_i^2(1-h_{ii})^{\delta_i}$ invece di $e_i^2$, dove $\delta_i = \min\{4, h_{ii}/\bar{h}\} = \min\{4, nh_{ii}/\sum h_{ii}\}$.

Le modifiche della matrice di White, peraltro, sono state proposte per migliorare le stime nei casi di campioni di ampiezza finita (Cribari-Neto 2004; in particolare HC4 è costruita in modo da contenere l'effetto di *outlier*) non conducono a risultati apprezzabilmente diversi con grandi campioni (Wooldridge 2002, p. 57).[6]

[6] Nel caso di HC1 ciò appare evidente. Per le altre varianti basta considerare che la somma degli $h_{ii}$ è uguale al numero $k$ dei parametri e la loro media è quindi uguale a $n/k$ (Kutner *et al.* 2005, pp. 398-399).

**Osservazione 3.7.** Nella sez. 2.6 si rilevava che, mentre negli studi sperimentali le variabili esplicative sono spesso deterministiche, in econometria sono aleatorie. Può essere utile tornare al semplice scenario dell'omoschedasticità per esplicitare una conseguenza della diversità degli approcci. Si assume comunque che $\mathrm{Cov}(\mathbf{x}, u) = 0$, quindi $\mathbb{V}[y] = \mathbb{V}[\mathbf{x}'\boldsymbol{\beta}] + \mathbb{V}[u]$. Se però le variabili esplicative non hanno variabilità, allora $\mathbb{V}[\mathbf{x}'\boldsymbol{\beta}] = 0$ e si ha:

$$\text{variabili esplicative deterministiche} \quad \Rightarrow \quad \mathbb{V}[y] = \mathbb{V}[u] = \sigma^2$$

ovvero $y$ sarebbe completamente determinata da $\mathbf{x}'\boldsymbol{\beta}$ se non fosse per una componente puramente accidentale. Quando invece le variabili esplicative sono anch'esse aleatorie, la variabilità di $y$ comprende anche quella delle esplicative:

$$\text{variabili esplicative aleatorie} \quad \Rightarrow \quad \mathbb{V}[y] = \mathbb{V}[\mathbf{x}'\boldsymbol{\beta}] + \mathbb{V}[u]$$

Ne segue, tra l'altro, che la variabilità di $y$ dipende anche dalla scelta delle esplicative. In generale, infatti, non esiste alcuna garanzia che si considerino tutte le variabili di controllo, rispetto alle quali sono possibili scelte diverse e può succedere, inoltre, che alcune variabili prima non osservabili lo diventino; la variabilità della variabile risposta può quindi cambiare da modello a modello.

## 3.3 Test di ipotesi e intervalli di confidenza

La stima della varianza approssimata di $\mathbf{b}$ consente di effettuare test di ipotesi e di calcolare intervalli di confidenza.

### 3.3.1 Test *z*

Si è visto che, nel modello lineare normale con ipotesi di omoschedasticità (sez. 2.4.4), si usano test $t$ in quanto la varianza $\sigma^2$ dell'errore non è nota e viene sostituita con una varianza campionaria corretta dei residui. In un approccio asintotico le differenze rispetto ad un test $z$ diventano trascurabili, mentre appare più rivelante l'abbandono dell'ipotesi di omoschedasticità.

La libreria `lmtest` di R contiene, tra altre, una funzione `coeftest()` che esegue test analoghi a quelli calcolati da `summary()` sul risultato di `lm()` (cfr. fig. 2.2), con le seguenti differenze (cfr. Zeileis 2004):

a) il parametro `df` (*degrees of freedom*) ha $n - k$ come valore di default e, se lo si accetta o si assegna un numero finito e positivo, viene calcolato un test $t$; con `df=Inf` si usa invece un'approssimazione normale;

b) il parametro `vcov.` (NB: con un punto finale, per distinguerlo dalla funzione `vcov()`) consente di passare una matrice di varianza e covarianza diversa da quella calcolata dalla funzione `lm()`.

La libreria `sandwich` consente di calcolare la matrice di White (anche le sue varianti) con una funzione `vcovHC()`.[7] La funzione usa per default la matrice HC3, ma si può usare quella di White assegnando `"HC0"` al parametro `type`.

[7] La libreria si chiama `sandwich` perché prodotti come quelli che compaiono nella matrice di White, $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, vengono detti "a forma di sandwich".

**Esempio 3.8.** Si vuole determinare se il salario delle donne è influenzato dalla condizione familiare, in particolare dall'età e dal numero dei figli. Le relative variabili esplicative sono:

- `age`: l'età anagrafica in anni;
- `kidslt6`: il numero dei figli di età minore di 6 anni;
- `kidsge6`: il numero dei figli di età compresa tra 6 e 18 anni.

Si prendono in considerazione anche altre variabili, che appaiono correlate almeno all'età anagrafica (variabili di controllo):

- `exper`: l'anzianità di lavoro;
- `expersq`: il quadrato dell'anzianità di lavoro (si ipotizza che intervengano negli anni avanzamenti di qualifica, quindi che l'effetto dell'anzianità sul salario non sia lineare);
- `educ`: il livello di istruzione, misurato con gli anni di frequentazione delle scuole.

Si sceglie il seguente modello per la popolazione:

$$\log(\texttt{wage}) = \beta_1 + \beta_2\texttt{exper} + \beta_3\texttt{expersq} + \beta_4\texttt{educ} + \beta_5\texttt{age} + \beta_6\texttt{kidslt6} + \beta_7\texttt{kidsge6} + u$$

si sceglie cioè di usare come variabile risposta il logaritmo del salario, `lwage`. Si carica il file `mroz.csv`[8] e si esegue la regressione lineare:

```
> mroz <- read.csv("mroz.csv")
> # seleziona le righe con inlf=1 (inlf: in labor force)
> mroz <- mroz[mroz$inlf==1,]
> reg <- lm(lwage ~ exper+expersq+educ+age+kidslt6+kidsge6, data=mroz)
```

I test sui singoli coefficienti sono riprodotti nella figura 3.1.

**Osservazione 3.9.** In questo e in molti degli esempi che seguono non si mostrano né si commentano i valori di $R^2$ e di $\overline{R}^2$, sia per non appesantire l'esposizione riportando l'output di `summary()`, sia per quanto sopra detto a pag. 14: si ottengono spesso valori relativamente piccoli – intorno a 0.15 in questo caso – e risultano più interessanti risultati relativi ai singoli coefficienti. In alcuni casi, peraltro, si useranno $R^2$ e $\overline{R}^2$ per valutare la misura in cui, aggiungendo o eliminando variabili esplicative, aumenta o diminuisce la quota spiegata della variabilità di $y$.

**Osservazione 3.10.** Il test $t$ o $z$ sui singoli coeffcienti sono utili, ma spesso abusati. Si deve ricordare che i test sottopongono a verifica un'ipotesi nulla del tipo $\beta_j = 0$ e che rifiutare l'ipotesi nulla vuol dire accettare che il valore "vero" di $\beta_j$ potrebbe diverso da zero, quindi anche... 0.001 (o meno; cfr. esempio 3.11). Occorre cautela soprattutto nell'approccio asintotico, in quanto in campioni di grandi dimensioni l'area di accettazione dell'ipotesi nulla è tanto più ristretta quanto maggiore è $n$ (cfr. Wonnacott e Wonnacott 1982, p. 219n; McCloskey e Ziliak 1996; Ziliak e McCloskey 2004). Risulta quindi più corretto valutare le stime puntuali in quanto *best guess*, gli *standard error* in quanto misure della precisione delle stime, e soprattutto gli intervalli di confidenza (Hansen 2010, p. 90).

[8] Scaricabile da http://web.mclink.it/MC1166/Econometria/mroz.csv. Si tratta di un adattamento del file `mroz.raw` proposto da Wooldridge (2002) e scaricabile dal sito del libro: si sono assegnati i nomi di colonna e si sono usati zeri per i dati mancanti (salario non rilevato per donne che non lavorano; vi erano infatti punti che R legge come non numerici, interpretando così le colonne `wage` e `lwage` come relative a dati qualitativi).

```
> library(sandwich)
> library(lmtest)
> coeftest(reg, df=Inf, vcov.=vcovHC(reg, type="HC0"))

z test of coefficients:

              Estimate  Std. Error z value  Pr(>|z|)
(Intercept) -0.42090796  0.31572069 -1.3332  0.182477
exper        0.03981902  0.01513251  2.6314  0.008504 **
expersq     -0.00078123  0.00040632 -1.9227  0.054519 .
educ         0.10783196  0.01351167  7.9807 1.456e-15 ***
age         -0.00146526  0.00588632 -0.2489  0.803418
kidslt6     -0.06071057  0.10522938 -0.5769  0.563983
kidsge6     -0.01459101  0.02910954 -0.5012  0.616199
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Figura 3.1.* Test sui singoli coefficienti con approssimazione normale e matrice di White.

### 3.3.2 Intervalli di confidenza

Date le stime $\hat{\theta}$ di un parametro incognito $\theta$ e della radice quadrata della sua varianza (dello *standard error*), un intervallo di confidenza $C_n$ viene definito come l'insieme dei valori cui $\theta$ appartiene con probabilità $(1-\alpha)\%$ per un qualche $\alpha$. Nell'approccio asintotico si usano i quantili della distribuzione normale e la stima della varianza approssimata, quindi:

$$C_n = \left[\mathbf{b}_j - c\sqrt{\hat{\mathbf{V}}_{jj}},\ \hat{\theta} + c\sqrt{\hat{\mathbf{V}}_{jj}}\right]$$

dove $c = 1.96$ se $\alpha = 0.05$, in quanto la probabilità che una variabile normale standard sia minore di $-1.96$ è 0.025, che sia maggiore di 1.96 è 0.975, quindi che sia compresa tra $-1.96$ e $1 - 96$ è $0.95 = 1 - 0.05$.

R fornisce una funzione `confint()` che, usando il risultato di `lm()`, calcola intervalli di confidenza basati sulla distribuzione $t$ e sull'ipotesi di omoschedasticità. Esite anche una funzione `confint.default()` che usa un'approssimazione normale, ma rimane vincolata all'ipotesi di omoschedasticità; la funzione `confintHC()`, proposta nella figura 3.3, calcola gli intervalli usando la matrice di White (o eventuali varianti).

**Esempio 3.11.** Partendo dalla regressione dell'esempio precedente, `confintHC()` calcola gli intervalli di confidenza mostrati nella figura 3.2. Se si confrontano gli intervalli con i risultati dei test $z$ (fig. 3.1), si può notare che:

a) quasi tutti gli intervalli lasciano dubbi sul segno dei coefficienti (quindi sulla stessa "direzione" degli effetti parziali!); si salvano quelli di `exper` e di `educ`, che risultano anche quelli statisticamente più significativi;

b) il coefficiente di `expersq` risulta moderatamente significativo (al 94.5%), ma il relativo intervallo è talmente stretto intorno allo zero che apparirebbe comunque avventato ipotizzare l'effettiva significatività di un valore non nullo.

### 3.3.3 Test di Wald

Si possono sottoporre a verifica più ipotesi contemporaneamente usando una *matrice di restrizioni* $\mathbf{R}$, per definire un'ipotesi nulla multipla del tipo:

$$\underset{r,k}{\mathbf{R}}\ \underset{k,1}{\boldsymbol{\beta}} - \underset{r,1}{\mathbf{q}} = \underset{r,1}{\mathbf{0}}$$

dove $\mathbf{R}$ è una $r \times 1$ ($r$ è il numero delle restrizioni) e $k$ colonne (il numero dei parametri). Ad esempio, dato il modello:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

a) se l'ipotesi nulla è:

$$H_0 : \beta_2 = \beta_3, \beta_4 = 5$$

si usa:

$$\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} - \begin{bmatrix} 0 \\ 5 \end{bmatrix} = \begin{bmatrix} \beta_2 - \beta_3 \\ \beta_4 \end{bmatrix} - \begin{bmatrix} 0 \\ 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

b) se l'ipotesi nulla è:

$$H_0 : \beta_2 + \beta_3 + \beta_4 = 1$$

si usa:

$$\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} - \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} \beta_2 + \beta_3 + \beta_4 \end{bmatrix} - \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} 0 \end{bmatrix}$$

Essendo lo stimatore di $\boldsymbol{\beta}$ approssimativamente normale per grandi campioni, è tale anche la sua trasformazione lineare $\mathbf{Rb} - \mathbf{q}$:

$$\mathbf{Rb} - \mathbf{q} \overset{a}{\sim} N\left(\mathbf{R}\boldsymbol{\beta} - \mathbf{q}, \mathbf{R}\hat{\mathbf{V}}_{\mathbf{b}}\mathbf{R}'\right)$$

Sotto ipotesi nulla si ha:

$$\mathbf{Rb} - \mathbf{q} \overset{a}{\sim} N(\mathbf{0}, \mathbf{R}\hat{\mathbf{V}}_{\mathbf{b}}\mathbf{R}')$$

---

```
> confintHC(reg)
                2.5 %  97.5 %
(Intercept) -1.03971 0.19789
exper        0.01016 0.06948
expersq     -0.00158 0.00002
educ         0.08135 0.13431
age         -0.01300 0.01007
kidslt6     -0.26696 0.14554
kidsge6     -0.07164 0.04246
```

*Figura 3.2.* Intervalli di confidenza (con approssimazione normale e matrice di White) dei parametri della regressione di cui all'esempio 3.8.

```
confintHC <- function (object, parm, level = 0.95, type = "HC0")
{
  cf <- coef(object)
  pnames <- names(cf)
  if (missing(parm))
    parm <- pnames
  else if (is.numeric(parm))
    parm <- pnames[parm]
  a <- (1 - level)/2
  a <- c(a, 1 - a)
  pct <- paste(format(100*a, trim = TRUE, scientific = FALSE, digits = 3), "%")
  fac <- qnorm(a)
  ci <- array(NA, dim = c(length(parm), 2L), dimnames = list(parm, pct))
  ses <- sqrt(diag(vcovHC(object, type = type)))[parm]
  ci[] <- cf[parm] + ses %o% fac
  ci <- round(ci, 5)
  ci
}
```

*Figura 3.3.* Funzione per il calcolo di intervalli di confidenza con approssimazione normale e matrice di White o sue varianti.

Si può standardizzare dividendo per la radice quadrata della varianza e definendo così la variabile:

$$\mathbf{Z} = (\mathbf{R}\hat{\mathbf{V}}_\mathbf{b}\mathbf{R}')^{-\frac{1}{2}}(\mathbf{Rb} - \mathbf{q})$$

Poiché $\mathbf{R}\hat{\mathbf{V}}_\mathbf{b}\mathbf{R}'$ è simmetrica, si ha:

$$\begin{aligned}\mathbf{Z}^2 = \mathbf{Z}'\mathbf{Z} &= (\mathbf{Rb} - \mathbf{q})'(\mathbf{R}\hat{\mathbf{V}}_\mathbf{b}\mathbf{R}')^{-\frac{1}{2}}(\mathbf{R}\hat{\mathbf{V}}_\mathbf{b}\mathbf{R}')^{-\frac{1}{2}}(\mathbf{Rb} - \mathbf{q}) \\ &= (\mathbf{Rb} - \mathbf{q})'(\mathbf{R}\hat{\mathbf{V}}_\mathbf{b}\mathbf{R}')^{-1}(\mathbf{Rb} - \mathbf{q})\end{aligned}$$

Si perviene così alla statistica test:

$$T = (\mathbf{Rb} - \mathbf{q})'(\mathbf{R}\hat{\mathbf{V}}_\mathbf{b}\mathbf{R}')^{-1}(\mathbf{Rb} - \mathbf{q}) \overset{a}{\sim} \chi^2_r$$

La libreria `car` di R comprende una funzione `linear.hypothesis()`, abbreviabile con `lht()`, che consente di definire restrizioni multiple. La funzione calcola per default test $F$ in omoschedasticità, ma si possono calcolare anche test $\chi^2$ con una matrice di White o sua variante. I parametri più rilevanti sono:

a) `hypothesis.matrix`: può essere una matrice $\mathbf{R}$, oppure una descrizione simbolica delle restrizioni (si rimanda alla guida della libreria per gli esempi);

b) `rhs`: un vettore $\mathbf{q}$ (nullo per default);

c) `test`: `"F"` o `"Chisq"`;

d) `vcov.` (con punto finale): una matrice di varianza e covarianza oppure una funzione per la sua stima, quale la `vcovHC()` della libreria `sandwich`.

**Esempio 3.12.** Per un semplice esempio di utilizzo, si può considerare un'ipotesi nulla relativa ad un solo coefficiente (figura 3.4). Si può notare che il *p-value* coincide con quello già calcolato (figura 3.1).

```
> lht(reg, "age=0", test="Chisq", vcov.=vcovHC(reg, type="HC0"))
Linear hypothesis test

Hypothesis:
age = 0

Model 1: lwage ~ exper + expersq + educ + age + kidslt6 + kidsge6
Model 2: restricted model

Note: Coefficient covariance matrix supplied.

  Res.Df Df Chisq Pr(>Chisq)
1    421
2    422 -1 0.062     0.8034
```

*Figura 3.4.* Un semplice esempio di utilizzo della funzione `linear.hypothesis()`.

```
> reg <- lm(lwage ~ exper+expersq+educ+age+kidslt6+kidsge6, data=mroz)
> reg2 <- lm(lwage ~ exper+expersq+educ, data=mroz)
> waldtest(reg, reg2, vcov=vcovHC(reg, type="HC0"), test="Chisq")
Wald test

Model 1: lwage ~ exper + expersq + educ + age + kidslt6 + kidsge6
Model 2: lwage ~ exper + expersq + educ
  Res.Df Df  Chisq Pr(>Chisq)
1    421
2    424 -3 0.5016     0.9185
```

*Figura 3.5.* Esempio di utilizzo della funzione `waldtest()` per un confronto tra un modello pieno e un modello ridotto.

La libreria `lmtest` contiene invece una funzione `waldtest()` che opera confrontando due o più modelli e accetta anch'essa i parametri `test` e `vcov` (senza punto finale).

**Esempio 3.13.** Restando alla regressione degli esempi precedenti, si nota che i coefficienti relativi alla condizione anagrafica e familiare risultano tutti non significativi (figura 3.1) e con intervalli di confidenza che non consentono di determinarne il segno (figura 3.2). Si può quindi sottoporre a verifica l'ipotesi nulla $H_0 : \beta_4 = \beta_5 = \beta_6$. Invece di costruire una matrice $\mathbf{R}$, si può eseguire una seconda regressione sul modello ridotto:

$$\log(\texttt{wage}) = \beta_1 + \beta_2 \texttt{exper} + \beta_3 \texttt{expersq} + \beta_4 \texttt{educ} + u$$

quindi passare a `waldtest()` i risultati di entrambe le regressioni (figura 3.5). Il *p-value* consente di accettare l'ipotesi nulla "i due modelli sono equivalenti".

### 3.3.4 Test *F*

Si è visto che, nel modello lineare normale (sez. 2.4.4), si usano test $F$ per sottoporre a verifica l'ipotesi nulla "tutti i coefficienti nulli tranne l'intercetta". Con R si possono

```
> F <- summary(reg)
> F
[omissis]
F-statistic: 13.19 on 6 and 421 DF,  p-value: 1.057e-13

> wF <- waldtest(reg)
> wF
[omissis]
  Res.Df Df      F    Pr(>F)
1    421
2    427 -6 13.191 1.057e-13 ***

> wChi <- waldtest(reg, test="Chisq")
> wChi
[omissis]
  Res.Df Df  Chisq Pr(>Chisq)
1    421
2    427 -6 79.144  5.368e-15 ***

> wChi$Chisq[2] / F$fstatistic[1]
value
    6
```

*Figura 3.6.* Confronto tra test $F$ e $\chi^2$.

eseguire analoghi test di Wald in vario modo; ad esempio:

a) `linear.hypothesis(reg, names(coef(reg))[-1], ...)`: il secondo parametro indica di usare i nomi di tutti i coefficienti tranne il primo, che vengono uguagliati a zero per default se non si avvalora il parametro `rhs`;

b) `waldtest(reg, ...)`: indicando un solo modello, questo viene confrontato col modello ridotto contenente la sola intercetta.

Se si usasse una matrice di varianza e covarianza omoschedastica, i risultati non sarebbero molto diversi da quelli che si otterrebbero con `summary()` (cfr. figura 2.2); in particolare i valori della statistica $F$ e della statistica di Wald risulterebbero coerenti e i relativi *p-value* presenterebbero differenze trascurabili.

Si ha infatti che, se $S \sim F_{r,s}$, allora $\lim_{s\to\infty} rS = T \sim \chi^2_r$. Ne segue che, se la statistica $T$ si distribuisce approssimativamente come un $\chi^2_r$, il rapporto $T/r$ si distribuisce approssimativamente come una $F_{r,n-k}$:

$$T \overset{a}{\sim} \chi^2_r \quad \Rightarrow \quad \frac{T}{r} \overset{a}{\sim} F_{r,n-k}$$

Il valore della statistica $F$ calcolato da `summary()` risulterebbe quindi pari al valore della statistica Wald diviso per il numero $r$ delle restrizioni (cfr. figura 3.6).

**Osservazione 3.14.** Hansen (2010, p. 93) sottolinea che test $F$ come quelli calcolati da `summary()` hanno senso solo con piccoli campioni, per valutare se una regressione può aspirare ad avere un qualche valore esplicativo. Con grandi campioni, invece, l'area di

accettazione di un'ipotesi nulla generale si riduce a tal punto che la relativa statistica $F$ risulta quasi sempre molto significativa, quindi inutile.[9]

## 3.4 Il problema delle variabili omesse

Nella pratica, l'ipotesi di esogeneità si scontra spesso con la mancanza di dati o con la difficoltà di una loro espressione quantitativa. Può esserne un esempio il modello:

$$\log(\texttt{wage}) = \beta_1 + \beta_2 \texttt{exper} + \beta_3 \texttt{expersq} + \beta_4 \texttt{educ} + \gamma\, \texttt{abil} + u$$

in cui risulta arduo disporre di misure dell'abilità e, quindi, di relativi dati. Come già visto, ciò comporta che, se si regredisse la variabile risposta solo sulle variabili disponibili, verrebbe meno la consistenza degli stimatori e non sarebbe possibile stimare gli effetti parziali. In tali casi, si dice che l'*equazione strutturale* (quella relativa al modello "vero") non è stimabile e si cerca, quindi, una *equazione stimabile*.

In generale si cerca di escludere le variabili non disponibili e di sostituirle con altre. Il prossimo capitolo è dedicato all'inclusione di *variabili strumentali*, mentre qui si illustra il metodo delle *variabili proxy*: una variabile proxy è una variabile che ha sulla variabile risposta un effetto paragonabile a quello della variabile mancante ed è a questa correlata al punto di poter ipotizzare che la variabile proxy agisce sulla variabile risposta "per procura" di quella mancante (una persona *proxy* è appunto una persona che agisce per procura).

Più formalmente, se l'equazione strutturale è:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \gamma q + u$$

se $q$ è correlata con qualche $x_j$ (quindi non può essere relegata nell'errore) ma non è disponibile, si può sostituire $q$ con una variabile proxy $z$ che soddisfi le seguenti condizioni:

a) *ridondanza*: se $q$ fosse disponibile, $z$ sarebbe inutile:

$$\mathbb{E}[y \mid \mathbf{x}, q, z] = \mathbb{E}[y \mid \mathbf{x}, q]$$

b) *correlazione forte*: $z$ deve essere correlata a $q$ in modo che, una volta inclusa $z$ nell'equazione, $q$ non sia più correlata con alcuna $x_j$:

$$L(q \mid 1, x_2, \ldots, x_k, z) = L(q \mid 1, z)$$

dove $L(a \mid b)$ indica la proiezione ortogonale di $a$ sullo spazio generato da $b$.

La seconda condizione ricorre all'aspetto geometrico dei modelli di regressione lineare (cfr. sez. 2.4.1). Sia $\mathbf{X}_{xz}$ una matrice avente $k+2$ colonne, una prima costituita da tutti 1, le altre per le variabili $x_2, \ldots, x_k$ e $z$. $\mathbf{X}_{xz}$ può essere vista come somma di due matrici $\mathbf{X}_x$, le cui colonne non nulle contengono $x_2, \ldots, x_k$, e $\mathbf{X}_z$, le cui colonne non nulle contengono 1 e $z$. Indicando con $\mathrm{Im}(\mathbf{X})$ lo spazio generato dalle colonne di una matrice $\mathbf{X}$, si ha (cfr. proposizione A.23):

$$\mathrm{Im}(\mathbf{X}_{xz}) = \mathrm{Im}(\mathbf{X}_x) + \mathrm{Im}(\mathbf{X}_z)$$

[9] Può bastare considerare che $F^* = \dfrac{ESS/(k-1)}{RSS/(n-k)}$ aumenta all'aumentare di $n$.

Se $q = \theta_1 + \theta_2 z + r$, e se $\theta_2 \neq 0$, si può proiettare $q$ sullo spazio generato da $\mathbf{X}_z$ e indicare tale proiezione con:

$$L(q \mid 1, z)$$

Indicando con $L(q \mid 1, x_2, \ldots, x_k, z)$ l'analoga proiezione sullo spazio generato da $\mathbf{X}_{xz}$, se si ha:

$$L(q \mid 1, x_2, \ldots, x_k, z) = L(q \mid 1, z)$$

ne segue che $\mathbb{E}[q] = \theta_1 + \theta_2 z$ può essere espresso come combinazione lineare dei vettori di una base sia di $\mathrm{Im}(\mathbf{X}_{xz})$ che di $\mathrm{Im}(\mathbf{X}_z)$, ma nel primo caso i coefficienti dei termini $x_j$ sarebbero tutti nulli. Quanto a $r = q - \mathbb{E}[q]$, se vale l'uguaglianza $r$ è ortogonale a tutto $\mathrm{Im}(\mathbf{X}_{xz})$, quindi per ogni $j = 2, \ldots, k$:

$$L(q \mid 1, x_2, \ldots, x_k, z) = L(q \mid 1, z) \quad \Rightarrow \quad \mathrm{Cov}(r, x_j) = 0,\ \mathrm{Cov}(q, x_j) = 0$$

Se valgono entrambe le condizioni, ponendo

$$q = \theta_1 + \theta_2 z + r \qquad\qquad \mathbb{E}[r] = 0,\ \mathrm{Cov}(z, r) = 0$$

si perviene all'equazione:

$$y = (\beta_1 + \gamma\theta_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \gamma(\theta_2 z) + (\gamma r + u)$$

che risulta stimabile in quanto viene rispettata l'ipotesi di esogeneità (in particolare, l'errore $\gamma r + u$ non è correlato con alcuna variabile esplicativa).

**Esempio 3.15.** Si vuole ragionare sul salario usando i dati del *National Longitudinal Survey* del 1980.[10] Si definisce l'equazione strutturale:

$$\begin{aligned}\log(\texttt{wage}) =& \beta_1 + \beta_2 \texttt{exper} + \beta_3 \texttt{tenure} + \beta_4 \texttt{married} + \beta_5 \texttt{south} \\ &+ \beta_6 \texttt{urban} + \beta_7 \texttt{black} + \beta_8 \texttt{educ} + \gamma \texttt{abil} + u\end{aligned}$$

dove:

- `wage` è il salario mensile, `lwage` il suo logaritmo;
- `exper` è l'anzianità di lavoro totale in anni;
- `tenure` è l'anzianità di lavoro nell'azienda;
- `married` vale 1 se il lavoratore è sposato;
- `south` vale 1 se il lavoratore vive negli stati del sud;
- `urban` vale 1 se il lavoratore vive in un'area metropolitana;
- `black` vale 1 se il lavoratore è nero;
- `educ` il livello di istruzione misurato con gli anni di frequentazione delle scuole;
- `abil` è l'abilità.

[10] File `http://web.mclink.it/MC1166/Econometria/nls80.csv`. Rispetto ai dati contenuti nel file scaricabile dal sito di Wooldridge (2002), si sono assegnati i nomi di colonna e si è usata la stringa `NA` per i dati mancanti.

Non è disponibile una misura dell'abilità, che non può essere scartata in quanto è facile supporre una sua correlazione almeno con `educ`. Sono disponibili i quozienti di intelligenza `iq`; si valuta che, se fossero disponibili dati circa l'abilità, il quozienge di intelligenza sarebbe inutile (ridondanza), si ipotizza che sia rispettata anche la seconda condizione. Si prova quindi ad eseguire una regressione sia sull'equazione strutturale senza la variabile `abil`, sia su un'equazione stimabile con `iq` al posto di `abil`:

