Acta Genet Med Gemellol 47: 57-71 (1998)

Breve Storia del Metodo Gemellare 2 - Le Attuali Formulazioni del Metodo

F. Lorenzi

Facoltà di Scienze Matematiche Fisiche e Naturali, Università degli Studi di Roma Tor Vergata, Roma, Italia

Il termine LISREL e l'acronimo di *LI*near *S*tructural *REL*ationship ed è nato inizialmente come nome di un software messo a punto dallo svedese Karl Joreskog e dai suoi collaboratori nei primi anni '70 per stimare, col metodo della massima verosimiglianza, i coefficienti strutturali dei modelli basati su sistemi di equazioni strutturali.

Tali modelli, nella elaborazione tramite il LISREL, rappresentano la sistemazione logica, prima ancora che statistica o computeristica, di tecniche di analisi multivariata le cui prime proposte risalgono all'ininzio del secolo; riconducendo ad un unico modello che ne costituisce una geniale sintesi, approcci ed itinerari scientifici fino ad allora distinti e non comunicanti, quali l'analisi fattoriale, i modelli causali e i modelli di misurazione. In particolare rappresentano in questo momento la più completa e sistematica risposta al problema di operazionalizzare in termini di ricerca e di verifica empirica, nel campo delle scienze sociali, la controversa, ma non per questo meno fondamentale, nozione di causalità. Essi sono quindi la reinterpretazione, sistemazione e soprattutto generalizzazione di quelli che negli anni '60 venivano chiamati i modelli causali e che nella prima metà degli anni '70 avevano conosciuto una notevole popolarità fra i sociologi soprattutto attraverso la tecnica della path analysis.

Ma nello stesso tempo i modelli di equazioni strutturali del LISREL, per il fatto di poter includere nel modello teorico e nella trattazione statistica anche variabili latenti, estendono la loro giurisdizione anche alla vasta famiglia delle tecniche di analisi fattoriale, fornendo a queste una nuova formulazione ed assieme ad essa una nuova legittimità agli occhi dei suoi vecchi (e giustificati) critici; ed offrono nel contempo alle scienze sociali (ed in particolare alla sociologia e psicometria) un terreno comune, derivato questo a sua volta dall'econometria (nell'ambito della quale sono nati infatti i modelli di equazioni strutturali). La principale caratteristica dei modelli di equazioni strutturali secondo il LISREL è infatti proprio la generalità del modello, al cui interno possono essere ricondotti, come casi speciali, una molteplicità di tecniche e di approcci che fino a questo momento avevano vissuto di vita autonoma, separata e non comunicante.

MODELLI DI EQUAZIONI STRUTTURALI

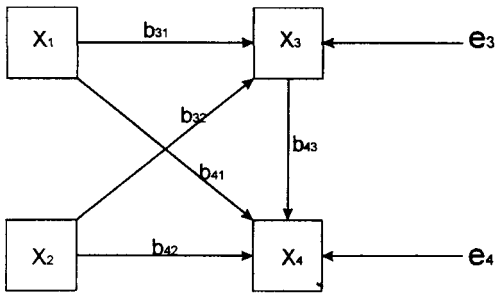
Per modello di equazioni strutturali si intende "*un modello stocastico nel quale ogni equazione rappresenta un legame causale, piuttosto che una mera associazione emprica*" [Goldberger, 1972]. L'unità costitutiva di un modello di equazioni strutturali è l'equazione di regressione che esprime attraverso la formalizzazione matematica la relazione esistente fra una variabile dipendente e diverse variabili indipendenti:

$$X_n = a + b_{n1} X_1 + \ldots + b_{nk} X_k + e_n$$

dove X_n è la variabile dipendente dalle $X_1 \ldots, X_k$ variabili indipendenti; i coefficienti $b_{n1} \ldots, b_{nk}$ si chiamano *coefficienti di regressione* ed esprimono il peso di ogni variabile indipendente nel suo impatto sulla dipendente; e_n e l'errore di predizione ed a l'intercetta. Nelle ipotesi che le variabili vengano espresse in termini di scarti dalle rispettive medie, ovvero che $E(X_i) = 0 \; \forall \; i = 1,\ldots, n$, la costante a viene a scomparire, per cui la nostra equazione diventa:

$$X_n = b_{n1} X_1 + \ldots + b_{nk} X_k + e_n$$

Per definizione di modello di equazioni strutturali il legame tra le variabili indipendenti e dipendenti è di causalità e come tale può essere rappresentato graficamente. Consideriamo il seguente modello definito sulle variabili X_1, X_2, X_3, X_4:



Le equazioni strutturali ad esso associate sono:

$$X_3 = b_{31} X_1 + b_{32} X_2 + e_3$$
$$X_4 = b_{41} X_1 + b_{42} X_2 + b_{43} X_3 + e_4$$

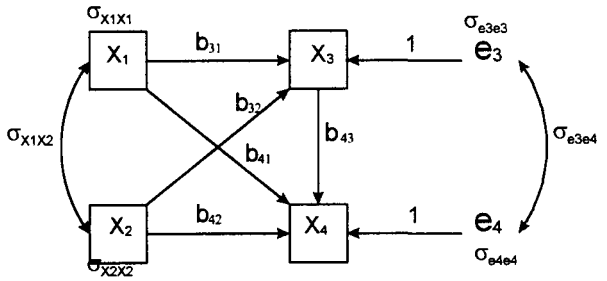
Questo sistema di equazioni, ognuna delle quali rappresenta un nesso causale, è quello che viene chiamato *modello di equazioni strutturali.* In una formulazione più generale, esso potrà essere rappresentato come segue:

$$X_1 = b_{12} X_2 + b_{13} X_3 + \ldots + b_{1k} X_k + \varepsilon_1$$
$$X_2 = b_{21} X_1 + b_{23} X_3 + \ldots + b_{2k} X_k + \varepsilon_2$$
$$\vdots$$
$$X_k = b_{k1} X_1 + b_{k2} X_2 + \ldots + b_{kk-1} X_{k-1} + \varepsilon_k$$

https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press

dove ε_i è *l'errore stocastico* della variabile X_i e i coefficienti b_{ij} si dicono *parametri strutturali* ed indicano quanto la variabile X_i dipende dalla variabile X_j.

