Lezione 15 (a cura di Giovanni Mariani)

Regressione Multivariata

Consideriamo $y_i....y_r$, con r = numero variabili risposta
$x_i....x_k$, con k = numero varibili esplicative

Nel modello di regressione Multivariata abbiamo più variabili risposta (tipicamente poche), in particolare avremo:

$\mathbf{y_{ij}}$ dove i= 1,2...n indica l'i-esimo soggetto
j=1,2.....r indica la j-esima variabile risposta

Avremo quindi un equazione lineare del tipo:

$$\mathbf{y_{ij}= \beta_{0j} +x_{i1}\beta_{1j}+x_{i2}\beta_{2j}+...+x_{ik}\beta_{kj}+\varepsilon_{ij}}$$

con: x_{ik} dove i indical'i-esimo soggetto
e k indica la k-esima covariata

E' come se replicassi un modello di regressione multipla r volte, con la differenza, come vedremo, che si fanno assunzioni diverse in merito ai termini di errore.

ESEMPIO
y_1=spesa alimentare
y_2= spesa per vacanze
in questo caso allora r=2, voglio spiegare le y rispetto a:
x_1= reddito famiglia
x_2= numero di figli
x_3= capofamiglia laureato
in questo caso k=3, avrò quindi 2*4=8 parametri, ovvero in genereale si ha:

parametri (coefficenti)= r*(k+1)

nel nostro esempio specifico avrò : $\beta_{01},\beta_{11},\beta_{21},\beta_{31}$ coefficenti relativi a y_{i1} e β_{02}, β_{12}, β_{22}, β_{32} coefficenti relativi a y_{i2}.

I coefficenti si interpretano nel modo usuale, tenendo a mente che il 2° indice ci dice a quale variabile risposta si riferiscono. Quindi β_{32} sarà l'effetto del capofamiglia laureato sulla spesa in vacanze della famiglia e β_{02} l'intercetta per y_2, ovvero quando tutte le covariate sono uguali a 0.

Ora, **l'assunzione di base del modello** è che per ogni i, **E[ε_i]=0**.
dove ε_i è un vettore colonna di dimensioni r*1 e contiene tutti i termini di errore per l'i-esimo soggetto, ovvero:

$$\underline{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ . \\ . \\ . \\ \varepsilon_{ir} \end{pmatrix}$$

con i=1,2,....,n
e la **Var[$\underline{\varepsilon}_i$]= Σ** matrice varianza covarianza, ammetto così che ci può essere una correlazione tra gli errori.

Ritornando al nostro ESEMPIO con r=2 avremo che:

$$\text{Var}[\underline{\varepsilon}_i] = \Sigma = \begin{pmatrix} \text{Var}[\varepsilon_{i1}] & \text{Cov}[\varepsilon_{i1,}\varepsilon_{i2}] \\ \text{Cov}[\varepsilon_{i1,}\varepsilon_{i2}] & \text{Var}[\varepsilon_{i2}] \end{pmatrix}$$

In Σ le covarianze non devono necessariamente essere uguali a 0, ammettiamo infatti che ci può essere correlazione, e nella pratica infatti spesso si trova una correlazione maggiore di 0. Questa è una importante differenza rispetto a analizzare separatamente le variabili risposta.
Nel nostro ESEMPIO trovare una covarianza positiva significa che se la famiglia i-esima spende di più per alimenti allora ci aspetteremo che spenda di più anche per le vacanze.

Possiamo esprimere la nostra equazione lineare in una forma più sintetica, in **notazione matriciale,** ovvero nella forma:

$$Y = X*B + E$$

con Y,X,B,E matrici
in particolare avremo che:

Y è una matrice di dimensioni n*r

$$Y = \begin{pmatrix} y_{11} & y_{12} & \dots & \dots & \dots & y_{1r} \\ y_{21} & y_{22} & \dots & \dots & \dots & y_{2r} \\ \dots & \dots & \dots & \dots & \dots & \\ \dots & \dots & \dots & \dots & \dots & \\ \dots & \dots & \dots & \dots & \dots & \\ y_{n1} & y_{n2} & \dots & \dots & \dots & y_{nr} \end{pmatrix}$$

contenente tutti i valori delle variabili risposta.