```
> nls80 <- read.csv("nls80.csv")
> regomitted <- lm(lwage ~ exper+tenure+married+south+urban+black+educ,
+ data=nls80)
> regproxy <- lm(lwage ~ exper+tenure+married+south+urban+black+educ+iq,
+ data=nls80)
```

Si possono confrontare i risultati unendo i test sui coefficienti e gli intervalli di confidenza delle due regressioni, con comandi del tipo:

```
> cbind(coeftest(reg, ...), confintHC(reg, ...))
```

Osservando la figura 3.7, si può notare che nella regressione con la variabile proxy il coefficiente di `educ` si riduce da 0.065 a 0.054 e aumenta lo *standard error*, anche se il suo intervallo di confidenza al 95% rimane coerente (estremi entrambi positivi) e piuttosto stretto.

## 3.5 Il problema degli errori di misura

Può succedere che una variabile sia osservabile, ma che i dati disponibili non siano pienamente attendibili. Si può pensare, per un esempio, ai risparmi delle famiglie: si tratta di un aggregato che potrebbe essere misurato con esattezza, ma se i dati disponibili derivano da risposte dei diretti interessati potrebbero essere imprecisi.

Se l'errore di misura riguarda la variabile risposta, il modello assume la forma:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + (u + \varepsilon)$$

dove $\varepsilon$ è l'errore di misura della variabile risposta. Appare naturale assumere che $\mathbb{E}[\varepsilon] = 0$, può esserlo meno assumere anche che $\varepsilon$ sia incorrelato con le variabili esplicative. Se ciò appare possibile, si può comunque procedere con la regressione. Si può solo notare che, se $u$ e $\varepsilon$ non sono correlati (come è spesso ragionevole assumere), la varianza complessiva dell'errore sarà somma delle loro varianze; si avranno quindi stime con un maggiore *standard error*, ma comunque consistenti.

In realtà l'errore di misura problematico è quello relativo alle variabili esplicative, che può presentarsi in due forme.

Sia $x_k$ una variabile affetta da un errore $\varepsilon_k$ e si disponga solo della sua misura $\tilde{x}_k$, con $\varepsilon_k = \tilde{x}_k - x_k$. Il modello diventa:

$$\begin{aligned} y &= \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \\ &= \beta_1 + \beta_2 x_2 + \cdots + \beta_k (\tilde{x}_k - \varepsilon_k) + u \\ &= \beta_1 + \beta_2 x_2 + \cdots + \beta_k \tilde{x}_k + (u - \beta_k \varepsilon_k) \end{aligned}$$

```
> round(cbind(
+ coeftest(regomitted, df=Inf, vcov.=vcovHC(regomitted, type="HC0")),
+ confintHC(regomitted)), 5)
            Estimate Std. Error  z value Pr(>|z|)    2.5 %   97.5 %
(Intercept)  5.39550    0.11264 47.89935  0.00000  5.17472  5.61627
exper        0.01404    0.00322  4.35484  0.00001  0.00772  0.02036
tenure       0.01175    0.00253  4.64721  0.00000  0.00679  0.01670
married      0.19942    0.03952  5.04552  0.00000  0.12195  0.27688
south       -0.09090    0.02725 -3.33644  0.00085 -0.14430 -0.03750
urban        0.18391    0.02700  6.81251  0.00000  0.13100  0.23682
black       -0.18835    0.03655 -5.15375  0.00000 -0.25998 -0.11672
educ         0.06543    0.00638 10.25271  0.00000  0.05292  0.07794

> round(cbind(
+ coeftest(regproxy, df=Inf, vcov.=vcovHC(regproxy, type="HC0")),
+ confintHC(regproxy)), 5)
            Estimate Std. Error  z value Pr(>|z|)    2.5 %   97.5 %
(Intercept)  5.17644    0.12064 42.90859  0.00000  4.93999  5.41289
exper        0.01415    0.00322  4.38964  0.00001  0.00783  0.02046
tenure       0.01140    0.00252  4.51376  0.00001  0.00645  0.01634
married      0.19976    0.03890  5.13520  0.00000  0.12352  0.27601
south       -0.08017    0.02760 -2.90424  0.00368 -0.13427 -0.02607
urban        0.18195    0.02661  6.83678  0.00000  0.12979  0.23411
black       -0.14313    0.03746 -3.82032  0.00013 -0.21655 -0.06970
educ         0.05441    0.00724  7.51747  0.00000  0.04022  0.06860
iq           0.00356    0.00095  3.73942  0.00018  0.00169  0.00542
```

*Figura 3.7.* Risultati di una regressione con variabile omessa e di un'altra con variabile proxy.

Se $\mathrm{Cov}(\tilde{x}_k, \varepsilon_k) = 0$ non c'è problema: l'ipotesi di esogeneità è rispettata e si ottengono stimatori consistenti, anche se con una maggiore varianza dell'errore.

Se, tuttavia, $\mathrm{Cov}(x_k, \varepsilon_k) = 0$, allora $\tilde{x}_k$ e $\varepsilon_k$ sono necessariamente correlate. Infatti, assumendo $\mathbb{E}[\varepsilon_k] = 0$ (se così non fosse, basterebbe aggiungere la media all'intercetta del modello),

$$\begin{aligned}\mathrm{Cov}(x_k, \varepsilon_k) &= \mathbb{E}[x_k \varepsilon_k] = 0\\ \mathrm{Cov}(\tilde{x}_k, \varepsilon_k) &= \mathbb{E}[\tilde{x}_k \varepsilon_k] = \mathbb{E}[(x_k + \varepsilon_k)\varepsilon_k] = \mathbb{E}[\varepsilon_k^2] = \sigma^2_{\varepsilon_k}\end{aligned}$$

ovvero la covarianza tra $\tilde{x}_k$ e $\varepsilon_k$ è uguale alla varianza dell'errore di misura. Una regressione porterebbe a stimatori non consistenti.

Si tratta quindi di valutare, caso per caso, se l'errore di misura è correlato al valore vero oppure al valore errato.

# Capitolo 4

# Le variabili strumentali

Se il modello della popolazione comprende $k-1$ variabili esogene (compresa, al solito, $x_1 = 1$) e una endogena:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$
$$\mathbb{E}[u] = 0 \qquad \mathrm{Cov}(x_j, u) = 0,\ j = 2, \ldots, k-1 \qquad \mathrm{Cov}(x_k, u) \neq 0$$

operando come nel capitolo precedente si otterrebbero stimatori non consistenti.

Si cerca quindi di prendere in considerazione ulteriori variabili, dette *variabili strumentali*,[1] che siano sia esogene che correlate con $x_k$: l'esogeneità assicura la consistenza, la correlazione fa sì che le nuovi variabili possano spiegare in buona parte la quota che verrebbe spiegata da $x_k$ della variabilità di $y$.

## 4.1 Una sola variabile strumentale

Una variabile $z_1$, diversa dalle $x_j$, può essere usata come variabile strumentale se soddisfa le seguenti due condizioni:

a) *esogeneità*:

$$\mathrm{Cov}(z_1, u) = 0$$

b) *correlazione parziale*: deve esistere una proiezione ortogonale di $x_k$ sullo spazio generato da *tutte* le esogene, compresa in particolare $z_1$:

$$x_k = \delta_1 + \delta_2 x_2 + \cdots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + r_k, \qquad \theta_1 \neq 0$$

ovvero $z_1$ deve risultare parzialmente correlata con $x_k$ al netto delle altre esogene.[2]

[1] Si usa spesso chiamare *strumenti* le nuove variabili, strumentali tutte le variabili esogene (sia quelle già presenti nel modello che quelle nuove). Qui si diranno strumentali solo le variabili aggiunte alle esogene già considerate.

[2] In altri termini, $z_1$ non deve risultare correlata a $x_k$ solo perché correlata con alcune delle $x_2, \ldots, x_{k-1}$ a loro volta sono correlate con $x_k$.

L'equazione della proiezione di $x_k$ viene detta *equazione in forma ridotta*. Sostituendo $x_k$ nell'equazione strutturale, si ottiene l'equazione in forma ridotta per $y$:

$$\begin{aligned} y &= \beta_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} \\ &\quad + \beta_k(\delta_1 + \delta_2 x_2 + \cdots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + r_k) + u \\ &= (\beta_1 + \beta_k \delta_1) + (\beta_2 + \beta_k \delta_2) x_2 \\ &\quad + (\beta_{k-1} + \beta_k \delta_{k-1}) x_{k-1} + (\beta_k \theta_1) z_1 + (u + \beta_k r_k) \\ &= \alpha_1 + \alpha_2 x_2 + \cdots + \alpha_{k-1} x_{k-1} + \lambda_1 z_1 + v \end{aligned}$$

Se una variabile $z_1$ rispetta le condizioni di esogeneità e di correlazione parziale, è possibile stimare $\boldsymbol{\beta}$ affiancando all'equazione strutturale quella in forma ridotta.

Sia $y = \mathbf{x}'\boldsymbol{\beta} + u$ l'equazione strutturale del modello originario, con $\mathbf{x} = (1, x_2, \ldots, x_k)$. Sia inoltre $\mathbf{z}$ il vettore di tutte le esogene: $\mathbf{z} = (1, x_2, \ldots, x_{k-1}, z_1)$. Si ha ovviamente $\mathbb{E}[\mathbf{z}u] = 0$, per l'esogeneità sia di $x_2, \ldots, x_k$ che di $z_1$.

Premoltiplicando l'equazione strutturale per $\mathbf{z}$ e calcolando i valori attesi si ha:

$$\mathbb{E}[\mathbf{z}y] = \mathbb{E}[\mathbf{z}\mathbf{x}'\boldsymbol{\beta}] + \mathbb{E}[\mathbf{z}u] = \boldsymbol{\beta}\mathbb{E}[\mathbf{z}\mathbf{x}']$$

Se $\mathbb{E}[\mathbf{z}\mathbf{x}']$, una matrice $k \times k$, ha rango pieno, il sistema di equazioni ammette un'unica soluzione:

$$\boldsymbol{\beta} = \mathbb{E}[\mathbf{z}\mathbf{x}']^{-1}\mathbb{E}[\mathbf{z}y]$$

Si giunge quindi, analogamente a quanto già visto nel capitolo precedente (sez. 3.2), allo stimatore consistente:

$$\mathbf{b}_{IV} = \left(n^{-1}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\mathbf{z}_i y_i\right) = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

Affinché $\mathbb{E}[\mathbf{z}\mathbf{x}']$ abbia rango pieno, deve essere rispettata la condizione di correlazione parziale: $x_k$ non deve risultare correlata solo con $x_2, \ldots, x_{k-1}$.[3]

**Osservazione 4.1.** Le condizioni che una variabile proxy deve soddisfare non sono verificabili, in quanto la variabile sostituita è, per definizione, non osservabile. Quando si considera una possibile variabile strumentale risulta non verificabile la condizione di esogeneità (cfr. osservazione 3.1), ma *la condizione di correlazione parziale può, e dovrebbe, essere verificata.*

[3] Se $\mathbf{X}$ e $\mathbf{Z}$ fossero matrici $4 \times 3$, si avrebbe:

$$\mathbf{Z}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_{12} & x_{22} & x_{32} & x_{42} \\ z_{11} & z_{21} & z_{31} & z_{41} \end{bmatrix} \begin{bmatrix} 1 & x_{12} & x_{13} \\ 1 & x_{22} & x_{23} \\ 1 & x_{32} & x_{33} \\ 1 & x_{42} & x_{43} \end{bmatrix}$$

$$= \begin{bmatrix} 1+1+1+1 & x_{12}+x_{22}+x_{32}+x_{42} & x_{13}+x_{23}+x_{33}+x_{43} \\ x_{12}+x_{22}+x_{32}+x_{42} & x_{12}^2+x_{22}^2+x_{32}^2+x_{42}^2 & x_{12}x_{13}+x_{22}x_{23}+x_{32}x_{33}+x_{42}x_{43} \\ z_{11}+z_{21}+z_{31}+z_{41} & z_{11}x_{12}+z_{21}x_{22}+z_{31}x_{32}+z_{41}x_{42} & z_{11}x_{13}+z_{21}x_{23}+z_{31}x_{33}+z_{41}x_{43} \end{bmatrix}$$

Se fosse $x_3 = \delta x_2$, la terza colonna sarebbe proporzionale alla seconda e la matrice non avrebbe rango pieno; prendendo la terza riga:

$$z_{11}x_{13} + z_{21}x_{23} + z_{31}x_{33} + z_{41}x_{43} = \delta(z_{11}x_{12} + z_{21}x_{22} + z_{31}x_{32} + z_{41}x_{42})$$

Se invece $x_3 = \delta x_2 + \theta z_1$, $\theta \neq 0$, la proporzionalità viene meno.

## 4.2 Più variabili strumentali

Se si dispone di più variabili strumentali $z_1, \ldots, z_m$, che soddisfino tutte i requisiti di esogeneità e correlazione parziale, l'equazione in forma ridotta per $x_k$ diventa:

$$x_k = \delta_1 + \delta_2 x_2 + \cdots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + \cdots + \theta_m z_m + r_k$$

Non si può procedere come nel caso di una sola strumentale, in quanto il vettore delle esogene

$$\mathbf{z} = (1, x_2, \ldots, x_{k-1}, z_1, \ldots, z_m)$$

che è un vettore di $l = (k-1) + m$ elementi, non è ora moltiplicabile per il vettore $\mathbf{x}$ delle variabili strutturali.

Tuttavia, l'equazione in forma ridotta di $x_k$ è un'equazione stimabile. Infatti:

$$\begin{aligned} r_k &= x_k - (\delta_1 + \delta_2 x_2 + \cdots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + \cdots + \theta_m z_m) \\ &= x_k - L(x_k \mid \mathbf{z}) \\ &= x_k - x_k^* \end{aligned}$$

ovvero $r_k$ è ortogonale alla proiezione di $x_k$ su $\mathbf{z}$, indicata con $x_k^*$, quindi è incorrelato con le esogene.

Inoltre $x_k^*$, essendo una combinazione lineare di esogene, è anch'essa un'esogena e può essere usata come unica variabile strumentale.

Si procede quindi in due passi, con una regressione detta 2SLS (*Two-Stage Least Squares*):

1) si stima $\hat{x}_k$ dalla sua equazione in forma ridotta usando $\mathbf{Z}$, la matrice $n \times (k-1+m)$ contenente le $n$ determinazioni di $\mathbf{z}$ nel campione estratto:

$$\begin{aligned} \mathbf{d}_{OLS} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}_k \\ \hat{\mathbf{x}}_k &= \mathbf{Z}\mathbf{d}_{OLS} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}_k = \mathbf{H}\mathbf{x}_k \end{aligned}$$

2) si crea una matrice $\hat{\mathbf{X}}$ sostituendo la colonna della matrice $\mathbf{X}$ contenente gli $x_{ik}$ con gli $\hat{x}_{ik}$ e si stimano i parametri $\boldsymbol{\beta}$ dell'equazione strutturale:

$$\mathbf{b}_{IV} = \left( \sum_{i=1}^{n} \hat{\mathbf{x}}_i \mathbf{x}' \right)^{-1} \left( \sum_{i=1}^{n} \hat{\mathbf{x}}_i y_i \right) = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

poiché $\hat{\mathbf{X}} = \mathbf{H}\mathbf{X}$,[4] ed essendo $\mathbf{H}$ simmetrica e idempotente, si ha:

$$\hat{\mathbf{X}}'\mathbf{X} = \mathbf{X}'\mathbf{H}'\mathbf{X} = \mathbf{X}'\mathbf{H}'\mathbf{H}\mathbf{X} = \hat{\mathbf{X}}'\hat{\mathbf{X}}$$

si può quindi usare la sola matrice $\hat{\mathbf{X}}$:

$$\mathbf{b}_{IV} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

[4] La matrice $\mathbf{H}$ proietta le colonne di $\mathbf{X}$ sullo spazio generato dalle colonne di $\mathbf{Z}$. Le $k-1$ colonne di $\mathbf{X}$ che sono già in $\mathbf{Z}$ (corrispondono alle strutturali esogene) rimangono immutate, la colonna con le osservazioni della variabile endogena viene modificata e contiene le relative stime.

```
> # primo stadio
> # - OLS dell'endogena su tutte le esogene
> olsreg <- lm(log(price) ~ log(income) + stax + etax, data=cigarettes)

> # secondo stadio
> # - creazione della matrice X
> n <- nrow(cigarettes)
> X <- matrix(c(rep(1,n),log(cigarettes$price),log(cigarettes$income)), nrow=n)
> # - creazione della matrice con le stime di log(price)
> Xhat <- X
> Xhat[,2] <- olsreg$fitted.values
> # - stima dei coefficienti beta
> solve(t(Xhat) %*% Xhat) %*% t(Xhat) %*% log(cigarettes$packs)
           [,1]
[1,]  9.8949555
[2,] -1.2774241
[3,]  0.2804048
```

*Figura 4.1.* Esecuzione separata dei due stadi di una regressione 2SLS. La stima dei coefficienti è identica a quella che si ottiene con la funzione `ivreg()`.

**Esempio 4.2.** Il dataset `cigarettes.csv`[5] contiene dati relativi al consumo di sigarette nei 48 stati continentali degli USA nel 1995:

- `state`: lo stato;
- `packs`: il numero pro capite di pacchetti di sigarette;
- `price`: il prezzo medio alla vendita;
- `income`: il reddito pro capite;
- `stax`: imposta media sulle vendite (*sales tax*, imposta *ad valorem* analoga all'IVA);
- `etax`: imposta media sulla produzione (*excise tax*, imposta specifica).

Gli importi delle ultime quattro variabili sono reali, non nominali (sono divisi per l'indice dei prezzi al consumo). Si muove dal modello:

$$\log(\texttt{packs}) = \beta_1 + \beta_2 \log(\texttt{price}) + \beta_3 \log(\texttt{income}) + u$$

ma si considera che il prezzo può essere effetto di fattori non considerati. Tra questi hanno sicuramente rilievo le imposte sulla produzione e sulle vendite, che peraltro appaiono verosimilmente esogene. Si definisce quindi un vettore di esogene contenente log(`income`), `stax` e `etax`. Per eseguire la regressione con R si può usare la funzione `ivreg()`, contenuta nella libreria `AER`:

```
> reg <- ivreg(log(packs) ~ log(price) + log(income) |
+ log(income) + stax + etax, data = cigarettes)
```

La funzione richiede come primo argomento una formula in cui si indichino prima le variabili strutturali e poi quelle esogene, separate da una barra verticale. I coefficienti $\mathbf{b}_{IV}$ che si ottengono sono:

[5] Adattato dal dataset `CigarettesSW` contenuto nella libreria `AER` di R e scaricabile da http://web.mclink.it/MC1166/Econometria/cigarettes.csv.

```
> coef(reg)
(Intercept)  log(price) log(income)
  9.8949555  -1.2774241   0.2804048
```

La figura 4.1 mostra come si potrebbe ottenere lo stesso risultato eseguendo separatamente i due stadi della regressione.

**Osservazione 4.3.** Nell'esempio precedente prezzi e redditi sono reali, non nominali. Ciò consente di ritenere le due variabili non correlate. Si potrebbe forse pensare di risolvere il problema della endogeneità dei prezzi semplicemente eliminandoli dal modello. Per quanto ovvio, si sottolinea che così si otterrebbero sì stimatori consistenti, ma diminuirebbe la quota spiegata della variabilità della variabile risposta (misurata da $R^2$). In questo caso la diminuzione sarebbe vistosa. La libreria `AER` contiene una versione di `summary()` specifica per il risultato di `ivreg()`; eseguendo `summary(reg)` sulla regressione descritta nell'esempio si ottengono $R^2 = 0.43$ e $\overline{R}^2 = 0.40$. Eseguendo una regressione OLS solo su log(`income`):

```
> lm(log(packs) ~ log(income), data=cigarettes)
```

si otterrebbero $R^2 = 0.038$ e $\overline{R}^2 = 0.017$.

# Capitolo 5

# Variabile risposta qualitativa

La variabile risposta potrebbe essere qualitativa (lavorare o non lavorare, comprare o non comprare un bene ecc.). In questi casi, viene espressa con un numero finito di possibili risultati e quello che interessa è determinare la probabilità di ciascuno in funzione di un vettore $\mathbf{x}$ di variabili esplicative. Nei modelli a risposta binaria vi sono due possibili risultati, $y = 1$ e $y = 0$, il primo dei quali si verifica con probabilità $p$ e l'altro con probabilità $q = 1 - p$.

## 5.1 Logit e probit

Nei modelli logit e probit Si usa sottindere un *modello a variabile latente*: si muove dai consueti vettori di esplicative $\mathbf{x}$ e di parametri $\boldsymbol{\beta}$, si definisce una variabile

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \qquad y = I(y^* > 0)$$

dove $I()$ è la funzione indicatrice, si pone:

$$P[y = 1 \mid \mathbf{x}] = P[y^* > 0 \mid \mathbf{x}] = P[\varepsilon > -\mathbf{x}'\boldsymbol{\beta} \mid \mathbf{x}] = 1 - G(-\mathbf{x}'\boldsymbol{\beta}) = G(\mathbf{x}'\boldsymbol{\beta})$$

dove $G$ è una funzione di ripartizione. Si usa anche indicare $P[y = 1 \mid \mathbf{x}]$ con $p(\mathbf{x})$:

$$p(\mathbf{x}) \equiv P[y = 1 \mid \mathbf{x}] = G(\mathbf{x}'\boldsymbol{\beta})$$

I due modelli si distinguono per la funzione di ripartizione adottata: in *probit* si tratta di $\Phi(z)$, la funzione di ripartizione della normale standard, mentre in *logit* si usa quella della distribuzione logistica:

$$\begin{aligned} \text{logit}: &\qquad G(z) = \frac{e^z}{1 + e^z} \\ \text{probt}: &\qquad G(z) = \Phi(z) \end{aligned}$$

Poiché $p(\mathbf{x})$ è una funzione di $\mathbf{x}'\boldsymbol{\beta}$, per determinare gli effetti parziali di una variabile esplicativa continua si deve ricorrere alla regola di derivazione delle funzioni composte:

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\mathbf{x}'\boldsymbol{\beta})\beta_j, \qquad g(z) = \frac{dG(z)}{dz}$$

Se la variabile esplicativa $x_h$ è binaria, invece, l'effetto parziale non è altro che:

$$G\left(\sum_{j=1}^{k}\beta_j x_j\right) - G\left(\sum_{j=1,\, j\neq h}^{k}\beta_j x_j\right)$$

Analogamente nel caso di variabili discrete a più di due valori.

In ogni caso, poiché $G$ non è lineare, l'effetto dipende dal punto in cui viene calcolato. Hanno quindi senso solo valutazioni degli effetti relativi di due esplicative.

Si usano stimatori di massima verosimiglianza, che sono per loro natura asintoticamente corretti, consistenti (a condizioni di regolarità soddisfatte dalle distribuzioni normale e logistica) e asintoticamente normali.

# Parte II

# Serie storiche

# Capitolo 6

# La regressione spuria

Nei dati *cross-section* le singole osservazioni sono contraddisinte da un indice $i$ che si riferisce alla $i$-esima unità del campione estratto. Nelle serie storiche le singole osservazioni si riferiscono a istanti o periodi di tempo diversi, quindi di usa un indice $t$. Sembra una modifica solo formale: se posso usare vari tipi di regressione su dati del tipo:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

sembrerebbe ovvio usare le stesse tecniche su dati del tipo:

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk}$$

Eppure, così facendo, si perviene a risultati paradossali, a volte palesemente assurdi.

Il problema venne posto con chiarezza da G. Udny Yule nel 1926 e ha poi trovato una soluzione solo gradualmente.[1]

## 6.1 Matrimoni religiosi e mortalità

Yule (1926) propose la correlazione tra la percentuale di matrimoni religiosi e la mortalità (numero di morti ogni 1000 abitanti) in Inghilterra e nel Galles dal 1866 al 1911.[2]

Il grafico delle due serie mostra una notevole somiglianza (figura 6.1). Yule calcolò la correlazione, trovando un valore decisamente elevato. Fece i calcoli a mano, ottenendo 0.9512; con R si ottiene un valore poco diverso: 0.9515. Se invece si prova ad eseguire una regressione, si ottiene un $R^2$ di 0.9054![3]

Dobbiamo pensare, commentava Yule, che si tratti di una correlazione spuria? Forse matrimoni religiosi e mortalità appaiono correlati tra loro solo perché entrambi correlati con un'altra variabile? Con un po' di fantasia e di buona volontà, aggiungeva, si può pensare che quest'altra variabile sia il progresso della scienza, che fa diminuire sia la mortalità che le manifestazioni religiose. In realtà, concludeva, è più ragionevole pensare che si tratti solo di una correlazione senza senso: «But most people would, I think, agree with me that the correlation is simply sheer nonsense».

[1] I capitoli della parte II si basano liberamente su Hamilton (1994), Hansen (2010) e Lucchetti (2008).

[2] I dati sono scaricabili da `http://web.mclink.it/MC1166/Econometria/yule.csv`. Per caricarli in R come serie storica si può usare il comando `read.ts("yule.csv", header=TRUE, sep=",", start=1866)`.

[3] Ovviamente, visto che $R^2$ è il quadrato del coefficiente di correlazione.


*Figura 6.1.* Percentuali di matrimoni religiosi e indici di mortalità (numero di morti per 1000 abitanti) in Inghilterra e nel Galles dal 1866 al 1911.

La conclusione può apparire opinabile, l'ipotesi di correlazione spuria forse meno fantasiosa di quanto Yule credeva, ma in seguito è emersa sempre più chiaramente la possibilità di individuare serie temporali indiscutibilmente indipendenti e tuttavia tali da esibire coefficienti $R^2$ di tutto rispetto. Una regressione che non tenesse conto di ciò rischierebbe di rivelarsi una *regressione spuria*.[4]

## 6.2 Processi stocastici

### 6.2.1 Con memoria

Prima di affrontare formalmente il problema, si possono considerare due semplici situazioni:

a) si lancia una moneta regolare $n$ volte;

b) si lancia una moneta e:

- se viene testa si fa un passo avanti a destra;
- se viene croce si fa un passo avanti a sinistra.

[4]Curiosamente, le regressioni "incaute" eseguite su serie storiche vengono dette *spurie*, mentre Yule parlava di correlazioni *senza senso* in opposizione a possibili loro interpretazioni come correlazioni spurie (anche se non usava questo termine).

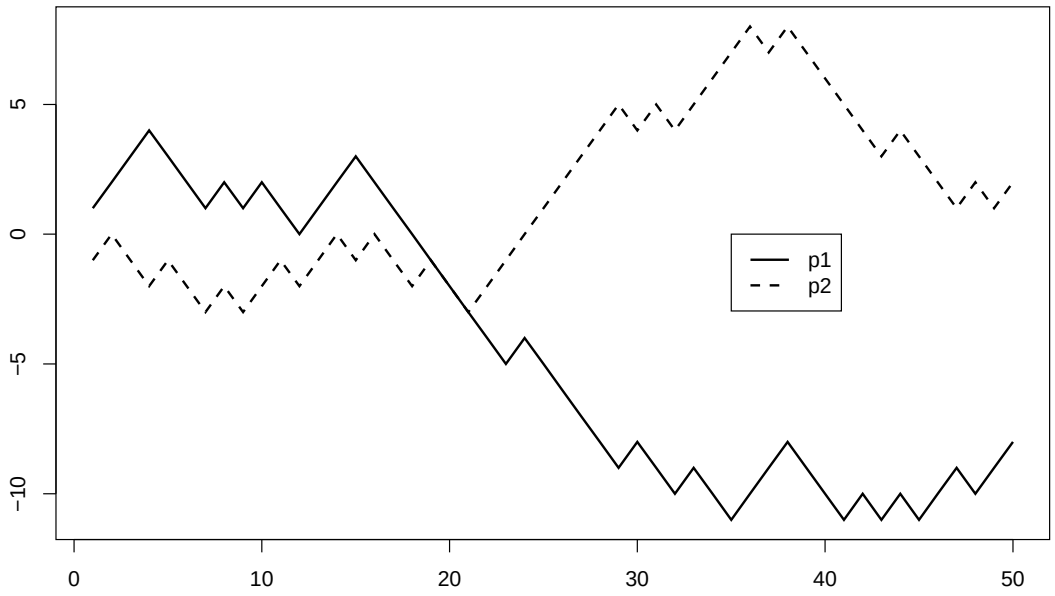


*Figura 6.2.* Due passeggiate governate dal lancio di una moneta: se viene testa si va in su, se viene croce si va in giù.

Il primo caso è una stilizzazione dei dati *cross-section*: vi sono $n$ realizzazioni di variabili aleatorie indipendenti e identicamente distribuite; realizzazioni che vengono indicizzate con una $i$ che va da 1 a $n$, ma possono essere mescolate a piacere. Il loro ordine non ha alcuna importanza.

Nel secondo caso la materia prima sembra identica (più lanci di una stessa moneta, tra loro indipendenti), ma gli spostamenti impongono un ordinamento che non può essere ignorato: la posizione in cui ci si trova dopo un lancio non dipende solo dal suo esito, ma anche dalla posizione in cui si era a seguito dei lanci precedenti. Si parla quindi di *processo stocastico*, o *aleatorio*, "con memoria" (a rigore, con *persistenza*).

Simulare una passeggiata governata dal lancio di una moneta è semplice. Simulandone due, si possono confrontare due processi assolutamente indipendenti:

```
> n <- 50
> m1 <- rbinom(n, 1, 0.5)
> m1[m1==0] <- -1
> p1 <- cumsum(m1)
> m2 <- rbinom(n, 1, 0.5)
> m2[m2==0] <- -1
> p2 <- cumsum(m2)
```

La figura 6.2 propone una rappresentazione grafica dei due processi. La correlazione tra `p1` e `p2` è maggiore di 0.82 e, se si esegue una regressione, si ottiene un $R^2$ pari a 0.675!

### 6.2.2 Senza memoria

L'ordine dei processi `p1` e `p2` è intoccabile, ma non basta a spiegare l'apparente correlazione tra processi indipendenti.

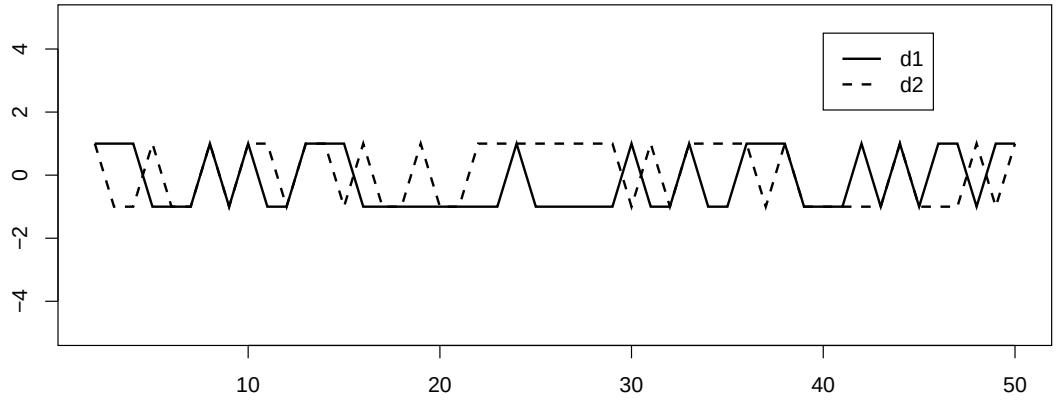


*Figura 6.3.* Grafici delle differenze tra due passi consecutive delle passeggiate mostrate nella figura 6.2.

Lasciando inalterato l'ordine, si possono esaminare le *differenze* tra i singoli passi delle due passeggiate:

```
> d1 <- diff(p1)
> d2 <- diff(p2)
```

Lo scenario cambia radicalmente (figura 6.3). Soprattutto, il coefficiente di correlazione e $R^2$ precipitano, rispettivamente, a 0.03 e 0.001. Cosa è successo?

Appare chiara l'esigenza di una struttura teorica di riferimento.

## 6.3 Definizioni

Si può definire un processo stocastico come una successione di variabili aleatorie. Più formalmente:

**Definizione 6.1.** Dato uno spazio di probabilità $(\Omega, \mathcal{A}, P)$, un *processo stocastico* (o *aleatorio*) con supporto $\mathcal{X}$ è una successione $I \to \mathcal{X}$ di variabili aleatorie a valori in $\mathcal{X}$ e si indica con $\{x_t\}_{-\infty}^{+\infty}$. Un processo stocastico viene detto *a tempo discreto* oppure *a tempo continuo* secondo la natura discreta o continua di $I$.

Qui si considereranno solo processi a tempo discreto con un insieme di indici $I = \mathbb{N}$. Si userà invece $T$ per indicare il numero di elementi in un campione estratto da un processo stocastico. Si userà inoltre $y_t$ per indicare sia un processo stazionario che un suo elemento.

Ciascun elemento $y_t$ di un processo stocastico ha una sua funzione di densità $f_{y_t}(y_t)$, nonché un suo valore atteso e una sua varianza: $\mathbb{E}[y_t] = \int_{-\infty}^{+\infty} y_t \, f_{y_t}(y_t) \, dy_t$. Ad esempio, se:

$$y_t = \alpha t + \varepsilon_t, \qquad \varepsilon_t \text{ iid}, \ \varepsilon_t \sim N(0, \sigma^2)$$

si ha:

$$\mathbb{E}[y_t] = \alpha t \qquad\qquad \mathbb{V}[y_t] = \mathbb{E}\left[(y_t - \alpha t)^2\right] = \mathbb{E}[\varepsilon^2] = \sigma^2$$

Si può anche considerare la covarianza tra due elementi di un processo stocastico; dal momento che si tratta della covarianza tra $y_t$ e un suo valore precedente (o successivo), viene detta *autocovarianza*.

**Definizione 6.2.** Dato un processo stocastico $y_t$, si dice *funzione di autocovarianza* tra due suoi elementi $y_t$ e $y_{t-j}$, e si indica con $\gamma$, la loro covarianza:

$$\gamma(t, t-j) = \operatorname{Cov}(y_t, y_{t-j}) = \mathbb{E}\Big[(y_t - \mathbb{E}[y_t])(y_{t-j} - \mathbb{E}[y_{t-j}]\Big] = \mathbb{E}[y_t y_{t-j}] - \mathbb{E}[y_t]\mathbb{E}[y_{t-j}]$$

Ad esempio, se:

$$y_t = c + \varepsilon_t, \qquad \varepsilon_t \text{ iid}, \ \varepsilon_t \sim N(0, \sigma^2)$$

si ha:

$$\gamma(t, t-j) \operatorname{Cov}(y_t, y_{t-j}) = \mathbb{E}[(y_t - c)(y_{t-j} - c)] = \mathbb{E}[\varepsilon_t \varepsilon_{t-j}] = 0$$

Si ha ovviamente $\gamma(t, t) = \mathbb{V}[y_t]$. Se l'autocovarianza è funzione solo di $j$, se cioè $\gamma(t, t-j) = \gamma(j)$ per ogni $t$, si definisce un'analoga funzione di autocorrelazione.

**Definizione 6.3.** Dato un processo stocastico $y_t$ con $\gamma(t, t-j) = \gamma(j)$, si dice *funzione di autocorrelazione* tra due suoi elementi, e si indica con $\rho(j)$, la loro correlazione:

$$\rho(j) = \frac{\gamma(t, t-j)}{\sqrt{\gamma(t,t)\gamma(t-j, t-j)}} = \frac{\gamma(j)}{\sqrt{\gamma(0)\gamma(0)}} = \frac{\gamma(j)}{\gamma(0)}$$

### 6.3.1 Persistenza

La covarianza e la correlazione tra due elementi di un processo stocastico costituiscono un indicatori della sua *persistenza* ("memoria").

**Definizione 6.4.** In un processo stocastico $y_t$ si ha *persistenza* se:

$$\mathbb{E}[y_t] \neq \mathbb{E}[y_t \mid \mathcal{F}_{t-1}] \qquad \mathcal{F}_{t-1} = \{y_{t-1}, y_{t-2}, \dots\}$$

$\mathcal{F}_{t-1}$ rappresenta la storia passata di $y_t$:[5] si ha persistenza se il valore atteso di $y_t$ cambia secondo che si conosca o no la sua storia passata. Si ha ovviamente $\mathbb{E}[y_t] = \mathbb{E}[y_t \mid \mathcal{F}_{t-1}]$ se $y_t$ è indipendente dagli $y_{t-j}$ e da loro funzioni.

Gli unici processi stocastici senza traccia di persistenza, infatti, sono le successioni di variabili aleatorie *indipendenti*. A rigore, la mancanza di correlazione implica l'indipendenza solo se le $y_t$ hanno distribuzione normale, ma nella pratica, data l'abbondanza di relazioni lineari, si usa spesso la correlazione come un indicatore della persistenza.

Diventa quindi importante distinguere i processi stocastici sulla base delle correlazioni tra loro distinti elementi.

### 6.3.2 Stazionarietà ed ergodicità

**Definizione 6.5.** Dato un processo stocastico $y_t$, si ha *stazionarietà debole*, o *di covarianza*, se:

- $\mathbb{E}[y_t] = \mu_y < \infty$;
- $\mathbb{V}[y_t] = \sigma_y^2 < \infty$;
- $\operatorname{Cov}(y_t, y_{t-j}) = \gamma(j) < \infty$.

[5] L'affermazione non brilla per rigore (cfr. Lucchetti 2008, pp. 5-6, poi Dall'Aglio 2003, p. 296), ma è sufficiente nel contesto di queste note.

Si ha cioè stazionarietà debole se tutte le variabili aleatorie hanno media, varianza e autocovarianza finite e costanti nel tempo; in particolare se l'autocovarianza è funzione di $j$, non di $t$.

Ad esempio, supponendo come sopra $\varepsilon_t$ iid e $\varepsilon_t \sim N(0, \sigma^2)$, il processo $y = \alpha t + \varepsilon_t$ non è stazionario, in quanto $\mathbb{E}[y_t] = \alpha t$ dipende da $t$. Il processo $y = c + \varepsilon_t$ è invece stazionario in quanto:

$$\mathbb{E}[y_t] = c$$
$$\mathrm{Cov}(y_t, y_{t-j}) = \mathbb{E}[\varepsilon_t \varepsilon_{t-j}] = \begin{cases} \sigma^2 & \text{se } j = 0 \\ 0 & \text{se } j \neq 0 \end{cases}$$

**Definizione 6.6.** Dato un processo stocastico $y_t$, si ha *stazionarietà forte*, o *stretta*, se la distribuzione congiunta di $(y_t, \dots, y_{t-k})$ è indipendente da $t$ per qualsiasi $k$.

In sostanza, un sottoinsieme di $k$ elementi di un processo stocastico è una variabile aleatoria $k$-dimensionale con una sua distribuzione congiunta che potrebbe dipendere da $t$. Si ha stazionarietà forte se, per qualsiasi $k$ e per qualsiasi sottoinsieme di ampiezza $k$, la distribuzione congiunta non dipende da $t$ ma è uguale a quella di un altro sottoinsieme di pari ampiezza i cui indici differiscano di un qualche $j$ da quelli del primo.

**Osservazione 6.7.** La stazionarietà debole non implica quella forte, in quanto considera solo variabili aleatorie doppie, ma nemmeno quella forte implica la debole, in quanto un processo potrebbe essere stazionario in senso forte ma non avere momenti. Tuttavia, se un processo è *gaussiano*, ovvero se la distribuzione congiunta di un qualsiasi sottoinsieme di suoi elementi è una normale multivariata, allora stazionarietà debole e forte coincidono. Data la pervasività dei processi gaussiani nelle applicazioni, si parla comunemente di stazionarietà senza aggettivi, intendendo con essa la stazionarietà debole.

**Definizione 6.8.** Un processo stocastico viene detto *ergodico* se è stazionario in covarianza e se:[6]

$$\sum_{j=0}^{\infty} |\gamma(j)| < \infty \quad \text{che implica} \quad \lim_{j \to \infty} \gamma(j) = 0$$

In sostanza, un processo è ergodico se, quanto più due suoi elementi sono lontani nel tempo, tanto meno sono correlati. Tale aspetto diventa importante non appena si passi dal processo stocastico come variabile aleatoria alle sue realizzazioni. Quando si osserva la realizzazione di un processo stocastico, infatti, si pongono alcuni problemi: si osserva solo un sottoinsieme finito di una realizzazione, non si può sapere né se un altro sottoinsieme presenterebbe le stesse caratteristiche, né se queste sarebbero presenti in altre realizzazioni. Se però un processo è ergodico, allora l'osservazione di una sua realizzazione "abbastanza lunga" è equivalente, ai fini inferenziali, all'osservazioni di diverse sue realizzazioni.[7] Si usano allo scopo i teoremi seguenti.

**Teorema 6.9.** *Se un processo stocastico $y_t$ è stazionario ed ergodico, una sua funzione $x_t = f(y_t, y_{t-1}, \dots)$ è a sua volta stazionaria ed ergodica.*

[6] A rigore questa è solo una versione dell'ergodicità per la media, ma non si può approfondire più di tanto in questa sede.

[7] Va notato che, mentre l'ipotesi di stazionarietà può essere verificata, almeno in alcuni contesti, quella di ergodicità non è verificabile se si dispone di una sola realizzazione di un processo (Lucchetti 2008, p. 5).

**Teorema 6.10** (Teorema ergodico)**.** *Se un processo stocastico $y_t$ è stazionario ed ergodico, con $\mathbb{E}[|y_t|] < \infty$, allora per $T \to \infty$:*

$$\mu = \frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{p} \mathbb{E}[y_t]$$

Da essi segue la possibilità di stime consistenti:

**Teorema 6.11.** *Se un processo stocastico $y_t$ è stazionario ed ergodico, con $\mathbb{E}[y_t^2] < \infty$, allora per $T \to \infty$:*
- $\hat{\mu} \xrightarrow{p} \mathbb{E}[y_t]$;
- $\hat{\gamma}(j) \xrightarrow{p} \gamma(j)$;
- $\hat{\rho}(j) \xrightarrow{p} \rho(j)$.

*Dimostrazione.* Il primo asserto segue direttamente dal teorema ergodico. Quanto al secondo:

$$\begin{aligned}
\hat{\gamma}(j) &= \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\mu})(y_{t-j} - \hat{\mu}) \\
&= \frac{1}{T}\sum_{t=1}^{T} y_t y_{t-j} - \frac{1}{T}\sum_{t=1}^{T} y_t\hat{\mu} - \frac{1}{T}\sum_{t=1}^{T} y_{t-j}\hat{\mu} + \hat{\mu}^2 = \frac{1}{T}\sum_{t=1}^{T} y_t y_{t-j} - 2\hat{\mu}^2 + \hat{\mu}^2 \\
&= \frac{1}{T}\sum_{t=1}^{T} y_t y_{t-j} - \hat{\mu}^2
\end{aligned}$$

La successione $y_t y_{t-j}$ è stazionaria ed ergodica per il teorema 6.9 e ha media finita per l'ipotesi $\mathbb{E}[y_t^2] < \infty$. Per il teorema ergodico:

$$\frac{1}{T}\sum_{t=1}^{T} y_t y_{t-j} \xrightarrow{p} \mathbb{E}[y_t y_{t-j}]$$

Quindi:

$$\hat{\gamma}(j) \xrightarrow{p} \mathbb{E}[y_t y_{t-j}] - \mu^2 = \gamma(j)$$

Il terzo asserto segue dal lemma di Slutsky (v. appendice C). □

Si vedrà nella sezione 7.5 come l'ergodicità venga utilizzata per consentire l'inferenza su processi autoregressivi stazionari.

### 6.3.3 *White noise* e *Random walk*

Il concetto di stazionarietà consente di stabilire un criterio per distinguere processi come quelli rappresentati nella figura 6.2 da altri simili a quelli rappresentati nella figura 6.3.

**Definizione 6.12.** Si dice *white noise* (*rumore bianco*), e si indica con $WN$, un processo stocastico i cui elementi $\varepsilon_t$ sono tali che:

$$\mathbb{E}[\varepsilon_t] = 0 \qquad \mathbb{V}[\varepsilon_t] = \sigma^2 \qquad \mathrm{Cov}(\varepsilon_t, \varepsilon_{t-j}) = 0,\ j \neq 0$$

Se $\varepsilon_t \sim N(0, \sigma^2)$, il processo viene detto *white noise gaussiano*.

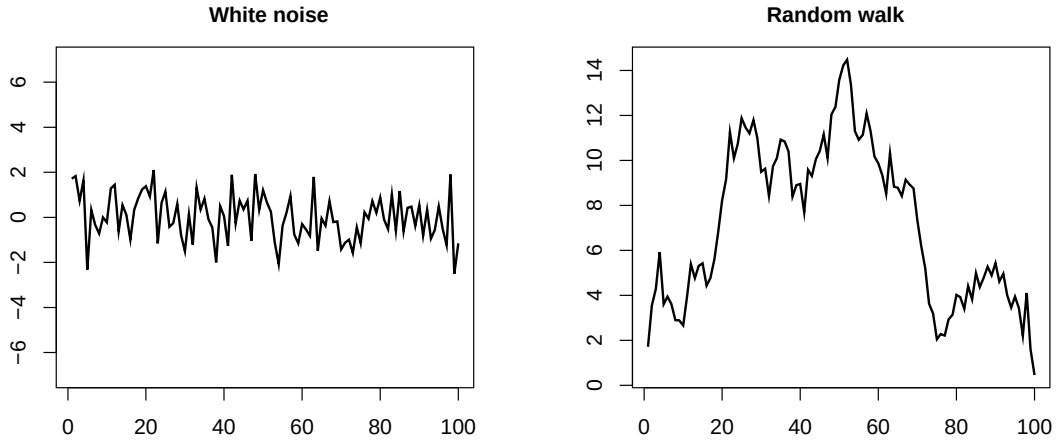


*Figura 6.4.* Esempi di processi *white noise* e *randow walk*.

Dal momento che il valore atteso è costante e finito e l'autocovarianza è nulla, un processo $WN$ è stazionario.

In un processo $WN$ gaussiano, inoltre, la mancanza di correlazione tra i termini comporta anche la loro indipendenza; si tratta quindi di processi senza traccia di persistenza.

**Definizione 6.13.** Si dice *randow walk* (*passeggiata aleatoria*) un processo stocastico del tipo:

$$y_t = y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim WN$$

La presenza di un elemento precedente nella definizione di $y_t$ cambia drasticamente le caratteristiche del processo. Supponendo $y_0 = 0$, si ha:

$$\begin{aligned} y_1 &= 0 + \varepsilon_1 = \varepsilon_1 \\ y_2 &= y_1 + \varepsilon_2 = \varepsilon_1 + \varepsilon_2 \\ &\dots \\ y_T &= \sum_{t=1}^{T} \varepsilon_t \end{aligned}$$

quindi:

$$\mathbb{E}[y_t] = 0 \qquad \mathbb{V}[y_t] = \mathbb{V}\left[\sum_{t=1}^{T} \varepsilon_t\right] = T\sigma^2$$

Essendo la varianza non costante, ma funzione di $t$, un processo *random walk* non è stazionario.

Creare con R processi *white noise* gaussiani e *random walk* è semplice:

```
> n <- 100
> wn <- rnorm(n)
> rw <- cumsum(wn)
```

La loro rappresentazione grafica (figura 6.4) mostra evidenti somiglianze con le figure 6.2 e 6.3. Si spiega così una parte del mistero: una successione di lanci di una moneta è un processo *stazionario*, una passeggiata governata dai lanci è un processo *non stazionario*.

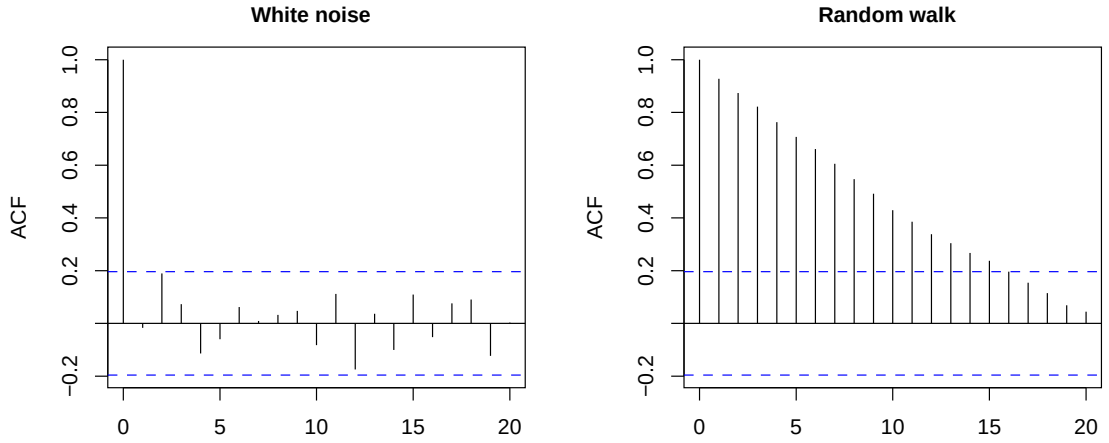


*Figura 6.5.* Autocorrelogrammi di un processo *white noise* e di un processo *random walk*.

Processi stazioniari e non stazionari presentano caratteristiche di persistenza molto diverse. La figura 6.5 è stata generata con la funzione `acf()` di R che, data una serie storica, produce i cosiddetti *autocorrelogrammi*, grafici che rappresentato l'autocorrelazione $\rho$ per diversi valori del ritardo $j$.

Come si vede, in entrambi i casi l'autocorrelazione è ovviamente pari a 1 per $j = 0$, ma poi nel *white noise* si riduce immediatamente per $j > 0$ ed oscilla con valori pressoché trascurabili entro una banda molto ristretta, mentre rimane elevata per molto tempo e decresce piuttosto lentamente nel *random walk*.

Ma che dire riguardo a matrimoni religiosi e mortalità?

### 6.3.4 Cointegrazione

Un processo stocastico stazionario viene anche detto *integrato di ordine* 0, $I(0)$.

Si è visto sopra che, calcolando le differenze tra i singoli passi di una passeggiata aleatoria, si giunge ad un processo stazionario. Formalmente, dato un *random walk*, intrinsecamente non stazionario, basta poco per costruire un processo stazionario:

$$y_t = y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim WN \qquad\qquad y_t - y_{t-1} = \varepsilon_t \sim WN$$

Se un processo è tale che la successione delle differenze tra un elemento e il precedente è stazionaria, il processo viene detto *integrato di ordine* 1, $I(1)$. È tale un processo *random walk*.

Dati due processi stocastici $I(1)$, potrebbe accadere che una loro combinazione lineare sia stazionaria. In tal caso, i due processi vengono detti *cointegrati*.

Ecco la soluzione del paradosso: due processi stocastici non stazionari potrebbero mostrare andamenti relativamente simili pur essendo totalmente indipendenti (come due *random walk*) e una regressione potrebbe far pensare a correlazioni in realtà inesistenti; se tuttavia i processi sono cointegrati, ma solo in questo caso, allora condividono un trend comune e un'analisi di regressione ha senso.

Ad esempio, se si riprendono i processi `p1` e `p2` (figura 6.2), un test di cointegrazione mostrerebbe che *non sono* cointegrati:[8]

```
> po.test(cbind(p1,p2))

Phillips-Ouliaris Cointegration Test

data:  cbind(p1, p2)
Phillips-Ouliaris demeaned = -7.5615, Truncation lag parameter = 0,
p-value = 0.15

