Multisample

Regressione

Se ho due variabili (X e Y) e faccio una regressione, ma scopro che esiste una terza variabile (dicotomica) che mi fa cambiare i parametri, suddividendo il campione in due gruppi... cosa posso fare?

1. Faccio analisi separate per i due gruppi, quindi confronto i singoli parametri standardizzati e commento quelli non standardizzati;

2. Uso le cosiddette "variabili *dummy*" (falso, fittizio).

Ad es. Nello spiegare gli omicidi nelle varie provincie italiane in base al tasso di disoccupazione trovo una differenza fra Nord e Sud (era un'esercitazione). Posso fare allora due analisi separate, oppure posso usare la tecnica dei *dummy*.

Usando una variabile dicotomica, l'equazione

$$y = b_0 + b_1 X + e$$

si trasforma in

$$y = b_0 + b_1 X + b_2 S + e$$

dove S vale 0 per il Nord e 1 per il Sud (o viceversa). Nella regressione lineare semplice o multipla, solo la variabile indipendente può essere dicotomica e si posso usare più variabili dicotomiche.

Se b_1 può essere pensata come "di quanto cambia Y al variare di X" , analogamente b_2 può essere considerato come un'*aggiustamento* che viene introdotto quando S vale 1, ma non quando vale 0.

Quindi è come se avessimo 2 equazioni:

$$\begin{aligned} Y = \ & b_0 + b_1 X + 0 + e \quad \text{per il Nord} \\ Y = \ & b_0 + b_1 X + b_2 + e \quad \text{per il Sud} \end{aligned}$$

Se ho una variabile categoriale divisa in 3 valori (Nord, Centro, Sud), creo due variabili dicotomiche:

S_1	S_2	
1	0	Nord
0	0	Centro
0	1	Sud

Quindi è come se avessimo 3 equazioni:

$$\begin{aligned} Y = \ & b_0 + b_1X + 0 + 0 + e \quad \text{per il Centro } (S_1 = S_2 = 0) \\ Y = \ & b_0 + b_1X + b_2 + 0 + e \quad \text{per il Nord } (S_2 = 0) \\ Y = \ & b_0 + b_1X + 0 + b_3 + e \quad \text{per il Sud } (S_1 = 0) \end{aligned}$$

b_2 è la differenza fra Nord e Centro; b_3 la differenza fra Sud e Centro.

Essendo delle costanti, quello che otteniamo è una sorta di intercetta (l'ipotesi è che tutte le rette siano delle parallele che differiscono solo per la costante).

Se invece le varie rette hanno "pendenze" diverse (graficamente sono intersezioni di rette), posso operare in modo simile.

Costruisco una variabile fittizia che moltiplica X per la dicotomica S (o le dicotomiche).

$$W = SX \qquad W = S_1X \text{ e } V = S_2X$$

Le equazioni diventeranno:

$$\begin{aligned} Y = \ & b_0 + b_1X + 0 + 0 + e \quad \text{per il Nord (S=W=0)} \\ Y = \ & b_0 + b_1X + b_2 + b_3W + e \quad \text{per il Sud} \end{aligned}$$

I parametri b_2, b_3 e simili sono parametri di regressione a tutti gli effetti e possono essere verificati con le solite formule di F o con i t.

Riepilogando:

normale	Y=a+bX+e
parallele	Y=a+cS+bX+e
intersezioni	Y=a+cS+bX+dW+e

Siccome il modello normale è un modello ridotto rispetto a quello parallelo che è ridotto rispetto a quello intersecato, useremo il modello gerarchicamente più complesso come R_f^2 e quelli meno complessi come R_r^2.

Lisrel

In Lisrel diventa un'analisi di due o più campioni contemporaneamente. [E' possibile anche fare 2 analisi separate dei 2 campioni e trovare 2 blocchi di parametri diversi].

