ASPETTI COMPUTAZIONALI NEI MODELLI STATISTICI LINEARI

L'analisi di regressione multipla

Si tratta di esaminare i problemi computazionali fondamentali per quanto riguarda l'analisi di regressione multipla.

Il modello è dato da:

$$\underset{(n \times 1)}{\underline{y}} = \underset{(n \times p)}{\mathbf{X}}\overset{(p \times 1)}{\underline{\beta}} + \underset{(n \times 1)}{\underline{\varepsilon}}$$

Lo stimatore per $\underline{\beta}$ è la soluzione delle equazioni normalizzate:

$$(\mathbf{X}'\mathbf{X})\hat{\underline{\beta}} = \mathbf{X}'\underline{\mathbf{y}}$$

vale a dire:

$$\hat{\underline{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{\mathbf{y}}$$

Problema:

Analizzare gli aspetti computazionali che riguardano:

- Il calcolo dello stimatore $\hat{\underline{\beta}}$
- la bontà di adattamento del modello di regressione
- una misura della variabilità di $\hat{\underline{\beta}}$

<u>Il problema dei minimi quadrati e le trasformazioni ortogonali</u>

IPOTESI:

a) $E(\underline{\varepsilon}) = \underline{0}$

b) $E(\underline{\varepsilon}\,\underline{\varepsilon}') = \sigma^2 \underline{I}_n \Rightarrow E(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$

c) X è un insieme di numeri certi

d) X ha rango $p < n$

n° dei parametri da stimare

n° osservazioni

Si trasforma, tramite le trasformazioni ortogonali, il problema originario in un problema equivalente la cui soluzione è identica all'originale ma comporta vantaggi numerici e computazionali

IL MODELLO DI REGRESSIONE MULTIPLA

Nel modello di regressione multipla la variabile dipendente Y è influenzata da più variabili indipendenti.

NB:

Se per il modello di regressione semplice ovvero per il modello di regressione con due variabili esplicative è ancora possibile calcolare le stime dei parametri senza ricorrere all'uso di un elaboratore nel caso di più di due variabili esplicative l'uso di tale strumento diventa indispensabile.

Obiettivo della regressione multipla

è quello di spiegare la variabile dipendente y in funzione delle n variabili indipendenti: $x_1, \ldots x_n$

ovvero di descrivere la dipendenza di y dalle n variabili x_i mediante una fu[nzione] lineare delle n variabili x_i

[dal punto di vista geometrico ciò corrisponde ad un iper-piano in un iperspazio ad n+1 dimensioni (analogamente al fatto che l'equazione di regressione con una variabile indipendente corrisponde ad una retta nel piano)]

In generale i punti individuati non giaceranno che in minima parte sull'iperpiano ma si discosteranno per una serie di elementi quali

errori di misurazione nelle variabili; (
variabili esplicative non incluse nel
modello, fattore di non-linearità ecc.

$\Rightarrow$

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

$i = 1, \ldots, n$

in forma compatta

$$\underline{Y} = X \underline{\beta} + \underline{\varepsilon}$$

$$\underline{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & X_{21} & \cdots & X_{p1} \\ 1 & X_{22} & \cdots & X_{p2} \\ 1 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & X_{2n} & \cdots & X_{pn} \end{bmatrix}$$

errori
(n (+1) di variabili esplicative

$$\underline{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \vdots \\ \beta_p \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

IPOTESI:

a)* $E(\underline{\varepsilon}) = \underline{0}$

b)* $E(\underline{\varepsilon}\,\underline{\varepsilon}') = \sigma^2 I_n \;{=}{\triangleright}\; E(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$

c)* X è un insieme di numeri certi (l'unica fonte di variabilità è data dal vettore $\underline{\varepsilon}$)

d)* X ha rango $p < n$ $\rightarrow$ n° di osservazi[oni]

p $\rightarrow$ n° di parametri da stimare

STIMA DEI PARAMETRI DEL MODELLO

Si applichi il metodo dei minimi quadrati per stimare i parametri.

Sia $\hat{\underline{\beta}} = \{\hat{\beta}_1, \ldots, \hat{\beta}_p\}'$

vettore colonna composto dalle stime d[ei]

Sotto le ipotesi prima scelte
Lo stimatore dei minimi quadrati,
è corretto : $E(\hat{\underline{\beta}}) = \underline{\beta}$

e a varianza minima $Var(\hat{\underline{\beta}}) = \sigma^2 (X'X)$

$\Rightarrow$ TH. DI GAUSS MARKOV

Gli stimatori dei minimi quadrati
SONO BLUE

Si consideri la formula standard:

$$\hat{\underline{\beta}} = (X'X)^{-1} X' \underline{Y}$$

APPROCCIO PER I CALCOLI:

.) CALCOLARE $(X'X)$, $(X'\underline{Y})$

.) INVERTIRE $(X'X) \longrightarrow (X'X)^{-1}$

.) FORMARE IL PRODOTTO:
$(X'X)^{-1}(X'\underline{Y})$

APPROCCIO VALIDO SE n e p SONO PICCOLI.

APPROCCIO FACILE DA IMPLEMENTA[RE]

(eccezione può derivare dall'inversione della matrice) CHE HA PERÒ ALCUNI LATI NEGATIVI.

OSS

•) $(X'X)$

sono matrici derivanti da prodotti interni di colonne e i prodotti interni sono fonte di errori.

$\Rightarrow$ meglio un approccio che minimizz[a] i prodotti interni.

•) Un altra fonte di errore e di inefficienza computazionale deriva dal fatto che vi è un passo in cui si deve effettuare l'inversione di una matrice.

$\Rightarrow$ INSTABILITÀ