Il fatto di lavorare con sistemi di equazioni comporta innanzitutto la perdita del significato di variabile dipendente ed indipendente, in quanto il ruolo di una qualunque variabile X; può variare da equazione ad equazione. Parleremo quindi di variabili *endogene* ed *esogene* dove le prime sono "interne" al modello e possono comparire all'interno delle varie equazioni come dipendenti o indipendenti; le seconde sono quelle "esterne" al modello ed intervengono sempre come variabili indipendenti. Nel nostro esempio X_1 e X_2 sono variabili esogene, mentre X_3 X_4 sono endogene. È necessaria anche una completa revisione del processo di stima dei parametri del modello. Mentre, infatti, nel caso del modello di regressione, caratterizzato da una sola equazione, è normalmente possibile procedere col metodo di stima dei minimi quadrati che tende a minimizzare la somma dei quadrati degli errori, nel caso invece di un modello costituito da più equazioni, dove le variabili indipendenti di un'equazione risultano le dipendenti di un'altra, si introducono enormi complicazioni nel processo di stima dei coefficienti b_{ij}. Noi non ci soffermeremo sulle complicazioni addotte: basti dire che una delle condizioni essenziali della stima dei minimi quadrati, quella dell'indipendenza fra gli errori ε e le variabili indipendenti X, normalmente non si verifica più. Osserviamo infine che, anche se l'interesse principale del ricercatore consiste nello stimare i valori dei parametri b, la struttura di un modello di equazioni strutturali è definita da altri due insiemi di parametri. In particolare si tratta delle varianze e covarianze delle variabili esogene X e delle varianze e covarianze degli errori ε, mentre sono per costruzione pari a zero le covarianze delle X con gli errori stocastici e pari a 1 i parametri che legano ogni errore ε alla propria variabile endogena. Il grafico completo del modello risulta quindi:



LE FASI DEL LISREL

Il punto di partenza di LISREL, cioè il dato empirico da cui l'intero procedimento muove, è dato dalla matrice di varianza-covarianza fra le variabili osservate. Il punto di arrivo è costituito dai parametri di un modello di equazioni strutturali che descrivono i nessi causali fra le variabili. È vero che partendo dai dati (matrice di covarianza osservata) nessuna relazione causale può essere provata, ma è altresì vero che, partendo da una certa relazione causale teorica (cioè ipotizzata), possiamo riprodurre qualche cosa di simile ai dati, cioè una matrice di covarianza teorica che confrontata con l'analoga matrice osservata ci permetterà di capire quanto il nostro modello è compatibile con i dati osservati. È a partire da tale ragionamento che nasce la logica che presiede al procedere del LISREL. Innan-

https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press

zitutto si stabilisce a priori. su base puramente teorica e quindi pre-empirica, il modello causale. Ciò comporta la definizione di un certo numero di parametri che diventano le incognite del modello da stimare. La loro stima avviene facendo interagire modello e dati, trovando cioè quei valori dei parametri che, una volta collocati nel modello, producono lo scarto minore fra matrice di covarianza prodotta dal modello e matrice di covarianza osservata sui dati. Lo stesso scarto è, nella fase successiva, alla base della procedura di falsificazione del modello: se troppo elevato il modello andrà respinto. Con una esposizione un po' più esplicita possiamo dire che LISREL procede secondo tre fasi.

La prima fase è quella della *formulazione del modello teorico*. Si tratta, nel nostro caso, di tradurre la teoria in un sistema di equazioni strutturali, definendo le variabili osservate, ipotizzando le eventuali latenti, stabilendo i legami causali fra le variabili e costruendo il modello complessivo in modo tale che possa essere matematicamente risolubile. Questa procedura porta, come punto conclusivo, alla definizione di un certo numero di parametri da stimare.

La seconda fase è quella della *stima dei parametri strutturali* del modello. Dalla fase teorica, si passa ai dati e, con il modello teorico da una parte e i dati dall'altra, mediante un processo iterativo di minimizzazione delle distanze tra i dati prodotti e quelli osservati, si stimano i parametri incogniti. In pratica si parte attribuendo ai parametri dei valori iniziali più o meno arbitrari, si vede quale matrice di covarianza questo modello produce, si misura la distanza tra questa matrice attesa e quella osservata e con procedure matematiche si minimizza questa distanza. Il processo si chiude quando ogni nuovo tentativo di ridurre ulteriormente lo scarto tra le matrici mediante una modifica dei valori dei parametri, non produce risultati migliori del tentativo precedente. I parametri ottenuti sono conclusivamente i migliori possibili compatibili sia con i dati che con il modello.

La terza fase è quella della *verifica del modello,* cioè del confronto tra modello teorico e dati osservati, per l'eventuale falsificazione del modello stesso. Questa si basa, come si è detto, sul confronto delle matrici di covarianza osservata ed attesa, cioè prodotta dal modello. Sappiamo che la distanza fra le due matrici è la minima compatibile con il modello; tuttavia può essere ancora troppo elevata per poter considerare il modello compatibile con i dati. Si apre a questo punto quella che potremmo chiamare la quarta fase di modifica del modello. Se il modello di partenza si è mostrato inadeguato a descrivere i dati osservati, esso va modificato ed il ciclo di verifica ricomincia.

