Capitolo 16

Un'introduzione alla metodologia avanzata

Questo ultimo capitolo introduce alcuni metodi statistici avanzati. Un testo introduttivo come questo non ha spazio per presentarli in dettaglio. Tuttavia, è probabile che un ricercatore nell'ambito delle scienze sociali faccia riferimento a questi metodi, ed è utile avere almeno una conoscenza rudimentale della loro natura e dei loro scopi. Anziché la presentazione dei dettagli tecnici, presenteremo una spiegazione in merito a (1) per cosa è utilizzato il metodo, e (2) i risultati che si possono ottenere e le loro interpretazioni.

16.1 Analisi per dati longitudinali*

I Paragrafi 12.6 e 12.7 hanno introdotto i metodi ANOVA per confrontare le medie di campioni *dipendenti*. Tali dati più comunemente derivano da studi che osservano i soggetti più volte nel corso del tempo, ovvero, da *studi longitudinali*. Per tali dati sono disponibili metodi specifici, che sono diversi per le ipotesi che fanno e per il modo in cui modellano la struttura di correlazione delle osservazioni ripetute.
I Paragrafi 12.6 e 12.7 hanno presentato le tradizionali *misure ripetute ANOVA*. Questo tipo di ANOVA ha delle limitazioni.

- Le osservazioni su tutti i soggetti devono essere fatte contemporaneamente.
- Il metodo non può trattare i dati mancanti. I soggetti con *qualche* eventuale osservazione mancante vengono eliminati dall'analisi. Questo può condurre a una significativa distorsione quando i dati mancanti sono numerosi.
- Si assume che le correlazioni tra le osservazioni sullo stesso soggetto abbiano una struttura *omoschedastica*, che implica una uguale variabilità nel tempo e uguale correlazione tra ciascuna coppia di osservazioni.

MANOVA: Analisi multivariata della varianza

Un tipo più generale di ANOVA si basa su assunzioni restrittive sulla struttura della correlazione. Esso tratta l'insieme di misure ripetute su un soggetto come un vettore multivariato di risposte. In tal caso, i metodi standard possono essere utilizzati per dati multivariati. Tali metodi possono verificare le ipotesi confrontando le medie senza fare alcuna assunzione sulla struttura di correlazione. I test sono chiamati MANOVA, abbreviazione di **analisi multivariata della varianza**. Il particolare il test MANOVA cui si riferisce la maggior parte software come *Wilks' lambda* è un *test del rapporto di verosimiglianza*, introdotto nel Paragrafo 15.3.

Tuttavia l'approccio MANOVA ha i suoi svantaggi. Il principale è che spesso perde potenza, in quanto si basa su ipotesi più deboli. Se le ipotesi per le misure ripetute tradizionali ANOVA non sono fortemente violate, tale metodo ha una maggiore potenza per verificare gli effetti. La MANOVA perde potenza perché richiede la stima di un gran numero di parametri. Inoltre, la MANOVA ha anche i primi due svantaggi delle tradizionali misure ripetute ANOVA: le osservazioni devono esse-

re effettuate contemporaneamente, e i soggetti con osservazioni mancanti vengono eliminati dall'analisi.

Modelli a effetti misti con effetti casuali

Un metodo sviluppato più recentemente rispetto ai tradizionali ANOVA e MANOVA non ha gli svantaggi appena discussi. Questo metodo permette diversi tipi di modellazione della struttura di correlazione per le risposte ripetute; consente di avere osservazioni in diversi tempi e di includere i soggetti nell'analisi anche quando alcune delle loro osservazioni sono mancanti. Alla base dei metodi vi è l'assunzione che le osservazioni siano *casualmente mancanti*. Ciò significa che la probabilità che un'osservazione sia mancante non dipende dal valore delle risposte non osservate. Come di consueto, l'inferenza assume la normalità, ma questa ipotesi diviene meno importante con campioni di dimensioni maggiori. Questo metodo utilizza un tipo alternativo di modello che include esplicitamente **effetti casuali** per i soggetti. Come spiegato nel Paragrafo 12.6, un modello per misure ripetute può includere una variabile dummy per ogni soggetto. Il coefficiente di una variabile dummy rappresenta l'effetto per un particolare soggetto. Per esempio, un effetto positivo significa che ogni osservazione per quel soggetto tende a essere superiore alla media osservata per i soggetti che condividono gli stessi valori degli altri predittori. Il fattore individuale comprende tutti questi effetti specifici del soggetto. Gli effetti sono chiamati *effetti casuali*, poiché i soggetti osservati sono considerati come un campione casuale di tutti i possibili soggetti che potevano essere campionati. La distinzione tra gli effetti casuali e altri parametri del modello è che gli effetti casuali sono trattati come variabili casuali non osservate piuttosto che come parametri. Cioè, i coefficienti nel modello per i soggetti si presume che provengano da una particolare distribuzione di probabilità, solitamente la distribuzione normale con varianza non nota. Questo è utile, perché diversamente il numero di parametri che devono essere stimati potrebbe essere enorme (quando n è grande) se ogni soggetto è trattato come parametro anziché come un effetto casuale.

I modelli includono anche gli *effetti fissi* (parametri ordinari) per le variabili predittrici (come il trattamento) per le quali le analisi utilizzano tutte le categorie di interesse. Poiché le variabili di classificazione sono una mistura di effetti casuali e fissi, il modello è chiamato **modello misto**.

Un aspetto interessante di questo approccio a effetti casuali è la libertà di usare diverse strutture di correlazione per modellare le misure ripetute su un soggetto. Una possibilità è la **scambiabilità** della struttura. Ciò presuppone che, considerando le risposte di un soggetto in tempi diversi, tutte le correlazioni a coppie siano uguali. In pratica, spesso le osservazioni più vicine nel tempo tendono a essere più correlate di osservazioni più lontane. La struttura **autoregressiva** è una via che permette di verificarlo. Con essa, se ρ indica la correlazione per le osservazioni in un tempo precedente, allora la correlazione è ρ^2 per le osservazioni considerate nei due tempi precedenti, ρ^3 per le osservazioni considerate nei tre tempi precedenti, e così via. È anche possibile utilizzare un approccio **non strutturato** che non fa alcuna ipotesi circa il modello di correlazione. In ciascuno di questi casi, le correlazioni sono esse stesse parametri da stimare utilizzando i dati.

Misure ripetute ANOVA a una via usando effetti casuali

Consideriamo prima il caso dell'ANOVA a una via con misure ripetute sui soggetti. Ovvero, il caso in cui l'interesse è rivolto alle osservazioni all'interno degli stessi soggetti, e non alle differenze tra i soggetti. Questo caso è stato considerato con le

misure ripetute tradizionali ANOVA nel Paragrafo 12.6. Per T tempi, il modello misto con effetti casuali è

$$E(y) = \alpha + s_i + \beta_1 t_1 + \beta_2 t_2 + \cdots + \beta_{T-1} t_{T-1}$$

dove s_i è un effetto casuale per il soggetto i, t_1 è una variabile dummy per il tempo 1, t_2 è una variabile dummy per il tempo 2, e così via. Il tempo è un effetto fisso. Quando $T = 2$, per esempio, l'obiettivo dello studio potrebbe essere quello di confrontare le medie prima e dopo la somministrazione di quanche trattamento. I parametri di maggior interesse da stimare sono gli effetti fissi $\{\beta_t\}$. Il termine relativo al soggetto nel modello corrisponde alla *intercetta casuale* che viene aggiunta al termine di intercetta ordinario.

Se ci attendiamo un trend lineare nel tempo, potremmo sostituire l'effetto fisso dei termini $T - 1$ per i T tempi con βt, quindi usando un coefficiente invece di $T - 1$ intercette temporali. I modelli misti possono avere allora un secondo tipo di effetto casuale per consentire alle pendenze di variare per soggetto intorno a una media di β, ovvero, permettere una *pendenza casuale* così come una intercetta casuale.

Il modello appena descritto per un confronto delle T medie all'interno dei soggetti, si estende per includere anche gli effetti tra soggetti, in aggiunta agli effetti all'interno (entro) i soggetti. Il prossimo esempio illustra questo aspetto.

ESEMPIO 16.1 Qualità della vita con trattamenti per la dipendenza da alcool

Gueorguieva e Krystal (2004) hanno analizzato i dati di uno studio clinico per verificare l'effetto di un particolare farmaco (naltrexone) oltre alla terapia psicosociale nel trattamento di 627 soggetti che soffrono di una grave dipendenza da alcool. La variabile risposta è il punteggio di soddisfazione per gli aspetti finanziari della propria vita, dato dalla media di quattro aspetti rilevati su una scala della qualità della vita, per la quale ogni elemento aveva risposte potenziali da 1 (terribile)a 7 (soddisfatto). Per ogni soggetto, questa risposta è stata osservata all'inizio e dopo 78 settimane. I tre trattamenti sono stati 12 mesi di farmaco, 3 mesi di farmaco seguito da 9 mesi di placebo, o 12 mesi di placebo. Questo è il fattore che differenzia i soggetti. La Tabella 16.1 mostra la media campionaria dei punteggi di soddisfazione per le 3×5 combinazioni di trattamento e tempo.

Tabella 16.1 Media campionaria del livello di soddisfazione per i soggetti che soffrono dalla dipendenza da alcool, per trattamento e tempo di misurazione. Il trattamento è un fattore tra gruppi e il tempo è un fattore entro i gruppi.

	Tempo				
Trattamento	**Iniziale**	**4 Settimane**	**26 Settimane**	**52 Settimane**	**78 Settimane**
Farmaco lungo-termine	3.9	4.0	4.3	4.5	4.4
Farmaco breve-termine	3.7	4.0	4.1	4.3	4.3
Placebo	3.6	3.9	3.9	4.2	4.3

Questo insieme di dati è simile a quello che abbiamo analizzato nel Paragrafo 12.7 mediante l'ANOVA a due fattori e le misure ripetute su uno di essi. Qui le misure ripetute si verificano per i cinque tempi.

Consideriamo il modello

$$E(y) = \alpha + s_i + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \beta_4 t_4 + \beta_5 d_1 + \beta_6 d_2 + \beta_7 t_1 \times d_1 + \beta_8 t_1 \times d_2 + \beta_9 t_2 \times d_1 + \beta_{10} t_2 \times d_2 + \beta_{11} t_3 \times d_1 + \beta_{12} t_3 \times d_2 + \beta_{13} t_4 \times d_1 + \beta_{14} t_4 \times d_2$$

dove s_i è l'effetto per il soggetto i, $\{t_1, t_2, t_3, t_4\}$ sono variabili dummy per i primi quattro tempi e $\{d_1, d_2\}$ sono variabili dummy per i due farmaci.

La tradizionale ANOVA per misure ripetute, utilizzando l'aggiustamento di Greenhouse-Geisser per tenere conto della violazione del presupposto di omoschedasticità, non mostra evidenza di un'interazione tra trattamento e tempo (P-value $=$ 0.69) e nessuna evidenza di un effetto del trattamento (P-value $= 0.80$). C'è una forte evidenza di un effetto tempo (P-value < 0.001), non sorprendente considerate le medie riportate nella Tabella 16.1. Concludiamo che anche se la soddisfazione sembra aumentare nel corso del tempo, il miglioramento è per il placebo altrettanto veloce quanto per il trattamento farmacologico. Tuttavia, i risultati completi per tutti e cinque i tempi erano disponibili solo per 211 dei 627 soggetti. Questa analisi ha utilizzato solo casi completi, ignorando così i dati per i 416 soggetti con osservazioni mancanti.

