

La regressione

(S. Terzi)

1. Retta di regressione (regressione lineare semplice)

Date n coppie di osservazioni (x_i, y_i) , $i=1, \dots, n$ si desidera fare un'interpolazione dei punti attraverso una retta:

$$y^* = a + bx$$

Naturalmente i punti non si troveranno già tutti sulla retta. Ci saranno degli scostamenti $(y_i - y_i^*)$ che chiamiamo residui (e_i).

In sostanza si assume che il modello vero che lega y ad x sia una retta ($y = a + bx$), ma che le osservazioni y siano affette da errore, per cui i punti non si trovano TUTTI sulla retta. Sarà in realtà: $y_i = a + bx_i + e_i$ (dove e_i è l'errore), e ci aspettiamo che in media gli errori si compensino (ovvero che abbiano media nulla). Inoltre, in un contesto inferenziale, si suppone anche che gli errori abbiano una varianza costante, e che quindi la loro variabilità non sia legata ai valori assunti dalla y o dalla x . Se così non fosse c'è il rischio che il modello (la retta) sia specificato male, per esempio che i punti non si trovino (quasi) su una retta ma su una parabola.

Il metodo dei minimi quadrati stima i parametri a e b minimizzando la somma dei quadrati dei residui:

$$\sum_i e_i^2 = \sum_i (y_i - a - bx_i)^2 = F$$

Derivando la funzione obiettivo F rispetto ai parametri incogniti a e b otteniamo:

$$(\partial F / \partial a) = 2 \sum_i (y_i - a - bx_i)(-1) = -2 \sum_i e_i = \mathbf{0}$$

$$(\partial F / \partial b) = 2 \sum_i (y_i - a - bx_i)(-x_i) = -2 \sum_i e_i(x_i) = \mathbf{0}$$

da cui, con vari passaggi si ottiene infine:

$$a^* = M_1(y) - bM_1(x)$$

$$b^* = (\text{cov}(x,y)) / \text{var}(x)$$

dove a^* e b^* sono i valori dei coefficienti a e b che minimizzano la funzione F.

Per misurare la **bontà di adattamento della retta**, si usa l'indice R^2 , che è al tempo stesso il quadrato del coefficiente di correlazione tra x ed y, ed il rapporto tra devianza spiegata dalla retta e devianza totale.

Al fine di definire un indice di bontà di adattamento, si parte dalla scomposizione della devianza della y .

$$\text{Dev}(y) = \text{dev}(y^*) + \text{dev}(e)$$

Condizione essenziale per ottenere tale scomposizione è che la somma dei residui si annulli. E cio' accade SOLO SE nella specificazione della retta abbiamo inserito anche l'intercetta.

Infatti la devianza della y è data da:

$$\sum_i (y_i - M_1(y))^2 \quad (1)$$

Se all'interno della parentesi tonda aggiungiamo e sottraiamo le y_i^* , si ha:

$$\begin{aligned} \text{Dev}(y) &= \sum_i (y_i - y_i^* + y_i^* - M_1(y))^2 = \\ &= \sum_i (y_i - y_i^*)^2 + \sum_i (y_i^* - M_1(y))^2 + \\ &\quad + 2\sum_i (y_i - y_i^*)(y_i^* - M_1(y)) \end{aligned}$$

Ricordando che abbiamo definito $e_i = (y_i - y_i^*)$, possiamo riscrivere la devianza totale come:

$$\text{dev}(y) = \sum_i e_i^2 + \sum_i (y_i^* - M_1(y))^2 + 2\sum_i e_i (y_i^* - M_1(y))$$

Essendo $\sum_i e_i = 0$ il primo addendo è proprio la devianza dei residui. Inoltre, sempre in virtù del fatto che $\sum_i e_i = 0$,

$\sum_i y_i = \sum_i y_i^*$, $M_1(y) = M_1(y^*)$ e quindi il secondo addendo e' proprio la devianza delle y^* .

Infine (ed e' questa la propriet  fondamentale), essendo $\sum_i e_i = 0$, il doppio prodotto si annulla.

Possiamo infatti scriverlo come:

$$\begin{aligned} 2\sum_i e_i (y_i^* - M_1(y)) &= 2\sum_i e_i (a^* + b^* x_i - M_1(y)) = \\ &= 2\sum_i e_i a^* + 2\sum_i e_i (b^* x_i) - 2\sum_i e_i M_1(y) = \\ &= 2a^* \sum_i e_i + 2\sum_i e_i (b^* x_i) - 2 M_1(y) \sum_i e_i = \\ &= 0 + 2b^* \sum_i e_i (x_i) - 0 \end{aligned}$$

Tenendo conto della derivata della funzione F rispetto a b, ovvero del fatto che $- 2\sum_i e_i(x_i) = 0$, si vede facilmente che tutto il doppio prodotto e' nullo.