Warning message:
In po.test(cbind(p1, p2)) : p-value greater than printed p-value
```

L'ipotesi nulla è la non cointegrazione, che non può essere rifiutata. Ne segue che qualsiasi regressione sarebbe *spuria*.

Eseguendo lo stesso test sulle serie dei matrimoni religiosi e della mortalità si otterrebbe un risultato diverso, quasi a pensare che, in quel caso, la *nonsense correlation* di Yule non fosse tanto *nonsense*, ma questo è tutto un altro discorso...

[8] La funzione `po.test()` è contenuta nella libreria `tseries`. Le statistiche test dei processi stocastici presentano spesso distribuzioni atipiche e i valori del *p-value* vengono quindi calcolati per interpolazione da tabelle; è questo il motivo per cui compare il messaggio che avverte che il *p-value* "vero" è maggiore di quello mosrato.

# Capitolo 7

# I processi ARMA

In questo capitolo si illustra in primo luogo l'*operatore ritardo* $L$, che viene definito con riferimento alla generica serie storica. L'operatore ritardo viene poi usato per definire i processi stocastici MA (*Moving Average*), AR (*AutoRegressive*) e ARMA (loro generalizzazione). Si esaminano le condizioni di stazionarietà di tali processi e si conclude con la possibilità di usare gli abituali stimatori e test di ipotesi in caso di stazionarietà.

In tutto il capitolo si intende $\varepsilon_t \sim WN$.

## 7.1 L: l'operatore ritardo

Un *operatore* su serie storiche è una funzione che trasforma una o più serie storiche in un'altra. Vi sono operatori su serie storiche molto simili a operatori familiari come la somma e il prodotto; ad esempio:

$$y_t = x_t + w_t \qquad\qquad y_t = \alpha x_t$$

Nel primo caso si definisce una serie storica il cui valore al tempo $t$ non è altro che la somma dei valori allo stesso tempo $t$ di $x$ e di $w$, nel secondo $y_t$ è il prodotto di una costante $\alpha$ per il valore che $x$ assume al tempo $t$. L'unica differenza rispetto alla somma e al prodotto di scalari è che qui si tratta di successioni infinite di somme e prodotti.

Risulta particolarmente utile l'*operatore ritardo*, indicato con $L$ (*Lag*), che trasforma una serie storica $\{x_t\}$ in un'altra $\{y_t\}$ tale che il valore di $y$ al tempo $t$ sia uguale a quello di $x$ al tempo $t-1$:

$$y_t = Lx_t = x_{t-1}$$

L'operatore ritardo è *lineare*, ovvero *additivo* e *omogeneo* (di grado 1):

$$L(x_t + w_t) = Lx_t + Lw_t \qquad\qquad L(\alpha x_t) = \alpha L x_t$$

L'operatore ritardo può essere applicato più volte; in questi casi, si usa indicare con un esponente il numero delle iterazioni. Ad esempio:

$$L^3 x_t = L(L(Lx_t)) = L(Lx_{t-1}) = Lx_{t-2} = x_{t-3}$$

In generale:

$$L^k x_t = x_{t-k}$$

Grazie a questa notazione, è possibile definire *polinomi* in $L$:

$$(1 + aL + bL^2 + cL^3)x_t = x_t + ax_{t-1} + bx_{t-2} + cx_{t-3}$$

## 7.2 MA: processi a media mobile

Un processo *a media mobile di ordine* $q$ è una successione di variabili aleatorie del tipo:

$$y_t = \varepsilon_t + c_1\varepsilon_{t-1} + \cdots + c_q\varepsilon_{t-q} = \sum_{n=0}^{q} c_n\varepsilon_{t-n}$$

e si indica con $MA(q)$. Si ha una *media mobile finita* se $q < \infty$, altrimenti una *media mobile infinita*.

Il processo viene detto "a media mobile di ordine $q$" perché $y_t$ è la somma di $\varepsilon_t$ e di una media ponderata dei $q$ valori precedenti più vicini di $\varepsilon$, media che cambia al variare di $t$. L'operatore ritardo consente una definizione più sintetica:

$$y_t = C(L)\varepsilon_t$$

dove $C(L) = 1 + c_1L + c_2L^2 + \cdots + c_qL^q$ è un polinomio di grado $q$ nell'operatore ritardo.

### 7.2.1 Medie mobili finite

Il valore atteso, la varianza e l'autocovarianza sono:

$$\mathbb{E}[y_t] = \mathbb{E}\left[\sum_{n=0}^{q} c_n\varepsilon_{t-n}\right] = \sum_{n=0}^{q} c_n\mathbb{E}[\varepsilon_{t-n}] = 0$$

$$\mathbb{V}[y_t] = \mathbb{E}[y_t^2] = \mathbb{E}\left[\left(\sum_{n=0}^{q} c_n\varepsilon_{t-n}\right)^2\right]$$

il quadrato di un polinomio è una somma di quadrati e di prodotti di coppie di termini diversi; essendo $\varepsilon_t \sim WN$, il valore attesso dei prodotti di termini diversi è nullo:

$$\mathbb{V}[y_t] = \mathbb{E}\left[\sum_{n=0}^{q} c_n^2\varepsilon_{t-n}^2\right] = \sum_{n=0}^{q} c_n^2\mathbb{E}[\varepsilon_{t-n}^2] = \sigma^2\sum_{n=0}^{q} c_n^2$$

$$\begin{aligned}\operatorname{Cov}(y_t, y_{t-j}) &= \mathbb{E}\left[\left(\sum_{m=0}^{q} c_m\varepsilon_{t-m}\right)\left(\sum_{n=0}^{q} c_n\varepsilon_{t-n-j}\right)\right] = \mathbb{E}\left[\sum_{m=0}^{q} c_m\left(\sum_{m=0}^{q} c_m\varepsilon_{t-m}\varepsilon_{t-n-j}\right)\right] \\ &= \sum_{m=0}^{q} c_m\left(\sum_{n=0}^{q} c_n\mathbb{E}[\varepsilon_{t-m}\varepsilon_{t-n-j}]\right)\end{aligned}$$

per le proprietà del $WN$, sono non nulli solo i termini in cui $t-m = t-n-j$, $m = n+j$, quindi:

$$\gamma(j) = \operatorname{Cov}(y_t, y_{t-j}) = \sigma^2\sum_{n=0}^{q} c_nc_{n+j}$$

Vale quindi il seguente teorema.

**Teorema 7.1.** *Un processo a media mobile finita $MA(q)$, $q < \infty$, è stazionario ed ergodico.*

*Dimostrazione.* Il valore atteso è nullo, la varianza e l'autocovarianza non dipendono da $t$ e sono sempre finite. Ciò dimostra la stazionarietà.

Quanto all'ergodicità, basta osservare che $\gamma(j) = 0$ non appena sia $j > q$. □

### 7.2.2 Medie mobili infinite

Per dimostrare le proprietà di una media mobile infinita occorre fissare alcuni risultati preliminari, poi avvalersi della rappresentazione del processo mediante l'operatore ritardo:

$$y_t = C(L)\varepsilon_t, \qquad C(L) = \sum_{n=0}^{\infty} c_n L^n$$

**Lemma 7.2.** *Se* $f(z) = \sum_{n=0}^{\infty} c_n z^n < \infty$ *per* $z \in D(0,r)$*, allora:*

$$\sum_{n=0}^{\infty} |c_n z^n| < \infty, \qquad z \in D(0,s),\ 0 < s < r$$

**Lemma 7.3.** *Una successione sommabile in valore assoluto è sommabile al quadrato (ma non viceversa):*

$$\sum_{n=0}^{\infty} |a_n| < \infty \quad \Rightarrow \quad \sum_{n=0}^{\infty} a_n^2 < \infty$$

**Teorema 7.4.** *Un processo a media mobile infinita* $MA(\infty)$ *è stazionario ed ergodico se la serie* $\sum_{n=0}^{\infty} c_n$ *converge.*

*Dimostrazione.* Nella dimostrazione del teorema precedente si è già visto che $\mathbb{E}[y_t] = 0$. Quanto a varianza e autocovarianza, ci si può limitare a considerare quest'ultima come caso più generale:

$$\mathrm{Cov}(y_t, y_{t-j}) = \sigma^2 \sum_{i=0}^{\infty} c_i c_{i-j}$$

Le proprietà algebriche di $C(L)$ possono essere esaminate sostituendo l'operatore $L$ con $z \in \mathbb{C}$, ottenendo la seguente serie di potenze:

$$f(z) = \sum_{n=0}^{\infty} c_n z^n < \infty, \qquad z \in D(0,r)$$

$D(0,r)$ è un disco di centro 0 e raggio $r$ e $r$ è il raggio di convergenza della serie (se $r = 0$ la serie non converge per $z \neq 0$).

Se e solo se $r > 1$, converge anche $f(1) = \sum_{n=0}^{\infty} c_n$. Inoltre, per il lemma 7.2 converge anche $\sum_{n=0}^{\infty} |c_n|$ su un disco $D(0,s)$, $0 < s < r$, e per il lemma 7.3 converge anche $\sum_{n=0}^{\infty} a_n^2$. Converge ovviamente anche una serie che non parta dal primo termine, come $\sum_{n=j}^{\infty} a_n^2$.

Ne segue, applicando la disuguaglianza di Cauchy-Schwarz:

$$\gamma(j) = \mathrm{Cov}(y_t, y_{t-j}) = \sigma^2 \sum_{n=j}^{\infty} c_n c_{n-j} \leq \sigma^2 \sqrt{\sum_{n=j}^{\infty} c_n^2 \sum_{n=j}^{\infty} c_{n-j}^2} < \infty$$

Questo dimostra la stazionarietà.

Quanto all'ergodicità, da

$$\gamma(j) = \sigma^2 \sum_{n=0}^{\infty} c_n c_{n+j}$$

segue, per le proprietà del valore assoluto:

$$|\gamma(j)| = \sigma^2 \left| \sum_{n=0}^{\infty} c_n c_{n+j} \right|$$

poi, per la disuguaglianza triangolare:

$$|\gamma(j)| \leq \sigma^2 \sum_{n=0}^{\infty} |c_n c_{n+j}|$$

$$\sum_{j=0}^{\infty} |\gamma(j)| \leq \sigma^2 \sum_{j=0}^{\infty} \sum_{n=0}^{\infty} |c_n c_{n+j}| = \sigma^2 \sum_{j=0}^{\infty} \sum_{n=0}^{\infty} |c_n| \cdot |c_{n+j}| = \sigma^2 \sum_{n=0}^{\infty} |c_n| \sum_{j=0}^{\infty} |c_{n+j}|$$

Per quanto già visto, le due serie a termini positivi convergono, quindi:

$$\sum_{j=0}^{\infty} |\gamma(j)| < \infty \quad \Rightarrow \quad \sum_{j=0}^{\infty} \gamma(j) < \infty \quad \Rightarrow \quad \gamma(j) \to 0 \qquad \square$$

## 7.3 AR: processi autoregressivi

Un processo *autoregressivo* è una successione di variabili aleatorie ciascuna delle quali è funzione delle precedenti:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \varepsilon_t$$

e si indica con $AR(p)$.

Nella definizione 6.4 si era indicata con $\mathcal{F}_{t-1}$ la storia passata di $y_t$: $y_{t-1}, y_{t-2}, \ldots$ Un processo $AR(p)$ è un modello di processi stocastici in cui risulta rilevante solo uno spezzone finito della storia passata, in quanto si intende $p < \infty$.

Processi di questo tipo vengono detti "autoregressivi" perché somigliano molto a modelli di regressione in cui le variabili esplicative sono il passato della variabile risposta, cui si aggiunge un errore $WN$ simile all'errore come inteso nel modello lineare normale (sez. 2.4.4).

Anche in questo caso l'operatore ritardo consente una definizione più sintetica. Da:

$$y_t - a_1 y_{t-1} - a_2 y_{t-2} - \cdots - a_p y_{t-p} = \varepsilon_t$$

si giunge a:

$$A(L) y_t = \varepsilon_t \qquad A(0) = 1$$

dove $A(L)$ è un polinomio di grado $p$ nell'operatore ritardo. Tale rappresentazione consente di indagare le proprietà del processo; a tale scopo, analogamente a quanto fatto nella dimostrazione del teorema 7.4, si studiano le proprietà algebriche del corrispondente polinomio $A(z)$, detto *polinomio caratteristico*.

**Teorema 7.5.** *Un processo autoregressivo $AR(p)$ è stazionario ed ergodico se e solo se le radici del polinomio caratteristico sono tutte fuori del cerchio unitario. In questo caso, il processo ammette la rappresentazione $MA(\infty)$:*

$$y_t = \sum_{n=0} c_n e_{t-n}$$

*dove i $c_n$ sono i coefficienti dell'espansione in serie di Taylor di $A(z)^{-1}$ intorno a zero.*

*Dimostrazione.* Da $A(L)y_t = \varepsilon_t$ si ricava:

$$y_t = A(L)^{-1}\varepsilon_t$$

Sostituendo l'operatore $L$ con $z \in \mathbb{C}$, per il Teorema Fondamentale dell'Algebra si ha:

$$A(z) = 1 - \sum_{n=1}^{p} a_n z^n = \prod_{n=1}^{l} \left(1 - \frac{z}{z_n}\right)^{m_n}$$

dove $z_n$ è una radice e $m_n$ la sua molteplicità algebrica (si può dividere per $z_n$ in quanto 0 non è una radice). Quanto al reciproco:

$$C(z) = A(z)^{-1} = \frac{1}{\prod_{n=1}^{l} \left(1 - \frac{z}{z_n}\right)^{m_n}}$$

si vede che le radici $z_n$ sono punti di singolarità:

$$\lim_{|z - z_n| \to 0} C(z) = \infty$$

Espandendo $C(z)$ in serie di Taylor intorno a 0, si ha:

$$C(z) = \sum_{n=0}^{\infty} \frac{C^{(n)}(0)}{n!} z^n = \sum_{n=0}^{\infty} c_n z^n$$

Ma questa è la stessa serie esaminata nella dimostrazione del teorema 7.4. Ne segue che, perché si abbia stazionarietà, il raggio di convergenza $r$ deve essere maggiore di 1. Il raggio di convergenza, peraltro, giunge fino alla prima singolarità: $r = \min_n |z_n|$, quindi le radici devono essere fuori del cerchio unitario. In tal caso, $|z| < |z_n|$, il processo ammette una rappresentazione $MA(\infty)$ stazionaria ed ergodica. □

### 7.3.1 Processi AR(1)

Se $p = 1$ il processo diventa:

$$y_t = a y_{t-1} + \varepsilon_t \quad \Rightarrow \quad A(L)y_t = (1 - a)y_t = \varepsilon_t$$

Partendo da un $y_0 = 0$ si avrebbe:

$$\begin{aligned} y_1 &= \varepsilon_1 \\ y_2 &= a y_1 + \varepsilon_2 = a\varepsilon_1 + \varepsilon_2 \\ y_3 &= a y_2 + \varepsilon_3 = a^2\varepsilon_1 + a\varepsilon_2 + \varepsilon_3 \\ &\dots \\ y_T &= \sum_{j=0}^{T} a^j \varepsilon_{T-j} \end{aligned}$$

In generale, $y_t = \sum_{j=0}^{\infty} a^j \varepsilon_{t-j}$.

Se il processo è stazionario, la sua rappresentazione $MA(\infty)$ è:

$$y_t = \sum_{n=0}^{\infty} c_n e_{t-n}, \qquad c_n = a^n$$

Il valore atteso è nullo. La varianza è:

$$\mathbb{V}[y_t] = \sigma^2 \sum_{n=0}^{\infty} c_n^2 = \sigma^2 \sum_{n=0}^{\infty} a^{2n} = \frac{\sigma^2}{1-a^2}$$

Infatti, l'unica radice del polinomio $1 - az$ è $z_1 = \frac{1}{a}$ e $|z_1| > 1$ comporta $|a| < 1$; in tal caso $= \sum_{n=0}^{\infty} a^{2n}$ è una serie geometrica convergente a $\frac{1}{1-a^2}$.

Quanto all'autocovarianza:[1]

$$\gamma(j) = \sigma^2 \sum_{n=j}^{\infty} c_n c_{n-j} = \sigma^2 \sum_{n=0}^{\infty} c_n c_{n+j} = \sigma^2 \sum_{n=0}^{\infty} a^n a^{n+j} = \sigma^2 \sum_{n=0}^{\infty} a^{2n+j}$$
$$= a^j \sigma^2 \sum_{n=0}^{\infty} a^{2n} = a^j \frac{\sigma^2}{1-a^2}$$

Infine, l'autocorrelazione:

$$\rho(j) = \frac{\gamma(j)}{\gamma(0)} = a^j$$

indica una persistenza tanto maggiore quanto maggiore è $|a|$, ma, poiché $|a| < 1$, decresce all'aumentare di $j$: un processo autoregressivo stazionario ha sì memoria infinita, ma il passato remoto gioca un ruolo di fatto irrilevante.

Al contrario, se fosse $|a| = 1$ il processo non avrebbe varianza costante ($y_t = y_{t-1} + \varepsilon_t$ è un *random walk*, in cui la varianza aumenta col tempo) ed "esploderebbe" se $|a| > 1$ (figura 7.1).

### 7.3.2 Processi AR(p)

Un aspetto interessante dei processi autoregressivi con $p > 1$ risiede nel fatto che tra le radici del polinonio caratteristico potrebbero esservi coppie di radici complesse coniugate; in questo caso, il processo assume un andamento ciclico (v. poi esempio 7.6).

Per il resto, le autocovarianze di un processo autoregressivo stazionario di ordine $p$:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \varepsilon_t$$

possono essere ricavate moltiplicando entrambi i membri per $y_{t-j}$ e prendendo i valori attesi:

$$\mathbb{E}[y_t y_{t-j}] = a_1 \mathbb{E}[y_{t-1} y_{t-j}] + \cdots + a_p \mathbb{E}[y_{t-p} y_{t-j}] + \mathbb{E}[\varepsilon_t \varepsilon_{t-j}]$$
$$\gamma(j) = \begin{cases} a_1\gamma(1) + \cdots + a_p\gamma(p) + \sigma^2 & \text{per } j = 0 \\ a_1\gamma(j-1) + \cdots + a_p\gamma(j-p) & \text{per } j = 1, 2, \ldots \end{cases}$$

[1] Nello sviluppo si usa un'espressione della covarianzia tale che $a$ risulti elevato ad un esponente $j$ inteso come non negativo (dati $y_t$ e $y_s$, $j = |t - s|$). Si può anche lasciare che $j$ "sembri" negativo, ma poi scrivere, come fanno alcuni testi, $a^{|j|}$.

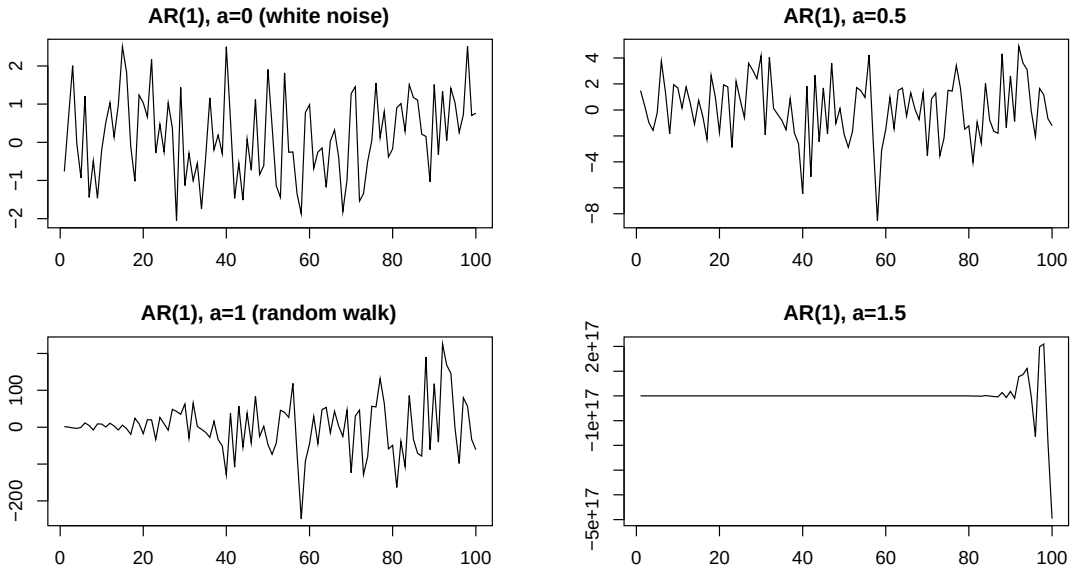


*Figura 7.1.* Andamento di processi $AR(1)$ per diversi valori del parametro $a$.

Dividendo per $\gamma(0)$ si ottengono le *equazioni di Yule-Walker* per le autocorrelazioni:

$$\rho(j) = a_1\rho(j-1) + \cdots + a_p\rho(j-p)$$

Le autocovarianze e le autocorrelazioni hanno quindi la forma di equazioni alle differenze (cfr. appendice B). Si può mostrare che il vettore:

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma(0) \\ \gamma(1) \\ \vdots \\ \gamma(p-1) \end{bmatrix}$$

è uguale ai primi $p$ elementi della prima colonna della matrice:

$$\sigma^2[\mathbf{I}_{p^2} - \mathbf{F} \otimes \mathbf{F}]^{-1}$$

dove $\mathbf{F}$ è la matrice definita nella sez. B.2 e $\otimes$ è il prodotto di Kronecker.[2]

**Esempio 7.6.** Nel processo $AR(2)$

$$y_t = 1.8y_{t-1} - 0.9y_{t-2} + \varepsilon_t$$

[2]Date due matrici $\underset{m,n}{\mathbf{A}}$ e $\underset{p,q}{\mathbf{B}}$, il loro prodotto di Kronecker è la matrice:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}$$

In R il prodotto di Kronecher si esegue con `A %x% B`.

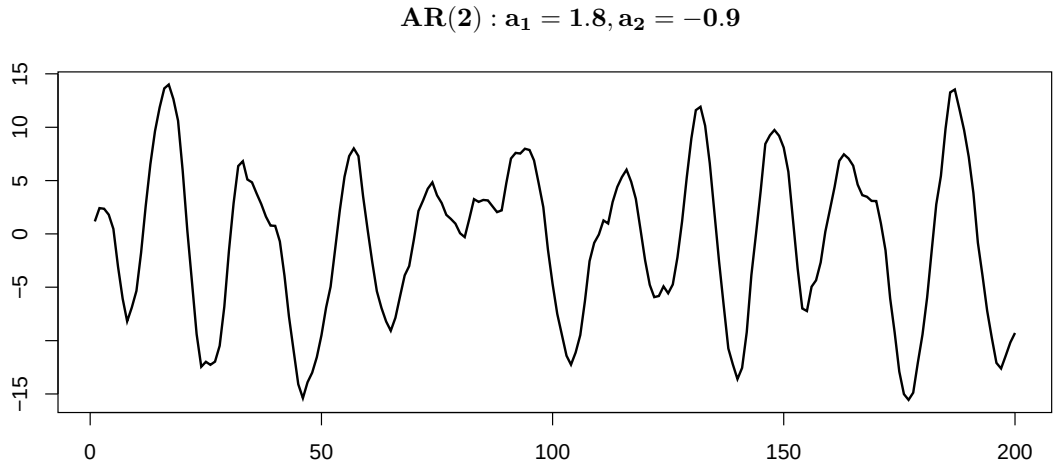


*Figura 7.2.* Un modello $AR(2)$ con coppia di radici coniugate del polinomio caratteristico mostra un andamento cicliclo.

le radici del polinomio caratteristico $1-1.8z+0.9z^2$ sono $z_{1,2} = 1 \pm \frac{i}{3}$, entrambe maggiori di 1 in valore assoluto.[3] La figura 7.2 mostra l'andamento ciclico del processo. Quanto alle autocovarianze, operando con R come nella figura 7.3 si ottiene:

$$\gamma(0) = \mathbb{V}[y_t] = 51.35 \qquad \gamma(1) = 48.65$$

[3] Se $z \in \mathbb{C}$, $|z| = |a+bi| = \sqrt{a^2+b^2}$. Quindi:

$$|1 + i/3| = \sqrt{1 + 1/9} = 1.054$$
$$|1 - i/3| = \sqrt{1 + 1/9} = 1.054$$

```
> a1 <- 1.8
> a2 <- -0.9
> F <- matrix(c(a1, a2, 1, 0), nrow=2, byrow=TRUE)
> FkF <- F %x% F
> I <- diag(4)
> M <- solve(I-FkF)
> M
         [,1]      [,2]      [,3]     [,4]
[1,] 51.35135 -43.78378 -43.78378 41.59459
[2,] 48.64865 -36.21622 -46.21622 39.40541
[3,] 48.64865 -46.21622 -36.21622 39.40541
[4,] 51.35135 -43.78378 -43.78378 42.59459
```

*Figura 7.3.* Calcolo delle autocovarianze del processo $y_t = 1.8y_{t-1} - 0.9y_{t-2} + \varepsilon_t$.

## 7.4 ARMA: una generalizzazione

La classe dei processi ARMA comprende sia i processi AR che i processi MA come casi particolari. Un processo $ARMA(p, q)$ è definito da:

$$\begin{aligned} y_t &= a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t + c_1 \varepsilon_{t-1} + \cdots + c_q \varepsilon_{t-q} \\ y_t - a_1 y_{t-1} - \cdots - a_p y_{t-p} &= \varepsilon_t + c_1 \varepsilon_{t-1} + \cdots + c_q \varepsilon_{t-q} \\ (1 - a_1 L - \cdots - a_p L^p) y_t &= (1 + c_1 L + \cdots + c_q L^q) \varepsilon_t \\ A(L) y_t &= C(L) \varepsilon_t \end{aligned}$$

Dato che qualsiasi processo $MA(q)$ è stazionario ed ergodico per $q$ finito, le proprietà di un modello $ARMA(p, q)$ dipendono solo dalla sua componente autoregressiva.

## 7.5 Inferenza

Tradizionalmente l'analisi delle serie storiche si basava sulla individuazione di tre componenti: *trend* (un andamento di fondo espresso spesso con una funzione polinomiale di grado non troppo elevato), *ciclo* (oscillazioni regolari di lungo periodo) e *stagionalità* (oscillazioni regolari di breve periodo). Ad esse si aggiungeva la consueta componente accidentale (l'errore).

Nei termini dell'approccio descritto in questo capitolo il trend è chiaramente non stazionario, mentre ciclo e stagionalità possono essere considerati componenti stazionarie in quanto oscillazioni a media 0.

Nel caso di *trend deterministico*, ad esempio $T(t) = \beta_1 + \beta_2 t$, processi del tipo $y_t = T(t) + u_t$, con $u_t$ stazionario a media 0, vendono detti *processi TS* (*Trend-Stationary*). In essi compaiono oscillazioni che tendono a smorzarsi sul trend di lungo periodo.

Viene invece detto *trend stocastico* un processo in cui compaia una componente non stazionaria, eventualmente accompagnata da altre componenti stazionarie. Può anche accadere che, pur non essendo stazionario $y_t = T(t) + u_t$, sia stazionaria la serie delle differenze prime $\Delta y_t$; in questo caso il processo viene detto $DS$ (*Difference-Stationary*). In ogni caso, se il trend è stocastico eventuali shock producono oscillazioni persistenti che non vengono riassorbite.

Quando si ha a che fare con un processo stazionario, i suoi parametri possono essere stimati via OLS. In particolare, nel caso di un processo $AR(p)$ stazionario, si può muovere dal modello:

$$y_t = \mathbf{x}'\boldsymbol{\beta} + \varepsilon_t$$

dove:

$$\mathbf{x} = \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

Disponendo di $T$ osservazioni, la relativa equazione diventa:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

dove $\mathbf{X}$ è una matrice $T \times p$ ciascuna riga $\mathbf{x}_t$ della quale contiene i valori osservati di $y_{t-j}$, $j = 1, \ldots, p$, per diversi valori di $t = 1, \ldots, T$, e lo stimatore OLS è:

$$\mathbf{b}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Le variabili in gioco non sono indipendenti, ma la consistenza e la normalità asintotica dello stimatore valgono in virtù di teoremi analoghi alla legge dei grandi numeri e al teorema del limite centrale per variabili iid.

**Teorema 7.7.** *Se $y_t$ è processo stazionario con media $\mu$ e autocovarianza $\gamma(j)$, la media campionaria $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$ soddisfa le seguenti proprietà:*[4]

*a)* $\bar{y} \xrightarrow{p} \mu$;

*b)* $\lim_{T\to\infty} \left(T \cdot \mathbb{E}[(\bar{y} - \mu)^2)]\right) = \sum_{j=-\infty}^{+\infty} \gamma(j)$.

Come si vede, il primo asserto coincide col teorema ergodico come sopra formulato (pag. 59).

**Teorema 7.8** (Anderson)**.** *Sia*

$$y_t = \mu + \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$$

*Se $\{\varepsilon_t\}$ è una successione di variabili aleatorie iid con varianza finita e se $\sum_{j=0}^{\infty} |a_j| < \infty$, allora:*

$$\sqrt{T}(\bar{y} - \mu) \xrightarrow{d} N\left(0, \sum_{j=-\infty}^{+\infty} \gamma_j\right)$$

### 7.5.1 Consistenza e normalità asintotica

**Teorema 7.9.** *Se il processo AR(p) $y_t = \sum_{n=1}^{p} a_n y_{t-n} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma^2)$, è stazionario ed ergodico, lo stimatore OLS*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \qquad \begin{cases} \mathbf{X} = \begin{bmatrix} y_{1,t-1} & \cdots & y_{1,t-p} \\ \cdots & \cdots & \cdots \\ y_{T,t-1} & \cdots & y_{T,t-p} \end{bmatrix} \\ \mathbf{y} = (y_1, \ldots, y_T) \end{cases}$$

*è consistente:*

$$\mathbf{b} \xrightarrow{p} \boldsymbol{\beta}, \qquad \boldsymbol{\beta} = (a_1, \ldots, a_p)$$

[4] A rigore, per la media campionaria vale la proprietà più forte della convergenza in media quadratica:

$$\forall \varepsilon > 0, \exists N : \mathbb{E}[(\bar{y}_T - \mu)^2] < \varepsilon \; \forall t \geq$$

*Dimostrazione.* Lo stimatore OLS può essere scritto nella forma:

$$\mathbf{b} = \boldsymbol{\beta} + \left( n^{-1} \sum_{t=1}^{T} x_t x_t' \right)^{-1} \left( n^{-1} \sum_{t=1}^{T} x_i u_t \right)$$

Una matrice $x_t x_t'$ è costituita dai seguenti elementi:

$$x_t x_t' = \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{bmatrix} \begin{bmatrix} y_{t-1} & y_{t-2} & \dots & y_{t-p} \end{bmatrix} = \begin{bmatrix} y_{t-1}^2 & y_{t-1}y_{t-2} & \dots & y_{t-1}y_{t-p} \\ y_{t-2}y_{t-1} & y_{t-2}^2 & \dots & y_{t-2}y_{t-p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{t-p}y_{t-1} & y_{t-p}y_{t-2} & \dots & y_{t-p}^2 \end{bmatrix}$$

e il suo valore atteso è, per la stazionarietà:

$$\mathbb{E}[\mathbf{x}_t \mathbf{x}_t'] = \begin{bmatrix} \mathbb{E}[y_{t-1}^2] & \mathbb{E}[y_{t-1}y_{t-2}] & \dots & \mathbb{E}[y_{t-1}y_{t-p}] \\ \mathbb{E}[y_{t-2}y_{t-1}] & \mathbb{E}[y_{t-2}^2] & \dots & \mathbb{E}[y_{t-2}y_{t-p}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[y_{t-p}y_{t-1}] & \mathbb{E}[y_{t-p}y_{t-2}] & \dots & \mathbb{E}[y_{t-p}^2] \end{bmatrix} < \infty$$

Inoltre, $\mathbb{E}[x_t \varepsilon_t] = 0$.

Quindi, per il teorema 7.7 e per il lemma di Slutsky (cfr. teorema 3.3):

$$\mathbf{b} \xrightarrow{p} \boldsymbol{\beta} + \mathbb{E}[\mathbf{x}\mathbf{x}']^{-1} \mathbb{E}[\mathbf{x}\mathbf{u}] = \boldsymbol{\beta}$$ □

**Teorema 7.10.** *Se il processo AR(p)* $y_t = \sum_{n=1}^{p} a_n y_{t-n} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma^2)$, *è stazionario ed ergodico, lo stimatore OLS*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

*è asintoticamente normale:*

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$$

*Dimostrazione.* Il teorema 7.8 mostra che la finitezza di $\mathbf{V}$ dipende dalla somma $\sum_{j=-\infty}^{+\infty} \gamma_j$, che è in effetti finita se il processo è stazionario ed ergodico, cioè se $\gamma(j) < \infty$ per ogni $j$ e $\sum_{j=0}^{\infty} |\gamma(j)| < \infty$. In questo caso, quindi, si può procedere analogamente a quanto visto nella dimostrazione del teorema 3.4. □

Se un processo non è stazionario, invece, consistenza e normalità asintotica non valgono più. In un processo *random walk*, ad esempio, l'unico parametro da stimare è il numero 1 e non vale più la "lenta" convergenza $\sqrt{T}(b_T - 1) \xrightarrow{d} N(0, V)$, ma si ha:

$$T(b_T - 1) \xrightarrow{p} 0$$

ovvero $b_T$ converge a 1 più rapidamente che nei processi stazionali (è *superconsistente*). Si dimostra, inoltre, che $T(b_T - \beta)$ converge ad una distribuzione definita in termini di integrali di moti browniani e i cui quantili sono stati calcolati attraverso simulazioni numeriche, la prima volta da Dickey e Fuller; viene quindi detta *distribuzione DF*.

È quindi necessario poter stabilire se processo è stazionario o non stazionario, in particolare se è $I(0)$ o $I(1)$.

### 7.5.2 Test di radice unitaria

Si tratta di test che usano come ipotesi nulla la non stazionarietà. In prima approssimazione, tali test si basano sulle semplici relazioni:

$$\begin{aligned} y_t &= a y_{t-1} + \varepsilon_t \\ y_t - y_{t-1} &= a y_{t-1} - y_{t-1} + \varepsilon_t \\ \Delta y_t &= \phi y_{t-1} + \varepsilon_t, \quad \phi = a - 1 \end{aligned}$$

Eseguita una regressione, si tratta di sottoporre a verifica l'ipotesi nulla "$y_t$ ha una radice unitaria", ovvero $\phi = 0$ ($a = 1$, il processo è un *random walk*).

La statistica test $\dfrac{\hat{\phi}}{\sqrt{\widehat{\mathbb{V}[\hat{\phi}]}}}$, tuttavia, non è distribuita né come una $t$ di Student, come nel modello lineare normale, né è asintoticamente normale, come accade negli altri modelli visti finora. Ai fini dei test, infatti, rileva la distribuzione sotto ipotesi nulla, che è la distribuzione DF cui tendono gli stimatori. Tali test vengono quindi detti *test DF*.

Si deve aggiungere che un processo $I(1)$ potrebbe non essere un *random walk*; potrebbe infatti presentare, al posto del *white noise* $\varepsilon_t$, un processo $AR(p)$ con persistenza di breve periodo. I test *ADF* (*Augmented Dickey-Fuller*) e *PP* (da Phillips e Perron, che lo hanno proposto) tengono conto di tale possibilità, facendo in modo che la distribuzione del test non risenta della memoria di breve periodo.[5]

### 7.5.3 Test di stazionarietà

Altri test seguono l'approccio inverso, scegliendo la stazionarità come ipotesi nulla. Il più noto è il *test KPSS* (da Kwiatkowski, Phillips, Schmidt e Shini), la cui idea di fondo è supporre un processo trend-stazionario, effettuare una regressione e verificare se i residui sono $I(0)$.[6]

Va notato che i test di radice unitaria e quelli di stazionarietà danno spesso, ma non sempre, risultati coerenti.

[5] La libreria `tseries` di R contiene le funzioni `adf.test()` e `pp.test()`.

[6] Con R si può usare la funzione `kpss.test()`, contenuta nella libreria `tseries`:

```
> wn <- rnorm(1000)  # white noise
> kpss.test(wn)
[...]
KPSS Level = 0.1015, Truncation lag parameter = 7, p-value = 0.1

Warning message:
In kpss.test(wn) : p-value greater than printed p-value
> rw <- cumsum(wn)   # random walk
> kpss.test(rw)
[...]
KPSS Level = 11.4318, Truncation lag parameter = 7, p-value = 0.01

Warning message:
In kpss.test(rw) : p-value smaller than printed p-value
```

### 7.5.4 La scomposizione di Beveridge-Nelson

Si è appena sottolineato che un *random walk* è solo un caso particolare di processo $I(1)$; possono esservi processi del tipo $y_t = x_t + u_t$ dove $x_t$ è un *random walk* ma $u_t$ non è un *white noise*.

In altri termini, dato un processo $I(1)$ si può stimare un modello ARMA sulle differenze prime, ma non è detto che queste descrivano un *white noise*; in realtà il processo di partenza potrebbe contenere, accanto alla componente non stazionaria, una componente $I(0)$ responsabile di oscillazioni di breve periodo.

La *scomposizione di Beveridge-Nelson* consente di separare le due componenti. Essa si basa su una semplice proprietà dei polinomi: dato un polinomio $C(z)$ di ordine $q$, è sempre possibile trovare un polinomio $C^*(z)$ di ordine $q-1$ tale che:

$$C(z) = C(1) + C^*(z)(1-z)$$

dove $C(1)$ non è altro che la somma dei coefficienti di $C(z)$.

Infatti, il polinomio $D(z) = C(z) - C(1)$ è ancora di grado $q$ e 1 è una sua radice; si ha quindi:

$$D(z) = C^*(z)(1-z) \quad \Rightarrow \quad C^*(z) = \frac{D(z)}{1-z} = \frac{C(z) - C(1)}{1-z}$$

Se $y_t$ e $I(1)$, allora $\Delta y_t$ è $I(0)$ e ammette una rappresentazione come media mobile:

$$\Delta y_t = C(L)\varepsilon_t$$

Scomponendo $C(L)$ si può scrivere:

$$\begin{aligned} \Delta y_t &= \left[C(1) + C^*(L)(1-L)\right]\varepsilon_t \\ &= C(1)\varepsilon_t + C^*(L)\varepsilon_t - C^*(L)\varepsilon_{t-1} \\ &= C(1)\varepsilon_t + C^*(L)\Delta\varepsilon_t \end{aligned}$$

Definendo un processo $r_t$ per cui valga $\Delta r_t = \varepsilon_t$, quindi un *random walk*, si giunge a:

$$y_t = C(1)r_t + C^*(L)\varepsilon_t$$

dove $C(1)r_t$ è un *random walk*, la *componente permanente* ad alta persistenza, mentre $C^*(L)\varepsilon_t$ è un processo $I(0)$, la *componente transitoria* a bassa persistenza.

**Esempio 7.11.** Dato un processo $y_t$, si valuta che si tratta di un processo $I(1)$ e si stima il modello $ARMA(1,1)$:

$$(1-aL)\Delta y_t = (1+cL)\varepsilon_t \quad \text{quindi} \quad C(L) = \frac{1+cL}{1-aL}$$

$C(1)$ non è altro che $\dfrac{1+c}{1-a}$. Quanto a $C^(L)$, svolgendo i semplici calcoli si ottiene:

$$C^*(L) = -\frac{a+c}{(1-a)(1-aL)}$$

e poi:

$$y_t = \frac{1+c}{1-a}r_t - \frac{a+c}{1-a}(1-aL)^{-1}\varepsilon_t$$

Quindi $y_t$ può essere rappresentato come combinazione di un *random walk* e di un processo $AR(1)$ tanto più persistente quanto maggiore è $|a|$.

# Capitolo 8

# I processi VAR

La serie storica del cambio euro/dollaro è un processo univariato. Le serie dei cambi euro/dollaro, euro/yen, /euro/sterlina è invece un processo multivariato. Se si considerano insieme $k$ processi $AR(p)$ univariati, si ottiene un processo $VAR(p)$ multivariato:

$$\underset{k,1}{\mathbf{y}_t} = \underset{k,k}{\mathbf{A}_1}\,\underset{k,1}{\mathbf{y}_{t-1}} + \underset{k,k}{\mathbf{A}_2}\,\underset{k,1}{\mathbf{y}_{t-2}} + \ldots \underset{k,k}{\mathbf{A}_p}\,\underset{k,1}{\mathbf{y}_{t-p}} + \underset{k,1}{\boldsymbol{\varepsilon}_t}$$

dove le $\mathbf{A}_j$ sono matrici di parametri e $\boldsymbol{\varepsilon}_t$ è un vettore di *white noise*. Ad esempio, per $k = 2$ e $p = 2$:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{1_{1,1}} & a_{1_{1,2}} \\ a_{1_{2,1}} & a_{1_{2,2}} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{2_{1,1}} & a_{2_{1,2}} \\ a_{2_{2,1}} & a_{2_{2,2}} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

Il *white noise* vettoriale, indicato con $VMN$ (*Vector White Noise*), è molto simile a quello univariato e presenta proprietà analoghe:

$$\mathbb{E}[\boldsymbol{\varepsilon}_t] = \mathbf{0} \qquad \boldsymbol{\Gamma}(j) = \mathbb{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_{t-j}] = \begin{cases} \mathbb{V}[\boldsymbol{\varepsilon}_t] = \boldsymbol{\Sigma} & \text{se } j = 0 \\ \mathbf{0} & \text{se } j \neq 0 \end{cases}$$

La matrice di varianze e covarianze $\boldsymbol{\Sigma}$ è simmetrica e definita positiva, ma non è necessariamente diagonale. Ciò vuol dire che qualsiasi $\varepsilon_{it}$ è incorrelato con qualsiai $\varepsilon$ presente nella storia passata ($j \neq 0$), ma potrebbe essere anche correlato con elementi contemporanei.

Esistono versioni vettoriali anche degli altri processi visti nel capitolo precedente, quindi $VMA(q)$ e $VARMA(p, q)$, ma sono nettamente più difficili da stimare. Nella pratica, quindi, si usano molto spesso i processi $VAR(p)$.

Il capitolo illustra la motivazione originaria dei processi VAR, le condizioni di stazionarietà e gli aspetti inferenziali.

## 8.1 Macroeconomia e realtà

Il titolo della sezione è quello dell'articolo di Sims (1980) che ha introdotto i modelli VAR. Sims mosse da un'analisi critica dei modelli basati su equazioni simultanee, del tipo:[1]

$$\begin{cases} y_t = c_t + i_t + u_{1t} \\ c_t = \beta_0 y_t + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 + u_{2t} \end{cases}$$

[1] La sezione si basa su Carlucci e Girardi (sd, pp. 2-4) e Lucchetti (2008, pp. 84-85).

in cui $y_t$ è il reddito, $c_t$ il consumo e $i_t$ l'investimento. Si tratta di modelli ricavati direttamente dalla teoria economica, detti quindi *strutturali*, e basati sulla distinzione tra variabili endogene ed esogene (in senso economico, non econometrico; cfr. cap. 3).

Sims mosse tre critiche:

a) la distinzione tra endogene e esogene è arbitraria; ad esempio, nel modello appena visto $i_t$ è un'esogena, ma si potrebbe aggiungere un'equazione in cui l'investimento fosse a sua volta dipendente dal reddito (ipotesi tutt'altro che avventata) e diventerebbe endogena;

b) ciascuna singola equazione costituisce un modello di equilibrio parziale e, come tale, si basa su ipotesi economiche che impongono una serie di vincoli del tipo *ceteris paribus*; tali vincoli, però, variano da equazione a equazione e possono risultare contraddittori quando si esamina il modello nel suo complesso;

c) il numero dei parametri è spesso maggiore del numero delle equazioni, rendendo i parametri non identificabili (non stimabili in modo univoco); il problema viene risolto adottando ipotesi economiche che consentano di introdurre restrizioni sui parametri, ad esempio di azzerarne alcuni; si tratta però di restrizioni poco credibili, in quanto derivate dalla teoria economica solo per risolvere un problema prettamente statistico.

Propose invece i VAR. In essi le singole equazioni, tutte in forma matriciale e tante quante sono le unità temporali considerate, non costituiscono modelli di equilibrio parziale, ma ciascuna variabile può dipendere *a priori* da ogni altra; si ha infatti un'unica variabile multipla, $\mathbf{y}_t$, che dipende da se stessa ritardata di $1, 2, \ldots, p$ unità temporali. Non vi è bisogno di basare le equazioni su ipotesi economiche (almeno non nella formulazione iniziale del modello), soprattutto non si deve ricorrere a ipotesi economiche per giustificare restrizioni motivate solo da considerazioni di tipo statistico.

All'inizio i modelli vennero etichettati come "a-teorici", e non sembrava un complimento, ma si sono poi molto diffusi e hanno ormai sostituito i sistemi di equazioni simultanee.

## 8.2 Condizioni di stazionarietà

Si può usare l'operatore ritardo anche per processi multivariati. Intendendo:

$$L\mathbf{y}_t = \mathbf{y}_{t-1}$$

un processo $VAR(p)$ può essere espresso nella forma:

$$\mathbf{A}(L)\mathbf{y}_t = \boldsymbol{\varepsilon}_t, \qquad \mathbf{A}(L) = \mathbf{I} - \mathbf{A}_1 L - \mathbf{A}_2 L^2 - \cdots - \mathbf{A}_p L^p$$

Il polinomio $\mathbf{I} - \mathbf{A}_1 L - \mathbf{A}_2 L^2 - \cdots - \mathbf{A}_p L^p$ è un polinomio matriciale. Analogamente a quanto visto nel capitolo 7, per studiare le proprietà algebriche del polinomio si sostituisce l'operatore $L$ con $z \in \mathbb{C}$.

Il determinante di $\mathbf{A}(z)$ viene detto *polinomio caratteristico*.

Vale per i $VAR(p)$ una condizione di stazionarietà analoga a quella stabilita dal teorema 7.5 per gli $AR(p)$.

**Teorema 8.1.** *Un processo VAR(p) è stazionario ed ergodico se e solo se le radici del polinomio caratteristico sono tutte fuori del cerchio unitario. In questo caso il processo ammette la rappresentazione VMA(∞):*

$$\mathbf{y}_t = \sum_{n=0}^{\infty} \mathbf{C}_n \boldsymbol{\varepsilon}_{t-n}$$

*dove i* $\mathbf{C}_n$ *sono i coefficienti dell'espansione in serie di Taylor di* $\mathbf{A}(z)^{-1}$ *intorno allo zero.*

*Dimostrazione.* Da $\mathbf{A}(L)\mathbf{y}_t = \varepsilon_t$ si ricava:

$$\mathbf{y}_t = \mathbf{A}(L)^{-1}\boldsymbol{\varepsilon}_t$$

Sostituendo l'operatore $L$ con $z \in \mathbb{C}$. Si ha:

$$\mathbf{A}(z)^{-1} = \frac{\operatorname{adj} \mathbf{A}(z)}{\det \mathbf{A}(z)}$$

Per il Teorema Fondamentale dell'Algebra,

$$\mathbf{C}(z) = \det \mathbf{A}(z) = \prod_{n=1}^{l} \left(1 - \frac{z}{z_n}^{m_n}\right)$$

dove $z_n$ è una radice e $m_n$ è la sua molteplicità algebrica (si può dividere per $z_n$ in quanto 0 non può essere una radice: $\det \mathbf{A}(0) = \det \mathbf{I} = 1$). Le radici $z_n$ sono punti di singolarità:

$$\lim_{|z-z_n|\to 0} \mathbf{C}(z) = \infty$$

Espandendo $\mathbf{C}(z)$ in serie di Taylor intorno a 0 si ha:

$$\mathbf{C}(z) = \sum_{n=0}^{\infty} \mathbf{c}_n z^n, \qquad \mathbf{c}_n = \frac{\mathbf{C}^{(n)}(0)}{n!}$$

Si tratta di una serie che definisce (con $L$ al posto di $z$) un processo $VMA(\infty)$:

$$\mathbf{y}_t = \mathbf{C}(L)\boldsymbol{\varepsilon}_t = \sum_{n=0}^{\infty} \mathbf{c}_n L^n \boldsymbol{\varepsilon}_{t-n}$$

Se e solo se le radici caratteristiche sono fuori del cerchio unitario si ha:

$$\mathbf{C}(1) = \sum_{n=0}^{\infty} \mathbf{c}_n < \infty$$

In questo caso, il processo $VAR(p)$ è rappresentabile come un processo $VMA(\infty)$ stazionario ed ergodico (Hamilton 1994). □

**Esempio 8.2.** Dato il processo:

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \mathbf{A} = \begin{bmatrix} 3 & 6 \\ 1 & 4 \end{bmatrix}$$

si ha:

$$\mathbf{A}(z) = \mathbf{I} - \mathbf{A}z = \begin{bmatrix} 1-3z & -6z \\ -z & 1-4z \end{bmatrix}, \qquad \det \mathbf{A}(z) = (1-3z)(1-4z) - 6z^2 = 1 - 7z + 6z^2$$

Le radici del polinomio caratteristico sono $z_1 = 1$ e $z_2 = \frac{1}{6}$. Ne risulta che il processo è esplosivo in quanto ha una matrice dentro il cerchio unitario.

## 8.3 Inferenza

L'eventuale non stazionarietà di un processo VAR ha un impatto minore rispetto a quanto accade con i processi univariati. Il processo $\mathbf{A}(L)\mathbf{y}_t = \boldsymbol{\varepsilon}_t$ è stazionario se le radici del polinomio caratteristico $\det \mathbf{A}(z)$ sono tutte maggiori di 1 in modulo; se qualche radice è pari a 1, tuttavia, possono verificarsi vari casi.

In particolare, se $\mathbf{A}(1)$ è una matrice nulla, si può utilizzare la scomposizione di Beveridge-Nelson (sez. 7.5.4):

$$\mathbf{A}(L) = \mathbf{A}(1) + \mathbf{A}^*(L)\Delta$$

che conduce a:

$$\mathbf{A}^*(L)\Delta\mathbf{y}_t = \boldsymbol{\varepsilon}_t$$

Se $\Delta\mathbf{y}_t$ risulta stazionario, si effettua l'analisi sulla serie delle differenze.

Se invece $\mathbf{A}(1)$ non è nulla, risulta comunque non invertibile. Ne derivano alcune proprietà che verranno esaminate nel capitolo successivo.

In ogni caso, il metodo OLS produce comunque stime consistenti dei parametri, anche se la presenza di radici unitarie può comportare distribuzioni limite non standard.

# Capitolo 9

# Cointegrazione

È possibile definire processi stocastici che siano combinazioni lineari di processi stocastici. La combinazione lineare di due processi stazionari è ancora un processo stazionario, quella di un processo $I(1)$ e un processo $I(0)$ è $I(1)$.[1] La combinazione lineare di due processi $I(1)$, invece, non è sempre $I(1)$; se risulta $I(0)$ i due processi vengono detti *cointegrati*.

Una combinazione lineare $I(0)$, se esiste, vuol dire che i due processi possono ciascuno tendere verso qualche asintoto, ma c'è comunque tra loro una relazione che vale sempre. In termini economici, potrebbe voler dire che esiste tra loro una *relazione di equilibrio* di lungo periodo.

## 9.1 Definizioni

Dati due processi stocastici $x_t \sim I(d)$ e $y_t \sim I(b)$, una loro combinazione lineare $z_t = x_t + ay_t$ è $I(c)$, dove:

$$\begin{cases} c = \max\{d, b\} & \text{se } d \neq b \\ c \leq \max\{d, b\} & \text{se } d = b \end{cases}$$

Se $d = b$ e $c < \max\{d, b\}$ (secondo caso con disuguaglianza stretta), si ha *cointegrazione*.

Si considera spesso il caso di due processi $I(1)$ per i quali esista una combinazione lineare $I(0)$. Ad esempio:

$$\begin{aligned} x_{1t} &= x_{1t-1} + \varepsilon_t \\ x_{2t} &= x_{1t} + u_t \end{aligned}$$

dove $\varepsilon_t$ e $u_t$ sono $I(0)$. I due processi sono chiaramente $I(1)$. Il processo

$$z_t = x_{2t} - x_{1t} = u_t$$

è altrettanto chiaramente $I(0)$. Quindi c'è cointegrazione.

Si può costruire un vettore $\mathbf{y}_t = (x_{1t}, x_{2_t})$ contenente i due processi $I(1)$ come elementi e scrivere:

$$z_t = \mathbf{y}_t' \boldsymbol{\beta} = u_t \qquad \mathbf{y}_t = \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

[1] Si può pensare ad un processo $ARMA(p, q)$, somma di un $MA(q)$ sempre stazionario e di un $AR(p)$ che può esserlo o non esserlo e, quindi, decide della stazionarietà della combinazione.

Il vettore $\boldsymbol{\beta}$ viene detto *vettore di cointegrazione*. Nel caso ce ne fossero più d'uno si parla di *matrice di correlazione* e il numero di vettori linearmente indipendenti viene detto *rango di correlazione*.

Un processo $I(1)$ multivariato per il quale esista almeno un vettore di cointegrazione viene detto *sistema cointegrato*.

## 9.2 Modelli a correzione d'errore

Dato un processo $VAR(1)$ composto di $n$ processi univariati:

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$$

sottraendo $\mathbf{y}_{t-1}$ da ambo i lati si ottiene $\mathbf{y}_t - \mathbf{y}_{t-1} = \mathbf{A}\mathbf{y}_{t-1} - \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$, ovvero:

$$\Delta\mathbf{y}_t = \mathbf{\Pi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \underset{n,n}{\mathbf{\Pi}} = \mathbf{A} - \mathbf{I}$$

Indicando con $r$ il rango della matrice $\mathbf{\Pi}$, ci sono tre possibilità:

**$r = n$:** la matrice $\mathbf{\Pi}$ è invertibile; ciò vuol dire che non vi sono radici unitarie in quanto, se vi fossero, non sarebbe invertibile nemmeno $\mathbf{A}(z) = \mathbf{I} - \mathbf{A}z$, che per $z = 1$ è l'opposta di $\mathbf{\Pi}$; quindi $\mathbf{y}_t$ è $I(0)$ (cfr. sez. 8.2) e non c'è cointegrazione;

**$r = 0$:** la matrice $\mathbf{\Pi}$ non è invertibile, ma è anche nulla, quindi $\mathbf{A} = \mathbf{I}$, $\mathbf{y}_t$ è un *random walk* multivariato e non c'è cointegrazione (cfr. sez. 8.3);

**$0 < r < n$:** $\mathbf{y}_t$ è un sistema cointegrato e $r$ è il rango di cointegrazione.

Il terzo caso è quello che interessa. Se $\mathbf{\Pi}$ non è a rango pieno, le sue colonne non sono linearmente indipendenti; deve quindi esistere una matrice $n \times r$, comunemente indicata con $\alpha$, le cui colonne siano una base dello spazio vettoriale generato dalle colonne di $\mathbf{\Pi}$. Ciascuna colonna di $\mathbf{\Pi}$ deve essere una combinazione lineare delle colonne di $\alpha$, cioè deve essere il prodotto di $\alpha$ per un vettore di $r$ elementi; l'intera $\mathbf{\Pi}$ sarà quindi uguale al prodotto di $\alpha$ per la trasposta di un'altra matrice $n \times r$, normalmente indicata con $\beta$. Riassumendo:

$$\mathbf{\Pi} = \alpha\beta'$$

Mentre $\alpha$ è una base priva delle ridondanze presenti in $\mathbf{\Pi}$, le colonne di $\beta$ operano le combinazioni lineari che conducono a $\mathbf{\Pi}$, ovvero al processo (ad un modello del processo che appare coerente con i dati osservati). $\beta$ è quindi la *matrice di cointegrazione* e si può scrivere:

$$\underset{n,1}{\Delta\mathbf{y}_t} = \underset{r,n}{\alpha}\ \underset{r,1}{\mathbf{z}_{t-1}} + \underset{n,1}{\boldsymbol{\varepsilon}_t}, \qquad \underset{r,1}{\mathbf{z}_t} = \underset{r,n}{\beta'}\ \underset{n,1}{\mathbf{y}_t}$$

$\mathbf{z}_t$ è un processo $I(0)$ e le sue singole realizzazioni rappresentano la serie storica delle oscillazioni di breve periodo intorno ad un equilibrio di lungo periodo. La matrice $\alpha$ viene detta *matrice dei pesi*, perché il suo elemento $ij$ indica l'effetto che il $j$-esimo elemento di $\mathbf{z}_{t-1}$ deve avere sulla $i$-esima variabile perché si ripristini l'equilibrio.

Il modello che si ottiene sostituendo $\mathbf{\Pi}\mathbf{y}_{t-1}$ con $\alpha\beta'\mathbf{y}_t = \alpha\mathbf{z}_{t-1}$ viene quindi detto *meccanismo a correzione d'errore*; il VAR viene riscritto come VECM: *Vector Error Correction Model*.

**Esempio 9.1.** Si può supporre di esaminare le serie storiche dei logaritmi del PIL, $y_t$, e dell'offerta reale di moneta, $m_t$. Si avrebbe un modello del tipo:

$$\mathbf{x}_t = \begin{bmatrix} y_t \\ m_t \end{bmatrix} = \mathbf{A} \begin{bmatrix} y_{t-1} \\ m_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

Secondo la teoria quantitativa della moneta:

$$MV = PY$$

da cui segue che la velocità di circolazione della moneta è data dal rapporto tra il PIL e l'offerta reale di moneta:

$$V = \frac{Y}{M/P}$$

Usando le minuscole per i logaritmi e $m$ per il logaritmo di $M/P$:

$$v = y - m$$

Supponendo che:

a) $y_t$ e $m_t$ siano processi $I(1)$;

b) $v_t$ sia $I(0)$, fluttuando intorno ad un valore centrale;

si potrebbe dire che $y_t$ e $m_t$ cointegrano e che il vettore di cointegrazione è $\beta = (1, -1)$, ovvero che

$$\mathbf{v}_t = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \mathbf{y}_t \\ \mathbf{m}_t \end{bmatrix} = \mathbf{y}_t - \mathbf{m}_t$$

è una *relazione di equilibrio*. Se vi è squilibrio, le variazioni del PIL e dell'offerta reale di moneta sono date da:

$$\begin{cases} \Delta y_t = \alpha_1 (y_{t-1} - m_{t-1}) + \varepsilon_{1t} \\ \Delta m_t = \alpha_2 (y_{t-1} - m_{t-1}) + \varepsilon_{2t} \end{cases}$$

ovvero: se $(y_{t-1} - m_{t-1})$, il logaritmo della velocità di circolazione, era troppo basso o troppo alto al tempo $t-1$, i coefficienti $\alpha_1$ e $\alpha_2$ dicono di quanto $y_t$ e $m_t$ sono aumentati o diminuiti per tendere a ripristinare l'equilibrio.

Si può dire che, mentre i VAR vennero inizialmente etichettati come "a-teorici" (sez. 8.1), i VECM hanno ristabilito e rifondato il collegamento tra analisi delle serie storiche e teoria economica.

## 9.3 Il teorema di rappresentazione di Granger

Il teorema di rappresentazione di Granger stabilisce la possibilità di rappresentare un sistema cointegrato anche nella forma della somma di un trend stocastico e di un processo $MA(\infty)$ stazionario.

**Teorema 9.2** (Teorema di rappresentazione di Granger)**.** *Dato un processo* $\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$*, se la matrice* $\boldsymbol{\Pi} = \mathbf{A} - \mathbf{I}$ *non ha rango pieno, quindi se* $\mathbf{y}_t$ *è un sistema cointegrato e ammette una rappresentazione VECM con* $\alpha\beta' = \boldsymbol{\Pi}$*,* $\mathbf{y}_t$ *ammette anche la rappresentazione:*

$$\mathbf{y}_t = \mathbf{C} \sum_{i=0}^{t} \boldsymbol{\varepsilon}_t + \sum_{n=0}^{\infty} c_n \boldsymbol{\varepsilon}_{t-n}, \qquad \mathbf{C} = \beta_\perp (\mathbf{M})^{-1} \alpha_\perp{}'$$

$\mathbf{C}\sum_{i=0}^{t}\boldsymbol{\varepsilon}_t$ è un trend stocastico, la parte non stazionaria, mentre $\sum_{n=0}^{\infty}c_n\boldsymbol{\varepsilon}_{t-n}$ è una $MA(\infty)$ stazionaria.

La matrici $\alpha_\perp$ e $\beta_\perp$ contengono le $n-r$ colonne che sono base dei sottospazi ortogonali a quelli generati dalle colonne, rispettivamente, di $\alpha$ e di $\beta$; ne segue che $\alpha'\alpha_\perp = \mathbf{0}$ e $\beta'\beta_\perp = \mathbf{0}$.

Vista la definizione di $\mathbf{C}$, moltiplicando entrambi i membri della rappresentazione per $\beta'$ si ottiene quindi:

$$\begin{aligned}\beta'\mathbf{y}_t &= \beta'\mathbf{C}\sum_{i=0}^{t}\boldsymbol{\varepsilon}_t + \beta'\sum_{n=0}^{\infty}c_n\boldsymbol{\varepsilon}_{t-n}\\ &= \beta'\beta_\perp(\mathbf{M})^{-1}{\alpha_\perp}'\sum_{i=0}^{t}\boldsymbol{\varepsilon}_t + \beta'\sum_{n=0}^{\infty}c_n\boldsymbol{\varepsilon}_{t-n}\\ &= \beta'\sum_{n=0}^{\infty}c_n\boldsymbol{\varepsilon}_{t-n}\end{aligned}$$

quindi $\beta'\mathbf{y}_t$ si conferma stazionario, come nella rappresentazione VECM.

Il teorema consente quindi di affermare che processi cointegrati condividono un trend stocastico comune, che conferisce senso al loro studio congiunto.

# Parte III

# Appendici

# Appendice A

# Complementi di algebra lineare

## A.1 Matrici inverse e inverse generalizzate

Come noto, data una matrice quadrata $\mathbf{A}$ di ordine $n$ a rango pieno, si dice sua *inversa* e si indica con $\mathbf{A}^{-1}$ una matrice tale che:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

La definizione di inversa può essere tuttavia resa più generale e così applicabile anche a matrici non quadrate e/o non di rango pieno.

**Definizione A.1.** Data una matrice $\underset{m,n}{\mathbf{A}}$, si dicono sua *inversa destra* una matrice $\underset{n,m}{\mathbf{A}}^{-R}$, sua *inversa sinistra* una matrice $\underset{n,m}{\mathbf{A}}^{-L}$ tali che:

$$\underset{m,n}{\mathbf{A}}\ \underset{n,m}{\mathbf{A}}^{-R} = \underset{m,m}{\mathbf{I}} \qquad\qquad \underset{n,m}{\mathbf{A}}^{-L}\ \underset{m,n}{\mathbf{A}} = \underset{n,n}{\mathbf{I}}$$

**Osservazione A.2.** Un'inversa destra di $\underset{m,n}{\mathbf{A}}$ esiste solo se $m \leq n$ e $\mathrm{rk}(\mathbf{A}) = m$, un'inversa sinistra solo se $n \leq m$ e $\mathrm{rk}(\mathbf{A}) = n$. Ciò in quanto la moltiplicazione di una matrice per un'altra non può aumentarne il rango: $\mathrm{rk}(\mathbf{AB}) \leq \min\{\mathrm{rk}(\mathbf{A}), \mathrm{rk}(\mathbf{B})\}$ (v. proposizione A.32), ma il risultato di una moltiplicazione per un'inversa destra o sinistra è, per definizione, una matrice identità di rango, rispettivamente, $m$ o $n$. Inoltre, se le inverse destra e sinistra esistono non sono uniche.

**Esempio A.3.** Date le seguenti tre matrici:

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 1 & 1 \\ -2 & -1 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} -5/18 & 1/9 & -13/18 \\ 1/2 & 0 & 1/2 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} -8/9 & 23/9 & -1/9 \\ 1 & -2 & 0 \end{bmatrix}$$

si verifica facilmente che $\mathbf{B}$ e $\mathbf{C}$ sono entrambe inverse sinistre di $\mathbf{A}$ e che le loro trasposte sono entrambe inverse destre della trasposta di $\mathbf{A}$:

$$\mathbf{BA} = \mathbf{CA} = \mathbf{I}_2 \qquad\qquad \mathbf{A}'\mathbf{B}' = \mathbf{A}'\mathbf{C}' = \mathbf{I}_2$$

**Esempio A.4.** In generale:

a) data una matrice $\underset{m,n}{\mathbf{A}}$ con $m > n$ e rango $r = n$, la matrice $\left(\underset{n,m}{\mathbf{A}}{}' \underset{m,n}{\mathbf{A}}\right)$ è una matrice simmetrica $n \times n$ di rango $n$, quindi è invertibile; un'inversa sinistra di $\mathbf{A}$ è:

$$(\underset{n,n}{\mathbf{A}}{}'\mathbf{A})^{-1} \underset{n,m}{\mathbf{A}}{}'$$

in quanto $[(\underset{n,m}{\mathbf{A}}{}'\mathbf{A})^{-1}\mathbf{A}'] \underset{m,n}{\mathbf{A}} = \underset{n,n}{\mathbf{I}}$; nell'esempio precedente, infatti, la matrice $\mathbf{B}$ era stata ottenuta proprio in questo modo;

b) analogamente, data una matrice $\underset{m,n}{\mathbf{A}}$ con $m < n$ e rango $r = m$, un'inversa destra sarà $\underset{n,m}{\mathbf{A}}{}'(\underset{m,m}{\mathbf{A}}\mathbf{A}')^{-1}$, in quanto $\underset{m,n}{\mathbf{A}}[\mathbf{A}'(\underset{n,m}{\mathbf{A}}\mathbf{A}')^{-1}] = \underset{m,m}{\mathbf{I}}$.

**Teorema A.5.** *Se* $\mathbf{A}$ *è una matrice quadrata di rango pieno, le sue inverse destra e sinistra coincidono e sono uniche. La matrice* $\mathbf{A}^{-L} = \mathbf{A}^{-R} = \mathbf{A}^{-1}$ *viene detta l'*inversa *di* $\mathbf{A}$.

**Definizione A.6.** Data una matrice $\underset{m,n}{\mathbf{A}}$, si dice sua *inversa generalizzata* una matrice $\underset{n,m}{\mathbf{A}}{}^{-}$ tale che:

$$\underset{m,n}{\mathbf{A}}\,\underset{n,m}{\mathbf{A}}{}^{-}\,\underset{m,n}{\mathbf{A}} = \underset{m,n}{\mathbf{A}}$$

**Osservazione A.7.** Se $\mathbf{A}$ ha un'inversa destra o sinistra, questa è anche una sua inversa generalizzata; infatti:

$$\mathbf{A}\mathbf{A}^{-R}\mathbf{A} = \mathbf{I}\mathbf{A} = \mathbf{A} \qquad\qquad \mathbf{A}\mathbf{A}^{-L}\mathbf{A} = \mathbf{A}\mathbf{I} = \mathbf{A}$$

Ne segue che l'inversa generalizzata non è unica, a meno che $\mathbf{A}$ sia quadrata e di rango pieno; in tal caso, infatti, $\mathbf{A}^{-R} = \mathbf{A}^{-L} = \mathbf{A}^{-1}$ e $\mathbf{A}\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}$, oppure, se $\mathbf{A}$ non è quadrata o non è di rango pieno, che l'inversa generalizzata sia tale da soddisfare le proprietà esposte nella definizione che segue.

**Definizione A.8.** Data una matrice $\underset{m,n}{\mathbf{A}}$, un'inversa generalizzata $\underset{n,m}{\mathbf{A}}{}^{+}$ tale che:

a) $\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$;

b) $\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+}$;

c) $\mathbf{A}\mathbf{A}^{+} = (\mathbf{A}\mathbf{A}^{+})'$;

d) $\mathbf{A}^{+}\mathbf{A} = (\mathbf{A}^{+}\mathbf{A})'$;

viene detta *pseudoinversa* (*di Moore-Penrose*).

**Esempio A.9.** La matrice $\mathbf{B}$ dell'esempio A.3 è la presudoinversa della matrice $\mathbf{A}$, come si verifica facilmente. Non lo è invece $\mathbf{C}$, in quanto $\mathbf{AC}$ non è simmetrica.

Una matrice può avere un'inversa destra o sinistra solo se è a rango pieno, ma si dimostra che ogni matrice ha una pseudo inversa di Moore-Penrose e, inoltre, che questa è unica.

**Osservazione A.10.** Per trovare la pseudoinversa di una matrice si può ricorrere alla *scomposizione ai valori singolari*, mediante la quale la matrice viene scomposta nel prodotto di tre matrici:

$$\underset{m,n}{\mathbf{A}} = \underset{m,m}{\mathbf{U}}\,\underset{m,n}{\mathbf{\Sigma}}\,\underset{n,n}{\mathbf{V}}{}'$$

dove:

– $\mathbf{U}$ è una matrice ortogonale le cui colonne sono autovettori di $\mathbf{AA}'$;
– $\mathbf{\Sigma}$ è una matrice "diagonale" (nel senso che $\sigma_{ij} = 0$ se $i \neq j$) i cui elementi $\sigma_{ii}$ sono i *valori singolari* di $\mathbf{A}'\mathbf{A}$, cioè le radici quadrate dei suoi autovalori;
– $\mathbf{V}'$ è la trasposta di una matrice ortogonale $\mathbf{V}$ le cui colonne sono autovettori di $\mathbf{A}'\mathbf{A}$.

La pseudoinversa di $\mathbf{\Sigma}$ è una matrice che ha come unici elementi non nulli i reciproci degli elementi non nulli di $\mathbf{\Sigma}$:

$$\underset{m,n}{\mathbf{\Sigma}} = \begin{bmatrix} \begin{bmatrix} \sigma_{11} & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & \sigma_{rr} \end{bmatrix} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \quad \underset{n,m}{\mathbf{\Sigma}^+} = \begin{bmatrix} \begin{bmatrix} 1/\sigma_{11} & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & 1/\sigma_{rr} \end{bmatrix} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}$$

e si ha:

$$\mathbf{\Sigma\Sigma}^+ = \begin{bmatrix} \begin{bmatrix} 1 & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & 1 \end{bmatrix} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}_{m\times m} \qquad \mathbf{\Sigma}^+\mathbf{\Sigma} = \begin{bmatrix} \begin{bmatrix} 1 & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & 1 \end{bmatrix} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}_{n\times n}$$

Si vede facilmente che pre/post moltiplicando $\mathbf{\Sigma}$ per $\mathbf{\Sigma}^+$ si ottengono matrici simmetriche e che $\mathbf{\Sigma\Sigma}^+\mathbf{\Sigma} = \mathbf{\Sigma}$ e $\mathbf{\Sigma}^+\mathbf{\Sigma\Sigma}^+ = \mathbf{\Sigma}^+$. Ricordando che l'inversa di una matrice ortogonale è la sua trasposta, la pseudoinversa di $\mathbf{A} = \mathbf{U\Sigma V}'$ è $\mathbf{A}^+ = \mathbf{V\Sigma}^+\mathbf{U}'$, infatti:

$$\mathbf{AA}^+\mathbf{A} = \mathbf{U\Sigma V}'\mathbf{V\Sigma}^+\mathbf{U}'\mathbf{U\Sigma V}' = \mathbf{U\Sigma\Sigma}^+\mathbf{\Sigma V}' = \mathbf{U\Sigma V}' = \mathbf{A}$$

## A.2 Matrici di proiezione

Come noto:

a) dato uno spazio vettoriale $V$, due suoi sottospazi $U$ e $W$ sono detti *ortogonali* se, comunque presi due vettori $\mathbf{u} \in U$ e $\mathbf{w} \in W$, si ha $\mathbf{u}'\mathbf{w} = \mathbf{w}'\mathbf{u} = 0$;
b) se $V = U \oplus W$, la somma diretta $U \oplus W$ viene detta *scomposizione ortogonale* di $V$, $U$ viene scritto anche come $W^\perp$ e $W$ come $U^\perp$, $U$ e $W$ vengono detti l'uno il *complemento ortogonale* dell'altro;
