La retta di regressione

Michele Impedovo

Uno dei temi nuovi e centrali per il rinnovamento dei programmi di matematica, che si impone in modo naturale quando si abbia a disposizione un qualunque strumento informatico, è quello di determinare la *miglior curva* che approssima una serie di dati osservati, solitamente forniti come punti

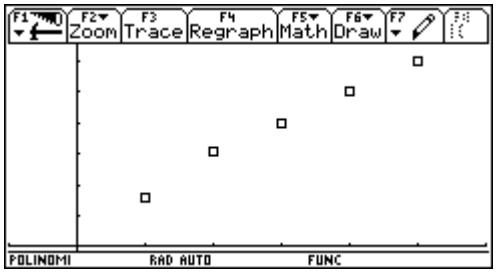
$$(x_i, y_i),\ i=1, \ldots n,$$

dove y è una grandezza che varia in funzione di x.
Si tratta di stabilire in modo ragionevole, sulla base delle informazioni disponibili, un buon modello (una retta, una curva esponenziale, una funzione potenza, eccetera) che si adatti ai punti. Una volta stabilito il tipo di funzione che si vuole adottare, occorre determinare la *miglior* funzione di quel tipo: un metodo per definire quale sia la *miglior funzione*, largamente utilizzato nella pratica scientifica, e di forte valenza concettuale è il metodo dei *minimi quadrati*.
Vediamo questo metodo nel caso più semplice, è quello della funzione lineare

$$x \to mx+q.$$

Supponiamo di avere n punti, per esempio i 5 punti

(1, 16), (2, 31), (3, 40), (4, 50), (5, 60).



e di voler determinare la funzione lineare che meglio si adatta. La definizione di *retta dei minimi quadrati*, o *retta di regressione* è la seguente: dati n punti (x_i, y_i), $i=1, \ldots n$, la retta di regressione è la retta di equazione

$$y = mx+q$$

che minimizza la *somma dei quadrati degli scarti*, cioè per la quale è minima la quantità (che è funzione di m e q)

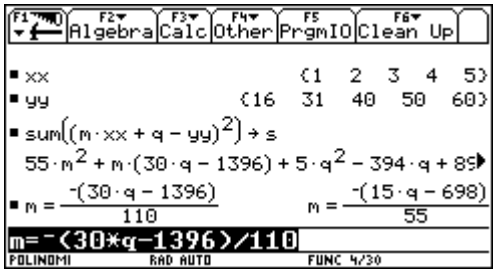
$$S(m, q) = \sum_{i=1}^{n} (mx_i + q - y_i)^2 \ .$$

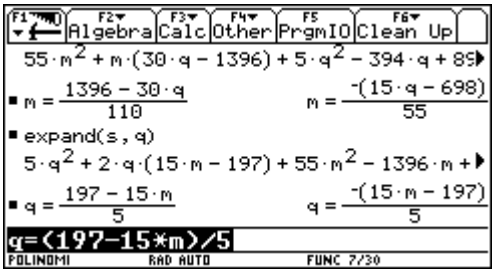
Perché proprio la somma dei quadrati degli scarti? La domanda ha implicazioni vaste, e non è questa la sede per una risposta esauriente. È relativamente facile convincere gli alunni che la somma degli scarti non è adatta a quantificare l'aderenza della retta agli n punti. Gli scarti possono essere infatti positivi o negativi e la loro somma può essere piccola in valore assoluto anche per rette palesemente inadatte a descrivere gli n punti. Per esempio, siano dati i tre punti allineati (1,1), (2,2), (3,3). Ovviamente la *miglior* retta è $y = x$; la somma degli scarti è nulla. Ma è nulla anche per qualunque retta passi per (2,2), quindi di equazione

$$y = m\,(x–2)+2.$$

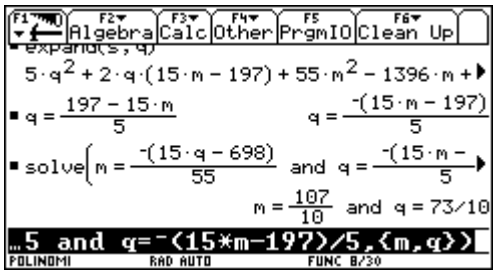
Dunque ci servono scarti che siano misurati da valori positivi; perché allora non usare la somma del valori assoluti degli scarti? È una scelta plausibile. Tuttavia la somma dei quadrati anziché dei valori assoluti si *sposa* in modo naturale con la media aritmetica: la media aritmetica di una sequenza di numeri gode della proprietà di **rendere minima** la somma dei quadrati degli scarti, mentre la mediana minimizza la somma dei valori assoluti. Si dimostra che la retta dei minimi quadrati passa per il *baricentro* dei punti, cioè il punto che ha per coordinate le medie aritmetiche delle ascisse e delle ordinate.