Con il modello misto, trattiamo s_i come effetto casuale per il soggetto i. Allora non è necessario ignorare i dati per i soggetti che presentano alcune osservazioni mancanti. Ciò si traduce in una maggiore potenza. Quando viene utilizzato un modello con un pattern non strutturato per le correlazioni, c'è una lieve evidenza di un effetto del trattamento (P-value $= 0.07$). Nessuna evidenza di interazione (P-value $= 0.63$) e un'evidenza molto forte di un effetto tempo (P-value < 0.001).

Questo esempio mostra il beneficio di un modello con un approccio misto in presenza di molte osservazioni mancanti. L'approccio tradizionale porterebbe a escludere molte informazioni, in questo modo si determina l'effetto del trattamento cambia da marginale ($P = 0.07$) a molto debole ($P = 0.80$).

Con i modelli misti, è possibile utilizzare i dati per prevedere gli effetti individuali. Questo ci permette di prevedere le tendenze temporali a livello di soggetto. Contrariamente agli approcci tradizionali, la risposta predetta in ogni intervallo temporale varia tra i soggetti sottoposti a uno stesso trattamento. Più in generale, nei modelli misti, gli effetti casuali possono rappresentare *gruppi* piuttosto che soggetti. Per esempio, in uno studio che campioni famiglie e osserva i soggetti in ogni famiglia, un gruppo è costituito da soggetti della stessa famiglia. Per alcuni valori delle variabili esplicative, due soggetti nella stessa famiglia tendono a essere più simili rispetto a due soggetti in famiglie diverse. L'identificazione delle famiglie nel modello utilizzando un effetto casuale per ogni famiglia tiene conto della correlazione tra soggetti all'interno della famiglia.

Sempre più software riescono ad adattare modelli misti assumendo varie strutture di correlazione per misure ripetute. Un esempio è PROC MIXED di SAS. L'analisi richiede alcune accortezze per evitare che i modelli siano inappropriati. In ogni caso, prima di utilizzare questi metodi sarebbe opportuna la guida di un esperto di statistica. Per maggiori dettagli, Fitzmaurice et al. (2004), Hedeker e Gibbons (2006), e Gueorguieva e Krystal (2004). Se non ci sono dati mancanti, i valori sono osservati contemporaneamente, e la correlazione comune sembra un'ipotesi plausibile, è più semplice utilizzare le tradizionali misure ripetute ANOVA. Se la dimensione del campione è piccola, questo approccio è ancora preferito quando le sue assunzioni non sono fortemente violate.

I precedenti esempi si riferiscono a risposte continue e modellano la media. Approcci simili sono stati sviluppati per risposte categoriali e modellano proporzioni usando i logit. Per esempio, è possibile includere effetti casuali in un modello di re-

gressione logistica per tener conto delle associazioni entro i soggetti negli studi con misure ripetute su una variabile binaria. Per una discussione sui modi per gestire risposte categoriali negli studi longitudinali, vedi Agresti (2007, Capitoli 9 e 10).

16.2 Modelli (gerarchici) multilivello*

Gli effetti casuali sono utili per vari tipi di modelli in aggiunta ai modelli per dati longitudinali. Un esempio è quello dei modelli gerarchici che descrivono le osservazioni che hanno una struttura annidata: le unità a un livello sono contenute all'interno di unità di un altro livello. I dati gerarchici sono comuni in taluni settori applicativi, come negli studi scolastici.

Per esempio, uno studio sui risultati ottenuti dagli studenti potrebbe misurare, per ogni studente, il risultato raggiunto in ogni esame considerando un insieme di esami. Gli studenti sono annidati all'interno delle scuole. Il modello potrebbe descrivere se la risposta attesa per un soggetto dipende dalle variabili esplicative e, al contempo, come la risposta attesa per una scuola dipende dalle variabili esplicative. Cioè, il modello potrebbe analizzare l'effetto delle caratteristiche di uno studente (per esempio i risultati negli esami precedenti) e delle caratteristiche della scuola che lo studente frequenta. Due osservazioni per lo stesso studente (su esami diversi) potrebbero essere più simili di due osservazioni per diversi studenti, così due studenti della stessa scuola potrebbero avere osservazioni più simili rispetto a due studenti provenienti da diverse scuole. Questo potrebbe verificarsi perché gli studenti all'interno di una scuola tendono a essere simili rispetto a diverse caratteristiche socio-economiche.

Modellazione delle osservazioni su due livelli

I modelli gerarchici contengono i termini per i diversi livelli di unità. Per l'esempio appena citato, il modello dovrebbe contenere i termini per prevedere una risposta attesa per lo studente e i termini per prevedere la risposta attesa all'interno di una scuola. Il livello 1 si riferisce a misurazioni a livello degli studenti e il livello 2 si riferisce a misurazioni a livello di scuola. Modelli aventi una struttura gerarchica di questo tipo sono detti *modelli multilivello*. I modelli multilivello spesso hanno un elevato numero di termini. Per limitare il numero di parametri, il modello considera i termini per le unità campionate sulle quali ci sono molteplici osservazioni come effetti casuali, piuttosto che come effetti fissi. Gli effetti casuali possono essere considerati nel modello a ogni livello della gerarchia.

Per esempio, gli effetti casuali per gli studenti e gli effetti casuali per scuole si riferiscono a differenti livelli del modello. Per il livello 1 gli effetti casuali possono tener conto della variabilità tra gli studenti con riferimento alle caratteristiche specifiche degli studenti che non sono state misurate dalle variabili esplicative. Queste potrebbero includere le abilità degli studenti e lo status socio-economico dei genitori. Gli effetti casuali a livello 2 rappresentano la variabilità tra le scuole dovuta alle caratteristiche specifiche della scuola, non misurate dalle variabili esplicative. Queste potrebbero includere la qualità del corpo docente, i problemi legati alla droga nella scuola, e le caratteristiche del distretto dal quale la scuola iscrive gli studenti.

ESEMPIO 16.2 Un modello a due livelli per il rendimento scolastico degli studenti

Uno studio sull'istruzione analizza i fattori che influenzano le prestazioni degli studenti su una prova di rendimento. Indichiamo il punteggio del test per lo studente t

nella scuola i con y_{it}. Prendiamo in considerazione un modello con due livelli, uno per gli studenti e uno per le scuole. Quando ci sono molte scuole e queste possono essere considerate un campione casuale delle scuole che lo studio avrebbe potuto analizzare, usiamo effetti casuali per le scuole. Il livello 1 (livello studente) dovrebbe includere un termine per la scuola i e le variabili esplicative con valori che variano tra studenti, come $x_1 =$ risultato su un test di profitto conseguito in un livello precedente, $x_2 =$ media dei voti, $x_3 =$ razza e $x_4 =$ se lo studente è stato mai bocciato. Il modello di livello 1 è

$$E(y_{it}) = \alpha + \delta_i + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

dove δ_i è l'effetto che accomuna ogni studente che frequenta la scuola i.
Il modello a due livelli (scuola) fornisce un predittore lineare per il termine di livello 2 δ_i nel modello di livello 1. Il modello a 2 livelli ha variabili esplicative che variano solo a livello scuola, come per esempio $w_1 =$ spesa per studente della scuola i, $w_2 =$ salario medio degli insegnanti e $w_3 =$ mediana del reddito delle famiglie nel distretto della scuola. Per esempio, il modello di livello 2 potrebbe essere

$$\delta_i = s_i + \gamma_1 w_1 + \gamma_2 w_2 + \gamma_3 w_3$$

Il termine s_i è un effetto casuale per la scuola i.
Sostituendo il modello a livello 2 nel modello a livello 1, otteniamo

$$E(y_{it}) = \alpha + s_i + \gamma_1 w_1 + \gamma_2 w_2 + \gamma_3 w_3 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Ogni studente nella scuola i condivide l'effetto casuale s_i. Questo è un modello misto, con un effetto casuale per le scuole ed effetti fissi costituiti dalle variabili esplicative per le scuole e le variabili esplicative per i soggetti. Il modello può essere adattato da un software per modelli misti o con un software specializzato per modelli multilivello (come HLM ed MLwiN).

Più in generale, gli effetti casuali possono essere considerati in ciascun livello del modello. Per esempio, supponiamo che ci siano diverse osservazioni per studente, come un punteggio per ogni esame in un'insieme di test. Il modello può includere effetti casuali per gli studenti e per le scuole. Più in generale, un modello multilivello può avere più di due livelli. Per ulteriori dettagli, vedi Gelman e Hill (2006), Raudenbush e Bryk (2002), Snijders e Bosker (1999).

16.3 Modelli di event history*

Alcuni studi hanno l'obiettivo di modellare una variabile risposta che rileva quanto tempo occorre attendere finché un certo tipo di evento si verifica. Per esempio, la variabile risposta potrebbe essere per quanto tempo una persona lavora prima di ritirarsi dal mondo del lavoro, l'età di una persona al primo matrimonio, il periodo di tempo prima che qualcuno appena uscito di prigione venga arrestato di nuovo o per quanto tempo una persona vive dopo la diagnosi di AIDS.

Come nella regressione tradizionale, i modelli interessati a studiare il tempo fino al verificarsi di un certo evento includono gli effetti delle variabili esplicative. Un modello per il periodo di tempo prima del nuovo arresto, per esempio, potrebbe utilizzare predittori quali il numero di arresti precedenti, lo status occupazionale, lo stato civile, l'età al momento del rilascio da un precedente arresto e il livello di istruzione.

La modellazione degli eventi che si verificano nel tempo utilizzando un insieme di variabili esplicative è chiamato analisi delle biografie o **event history analysis**. Il primo sviluppo dei modelli di event history fu portato avanti nel 1980 nell'ambito della biostatistica per modellare il periodo di sopravvivenza di un paziente dopo che

lo stesso era stato sottoposto a un particolare trattamento medico. In tale contesto, l'analisi è chiamata **analisi di sopravvivenza**. Per esempio, l'analisi di sopravvivenza può modellare il tempo di sopravvivenza fino alla morte di un paziente che ha avuto un trapianto di cuore, utilizzando variabili esplicative quali l'età al momento dell'operazione, lo stato generale di salute, un indice di massa corporea, e se il paziente è o è stato un fumatore.

Dati censurati e covariate che variano nel tempo

Nelle analisi di event history, il dato per ogni soggetto è costituito dal tempo intercorso finché l'evento di interesse si verifica. Esistono due fattori di complicazione, che non sono considerati nei modelli di regressione tradizionali.

Primo, per alcuni soggetti, l'evento non si verifica alla fine del periodo di osservazione dello studio. Non siamo in grado di osservare il tempo in cui si verifica l'evento per questi soggetti, ma solo i limiti inferiori dei tempi. Per esempio, uno studio sugli effetti di diverse variabili esplicative sulla età di pensionamento può utilizzare un campione di adulti di almeno 65 anni. Alcuni soggetti di questo campione possono non essere ancora andati in pensione. Se una persona di 68 anni non è ancora andata in pensione, si sa solo che la variabile risposta (età pensionabile) assume un valore di almeno 68.

Un'osservazione di cui conosciamo solo la regione di valori in cui essa ricade, si dice essere **censurata**. I metodi di event history hanno modalità particolari di trattamento dei dati censurati. Ignorare i dati censurati e stimare modelli utilizzando solo i dati per i soggetti con risposte osservate, può comportare una distorsione grave nella stima dei parametri.

In secondo luogo, alcune variabili esplicative possono assumere valori diversi nel tempo. Per esempio, consideriamo uno studio sulla recidività criminale che modelli il tempo intercorso fino a quando si verifica un nuovo arresto. Ogni mese si può osservare se una persona è stata nuovamente arrestata (l'evento di interesse) e utilizzare come variabili esplicative se il soggetto lavora e se il soggetto è sposato o vive con un partner. Per un soggetto specifico, il valore di queste variabili esplicative potrebbe variare nel tempo. Una variabile esplicativa che varia nel tempo è chiamata **covariata tempo-dipendente**. I metodi avanzati per adattare modelli di event history possono trattare sia le covariate dipendenti dal tempo che quelle indipendenti dal tempo.