FASE I: LA FORMULAZIONE DEL MODELLO

Notazioni

Introduciamo rapidamente le notazioni del LISREL. I simboli utilizzati possono essere raggruppati in quattro categorie:

Le variabili	
latenti endogene	η;
latenti esogene	ξ;
osservate endogene	Y;
osservate esogene	X;

https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press

Gli errori stocastici	
delle variabili η	ζ;
delle variabili Y	ε;
delle variabili X	δ;
I coefficienti strutturali	
fra le variabili η e Y	λ^x;
fra le variabili ξ e X	λ^y;
fra le variabili η e η	β;
fra le variabili ξ e η	γ;
Le varianze-covarianze	
fra le variabili ξ	Φ;
fra gli errori ζ	Ψ;
fra gli errori ε	θ^{ε};
fra gli errori δ	θ^{δ};

Rappresentazione grafica

I criteri che presiedono alla rappresentazione grafica di un modello secondo LISREL, sono i seguenti:

1. Le variabili latenti sono racchiuse in un cerchio od ellisse, mentre quelle osservate sono racchiuse in un quadrato o rettangolo; gli errori stocastici sono rappresentati con la lettera corrispondente, senza essere cerchiati.
2. Il legame causale diretto fra due variabili viene indicato con una freccia orientata dalla variabile causa (indipendente) a quella effetto (dipendente). La correlazione fra due variabili senza che sia fornita una interpretazione causale viene invece indicata con una freccia a due direzioni che collega le due variabili; l'assenza di frecce indica l'assenza di relazione.
3. La forza della relazione fra le variabili viene indicata riportando il valore del coefficiente relativo in corrispondenza della freccia; l'assenza di tale valore sta a significare che il coefficiente è assunto pari a 1. Se il parametro strutturale è espresso non in termini numerici, ma simbolici, esso presenterà due indici: il primo che si riferisce alla variabile di arrivo, il secondo alla variabile di partenza.

Dal diagramma si possono quindi ricavare le corrispondenti equazioni strutturali, secondo i seguenti criteri:

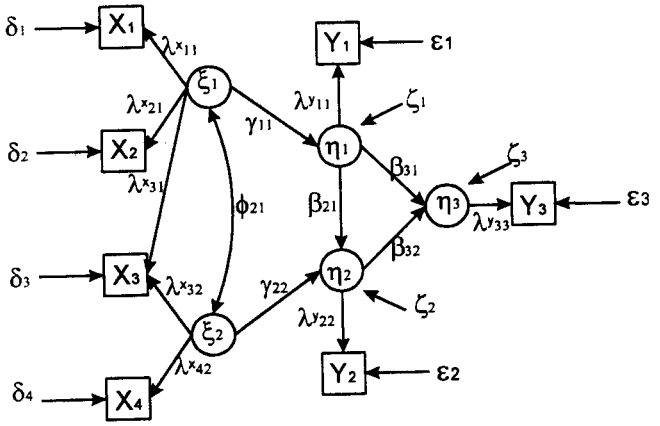
1. Per ogni variabile dipendente va scritta una equazione. La variabile dipendente ne costituirà il primo membro.
2. Il secondo membro dell'equazione è dato dalla somma di tanti addendi quante sono le variabili che agiscono causalmente sulla variabile dipendente contenuta nel primo membro; tali addendi sono costituiti ciascuno dal prodotto della variabile indipendente per il coefficiente associato alla relazione, in più, come addendo conclusivo, va aggiunto l'errore stocastico.

Il modello

Il modello LISREL è costituito da tre parti, ognuna riassumibile in una equazione base:

1. il modello per le relazioni casuali fra le variabili latenti;
2. il modello per la misurazione delle variabili endogene;
3. il modello per la misurazione delle variabili esogene.

Analizziamo da vicino le singole parti con il seguente esempio:



$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1$$
$$\eta_2 = \beta_{21}\eta_1 + \gamma_{22}\xi_2 + \zeta_2$$
$$\eta_3 = \beta_{31}\eta_1 + \beta_{32}\eta_2 + \zeta_3$$

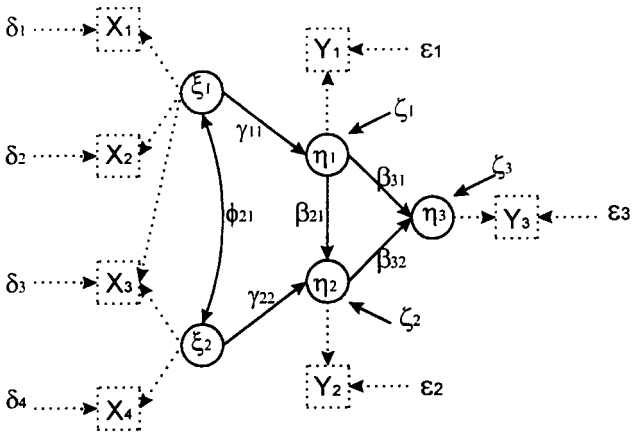
$$X_1 = \lambda^x_{11}\xi_1 + \delta_1$$
$$X_2 = \lambda^x_{21}\xi_1 + \delta_2$$
$$X_3 = \lambda^x_{31}\xi_1 + \lambda^x_{32}\xi_2 + \delta_3$$
$$X_4 = \lambda^x_{42}\xi_2 + \delta_4$$

$$Y_1 = \lambda^y_{11}\eta_1 + \varepsilon_1$$
$$Y_2 = \lambda^y_{22}\eta_2 + \varepsilon_2$$
$$Y_3 = \lambda^y_{33}\eta_3 + \varepsilon_3$$