Tuttavia, poiche' l'intera scomposizione della devianza si basa sul fatto che $\sum_i e_i = 0$, se nell'espressione della retta omettiamo l'intercetta i residui non dovranno soddisfare tale condizione; quindi **la scomposizione della devianza totale in devianza dei residui + devianza della retta e' valida solo in presenza di un'intercetta.**

2. Estensione

Supponiamo di voler interpolare non più con la retta ma con una parabola:

$$y^{**} = c_0 + c_1x + c_2x^2$$

Chiamiamo f i residui ($f_i = y_i - y_i^{**}$).

Possiamo ottenere le stime dei coefficienti minimizzando la somma dei quadrati dei residui.

Anche in questo caso varrà una scomposizione della devianza della y . Sarà:

$$\text{Dev}(y) = \text{dev}(y^{**}) + \text{dev}(f)$$

In virtù del metodo di stima prescelto (la minimizzazione della devianza dei residui), $\text{dev}(f)$ sarà non superiore a $\text{dev}(e)$, e pertanto $\text{dev}(y^{**})$ sarà non superiore a $\text{dev}(y^*)$. (Questo perché essendo la funzione da minimizzare data da $\sum_i f_i^2$, qualora l'inserimento del termine x^2 non aiuti a migliorare l'interpolazione, troveremo $c_2 = 0$; $c_0 = a$ e $c_1 = b$; la parabola coinciderà con la retta e $\sum_i f_i^2 = \sum_i e_i^2$. In tutti gli altri casi, se la parabola rappresenta un miglioramento rispetto alla retta, sarà: $\text{dev}(f) < \text{dev}(e)$, e $\text{dev}(y^{**}) > \text{dev}(y^*)$).

In sostanza aumentando il numero di parametri della funzione interpolante, la devianza residua o diminuisce o resta costante.

3. La regressione multipla

Si abbiano k variabili esplicative: x_1, x_2, \dots, x_k (linearmente indipendenti) e una variabile dipendente y . Per ciascuna variabile abbiamo n osservazioni. Quindi per la i -esima unità abbiamo un vettore di osservazioni:

$(y_i, x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ik})$, e viceversa, ciascuna variabile è rappresentata da un vettore di n componenti, appartenenti quindi allo spazio \mathbb{R}^n .

Desideriamo proiettare il vettore y nel sottospazio (di dimensione $k+1$) generato dalle variabili esplicative (+ l'intercetta):

$$y^* = b_0 + b_1x_1 + \dots + b_jx_j + \dots + b_kx_k \quad (2)$$

Definendo i residui $e_i = (y_i - y_i^*)$, il metodo dei minimi quadrati fornirà come stime dei coefficienti b_j ($j=0, \dots, k$) quei valori b_j^* che rendono minima $\sum_i e_i^2$. Essendo $\sum_i e_i^2$ una quantità non negativa, il suo unico punto stazionario sarà un minimo. Non è quindi necessario controllare le condizioni del secondo ordine.

Nel modello di regressione multipla, la variabile dipendente y è rappresentata da un vettore di n elementi appartenente quindi a \mathbb{R}^n . Cerchiamo poi il sottospazio generato dai soli vettori x_1, x_2, \dots, x_k (con l'aggiunta di un ulteriore vettore di tutti 1 per tenere conto dell'intercetta). Se tali vettori sono linearmente indipendenti essi saranno una base del sottospazio, che avrà quindi dimensione $k+1$. In linea di massima, y non apparterrà a questo sottospazio (a meno che non sia già esattamente una combinazione lineare delle variabili x_1, x_2, \dots, x_k); vi apparterrà invece y^* . Possiamo pensare che y^* sia la proiezione di y nel sottospazio delle variabili esplicative.

Per procedere è opportuno utilizzare **una notazione matriciale**.

Sia \mathbf{X} la matrice dei dati. Le sue colonne contengono i vettori x_1, x_2, \dots, x_k . Poiché il nostro scopo è ridefinire la (2) come prodotto tra una matrice \mathbf{X} ed un vettore di coefficienti \mathbf{b} , e poiché la (2) contiene un'intercetta, inseriamo nella matrice \mathbf{X} una prima colonna di 1.

Siano:

$\mathbf{X} = [\mathbf{1}, x_1, x_2, \dots, x_k]$, di dimensione $n \times (k+1)$.

$\mathbf{b} = (b_0, b_1, \dots, b_k)'$, di dimensione $(k+1) \times 1$.