Il tasso di accadimento di un evento

Abbiamo considerato il *periodo di tempo* finché un particolare evento si verifica come la variabile risposta di interesse. I più conosciuti modelli di event history descrivono, comunque, il **tasso** di accadimento dell'evento.

Si consideri, per esempio, uno studio sui problemi di salute dei soggetti ricoverati presso una casa di cura. La risposta è il periodo di tempo intercorso dopo il ricovero prima che un soggetto richieda una particolare cura medica speciale che necessita il ricovero in un ospedale. Per un determinato valore delle variabili esplicative, il campione contiene cinque soggetti. Il tempo intercorso prima della richiesta di cure mediche speciali è pari 0.5 anni per il primo soggetto, 0.2 anni per il secondo, 1.3 anni per il terzo e 0.1 anno per il quarto. Il quinto soggetto è una osservazione censurata, non richiede alcuna assistenza medica speciale durante 0.4 anni, e quando il periodo di osservazione si è concluso era ancora nella casa di cura. Quindi, per questi cinque soggetti, il numero totale di avvenimenti dell'evento di interesse è 4, e il tempo totale di osservazione è di $(0.5+0.2+1.3+0.1+0.4) = 2.5$ anni. Il tasso di accadimento campionario è $4/2.5 = 1.6$, ovvero, il tasso di accadimento è uguale a 1.6 eventi per anno. Il numero di eventi del campione è pari 1.6 volte la quantità totale di tempo

per il quale l'intero campione di soggetti è stato sotto osservazione. La formula del modello si riferisce al tasso di accadimento dell'evento piuttosto che al tempo trascorso prima del verificarsi dell'evento. In letteratura, il tasso è solitamente chiamato **tasso di rischio** o **hazard rate** e indicato con h. Il calcolo di cui sopra per un hazard rate del campione assume implicitamente che questo tasso sia costante nel tempo. In realtà questa assunzione spesso non è realistica. I modelli possono permettere al tasso di rischio di essere dipendente dal tempo così come dai valori delle variabili esplicative.

Il modello a rischi proporzionali

Sia $h(t)$ il tasso di rischio al tempo t, per esempio t anni dopo il ricovero presso una casa di cura. Il modello per il tasso di rischio e un insieme di variabili esplicative è

$$\log h(t) = \alpha(t) + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Il modello si applica al log del tasso di rischio, perché il tasso di rischio deve essere positivo. Le funzioni lineari per il tasso di rischio hanno lo svantaggio che potrebbero fornire un valore previsto negativo, come il modello di probabilità lineare per una variabile risposta binaria. Nella forma del modello sopra descritto, le variabili esplicative sono indipendenti, ma potrebbero essere anche dipendenti dal tempo.

L'intercetta $\alpha(t)$ può dipendere dal tempo t e questo permette al tasso di rischio di variare nel tempo. Di solito l'obiettivo primario è la stima degli effetti delle variabili esplicative sul tasso di rischio, non sulla modellazione della dipendenza del tasso di rischio dal tempo. Lo studio geriatrico, per esempio, potrebbe stimare gli effetti sul tasso di rischio dei fattori sesso ed età del soggetto. Per questo motivo, è comune permettere che il parametro $\alpha(t)$ sia una funzione arbitraria e comunque non specificata. Il principale obiettivo dell'analisi è la stima di $\{\beta_j\}$ per fare inferenza sugli effetti delle covariate.

Come interpretiamo gli effetti delle covariate? Consideriamo l'effetto sul tasso di rischio di un incremento unitario di x_1 tenendo gli altri predittori costanti. Indichiamo i tassi di rischio in corrispondenza dei valori di x_1 e $x_1 + 1$ rispettivamente con h_1 e h_2. Allora

$$\begin{aligned}\log h_2(t) - \log h_1(t) &= [\alpha(t) + \beta_1(x_1 + 1) + \cdots + \beta_k x_k] \\ &\quad - [\alpha(t) + \beta_1 x_1 + \cdots + \beta_k x_k] = \beta_1\end{aligned}$$

Quindi β_1 è la variazione del log del tasso di rischio per un cambiamento unitario di x_1, tenendo gli altri predittori fissi. Ma $\log h_2(t) - \log h_1(t) = \log[h_2(t)/h_1(t)]$ ed elevando a potenza entrambi i membri, $h_2(t)/h_1(t) = e^{\beta_1}$, ovvero

$$h_2(t) = e^{\beta_1} h_1(t)$$

Quindi, l'aumento unitario di x_1 ha l'effetto di moltiplicare il tasso di rischio di e^{β_1}. Gli effetti sono moltiplicativi, come in altri modelli che utilizzano i logaritmi per rendere lineari le funzioni (come la regressione esponenziale e i modelli di regressione logistica).

L'equazione $h_2(t) = e^{\beta_1} h_1(t)$ mostra che il tasso di rischio per un dato valore delle variabili esplicative è proporzionale al tasso di rischio considerando un altro valore di queste. La stessa costante di proporzionalità si applica in ogni tempo t. Per questa proprietà, questa forma di modello è chiamata **modello a rischi proporzionali**. È semplice interpretare gli effetti per tali modelli, perché l'effetto di qualsiasi variabile esplicativa sul tasso di rischio è identico in ogni tempo.

Nel 1972 lo statistico britannico David Cox ha proposto e ha mostrato come stimare questa forma di modello a rischi proporzionali, in cui la dipendenza del tasso di rischio nel tempo, attraverso $\alpha(t)$, è arbitraria. Il modello è chiamato **Modello a rischi proporzionali di Cox**. È un modello *non parametrico*, nel senso che non fa alcuna ipotesi circa la distribuzione di probabilità del tempo per l'evento, ma si concentra sugli effetti delle variabili esplicative. Modelli più specializzati del modello a rischi proporzionali (o altre forme) formulano ipotesi parametriche su questa distribuzione. Ciò è utile se la distribuzione del tempo fino all'accadimento dell'evento è un obiettivo importante dello studio, come, per esempio, per l'analisi del tempo di rottura dei componenti elettronici.

ESEMPIO 16.3 Modellazione del tempo alla separazione coniugale

Un articolo sulla modellazione delle dinamiche familiari con le tecniche di event history ha analizzato negli USA i dati dell'Indagine Nazionale delle Famiglie e delle Convivenze. È stato intervistato un campione probabilistico di circa 13.000 soggetti, e successivamente 10.000 di questi soggetti sono stati intervistati dopo circa sei anni. Lo scopo è stato quello di analizzare i fattori che influenzano il tasso di rischio della separazione coniugale. L'esito per ogni soggetto, sposato all'inizio dello studio, è il numero di mesi dal matrimonio fino alla separazione. Le persone che sono ancora sposate o vedove alla fine dello studio sono osservazioni censurate.

La Tabella 16.2 sintetizza l'adattamento del modello. L'ultima colonna della tabella mostra le stime elevate a potenza dei parametri di regressione, che forniscono i tassi di rischio. Per esempio, poiché $e^{0.353} = 1.42$, la stima del tasso di separazione per i neri era 1,42 volte rispetto al tasso dei bianchi. Questo è il più importante degli effetti indicati nella tabella.

Tabella 16.2 Effetti stimati sul tasso di rischio per la separazione coniugale, basato sul modello a rischi proporzionali di Cox.

Variabile	Stima	Std. Error	*P*-Value	e^b
Età al matrimonio	−0.086	0.0050	0.000	0.917
Anno del matrimonio	0.048	0.0017	0.000	1.049
Razza (nera = 1)	0.353	0.0423	0.000	1.423
Genere (maschio = 1)	−0.065	0.0375	0.083	0.937

Fonte: T. B. Heaton e V. R. A. Call, Journal of Marriage & Family, vol. 57, 1995, p. 1078.

Come nella regressione logistica, i test di significatività dei parametri del modello possono utilizzare le statistiche di Wald o le statistiche del rapporto di verosimiglianza. Per esempio, per H_0: nessun effetto di genere, la statistica test di Wald è $z = -0.065/0.0375 = -1.73$ con P-value = 0.083. Equivalentemente, il quadrato di questa statistica è un chi-quadro con $df = 1$.

Attualmente esistono diversi software per adattare i modelli di event history, come per esempio l'opzione SOPRAVVIVENZA con REGRESSIONE DI COX come sub-opzione nel menu ANALIZZA di SPSS. Vedi Allison (1984), DeMaris (2004, Cap. 11), e Yamaguchi (1991) per un'introduzione su questo argomento.

16.4 Analisi causale o path analysis*

L'analisi causale o *path analysis* utilizza i modelli di regressione per rappresentare la teoria sottostante alle relazioni causali tra un insieme di variabili. Statisticamente, è semplicemente una analisi di regressione. Ci sono vantaggi a condurre le analisi all'interno del contesto analitico della path analysis. Il vantaggio principale è che il ricercatore deve specificare in modo esplicito le presunte relazioni causali tra le variabili e questo può contribuire allo sviluppo di teorie rilevanti sulle relazioni tra variabili.

L'associazione è una caratteristica delle relazioni causa-effetto. Come discusso nel Paragrafo 10.1, tuttavia, ciò non implica il nesso di causalità. Due variabili che sono causalmente dipendenti da una terza variabile potrebbero essere tra loro associate. Tuttavia, nessuna delle due è causa dell'altra, e l'associazione scompare quando la terza variabile è controllata. La path analysis utilizza i modelli di regressione che includono opportune variabili di controllo.

Una variabile esplicativa x è una possibile causa della variabile risposta y se si verifica il corretto ordine temporale e se cambiamenti in x danno luogo a cambiamenti in y, anche quando tutte le altre variabili rilevanti sono controllate.

Se l'associazione tra due variabili scompare tenendo sotto controllo una terza variabile, non esiste un nesso causale diretto tra loro. Se l'associazione non scompare, però, il rapporto non è necessariamente di causazione. L'associazione potrebbe scomparire quando altre variabili sono controllate. Così possiamo dimostrare la non causalità ma non si può mai dimostrare causalità. Un'ipotesi di nesso causale è sostenuta, però, se l'associazione rimane dopo che i controlli sono stati introdotti.

I diagrammi causali o path diagrams

Spiegazioni teoriche delle relazioni causa-effetto spesso ipotizzano un sistema di relazioni in cui alcune variabili, ritenute essere causate da altre, potrebbero a loro volta avere effetti su altre variabili ancora. Un unico modello di regressione multipla è insufficiente per tale sistema, dal momento che può gestire solo una singola variabile di risposta. La path analysis utilizza il numero necessario di modelli di regressione per includere tutte le relazioni proposte nella spiegazione teorica.

ESEMPIO 16.4 Cause del livello di istruzione

Supponiamo che una teoria specifichi la seguente:

1. Il livello di istruzione di un soggetto dipende da diversi fattori, tra cui l'intelligenza del soggetto, la motivazione del soggetto e il reddito dei genitori.
2. La motivazione del soggetto dipende da diversi fattori, tra cui il livello di intelligenza generale e dal livello di istruzione dei genitori.
3. Il reddito dei genitori dipende in parte dal loro livello di istruzione.

La Figura 16.1 mostra una sintesi grafica dei legami appena illustrati. La figura è chiamata **diagramma causale** o **path diagram**. Tali schemi generalizzano gli schemi causali introdotti nel Capitolo 10.

Nei diagrammi causali, una relazione causa-effetto è rappresentata da un freccia verso l'effetto (la risposta) e parte dalla variabile causale (esplicativa). Le variabili risposta delle equazioni di regressione sono le variabili a cui le frecce puntano.

