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.

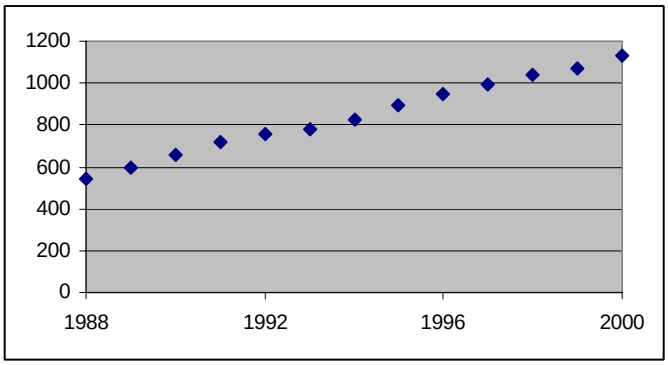
Abstract. Starting with a numerical table displaying two quantities X and Y, it is natural to wonder which reasonable function could describe or approximate the relationship between X and Y. A largely used way to try and solve the problem is the least squares method. In this paper I present a teaching 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:

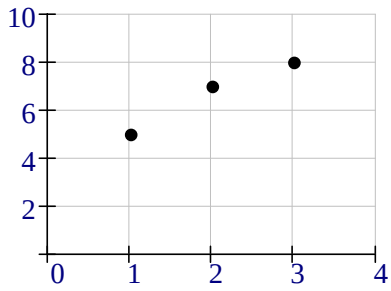
– modelli lineari: $x \to ax+b$
– modelli potenza: $x \to ax^b$
– modelli esponenziali: $x \to ab^x$

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)?



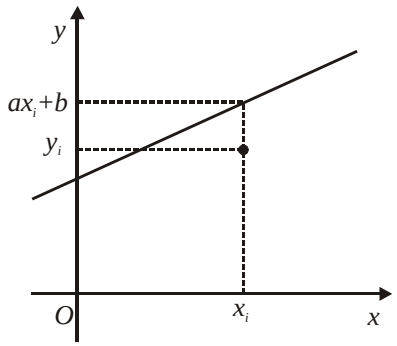
La formulazione del problema può sembrare un po' *naif*, ma è invece il risultato di diversi aggiustamenti; l'obiettivo è trasmettere un significato condiviso. Per me è sempre stato fonte di sorpresa scoprire che gli studenti non hanno dubbi sul significato del problema: si mettono al lavoro senza chiedere 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.

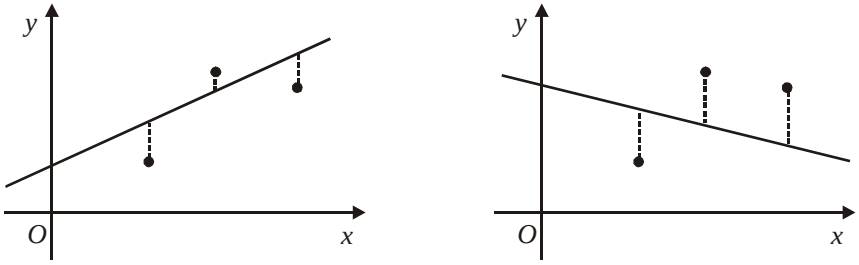
Siano dati n punti di ascisse $X := \{x_1, \ldots, x_n\}$ e ordinate $Y := \{y_1, \ldots, y_n\}$; si consideri, per ogni punto (x_i, y_i), la differenza

$$ax_i+b-y_i$$

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



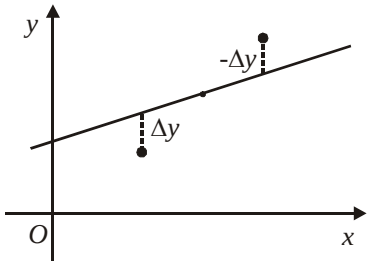
L'intuizione suggerisce che in qualche modo si debba rendere minima una funzione che coinvolge gli *n* scarti. Per esempio, tra le seguenti rette (riferite alla stessa terna di punti) nessuno studente ha dubbi sul fatto che la prima è "migliore" della seconda. Si tratta di quantificare questa intuizione.