$\mathbf{y} = (y_1, \dots, y_i, \dots, y_n)'$ di dimensione $n \times 1$;

$\mathbf{y}^* = (y^*_1, \dots, y^*_i, \dots, y^*_n)'$ di dimensione $n \times 1$;

$\mathbf{e} = (e_1, \dots, e_i, \dots, e_n)'$ di dimensione $n \times 1$.

La (2) può essere riscritta in notazione matriciale come:

$$\mathbf{y}^* = \mathbf{Xb}$$

e la funzione obiettivo da minimizzare sarà:

$$F = \sum_i e_i^2 = \mathbf{e}'\mathbf{e}$$

Essendo anche $\mathbf{e} = \mathbf{y} - \mathbf{y}^* = \mathbf{y} - \mathbf{X}\mathbf{b}$, possiamo scrivere la funzione obiettivo anche come:

$$F = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

Infine, tenendo conto che $\mathbf{y}'\mathbf{X}\mathbf{b}$ e' il trasposto di $\mathbf{b}'\mathbf{X}'\mathbf{y}$ e che si tratta di scalari (cioe' tenendo conto che si tratta di due addendi uguali), abbiamo:

$$F = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

Derivando la funzione F rispetto al vettore dei coefficienti b ed uguagliando a zero si ottiene:

$$\partial F / \partial \mathbf{b} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}$$

da cui:

$\mathbf{X}'\mathbf{X}\mathbf{b}^* = \mathbf{X}'\mathbf{y}$ e quindi:

$$\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

purche' la matrice $(\mathbf{X}'\mathbf{X})$ sia invertibile, cioe' di rango pieno $k+1$.

4. Interpretazione

Siamo partiti dal definire $\mathbf{y}^* = \mathbf{X}\mathbf{b}^*$, se ora sostituiamo $\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$, otteniamo:

$$\mathbf{y}^* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Analogamente, essendo $\mathbf{e} = \mathbf{y} - \mathbf{y}^*$, sarà anche :

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'] \mathbf{y} = \\ &= [\mathbf{I} - \mathbf{H}]\mathbf{y} = \mathbf{M}\mathbf{y} \end{aligned}$$

Sia la matrice \mathbf{H} che la matrice \mathbf{M} sono matrici simmetriche ($\mathbf{H} = \mathbf{H}'$; $\mathbf{M} = \mathbf{M}'$) ed idempotenti ($\mathbf{H}\mathbf{H} = \mathbf{H}$; $\mathbf{M}\mathbf{M} = \mathbf{M}$).

Le matrici simmetriche ed idempotenti si chiamano **proiettori**. Inoltre hanno la proprietà di avere il rango uguale alla loro traccia.

In pratica la matrice \mathbf{H} trasforma \mathbf{y} in \mathbf{y}^* , cioè lo proietta nel sottospazio generato dalle variabili esplicative.

Viceversa, la matrice \mathbf{M} trasforma \mathbf{y} in \mathbf{e} . Anche in questo caso si tratta di una proiezione.

Occorre aprire nuovamente una parentesi riguardante gli spazi vettoriali....

Proiezione ortogonale

Sia S è uno spazio lineare di dimensione n , sia S_1 un suo sottospazio.

Ogni vettore \mathbf{x} appartenente ad S si può decomporre nella somma di un vettore \mathbf{x}_1 appartenente a S_1 e di un vettore \mathbf{x}_2 ortogonale ad \mathbf{x}_1 . (Due vettori si dicono ortogonali quando il loro prodotto interno è nullo). Il vettore \mathbf{x}_1 si chiama *proiezione ortogonale* di \mathbf{x} in S_1 .

Vale inoltre la relazione pitagorica:

$$||\mathbf{x}||^2 = ||\mathbf{x}_1||^2 + ||\mathbf{x}_2||^2$$

L'insieme di tutti i possibili vettori \mathbf{x}_2 ortogonali ai vettori che costituiscono la base di S_1 costituiscono un sottoinsieme di S che si chiama complemento ortogonale di S_1 . La dimensione del complemento ortogonale di S_1 è data dalla differenza tra la dimensione di S e la dimensione di S_1 .

Tornando alle matrici H ed M, $\mathbf{y} = \mathbf{y}^* + \mathbf{e}$;

- \mathbf{y}^* appartiene ad un sottospazio di S. Il sottospazio (S_{k+1}) generato dai vettori $\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$.
- \mathbf{y}^* ed \mathbf{e} sono ortogonali: $\mathbf{y}^{*\prime} \mathbf{e} = \mathbf{y}' \mathbf{H}' \mathbf{M} \mathbf{y} = \mathbf{y}' \mathbf{H} (\mathbf{I} - \mathbf{H}) \mathbf{y} =$
 $= \mathbf{y}' \mathbf{H} \mathbf{y} - \mathbf{y}' \mathbf{H} \mathbf{H} \mathbf{y} = \mathbf{0}$
- \mathbf{y}^* e' la proiezione ortogonale di \mathbf{y} nel sottospazio S_{k+1} .
- \mathbf{e} appartiene al complemento ortogonale di S_{k+1} .