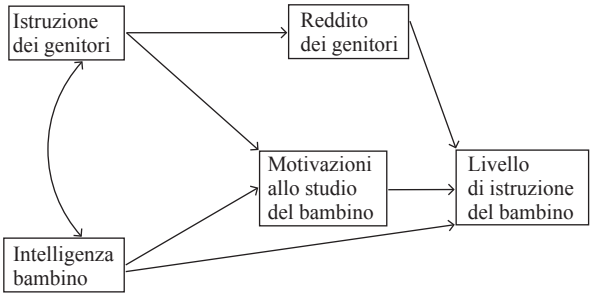

Figura 16.1 Esempio di diagramma causale per il livello di istruzione.

Le variabili esplicative per un'equazione con una particolare variabile risposta sono quelle variabili con le frecce che puntano verso quella variabile di risposta. Nella Figura 16.1, il reddito dei genitori è modellato come dipendente dalla istruzione dei genitori; il livello di istruzione del bambino come dipendente dal reddito dei genitori, dalla sua intelligenza, e dalla sua motivazione; e la motivazione del bambino come dipendente dal livello di istruzione dei genitori e dall'intelligenza del bambino. Ciò indica che le due variabili possono essere associate, ma il modello non risolve il rapporto di causalità (se presente).

I coefficienti causali o path coefficients

Solitamente in un diagramma causale, su ogni freccia è scritto un valore. Questi valori sono i coefficienti di regressione standardizzati per l'equazione di regressione della variabile risposta indicata dalle frecce. Nel contesto dell'analisi causale, essi sono chiamati **coefficienti causali**. La Figura 16.1 ha tre insiemi di coefficienti da stimare, in quanto fa riferimento a tre variabili risposta separate.

Indichiamo le variabili standardizzate in questa figura con E, A e I per il livello di istruzione del bambino, la motivazione e l'intelligenza, e con Pe e Pi il livello di istruzione dei genitori e il loro reddito. Inoltre, per le due variabili x e y, sia β^*_{yx} il coefficiente di regressione standardizzato per l'effetto di x su y. Allora la Figura 16.1 corrisponde alle tre equazioni di regressione

$$E(E) = \beta^*_{EI} I + \beta^*_{EA} A + \beta^*_{EPi} Pi \tag{1}$$

$$E(A) = \beta^*_{AI} I + \beta^*_{APe} Pe \tag{2}$$

$$E(Pi) = \beta^*_{PiPe} Pe \tag{3}$$

Per esempio, il coefficiente causale che collega il livello di istruzione del genitore alla motivazione del bambino è la stima del coefficiente di regressione standardizzato β^*_{APe} del modello di regressione multipla (2) che considera la motivazione del bambino come variabile di risposta e l'istruzione dei genitori e l'intelligenza del bambino come variabili esplicative. Il reddito dei genitori, in questo modello, dipende solo dalla loro istruzione [vedi (3)]. Il coefficiente causale per quella freccia è il coefficiente di regressione bivariata standardizzata, ovvero in coefficiente di correlazione tradizionale.

I coefficienti causali indicano la direzione e l'importanza relativa degli effetti delle variabili esplicative, tenendo costanti le altre variabili. La loro interpretazione è quella dei coefficienti della regressione multipla standardizzata (Paragrafo 11.8).

Per esempio, un valore di 0.40 significa che un aumento di una deviazione standard nella variabile esplicativa corrisponde a un aumento previsto di 0.40 deviazioni standard della variabile risposta, controllando per le altre variabili esplicative del modello.

Nel diagramma causale, per ogni variabile risposta è associato un **residuo**. Questo rappresenta la variazione non spiegata dalle sue variabili esplicative. Ogni variabile residuale rappresenta la quota restante $(1 - R^2)$ della variazione non spiegata, dove R^2 denota il valore di R-quadro per l'equazione di regressione per quella variabile risposta. Il suo coefficiente causale è pari a $\sqrt{1 - R^2}$.

Effetti diretti e indiretti

Molti modelli causali includono variabili dipendenti da altre che, allo stesso tempo, sono cause di altre variabili risposta. Queste variabili sono **variabili intervenienti** (Paragrafo 10.3), poiché ricorrono in sequenza tra altre variabili. Nella Figura 16.1, la motivazione del bambino interviene tra l'intelligenza del bambino e il livello di istruzione dello stesso. Se questa teoria causale è corretta, l'intelligenza influenza il livello di istruzione attraverso il suo effetto sulla motivazione. Un effetto di questo tipo, che opera attraverso una variabile interveniente, si dice che sia **indiretto**.

La Figura 16.1 suggerisce anche che l'intelligenza del bambino ha un effetto **diretto** sul livello di istruzione, in aggiunta al suo effetto attraverso la motivazione. Una motivazione importante per utilizzare l'analisi causale è che studia gli effetti diretti e indiretti di una variabile.

D'altra parte, la Figura 16.1 suggerisce che il livello di istruzione dei genitori non ha un effetto diretto sul livello di istruzione del bambino. Lo influenza solo attraverso i suoi effetti sul reddito dei genitori e la motivazione del bambino. Quindi, se aggiungiamo il livello di istruzione dei genitori come predittore al modello di regressione multipla (1) per la variabile risposta E, il suo effetto non dovrebbe essere significativo quando il reddito dei genitori e la motivazione del bambino sono nel modello.

L'analisi di regressione condotta come parte dell'analisi causale rivela se esistono evidenze significative dei diversi effetti. Se l'intelligenza influenza direttamente il livello di istruzione, così come indirettamente attraverso il suo effetto sulla motivazione, allora tutti e tre i coefficienti parziali dovrebbero essere significativi. L'effetto diretto dovrebbe essere verificato da un effetto parziale significativo per l'intelligenza nel modello di regressione multipla (1) contenente l'intelligenza, la motivazione, e il reddito dei genitori come predittori del livello di istruzione.

L'effetto indiretto dovrebbe essere rilevato da un effetto parziale significativo per la motivazione in quel modello e un effetto significativo parziale per l'intelligenza sulla motivazione nel modello di regressione multipla (2), contenente anche il livello di istruzione dei genitori come fattore predittivo per la motivazione.

Nello svolgere l'analisi di regressione, se troviamo una relazione causale non significativa, possiamo eliminarla dal diagramma ed eseguire l'analisi più appropriata, procedendo stimando nuovamente i coefficienti delle rimanenti relazioni causali.

Per piccoli campioni, però, occorre tenere presente che un effetto potrebbe non essere significativo, anche se il campione è di dimensioni sufficienti. Per condurre una sofisticata analisi causale analizzando le diverse associazioni dirette e indirette con qualsiasi grado di precisione è richiesto un campione di grandi dimensioni.

ESEMPIO 16.5 Diagramma causale per il livello di istruzione

La Figura 16.2 mostra il diagramma causale della Figura 16.1 con i coefficienti aggiunti. Le variabili residuali per le tre variabili risposta sono indicate con R_1, R_2, e R_3. Se il 28% del livello di istruzione dei bambini è spiegato dai tre predittori, per esempio, allora il coefficiente causale della variabile residuale R_1 per il livello di istruzione del bambino dovrebbe essere $\sqrt{1-R^2} = \sqrt{1-0.28} = 0.85$.

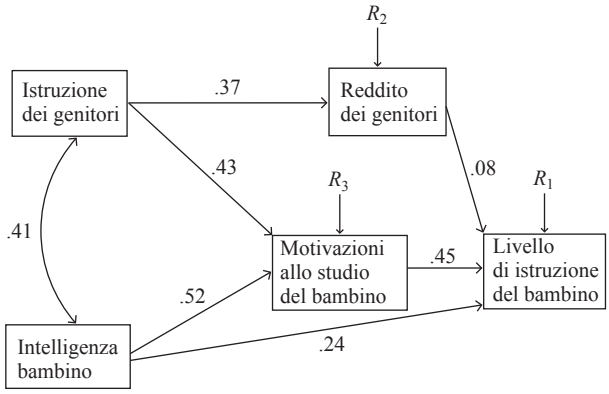

Figura 16.2 Diagramma causale per il livello di istruzione, con i coefficienti causali aggiunti.

La Figura 16.2 mostra che dei tre predittori diretti del livello di istruzione del bambino, il livello di motivazione del bambino ha l'effetto parziale più forte.

L'intelligenza del bambino ha un effetto indiretto moderato attraverso un aumento della motivazione, nonché un effetto diretto. Il reddito dei genitori non è così direttamente importante come la motivazione del bambino o l'intelligenza, ma il livello di istruzione dei genitori ha un effetto rilevante sulla motivazione del bambino. Tali conclusioni sono altamente incerte se i coefficienti causali risentono della presenza di errori di campionamento non trascurabili.

Scomposizioni causali

Un diagramma causale è un modo per ipotizzare la causa responsabile di una associazione tra due variabili. Uno dei risultati fondamentali dell'analisi causale è quello di scomporre la correlazione tra le due variabili in componenti che trattano le diverse relazioni tra queste due variabili.

È più facile illustrare questa idea usando un semplice diagramma causale. Per tre variabili, consideriamo il modello di una relazione a catena, introdotto nel Paragrafo 10.3. Specificatamente,

$$x \longrightarrow z \longrightarrow y$$

Secondo questo modello, la correlazione tra x e y è spiegata con la variabile interveniente, z. Controllando per z, questa associazione dovrebbe scomparire.
La correlazione parziale (Paragrafo 11.7)

$$\rho_{xy \cdot z} = \frac{\rho_{xy} - \rho_{zx}\rho_{zy}}{\sqrt{\left(1-\rho_{zx}^2\right)\left(1-\rho_{zy}^2\right)}}$$

misura l'associazione tra x e y, controllando per z. Per la correlazione parziale affinché $\rho_{xy\cdot z}$ sia uguale a 0, è necessario che

$$\rho_{xy} = \rho_{zx}\rho_{zy}$$

Cioè, la correlazione tra x e y si decompone nella correlazione tra la variabile interveniente e x per la correlazione tra la variabile interveniente e y.

Una generalizzazione di questa formula vale per i diagrammi causali più complessi. Specificatamente, sia $\beta^*_{z_i x}$ il coefficiente causale per il modello nel quale z_i è una variabile risposta e x è un suo predittore. Supponiamo che $\{z_i\}$ è anche un predittore di y in un altro modello. Allora la correlazione tra x e y si decompone in

$$\rho_{xy} = \sum_i \beta^*_{z_i x}\rho_{z_i y}$$

dove la somma interessa tutte variabili z_i che hanno una relazione diretta con y. L'espressione più semplice $\rho_{xy} = \rho_{zx}\rho_{zy}$ fornita precedentemente per la relazione a catena $x \longrightarrow z \longrightarrow y$ corrisponde al caso con una sola variabile z_i, che, pertanto, viene chiamata z. In tal caso, poiché x è l'unica variabile nel modello che predice z, il coefficiente causale di x su z è la correlazione tra loro.

Quanto è utile la decomposizione generale? L'equazione predice ciò che la correlazione *dovrebbe* essere se il diagramma causale fosse corretto. Per i dati del campione, possiamo calcolare la correlazione predetta da questa formula sostituendo le stime del campione nel lato destro. Confrontiamo questa correlazione predetta alla correlazione del campione. Se la differenza tra i due non può essere spiegata dall'errore di campionamento, allora i risultati confutano l'ipotesi causale che il diagramma rappresenta.

Per il modello a catena, per esempio, r_{xy} dovrebbe essere vicino a $r_{zx}r_{zy}$; cioè, la correlazione parziale $r_{xy\cdot z}$ dovrebbe essere vicino allo zero. Il test t per H_0: $\rho_{xy\cdot z} = 0$ (Paragrafo 11.7) è un modo per verificare il modello.

