Regressione: un'ipotesi di percorso didattico

Michele Impedovo

Riassunto. Nota una tabella di dati relativi alle osservazioni di due grandezze X e Y, è naturale formulare ipotesi su quale possa essere una ragionevole funzione che rappresenti o che approssimi la relazione tra X e Y. Si tratta, in un certo senso, di capovolgere il tradizionale "studio di funzioni". Il metodo dei minimi quadrati è una risposta largamente condivisa a tale problema. In questo articolo viene presentata un'ipotesi di percorso didattico che prende le mosse dal modello più semplice di regressione (quello lineare), per trattare poi le funzioni polinomiali, le funzioni potenza, le funzioni esponenziali, e accennare infine al problema più generale della regressione non lineare.

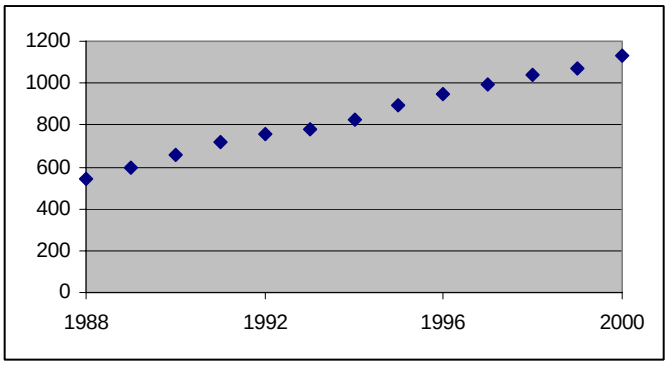
Abstract. Starting with a numerical table displaying two quantities X and Y, it is natural to wonder which reasonable function could describe or approximate the relationship between X and Y. A largely used way to try and solve the problem is the least squares method. In this paper I present a teaching journey across linear and polynomial regression, through power and exponential regression and end up mentioning the general problem of non linear regression.

Michele Impedovo
Università Bocconi di Milano
michele.impedovo@uni-bocconi.it

La retta di regressione: il problema

La tabella seguente riporta i dati relativi all'andamento del PIL (Prodotto Interno Lordo, in miliardi di Euro) in Italia dal 1988 al 2000 (fonte ISTAT, www.istat.it).

1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
546	598	660	720	759	782	827	894	951	994	1039	1072	1129



Il grafico mostra un andamento sostanzialmente lineare. Sorge naturale la domanda: qual è una buona funzione lineare $f(x) := ax+b$ che approssima i dati?

Si tratta di un problema che non dovrebbe mancare nella preparazione scientifica di base di qualunque studente. La commissione UMI (Unione Matematica Italiana), della quale ho fatto parte, ha recentemente prodotto una proposta di curriculum (vedi *Matematica 2003* al sito UMI http://www.dm.unibo.it/umi) per la scuola secondaria, e ha lanciato la parola d'ordine "la matematica per il cittadino": l'attività matematica non deve solo essere propedeutica a studi superiori, anzi, il curriculum deve essere pensato per tutti gli studenti, e in particolar modo per coloro che non proseguiranno con studi scientifici. In questa prospettiva assume particolare rilevanza il problema di stimare l'andamento dei dati di una tabella; in particolare all'uscita dalle scuole superiori ogni studente dovrebbe saper riconoscere e analizzare i seguenti fondamentali modelli di variazione:

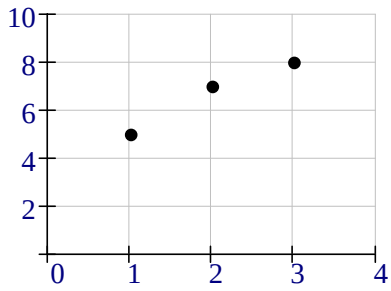
– modelli lineari: $x \to ax+b$
– modelli potenza: $x \to ax^b$
– modelli esponenziali: $x \to ab^x$

Le osservazioni che seguono vogliono essere una proposta di percorso didattico in questa direzione.