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:

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

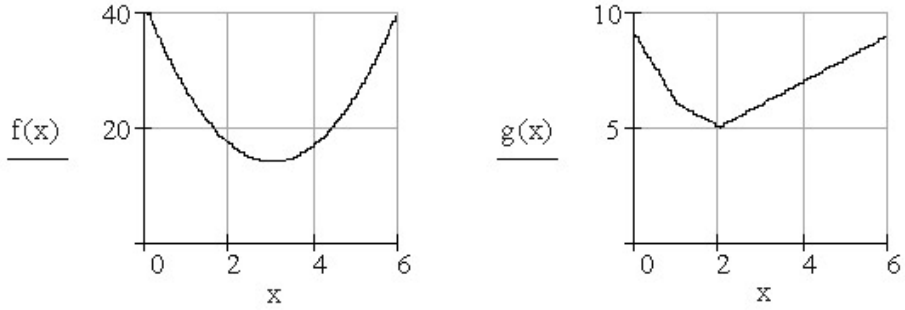
e il vertice ha ascissa

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

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

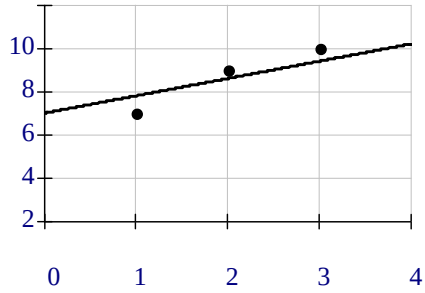
$$S(a, b) := \sum_{i=1}^{n}\left(ax_i + b - y_i\right)^2.$$

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

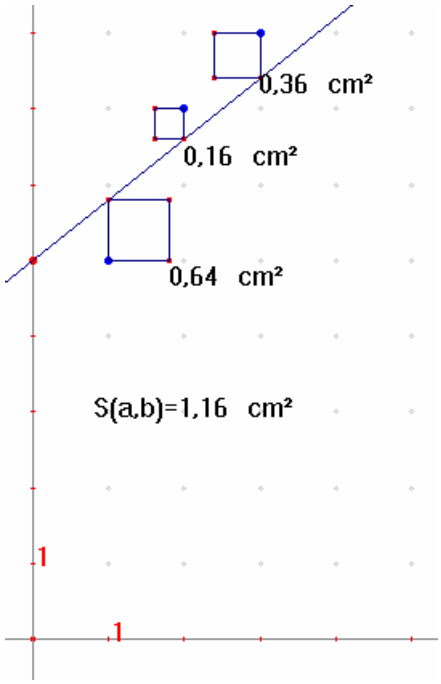
Si tratta di una definizione convenzionale, ma convincente.

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.



Se le attività precedenti hanno fatto prendere confidenza con la definizione, ora non ci resta che procedere al calcolo simbolico di a e b in funzione di $x_1, \ldots, x_n$ e $y_1, \ldots, y_n$. Questo lavoro, che è un semplice calcolo algebrico, può (e deve) essere eseguito almeno una volta con carta e penna, proprio per poter in seguito utilizzare le tecnologie con consapevolezza.

Lavorando sull'esempio precedente, la funzione da minimizzare è

$$S(a, b) := (a+b-5)^2+(2a+b-7)^2+(3a+b-8)^2.$$

Essa esprime, in funzione delle variabili incognite a e b, la somma dei quadrati degli scarti del modello lineare $x \to ax+b$ dai dati osservati.