Ma in un multisample, vediamo se gli stessi parametri possono funzionare abbastanza bene (χ^2 non significativo) in entrambi i gruppi. Dal momento che i dati sono espressi come scarti dalla media non avremo le intercette.

Usiamo un nuovo sottocomando `NG=` da mettere nel comando `DA`.

Prima descriviamo il primo campione e creiamo il modello del primo campione, poi (dopo `OU`) descriviamo il secondo campione, indicando SOLO le cose che cambiano.

```
NORD   !sottolineate le diversità
DA NO=61 NI=7 MA=CM NG=2
CM=NORD.COV
LA; ...
SE; ...
MO NY=3 NX=.... GA=FU,FR
FR ....
FI ....
OU

SUD
DA NO=34 NI=7
CM=SUD.COV
...
MO .... GA=IN BE=IN PS=IN
PD; OU
```

`NG=2` significa che abbiamo 2 gruppi (blocchi DA-OU); `=IN` significa INVARIANT (invariante, non cambia); `PD` funziona solo dopo l'ultimo `DA`.

In questo caso, l'esatto identico modello viene verificato sui due campioni diversi. Il χ^2 sarà uno solo (la somma dei due parziali), i gradi di libertà considerano tutti i parametri dei due modelli (gruppi x covarianze - parametri da stimare)

$$NG\frac{o(o+1)}{2} - t$$

Non sono obbligato a porre tutto come invariante. Posso anche stimare i modelli separatamente, poi verificare che abbiano senso anche messi assieme.

Succede soprattutto con le analisi fattoriali.

Ipotizziamo due campioni (scuola pubblica e scuola privata), su cui sono misurate quattro osservate (VERBAL40 VERBAL50 MATH35 MATH25). Ipotizziamo che esistano due fattori (VERBAL e MATH) e che sia possibile stimare delle saturazioni identiche nei due gruppi.

Se fosse così, potremmo poi calcolare i punteggi fattoriali che potremmo poi confrontare fra i due gruppi (ad es. con ANOVA). Se le due analisi fattoriali avessero saturazioni diverse, non potremmo confrontare fra loro i punteggi fattoriali.

```
Gruppo 1 ! diversità in sottilineato
da ni=4 NO=865 NG=2
LA; VERBAL40 VERBAL50 MATH35 MATH25
cm=EX10_1.COV
mo nx=4 nk=2 LX=fi PH=SY,FI TD=DI,FR
lK; Verbal Math
PA LX
1 0
1 0
0 1
0 1
FR ph 2,1
va 1 ph 1,1 ph 2,2
ou

Gruppo 2
da ni=4 NO=900
```

```
cm=EX10_2.COV
mo LX=IN PH=IN
pd; ou
```

Prima di effettuare le analisi, indica che:

```
Parameter Specifications

 LAMBDA-X EQUALS LAMBDA-X IN THE FOLLOWING GROUP
 PHI EQUALS PHI IN THE FOLLOWING GROUP
```

Dopo il primo gruppo, fornisce l'adattamento:

```
          Group Goodness of Fit Statistics

        Contribution to Chi-Square = 5.48
     Percentage Contribution to Chi-Square = 50.40
```

e dopo l'ultimo gruppo fornisce tutti gli indici di adattamento:

```
       Global Goodness of Fit Statistics

             Degrees of Freedom = 7
   Minimum Fit Function Chi-Square = 10.87 (P = 0.14)
Normal Theory Weighted LS Chi-Sq. = 10.90 (P = 0.14)
```

Come vedete i gradi di libertà sono:

$$2 \times \frac{4 \times 5}{2} - (4\lambda + 8\theta^{\delta} + 1\phi) = 20 - 13 = 7$$

E il modello globale va abbastanza bene, per cui possiamo ipotizzare le stesse saturazioni per entrambi i gruppi.

In ogni caso, si dovrebbe iniziare verificando lo stesso modello per tutti i gruppi. Se non funziona, si stimano separatamente i modelli e poi si copiano nel multisample. Dovrebbero essere per lo meno "nested".