Viceversa, anche \mathbf{e} puo' vedersi come proiezione ortogonale di \mathbf{y} .

La dimensione del sottospazio S_{k+1} coincide con il rango della matrice H. La dimensione del complemento ortogonale coincide con il rango della matrice M. Essendo $\text{rango}(\mathbf{H}) = \text{traccia}(\mathbf{H}) = k+1$, S_{k+1} ha effettivamente dimensione $k+1$. Mentre il suo complemento ortogonale ha dimensione $= \text{tr}(\mathbf{I} - \mathbf{H}) = n - k - 1$, ma anche $\text{dim}(\mathbf{S}) - \text{dim}(\mathbf{S}_{k+1})$.

Anche in questo caso, per valutare la **bontà di adattamento** si può ricorrere alla scomposizione della devianza e all' R^2 .

Infatti, essendo:

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'^*\mathbf{y}^* + \mathbf{e}'\mathbf{e}$$

dividendo tutto per n e immaginando di sottrarre ($M_1^2(y)$) ad entrambi i membri dell'uguaglianza e' anche:

$$\text{dev}(y) = \text{dev}(y^*) + \text{dev}(e)$$

pertanto possiamo definire:

$$R^2 = \text{dev}(y^*)/\text{dev}(y) = 1 - (\text{dev}(e)/\text{dev}(y))$$

$$0 \leq R^2 \leq 1.$$

5 Validazione del modello : quali e quante variabili esplicative ?

Come abbiamo notato in precedenza, nel confronto tra retta e parabola di regressione, anche nel caso del piano di regressione, se aggiungiamo un'ulteriore variabile esplicativa, definendo p.es

$$y^{**} = c_0 + c_1X_1 + \dots + c_jX_j + \dots + c_kX_k + c_{k+1}X_{k+1}$$

e

$$f_i = (y_i - y_i^{**})$$

sarà sempre:

$$\sum_i f_i^2 \leq \sum_i e_i^2, \text{ e quindi:}$$

$$\text{dev}(y^{**}) \geq \text{dev}(y^*)$$

Pertanto all'aumentare del numero di variabili esplicative, tende ad aumentare l' R^2 .

Per "bilanciare" in parte questo aumento e preferire rappresentazioni più parsimoniose a meno che l'incremento di devianza non sia effettivamente molto significativo, si usa spesso una versione "corretta" dell' R^2 .

$$R^2_c = 1 - [\text{dev}(e)/(n-k-1)] \times [(n-1)/\text{dev}(y)]$$

E quindi anche:

$$R^2_c = R^2 - k(1-R^2)/(n-k-1)$$

In sostanza se k è elevato rispetto ad n (e quindi se $k/(n-k-1)$ è elevato) R^2 risulterà molto diminuito per effetto della correzione; viceversa, se rispetto ad n , k è "trascurabile" la correzione risulterà molto piccola ($R^2_c \approx R^2$).

Che succede se la matrice X non ha rango pieno ?

Dal punto di vista geometrico, se le colonne della matrice X non sono linearmente indipendenti, non costituiscono una base del sottospazio su cui vogliamo proiettare y . Tale sottospazio potrà al massimo avere dimensione k , o anche inferiore a k , e lo individueremo eliminando da X le colonne che contengono vettori linearmente dipendenti da altri.

In concreto, se X non ha rango pieno, significa che non riusciamo ad ottenere le stime del vettore dei coefficienti. Bisognerà ridurre il numero di colonne di X finché tale matrice non avrà rango pieno di colonna.

Tale problema è noto come "multicollinearità". Perfetta.

Alcune variabili sono combinazione lineare di altre e vanno eliminate.

Esiste un altro caso che pure merita cautela ed e' il caso in cui X , pur avendo rango pieno, ha un determinante molto prossimo a zero. Si parla in tal caso di collinearità non perfetta tra le variabili. Le conseguenze sono che le stime dei coefficienti (ottenute invertendo la matrice $X'X$), risultano tutte divise per il determinante della matrice $X'X$. E quindi assumeranno valori non solo molto elevati, ma molto sensibili agli arrotondamenti.... Avremo in tal caso delle stime molto instabili, che assumono valori molto diversi se variamo il numero di cifre decimali a cui arrotondiamo.