Il modello strutturale

Il modello strutturale tratta della struttura delle relazioni causali esistente tra le variabili latenti; è questa dunque la parte causale del modello, contrapposta a quella di misura. Consideriamo dal nostro esempio il sottografo relativo alle variabili latenti e le equazioni strutturali associate:

https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press



$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1$$
$$\eta_2 = \beta_{21}\eta_1 + \gamma_{22}\xi_2 + \zeta_2$$
$$\eta_3 = \beta_{31}\eta_1 + \beta_{32}\eta_2 + \zeta_3$$

ovviamente possiamo riscrivere tali equazioni facendo comparire tutte le variabili:

$$\eta_1 = 0\eta_1 + 0\eta_2 + 0\eta_3 + \gamma_{11}\xi_1 + 0\xi_2 + \gamma_{11}\xi_1 + \zeta_1$$
$$\eta_2 = \beta_{21}\eta_1 + 0\eta_2 + 0\eta_3 + 0\xi_1 + \gamma_{22}\xi_2 + 0\xi_1 + \zeta_2$$
$$\eta_3 = \beta_{21}\eta_1 + \beta_{32}\eta_2 + 0\eta_3 + 0\xi_1 + 0\xi_2 + 0\xi_1 + \zeta_3$$

e dunque sintetizzarle nella forma matriciale

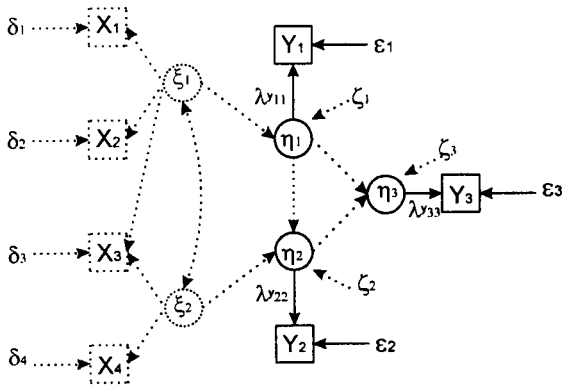
$$\eta = B\eta + \Gamma\xi + \zeta$$

In particolare il vettore η contiene le variabili latenti endogene; ζ i rispettivi errori; il vettore ξ le variabili latenti esogene e le matrici B e Γ hanno per coefficienti rispettivamente i parametri tra le variabili η e η e le variabili η e ξ. Questa parte del modello, per essere completamente specificata, necessita di altre due matrici: si tratta della matrice Φ di covarianza fra le variabili esogene ξ e la matrice Ψ di covarianza tra gli errori ζ.

Il modello di misurazione per le variabili endogene

Le due restanti equazioni base del modello sono relative ai legami tra le variabili osservate. Esse affrontano quindi non il problema della causazione, ma quello della misurazione. Iniziamo dalle variabili endogene: il legame fra le variabili latenti e quelle osservate da luogo alla seconda equazione base del LISREL. Consideriamo dal nostro esempio il sottografo relativo alle variabili endogene e le equazioni strutturali associate:

https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press



$$Y_1 = \lambda^y_{11}\eta_1 + \varepsilon_1$$
$$Y_2 = \lambda^y_{22}\eta_2 + \varepsilon_2$$
$$Y_3 = \lambda^y_{33}\eta_3 + \varepsilon_3$$

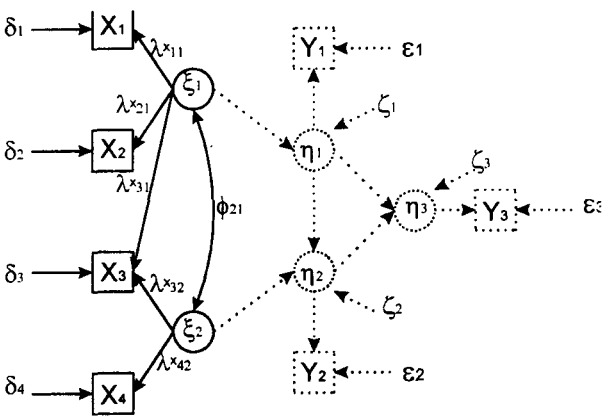
che in forma matriciale diventano:

$$Y = \Lambda_y \eta = \varepsilon$$

dove Λ_y è la matrice dei coefficienti strutturali tra η e Y. La matrice di covarianza degli errori ε viene indicata con Θ_ε.

Il modello di misurazione delle variabili esogene

Le caratteristiche di questa parte del modello sono del tutto analoghe a quelle della sezione precedente, riferite alle variabili esogene. Consideriamo dal nostro esempio il sottografo relativo alle variabili endogene e le equazioni strutturali associate:



https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press

$$X_1 = \lambda^x_{11}\xi_1 + \delta_1$$
$$X_2 = \lambda^x_{21}\xi_1 + \delta_2$$
$$X_3 = \lambda^x_{31}\xi_1 + \lambda^x_{32} + \delta_3$$
$$X_4 = \lambda^x_{42}\xi_2 + \delta_4$$

che in forma matriciale diventano:

$$X = \Lambda_x\xi = \delta$$

dove Λ_x è la matrice dei coefficienti strutturali tra le variabili esogene osservate e latenti. La matrice di covarianza degli errori δ viene indicata con Θ_δ.

