La regressione

(S. Terzi)

1. Retta di regressione (regressione lineare semplice)

Date n coppie di osservazioni (x_i,y_i), i=1,…,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 e' 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 cosi' non fosse c'e' 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:

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

Derivando la funzione obiettivo F rispetto ai parametri incogniti a e b otteniamo:
$(\partial F/\partial a) = 2\Sigma_i (y_i - a - bx_i)(-1) = \mathbf{-2\Sigma_i e_i = 0}$
$(\partial F/\partial b) = 2\Sigma_i (y_i - a - bx_i)(-x_i) = \mathbf{-2\Sigma_i e_i (x_i) = 0}$

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 e' 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.

$$Dev(y) = dev(y^*) + dev(e)$$

Condizione essenziale per ottenere tale scomposizione e' 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:

$$\Sigma_i(y_i - M_1(y))^2 \qquad (1)$$

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

$$Dev(y) = \Sigma_i(y_i - y_i^* + y_i^* - M_1(y))^2 =$$

$$= \Sigma_i(y_i - y_i^*)^2 + \Sigma_i(y_i^* - M_1(y))^2 +$$

$$+2\Sigma_i(y_i - y_i^*)(y_i^* - M_1(y))$$

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

$$dev(y) = \Sigma_i e_i^2 + \Sigma_i(y_i^* - M_1(y))^2 + 2\Sigma_i e_i (y_i^* - M_1(y))$$

Essendo $\mathbf{\Sigma_i e_i = 0}$ il primo addendo e' proprio la devianza dei residui. Inoltre, sempre in virtu' del fatto che $\mathbf{\Sigma_i e_i = 0,}$

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

Infine (ed e' questa la proprietà fondamentale), essendo $\mathbf{\Sigma_i e_i = 0}$, il doppio prodotto si annulla.
Possiamo infatti scriverlo come:

$2\Sigma_i e_i \, (y_i^* - M_1(y)) = 2\Sigma_i e_i \, (a^* + b^* x_i - M_1(y)) =$

$= 2\Sigma_i e_i \, a^* + 2\Sigma_i e_i \, (b^* x_i) - 2\Sigma_i e_i \, M_1(y) =$

$= 2a^* \Sigma_i e_i + 2\Sigma_i e_i \, (b^* x_i) - 2 \, M_1(y) \, \Sigma_i e_i =$

$= 0 + 2b^* \Sigma_i e_i \, (x_i) - 0$

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

Tuttavia, poiche' l'intera scomposizione della devianza si basa sul fatto che $\mathbf{\Sigma_i e_i = 0,}$ 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 piu' 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:
Dev(y) = dev(y**) + dev(f)

In virtu' del metodo di stima prescelto (la minimizzazione della devianza dei residui), dev(f) sarà non superiore a dev(e), e pertando dev(y*) sarà non superiore a dev(y**). (Questo perche' essendo la funzione da minimizzare data da $\Sigma_i f_i^2$, 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 $\Sigma_i f_i^2 = \Sigma_i e_i^2$. In tutti gli altri casi, se la parabola rappresenta un miglioramento rispetto alla retta, sarà: dev(f)< dev(e), e
dev(y**) > 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, …. 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},….,x_{ij},…,x_{ik}), e viceversa, ciascuna variabile e' rappresentata da un vettore di n componenti, appartenenti quindi allo uno spazio R^n.
Desideriamo proiettare il vettore y nel sottospazio (di dimesione k+1) generato dalle variabili esplicative (+ l'intercetta):

$$y^* = b_0 + b_1x_1 + \ldots + b_jx_j + \ldots + b_kx_k \qquad (2)$$

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

Parentesi
Cos'e' uno spazio, cos'e' un sottospazio, cos'e' la dimensione.

Riprendendo le dispense di Algebra, troviamo:

Definizione di spazio vettoriale

Sia S un insieme di vettori di ordine n. S è detto *spazio lineare* (o vettoriale) se e' un insieme chiuso rispetto alle operazioni somma e prodotto per uno scalare.

Esempio di spazio vettoriale: l'insieme di tutti i vettori ad n componenti (spazio che chiamiamo R^n).

1) **Sottospazio**

Un sottoinsieme S_o di uno spazio vettoriale S e' detto *sottospazio* di S se risulta chiuso rispetto alle operazioni di somma e prodotto per uno scalare.

Definizione di insieme generatore di uno spazio vettoriale.

Dati p vettori di ordine n, $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$, l'insieme di tutte le loro combinazioni lineari

$$c_1\mathbf{x_1} + c_2\mathbf{x_2} + \ldots\ldots + c_p\mathbf{x_p}$$

al variare dei coefficienti c_i è uno spazio vettoriale S. Si dice anche che lo spazio vettoriale S è generato dai vettori $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$, o anche che i vettori $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$, sono un insieme generatore di S.

