Calcolo dell'errore sui coefficienti in una regressione lineare.

Giacomo Torzo e Michele Impedovo

Se due grandezze x e y sono legate teoricamente da una formula, $y = f(x, a_i)$, con M parametri a_i che esprimono grandezze per noi interessanti, e se otteniamo sperimentalmente N misure indipendenti di coppie di valori $\{x_j, y_j\}$, possiamo usare il *metodo dei minimi quadrati* per determinare i valori più attendibili dei parametri a_i Questa situazione è frequente nella pratica di laboratorio, e il caso più comune è quello in cui la relazione teorica sia *lineare tra x e y:* in tal caso il metodo viene detto *regressione lineare.*

Il metodo dei minimi quadrati (che vale per una funzione f qualsiasi) assume che la miglior curva sia quella corrispondente ai valori $\{a_i\}$ che rendono minima la somma dei quadrati delle differenze tra valori misurati y_j e valori interpolati $f(x_j, a_i)$. Questa assunzione parte dall'ipotesi (del resto spesso valida) che una delle due grandezze, per esempio x, sia *misurata in modo molto più preciso*, così da poterla considerare priva di errore.

Poiché le misure sono affette dagli errori σ_x e σ_y, la curva che meglio li interpola è quella che *distribuisce i punti ugualmente sopra e sotto,* cioè quella che minimizza la somma dei *quadrati delle distanze* dei punti dalla curva, ove tali "distanze" andrebbero misurate *lungo una direzione* la cui orientazione dipende dalla scala degli assi e dal rapporto σ_x/σ_y. Seguire questa impostazione è però piuttosto complicato[1] così che di solito si ricorre alla procedura molto più semplice giustificata dalla assunzione $\sigma_x = 0$.

In tal caso la direzione lungo la quale misurare la distanza del punto sperimentale dalla curva interpolante è quella dell'asse y.

La distanza del punto j-simo dal valore previsto, $f(x_j, a_i)$, è data da $\left|y_j - f(x_j, a_i)\right|$. Se "misuriamo" questa distanza in unità del relativo errore, σ_j, quadrando e sommando su tutte le misure, otteniamo la funzione "chi quadro":

$$\chi^2 = \sum_j \left[\frac{y_j - f(x_j, a_i)}{\sigma_j}\right]^2 \qquad [1]$$

Per calcolare i valori dei parametri a_i che minimizzano la funzione χ^2, calcoliamo le derivate parziali di χ^2 rispetto ai parametri a_i, le poniamo uguali a zero e risolviamo il sistema di equazioni:

$$\frac{\partial \chi^2}{\partial a_i} = -2\sum_j \left[\frac{1}{\sigma_j^2}\frac{\partial f(x_j, a_i)}{\partial a_i}\left[y_j - f(x_j, a_i)\right]\right] = 0 \qquad [2]$$

[1] Per una trattazione dettagliata di questo problema si veda Ralph H. Bacon, *The best straight line among the points* American Journal of Physics, **21**, 428-446 (1953).

Supposto che la relazione $f(x, a_i)$ sia la relazione lineare $y = ax + b$, e che gli errori sulle y_j siano *tutti uguali* e pari a σ_y otteniamo, con il metodo di Cramer:

$$a = \frac{N\sum x_j y_j - \sum x_j \sum y_j}{\Delta}$$
$$b = \frac{\sum x_j^2 \sum y_j - \sum x_j \sum x_j y_j}{\Delta} \qquad [3]$$
$$\Delta = N\sum x_j^2 - (\sum x_j)^2$$

Queste relazioni forniscono i valori stimati di a e b .

Come si vede, l'ipotesi che le incertezze sui valori delle y siano tutte uguali semplifica il calcolo, e rende i valori di a e b indipendenti dal valore di σ_y. Tuttavia è abbastanza ovvio che l'incertezza su a e b debba dipendere anche da σ_y. In assenza di informazioni indipendenti, normalmente si ricava σ_y una volta calcolata la retta di regressione come valore quadratico medio degli scarti dal valore interpolato linearmente:

$$\sigma_y = \sqrt{\frac{\sum (y_j - b - ax_j)^2}{N-2}} \qquad [4]$$

Una volta noto σ_y le incertezze sui valori possono essere stimate a partire dalle equazioni [3] con la formula generale di propagazione dell'errore.

$$\sigma_a = \sigma_y\sqrt{\frac{N}{\Delta}}$$
$$\sigma_b = \sigma_y\sqrt{\frac{\sum x_j^2}{\Delta}} \qquad [5]$$

Il calcolo eseguito su TI-89/92.