Poiché S è una funzione quadratica di a e b, non occorrono le derivate per minimizzarla; è sufficiente sapere che una funzione quadratica (di una variabile)

$$x \to Ax^2 + Bx + C,$$

con $A > 0$, assume valore minimo globale nell'ascissa del vertice

$$x = \frac{-B}{2A}.$$

Ragioniamo in questo modo: se pensiamo b fissato, S è una funzione quadratica di a (con coefficiente direttivo positivo)

$$S(a, b) = 14a^2+2(6b-43)a+(3b^2-40b+138)$$

e dunque, qualunque sia b, assume valore minimo nel vertice:

$$a = (43-6b)/14$$

In modo analogo per b:

$$b = (20-6a)/3$$

Risolvendo il sistema lineare di queste due equazioni si ottiene $a = 3/2$, $b = 11/3$ e la retta di regressione è dunque

$$y = \frac{3}{2}x+\frac{11}{3}.$$

Risulta $S(3/2, 11/3) = 1/6 \approx 0.167$ e non è possibile far di meglio: per nessuna coppia di valori a, b è possibile che la somma dei quadrati degli scarti sia minore di 1/6. La funzione lineare $f(x) := \frac{3}{2}x + \frac{11}{3}$ è dunque la "miglior" funzione lineare, secondo il metodo dei minimi quadrati.

La retta di regressione e il calcolo differenziale

Dal punto di vista del calcolo differenziale in più variabili il problema di minimizzare la funzione

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

è un problema "facile", perché S è una funzione quadratica in a e b e dunque il sistema delle derivate parziali di S, rispetto ad a e a b, uguagliate a 0

$$\begin{cases} S'_a(a,b) = 0 \\ S'_b(a,b) = 0 \end{cases}$$

è lineare rispetto ad a e b; la sua risoluzione è affidata ad algoritmi di "bassa" complessità computazionale che forniscono in forma simbolica la soluzione in funzione dei dati x_k e y_k. Vediamo i calcoli in dettaglio.

$$\begin{cases} 2\sum_{i=1}^{n}\left(ax_i+b-y_i\right)x_i=0 \\ 2\sum_{i=1}^{n}\left(ax_i+b-y_i\right)=0 \end{cases};$$

dividendo per 2 ed espandendo:

$$\begin{cases} a\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}x_iy_i \\ a\sum_{i=1}^{n}x_i+nb=\sum_{i=1}^{n}y_i \end{cases}$$

da cui

$$a=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2} \qquad b=\frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}x_iy_i\sum_{i=1}^{n}x_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

Le espressioni simboliche sembrano inguardabili, ma possiamo scriverle in una forma più significativa. Se indichiamo con

– $\overline{X}$: la media di $X := \{x_1, \ldots, x_n\}$, cioè $\frac{1}{n}\sum_{i=1}^{n}x_i$

– $\overline{Y}$: la media di $Y := \{y_1, \ldots, y_n\}$, cioè $\frac{1}{n}\sum_{i=1}^{n}y_i$

– $\overline{X^2}$: la media di $X^2 := \left\{x_1^2,\ldots,x_n^2\right\}$, cioè $\frac{1}{n}\sum_{i=1}^{n}x_i^2$

– $\overline{XY}$: la media di $XY := \{x_1y_1, \ldots, x_ny_n\}$, cioè $\frac{1}{n}\sum_{i=1}^{n}x_iy_i$

allora dividendo per n il sistema diventa

$$\begin{cases} a\overline{X^2}+b\overline{X}=\overline{XY} \\ a\overline{X}+b=\overline{Y} \end{cases}.$$

La seconda equazione ci dice una cosa importante (e auspicabile): il *baricentro*

$$\left(\overline{X}, \overline{Y}\right)$$

cioè il punto le cui coordinate sono la media delle ascisse e la media delle ordinate, appartiene necessariamente alla retta di regressione, perché soddisfa l'equazione $y = ax + b$. In effetti ci saremmo stupiti del contrario. Un percorso didattico alternativo potrebbe dare per buona questa proprietà del baricentro (che consente di esprimere b in funzione di a) e ridursi così a minimizzare la funzione quadratica $S(a)$ della sola variabile a.

