Regressione lineare

Giovanni Organtini

6 novembre 2006

Supponiamo di aver misurato una grandezza fisica y (ad esempio, il pH di una soluzione) in funzione di un'altra grandezza fisica x (come, ad esempio, la concentrazione di ioni idrogeno $[H^+]$).

Poniamo anche il caso che esista un modello teorico che lega le due grandezze fisiche secondo una relazione funzionale del tipo

$$y = f(x)\,.$$

Nel caso dell'esempio ci si aspetta che

$$\mathrm{pH} \simeq -\log_{10}\left|\left[H^+\right]\right|$$

per cui, riportando su un grafico i valori della grandezza fisica y in funzione di x ci si aspetterebbe di vedere una curva. Di norma è piú semplice osservare relazioni di tipo lineare, perciò conviene procedere prima a *linearizzare* la relazione. Nel caso specifico, riportando sul grafico i valori di y in funzione dei valori di $x' = \log_{10} x$, la relazione aspettata assume la forma

$$y = Ax' + B\,,$$

con $A = -1$ e $B = 0$.

Attraverso i dati sperimentali si possono ricavare i valori misurati di A e B e, in questo caso, verificare le previsioni del modello. Per mostrare come questo sia possibile chiamiamo $\{x_i\}$, $i = 1\ldots, N$ un insieme di N dati sperimentali ottenuti misurando una grandezza fisica x e $\{y_i\}$, $i = 1\ldots, N$ l'insieme dei dati sperimentali relativi alla misura della grandezza fisica y in corrispondenza di ciascun valore di x_i. Facciamo l'ipotesi dunque che $y = Ax + B$, e cerchiamo di determinare i valori dei parametri A e B che meglio descrivono i dati sperimentali.

La retta che si avvicina di piú ai dati sperimentali sarà quella che per cui la somma delle distanze da ciascun punto sperimentale è minima.

Un modo di definire tale distanza è

$$\Delta = \sum_{i=1}^{N} |y_i - Ax_i - B|\,. \tag{1}$$

Questa definizione contiene però l'operatore di modulo che è scomodo da trattare, perciò è piú conveniente definire la distanza come

$$\Delta' = \sum_{i=1}^{N} (y_i - Ax_i - B)^2 \,. \tag{2}$$

Dunque la distanza Δ' si ottiene come somma delle distanze $d_i = (y_i - Ax_i - B)$ elevate al quadrato. Ciascuna di queste distanze d_i in realtà è nota con un'indeterminazione δ che si può valutare, usando la formula di propagazione degli errori, come

$$\delta_i^2 = \delta y_i^2 + A\delta x_i^2 \,,$$

dove δy_i e δx_i sono, rispettivamente, le indeterminazioni delle misure y_i e x_i. È chiaro che, nell'eseguire la somma delle d_i, i punti sperimentali affetti da grandi errori possono far tendere la retta ad assumere una pendenza e un'intercetta non corretti. Un risultato piú realistico si otterrebbe *pesando* ciascuna distanza elementare d_i con il suo errore, definendo una nuova *distanza pesata*

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - Ax_i - B)^2}{\delta_i^2} \,.$$

In questo modo a χ^2 contribuiscono maggiormente i punti con errore piccolo, che dunque contano di piú. Se, come spesso accade, il contributo dell'errore δx_i a δ_i è trascurabile, si giunge alla definizione piú comune di questa variabile:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - Ax_i - B)^2}{\delta y_i^2} \,. \tag{3}$$

Se troviamo i valori di A e B che rendono minimo il χ^2, abbiamo trovato la retta che si avvicina di piú ai dati sperimentali. Per fare questo in genere sono necessari metodi numerici che si possono applicare con l'ausilio di un computer.

Se assumiamo, per semplicità, che tutte le misure abbiano la stessa indeterminazione, cioè che $\delta y_i = \sigma \ \forall i$, allora il minimo si può trovare facilmente per via analitica.

Minimizzare tale distanza significa imporre che le derivate della funzione $\chi^2 = \chi^2(A, B)$ rispetto ad A e B siano nulle, cioè, ricordando che σ è una costante,

$$\begin{cases} 0 = \dfrac{\partial \chi^2}{\partial A} = -2 \displaystyle\sum_{i=1}^{n} (y_i - Ax_i - B)\, x_i \,, \\ 0 = \dfrac{\partial \chi^2}{\partial B} = -2 \displaystyle\sum_{i=1}^{n} (y_i - Ax_i - B) \,. \end{cases} \tag{4}$$

Dalla seconda equazione si ricava

$$B = \frac{1}{N}\left(\sum_{i=1}^{n} y_i - A\sum_{i=1}^{n} x_i\right) . \tag{5}$$

Sostituendo nella prima equazione si ottiene dunque

$$A = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{N}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 + \frac{1}{N}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i} . \tag{6}$$

Si noti che le dimensioni della grandezza fisica A sono quelle di $\left[\frac{y}{x}\right]$, mentre quelle di B sono le stesse di y.

I parametri A e B, essendo calcolati a partire da grandezze fisiche, sono anch'essi da considerarsi grandezze fisiche e di essi si deve calcolare l'errore. Il calcolo si può eseguire con la tecnica della propagazione dell'errore. Il risultato è:

$$\sigma_A^2 = \frac{N\sigma^2}{N\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{7}$$

$$\sigma_B^2 = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{N\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{8}$$

Il risultato si può usare come tale (non si conoscono i valori di A e di B e in questo modo si determinano) oppure per essere confrontato con un valore predetto da un modello. Nel caso del nostro esempio occorrerebbe verificare che A sia compatibile con il valore -1 e che B sia compatibile con 0.