Restrizioni

Conclusivamente il modello LISREL necessita per la sua completa specificazione di otto matrici: 4 di coefficienti (B, Γ, Λ_y, Λ_x) e di 4 matrici di covarianza (Φ, Ψ, Θ_ε, Θ_δ). Inoltre le equazioni poggiano sulle seguenti tre assunzioni:

1. Le variabili sono misurate in termini di scarti dalle loro medie cioè:
   $E(\eta) = E(\xi) = 0;$ $E(Y) = E(X) = 0;$ $E(\zeta) = E(\varepsilon) = E(\delta) = 0;$
2. Le variabili indipendenti e gli errori sono fra loro incorrelati, ovvero nella stessa equazione
   $E(\xi\zeta) = 0;$ $E(\eta\varepsilon) = 0;$ $E(\xi\delta) = 0$
   e fra equazioni:
   $E(\eta\delta) = 0;$ $E(\xi\varepsilon) = 0;$
3. Gli errori delle diverse equazioni sono fra loro incorrelati:
   $E(\zeta\varepsilon) = 0;$ $E(\zeta\delta) = 0;$ $E(\varepsilon\delta) = 0;$
4. Nessuna delle equazioni strutturali deve essere ridondante, cioè B è non singolare. Ovvero: le equazioni del modello che esprimono le varie η devono essere fra loro indipendenti, il che significa che nessuna variabile endogena η può essere combinazione lineare delle altre.

FASE II: LA STIMA DEI PARAMETRI DEL MODELLO

Abbiamo visto come dal diagramma teorico si arriva alla definizione delle otto matrici. Nella formulazione del modello ad alcune componenti vengono assegnati dei valori fissi, mentre gli altri parametri rimangono liberi (incogniti) e vanno stimati per poter valutare quantitativamente i nessi causali tra le variabili. In questo paragrafo spiegheremo come vengono stimati questi parametri liberi e come poterci accertare della bontà del modello.

Come abbiamo già detto, i dati in nostro possesso sono costituiti dalla matrice di covarianza tra le variabili osservate X e Y. Vogliamo innanzitutto dimostrare che esiste un legame algebrico tra il modello teorico (le 8 matrici) e la matrice di covarianza delle X

https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press

e Y, quindi perverremo ad una stima dei parametri incogniti. Osserviamo che dimostrare che la matrice di covarianza delle X e Y può essere espressa in funzione delle 8 matrici significa che, se il modello cambia, cambierà anche la matrice di covarianza attesa. In altre parole il modello teorico che abbiamo costruito implica una unica certa matrice attesa di covarianza. Dedichiamoci dunque alla dimostrazione. La matrice di covarianza è del tipo:

$$\Sigma = \begin{array}{c|c|c} & X & Y \\ \hline X & \Sigma_{XX} & \Sigma_{XY}^T \\ \hline Y & \Sigma_{XY} & \Sigma_{YY} \\ \hline \end{array}$$

Dividiamo la dimostrazione in tre parti dedicate rispettivamente alla sottomatrice di covarianza Σ_{XX}, Σ_{YY} e Σ_{XY}. Utilizzeremo le tre equazioni base del LISREL e ricaveremo le matrici di covarianza dalla formula $\Sigma_{xx} = E(XX^T)$. Faremo riferimento anche alle proprietà del modello elencate nel paragrafo delle restrizioni che in sostanza garantiscono l'incorrelazione tra gli errori, e tra le variabili indipendenti e gli errori. Per comodità riscriviamo di seguito le equazioni base trovate:

1. $\eta = \beta\eta + \Gamma\xi + \zeta$
2. $Y = \Lambda_y\eta + \varepsilon$
3. $X = \Lambda_x\xi + \Delta$

Covarianza tra le variabili esogene X

Tenendo conto della terza equazione base

$$\begin{aligned}
&\Sigma_{xx} = E(XX^T) = E[(\Lambda_x\xi + \delta)(\lambda_x\xi + \delta)^T] = \\
&= E[(\Lambda_x\xi + \delta)(\xi^T\Lambda_x^T + \delta^T) = \\
&= E[(\Lambda_x\xi\xi^T\Lambda_x^T + \delta\xi^T\Lambda_x^T + \Lambda_x\xi\delta^T + \delta\delta^T) = \\
&= E(\Lambda_x\xi\xi^T\Lambda_x^T) + E(\delta\xi^T\Lambda_x^T) + E(\Lambda_x\xi\delta^T) + E(\delta\delta^T) = \\
&= \Lambda_x E(\xi\xi^T)\Lambda_x^T + E(\delta\xi^T)\Lambda_x^T + \Lambda_x E(\xi\delta^T) + E(\delta\delta^T)
\end{aligned}$$

Ora $E(\xi\xi^T) = \Sigma\xi\xi = \Phi$, $E(\delta\delta^T) = \Sigma\delta\delta = \Theta\delta$, mentre gli altri addendi sono nulli per la proprieta di incorrelazione, da cui:

$$\Sigma_{xx} = \Lambda_x\Phi\Lambda_x^T + \Theta\delta$$

Covarianza tra le variabili endogene Y

Allo stesso modo, tenendo conto della seconda equazione base

$$\begin{aligned}
&\Sigma_{YY} = E(YY^T) = E[(\Lambda_y\eta + \varepsilon)(\Lambda_y\eta + \varepsilon)^T] = \\
&= E[(\Lambda_y\eta + \varepsilon)(\eta^T\Lambda_y^T + \varepsilon^T)] = \\
&= E(\Lambda_y\eta\eta^T\Lambda_y^T + \varepsilon\eta^T\Lambda_y^T + \Lambda_y\eta\varepsilon^T + \varepsilon\varepsilon^T) = \\
&= E(\Lambda_y\eta\eta^T\Lambda_y^T) + E(\varepsilon\eta^T\Lambda_y^T) + E(\Lambda_y\eta\varepsilon^T) + E(\varepsilon\varepsilon^T) = \\
&= \Lambda_y E(\eta\eta^T)\Lambda_y^T + E(\varepsilon\eta^T)\Lambda_y^T + \Lambda_y E(\eta\varepsilon^T) + E(\varepsilon\varepsilon^T) = \\
&= \Lambda_y E(\eta\eta^T)\Lambda_y^T + \Theta\varepsilon
\end{aligned}$$

https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press

Ora dalla prima equazione base ricaviamo che

$$\eta = B\eta + \Gamma\xi + \zeta \Leftrightarrow (I - B)\eta = \Gamma\xi + \zeta \Leftrightarrow \eta = (I - B)^{-1}(\Gamma\xi + \zeta).$$