La retta di regressione: un possibile percorso didattico

Provate a proporre agli studenti, senza alcuna preparazione preliminare, il seguente problema, da svolgere a piccoli gruppi.

Qual è la miglior retta che "passa" per i punti (1,5), (2,7), (3,8)?



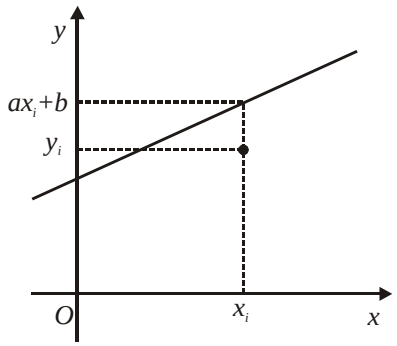
La formulazione del problema può sembrare un po' *naif*, ma è invece il risultato di diversi aggiustamenti; l'obiettivo è trasmettere un significato condiviso. Per me è sempre stato fonte di sorpresa scoprire che gli studenti non hanno dubbi sul significato del problema: si mettono al lavoro senza chiedere quale sia la *definizione* di "miglior retta" ed escogitano le più disparate strategie. Per loro la "miglior retta" esiste, si tratta soltanto di scovarla; nei loro lavori compare spesso l'idea di *media* (per esempio alcuni calcolano la media delle pendenze e delle intercette delle rette passanti per ogni coppia di punti).

La definizione di retta di regressione non è banale (si basa sull'idea di *minimo*), ma ha una valenza concettuale notevole e quindi vale la pena di introdurla nei curriculum, sia perché si adatta bene all'intuizione degli studenti, sia perché è costruttiva e fornisce direttamente un metodo di calcolo per la pendenza a e l'intercetta b. Vediamo come passare dall'intuizione alla definizione.

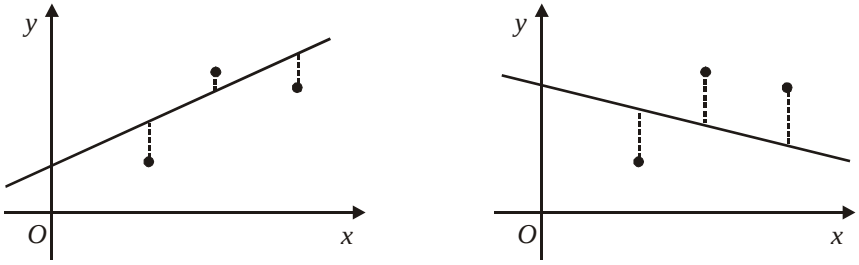
Siano dati n punti di ascisse $X := \{x_1, \ldots, x_n\}$ e ordinate $Y := \{y_1, \ldots, y_n\}$; si consideri, per ogni punto (x_i, y_i), la differenza

$$ax_i + b - y_i$$

che fornisce lo scarto tra l'ordinata "teorica" ax_i+b sulla retta incognita e l'ordinata "osservata" y_i.



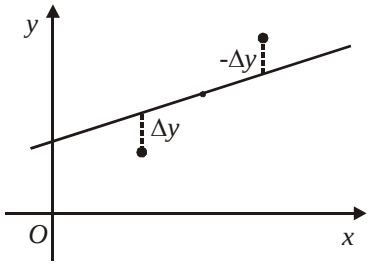
L'intuizione suggerisce che in qualche modo si debba rendere minima una funzione che coinvolge gli *n* scarti. Per esempio, tra le seguenti rette (riferite alla stessa terna di punti) nessuno studente ha dubbi sul fatto che la prima è "migliore" della seconda. Si tratta di quantificare questa intuizione.



Un'ipotesi potrebbe essere quella di definire "miglior retta" quella che rende più piccola possibile la *somma* degli scarti

$$\sum_{i=1}^{n}\left(ax_i+b-y_i\right);$$

ma tale definizione condurrebbe ad un paradosso. Infatti ogni scarto può essere positivo o negativo; se consideriamo due soli punti, la miglior retta *deve* essere quella che passa per essi (somma degli scarti: 0 + 0 = 0), invece qualunque retta che passi per il loro punto medio realizza anch'essa una somma degli scarti uguale a 0 ($\Delta y - \Delta y = 0$).