Possiamo ora scrivere la soluzione mediante un'espressione simbolica più semplice:

$$a = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2}, \qquad b = \overline{Y} - a\overline{X}\ .$$

Per chi desidera qualche suggestione statistica, a è il rapporto tra la covarianza di X e Y e la varianza di X:

$$a = \frac{\mathrm{Cov}\left(X, Y\right)}{\mathrm{Var}\left(X\right)}.$$

La retta di regressione con Excel

Ora che sappiamo come calcolare pendenza e intercetta della retta di regressione, vogliamo utilizzare il foglio elettronico per ottenere rapidamente lo stesso risultato. Sfruttiamo il primo esempio, quello del PIL italiano. I comandi

=PENDENZA(intervalloY; intervalloX)
=INTERCETTA(intervalloY; intervalloX)

forniscono rispettivamente i valori dei parametri a e b della retta di regressione $y = ax + b$ (nella versione inglese i comandi sono rispettivamente SLOPE e INTERCEPT).

La figura seguente mostra il calcolo di a e b nelle celle D1 e D2:

D1: =PENDENZA(B1:B13;A1:A13)
D2: =INTERCETTA(B1:B13;A1:A13)

D1 =PENDENZA(B1:B13;A1:A13)

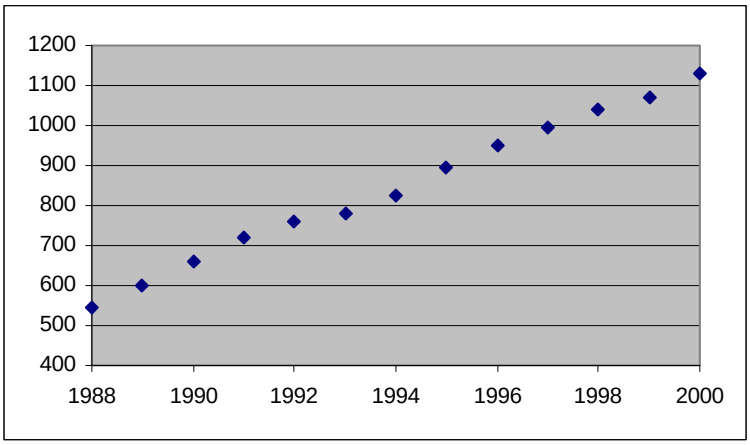
	A	B	C	D	E
1	1988	546	a =	47.81	
2	1989	598	b =	-94496	
3	1990	660			
4	1991	720			
5	1992	759			
6	1993	782			
7	1994	827			
8	1995	894			
9	1996	951			
10	1997	994			
11	1998	1039			
12	1999	1072			
13	2000	1129			