Quindi

$$\begin{aligned}
\Sigma_{\eta\eta} &= E(\eta\eta^T) = E[((I - B)^{-1}(\Gamma\xi + \zeta))((I - B)^{-1}(\Gamma\xi + \zeta))^T] = \\
&= E[((I - B)^{-1}(\Gamma\xi + \zeta))(\xi^T\Gamma^T + \zeta^T)(I - B)^{-1T}] = \\
&= (I - B)^{-1} E(\Gamma\xi\xi^T\Gamma^T + \zeta\xi^T\Gamma^T + \Gamma\xi\zeta^T + \zeta\zeta^T)(I - B)^{-1T} = \\
&= (I - B)^{-1}[E(\Gamma\xi\xi^T\Gamma^T) + E(\zeta\xi^T\Gamma^T) + E(\Gamma\xi\zeta^T) + E(\zeta\zeta^T)](I - B)^{-1T} = \\
&= (I - B)^{-1}[\Gamma E(\xi\xi^T)\Gamma^T + E(\zeta\xi^T)\Gamma^T) + \Gamma E(\xi\zeta^T) + E(\zeta\zeta^T)](I - B)^{-1T} = \\
&= (I - B)^{-1}[\Gamma\Phi\Gamma^T + \Psi](I - B)^{-1T}
\end{aligned}$$

Per cui

$$\Sigma_{YY} = \Lambda_y (I - B)^{-1}[\Gamma\Phi\Gamma^T + \Psi](I - B)^{-1T}\Lambda_y^T + \Theta\varepsilon$$

Covarianza tra le variabili endogene Y e quelle esogene X

$$\begin{aligned}
\Sigma_{xy} &= E(xy^T) = E[(\Lambda_x\xi + \delta)(\Lambda_y\eta + \varepsilon)^T] = \\
&= \Lambda_x E(\xi\eta^T)\Lambda_y = \Lambda_x E(\xi\varepsilon^T) + E(\delta\eta^T)\Lambda_y^T + E(\delta\varepsilon^T) = \\
&= \Lambda_x E(\xi\eta^T)\Lambda_y
\end{aligned}$$

in quanto le variabili indipendenti sono incorrelate con gli errori e gli errori sono incorrelati tra loro.

In particolare:

$$\begin{aligned}
E(\xi\eta^T) &= E[\xi((I - B)^{-1}(\Gamma\xi + \zeta))^T] = E[\xi(\xi^T\Gamma^T + \zeta^T)(I - B)^{-1T}] = \\
&= [E(\xi\xi^T)\Gamma^T + E(\xi\zeta^T)](I - B)^{-1T}(\Phi\Gamma^T(I - B)^{-1T}
\end{aligned}$$

Dunque

$$\Sigma_{xy} = \Lambda_x\Phi\Gamma^T(I - B)^{-1T}\Lambda_y$$

Con questo si conclude la dimostrazione. Riassumendo la matrice di covarianza attesa sarà

$$\Sigma = \begin{array}{c|cc} & X & Y \\ \hline X & \Sigma_{XX} & \Sigma_{XY}^T \\ Y & \Sigma_{XY} & \Sigma_{YY} \end{array}$$

dove

$$\begin{aligned}
\Sigma_{xx} &= \Lambda_x\Phi\Lambda_x^T + \Theta\delta \\
\Sigma_{YY} &= \Lambda_y(I - B)^{-1}[\Gamma\Phi\Gamma^T + \Psi](I - B)^{-1T}\Lambda_y^T + \Theta\varepsilon \\
\Sigma_{xy} &= \Lambda_x\Phi\Gamma^T(I - B)^{-1T}\Lambda_y
\end{aligned}$$

https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press

FASE III: LA STIMA DEI PARAMETRI STRUTTURALI

Il metodo di stima che il LISREL usa è quello chiamato della massima verosimiglianza. Il criterio di massima verosimiglianza è un criterio generale, che permette di stimare i parametri incogniti della popolazione individuando quei parametri che generano la più elevata probabilità per i dati campionari di essere osservati. Nel nostro caso specifico esso consiste nell'individuare, data una certa matrice di covarianza S, qual è la probabilità che questa matrice derivi da una certa matrice teorica Σ; e permette, premesso che siano liberi alcuni parametri del modello, di determinare quali valori attribuire a tali parametri affinché la probabilità che S derivi da Σ sia massima possibile. Per poter proseguire dobbiamo a questo punto essere in grado di calcolare la probabilità di ottenere un certo S dato Σ.

Va inoltre detto che il processo di stima è iterativo: il punto di partenza è costituito da prime stime dei parametri ottenute secondo il metodo dei minimi quadrati a due stadi e termina quando ogni tentativo di massimizzare ulteriormente la funzione di verosimiglianza non produce risultati migliori (o più apprezzabili) dei precedenti.