Le calcolatrici grafiche Texas eseguono automaticamente la regressione lineare [3] fornendo i valori dei coefficienti a e b ma non la stima della loro incertezza (σ_a e σ_b) che richiede il calcolo delle funzioni [4] e [5].
Un programma di poche righe, che esegue tale operazione sulle calcolatrici grafiche (TI-89/92) usando due liste di valori (Lx e Ly), viene qui portato come esempio.
Si è introdotta nel programma la possibilità di accettare o rifiutare il valore di σ_y calcolato mediante la [4], per consentire di usare eventualmente un valore maggiore suggerito da una stima indipendente della accuratezza dei valori y.

```
regres(lx,ly)
Prgm
Local s,d,da,db,st
ClrIO
LinReg lx,ly
regeq(x)→y1(x)
lx→xx:ly→yy
NewPlot 1,1,xx,yy
ZoomData
Pause
setMode("DisplayDigits","Float 4")
RegCoef[1]→a
RegCoef[2]→b
Disp "Y=aX+b= "&string(a)&"X+"&string(b)
TwoVar lx,ly
nStat*Σx²-Σx^2→d
√(sum((ly-b-a*lx)^2/(nStat-2)))→s
string(s)→st
Dialog
Title "Yerr ="&string(s)
Request "change Yyerr ?",st
EndDlog
expr(st)→s
√(nStat/d)→da
√(Σx²/d)→db
Disp "Yerr  ="&string(s)
Disp "a = "&string(a)
Disp "Da = "&string(da*s)
Disp "b = "&string(b)
Disp "Db = "&string(db*s)
EndPrgm
```

Può infatti accadere che una serie di punti sperimentali risulti casualmente molto *bene allineata*, anche se l'incertezza su *y* è abbastanza grande. Il risultato fornito dalle formule [4] e [5] viene allora *sottostimato* (nel caso particolare in cui i punti siano per caso tutti esattamente allineati l'incertezza calcolata per σ_y e quindi per σ_a e σ_b risulta infatti ovviamente nulla). In questi casi il valore di σ_y calcolato automaticamente dalla dispersione dei punti sperimentali va sostituito con il valore stimato in modo indipendente, in base alla conoscenza delle modalità in cui è stata eseguita la misura di y.

Un esempio.

Come esempio di applicazione possiamo considerare il calcolo della accelerazione come pendenza della retta di regressione in un grafico (velocità-tempo) ottenuto sperimentalmente per un moto che sappiamo essere uniformemente accelerato. Supponiamo di usare l'interfaccia CBL con CBR (o sensore di distanza) per registrare il moto di un oggetto che scende lungo un piano inclinato.
Usando il programma PHYSICS in modo "TIME GRAPH" e "NON-LIVE DISPLAY" otteniamo un file DATA in cui i tempi sono nella colonna c1, le posizioni nella colonna c4 e le velocità nella colonna c5 .
Siamo interessati a calcolare l'accelerazione come pendenza della retta che interpola i punti definiti dalle liste Ly=velocità e Lx=tempo.
Una volta ottenuto, mediante il menù "SELECT REGION", un file DATA che contiene solo i punti (t,v) che si vogliono interpolare con una retta, (supponiamo di salvare questo file con il nome `test`), dobbiamo costruire due liste, che ad esempio chiameremo Lx e Ly, a partire dai

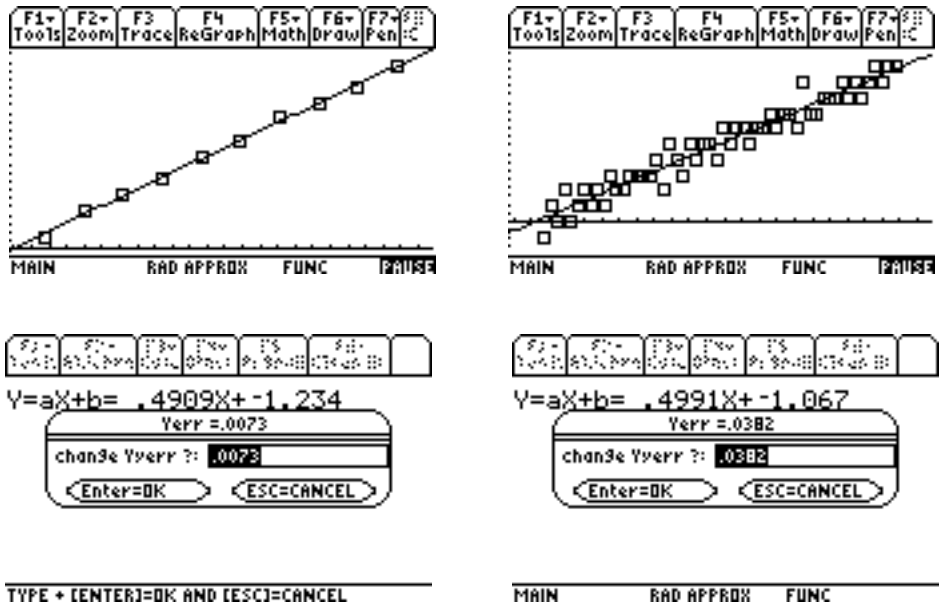
dati contenuti nel file test. Questo si ottiene, dato che i dati di tempo sono contenuti nella colonna c1 e quelli di velocità nella colonna c5, digitando nell'ambiente HOME le due righe di comando `test[1]→Lx` e `test[5]→Ly`, ove l'assegnazione (→) si ottiene premendo il tasto STO.

