TIPS E RIFLESSIONI SULLA LA REGRESSIONE MULTILINEARE

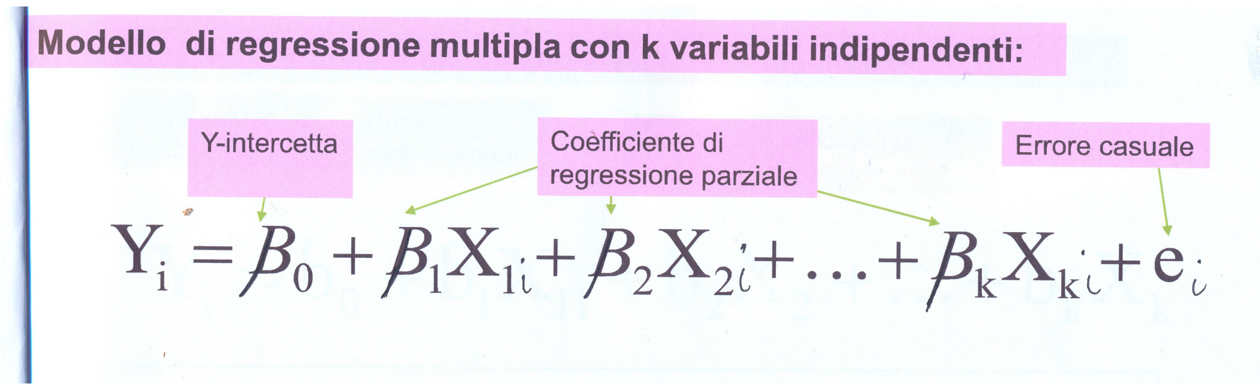
Introduzione

Tutti i fenomeni collettivi complessi (economici, demografici, meteorologici, naturali, quelli legati all'inquinamento …) sono strettamente collegati, all'interno e all'esterno, ad una complessa rete di relazioni, per cui studiare un fenomeno con una relazione a due sole variabili è spesso senza senso, anche se, e ne siamo certi, l'usare più variabili, non risolverà fino in fondo il problema, pur aumentando la *verisimiglianza popperiana*.

Una volta ammesso che il fenomeno ha relazioni interne fra variabili anche complesse, scelta la variabile dipendente y (variabile *risposta),* es., prezzi, consumi, crescita..., ci poniamo il problema di come essa vari in media al variare di altri caratteri che scegliamo (*variabili indipendenti esplicative o regressori)*.

Questo studio, che è una rappresentazione matematica della realtà o *modello della realtà*, viene denominato *regressione lineare multipla*, se pensiamo che ciascun valore osservato della variabile dipendente o risposta sia esprimibile come funzione lineare dei corrispondenti valori delle *variabili indipendenti esplicative,* più un termine residuo **ε** per indicare che il modello *non può* riprodurre la *realtà*. La *linearità* quindi è relativa non alle variabili esplicative (lineari o non lineari), ma ai *coefficienti* della relazione. **Naturalmente tutti i modelli sono falsi, ma spesso utili.**

Si voglia trovare così la relazione fra k variabili indipendenti ($x_1, x_2, \dots x_k$) e la variabile dipendente y. Il modello è un'equazione del tipo:



Di Y e delle variabili $X_1, X_2, \dots X_k$ ho n valori sperimentali (serie di dati) per ciascuna variabile. In termini matriciali :

1 - **Y** (i suoi elementi sono indicati con il generico y_i) è un vettore di osservazioni (n righe x 1 colonna) sulla variabile dipendente o risposta.

2 - **X** è una matrice (n x k) di osservazioni sui K regressori o variabili indipendenti o esplicative. La matrice contiene anche una colonna supplementare (la prima) composta da n valori tutti uguali a uno in corrispondenza dell'intercetta del modello. Il modello geometricamente coriisponde ad un iperpia a k dimensioni.