LA VALUTAZIONE E IL MIGLIORAMENTO DEL MODELLO

Stimati i parametri liberi e prodotta la matrice di covarianza attesa Σ, torniamo alla matrice iniziale in quanto è sullo scarto $S - \Sigma$, che viene chiamato residuo, che fondiamo il nostro test di falsificazione del modello. Se lo scarto è eccessivo allora il modello non può essere considerato compatibile con i dati: troppo distanti sono le matrici di covarianza. Se lo scarto invece può essere addebitabile ad oscillazioni stocastiche allora il modello non risulta falsificato e non viene respinto. A questa prima e prioritaria fase di valutazione dell'adattamento del modello ai dati segue una seconda fase: quella del miglioramento del modello sulla base dell'analisi dei parametri stimati. Si tratta, anche in questo caso, di una procedura iterativa fino a che il modello non è più migliorabile. Di seguito affronteremo le diverse fasi dell'intero processo. Tuttavia, poiché questa analisi si basa sui residui delle covarianze, dobbiamo capire meglio il legame della matrice di covarianza Σ con i parametri del modello.

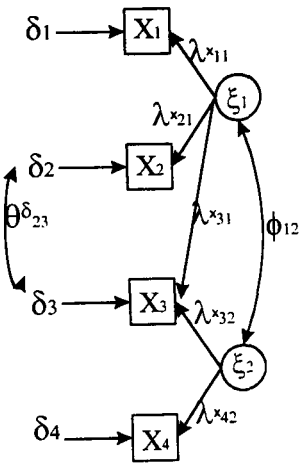
Le covarianze espresse in funzione dei parametri

Esistono due regole pratiche di espressione delle varianze e covarianze in termini dei parametri del modello dette anche regole di scomposizione. Definiamo innanzitutto il percorso tra due variabili come il tracciato fatto di sequenze di frecce direzionali che unisce due variabili.

Prima regola di scomposizione: la covarianza fra due variabili può essere scomposta in tanti addendi quanti sono i percorsi che la collegano; ogni addendo è dato dal prodotto dei coefficienti incontrati sul percorso. Se il percorso passa per due variabili indipendenti, nella formula compare la covarianza delle due; se il percorso ne attraversa solo una nella formula ne compare la varianza.

Seconda regola di scomposizione: la varianza di una variabile dipendente si scompone in varianza spiegata e varianza non spiegata. La seconda è la varianza dell'errore; la

prima è data da tanti addendi quante sono le variabili agenti causalmente in modo diretto su quella variabile. Ogni addendo è dato dal legame diretto fra le due variabili moltiplicato per la somma di tutti i legami diretti e indiretti fra le due variabili.



$$X_1 = \lambda^x_{11}\xi + \delta_1$$
$$X_2 = \lambda^x_{21}\xi_1 + \delta_2$$
$$X_3 = \lambda^x_{31}\xi_1 + \lambda^x_{32}\xi_2 + \delta_3$$
$$X_4 = \lambda^x_{42}\xi_2 + \delta_4$$

A titolo di esempio consideriamo il modello in figura e calcoliamo la covarianza tra la variabile X_1 e X_3 e la varianza di X_3 applicando le regole di scomposizione:

$$\text{Var}\,(X_3) = \sigma_{33} = \lambda^x_{31}\,(\lambda^x_{31} + \lambda^x_{32}\,\Phi_{21}) + \lambda^x_{32}\,(\lambda^x_{32} + \Phi_{21}\lambda^x_{31}) + \theta^\delta_{33}$$
$$\text{Cov}\,(X_1X_3) = \sigma_{13} = \lambda^x_{11}\lambda^x_{31} + \lambda^x_{11}\,\Phi_{21}\lambda^x_{32}) + \lambda^x_{11}\,(\lambda^x_{31} + \Phi_{21}\lambda^x_{32})$$

Verifichiamo esplicitamente che queste espressioni equivalgono a quelle trovate dalla formula per il calcolo della matrice di covarianza

$$\Sigma_{xx} = \Lambda_x\Phi\Lambda_x^T + \Theta\delta$$

$$\Lambda_x = \begin{vmatrix} \lambda^x_{11} & 0 \\ \lambda^x_{21} & 0 \\ \lambda^x_{31} & \lambda^x_{32} \\ 0 & \lambda^x_{42} \end{vmatrix} \qquad \Phi = \begin{vmatrix} 1 & \Phi_{12} \\ \Phi_{21} & 1 \end{vmatrix} \qquad \Phi_\delta = \begin{vmatrix} \theta^\delta_{11} & 0 & 0 & 0 \\ 0 & \theta^\delta_{22} & \theta^\delta_{32} & 0 \\ 0 & \theta^\delta_{32} & \theta^\delta_{33} & 0 \\ 0 & 0 & 0 & \theta^\delta_{44} \end{vmatrix}$$

$$\Sigma_{xx} = \begin{vmatrix} \lambda^{x\,2}_{11} + \theta^\delta_{11} & & & \\ \lambda^x_{11}\lambda^x_{21} & \lambda^x_{21} + \theta^\delta_{22} & & \\ \lambda^x_{11}\,(\lambda^x_{31} + \lambda^x_{32}\,\Phi_{12}) & \lambda^x_{21}\,(\lambda^x_{32}\Phi_{12} + \lambda^x_{32}) + \theta^d_{32} & \lambda^{x2}_{31} + 2\lambda^x_{32}\lambda^x_{32}\Phi_{12} + \lambda^{x\,2}_{32} + \theta^d_{33} & \\ \lambda^x_{11}\lambda^x_{42}\Phi & \lambda^x_{21}\lambda^x_{42}\Phi_{12} & \lambda^x_{42}\,(\lambda^x_{32} + \lambda^x_{31}\Phi_{12}) & \lambda^x_{42} + \theta^d_{44} \end{vmatrix}$$