Poi basta digitare il comando `regres(Lx,Ly)` per eseguire il programma, che dapprima mostra i punti sperimentali, e poi (su comando ENTER) scrive il valore calcolato dalla dispersione dei punti (Yerr) di σ_y e chiede se tale valore ci va bene. Se vogliamo (in base alla conoscenza del modo in cui abbiamo ricavato i valori sperimentali) modificare il valore di *Yerr*, possiamo digitare nella finestra di dialogo il nuovo valore. Premendo ancora ENTER si ottengono i valori di a e di b insieme ai valori calcolati per σ_a e σ_b. In alternativa alla costruzione delle liste Lx e Ly si può usare il comando `regres(test[1],test[5])`.

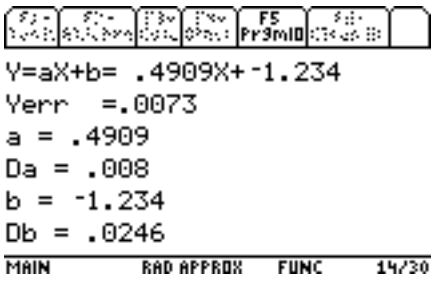
Come esempio concreto mostriamo cosa si ottiene con i dati relativi alla discesa di un carrello lungo un piano inclinato di α=3.5 gradi sulla orizzontale (accelerazione prevista $a=g\sin\alpha$=0.599 m/s^2).

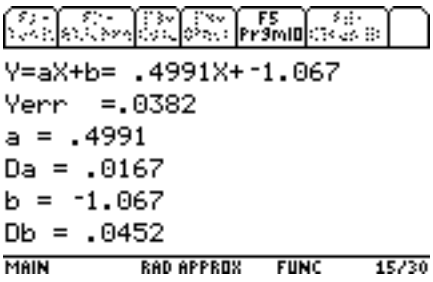
Da una stima a priori della precisione[2] del sonar valutiamo l'incertezza nella misura della posizione del carrello pari a 1 mm, e quindi l'incertezza nella misura della velocità (assumendo trascurabile l'errore nella misura del tempo di campionamento Δt) vale *Yerr* $=\sigma_v=0.001/\Delta t$ (m/s).

Facciamo due misure con tempi di campionamento diversi (Δt=0.1s e Δt=0.02s) ma per lo stesso tempo totale (1 s).

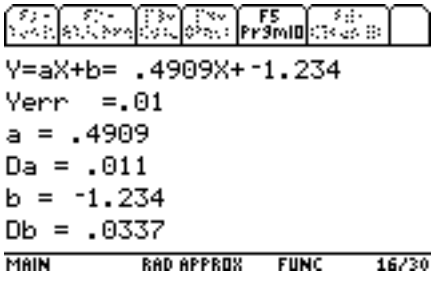


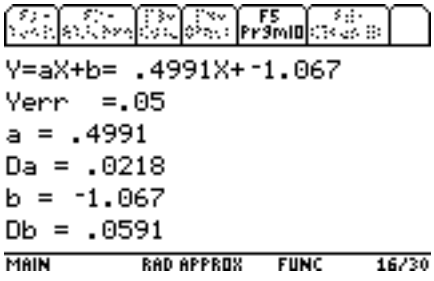
[2] Qui ci occupiamo solo degli errori *casuali*, tralasciando l'effetto dell'errore *sistematico* dovuto alla dipendenza della velocità del suono dalla temperatura ambiente che provoca una incertezza percentuale costante sui valori di distanza misurati, e che incide sulla *accuratezza* delle misure, non sulla loro *precisione*.





Questi dati mostrano che la frequenza di campionamento di 50Hz (Δt=0.02s) fa acquisire molti dati senza migliorare la precisione nella determinazione della pendenza rispetto alla frequenza di campionamento di 10Hz (Δt=0.1s): l'errore sulla pendenza *stimato in base alla sola dispersione dei punti*, passa da *Da*=0.8cm/s^2 a *Da*=1.7cm/s^2 .
I valori *Yerr* ricavati per σ_v dalla dispersione dei dati in entrambi i casi qui portati come esempio non si discostano molto da quelli calcolati a priori. Non ci troviamo quindi di fronte a dati casualmente eccessivamente allineati. Possiamo tuttavia ripetere il calcolo introducendo nella finestra di dialogo della subroutine `regres()` i valori *stimati a priori* (*Yer*r=0.01 e *Yerr*=0.05 rispettivamente) e questo modifica leggermente l'incertezza su *a* rispettivamente in *Da*=1 cm/s^2 e *Da*=2 cm/s^2 .





In questo esempio la stima dell'errore nel calcolo della accelerazione ci consente di concludere che il valore misurato (*a*=0.49±0.02 m/s^2) non è compatibile con il valore atteso in base al modello (*a*=0.599 m/s^2), e ci spingerà a prendere in considerazione la possibile presenza di effetti che avevamo trascurato, quali l'attrito, o una errata calibrazione del sonar. Una regressione lineare su dati relativi a *salita* e *discesa* potrà dare soluzione a questo problema, e fornire anche una misura del coefficiente di attrito.