1) Base di uno spazio lineare

Si supponga che ogni vettore **x** di ordine n, appartenente ad S, sia esprimibile in un solo modo (univocamente) come combinazione lineare di p vettori $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$, appartenenti ad S. Si dice allora che i p vettori $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$, costituiscono una *base* di S.

Si puo' dimostrare che un insieme generatore di S e' una base di S se e solo se è costituito da vettori linearmente indipendenti.

Si tenga però presente che uno stesso spazio può avere diverse basi.

2) Dimensione di uno spazio lineare

Si può dimostrare che tutte le basi di uno stesso spazio lineare hanno lo stesso numero di elementi. Si chiama *dimensione* dello spazio lineare, il massimo numero di vettori linearmente indipendenti appartenenti ad S.

Si puo' dimostrare che se S_0 è un sottospazio di S non coincidente con esso, allora

$$\dim(S_o) < \dim(S)$$

3) **Spazio $\Re^n$.**

Intendiamo la totalità dei vettori ad n componenti. E' uno spazio vettoriale di dimensione n. Cio' implica che n+1 vettori in $\Re^n$ sono sempre linearmente dipendenti.

Nel modello di regressione multipla, la variabile dipendente y e' rappresentata da un vettore di n elementi appartenente quindi a R^n . Cerchiamo poi il sottospazio generato dai soli vettori x_1 ,x_2, …. 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, …. x_k); vi apparterrà invece y*. Possiamo pensare che y* sia la proiezione di y nel sottospazio delle variabili esplicative.

Per procedere e' opportuno utilizzare **una notazione matriciale.**
Sia **X** la matrice dei dati. Le sue colonne contengono i vettori $\mathbf{x_1}$,$\mathbf{x_2}$, …. $\mathbf{x_k}$. Poiche' il nostro scopo e' ridefinire la (2) come prodotto tra una matrice **X** ed un vettore di coefficienti **b**, e poiche' la (2) contiene un'intercetta, inseriamo nella matrice X una prima colonna di 1.
Siano:

X = [**1**, $\mathbf{x_1}$,$\mathbf{x_2}$, …. $\mathbf{x_k}$], di dimensione nx$(k+1)$.

b = $(b_0 , b_1,\ldots b_k)$', di dimensione $(k+1)$x1.

y = $(y_1,\ldots,y_i,\ldots,y_n)$' di dimensione nx1;
y* = $(y^*_1,\ldots,y^*_i,\ldots,y^*_n)$' di dimensione nx1;
e = $(e_1,\ldots,e_i,\ldots,e_n)$' di dimensione nx1.

La (2) puo' essere riscritta in notazione matriciale come:

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

e la funzione obiettivo da minimizzare sarà:

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

Essendo anche **e** = **y – y*** = **y – Xb**, possiamo scrivere la funzione obiettivo anche come:

F = (**y-Xb**)’(**y-Xb**) = **y’y – y’Xb – b’X’y + b’X’Xb**

Infine, tenendo conto che **y’Xb** e’ il trasposto di **b’X’y** e che si tratta di scalari (cioe’ tenendo conto che si tratta di due addendi uguali), abbiamo:

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

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

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

da cui:

X’Xb* = **X’y** e quindi:

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

purche’ la matrice **(X’X)** sia invertibile, cioe’ di rango pieno k+1.

4. Interpretazione

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

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

Analogamente, essendo **e** = **y** – **y***, sarà anche :

$$\mathbf{e = y - X(X'X)^{-1} X'y = [I - X(X'X)^{-1} X'] y =}$$
$$\mathbf{= [I - H]y = My}$$

Sia la matrice H che la matrice M sono matrici simmetriche (H = H'; M = M') ed idempotenti (HH = H; MM = 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 H trasforma y in y*, cioe' lo proietta nel sottospazio generato dalle variabili esplicative.
Viceversa, la matrice M trasforma y in 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 **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 e' nullo).Il vettore $\mathbf{x}_1$ si chiama *proiezione ortogonale* di **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 e' data dalla differenza tra la dimensione di S e la dimensione di S_1.

Tornando alle matrici H ed M, **y = y* + e;**

- y* appartiene ad un sottospazio di S. Il sottospazio (S_{k+1}) generato dai vettori , **1**, $\mathbf{x_1}$, $\mathbf{x_2}$, …., $\mathbf{x_k}$.

- **y*** ed **e** sono ortogonali: y*'e = y'H'My = y'H(I-H)y =

  = y'Hy – y'HHy = 0

- **y*** e' la proiezione ortogonale di **y** nel sottospazio S_{k+1}.
- **e** appartiene al complemento ortogonale di S_{k+1}.

Viceversa, anche **e** puo' vedersi come proiezione ortogonale di 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 rango(H) = traccia(H) = k+1, S_{k+1} ha effettivamente dimensione k+1. Mentre il suo complemento ortogonale ha dimensione = tr(I-H) = n-k-1 ,ma anche dim(S) – dim(S_{k+1}).