$S(m, q)$ è un polinomio di secondo grado in m e q; per minimizzare $S(m, q)$ non occorrono le derivate: se si pensa S come polinomio in m (e q come parametro), il grafico di $S(m)$ è una parabola con la concavità verso l'alto; il valore di m che rende minimo S è l'ascissa del vertice.
Ordinando il polinomio prima rispetto ad m e poi rispetto a q si ottengono due relazioni lineari tra m e q.



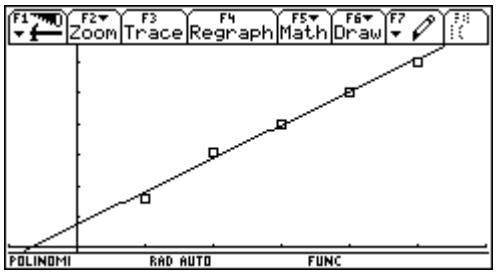


Risolvendo il sistema delle due equazioni si ottiene la soluzione:



$m = 107/10$, $q = 73/10$. La funzione lineare cercata è dunque:

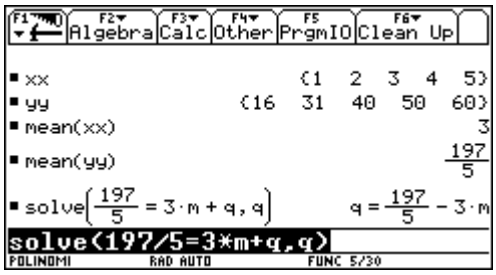
$$x \to 10.7x + 7.3.$$



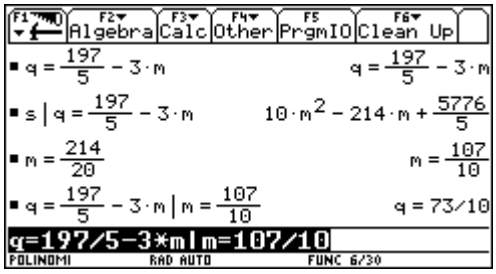
Un altro modo di ottenere m e q è quello di assumere (oppure di imporre) una proprietà importante della retta di regressione: il fatto che essa passi comunque per il baricentro dei punti, che è il punto

$$\left(3, \frac{197}{5}\right).$$

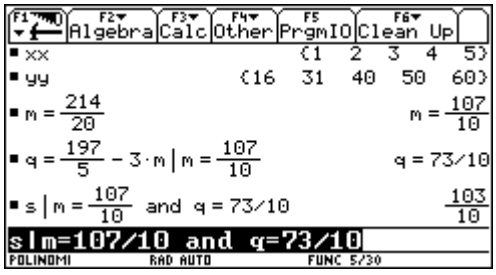
Possiamo allora esprimere q in funzione di m e ridurci ad una sola incognita.



Ora S risulta essere un polinomio di secondo grado nella sola m: l'ascissa del vertice è il minimo per S.

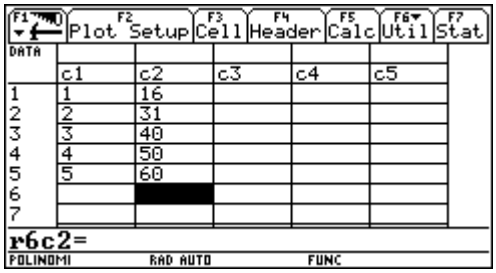


Calcoliamo per questa retta la somma dei quadrati degli scarti.

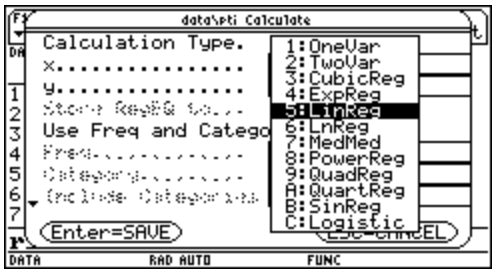


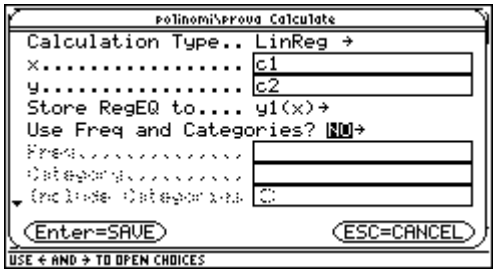
Risulta *S*=10.3: non è possibile, con una retta, fare di meglio.