https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press

Misure di adattamento complessivo del modello

Come si è detto, le misure di adattamento del modello ai dati sono tutte funzioni del residuo, cioè dello scarto fra S e Σ. Ci sono vari modi per misurare questo scarto, comunque si può dimostrare che se il modello è corretto ed il campione sufficientemente grande, allora la funzione di adattamento si distribuisce secondo la distribuzione del χ^2 con df gradi di libertà, dove df = 1/2 (p+q) (p+q+1) – t essendo t il numero dei parametri liberi, p il numero delle variabili Y e q quello delle variabili X. In particolare

df = n° varianze-covarianze – n° parametri liberi = n° dei parametri fissi

Si può quindi vedere come i gradi di libertà misurino anche la parsimoniosità del modello: maggiore è il numero di gradi di libertà maggiore è la capacità di semplificazione della realtà da parte del modello. In questo senso un ricercatore, nel processo di miglioramento del modello deve andare nella direzione della semplificazione del modello e quindi dell'innalzamento dei gradi di libertà. Queste considerazioni devono entrare anche nel criterio di valutazione complessiva: fra due modelli con lo stesso livello di significatività T il ricercatore privilegerà quello più parsimonioso. Per questo motivo è stata proposta, come misura di valutazione complessiva non il semplice valore T ma il rapporto

T/df = indice di bontà

che tiene conto non solo dell'adattamento fra S e Σ ma anche della parsimoniosità del modello. Accettabile per la non falsificazione del modello è un rapporto compreso tra 1 e 3.

Il miglioramento del modello

I processi di miglioramento del modello seguono tre direzioni principali:

1. esclusione di alcuni parametri;
2. inclusione di nuovi parametri;
3. riformulazione del modello.

Entrano in gioco ora l'analisi dettagliata dei singoli parametri, degli scarti e delle covarianze. Iniziamo con l'analisi dei parametri strutturali evidenziando quelli stimati con valori molto bassi, ovvero non significativamente diversi da zero. Il valore stimato può in questo caso essere diverso da zero a causa delle oscillazioni stocastiche per cui possiamo escluderlo dai parametri liberi e fissarlo pari a zero.

Al contrario, possiamo cercare di migliorare il modello, liberando alcuni parametri che avevamo posto fissi sulle basi teoriche della formulazione del modello. Osserviamo che sia l'eliminazione che la liberazione dei parametri dovrà essere effettuata un parametro alla volta, con successiva stima del modello, in quanto la modifica anche di un solo parametro comporta una variazione anche significativa di tutti gli altri.

Infine approfondiamo il terzo criterio di miglioramento quello della riformulazione del modello che si basa sull'analisi dei singoli scarti tra le covarianze $s_{ij} - \sigma_{ij}$. All'origine di questa procedura sta il meccanismo di scomposizione delle covarianze: la covarianza

https://doi.org/10.1017/S0001566000000386 Published online by Cambridge University Press

tra due variabili e il prodotto di tutti i percorsi esistenti tra queste due. Se il modello sottoposto a stima non include tutti i percorsi effettivamente esistenti tra le variabili i e j, allora la covarianza σ_{ij} stimata dal modello risulterà inferiore a quella osservata s_{ji} e le due variabili presenteranno un residuo positivo $s_{ij} - \sigma_{ij} > 0$. Nello stesso tempo gli altri percorsi fra i e j risulteranno avere parametri sovradimensionati rispetto a quelli reali, in quanto la procedura di stima cerca di avvicinare il più possibile le covarianze stimate a quelle osservate. Mancando nel modello alcuni dei percorsi fra le variabili i e j parte della covarianza dovuta ai percorsi mancanti verrà assorbita dai percorsi esistenti, con un gonfiamento dei rispettivi parametri. Una volta individuati i residui elevati abbiamo a disposizione tre vie per eliminarli:

1. introdurre fra le variabili presenti dei legami aggiuntivi che in modo diretto o indiretto coinvolgano le variabili con il residuo elevato;
2. introdurre delle nuove variabili latenti che agiscano quelle affette dai residui elevati;
3. introdurre dei legami tra gli errori delle variabili con i residui elevati.

Ringraziamenti: Ringrazio vivamente il Dott. G. Brenci per il suo prezioso aiuto, le informazioni storiche sulla gemellologia ed il supporto ricevuto nel corso delle mie ricerche.

BIBLIOGRAFIA

1. Azzalini (1994): Inferenza statistica. Un'introduzione basata sul concetto di verosomiglianza. Springer-Verlag.
2. Boomsma DI, Molenaar PCM (1987): Constrained maximum likehood analysis of familial resemblance of twins and their parents. Acta Genet Med Gemell 36: 29-39.
3. Corbetta P (1992): *Metodi di analisi multivariata per le scienze sociali. Il Mulino.*
4. Eaves LJ, Krysyna AL, Young PA, Martin NG (1978): Model-fitting approaches to the analysis of human behaviour. Heredity 41: 249-320.
5. Heath AC (1987): The analysis of marital interaction in cross-sectional twin data. Acta Genet Med Gemell 36: 41-49.
6. Huitema BE: The analysis of covariance and alternatives. J. Wiley & sons, New York, Chichester, Bribane, Toronto.
7. Lathrope GM, Lalovel JM, Jacquard A (1984): Path analysis of family resemblance and gene-environment interaction. Biometrics 40: 611-625.
8. McArdle JJ (1986): Latent variable growth within behaviour genetic models. Behav Genet 16: 163-200.
9. Mood AM, Graybill FA, Boes DC (1988): Introduzione alla statistica. McGraw Hill.
10. Morrison DF (1967): Multivariate statistical methods. McGraw Hill.

Corrispondenza: Dr.ssa Francesca Lorenzi, Via dei Carraresi 18/E, 00164 Rome, Italy; e-mail francesca.lorenzi@usa.net.