In sintesi, i passaggi base in un'analisi causale sono i seguenti:

1. Impostare una teoria preliminare da verificare, disegnando il diagramma causale senza i coefficienti causali.
2. Condurre l'analisi di regressione per stimare i coefficienti causali e i coefficienti residui.
3. Valutare il modello, verificare se i risultati del campione concordano con esso. Quindi riformulare il modello, eventualmente eliminando le relazioni non significative. Il modello rivisto potrebbe essere la base ulteriori ricerche. Occorre specificare un modello che interpreti il diagramma, modificare e stimare nuovamente i coefficienti causali per quel diagramma.

Un avvertimento sui modelli causali

Dobbiamo aggiungere un importante avvertimento. Affinché la formula della scomposizione dell'analisi causale sia valida, dobbiamo presumere che le variabili non osservate che rappresentano la variabilità residua per ciascuna variabile risposta siano incorrelate con i predittori del modello di regressione per quella risposta. Nelle Figure 16.1 e 16.2, per esempio, tutte le altre variabili che influenzano il livello di istruzione del bambino si presume che siano non correlate con il reddito dei genitori, con la motivazione del bambino e con l'intelligenza del bambino. In pratica, è in dubbio che questo sia esattamente vero.

La realtà, per lo studio degli scienziati sociali non è mai così semplice come nello schema di un diagramma causale. Tale diagramma è una approssimazione grossolana per la realtà. Il vero diagramma dovrebbe essere molto complesso. Un gran numero

di variabili probabilmente giocano un ruolo, con relazioni a coppie, che coinvolgono quasi tutte le variabili.

Un'osservazione analoga vale per i modelli di regressione. Le stime dei parametri nelle equazioni di previsione si riferiscono alle variabili del modello. Se inseriamo altre variabili esplicative che influenzano la variabile risposta, gli effetti stimati cambiano, perché, senza dubbio, le variabili aggiunte sono in qualche modo correlate con i predittori originariamente inseriti nel modello. Questa è una fondamentale caratteristica di tutta la ricerca nelle scienze sociali. Al di là del risultato finale osservato, si potrebbe sempre sostenere che si sarebbero ottenuti risultati diversi con l'inclusione di altre variabili nel modello.

Infine, anche se i dati sono consistenti con un particolare diagramma causale, ciò non implica che il sistema causale rappresentato dal diagramma funzioni veramente. I metodi statistici non possono verificare direttamente l'ordine causale ipotizzato. L'analisi causale non implica causazione da un'associazione, ma si limita a prevedere la struttura per la rappresentazione e la stima degli effetti causali assunti. Per ulteriori dettagli sull'analisi causale, vedi Duncan (1966), Freedman (2005, Capitolo 5), Land (1969) e Pedhazur (1997, Capitolo 18). Per la discussione di alcune delle criticità nel tentativo di utilizzare l'analisi di regressione per scoprire le relazioni causali, si consigliano i libri di Berk (2009), Freedman (2005) e Pedhazur (1997).

16.5 Analisi fattoriale*

L'analisi fattoriale è un metodo statistico multivariato utilizzato per un'ampia varietà di scopi. Questi includono:

1. modelli rivelatori di interrelazioni tra variabili;
2. individuazione di gruppi di variabili, ognuna delle quali contiene variabili fortemente intercorrelate e quindi talvolta ridondanti;
3. riduzione di un gran numero di variabili in un numero minore statisticamente incorrelate, i **fattori** dell'analisi fattoriale.

Il terzo utilizzo è utile per trattare diverse variabili fortemente correlate tra loro. Per esempio, supponiamo che in un modello di regressione multipla ci sia una forte multicollinearità, parzialmente dovuta al gran numero di variabili predittrici utilizzate per valutare ogni elemento di interesse. L'analisi fattoriale può trasformare un insieme di variabili esplicative altamente correlate, in indicatori dello stesso tipo con uno o due fattori aventi quasi lo stesso potere predittivo. Ogni fattore è una combinazione artificiale delle variabili originarie, in che modo questo sia utile dipende dalla interpretabilità dei fattori.

Il modello fattoriale analitico

Il modello fattoriale esprime i valori attesi di un insieme di variabili manifeste $x_1, \ldots, x_k$ come una funzione lineare di un insieme di variabili non osservabili, chiamate **fattori**. Il numero di fattori è indicato con m. Questo deve essere inferiore al numero di variabili k. Il processo utilizza le variabili *standardizzate* e infatti si basa sulla matrice di correlazione delle variabili.

Il modello si compone di k equazioni; ognuna esprime una variabile standardizzata in termini di m fattori. In termini generali, il modello prevede che le variabili manifeste possano essere sostituite da un insieme ridotto di fattori. I fattori del modello dell'analisi fattoriale sono variabili artificiali, ovvero non osservate.

In statistica, le variabili non osservabili sono indicate come **variabili latenti**.

La correlazione di una variabile con un fattore è chiamata **peso** della variabile su quel fattore. Dopo aver condotto un'analisi fattoriale, il software mostra una matrice

con una riga per ogni variabile e una colonna per ogni fattore che mostra questi pesi fattoriali. La somma dei pesi al quadrato per una variabile è chiamata **comunalità** della variabile e rappresenta la proporzione della sua variabilità spiegata dai fattori. Teoricamente, si dovrebbe osservare una elevata comunalità per ogni variabile, usando pochi fattori.

Il processo di adattamento può anche fornire equazioni che esprimono i fattori come funzioni lineari delle variabili osservate. I coefficienti nelle equazioni dipendono dalle correlazioni campionarie tra le coppie di variabili. Per esempio, il primo fattore potrebbe essere messo in relazione a sette variabili osservate standardizzate dalla funzione

$$f_1 = 0.93x_1 + 0.78x_2 - 0.11x_3 + 0.02x_4 + 0.14x_5 - 0.06x_6 - 0.18x_7$$

Questa equazione indica che f_1 principalmente sintetizza l'informazione fornita da x_1 e x_2. L'equazione fattoriale converte i valori sulle k variabili per ogni soggetto in un più piccolo insieme di punteggi sugli m fattori.

Adattamento del modello di analisi fattoriale

Il ricercatore seleziona il numero di fattori da ritenersi adeguato per spiegare le relazioni tra le variabili osservate. Egli spesso può avere una buona intuizione su questo numero osservando la matrice di correlazione con riferimento alle variabili manifeste. Se diversi insiemi di variabili si raggruppano, con forti correlazioni tra coppie di variabili all'interno di ogni insieme e basse correlazioni tra le variabili dei diversi insiemi, allora si potrebbe scegliere un numero di fattori pari al numero dei gruppi.

Una forma **esplorativa** dell'analisi fattoriale ricerca il numero appropriato di fattori. Una guida in tal senso è fornita dagli **autovalori**. L'autovalore per un particolare fattore riassume la percentuale di variabilità delle variabili spiegata da tale fattore. I fattori vengono aggiunti nel modello fino a quando l'introduzione di ulteriori fattori fornisce un miglioramento nella variabilità spiegata. Un'analisi più strutturata chiamata **confermativa** preseleziona un particolare valore per il numero dei fattori. Si può anche assumere una particolare struttura per i pesi fattoriali, come fissare alcuni di loro pari a 0.

Il modello assume che le variabili risposta abbiano una distribuzione **normale multivariata**. In particolare, questo implica che ogni singola variabile abbia una distribuzione normale e che la relazione di regressione tra ciascuna coppia di variabili sia lineare. In pratica, questo non è realistico. C'è una tendenza generalizzata nell'utilizzare questo metodo indipendentemente dalla forma delle distribuzioni, ma queste forti assunzioni dovrebbero imporre cautela sull'utilizzo del metodo con variabili fortemente non normali (per esempio, binarie) o senza un attento controllo dell'effetto di eventuali valori anomali sulle conclusioni finali. Con la maggior parte delle analisi fattoriali esplorative sono presenti così tanti parametri (pesi fattoriali) che non esiste una soluzione unica. Si dice allora che i parametri sono *non identificati*. Per esempio, dati i valori di due variabili e un fattore, ci sono molte possibili soluzioni per le stime dei parametri in

$$f_1 = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Soluzioni diverse possono dare diverse stime dei pesi fattoriali ma corrispondere allo stesso adattamento. Dopo aver ottenuto una soluzione iniziale per i pesi fattoriali, le procedure analitiche fattoriali trattano ogni riga di m pesi fattoriali come un punto nello spazio m-dimensionale e possono ruotare le stime per ottenere fattori più significativi con la struttura fattoriale più semplice. Scopo della rotazione è di portare molti pesi della variabile vicino a 0, in modo che ciascuna variabile sia fortemente

correlata con solo uno o due fattori. Questo rende più semplice l'interpretazione di ciascun fattore in funzione di un particolare sottoinsieme di variabili. La soluzione ruotata riproduce le correlazioni osservate tra le variabili osservate altrettanto bene come la soluzione originaria. Spesso, un fattore è fortemente legato a tutte le variabili. Idealmente, dopo la rotazione, la struttura dei pesi fattoriali potrebbe apparire come mostrato nella Tabella 16.3. Gli 0 nella tabella rappresentano pesi fattoriali che non sono significativamente diversi da zero. Il primo fattore è associato con tutte le variabili, il secondo fattore fornisce informazioni contenute in x_1, x_2, x_3, e il terzo fattore fornisce informazioni contenute in x_4, x_5, x_6.

Tabella 16.3 Semplice struttura di pesi fattoriali di sette variabili su tre fattori.

		Fattore		
		1	**2**	**3**
Variabile	1	*	*	0
	2	*	*	0
	3	*	*	0
	4	*	0	*
	5	*	0	*
	6	*	0	*
	7	*	0	0

*Indica un peso significativamente diverso da zero.

Nella sua forma più semplice, il processo di stima deriva i fattori in modo che la correlazione sia uguale a zero tra ogni coppia. È anche possibile usare rotazioni per le quali i risultanti fattori siano correlati (cioè, rotazioni *non ortogonali*). Spesso questo è più plausibile per le applicazioni delle scienze sociali.

ESEMPIO 16.6 L'analisi fattoriale delle variabili elettorali

Le correlazioni nella Tabella 16.4 si riferiscono alle otto variabili seguenti, misurate in una elezione in 147 distretti a Chicago.

1. Percentuale di voti per il candidato Democratico nella elezione del sindaco
2. Percentuale di voti per il candidato Democratico nelle elezioni presidenziali
3. Percentuale di voti a tutti i candidati di ogni partito (voto di lista)
4. Costo mediano dell'affitto
5. Percentuale di proprietari di abitazioni
6. Percentuale di disoccupati
7. Percentuale di trasferimenti nell'ultimo anno
8. Percentuale di chi ha completato più di dieci anni di scuola

La tabella mostra che le variabili 1, 2, 3 e 6 sono fortemente positivamente correlate, come le variabili 4, 7 e 8. Questo suggerisce che due fattori possono rappresentare queste otto variabili. Adattando un modello fattoriale usando la soluzione **fattori principali** con due fattori produce i pesi fattoriali stimati della Tabella 16.5.

Tabella 16.4 Matrice di correlazione per otto variabili misurate per 147 distretti nelle elezioni di Chicago.

		Variabile Numero							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Variabile Numero	1	1.0							
	2	0.84	1.0						
	3	0.62	0.84	1.0					
	4	−0.53	−0.68	−0.76	1.0				
	5	0.03	−0.05	0.08	−0.25	1.0			
	6	0.57	0.76	0.81	−0.80	0.25	1.0		
	7	−0.33	−0.35	−0.51	0.62	−0.72	−0.58	1.0	
	8	−0.66	−0.73	−0.81	0.88	−0.36	−0.84	0.68	1.0

Fonte: Ristampato da Harman (1967, pp. 165–166) con il permesso di University of Chicago Press.