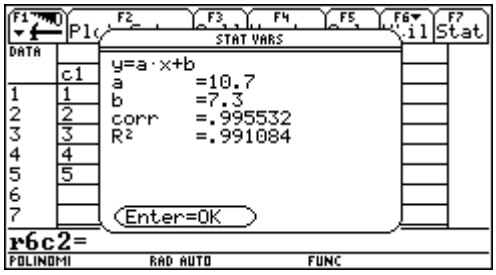
Questo processo, illustrato passo-passo, è utile per mostrare agli alunni i fondamenti teorici. Con un numero limitato di punti l'intero calcolo può essere effettuato con carta e penna. Naturalmente quando il numero di punti è elevato non ha più senso utilizzare carta e penna. La TI-92 mette a disposizione nell'ambiente Data/Matrix Editor la possibilità sia di tracciare un grafico a dispersione dei dati, sia di calcolare la curva di regressione che si vuole adottare, scegliendola tra diverse famiglie di funzioni (lineare, quadratica, cubica, esponenziale, potenza, logaritmica, …).
Cerchiamo allora di confermare il risultato già ottenuto utilizzando direttamente la TI-92.
Costruiamo la tabella dei punti.



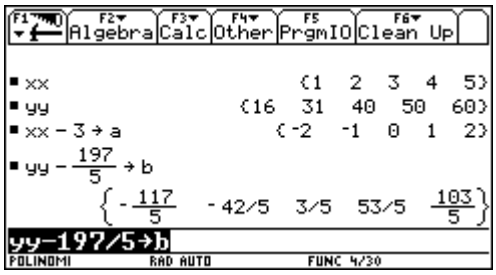
Con F5, Calc possiamo scegliere il tipo di funzione da assumere come modello, nel nostro caso scegliamo una funzione lineare (LINREG).







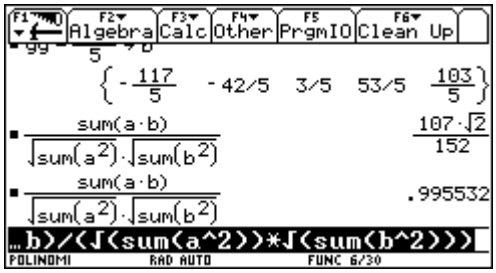
Il risultato è confermato. Inoltre ci viene fornito il valore del *coefficiente di correlazione lineare* (corr) e il suo quadrato (R^2). La correlazione lineare è relativamente alta.
Il coefficiente di correlazione lineare è un numero compreso tra –1 e 1, è negativo o positivo a seconda che si tratti di una decrescita o una crescita, vale 0 in caso di assenza di "linearità" nei dati, vale 1 o –1 quando i punti sono allineati. Uno degli aspetti più interessanti della statistica è proprio questo: è possibile *misurare* il grado di linearità che possiedono i dati grezzi, e *quantificare* l'adattabilità dei dati ad un andamento lineare. Come è definito il coefficiente di correlazione lineare?
Occorre innanzitutto riferire i dati alla loro *media*.



Si ottengono due nuovi vettori **a** e **b**. Il coefficiente di correlazione lineare non è altro che il **coseno** di tali vettori, cioè il rapporto tra il loro prodotto scalare e il prodotto delle loro norme:

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} = \frac{\sum \mathbf{a}_i \mathbf{b}_i}{\sqrt{\sum \mathbf{a}_i^2} \sqrt{\sum \mathbf{b}_i^2}}$$

Questo fatto non deve sorprenderci: così come il coseno di due vettori nel piano o nello spazio è un numero reale compreso tra –1 e 1, e in qualche modo misura (attraverso ilo coseno) l'**angolo** tra i due vettori, cioè la loro "distanza angolare", cioè ancora il fatto che siano disposti lungo la stessa direzione, nello stesso modo si può calcolare l'"angolo" (o meglio il suo coseno) tra due vettori qualsiasi, e il risultato ci dà informazioni su quanto i due vettori siano oppure no "indipendenti", siano oppure no linearmente correlati. Evidentemente il concetto di *angolo* è molto più ricco di quanto siamo abituati a pensare, è più ricco del semplice significato geometrico.
Calcoliamo ora finalmente cos(**a**, **b**).