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

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

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

$$dev(y) = dev(y^*) + dev(e)$$

pertanto possiamo definire:

$$R^2 = dev(y^*)/dev(y) = 1 - (dev(e)/dev(y))$$

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

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

$y^{**} = c_0 + c_1x_1 + \ldots + c_jx_j + \ldots + c_kx_k + c_{k+1}x_{k+1}$
e
$f_i = (y_i - y_i^{**})$
sarà sempre:
$\Sigma_i f_i^2 \leq \Sigma_i e_i^2$, e quindi:

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

Pertanto all'aumentare del numero di variabili esplicative, tende ad aumentare l'R^2.
Per "bilanciare" in parte questo aumento e preferire rappresentazioni piu' parsimoniose a meno che l'incremento di devianza non sia effettivamente molto significativo, si usa spesso una versione "corretta" dell'R^2.

$$R^2_c = 1 - [dev(e)/(n-k-1)]x[(n-1)/dev(y)]$$

E quindi anche:
$R^2_c = R^2 – k(1-R^2)/(n-k-1)$

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

Un possibile modo per individuare il miglior insieme di variabili esplicative partendo da una matrice che ne contiene k potrebbe essere quello di fare una regressione per tappe successive (step-wise). Se procediamo "per inclusione" ad ogni tappa successiva possiamo aggiungere la variabile che porta il maggior incremento all'R^2.
Potremmo partire dalla retta, utilizzando come unica variabile esplicativa quella x_j che abbia il maggior coefficiente di correlazione con la y. E via via aggiungere altre variabili fin tanto che la loro inclusione comporti un incremento dell' R^2_c .

Esiste un altro valore di riferimento utile per valutare la "*significatività*" di ciascuna variabile esplicativa. Una variabile x_j si ritiene "significativa" se siamo in grado di rifiutare l'ipotesi che il suo coefficiente b_j sia nullo. Il **p-value** ci misura la probabilità di commettere errore ritenendo non nullo quel coefficiente. In sostanza se il p-value associato al coefficiente b_j e' basso ($\leq 0{,}05$) a considerare "significativa" la variabile x_j abbiamo una bassa probabilità di sbagliare. E quindi la teniamo nel modello. Viceversa, se il p-value risulta alto la variabile e' non significativa e va tolta dal modello.

N.B. Il contesto in cui si fanno tali considerazioni e' un contesto *inferenziale*. Si assume che ci sia un modello "vero" con coefficienti b_j, ed un modello "stimato" con coefficienti b^*_j. La stima b^*_j potrà risultare $\neq 0$, pur essendo il parametro "vero" = 0. Il p-value misurerà la probabilità di commettere errore qualora si tenga nel modello la variabile ad esso associata, solo nel caso in cui siano soddisfatte alcune ipotesi di base. Tali ipotesi in sostanza prevedono che la y abbia una distribuzione Normale multivariata e che la componente di errore abbia media nulla e varianza costante $\forall i = 1,\ldots,n$. Al di fuori di qeusto contesto i p-values non hanno alcun fondamento teorico.

Tenendo conto di questo secondo criterio per la scelta delle variabili esplicative, possiamo inizialmente includere nel modello tutte e k le variabili. Calcoliamo i parametri ed i p-values. Eliminiamo la variabile con il p-value piu' alto

(purche' maggiore di 0,05) e ristimiamo tutto il modello. Se una variabile ha p-value elevato, la eliminiamo e stimiamo di nuovo l'intero modello….
Si continua cosi' finche' tutti i p-values risultano "bassi".
A quel punto abbiamo ottenuto il modello "significativo", e ne calcoliamo l'R^2.

Esistono in realtà vari metodi e vari indicatori che ci aiutano sia nella scelta del modello che nella valutazione della conformità del modello prescelto ad alcune ipotesi di partenza. Tutto cio' si studia nell'ambito dell'inferenza statistica e – ancor piu' – nell'ambito dell'econometria.
Si tratta di effettuare alcuni test d'ipotesi preliminari (p.es test di Normalità), di analizzare poi i residui del modello stimato (facendo anche in questo caso alcuni test di ipotesi).
Sulla base dell'analisi dei residui e' possibile concludere che alcune ipotesi di partenza (ipotesi "classiche") non sono soddisfatte, nel qual caso si procede ad una stima del modello con minimi quadrati generalizzati. Una volta validato il modello si puo' procedere alla scelta di quali variabili includervi sulla base dei p-values e di considerazioni che riguardano l'R^2.

6. 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 finche' tale matrice non avrà rango pieno di colonna.

Tale problema e' noto come "multicollinearità ". Perfetta. Alcune variabili sono combinazione lineare di altre e vanno eliminate. Oppure, l'insieme delle k variabili e' un insieme linearmente dipendente. In quest'ultimo caso, se non si riesce a capire quale variabile sia linearmente dipendente dalle altre, potremmo trasformarle tutte prendendo le prime componenti principali e definendo y come combinazione lineare delle prime CP.

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.

Qualora volessimo fare dei test d'ipotesi, questi risulterebbero poco attendibili. In particolare perdono totalmente di attendibilità i p-values.