La tabella dei pesi fattoriali ha $k = 8$ righe, una per ogni variabile osservata, e $m = 2$ colonne, una per ogni fattore.

Tabella 16.5 Pesi fattoriali per una soluzione a due fattori per le correlazioni nella Tabella 16.4.

		Pesi		
		Fattore 1	**Fattore 2**	**Comunalità**
Variabile Numero	1	0.69	−0.28	0.55
	2	0.88	−0.48	1.00
	3	0.87	−0.17	0.79
	4	−0.88	−0.09	0.78
	5	0.28	0.65	0.50
	6	0.89	0.01	0.79
	7	−0.66	−0.56	0.75
	8	−0.96	−0.15	0.94

Il primo fattore è detto **bipolare** perché contiene pesi alti positivi e alti negativi. Le correlazioni positive si verificano con le variabili 1, 2, 3 e 6, per le quali i punteggi alti tendono a verificarsi in distretti con voto fortemente Democratico. Forse questo fattore riassume il tradizionale voto Democratico. Il fattore 2, che è altamente e positivamente correlato con la variabile 5 e negativamente correlato con la variabile 7, è interpretato come una misura di stabilità della dimora. Allorché il punteggio sul fattore 2 per un distretto aumenta, la percentuale di proprietari di casa tende ad aumentare, e la percentuale di coloro che si sono trasferiti nel precedente anno tende a diminuire. La Figura 16.3 riporta i pesi delle variabili sui due fattori. Ogni punto nella Figura 16.3 rappresenta una particolare variabile. Per esempio, il punto etichettato con 4 ha come coordinata x il peso della variabile 4 sul fattore 1 (-0.88) e come coordinata y il peso della variabile 4 sul fattore 2 (-0.09). Il piano mostra che le variabili 1, 2, 3 e 6 si raggruppano, avendo simili coppie di pesi. Inoltre, le variabili 4, 7 e 8 si raggruppano. I valori relativamente grandi per le comunalità indicano che i fattori spiegano la maggior parte della variabilità delle variabili originarie.

Successive analisi di questi dati potrebbero sostituire le otto variabili con questi due fattori. Essi sembrano avere una chiara interpretazione.

Sono incorrelati, in modo che nessuna ridondanza si verifichi quando entrambi sono utilizzati nell'analisi di regressione.

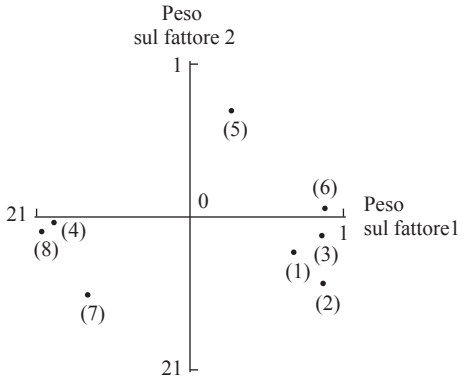

Figura 16.3 Piano dei pesi dalla Tabella 16.5 delle otto variabili sui due fattori.

Due equazioni esprimono ogni fattore in termini delle otto variabili e queste equazioni forniscono punteggi sui due fattori per i 147 distretti.

Un'analisi fattoriale confermativa prevede che sia ipotizzata una struttura ben specificata prima di analizzare i dati. Per esempio, si potrebbe specificare una struttura come nella Tabella 16.3, in cui alcuni pesi fattoriali sono vincolati a essere pari a 0. Questo rende più semplice l'interpretazione dei fattori finali. I test chi-quadro vengono utilizzati per il controllo di una particolare struttura, per verificare se un insieme di parametri assume certi valori prefissati.

Modelli a classi latenti per risposte categoriche

Modelli analoghi di analisi fattoriale sono stati sviluppati per variabili risposta categoriali. Il più semplice è il **modello a classi latenti**; esso prevede che vi sia una variabile latente categoriale non osservata che spiega le associazioni tra le variabili osservate. Condizionatamente alla classe latente di appartenenza, di una determinata unità, le risposte fornite alle variabili osservate sono statisticamente indipendenti.

ESEMPIO 16.7 Modello a classi latenti per l'atteggiamento verso l'aborto

La GSS chiede ai soggetti se sono favorevoli o contrari alla legalizzazione dell'aborto sotto varie condizioni, per esempio ogni volta che la donna lo vuole (variabile ABANY), quando il bambino avrebbe avuto un difetto di nascita (ABDEFECT), quando la donna non vuole altri figli (ABNOMORE), quando la salute della madre è in pericolo (ABHLTH), quando la donna è troppo povera per avere più figli (ABPOOR), quando la donna è incinta a causa di uno stupro (ABRAPE) e quando la donna è single (ABSINGLE). Si può supporre che una variabile latente sottostante descrive l'atteggiamento di base verso l'aborto legalizzato tale che, dato il valore di tale variabile latente, le risposte su queste variabili siano condizionatamente indipendenti.

Un possibile modello a classi latenti avrebbe una sola variabile latente con tre categorie. Ciò porterebbe a ipotizzare che vi è una classe per coloro che si oppongono quasi sempre all'aborto legalizzato indipendentemente dalla situazione, una seconda classe per coloro che quasi sempre sono favorevoli all'aborto legalizzato e una terza classe per coloro le cui risposte dipendono dalla situazione.

Questo modello di base a classi latenti si estende in vari modi. Per esempio, ci potrebbero essere due fattori latenti, ognuno con le proprie categorie, oppure un *modello a variabili latenti* con una caratteristica latente che varia continuamente e si assume che abbia una distribuzione normale.

Origine e dibattito

L'analisi fattoriale fu originariamente sviluppata all'inizio del ventesimo secolo dagli psicometrici nel tentativo di costruire un fattore o dei fattori per misurare l'intelligenza. Charles Spearman postulò l'esistenza di un unico fattore che misura l'intelligenza in generale. In seguito, L.L. Thurstone e altri ipotizzarono un insieme di fattori di gruppo, ciascuno dei quali potrebbe essere misurato con un insieme di test di natura analoga. Per uno sguardo storico e critico sull'utilizzo dell'analisi fattoriale per la misurazione dell'intelligenza, vedi Gould (1981).

Un pericolo che si corre con l'analisi fattoriale è l'errore di *reificazione*, agendo come se un fattore veramente misuri una caratteristica di nostro interesse. In realtà, non sappiamo se ciò accada.

Inoltre, ci sono pericoli statistici utilizzando questo metodo. In ogni analisi che riguarda variabili non misurate come questi fattori artificiali è possibile identificare i modelli in una matrice di pesi fattoriali come suggeriscono alcune interpretazioni per i fattori, quando in realtà quei risultati sono in gran parte dovuti a errori di campionamento. Un controllo che si può fare è di suddividere i dati in modo casuale in due parti e quindi condurre un'analisi fattoriale su ciascuno. Se i risultati sembrano inconsistenti in qualche modo, le previsioni dovrebbero essere tentativi molto incerti e servono principalmente per suggerire modelli di controllo con altri insiemi di dati. Per lungo tempo, la base statistica dell'analisi fattoriale è stata incerta e non era possibile, per esempio, riportare errori standard validi per i pesi fattoriali. Recentemente, sono stati sviluppati metodi di massima verosimiglianza che migliorano molti dei vecchi metodi per lo svolgimento dell'analisi fattoriale. C'è stata anche una crescente attenzione sull'utilizzo dell'analisi fattoriale, più per motivi confermativi che esplicativi, come descritto nel paragrafo seguente. Questo costringe il ricercatore a pensare più attentamente a una struttura fattoriale ragionevole prima di eseguire l'analisi. Allora le conclusioni spurie sono meno probabili. Per ulteriori dettagli, vedi Afifi et al. (2003), DeMaris (2002), Hagenaars e McCutcheon (2006), Harman (1967) e Thompson (2004).

16.6 Modelli a equazioni strutturali*

Un modello molto generale combina gli elementi dell'analisi causale e dell'analisi fattoriale. Il modello è chiamato **modello della struttura di covarianza** poiché esso tenta di spiegare le varianze e le correlazioni tra le variabili osservate. Questa spiegazione prende la forma di un modello causale relativo a un sistema di fattori, alcuni dei quali potrebbero essere creati nell'analisi fattoriale e alcuni dei quali potrebbero essere variabili osservate.

I modelli della struttura di covarianza hanno due componenti; la prima è un **modello di misurazione**, assomiglia a una analisi fattoriale che deriva un insieme di fattori non osservati dalle variabili manifeste; la seconda componente è il **modello a equazioni strutturali**.

Modello di misurazione

Il modello di misurazione specifica come le variabili osservate sono in relazione a un insieme di fattori non osservati, le **variabili latenti**. Questa parte dell'analisi assomiglia a un'analisi fattoriale, tranne che per la struttura fortemente specificata del modello. Il modello di misurazione assegna ogni variabile latente, a priori, a un insieme specifico di variabili osservate. Questo si ottiene forzando determinati pesi fattoriali a essere uguali a 0, in modo che le variabili latenti siano incorrelate con le altre variabili. Il modello di misurazione assume che le variabili osservate, essendo soggette a errori di misurazione e a problemi di validità e di affidabilità, siano indicatori imperfetti dei concetti di vero interesse.

Per esempio, uno studio potrebbe utilizzare le risposte a un insieme di domande su un questionario sugli atteggiamenti razzisti come indicatori grezzi di razzismo. L'analisi fattoriale potrebbe produrre una singola variabile latente che è una misura complessiva di razzismo, migliore di qualsiasi singola risposta. Lo scopo di creare variabili latenti è quello di operativizzare le caratteristiche difficili da misurare, come i pregiudizi, l'ansia e l'atteggiamento conservatore in politica.

Modello a equazioni strutturali

Il modello a equazioni strutturali utilizza i modelli di regressione per specificare le relazioni causali tra le variabili latenti. Una o più variabili latenti vengono identificate come variabili risposta e le altre sono identificate come variabili esplicative. Le variabili risposta latenti possono essere messe in relazione con le variabili esplicative latenti, nonché con altre variabili risposta latenti. A differenza dell'analisi causale, questo approccio consente l'adattamento di modelli con doppio senso di causalità, in cui le variabili latenti possono essere regredite a vicenda.

ESEMPIO 16.8 Modello della struttura di covarianza per l'intelligenza, SES e il profitto

La Figura 16.4, sulla base dell'esempio di Pedhazur (1997), illustra un modello della struttura di covarianza. Il modello analizza gli effetti dell'intelligenza e dello status socioeconomico sul profitto. Le variabili osservate sono gli indicatori di intelligenza, x_1 = punteggio di Wechsler e x_2 = punteggio IQ di Stanford-Binet; gli indicatori dello status socioeconomico, x_3 = istruzione del padre, x_4 = istruzione della madre

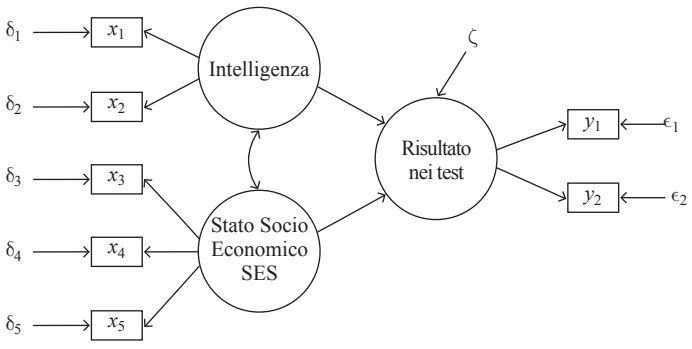

Figura 16.4 Un modello della struttura di covarianza per profitto, intelligenza e status socioeconomico variabili latenti. Si basa su cinque variabili esplicative osservate e due variabili risposta osservate.

e x_5 = reddito totale dei genitori; gli indicatori di successo, y_1 = punteggio verbale e y_2 = punteggio quantitativo nei test di profitto. Gli indicatori di profitto sono le variabili risposta.