Avremmo così infinite "migliori rette" e la cosa non ci garba: dunque non vogliamo che gli scarti si annullino a vicenda, ci serve una funzione che contempli solo "scarti" positivi; i matematici, quando vogliono rendere positiva una quantità, ricorrono a due possibili scelte: ne prendono il valore assoluto, oppure il quadrato. Ecco nei due casi la somma da minimizzare:

$$\sum_{i=1}^{n} \left| ax_i + b - y_i \right|, \qquad \text{oppure} \qquad \sum_{i=1}^{n} \left(ax_i + b - y_i \right)^2 .$$

A ben pensarci, lo stesso problema si presenta con la definizione di media aritmetica di n numeri $z_1, \ldots, z_n$:

$$m := \frac{1}{n} \sum_{i=1}^{n} z_i .$$

Infatti la media gode di due fondamentali proprietà:

– m è il numero che annulla la somma degli scarti:

$$\sum_{i=1}^{n} \left(z_i - m \right) = \left(\sum_{i=1}^{n} z_i \right) - nm = \sum_{i=1}^{n} z_i - n \frac{1}{n} \sum_{i=1}^{n} z_i = 0.$$

– m è il numero che rende minima la somma dei quadrati degli scarti, nel senso che la funzione

$$f(x) := \sum_{i=1}^{n} \left(z_i - x \right)^2$$

assume valore minimo in m; infatti $f(x)$ è una funzione quadratica il cui grafico è una parabola con la concavità rivolta verso l'alto:

$$f(x) = \sum_{i=1}^{n} z_i^2 - 2x \sum_{i=1}^{n} z_i + nx^2$$

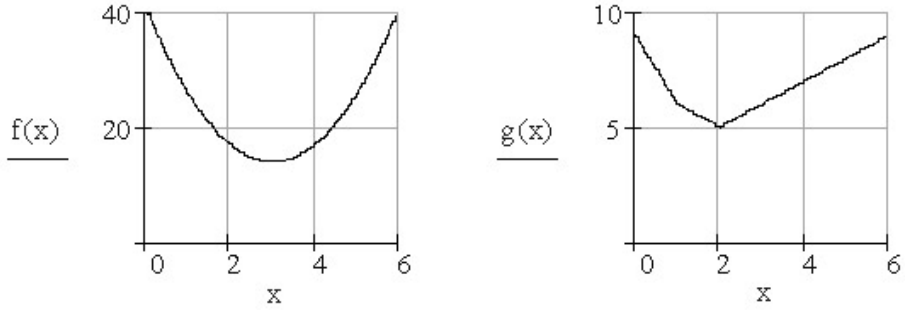
e il vertice ha ascissa

$$\frac{1}{n}\sum_{i=1}^{n} z_i = m.$$

La seconda proprietà spiega perché l'indice di posizione *media aritmetica* si "sposa" con l'indice di dispersione *somma dei quadrati* degli scarti. In modo analogo si dimostra che la *mediana* si sposa con la *somma dei valori assoluti* degli scarti: infatti la mediana di $z_1, \ldots, z_n$ è il numero rende minima la funzione

$$g(x) := \sum_{i=1}^{n} \left| z_i - x \right|.$$

I grafici seguenti mostrano le funzioni $f(x)$ (somma dei quadrati degli scarti) e $g(x)$ (somma dei valori assoluti degli scarti) per la terna $Z := [1, 2, 6]$, che ha media $m = 3$ e mediana $M = 2$.



Il fatto che la media aritmetica minimizzi la somma dei quadrati degli scarti costituisce un buon motivo per misurare la dispersione di n numeri dalla media m mediante la varianza VAR e la deviazione standard σ:

$$\text{VAR} := \frac{1}{n}\sum_{i=1}^{n}\left(z_i - m\right)^2 \qquad\qquad \sigma := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(z_i - m\right)^2}$$

Siamo finalmente pronti per la definizione di "miglior retta".