X è una matrice di dimensioni n*(k+1), è la matrice del disegno:

$$X=\begin{pmatrix} 1 & x_{11} & x_{12} & \dots & \dots & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & \dots & \dots & x_{2k} \\ 1 & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & . \, . & \dots & x_{nk} \end{pmatrix}$$

B è una matrice di dimensione (k+1)*r, in cui ad ogni colonna corrisponde una variabile risposta, matrice dei coefficenti:

$$B=\begin{pmatrix} \beta_{01} & \beta_{02} & \dots & \dots & \dots & \beta_{0r} \\ \beta_{11} & \beta_{12} & \dots & \dots & \dots & \beta_{1r} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \beta_{k1} & \beta_{k2} & \dots & . \, . & \dots & \beta_{kr} \end{pmatrix}$$

E è una matrice di dimensione n*r, per ogni riga a tutti gli errori dell i-esimo soggetto (deve essere coerente con la Y)

$$E=\begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \dots & \dots & \dots & \varepsilon_{1r} \\ \varepsilon_{21} & \varepsilon_{22} & \dots & \dots & \dots & \varepsilon_{2r} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \varepsilon_{n1} & \varepsilon_{n2} & \dots & . \, . & \dots & \varepsilon_{nr} \end{pmatrix}$$

si può scrivere anche come vettore colonna dei vettori $\underline{\varepsilon_i}$' trasposti.

Ora se vogliamo ricavarci l'equazione per il singolo soggetto i avremo:

$$y_{ij}= \underline{x}_i'*\underline{\beta}_j+\varepsilon_{ij}$$

con $\underline{x}_i$' vettore riga trasposto, $\underline{\beta}_j$ vettore colonna e ε_{ij} vettore riga.

Per **stimare il modello** utilizziamo il metodo OLS (minimi quadrati) avremo:

$$Y_{hat}=X*B$$

dove B sono coefficenti che io ipotizzo e Y_{hat} matrice di dimensione n*r.
Per ESEMPIO se prendiamo: y_{hat52} = 3000, indica la previsione per la quinta famiglia per le spese per le vacanze.

Andando a confrontare la mia previsione con il valore osservato posso vedere l'errore di previsione:

$$Y-Y_{hat}= Y-X*B$$

con Y-X*B è la matrice degli errori di previsione, questa matrice nel nostro ESEMPIO mi dice per ogni famiglia e per ogni tipo di spesa l'errore di previsione.

Per avere una **misura complessiva degli error**i definisco S, come:

$$S= ||Y-X*B||^2$$

utilizzo l'operatore *norma* (somma degli elementi al quadrato sotto radice, se considero la matrice generica A allora $||A||=(a_{11}^2+a_{12}^2+....)^{1/2}$).
Ora facendo la norma al quadrato elimino la radice per cui in pratica è come se facessi una sommatoria:

$$S= \Sigma_i\Sigma_j(y_{ij}-y_{hat\ ij})^2$$

avrò quindi la sommatoria rispetto alle famiglie (Σ_i) e rispetto alle variabili risposta(Σ_j).
Il passo successivo è minimizzare rispetto a B la quantità S che è funzione di B, quindi:

$$\min_B S(B) \text{ la cui soluzione è}$$
$$B_{hat}=(X'X)^{-1}X'Y$$

con X matrice di dimensione n*(k+1)
X' matrice di dimensione (k+1)*n
Y matrice di dimensione n*r
e quindi B_{hat} matrice di dimensione (k+1)*r.

Ora per capire se cè **correlazione tra gli errori** bisogna stimare la matrice di varianza e covarianza degli errori Σ. Andremo quindi a calcolare la matrice dei residui:

$$E_{hat}= Y-Y_{hat}$$

NB: ora la matrice Y_{hat} non è più quella con i coefficenti ipotizzati bensì quella con i coefficenti B_{hat} ottimali, ricavati col metodo dei minimi quadrati ($Y_{hat}=X*B_{hat}$).

Quindi avremo la Σ stimata di dimensione r*r:

$$\Sigma_{hat}=(E_{hat}' * E_{hat})/(n-(k+1)= \begin{pmatrix} Var_{hat}[\varepsilon_{i1}] & Cov_{hat}[\varepsilon_{i1,}\varepsilon_{i2}] \\ Cov_{hat}[\varepsilon_{i1,}\varepsilon_{i2}] & Var_{hat}[\varepsilon_{i2}] \end{pmatrix}$$

Per fare **inferenza** (verifica delle ipotesi e intervalli di confidenza) devo assumere una distribuzione per gli errori, tale distribuzione sarà un normale, in particolare avremo che:

$$\text{il vettore } \underline{\varepsilon}_i \sim N_r(0,\Sigma).$$