Nella Figura 16.4, i rettangoli rappresentano le variabili osservate e i cerchi rappresentano le variabili latenti. Una variabile latente di intelligenza è definita solamente da x_1 e x_2, gli indicatori di intelligenza. Una variabile latente di status socioeconomico è definita solo da x_3, x_4 e x_5, i suoi indicatori. Una variabile latente di profitto riguarda solo y_1 e y_2, gli indicatori di profitto. La figura mostra la dipendenza delle variabili osservate sulle variabili latenti. I legami causali tra le variabili latenti indicano che il profitto è direttamente dipendente dall'intelligenza e lo status socioeconomico e che esiste una associazione tra l'intelligenza e lo status socioeconomico.

Come in un qualsiasi modello di regressione, una variabile non è completamente determinata dalle altre nel sistema. Nella Figura 16.4, i termini δ (delta) ed ε (epsilon), che puntano alle variabili osservate, sono termini di errore rappresentando la variazione di queste variabili che non è spiegata dalle variabili latenti del modello di errore di misurazione. (Ricordiamo la rappresentazione dei modelli di regressione con termini di errore mostrati nel Paragrafo 9.6.) Il simbolo ζ (zeta) rappresenta la variazione non spiegata nel modello a equazioni strutturali, la variabile latente profitto non essendo completamente determinata dalle variabili latenti intelligenza e status socioeconomico.

Casi speciali di modelli della struttura di covarianza

I modelli della struttura di covarianza hanno caratteristiche interessanti di flessibilità e generalità. Un parametro di regressione può essere posto uguale a un valore fisso, per esempio 0. Viene quindi chiamato parametro **fisso**, ovvero un parametro può eguagliare un altro parametro nel sistema. Viene quindi chiamato parametro **vincolato**. Oppure può essere del tutto sconosciuto, parametro **libero**.

Nella Figura 16.4, nel modello di misurazione i pesi fattoriali degli indicatori di intelligenza x_1 e x_2 sullo status socioeconomico e sulle variabili latenti del profitto sono uguali a 0. Quindi questi pesi fattoriali sono parametri fissi. Allo stesso modo, i pesi fattoriali degli indicatori relativi allo status socioeconomico sulle variabili latenti intelligenza e profitto sono pari 0, e i pesi fattoriali degli indicatori di profitto sulle variabili latenti status socioeconomico e intelligenza sono uguali a 0. Di contro, nella parte strutturale dell'equazione del modello, i coefficienti di regressione dell'intelligenza e dello status socioeconomico sul profitto sono parametri liberi. Per trattare una variabile osservata come perfettamente misurata, assumiamo che la corrispondente variabile latente sia identica a quella osservata. I modelli di regressione tradizionali sono casi speciali di modelli della struttura di covarianza che hanno una sola variabile risposta e trattano tutte le variabili come perfettamente misurate.

Supponiamo di trattare tutte le variabili osservate come variabili risposta e concentriamoci su come rappresentarle attraverso un insieme di variabili latenti. Allora il modello fornisce un tipo strutturato di analisi fattoriale.

L'analisi è **confermativa**: ha lo scopo di confermare uno schema di pesi fattoriali per le variabili latenti pre-specificate. Per esempio, in un articolo che riguarda il razzismo, l'atteggiamento conservatore in politica, le azioni positive, e la raffinatezza intellettuale,[1] gli autori hanno creato un fattore da cinque indicatori di razzismo. Molte altre variabili sono state misurate nello studio (come il livello di istruzione), ma queste variabili sono state vincolate ad avere pesi fattoriali pari a 0 sul razzi-

[1] J. Sidanius et al., *Journal of Personality and Social Psychology*, vol. 70, 1996, p. 476.

smo. Un test del chi-quadro ha indicato che i cinque indicatori di razzismo sono stati adeguatamente rappresentati da un unico fattore.

L'analisi fattoriale confermativa contrasta con la natura **esplorativa** dell'analisi fattoriale tradizionale. In un'analisi fattoriale esplorativa, come quella dell'Esempio 16.6, non giudichiamo il numero di fattori importanti e le loro relazioni con le variabili osservate, se non dopo aver visto la matrice dei pesi fattoriali. Con l'analisi fattoriale esplorativa, i risultati potrebbero sembrare interessanti, ma in realtà potrebbero riflettere errori di campionamento.

Adattamenti di modelli della struttura di covarianza

La **covarianza** tra due variabili x e y è la media del prodotto delle loro deviazioni dalle loro medie.

Il suo valore è definito come

$$\text{Cov}(x,y) = E[(x-\mu_x)(y-\mu_y)] = \rho_{xy}\sigma_x\sigma_y$$

cioè, la covarianza è completamente determinata dalla correlazione e dalle deviazioni standard marginali. Una **matrice di covarianza** sintetizza le covarianze per ogni coppia in un insieme di variabili. L'argomento nella riga i e colonna j è la covarianza tra le variabili i e j. Se usiamo le variabili standardizzate, allora le deviazioni standard sono uguali a 1 e la matrice di covarianza è identica alla matrice di correlazione.

La *massima verosimiglianza* è il metodo standard per adattare i modelli della struttura di covarianza e per stimare i parametri. Il software per la stima dei modelli utilizza la covarianza campionaria tra le variabili osservate per stimare i parametri nel modello di misurazione e nel modello a equazioni strutturali. I parametri nel modello a equazioni strutturali sono solitamente quelli di interesse finale. Nel precedente esempio, questi includono i coefficienti di regressione delle variabili latenti intelligenza e status socioeconomico sulla variabile latente profitto. Come nella regressione tradizionale, l'inferenza presuppone variabili risposta normalmente distribuite. Una stima del parametro divisa per il suo errore standard è approssimativamente un test z di significatività. L'interpretazione dell'importanza di una stima dipende dal fatto che le variabili siano misurate in unità originarie o in forma standardizzata.

A meno che il modello non specifichi un numero di parametri sufficientemente elevato, i parametri non sono **identificabili**. Questo significa che non ci sono stime uniche, una situazione che si verifica sempre con l'analisi fattoriale tradizionale. Allora è meglio fissare un numero minimo di parametri, in modo che questo non accada. I software forniscono una guida per il raggiungimento della identificabilità. Se non viene raggiunta, occorre provare a impostare ulteriori pesi fattoriali pari a zero o sostituire un fattore con variabili osservate nella parte strutturale dell'equazione.

I modelli della struttura di covarianza richiedono software specializzati.

LISREL è un noto programma che adatta il modello usando la massima verosimiglianza. Il nome del software è un acronimo per *linear structural relationships*. Un altro programma noto è EQS.

Verifica del modello adattato

Il modello della struttura di covarianza e la combinazione dei parametri fissi, vincolati e liberi, determinano un particolare modello che la vera matrice di covarianza delle variabili osservate dovrebbe soddisfare. Come possiamo controllare l'adattamento del modello della struttura di covarianza? Un modo utilizza un test chi-quadro per grandi campioni per confrontare la matrice di covarianza del campione con la matrice di covarianza stimata dal modello. La statistica test misura quanto è vicina la matrice di

covarianza campionaria al suo valore stimato, se il modello fosse credibile. Più grande è la statistica, peggiore è l'adattamento. Tale test di bontà di adattamento fornisce solo una guida approssimativa. In primo luogo, il test assume normalità multivariata delle variabili osservate, che nel migliore dei casi, è una grezza approssimazione della realtà. In secondo luogo, come altri test chi-quadro, il test di adattamento è un test *globale*. Se il modello si adatta male, ciò non indica cosa causa la mancanza di adattamento. Può essere più informativo visualizzare i residui standardizzati che confrontano i singoli elementi della matrice di varianza e covarianza campionaria basata sul modello. In terzo luogo, come in ogni prova, occorre tenere a mente la dipendenza dei risultati dall'ampiezza del campione. Un risultato potrebbe essere statisticamente significativo, per n grande, senza essere praticamente significativo. Un modo alternativo e più informativo per verificare l'adattamento è quello utilizzato nei modelli di regressione tradizionali: confrontare un dato modello con un modello più complesso con una struttura aggiuntiva che può essere rilevante. Per verificare se il modello più complesso fornisce un adattamento migliore del modello ridotto, la statistica test utilizza la differenza in termini bontà di adattamento del chi-quadrato per i due modelli.

Aspetti positivi dei modelli della struttura di covarianza rispetto all'analisi fattoriale non strutturata sono che (1) i modelli costringono i ricercatori a fornire basi teoriche per le loro analisi, (2) i metodi inferenziali forniscono una verifica dell'adattamento del modello teorico ai dati. Comunque, il modello è complesso. Qualsiasi modello con variabili latenti può richiedere un campione di grandi dimensioni per ottenere buone stime degli effetti, anche per un sistema relativamente modesto di variabili come quello rappresentato nella Figura 16.4. In sintesi, i modelli della struttura di covarianza forniscono un quadro di riferimento versatile per lo svolgimento di una serie di analisi utili nelle scienze sociali. Il loro utilizzo, da parte degli scienziati sociali, è aumentato negli ultimi anni. Tuttavia, la complessità del modello implica che i risultati di interesse siano molto approssimativi a causa delle molte fonti di variabilità. Si consiglia la guida di un esperto di statistica o di un metodologo delle scienze sociali ben addestrato, prima di utilizzare questo metodo. Per ulteriori dettagli, vedi Bentler (1980), Bollen (1989), DeMaris (2002), Jöreskog e Sörbom (1997), Long (1983) e Pedhazur (1997).

16.7 Catene di Markov*

I ricercatori talvolta sono interessati alle sequenze delle risposte nel tempo. Uno studio sui modelli di voto nelle elezioni presidenziali potrebbe analizzare i dati in cui i soggetti indicano il partito per il quale hanno votato in ciascuna delle ultime elezioni. Una sequenza di osservazioni che varia casualmente è chiamato **processo stocastico**. I possibili valori del processo a ogni passo sono gli **stati**. Per esempio, i possibili stati per il voto in un'elezione potrebbe essere Democratico, Repubblicano, Altro, Non-Voto. I modelli stocastici descrivono le sequenze di osservazioni su una variabile.

Uno dei più semplici processi stocastici è la **catena di Markov**. Esso è appropriato se, dato il comportamento del processo ai tempi $t, t-1, t-2, \ldots, 1$, la distribuzione di probabilità dell'esito al tempo $t+1$ dipende solo dall'esito al tempo t. In altre parole, dato il risultato al tempo t, il risultato al tempo $t+1$ è statisticamente indipendente dal risultato in ogni tempo precedente al tempo t.

La proprietà delle catene di Markov *non è* che lo stato al tempo $t+1$ è *independente* dagli stati ai tempi $t-1$, $t-2$ e così via; piuttosto che *condizionatamente* al valore del processo al tempo t, essi siano indipendenti.

Se $y_1, y_2, \ldots$ denotano gli stati successivi della catena, y_{t+1} potrebbe essere associato con $y_{t-1}, y_{t-2}, \ldots$, ma condizionatamente a y_t, y_{t+1} è statisticamente indipen-

dente da $y_{t-1}, y_{t-2}, \ldots$. Le associzioni potrebbero esistere, ma non le associazioni condizionate (i.e., tra y_{t+1} e y_{t-1}, controllando per y_t).