Come si vede, si ottiene lo stesso valore fornito direttamente da LINREG.

Una generalizzazione del procedimento di ricerca dei parametri *m* e *q* dell'equazione
$y = mx+q$
della retta di regressione conduce alle notevoli formule

$$\begin{cases} m = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ q = \bar{y} - m\bar{x} \end{cases}$$

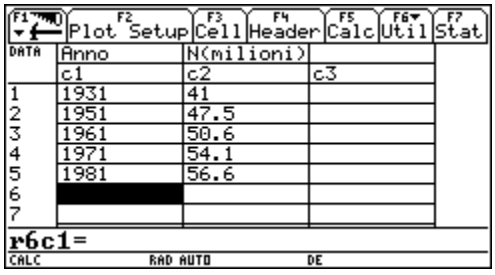
dove $\bar{x}$ e $\bar{y}$ sono rispettivamente la media aritmetica delle ascisse e delle ordinate degli *n* punti, e il punto $(\bar{x}, \bar{y})$ è il *baricentro* della distribuzione.

I due coefficienti *m* e *q* della funzione lineare

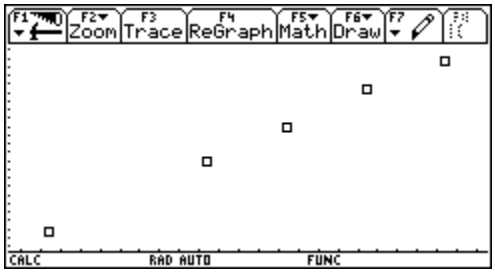
$$x \to mx+q$$

avranno per tutto il triennio un significato geometrico importantissimo: *m* è la **pendenza** costante della funzione (*pendenza* è un'espressione migliore di *coefficiente angolare*, che è lunga e in definitiva sbagliata se le grandezze rappresentate sui due assi non sono omogenee: per esempio nel piano *spazio-tempo* della fisica le unità di misura sono arbitrarie, e quindi sono arbitrari anche gli angoli), cioè l'incremento costante di *y* per un incremento unitario di *x*; *q* è il valore che la funzione assume per *x*=0.

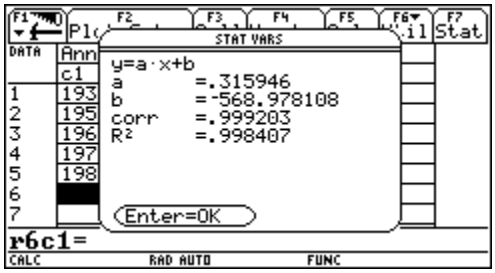
Applichiamo ora quanto visto ad un esempio significativo di crescita lineare. Nella tabella seguente sono riportati il numero di residenti in Italia come risulta dai censimenti ufficiali dal 1931 al 1981 (dati ISTAT).



Vediamo il grafico:



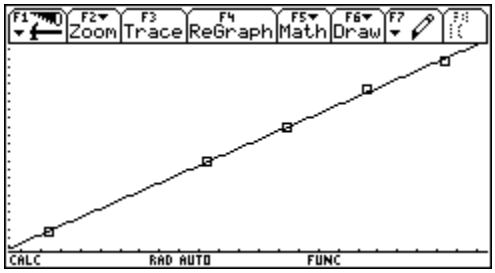
e ricaviamo l'equazione della retta di regressione lineare.



La funzione lineare è dunque del tipo

$$N\colon t \to 0.316t - 569$$

e si adatta molto bene ai dati.



La pendenza della retta di regressione (la cui unità di misura è "numero di abitanti, in milioni, all'anno") ci fornisce l'incremeno medio annuo della popolazione: circa 316.000 unità.
Da questo modello possiamo ragionevolmente *interpolare*, cioè ricavare valori attendibili in un istante compreso tra 1931 e 1981.
L'*estrapolazione* invece (cioè il ricavare valori esterni all'intervallo considerato) si rivela fallimentare nel nostro esempio. Secondo il modello lineare avremmo dovuto aspettarci al censimento del 1991 un numero di abitanti pari a $N(1991)=60.1$ milioni di abitanti; invece nel decennio 81-91 il calo demografico ha modificato radicalmente il tasso di crescita della popolazione italiana, che al censimento del 1991 è risultata essere pari a 56.8 milioni di abitanti, registrando praticamente una *crescita zero*.