Dati n punti di ascisse $X := \{x_1, \ldots, x_n\}$ e ordinate $Y := \{y_1, \ldots, y_n\}$, la retta di regressione di Y rispetto a X è la funzione lineare $y = ax + b$ che minimizza la funzione

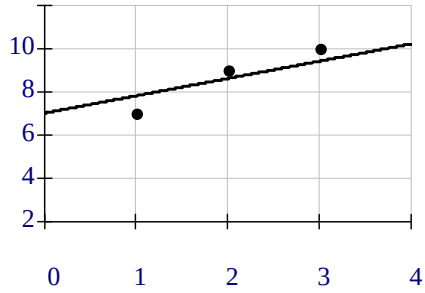
$$S(a, b) := \sum_{i=1}^{n}\left(ax_i + b - y_i\right)^2.$$

Tale funzione lineare è chiamata *retta di regressione* o anche *retta dei minimi quadrati*.

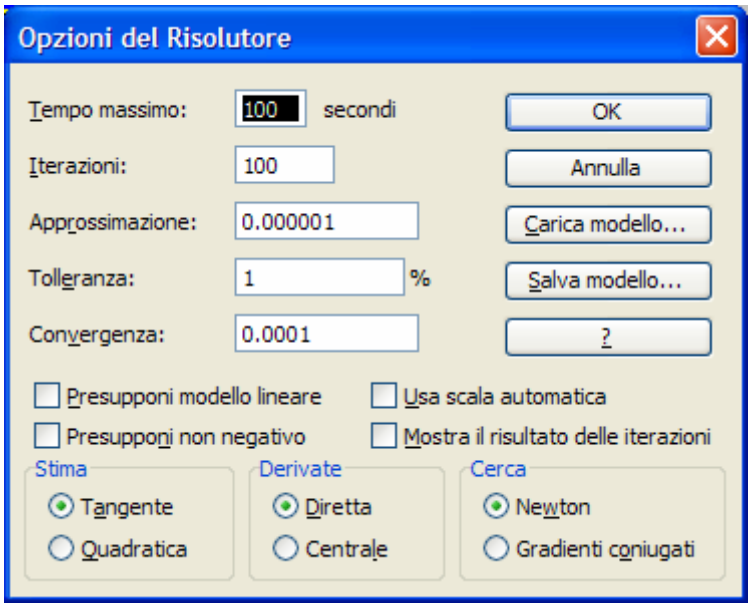
Si tratta di una definizione convenzionale, ma convincente.

Un buon modo per consolidare il valore semantico di questa definizione è quello di proporre agli studenti la seguente attività: si costruisce un foglio elettronico come il seguente, in cui si impostano due valori ragionevoli per a e b, si riportano le ascisse e le ordinate dei punti (colonne C e D), si calcola per ciascun punto il quadrato dello scarto (colonna E) e infine se ne calcola la somma $S(a, b)$ (cella F2); vince chi, modificando opportunamente le celle A2 e B2, trova il valore minore nella cella F2. Nella figura seguente viene calcolato $S(0.8, 5) = 1.16$.

	A	B	C	D	E	F
1	**a**	**b**	**x**	**y**	**(ax+b-y)^2**	**S(a,b)**
2	0,8	5	1	5	0,64	1,16
3			2	7	0,16	
4			3	8	0,36	
5						



Un altro modo, ugualmente efficace, consiste nel far costruire agli studenti un foglio di Cabri in cui, per ciascun punto, si costruisce il quadrato che ha per lato lo scarto, e si sommano le aree.



Conclusioni

Nel percorso delineato la ricerca della “miglior funzione” che approssima i dati osservati di due grandezze può risultare un’attività ricca dal punto di vista didattico, e ben si adatta alla parola d’ordine “matematica per il cittadino”, perché permette di esplorare e consolidare la padronanza dei fondamentali modelli lineare, potenza, esponenziale. Inoltre permette di affrontare da un punto di vista generale il problema della risoluzione di un’equazione o di un sistema non lineare, cogliendone le difficoltà implicite.

Infine può mostrare una spendibilità culturale della matematica che troppo spesso manca nei nostri curriculum.