ESEMPIO 16.9 Modellazione della mobilità della classe sociale

Uno studio sulla mobilità della classe sociale dei maschi considera un periodo di tre generazioni, contrassegnato da nonno, padre e figlio. Lo studio segue una discendenza familiare considerando la sequenza dei figli primogeniti a 40 anni. In ogni generazione, i possibili stati del processo sono Superiore, Medio e Inferiore.
Supponiamo che questo processo si comporti come una catena di Markov. Quindi, per esempio, per tutti i padri in una data classe (per esempio superiore), la classe sociale del figlio è statisticamente indipendente della classe sociale del nonno. Utilizzando la barra verticale | per rappresentare *dato* o *condizionato a*, le seguenti quattro probabilità sarebbero identiche:

Pr(figlio in M | padre in S, nonno in I)
Pr(figlio in M | padre in S, nonno in M)
Pr(figlio in M | padre in S, nonno in S)
Pr(figlio in M | padre in S)

Probabilità di transizione

La probabilità considerata nell'esempio sopra è chiamata **probabilità di transizione** di muoversi dalla classe superiore alla classe media in una generazione. La indichiamo con P_{SM}. Un modello a catena di Markov studia problemi come:

- Qual è la probabilità di passare da uno specifico stato a un altro in una determinato intervallo di tempo?
- Quanto tempo occorre, in media, per passare da uno specifico stato a un altro?
- Le probabilità di transizione tra ogni coppia di stati sono costanti nel tempo? Se lo sono, il processo si dice che abbia **probabilità di transizione costante**.
- Il processo è una catena di Markov o la struttura di dipendenza è più complessa?

Le proprietà di una catena di Markov dipendono dalle probabilità di transizione. Queste sono studiate con la *matrice di transizione delle probabilità*, indicata da **P**. Per una catena con s-stati, questa matrice è una tabella $s \times s$.

L'elemento nella cella in corrispondenza della riga i e della colonna j è la probabilità che, dato che la catena è attualmente nello stato i, nel tempo successivo sia nello stato j.

La Tabella 16.6 mostra il formato per una matrice delle probabilità di transizione per l'esempio sulla mobilità sociale, con un insieme di potenziali probabilità di transizione. Le etichette di riga si riferiscono alla classe del padre, e le etichette delle colonne si riferiscono alla classe del figlio. Dalla tabella emerge che, dato che il padre è nella classe superiore, allora la probabilità che il figlio sia nella classe superiore è $P_{SS} = 0.45$, che il figlio sia nella classe media è $P_{SM} = 0.48$, e che il figlio sia nella classe inferiore è $P_{SI} = 0.07$. La somma delle probabilità entro ciascuna riga della matrice è uguale a 1.0.

In pratica, stimiamo le probabilità di transizione dalla proporzione campionaria delle transizioni da uno stato all'altro. Se ci sono 200 coppie padre-figlio con il padre nella classe superiore, e se per 90 di queste coppie il figlio è nella classe superiore, allora $\hat{P}_{SS} = 90/200 = 0.45$.

Nelle applicazioni nelle scienze sociali, di solito è irrealistico aspettarsi che le probabilità di transizione siano stazionarie. Questo limita l'utilità di semplici modelli a catena di Markov.

Tabella 16.6 Esempio di formato della matrice della probabilità di transizione P.

$$\begin{array}{c} \\ \text{Tempo } t \\ \\ \end{array} \begin{array}{c} \\ \begin{array}{ccc} \text{S} & \text{M} & \text{I} \end{array} \\ \begin{array}{c} \text{S} \\ \text{M} \\ \text{I} \end{array} \begin{pmatrix} P_{SS} & P_{SM} & P_{SI} \\ P_{MS} & P_{MM} & P_{MI} \\ P_{IS} & P_{IM} & P_{II} \end{pmatrix} \end{array} \begin{array}{c} \\ \text{Tempo } t+1 \\ \\ \end{array} = \begin{array}{c} \\ \begin{array}{ccc} \text{S} & \text{M} & I \end{array} \\ \begin{array}{c} \text{S} \\ \text{M} \\ \text{I} \end{array} \begin{pmatrix} 0.45 & 0.48 & 0.07 \\ 0.05 & 0.70 & 0.25 \\ 0.01 & 0.50 & 0.49 \end{pmatrix} \end{array}$$

Per dati campionari, sono disponibili i test chi-quadro per verificare le ipotesi di dipendenza di Markov e per la stazionarietà della probabilità di transizione. Anche se il modello a catena di Markov è troppo semplicistico per essere utilizzato, spesso è un componente di un modello più complesso e realistico. Per esempio, vedi Bartholomew (1982), Goodman (1962) e Scheaffer e Young (2008).

Problemi

16.1 Riassumere i vantaggi dell'utilizzare un modello con effetti casuali per analizzare i dati con misure ripetute rispetto al tradizionale utilizzo di misure ripetute ANOVA.

16.2 Spiegare cosa si intende con il termine *modello misto*, e spiegare la distinzione tra un *effetto fisso* e un *effetto casuale*.

16.3 Il Paragrafo 16.1 ha evidenziato che i modelli misti possono trattare i dati anche quando sono presenti osservazioni mancanti. In ogni modo, il metodo assume che i dati siano *casualmente mancanti*. Nell'Esempio 16.1, supponiano che i soggetti che escono dallo studio divengano, nel tempo, meno soddisfatti finanziariamente.

a. Spiegare perché l'assunzione di dati casualmente mancanti sarebbe violata.
b. Spiegare perché l'effetto tempo potrebbe essere sovrastimato usando solo i dati osservati.
c. Descrivere un pattern di uscita per cui le stime dell'effetto trattamento siano distorte.

16.4 Spiegare a cosa serve un modello multilivello. Descrivere un'applicazione in cui questo tipo di modello sarebbe utile.

16.5 Con riferimento alla Tabella 16.2 interpretare gli effetti stimati del genere sul tasso di rischio. Saggiare l'effetto della razza e interpretarlo.

16.6 Uno studio sulla recidività seleziona un campione di osservazioni di persone che sono state rilasciate dal carcere nel 2000. La variabile risposta, misurata quando le osservazioni sono osservate nuovamente nel 2008, è il numero di mesi intercorsi finché la persona è stata nuovamente arrestata. Nel contesto di questo studio, spiegare cosa si intende per osservazione *censurata*.

16.7 Studiando l'effetto della razza sul licenziamenti nella burocrazia federale, uno studio[2] ha usato la event history analysis per modellare il tasso di rischio con riferimento al licenziamento. Nell'analisi, utilizzando un campione di dimensione 2141, gli autori hanno riportato un $P < 0.001$ nei test di significatività per gli effetti parziali di razza ed età. Essi hanno riportato un effetto stimato sul rischio pari a $e^{\hat{\beta}} = 2.13$ per i neri (variabile binaria razza). Spiegare come interpretare il risultato.

16.8 Sia I = reddito annuo, E = livello di istruzione, J = numero di anni di esperienza lavorativa, M = motivazione, A = età, G = genere, e P = livello di istruzione raggiunto dai genitori. Costruire un diagramma causale che mostra la tua opinione sulle relazioni probabili tra queste variabili. Specificare i modelli di regressione necessari per stimare i coefficienti causali per questo schema.

16.9 I dati di fonte Nazioni Unite sono disponibili per molte nazioni su B = tasso di natalità, G = prodotto interno lordo pro capite, L = percentuale di alfabetizzazione, T = percentuale di case aventi una televisione, e C = percentuale uso di contraccettivi. Disegnare un diagramma causale relativo a queste variabili. Specificare quali modelli di regressione occorre adattare per stimare i coefficienti causali per il diagramma.

16.10 Il dataset "Statewide crime" dell'anno 2005 disponibile presso il sito Web contiene i dati sul tasso di omicidi, percentuale di urbanizzazione, percentuale di diplomati, e la percentuale individui in povertà. Non utilizzare l'osservazione per D.C. Costruisci un diagramma causale realistico per queste variabili. Adattando modelli appropriati per questi dati, stima i coefficienti e costruisci il diagramma causale finale. Interpreta i risultati.

[2] C. Zwerling e H. Silver, *American Sociological Review*, vol. 57, 1992, p. 651.

16.11 Con riferimento all"Esempio 1 del Capitolo 11, sui dati per le 67 contee della Florida su $y =$ tasso di criminalità, $x_1 =$ percentuale di diplomati e $x_2 =$ percentuale di soggetti che vivono in un ambiente urbano. Considera il modello causale spurio per l'associazione tra tasso di criminalità e percentuale di diplomati, controllando per la percentuale di urbanizzazione. Usando il dataset "Florida crime" disponibile sul sito web, determina se i dati sono coerenti con questo modello.

16.12 Con riferimento al dataset Statewide crime dell'anno 2005 disponibile sul sito web del testo (edizione USA).

a. Conduci una analisi fattoriale. Quanti fattori sembrano appropriati? Interpreta i fattori, usando i pesi fattoriali stimati.
b. Rimuovi l'osservazione per D.C. e ripetere. Quanto sono sensibili i pesi fattoriali stimati e la vostra identificazione di fattori a quella osservazione?

16.13 Costruisci un diagramma che rappresenta il seguente modello di struttura di covarianza: tre variabili risposta manifeste sono descritte da una sola variabile latente, e tale variabile latente è regredita su quattro predittori osservati.

16.14 Costruisci un diagramma che rappresenta il seguente modello della struttura di covarianza, per le variabili misurate per ogni stato. La variabile risposta latente è basata su due indicatori osservati, il tasso di criminalità violenta e il tasso di omicidi. Le due variabili predittive per tale variabile latente sono i valori osservati della percentuale di residenti in povertà e la percentuale di famiglie monoparentali. Questi sono trattati come perfettamente misurati.

16.15 Con riferimento al precedente esercizio. Usando il software, adattare questo modello al dataset "Statewide crime 2" disponibile sul sito web. Interpreta i risultati.

16.16 Costruisci un diagramma che rappresenta un modello della struttura di covarianza nella seguente situazione: nel modello di misurazione, un singolo fattore rappresenta il tasso di criminalità violenta e il tasso di omicidi e un singolo fattore rappresenta la percentuale dei diplomati delle scuole superiori, percentuale in famiglie povere, e la percentuale di famiglie monoparentali. Nel modello di equazioni strutturali, il primo fattore dipende dal secondo fattore cosi come dalla percentuale di residenti nelle aree urbane.

16.17 Costruisci un diagramma che rappresenta un modello della struttura di covarianza per il seguente caso. Un fattore di religiosità si basa su due indicatori della GSS circa la frequenza con cui si recitano le preghiere e la presenza alle funzioni religiose. Un fattore di educazione si basa su due indicatori della GSS circa il livello di istruzione e di istruzione dei genitori. Un fattore di conservatorismo politico si basa su due indicatori della GSS circa l'ideologia politica. Un fattore di attivismo del governo si basa su tre indicatori della GSS circa la misura in cui il governo dovrebbe essere coinvolto nel ridurre le disparità di reddito e nell'aiutare i membri più poveri della società. Il modello a equazioni strutturali prevede il fattore di conservatorismo politico utilizzando il fattore educazione e il fattore religiosità, predice il fattore di attivismo del governo da altri tre fattori, e permette un'associazione tra l'educazione e i fattori di religiosità.

16.18 Una variabile è misurata in tre tempi, y_1 al tempo 1, y_2 al tempo 2, e y_3 al tempo 3. Supponiamo che la catena di relazione funzioni, con y_1 che influenza y_2, che a sua volta influenza y_3. Questo sequenza di osservazioni soddisfa la dipendenza di Markov? Spiegare.

16.19 Che cosa c'è di sbagliato in questa affermazione? "Per un modello a catena di Markov, y_t è indipendente da y_{t-2}."