3 – **ß** è un vettore con (k x 1) *parametri incogniti*; il primo è $ß_0 = 1$

4 - **ε** (**e** nell'espressione del modello) è un vettore (n x 1) di disturbi stocastici o termini dell'errore.

Le matrici e i vettori sono così definiti

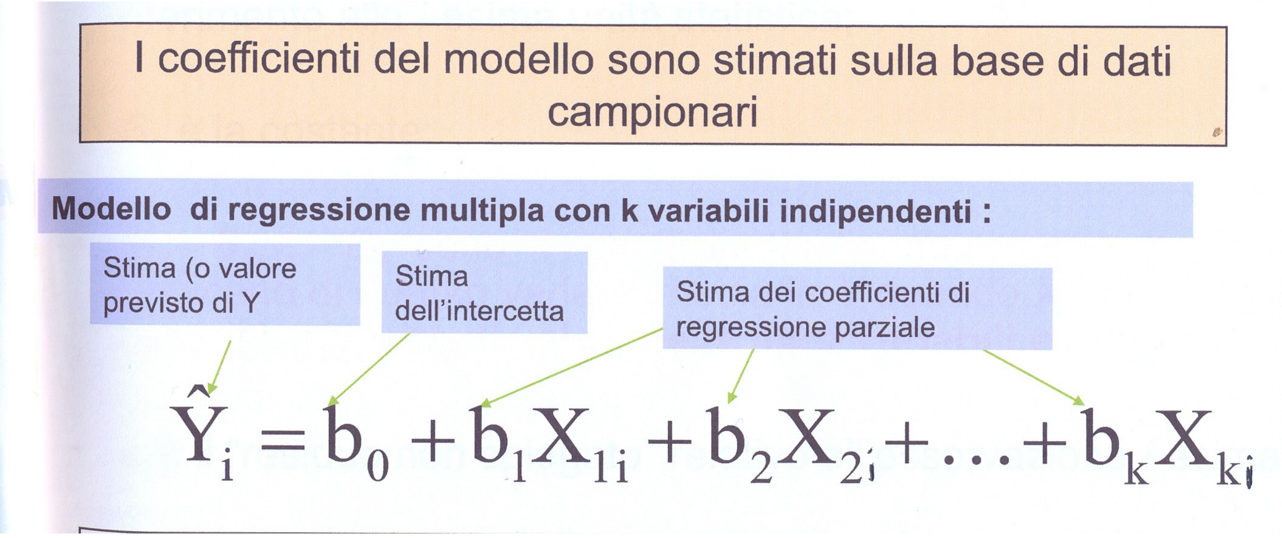
$$\underset{(n\times 1)}{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ y_n \end{pmatrix} \qquad \underset{(n\times(k+1))}{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & x_{2k} \\ & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{nk} \end{pmatrix}$$

$$\underset{(k\times 1)+1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{pmatrix} \qquad \underset{(n\times 1)}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

N.B.

La matrice X ha la prima colonna unitaria nel caso in cui si consideri un modello con intercetta β nel sistema di riferimento multidimensionale

Il fatto di non conoscere gli n errori (e o **ε)** impedisce di fatto la conoscenza del valore reale dei coefficienti **ß**; ci accontenteremo di ricavare solo le loro stime **b**. In tal caso lavoreremo su la seguente espressione semplificata del modello:



Si ha così un sistema di n equazioni esprimibile nel linguaggio matriciale con l'espressione condensata:

Y =X b

La scelta e il tipo delle variabili esplicative (x, x^2, 1/x ecc.) viene suggerita e dalle teorie relative al fenomeno in gioco e dal buonsenso. Si può anche usare una serie di scatterplot (matrice di scatterplot) fra ciascuna coppia di variabili osservando le loro relazioni (dirette o inverse, lineari o non lineari, l'eventuale presenza di outliers e la forza della loro relazione.

Nel prossimo tip faremo degli esempi applicativi.