c) se $U$ è un sottospazio di $\mathbb{R}^n$, $U \oplus U^\perp = \mathbb{R}^n$;
d) se i vettori di una base di uno spazio vettoriale sono tra loro a due a due ortogonali, la base viene detta *ortogonale*;
e) se i vettori di uba base di uno spazio vettoriale sono tra loro a due a due ortogonali e hanno norma unitaria, la base viene detta *ortonormale*.

**Esempio A.11.** Prima di procedere, potrebbe essere utile qualche esempio basato sui familiari spazi $\mathbb{R}^n$. Se $U \subset \mathbb{R}^2$ è uno spazio ad una dimensione, può essere l'insieme delle rette proporzionali al vettore unitario $\mathbf{e}_1 = (1, 0)$ (l'asse delle ascisse); il suo complemento ortogonale è il sottospazio $W$ delle rette proporzionali al vettore $\mathbf{e}_2 = (0, 1)$; la somma diretta dei due sottospazi è il piano $\mathbb{R}^2$, con base ortonormale $\{(1, 0), (0, 1)\}$. Analogamente, se $U \subset \mathbb{R}^3$ è uno spazio a due dimensioni con base $\{\mathbf{e}_1 = (1, 0, 0), \mathbf{e}_2 = (0, 1, 0)\}$ può essere visto come il piano $xy$, i cui punti hanno ascissa $x$, ordinata $y$ e quota nulla; il suo complemento ortogonale è il sottospazio $W$ delle rette proporzionali al vettore

$\mathbf{e}_3 = (0, 0, 1)$; la loro somma diretta è lo spazio tridimensionale $\mathbb{R}^3$ con base ortonormale $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

**Definizione A.12.** Dati lo spazio vettoriale $\mathbb{R}^n$ e una sua scomposizione ortogonale $\mathbb{R}^n = U \oplus U^\perp$, si dice *scomposizione ortogonale* di un vettore $\mathbf{v} \in \mathbb{R}^n$ la sua espressione come somma di due vettori $\mathbf{v}_1 \in U$ e $\mathbf{v}_2 \in U^\perp$:

$$\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 \qquad \mathbf{v}_1 \in U,\ \mathbf{v}_2 \in U^\perp$$

**Definizione A.13.** Dati lo spazio vettoriale $\mathbb{R}^n$ e una sua scomposizione ortogonale $\mathbb{R}^n = U \oplus U^\perp$, si dice *matrice di proiezione* sullo spazio $U$ una matrice quadrata $\mathbf{P}$ tale che:

a) $\mathbf{Pv} \in U$ per ogni $\mathbf{v} \in \mathbb{R}^n$;

b) $\mathbf{Pv} = \mathbf{v}$ per ogni $\mathbf{v} \in U$.

In altri termini, una matrice di proiezione trasforma qualsiasi vettore di $\mathbb{R}^n$ in un vettore di $U$ e lascia immutato un vettore che già appartenga a $U$. È quadrata in quanto trasforma vettori di $\mathbb{R}^n$ in vettori di $\mathbb{R}^n$.

**Osservazione A.14.** Dalla definizione di matrice di proiezione segue che $\mathbf{PPv} = \mathbf{Pv}$ (da destra verso sinistra: $\mathbf{Pv}$ trasforma $\mathbf{v}$ in un vettore di $U$; la successiva moltiplicazione per $\mathbf{P}$ lascia immutato il risultato); segue cioè che una matrice di proiezione è una matrice *idempotente*: $\mathbf{P}^2 = \mathbf{P}$.

**Osservazione A.15.** La matrice identità $\mathbf{I}$ è chiaramente idempotente. Se $\mathbf{P}$ è una matrice idempotente, è tale anche $\mathbf{I} - \mathbf{P}$. Infatti:

$$(\mathbf{I} - \mathbf{P})^2 = \mathbf{I}^2 - \mathbf{IP} - \mathbf{PI} + \mathbf{P}^2 = \mathbf{I} - 2\mathbf{P} + \mathbf{P} = \mathbf{I} - \mathbf{P}$$

**Definizione A.16.** Se $\mathbf{P}$ è una matrice di proiezione su $U \subset \mathbb{R}^n$, $\mathbb{R}^n = U \oplus U^\perp$ e se $\mathbf{I} - \mathbf{P}$ è una matrice di proiezione su $u^\perp$, allora $\mathbf{P}$ viene detta *matrice di proiezione ortogonale* su $U$.

**Osservazione A.17.** Una matrice di proiezione ortogonale $\mathbf{P}$, oltre ad essere idempotente, è anche simmetrica. Infatti, per qualsiasi $\mathbf{v} \in \mathbb{R}^n = U \oplus U^\perp$, essendo $\mathbf{Pv} \in U$ e $(\mathbf{I} - \mathbf{P})\mathbf{v} \in U^\perp$ si deve avere:

$$(\mathbf{Pv})'(\mathbf{I} - \mathbf{P})\mathbf{v} = \mathbf{v}'\mathbf{P}'(\mathbf{I} - \mathbf{P})\mathbf{v} = 0$$

Potendo $\mathbf{v}$ essere un qualsiasi vettore di $\mathbb{R}^n$, deve risultare:

$$\mathbf{P}'(\mathbf{I} - \mathbf{P}) = \mathbf{P}' - \mathbf{P}'\mathbf{P} = \mathbf{O}$$

Ciò è possibile se e solo se $\mathbf{P}'\mathbf{P} = (\mathbf{P}')^2 = \mathbf{P}'$, cioè se e solo se $\mathbf{P} = \mathbf{P}'$.

**Esempio A.18.** Sia $\{\mathbf{u}_1 = (1, 0, 0), \mathbf{u}_2 = (1, 1, 0)\}$ una base di $U \subset \mathbb{R}^3$. Le matrici:

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{I} - \mathbf{P} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

sono entrambe idempotenti. $\mathbf{P}$ proietta qualsiasi vettore di $\mathbb{R}^3$ in $U$. Ad esempio, se $\mathbf{v} = (2, 1, 1)$, $\mathbf{Pv} = (3, 2, 0)$, che appartiene evidentemente a $U$: $\mathbf{Pv} = \mathbf{u}_1 + 2\mathbf{u}_2$. $\mathbf{I} - \mathbf{P}$ proietta invece $\mathbf{v}$ in uno spazio che non è ortogonale a $U$, infatti $(\mathbf{I} - \mathbf{P})\mathbf{v} = (-1, -1, 1)$ e $\mathbf{u}_1'\mathbf{v} = -1$, $\mathbf{u}_2'\mathbf{v} = -2$.

**Esempio A.19.** Sia $\{\mathbf{u}_1 = (1,0,0), \mathbf{u}_2 = (1,1,0)\}$ una base di $U \subset \mathbb{R}^3$. Le matrici:

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{I} - \mathbf{P} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

sono entrambe simmetriche oltre che idempotenti. $\mathbf{P}$ proietta qualsiasi vettore di $\mathbb{R}^3$ in $U$. Ad esempio, se $\mathbf{v} = (2,1,1)$, $\mathbf{Pv} = (2,1,0) = \mathbf{u}_1 + \mathbf{u}_2$. $\mathbf{I} - \mathbf{P}$ proietta $\mathbf{v}$ in uno spazio ortogonale a $U$, infatti $(\mathbf{I}-\mathbf{P})\mathbf{v} = (0,0,1)$ è ortogonale sia a $\mathbf{u}_1$ che a $\mathbf{u}_2$, quindi a tutte le loro combinazioni lineari (a tutti gli elementi di $U$). $\mathbf{P}$ è quindi una matrice di proiezione ortogonale.

**Osservazione A.20.** Dati uno spazio vettoriale $V$ ed un suo sottospazio $U$, esistono molte matrici di proiezione su $U$, ma una sola matrice di proiezione ortogonale su $U$; esiste, cioè, una sola matrice di proiezione $\mathbf{P}$ tale che $\mathbf{I} - \mathbf{P}$ sia una matrice di proiezione su $U^\perp$.

**Proposizione A.21.** *Una matrice idempotente ha come autovalori solo 1 e/o 0.*

*Dimostrazione.* Sia $\mathbf{A}$ una matrice idempotente e sia $\mathbf{v}$ un vettore di tanti elementi quante sono le colonne di $\mathbf{A}$. Per la definizione di autovalore e autovettore, si ha $\mathbf{Av} = \lambda\mathbf{v}$, ma anche $\mathbf{A}^2\mathbf{v} = \mathbf{A}(\mathbf{Av}) = \mathbf{A}(\lambda\mathbf{v}) = \lambda^2\mathbf{v}$. Essendo $\mathbf{A}$ idempotente:

$$\mathbf{A}^2\mathbf{v} = \mathbf{Av} \;\Rightarrow\; \lambda^2\mathbf{v} = \lambda\mathbf{v} \;\Rightarrow\; (\lambda^2 - \lambda)\mathbf{v} = 0 \;\Rightarrow\; \lambda(\lambda - 1) = 0 \;\Rightarrow\; \lambda \in \{0,1\} \qquad \square$$

**Proposizione A.22.** *Il rango di una matrice idempotente è uguale alla sua traccia.*

*Dimostrazione.* Per la proposizione precedente, una matrice idempotente è simile ad una matrice diagonale avente solo 1 e/o 0 sulla diagonale principale e il cui rango è quindi uguale alla sua traccia, cioè al numero degli 1 sulla diagonale principale. Ma matrici simili hanno la stessa traccia e lo stesso rango, quindi per qualsiasi matrice idempotente il rango è uguale alla traccia. $\square$

## A.3 Immagine di una matrice

È noto che una qualsiasi matrice può essere considerata come associata ad un'applicazione lineare e che, quindi, si usa parlare di *immagine* di una matrice; ad esempio, data un'applicazione lineare $L : \mathbb{R}^n \to \mathbb{R}^m$, ad essa può essere associata una matrice $\underset{m,n}{\mathbf{A}}$ tale che, per ogni $\mathbf{v} \in \mathbb{R}^n$, $L(\mathbf{v}) = \mathbf{Av} \in \mathbb{R}^m$. L'immagine di una matrice è quindi l'insieme di tutti i vettori $\mathbf{Av}$, che coincide con l'immagine dell'applicazione associata.

È noto anche che, essendo il prodotto $\mathbf{Av}$ una combinazione lineare delle colonne di $\mathbf{A}$ (di cui gli elementi di $\mathbf{v}$ sono i coefficienti), la dimensione dell'immagine di una matrice è uguale al suo rango e che questo è uguale non solo al numero delle colonne linearmente indipendenti, ma anche al numero delle righe linearmente indipendenti (quindi il rango di una matrice e della sua trasposta sono uguali).

**Proposizione A.23.** *Data una matrice* $\underset{m,p}{\mathbf{A}} : \underset{m,n-p}{\mathbf{B}}$*, cioè una matrice di* $m$ *righe le cui prime* $p$ *colonne siano costituite dalla matrice* $\mathbf{A}$ *e le restanti* $n - p$ *dalla matrice* $\mathbf{B}$*, si ha:*

$$\operatorname{Im}(\mathbf{A} : \mathbf{B}) = \operatorname{Im}(\mathbf{A}) + \operatorname{Im}(\mathbf{B}) \qquad \dim \operatorname{Im}(\mathbf{A} : \mathbf{B}) \leq \dim \operatorname{Im}(\mathbf{A}) + \dim \operatorname{Im}(\mathbf{B})$$

*Dimostrazione.* Segue dalla definizione di immagine di una matrice: l'immagine di $\mathbf{A} : \mathbf{B}$ è lo spazio generato dalle sue colonne ed è quindi lo spazio generato dall'unione delle colonne di $\mathbf{A}$ e di quelle di $\mathbf{B}$, è quindi la somma delle immagini delle due matrici sue componenti.

Inoltre, alcune delle $\mathrm{rk}(\mathbf{A})$ colonne linearmente indipendenti di $\mathbf{A}$ potrebbero risultare linearmente dipendenti da alcune delle $\mathrm{rk}(\mathbf{B})$ colonne linearmente indipendenti di $\mathbf{B}$, e viceversa, da cui la disuguaglianza delle dimensioni. □

**Proposizione A.24.** *Date due matrici* $\mathbf{A}$ *e* $\mathbf{B}$*, l'immagine del prodotto* $\mathbf{AB}$ *è un sottoinsieme dell'immagine di* $\mathbf{A}$*:*

$$\mathrm{Im}(\mathbf{AB}) \subseteq \mathrm{Im}(\mathbf{A})$$

*Dimostrazione.* $\mathbf{ABv} = \mathbf{A}(\mathbf{Bv}) \subseteq \mathrm{Im}(\mathbf{A})$. □

**Proposizione A.25.** *L'immagine di una matrice* $\mathbf{A}$ *è uguale all'immagine del suo prodotto per la sua trasposta e sono uguali anche i ranghi.*

$$\mathrm{Im}(\mathbf{AA}') = \mathrm{Im}(\mathbf{A}) \qquad \mathrm{rk}(\mathbf{AA}') = \mathrm{rk}(\mathbf{A})$$

*Dimostrazione.* Per l'uguaglianza delle immagini si tratta di dimostrare che valgono sia $\mathrm{Im}(\mathbf{A}) \subseteq \mathrm{Im}(\mathbf{AA}')$ che $\mathrm{Im}(\mathbf{AA}') \subseteq \mathrm{Im}(\mathbf{A})$. La seconda inclusione segue dalla proposizione precedente.

Se $\mathbf{v}$ è un vettore appartenente al complemento ortogonale di $\mathrm{Im}(\mathbf{AA}')$, $\mathbf{v}$ appartiene anche al complemento ortogonale di $\mathrm{Im}(\mathbf{A})$:

$$\begin{aligned} \mathbf{v} \in \mathrm{Im}(\mathbf{AA}')^{\perp} \quad &\Rightarrow \quad \mathbf{v}'\mathbf{AA}' = \mathbf{0} \quad \Rightarrow \quad \mathbf{v}'\mathbf{AA}'\mathbf{v} = \mathbf{0} \quad \Rightarrow \quad \|\mathbf{Av}\| = \mathbf{0} \\ &\Rightarrow \quad \mathbf{Av} = \mathbf{0} \quad \Rightarrow \quad \mathbf{v} \in \mathrm{Im}(\mathbf{A})^{\perp} \end{aligned}$$

Ne segue $\mathrm{Im}(\mathbf{AA}')^{\perp} \subseteq \mathrm{Im}(\mathbf{A})^{\perp}$, quindi si ha anche $\mathrm{Im}(\mathbf{A}) \subseteq \mathrm{Im}(\mathbf{AA}')$. L'uguaglianza dei ranghi segue da quella delle immagini. □

**Proposizione A.26.** *Date due matrici* $\mathbf{A}$ *e* $\mathbf{C}$ *con lo stesso numero di righe,* $\mathrm{Im}(\mathbf{C})$ *è un sottospazio di* $\mathrm{Im}(\mathbf{A})$ *solo se* $\mathbf{C} = \mathbf{AB}$*, dove* $\mathbf{B}$ *sia una matrice moltiplicabile per* $\mathbf{A}$ *e con lo stesso numero di colonne di* $\mathbf{C}$*:*

$$\mathrm{Im}(\underset{m,p}{\mathbf{C}}) \subseteq \mathrm{Im}(\underset{m,n}{\mathbf{A}}) \quad \Rightarrow \quad \underset{m,p}{\mathbf{C}} = \underset{m,n}{\mathbf{A}}\ \underset{n,p}{\mathbf{B}}$$

*Dimostrazione.* $\mathrm{Im}(\mathbf{C})$ è lo spazio generato dalle colonne di $\mathbf{C}$. Perché questo sia incluso nell'immagine di $\mathbf{A}$, per ciascuna colonna $\mathbf{c}_i$ di $\mathbf{C}$ deve esservi un vettore $\mathbf{b}_i$ tale che $\mathbf{Ab}_i = \mathbf{c}_i$. Quindi $\mathbf{C} = \{\mathbf{c}_1 : \cdots : \mathbf{c}_p\}$ deve essere uguale a $\mathbf{AB}$ con $\mathbf{B} = \{\mathbf{b}_1 : \cdots : \mathbf{b}_p\}$. □

**Proposizione A.27.** *Date due matrici* $\mathbf{A}$ *e* $\mathbf{B}$*, se* $\mathrm{Im}(\mathbf{B}) \subseteq \mathrm{Im}(\mathbf{A})$ *allora* $\mathbf{AA}^{-}\mathbf{B} = \mathbf{B}$*, quale che sia l'inversa generalizzata di* $\mathbf{A}$*. Analogamente, se* $\mathrm{Im}(\mathbf{B}') \subseteq \mathrm{Im}(\mathbf{A}')$ *allora* $\mathbf{BA}^{-}\mathbf{A} = \mathbf{B}$*.*

*Dimostrazione.* Se $\mathrm{Im}(\mathbf{B}) \subseteq \mathrm{Im}(\mathbf{A})$, per la proposizione precedente esiste una matrice $\mathbf{M}$ tale che $\mathbf{B} = \mathbf{AM}$, quindi:

$$\mathbf{AA}^{-}\mathbf{B} = \mathbf{AA}^{-}\mathbf{AM} = \mathbf{AM} = \mathbf{B}$$

Se invece $\mathrm{Im}(\mathbf{B}') \subseteq \mathrm{Im}(\mathbf{A}')$, esiste una matrice $\mathbf{N}$ tale che $\mathbf{B}' = \mathbf{A}'\mathbf{N}'$ e $\mathbf{B} = (\mathbf{N}')'(\mathbf{A}')' = \mathbf{N}\mathbf{A}$, quindi:

$$\mathbf{B}\mathbf{A}^{-}\mathbf{A} = \mathbf{N}\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{N}\mathbf{A} = \mathbf{B}$$ □

**Proposizione A.28.** *Date tre matrici* $\mathbf{A}, \mathbf{B}, \mathbf{C}$*, si ha* $\mathrm{Im}(\mathbf{B}') \subseteq \mathrm{Im}(\mathbf{A}')$ *e* $\mathrm{Im}(\mathbf{C}) \subseteq \mathrm{Im}(\mathbf{A})$ *se e solo se* $\mathbf{B}\mathbf{A}^{-}\mathbf{C}$ *è invariante rispetto alla scelta dell'inversa generalizzata di* $\mathbf{A}$*.*

*Dimostrazione.* Se $\mathrm{Im}(\mathbf{B}') \subseteq \mathrm{Im}(\mathbf{A}')$ e $\mathrm{Im}(\mathbf{C}) \subseteq \mathrm{Im}(\mathbf{A})$, allora per la proposizione A.26 esistono due matrici $\mathbf{M}$ e $\mathbf{N}$ tali che $\mathbf{B} = \mathbf{N}\mathbf{A}$ e $\mathbf{C} = \mathbf{A}\mathbf{M}$. Se $\mathbf{A}_1^{-}$ e $\mathbf{A}_2^{-}$ sono due inverse generalizzate di $\mathbf{A}$, si ha:

$$\begin{aligned}\mathbf{B}\mathbf{A}_1^{-}\mathbf{C} - \mathbf{B}\mathbf{A}_2^{-}\mathbf{C} &= \mathbf{N}\mathbf{A}\mathbf{A}_1^{-}\mathbf{A}\mathbf{M} - \mathbf{N}\mathbf{A}\mathbf{A}_2^{-}\mathbf{A}\mathbf{M} = \mathbf{N}(\mathbf{A}\mathbf{A}_1^{-}\mathbf{A} - \mathbf{A}\mathbf{A}_2^{-}\mathbf{A})\mathbf{M} \\ &= \mathbf{N}(\mathbf{A} - \mathbf{A})\mathbf{M} = \mathbf{O}\end{aligned}$$

Si può dimostrare anche l'implicazione inversa. □

**Proposizione A.29.** *Il prodotto di due matrici* $\mathbf{A}$ *e* $\mathbf{B}$ *è nullo se e solo se l'immagine dell'una è inclusa nel complemento ortogonale dell'immagine dell'altra:*

$$\mathrm{Im}(\mathbf{B}'\mathbf{A}) = \mathbf{O} \quad \Leftrightarrow \quad \mathrm{Im}(\mathbf{B}) \subseteq \mathrm{Im}(\mathbf{A})^{\perp}$$

*Dimostrazione.* Se $\mathbf{v}$ è un elemento dell'immagine di $\mathbf{B}$, esiste un vettore $\mathbf{u}$ tale che $\mathbf{B}\mathbf{u} = \mathbf{v}$; se $\mathbf{w}$ è un elemento dell'immagine di $\mathbf{A}$, esiste un vettore $\mathbf{x}$ tale che $\mathbf{A}\mathbf{x} = \mathbf{w}$ e si ha:

$$\mathbf{v}'\mathbf{w} = \mathbf{u}'\mathbf{B}'\mathbf{A}\mathbf{x} = 0$$

ovvero $\mathbf{v} \in \mathrm{Im}(\mathbf{A})^{\perp}$. □

**Proposizione A.30.** *Se una matrice* $\mathbf{A}$ *ha* $m$ *righe, allora la dimensione dell'immagine del suo complemento ortogonale è* $m - \mathrm{rk}(\mathbf{A})$*.*

*Dimostrazione.* Si può vedere $\mathbf{A}$ come associata all'applicazione $L : \mathbb{R}^n \to \mathbb{R}^m$. L'immagine di $\mathbf{A}$ è un sottospazio di $\mathbb{R}^m$ di dimensione pari al rango di $\mathbf{A}$; essendo $\mathbb{R}^m = \mathrm{Im}(\mathbf{A}) \oplus \mathrm{Im}(\mathbf{A})^{\perp}$, la dimensione di $\mathrm{Im}(\mathbf{A})^{\perp}$ è $m - \mathrm{rk}(\mathbf{A})$. □

**Proposizione A.31.** *Date due matrici* $\mathbf{A}$ *e* $\mathbf{B}$*, se* $\mathrm{Im}(\mathbf{A}) \subseteq \mathrm{Im}(\mathbf{B})$ *e* $\mathrm{rk}(\mathbf{A}) = \mathrm{rk}(\mathbf{B})$ *allora* $\mathrm{Im}(\mathbf{A}) = \mathrm{Im}(\mathbf{B})$*.*

*Dimostrazione.* Se ciascun elemento di $\mathbf{A}$ è anche elemento di $\mathbf{B}$, ciò vale anche per gli elementi delle basi; poiché l'uguaglianza dei ranghi implica l'uguaglianza delle dimensioni, quindi delle numerosità delle basi, le due immagini hanno le stesse basi, quindi sono uguali. □

**Proposizione A.32.** *Date due matrici* $\mathbf{A}$ *e* $\mathbf{B}$*,* $\mathrm{rk}(\mathbf{A}\mathbf{B}) \leq \min\{\mathrm{rk}(\mathbf{A}), \mathrm{rk}(\mathbf{B})\}$*.*

*Dimostrazione.* Per la proposizione A.24, $\mathrm{Im}(\mathbf{A}\mathbf{B}) \leq \mathrm{Im}(\mathbf{A})$, quindi $\mathrm{rk}(\mathbf{A}\mathbf{B}) \leq \mathrm{rk}(\mathbf{A})$ e, analogamente, $\mathrm{rk}(\mathbf{A}\mathbf{B}) = \mathrm{rk}(\mathbf{B}'\mathbf{A}') \leq \mathrm{rk}(\mathbf{B}') = \mathrm{rk}(\mathbf{B})$. □

**Proposizione A.33.** *Date due matrici* $\mathbf{A}$ *e* $\mathbf{B}$*,* $\mathrm{rk}(\mathbf{A} + \mathbf{B}) \leq \mathrm{rk}(\mathbf{A}) + \mathrm{rk}(\mathbf{B})$*.*

*Dimostrazione.* Si ha:

$$\text{rk}(\mathbf{A}+\mathbf{B}) \leq \text{rk}(\mathbf{A}:\mathbf{B}) \leq \text{rk}(\mathbf{A}) + \text{rk}(\mathbf{B})$$

La prima disuguaglianza vale in quanto $\mathbf{A}+\mathbf{B}$ ha un numero di colonne pari alla metà di quello di $\mathbf{A}:\mathbf{B}$, la seconda per la proposizione A.23. □

Segue un risultato di particolare interesse per i modelli lineari.

## A.4 Proiezione ortogonale sull'immagine di una matrice

**Proposizione A.34.** *Data una matrice* $\mathbf{A}$*, la matrice* $\mathbf{A}\mathbf{A}^-$ *è una matrice di proiezione su* $\text{Im}(\mathbf{A})$*. Inoltre, la matrice di proiezione ortogonale su* $\text{Im}(\mathbf{A})$ *è* $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$*.*

*Dimostrazione.* Sia $\mathbf{A}$ una matrice $n \times p$. $\underset{n,p}{\mathbf{A}}\ \underset{p,n}{\mathbf{A}^-}$ è una matrice di proiezione su $\text{Im}(\mathbf{A}) \subseteq \mathbb{R}^n$ in quanto:

a) dato un vettore $\mathbf{v}$, per la proposizione A.24 $\text{Im}(\mathbf{A}\mathbf{A}^-) \subseteq \text{Im}(\mathbf{A})$, quindi:

$$(\mathbf{A}\mathbf{A}^-)\mathbf{v} \in \text{Im}(\mathbf{A})$$

b) se $\mathbf{v}$ stesso appartiene a $\text{Im}(\mathbf{A})$, esiste un $\mathbf{x}$ tale che $\mathbf{v} = \mathbf{A}\mathbf{x}$, quindi:

$$(\mathbf{A}\mathbf{A}^-)\mathbf{v} = \mathbf{A}\mathbf{A}^-\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x} = \mathbf{v}$$

Quanto a $\underset{n,p}{\mathbf{A}}(\underset{p,n}{\mathbf{A}'}\ \underset{n,p}{\mathbf{A}})^-\underset{p,n}{\mathbf{A}'}$, per la proposizione A.25 e per la simmetria di $\mathbf{A}'\mathbf{A}$:

$$\text{Im}(\mathbf{A}') = \text{Im}(\mathbf{A}'\mathbf{A}) = \text{Im}[(\mathbf{A}'\mathbf{A})']$$

e, per la proposizione A.27:

$$\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{A} = \mathbf{A}$$

Quindi $(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ è un'inversa generalizzata di $\mathbf{A}$ e $\mathbf{P} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ è una matrice di proiezione. Per un qualsiasi vettore $\mathbf{v} \in \text{Im}(\mathbf{A}) \subseteq \mathbb{R}^n$ esiste un $\mathbf{x} \in \mathbb{R}^p$ tale che $\mathbf{A}\mathbf{x} = \mathbf{v}$; se $\mathbf{y} \in \text{Im}(\mathbf{A})^\perp$, $\mathbf{v}'\mathbf{y} = (\mathbf{A}\mathbf{x})'\mathbf{y} = \mathbf{x}'\mathbf{A}'\mathbf{y} = 0$, ovvero $\mathbf{A}'\mathbf{y} = \mathbf{0}$, quindi:

$$\mathbf{P}\mathbf{y} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{y} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{0} = \mathbf{0} \qquad (\mathbf{I}_n - \mathbf{P})\mathbf{y} = \mathbf{y}$$

Inoltre, per qualsiasi vettore $\mathbf{v} \in \mathbb{R}^n$ si ha, ancora per la proposizione A.27:

$$\mathbf{A}'(\mathbf{I}_n - \mathbf{P})\mathbf{v} = [\mathbf{A}' - \mathbf{A}'\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}']\mathbf{v} = [\mathbf{A}' - \mathbf{A}']\mathbf{v} = \mathbf{0} \quad \Rightarrow \quad (\mathbf{I}_n - \mathbf{P})\mathbf{v} \in \text{Im}(\mathbf{A})^\perp$$

Quindi $\mathbf{P}$ è la matrice di proiezione ortogonale su $\in (\mathbf{A})$. □

Se $\mathbf{A}$ è una matrice di riparametrizzazione a rango pieno, $\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ è la matrice hat $\mathbf{H}$, che è appunto la matrice di proiezione ortogonale su $\text{Im}(\mathbf{A})$.

# Appendice B

# Equazioni alle differenze

## B.1 Equazioni alle differenze del primo ordine

**Definizione B.1.** Se $y$ è una variabile che assume valori diversi nel tempo, indicando con $y_t$ il suo valore al tempo $t$, si dice *equazione alle differenze lineare del primo ordine* un'equazione del tipo:

$$y_t = \phi y_{y-1} + w_t \tag{B.1}$$

in cui $w_t$ è un termine di una successione $\{w_0, w_1, w_2, \dots\}$.

Un'equazione alle differenze descrive lo stato di un sistema al variare del tempo.

**Definizione B.2.** Data un'equazione alle differenze lineare del primo ordine, si dice *moltiplicatore dinamico* l'effetto di un cambiamento di $w_t$ sul valore di $y_{t+j}$.

**Proposizione B.3.** *Il moltiplicatore dinamico di un'equazione alle differenze lineare del primo ordine dipende solo dal coefficiente $\phi$ e dal numero $j$ di periodi compresi tra $t$ e $t + j$.*

*Dimostrazione.* Per determinare l'effetto su $y_t$ di un cambiamento nel valore di $w_0$ si può adottare una *sostituzione ricorsiva*; ipotizzando dato $y_{-1}$:

$$\begin{aligned} y_0 &= \phi y_{-1} + w_0 \\ y_1 &= \phi y_0 + w_1 \\ y_2 &= \phi y_1 + w_2 \\ &\vdots \\ y_t &= \phi y_{t-1} + w_t \end{aligned}$$

da cui:

$$\begin{aligned} y_t &= \phi y_{t-1} + w_t = \phi(\phi y_{t-2} + w_{t-1}) + w_t \\ &= \phi^2 y_{t-2} + \phi w_{t-1} + w_t = \phi^2(\phi y_{t-3} + w_{t-2}) + \phi w_{t-1} + w_t \\ &= \phi^3 y_{t-3} + \phi^2 w_{t-2} + \phi w_{t-1} + w_t \\ &\dots \\ &= \phi^{t+1} y_{-1} + \phi^t w_0 + \phi^{t-1} w_1 + \phi^{t-2} w_2 + \cdots + \phi w_{t-1} + w_t \end{aligned}$$

e quindi il moltiplicatore dinamico è:

$$\frac{\partial y_t}{\partial w_0} = \phi^t$$

Generalizzando, se si fosse partiti dal tempo $t$ per arrivare a $y_{t+j}$, ipotizzando dato $y_{t-1}$, si sarebbe avuto:

$$y_{t+j} = \phi^{j+1} y_{t-1} + \phi^j w_t + \phi^{j-1} w_{t+1} + \cdots + \phi w_{t+j-1} + w_{t+j}$$

quindi:

$$\frac{\partial y_{t+j}}{\partial w_t} = \phi^j$$ □

**Proposizione B.4.** *Se $|\phi| < 1$ il moltiplicatore dinamico tende a zero, tende invece a infinito se $|\phi| > 1$. Se $\phi = 1$ allora $y_{t+j}$ è la somma di $y_{t-1}$ e degli $j+1$ termini $w_t, \ldots, w_{t+j}$.*

*Dimostrazione.* Segue dalle proprietà della funzione potenza. □

Il limite del moltiplicatore dinamico per $t \to \infty$ esprime la stabilità, o meno, del sistema descritto da un'equazione alle differenze: se il moltiplicatore tende a zero il sistema è *stabile* (l'impulso iniziale viene progressivamente smorzato), se tende a infinito il sistema è *esplosivo* (l'impulso iniziale viene sempre più amplificato).

## B.2 Equazioni alle differenze di ordine $p$

**Definizione B.5.** Se $y$ è una variabile che assume valori diversi nel tempo, indicando con $y_t$ il suo valore al tempo $t$, si dice *equazione alle differenze lineare di ordine $p$* un'equazione del tipo:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + w_t \tag{B.2}$$

in cui $w_t$ è un termine di una successione $\{w_0, w_1, w_2, \ldots\}$.

Risulta comodo riscrivere la (B.2) in forma matriciale. Ponendo:

$$\boldsymbol{\xi}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \qquad \mathbf{v}_t = \begin{bmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Si può scrivere:

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t \tag{B.3}$$

ovvero:

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-(p-1)} \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Si tratta di un sistema di $p$ equazioni, la prima delle quali è la (B.2), la seconda semplicemente $y_{t-1} = t_{t-1}$, la terza $y_{t-2} = y_{t-2}$, la $p$-esima $y_{t-p+1} = y_{t-p+1}$.

**Proposizione B.6.** *Il moltiplicatore dinamico di un'equazione alle differenze lineare di ordine $p$ dipende solo dalla matrice $\mathbf{F}$ e dal numero $j$ di periodi compresi tra $t$ e $t+j$.*

*Dimostrazione.* Procedendo ricorsivamente come nella dimostrazione della proposizione B.3, si ottiene:

$$\boldsymbol{\xi}_t = \mathbf{F}^{t+1}\boldsymbol{\xi}_{-1} + \mathbf{F}^t\mathbf{v}_0 + \mathbf{F}^{t-1}\mathbf{v}_1 + \mathbf{F}^{t-2}\mathbf{v}_2 + \cdots + \mathbf{F}\mathbf{v}_{t-1} + \mathbf{v}_t$$

ovvero:

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \mathbf{F}^{t+1}\begin{bmatrix} y_{-1} \\ y_{-2} \\ y_{-3} \\ \vdots \\ y_{-p} \end{bmatrix} + \mathbf{F}^t\begin{bmatrix} w_0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \mathbf{F}^{t-1}\begin{bmatrix} w_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + \mathbf{F}^1\begin{bmatrix} w_{t-1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Indicando con $f_{rc}^{(k)}$ l'elemento $(r, c)$ della matrice $\mathbf{F}^k$, la prima equazione di tale sistema è:

$$\begin{aligned} y_t = {} & f_{11}^{(t+1)}y_{-1} + f_{12}^{(t+1)}y_{-2} + \cdots + f_{1p}^{(t+1)}y_{-p} \\ & + f_{11}^{(t)}w_0 + f_{11}^{(t-1)}w_1 + \cdots + f_{11}^{(1)}w_{t-1} + w_t \end{aligned}$$

Generalizzando:

$$\boldsymbol{\xi}_{t+j} = \mathbf{F}^{j+1}\boldsymbol{\xi}_{t-1} + \mathbf{F}^j\mathbf{v}_t + \mathbf{F}^{j-1}\mathbf{v}_{t+1} + \mathbf{F}^{j-2}\mathbf{v}_{t+2} + \cdots + \mathbf{F}\mathbf{v}_{t+j-1} + \mathbf{v}_{t+j}$$

da cui:

$$\begin{aligned} y_{t+j} = {} & f_{11}^{(j+1)}y_{t-1} + f_{12}^{(j+1)}y_{t-2} + \cdots + f_{1p}^{(j+1)}y_{t-p} \\ & + f_{11}^{(j)}w_t + f_{11}^{(j-1)}w_{t+1} + \cdots + f_{11}^{(1)}w_{t+j-1} + w_{t+j} \end{aligned}$$

Il moltiplicatore dinamico è quindi:

$$\frac{\partial y_{t+j}}{\partial w_t} = f_{11}^{(j)}$$

□

**Proposizione B.7.** *Se la matrice $\mathbf{F}$ è diagonalizzabile, lo scalare $f_{11}^{(j)}$ è una media ponderata dei suoi autovalori, ciascuno elevato alla $j$-esima potenza.*

*Dimostrazione.* Se $\mathbf{F}$ è diagonalizzabile, si ha $\mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1}$ dove $\mathbf{T}$ è una matrice invertibile le cui colonne sono gli autovettori di $\mathbf{F}$ e $\boldsymbol{\Lambda}$ è una matrice diagonale con elementi i relativi autovalori. Inoltre, $\mathbf{F}^j = \mathbf{T}\boldsymbol{\Lambda}^j\mathbf{T}^{-1}$, in quanto:

$$\begin{aligned} \mathbf{F}^j &= \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1}\mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1}\cdots\mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1} \\ &= \mathbf{T}\boldsymbol{\Lambda}\boldsymbol{\Lambda}\cdots\boldsymbol{\Lambda}\mathbf{T}^{-1} \\ &= \mathbf{T}\boldsymbol{\Lambda}^j\mathbf{T}^{-1} \end{aligned}$$

Indicando con $t_{ij}$ il generico elemento di $\mathbf{T}$, con $t^{ij}$ quello di $\mathbf{T}^{-1}$:

$$\mathbf{F}^j = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1p} \\ t_{21} & t_{22} & \cdots & t_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ t_{p1} & t_{p2} & \cdots & t_{pp} \end{bmatrix} \begin{bmatrix} \lambda_1^j & 0 & \cdots & 0 \\ 0 & \lambda_2^j & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_p^j \end{bmatrix} \begin{bmatrix} t^{11} & t^{12} & \cdots & t^{1p} \\ t^{21} & t^{22} & \cdots & t^{2p} \\ \vdots & \vdots & \cdots & \vdots \\ t^{p1} & t^{p2} & \cdots & t^{pp} \end{bmatrix}$$
$$= \begin{bmatrix} t_{11}\lambda_1^j & t_{12}\lambda_2^j & \cdots & t_{1p}\lambda_p^j \\ t_{21}\lambda_1^j & t_{22}\lambda_2^j & \cdots & t_{2p}\lambda_p^j \\ \vdots & \vdots & \cdots & \vdots \\ t_{p1}\lambda_1^j & t_{p2}\lambda_2^j & \cdots & t_{pp}\lambda_p^j \end{bmatrix} \begin{bmatrix} t^{11} & t^{12} & \cdots & t^{1p} \\ t^{21} & t^{22} & \cdots & t^{2p} \\ \vdots & \vdots & \cdots & \vdots \\ t^{p1} & t^{p2} & \cdots & t^{pp} \end{bmatrix}$$

Pertanto, l'elemento $(1,1)$ di $\mathbf{F}^j$ è:

$$f_{11}^{(j)} = (t_{11}t^{11})\lambda_1^j + (t_{12}t^{21})\lambda_2^j + \cdots + (t_{1p}t^{p1})\lambda_p^j$$

Dal momento che $\mathbf{T}\mathbf{T}^{-1} = \mathbf{I}$, la somma dei prodotti $t_{ij}t^{ji}$ è uguale a 1. Si ha quindi:

$$f_{11}^{(j)} = c_1\lambda_1^j + c_2\lambda_2^j + \cdots + \lambda_p^j \qquad c_1 + \cdots + c_p = 1$$

con $c_i = t_{1i}t^{i1}$. □

**Proposizione B.8.** *Se la matrice* $\mathbf{F}$ *è diagonalizzabile, i pesi* $c_i$ *sono dati da:*

$$c_i = \frac{\lambda_i^{p-1}}{\prod_{k=1,k\neq i}^{p} (\lambda_i - \lambda_k)}$$

Se gli autovalori di $\mathbf{F}$ sono tutti reali, il sistema è stabile se essi sono tutti minori di uno in valore assoluto, esplosivo se almeno uno è maggiore di 1 in valore assoluto.

**Esempio B.9.** La funzione R `lde.dm()`, proposta nella figura B.1, accetta come argomenti un vettore di coefficienti $\phi_i$ e, opzionalmente, un numero di ritardi per default pari a 40. Un terzo parametro opzionale consente di evitare la produzione del grafico dei moltiplicatori dinamici al crescere dei ritardi. La funzione calcola gli autovalori, il loro valore assoluto e i coefficienti $c_i$. Se l'equazione è:

$$y_t = 0.6y_{t-1} + 0.2y_{t-2} + w_t$$

si ottiene:

```
> lde.dm(c(0.6, 0.2))
$lambda
[1]  0.8385165 -0.2385165
$mod
[1] 0.8385165 0.2385165
$c
[1] 0.778543 0.221457
```

```
lde.dm <- function(phi, j=40, plot=TRUE) {
  stopifnot(is.numeric(phi))
  p <- length(phi)
  F <- diag(1, nrow=p-1)
  F <- cbind(F, rep(0, p-1))
  F <- rbind(phi, F)
  eig <- eigen(F)
  lambda <- eig$values
  mod <- Mod(lambda)
  T <- eig$vectors;
  T1 <- solve(T)
  c <- numeric(p)
  for (i in 1:p)
    c[i] <- T[1,i] * T1[i,1]
  if (plot) {
    f11 <- numeric(j)
    for (i in 1:j)
      suppressWarnings(f11[i] <- as.real(sum(c * lambda^i)))
    plot(f11, type="h", lwd=5, xlab="Lag", ylab="Multiplier",
      main=paste("phi", 1:p, " = ", phi, sep="", collapse=", "))
  }
  return(list(lambda=lambda, mod=mod, c=c))
}
```

*Figura B.1.* Funzione `lde.dm()`.

Gli autovalori sono tutti minori di 1 in valore assoluto, quindi il sistema è stabile (figura B.2, primo grafico dall'alto). Se invece l'equazione è:

$$y_t = 0.6y_{t-1} + 0.8y_{t-2} + w_t$$

si ottiene:

```
> lde.dm(c(0.6, 0.8))
$lambda
[1]  1.2433981 -0.6433981
$mod
[1] 1.2433981 0.6433981
$c
[1] 0.6589997 0.3410003
```

Ora un autovalore è maggiore di 1, quindi il sistema è esplosivo (figura B.2, secondo grafico dall'alto).

Se alcuni autovalori di $\mathbf{F}$ sono complessi, essi compaiono a coppie (un complesso e il suo coniugato). Per elevare a potenza un autovalore complesso, lo si scrive nella forma:

$$\lambda_i = r[\cos(\theta) + i \cdot \sin(\theta)]$$

dove $r = |\lambda_i|$, e si ha:

$$\lambda_i^j = r^j[\cos(j\theta) + i \cdot \sin(j\theta)]$$

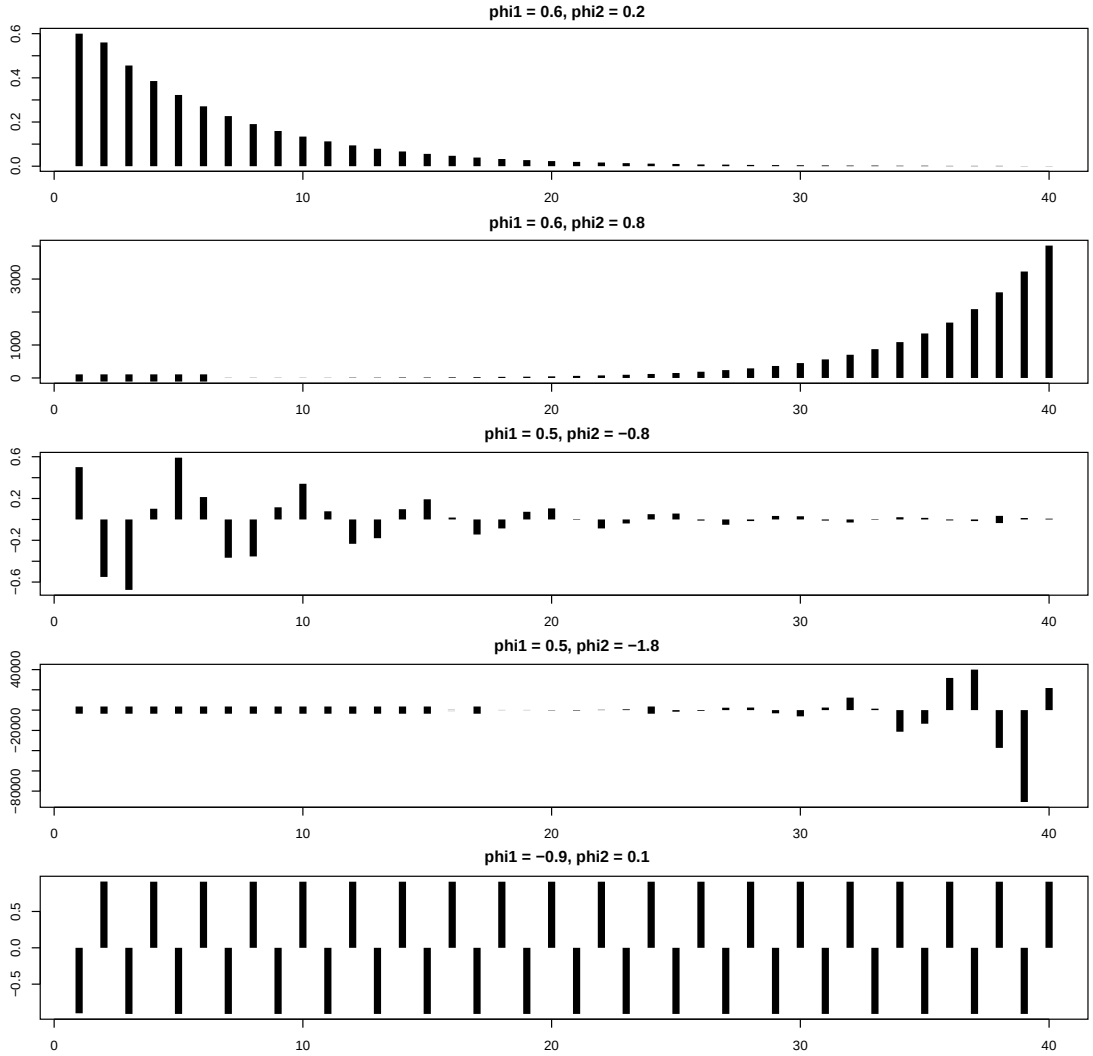


*Figura B.2.* Moltiplicatori dinamici tendenti a zero, all'infinito o periodici al variare dei coefficienti $\phi_i$.

il cui coniugato è:

$$\lambda_{i+1}^{j} = r^{j}[\cos(j\theta) - i \cdot \sin(j\theta)]$$

Il moltiplicatore dinamico diventa, nel caso di un'equazione del secondo ordine con due autovalori complessi coniugati:

$$\begin{aligned}\frac{\partial y_{t+j}}{\partial w_t} &= c_1\lambda_1^j + c_2\lambda_2^j = c_1 r^j[cos(j\theta) + i \cdot \sin(j\theta)] + c_1 r^j[cos(j\theta) - i \cdot \sin(j\theta)] \\ &= (c_1 + c_2)r^j \cos(j\theta) + i \cdot (c_1 - c_2)r^j \sin(j\theta)\end{aligned}$$

Per la proposizione B.8, se $\lambda_1$ e $\lambda_2$ sono complessi coniugati sono tali anche $c_1$ e $c_2$, ovvero:

$$c_1 = \alpha + \beta i \qquad\qquad c_2 = \alpha - \beta i$$

Il moltiplicatore dinamico è quindi un numero reale:

$$\begin{aligned} c_1\lambda_1^j + c_2\lambda_2^j &= [(\alpha+\beta i)+(\alpha-\beta i)]r^j\cos(j\theta) + i\cdot[(\alpha+\beta i)-(\alpha-\beta i)]r^j\sin(j\theta) \\ &= [2\alpha]r^j\cos(j\theta) + i\cdot[2\beta i]r^j\sin(j\theta) \\ &= 2\alpha r^j\cos(j\theta) - 2\beta r^j\sin(j\theta) \end{aligned}$$

Come nel caso di autovalori reali si guarda al loro valore assoluto, nel caso di autovalori complessi di guarda al loro modulo $r = |\lambda_i|$: se il modulo maggiore è minore di 1 il sistema è stabile, se è maggiore di 1 il sistema è esplosivo.

**Esempio B.10.** Se l'equazione è:

$$y_t = 0.5y_{t-1} - 0.8y_{t-2} + w_t$$

con la funzione `lde.dm()` si ottiene:

```
> lde.dm(c(0.5,-0.8))
$lambda
[1] 0.25+0.8587782i 0.25-0.8587782i
$mod
[1] 0.8944272 0.8944272
$c
[1] 0.5-0.1455556i 0.5+0.1455556i
```

Il modulo dei due autovalori è minore di 1, quindi il sistema è stabile (figura B.2, terzo grafico dall'alto). Se invece l'equazione è:

$$y_t = 0.5y_{t-1} - 1.8y_{t-2} + w_t$$

si ottiene:

```
> lde.dm(c(0.5,-1.8))
$lambda
[1] 0.25+1.318143i 0.25-1.318143i
$mod
[1] 1.341641 1.341641
$c
[1] 0.5-0.0948304i 0.5+0.0948304i
```

Ora il modulo degli autovalori è maggiore di 1 e il sistema esplode (figura B.2, quarto grafico dall'alto).

Se il modulo maggiore degli autovalori, reali o complessi, è uguale a 1 il moltiplicatore dinamico è periodico.

**Esempio B.11.** Se l'equazione è:

$$y_t = -0.9y_{t-1} + 0.1y_{t-2} + w_t$$

si ottiene:

```
> lde.dm(c(-0.9,0.1))
$lambda
[1] -1.0  0.1
$mod
[1] 1.0 0.1
$c
[1] 0.9090909 0.0909091
```

Gli autovalori sono reali, il maggior valore assoluto è uguale a 1 e i moltiplicatori mostrano un andamento periodico (figura B.2, primo grafico dal basso).

Infine, se la matrice $\mathbf{F}$ non è diagonalizzabile i risultati precedenti possono essere generalizzati usando la decomposizione di Jordan:

$$\mathbf{F} = \mathbf{M}\mathbf{J}\mathbf{M}^{-1} \qquad \mathbf{F}^j = \mathbf{M}\mathbf{J}^j\mathbf{M}^{-1}$$

dove $\mathbf{J}$ è la forma canonica di Jordan della matrice $\mathbf{F}$.

# Appendice C

# Richiami di probabilità e di statistica

## C.1 Variabili aleatorie multidimensionali

Dato uno spazio campionario $\Omega$, se ad ogni evento elementare $\omega \in \Omega$ viene associata una $n$-upla di numeri reali $(X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$, si ha una *variabile aleatoria n-dimensionale*.

In generale, la *funzione di ripartizione* di una variabile aleatoria multidimensionale $\mathbf{X}$ è:

$$F_{\mathbf{X}}(\mathbf{x}) = P[X_1 < x_1, X_2 < x_2, \ldots, X_n < n] \qquad \forall \mathbf{x} \in \mathbb{R}^n$$

Nel caso di una variabile aleatoria doppia $(X, Y)$ assolutamente continua:

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v)\, du\, dv$$

dove $f_{X,Y}(x, y)$ è la *funzione di densità* della v.a. con funzione di ripartizione $F(x, y)$.

Considerando la sola componente $X$ di una v.a. doppia $(X, Y)$, la sua funzione di ripartizione è

$$F_X(x) = P[X < x] = P[X < x, Y < +\infty] = \int_{-\infty}^{x} du \int_{-\infty}^{+\infty} f_{X,Y}(u, v)\, dv$$

Poiché in generale, per una v.a. assolutamente continua, $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$, le *funzioni di densità marginale* delle componenti di $(X, Y)$ sono

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dy \qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dx$$

Sempre nel caso di una v.a. doppia $(X, Y)$, la *densità condizionata* di $X$ dato l'evento $Y = y$ è

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

## C.2 Aspettativa condizionata

Date due variabili aleatorie assolutamente continue $X$ e $Y$ definite nello stesso spazio di probabilità, $\mathbb{E}[X \mid Y]$ è anch'essa una variabile aleatoria, detta *aspettativa* (o *media*) *condizionata*, e assume i valori:

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x \mid y)\, dx = \int_{-\infty}^{+\infty} x \frac{f_{X,Y}(x,y)}{f_Y(y)} dx$$

Da notare che, mentre $\mathbb{E}[X \mid Y]$ è una variabile aleatoria, $\mathbb{E}[X \mid Y = y]$ è un numero (il valore atteso di $X$ *dato* $Y = y$). In altri termini, $\mathbb{E}[X \mid Y]$ è una variabile aleatoria in quanto funzione della variabile aleatoria $Y$.

L'aspettativa condizionata gode delle seguenti proprietà:

a) se $a$ è una qualsiasi costante, $\mathbb{E}[a \mid Y] = a$;

b) $\mathbb{E}[aX + bZ \mid Y] = a\mathbb{E}[X \mid Y] + b\mathbb{E}[Z \mid Y]$ (linearità);

c) $\mathbb{E}[X \mid Y] \geq 0$ se $X \geq 0$;

d) $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$ se $X$ e $Y$ sono indipendenti, infatti in questo caso:

$$\begin{aligned}\mathbb{E}[X \mid Y = y] &= \int_{-\infty}^{+\infty} x \frac{f_{X,Y}(x,y)}{f_Y(y)} dx \overset{ind}{=} \int_{-\infty}^{+\infty} x \frac{f_X(x) f_Y(y)}{f_Y(y)} dx \\ &= \int_{-\infty}^{+\infty} x\, f_X(x) dx = \mathbb{E}[X]\end{aligned}$$

e) $\mathbb{E}[g(Y)X \mid Y] = g(Y)\mathbb{E}[X \mid Y]$, in particolare $\mathbb{E}[g(Y) \mid Y] = g(Y)$; infatti, dato $Y = y$ è dato anche $g(y)$, i valori di $g(Y)X$ sono $g(y)x$ e variano al variare di $x$:

$$\begin{aligned}\mathbb{E}[g(Y)X \mid Y = y] &= \int_{-\infty}^{\infty} g(y) x f_{X|Y}(x \mid y)\, dx \\ &= g(y) \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\, dx = g(y)\mathbb{E}[X \mid Y = y]\end{aligned}$$

Sono inoltre particolarmente importanti le leggi dell'aspettativa totale e della varianza totale.

### C.2.1 Legge dell'aspettativa totale (LTE)

La *legge dell'aspettativa totale* (*LTE*, *Law of Total Expectation*) stabilisce che:

$$\mathbb{E}\Big[\mathbb{E}[X \mid Y]\Big] = \mathbb{E}[X]$$

Infatti,

$$\mathbb{E}\Big[\mathbb{E}[X \mid Y]\Big] = \int_{-\infty}^{+\infty} \mathbb{E}[X \mid Y = y] f_Y(y)\, dy$$

dove

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x \mid y)\, dx$$

Si ha quindi:

$$\mathbb{E}[X] = \int x f_X(x)\, dx$$

vedendo $f_X$ come marginale di una congiunta:

$$
\begin{aligned}
&= \int x \left( \int f_{X,Y}(x, y)\, dy \right) dx \\
&= \int \int x f_{X|Y}(x \mid y) f_Y(y)\, dx\, dy \\
&= \int \left( \int x f_{X|Y}(x \mid y) dx \right) dy \\
&= \int \mathbb{E}[X \mid Y = y] f_Y(y)\, dy = \mathbb{E}\Big[\mathbb{E}[X \mid Y]\Big]
\end{aligned}
$$

Informalmente, la legge dice che il valore atteso totale di $X$ è uguale alla somma dei valori attesi di $X \mid Y$ per tutti i diversi possibili valori che $Y$ può assumere, ciascuno ponderato con la propria probabilità.

### C.2.2 Legge della varianza totale (LTV)

La *legge della varianza totale* (*LTV*, *Law of Total Variance* stabilisce che:

$$\mathbb{V}[X] = \mathbb{E}\Big[\mathbb{V}[X \mid Y]\Big] + \mathbb{V}\Big[\mathbb{E}[X \mid Y]\Big]$$

Infatti:

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \overset{LTE}{=} \mathbb{E}\Big[\mathbb{E}[X^2 \mid Y]\Big] - \mathbb{E}\Big[\mathbb{E}[X \mid Y]\Big]^2$$

riscrivendo il momento secondo in termini della varianza e del momento primo:

$$= \mathbb{E}\Big[\mathbb{V}[X \mid Y] + \mathbb{E}[X \mid Y]^2\Big] - \mathbb{E}\Big[\mathbb{E}[X \mid Y]\Big]^2$$

per la linearità di $\mathbb{E}$:

$$
\begin{aligned}
&= \mathbb{E}\Big[\mathbb{V}[X \mid Y]\Big] + \Big(\mathbb{E}\Big[\mathbb{E}[X \mid Y]^2\Big] - \mathbb{E}\Big[\mathbb{E}[X \mid Y]\Big]^2\Big) \\
&= \mathbb{E}\Big[\mathbb{V}[X \mid Y]\Big] + \mathbb{V}\Big[\mathbb{E}[X \mid Y]\Big]
\end{aligned}
$$

## C.3 La funzione caratteristica di una variabile aleatoria

Data una variabile aleatoria $X$ con funzione di ripartizione $F_X(x) = P[X < x]$, la *funzione caratteristica* della v.a. $X$ è una funzione $\Phi_X : \mathbb{R} \to \mathbb{C}$ definita da:[1]

$$\Phi_X(t) = \mathbb{E}[e^{itX}] = \int_{-\infty}^{+\infty} e^{itx} f_X(x)\, dx$$

Esiste una corrispondenza biunivoca tra la funzione caratteristica e la funzione di ripartizione di una qualsiasi variabile aleatoria. La funzione caratteristica presenta, tra altre, le seguenti proprietà:

[1] La definizione vale solo nel caso la funzione di densità di $X$ esista; in caso contrario, si deve ricorrere a un integrale di Riemann-Stieltjes: $\int_\Omega e^{itx} dF_X(x)$.

a) la f.c. è sempre minore o uguale a 1; è uguale a 1 per $t = 0$:

$$|\Phi_X(t)| \leq 1 \qquad \Phi_X(0) = 1$$

b) la sua derivata $n$-esima in 0 è uguale al momento $n$-esimo di $X$ moltiplicato per $i^n$:

$$\frac{d^n}{dt^n}\Phi_X(t)\Big|_{t=0} \equiv \Phi_X^{(n)}(0) = i^n \mathbb{E}[X^n]$$

c) la f.c. di una trasformazione affine di $X$, $Z = aX + b$, è

$$\Phi_Z(t) = \mathbb{E}[e^{itZ}] = \mathbb{E}[e^{itaX}e^{itb}] = e^{itb}\Phi_X(at)$$

d) la f.c. di una v.a. degenere $C$, con $P[C = c] = 1$, è

$$\Phi_C(t) = \mathbb{E}[e^{itc}] = e^{itc}$$

e) la f.c. di una v.a. normale $X \sim N(\mu, \sigma^2)$ è

$$\Phi_X(t) = e^{\mu it - \frac{t^2\sigma^2}{2}}$$

quindi quella di una normale standard $Z \sim N(0, 1)$ è

$$\Phi_Z(t) = e^{-\frac{t^2}{2}}$$

## C.4 Successioni di variabili aleatorie

### C.4.1 Convergenza in distribuzione e in probabilità

Si dice che una successione $X_n$ di variabili aleatorie con funzioni di ripartizione $F_n$ *converge in distribuzione* alla v.a. $X$ con f.r. $F$, e si scrive $X_n \xrightarrow{d} X$, se esiste il limite

$$\lim_{n\to\infty} F_n(x) = F(x)$$

Dal momento che $F_X(x) = P[X < x]$, la convergenza in distribuzione comporta che, al crescere di $n$, la probabilità che la successione assuma valori minori di $x$ diventa sempre più simile alla probabilità che $X$ assuma valori minori di $x$, ma non che $X_n$ e $X$ tendano ad assumere gli stessi valori.

Si dice invece che una successione $X_n$ di variabili aleatorie *converge in probabilità* alla variabile aleatoria $X$, e si scrive $X_n \xrightarrow{p} X$ oppure $\operatorname{plim} X_n = X$, se

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} P[|X_n - X| < \varepsilon] = 1$$

oppure, equivalentemente, se

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} P[|X_n - X| > \varepsilon] = 0$$

La convergenza in probabilità riguarda quindi i valori assunti da $X_n$ e $X$. Se vale, si ha che all'aumentare di $n$ aumenta sempre più la probabilità che i valori assunti da $X_n$ e quelli assunti da $X$ differiscano meno di un $\varepsilon$, per quanto piccolo sia $\varepsilon$.[2]

[2] Si dice anche che una successione $X_n$ di variabili aleatorie *converge quasi certamente* alla v.a. $X$, e si scrive $X_n \xrightarrow{q.c.} X$, se

$$P[\lim_{n\to\infty} X_n = X] = 1$$

Se vale la convergenza quasi certa, all'aumentare di $n$ le variabili $X_n$ e $X$ tendono a differire solo per eventi di probabilità nulla.

La convergenza in probabilità implica la convergenza in distribuzione:

$$X_n \xrightarrow{p} X \quad \Rightarrow \quad X_n \xrightarrow{d} X$$

mentre la convergenza in distribuzione implica quella in probabilità solo nel caso di convergenza in distribuzione ad una variabile aleatoria degenere:

$$X_n \xrightarrow{d} c \quad \Rightarrow \quad X_n \xrightarrow{p} c$$

Inoltre:

**Teorema di Slutsky.** *Se $X_n$ e $Y_n$ sono due successioni di variabili aleatorie tali che $X_n$ converge in distribuzione ad una variabile aleatoria $X$, $X_n \xrightarrow{d} X$, e $Y_n$ converge in probabilità ad una costante reale $c$, $Y_n \xrightarrow{p} c$, allora:*

- $X_n + Y_n \xrightarrow{d} X + c$;
- $X_n Y_n \xrightarrow{d} cX$;
- $X_n / Y_n \xrightarrow{d} X/c$, *se* $c \neq 0$.

**Lemma di Slutsky.** *Date una successione $X_n$ di variabili aleatorie $k$-dimensionali e una funzione $g : \mathbb{R}^k \to \mathbb{R}$ è una funzione continua in $X \in \mathbb{R}^k$, se $X_n \xrightarrow{p} X$ allora $g(X_n) \xrightarrow{p} g(X)$:*

$$X_n \xrightarrow{p} X \quad \Rightarrow \quad g(X_n) \xrightarrow{p} g(X)$$

### C.4.2 La legge dei grandi numeri

**Teorema C.1** (Legge dei grandi numeri, *LLN* (Law of Large Numbers))**.** *Data una successione $X_n$ di variabili aleatorie indipendenti e identicamente distribuite, con $\mathbb{E}[X] = \mu < \infty$, indicando con $S_n$ la somma dei primi $n$ termini e con $\bar{X}_n = \dfrac{S_n}{n}$ la loro media, si ha:*

$$\bar{X_n} \xrightarrow{p} \mu \qquad \textit{ovvero} \qquad \forall \varepsilon > 0,\ \lim_{n\to\infty} P\left[\left|\bar{X}_n - \mu\right| < \varepsilon\right] = 1$$

*Dimostrazione.* La LLN si dimostra agevolmente nel caso si assuma anche $\mathbb{V}[X] < \infty$. Infatti, la funzione caratteristica di $S_n = \sum_{j=1}^{n} X_j$ è

$$\begin{aligned}
\Phi_{S_n}(t) &= \mathbb{E}[e^{itX_1 + itX_2 + \cdots + itX_n}] = \mathbb{E}[e^{itX_1} e^{itX_2} \cdots e^{itX_n}] \\
&\overset{ind}{=} \mathbb{E}[e^{itX_1}]\mathbb{E}[e^{itX_2}] \cdots \mathbb{E}[e^{itX_n}] \\
&\overset{id}{=} \mathbb{E}[e^{itX}]^n = \Phi_X(t)^n
\end{aligned}$$

Passando per il logaritmo, $\log \Phi_X(t)^n = n \log \Phi_X(t)$. Sviluppando $\log \Phi_X(t)$ secondo Taylor nell'intorno di 0:

$$\begin{aligned}
\log \Phi_X(t)\Big|_{t=0} &= \log \Phi_X(0) + \frac{\Phi_X^{(1)}(0)}{\Phi_X(0)} t + \frac{\Phi_X(0)\Phi_X^{(2)}(0) - \Phi_X^{(1)}(0)^2}{\Phi_X(0)^2} \frac{t^2}{2} + R(t)t^3 \\
&= 0 + i\mathbb{E}[X]t + (i^2\mathbb{E}[X^2] - (i\mathbb{E}[X])^2)\frac{t^2}{2} + R(t)t^3 \\
&= i\mathbb{E}[X]t + (-\mathbb{E}[X^2] + \mathbb{E}[X]^2)\frac{t^2}{2} + R(t)t^3 \\
&= i\mathbb{E}[X]t - \mathbb{V}[X]\frac{t^2}{2} + R(t)t^3
\end{aligned}$$

Si ha quindi:

$$\log \Phi_{S_n}(t) = \log \Phi_X(t)^n = n\left(i\mathbb{E}[X]t - \mathbb{V}[X]\frac{t^2}{2} + R(t)t^3\right)$$

Poiché $\Phi_{aX}(t) = \Phi_X(at)$, per $\bar{X}_n$ si ha:

$$\begin{aligned}\Phi_{\bar{X}_n}(t) &= \Phi_{\frac{1}{n}S_n}(t) = \Phi_{S_n}\left(\frac{t}{n}\right)\\ \log \Phi_{\bar{X}_n}(t) = \log \Phi_{S_n}\left(\frac{t}{n}\right) &= n\left(i\mathbb{E}[X]\frac{t}{n} - \mathbb{V}[X]\frac{t^2}{2n^2} + R\left(\frac{t}{n}\right)\frac{t^3}{n^3}\right)\\ &= i\mathbb{E}[X]t - \mathbb{V}[X]\frac{t^2}{2n} + R\left(\frac{t}{n}\right)\frac{t^3}{n^2}\end{aligned}$$

Ora:

a) se $\mathbb{E}[X] = \mu = 0$,

$$\begin{aligned}\log \Phi_{X_n}(t) &= -\mathbb{V}[X]\frac{t^2}{2n} + R\left(\frac{t}{n}\right)\frac{t^3}{n^2}\\ \lim_{n\to\infty} \log \Phi_{X_n}(t) = 0 \quad &\Rightarrow \quad \lim_{n\to\infty} \Phi_{X_n}(t) = \mathbb{E}[e^{itX_n}] = 1\end{aligned}$$

Questo vuol dire che la funzione caratteristica di $\bar{X}_n$ converge a quella di una variabile aleatorie degenere $X = 0$, ovvero $\bar{X}_n \xrightarrow{d} 0$, e ciò implica:

$$\bar{X}_n = \frac{S_n}{n} \xrightarrow{p} 0 = \mu$$

b) se $\mathbb{E}[X] = \mu \neq 0$, ponendo $Y_n = X_n - \mu$, $\mathbb{E}[Y] = \mathbb{E}[X] - \mu = 0$, si perviene in modo analogo a:

$$\sum_{j=1}^{n}\frac{Y_j}{n} = \sum_{j=1}^{n}\frac{X_j - \mu}{n} = \sum_{j=1}^{n}\frac{X_j}{n} - \mu = \bar{X}_n - \mu \xrightarrow{p} 0 \quad \Rightarrow \quad \bar{X}_n \xrightarrow{p} \mu \qquad \square$$

La *LLN* dice che, quando l'ampiezza di un campione è sufficientemente elevata, allora, per quanto piccolo si possa scegliere $\varepsilon$, la probabilità che la media campionaria si trovi nell'intervallo $\mu \pm \varepsilon$ tende a 1. Ciò non vuol dire che $\bar{X}_n$ sia realmente vicino a $\mu$, ma solo che questo avviene con probabilità molto elevata.

### C.4.3 Il teorema del limite centrale

**Teorema C.2** (Teorema del limite centrale, *CLT* (Central Limit Theorem))**.** *Data una successione $X_n$ di variabili aleatorie indipendenti e identicamente distribuite, con $\mathbb{E}[X] = \mu < \infty$ e $\mathbb{V}[X] = \sigma^2 < \infty$, indicando con $S_n$ la somma dei primi n termini, si ha:*

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{n}} \xrightarrow{d} N(0, \mathbb{V}[X])$$

*Dimostrazione.* Sia $\mathbb{E}[X] = 0$, quindi $\mathbb{E}[S_n] = 0$. Sia $U_n = \frac{S_n}{\sqrt{n}}$. Procedendo come nella dimostrazione della $LLN$ si ottiene:

$$\Phi_{U_n}(t) = \Phi_{S_n}\left(\frac{t}{\sqrt{n}}\right)$$

nonché:

$$\begin{aligned}\log \Phi_{U_n}(t) = \log \Phi_{S_n}\left(\frac{t}{\sqrt{n}}\right) &= n\left(-\mathbb{V}[X]\frac{t^2}{2(\sqrt{n})^2} + R(t/\sqrt{n})\frac{t^3}{(\sqrt{n})^3}\right) \\ &= -\mathbb{V}[X]\frac{t^2}{2} + R(t/\sqrt{n})\frac{t^3}{n^{1/2}}\end{aligned}$$

quindi:

$$\lim_{n\to\infty} \log \Phi_{U_n}(t) = \lim_{n\to\infty} \log \Phi_{S_n}\left(\frac{t}{\sqrt{n}}\right) = -\frac{t^2\mathbb{V}[X]}{2}$$

da cui:

$$\lim \Phi_{U_n}(t) = e^{-\frac{t^2\mathbb{V}[X]}{2}}$$

Ma questa è la funzione caratteristica di una v.a. normale con media nulla e varianza $\mathbb{V}[X]$, quindi:

$$U_n \xrightarrow{d} N(0, \mathbb{V}[X])$$

Se $\mathbb{E}[S_n] \neq 0$, basta sostituire $S_n$ con gli scarti $S_n - \mathbb{E}[S_n]$ e si ha:

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{n}} \xrightarrow{d} N(0, \mathbb{V}[X])$$ □

# Bibliografia


Carlucci, F. e Girardi, A. (s.d.), «Modelli autoregressivi vettoriali», http://dep.eco.uniroma1.it/~carlucci/docs/Modulo10-01.pdf.

Cottrell, A. e Lucchetti, R. (2010), «Gretl User's Guide», manuale utente distribuito con gretl 1.8.1, http://gretl.sourceforge.net/.

Cribari-Neto, F. (2004), «Asymptotic inference under heteroskedasticity of unknown form», *Computational Statistics & Data Analysis*, Vol. 45 (2), pp. 215–233, http://dx.doi.org/10.1016/S0167-9473(02)00366-3.

Dall'Aglio, G. (2003), *Calcolo delle probabilità*, Zanichelli, Bologna.

Engle, R. F. e Granger, C. W. J. (1987), «Co-Integration and Error Correction: Representation, Estimation, and Testing», *Econometrica*, Vol. 55 (2), pp. 251–276, http://www.jstor.org/stable/1913236.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press, Princeton (NJ).

Hansen, B. E. (2010), «Econometrics», draft graduate textbook, http://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf.

Johansen, S. (1991), «Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models», *Econometrica*, Vol. 59 (6), pp. 1551–1580, http://www.jstor.org/stable/2938278.

Kutner, M. H. *et al.* (2005), *Applied Linear Statistical Models*, McGraw-Hill, New York (NY).

Lucchetti, R. (2008), «Appunti di analisi delle serie storiche», http://www.econ.univpm.it/lucchetti/didattica/matvario/procstoc.pdf.

McCloskey, D. N. e Ziliak, S. T. (1996), «The Standard Error of Regressions», *Journal of Economic Literature*, Vol. 34 (1), pp. 97–114, http://www.jstor.org/stable/2729411.

Sims, C. A. (1980), «Macroeconomics and Reality», *Econometrica*, Vol. 48 (1), pp. 1–48, http://www.jstor.org/stable/1912017.

Wonnacott, T. H. e Wonnacott, R. J. (1982), *Introduzione alla statistica*, Franco Angeli, Milano.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge (MA).

Yule, G. U. (1926), «Why do we Sometimes get Nonsense-Correlations between Time-Series?–A Study in Sampling and the Nature of Time-Series», *Journal of the Royal Statistical Society*, Vol. 89 (1), pp. 1–63, http://www.jstor.org/stable/2341482.

Zeileis, A. (2004), «Econometric Computing with HC and HAC Covariance Matrix Estimators», *Journal of Statistical Software*, Vol. 11 (10), pp. 1–17, http://www.jstatsoft.org/v11/i10.

Ziliak, S. T. e McCloskey, D. N. (2004), «Size matters: the standard error of regressions in the American Economic Review», *Journal of Socio-Economics*, Vol. 33 (5), pp. 527–546, http://dx.doi.org/10.1016/j.socec.2004.09.024.