La retta di regressione è dunque la funzione lineare

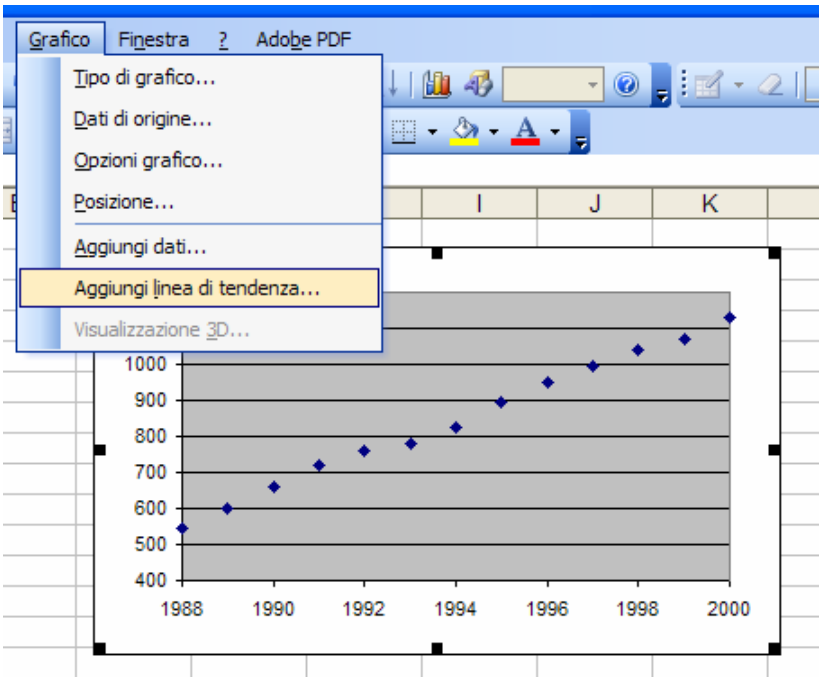
$$f(x) := 47.8x–94496.$$

Il significato della pendenza è chiaro: in media il PIL è aumentato di circa 47.8 miliardi di Euro all'anno. Il significato dell'intercetta è un po' comico: rappresenterebbe, in condizioni di linearità, il PIL italiano nell'anno 0.

Possiamo ottenere direttamente grafico ed equazione della retta di regressione utilizzando il comando "Aggiungi linea di tendenza"; costruiamo innanzitutto il grafico per punti dei dati osservati, con il comando Inserisci, Grafico, Dispersione, dopo aver selezionato l'intervallo di celle A1:B13.

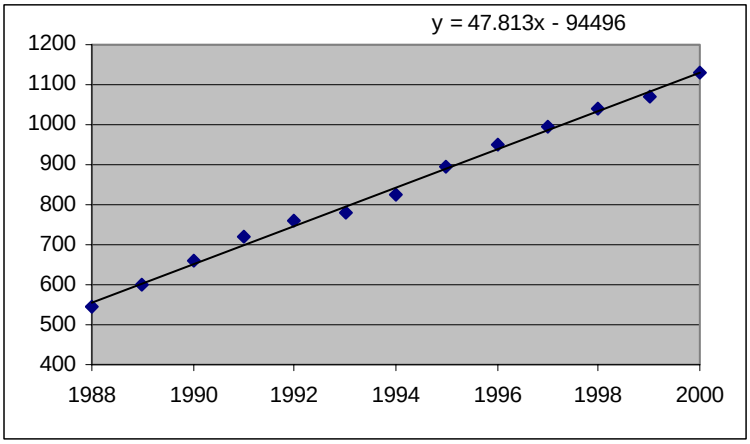


Ora selezioniamo il grafico: nella barra dei menù compare la voce “Grafico”. Clicchiamo su Aggiungi linea di tendenza.



Nella scheda Tipo scegliamo “Lineare” e nella scheda “Opzioni” selezioniamo “Visualizza l’equazione sul grafico”.

Il risultato è il seguente.



Polinomi di regressione

Fin qui abbiamo applicato il metodo dei minimi quadrati a funzioni lineari $f(x) := ax + b$, con due parametri. Il metodo dei minimi quadrati si può applicare in linea di principio a qualsiasi famiglia di funzioni $f(x)$, con un numero qualsiasi di parametri, per esempio ad una funzione polinomiale di grado m (e quindi con $m+1$ parametri).

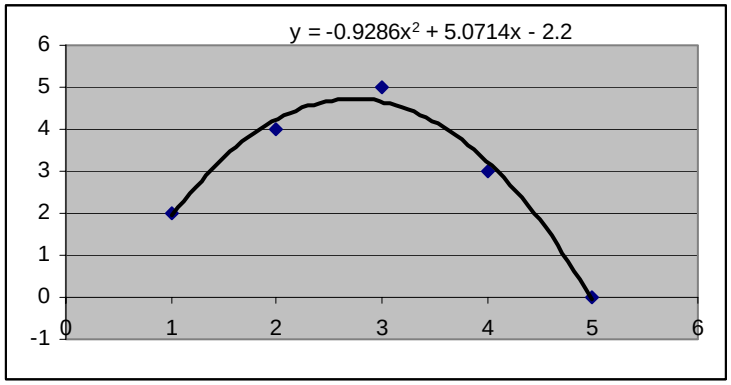
Esempio. Determiniamo, con il metodo dei minimi quadrati, la miglior funzione quadratica $f(x) := ax^2 + bx + c$ che approssima i punti le cui coordinate sono raccolte nei seguenti vettori:

$$X := [1, 2, 3, 4, 5], \qquad Y := [2, 4, 5, 3, 0].$$

La somma dei quadrati degli scarti è data dalla funzione

$$S(a, b, c) := \sum_{i=1}^{5} \left(ax_i^2 + bx_i + c - y_i \right)^2 .$$

Attenzione: f è *quadratica* rispetto ad x ma è *lineare* rispetto ai parametri a, b, c (si parla in questo caso di *regressione lineare*); come prima, la funzione S è quadratica nelle variabili a, b, c, le derivate parziali di S rispetto ad a, b, c sono lineari e il sistema delle derivate parziali uguagliate a 0 è lineare. La figura seguente mostra la risoluzione con il comando “Aggiungi linea di tendenza” di tipo polinomiale di grado 2 (Excel calcola il polinomio di regressione fino al grado 6).



In generale, se indichiamo con $f(a_1, \ldots, a_p, x)$ la funzione che si vuole adottare come modello, dove $a_1, \ldots, a_p$ sono i parametri incogniti, allora il metodo dei minimi quadrati porta a minimizzare la funzione

$$S(a_1, \ldots, a_p) := \sum_{i=1}^{n} \left(f\left(a_1, a_2, \ldots, a_p, x_i\right) - y_i \right)^2 ,$$

e quindi a risolvere il sistema

$$\begin{cases} S_{a_1}'\left(a_1, \ldots, a_p\right) = 0 \\ \ldots \\ S_{a_p}'\left(a_1, \ldots, a_p\right) = 0 \end{cases} .$$

Se il modello f è polinomiale allora tale sistema è lineare e la sua risoluzione non offre particolari difficoltà.

Se il modello di funzione *f* che si vuole adottare non è lineare rispetto ai parametri (si parla allora di *regressione non lineare*), il sistema delle derivate parziali di *S* uguagliate a 0 può essere di ardua risoluzione. Torneremo tra poco sull'argomento.

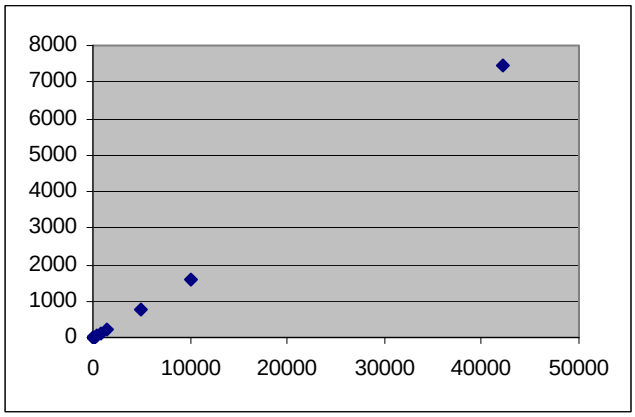
La regressione potenza

La tabella seguente mostra i record mondiali maschili di atletica leggera (gennaio 2006) nelle discipline olimpiche di corsa, dai 100 metri alla maratona.

distanza (m)	tempo (s)
100	9.77
200	19.32
400	43.18
800	101.11
1500	206.00
5000	757.35
10000	1577.53
42195	7440.55

Prima di vedere il grafico dei punti: che ne dite di un modello lineare? No, eh? Anche gli studenti sono concordi: qui la linearità non c'entra, non si può correre la maratona alla stessa velocità media dei 100 m. Allora che grafico ci aspettiamo? Ci aspettiamo una concavità verso l'alto, cioè i tempi impiegati dovrebbero crescere più rapidamente delle distanze percorse.

Beh, probabilmente il grafico vi stupirà. Eccolo.



Sorpresi? I punti non sono allineati, ma … poco ci manca! Eppure tutti siamo convinti che in questo contesto il modello lineare non sia sensato. Il fatto è che le ascisse dei punti non sono distribuite in modo uniforme; c'è una grande ressa vicino all'origine e poi c'è un salto enorme tra i 10000 metri e la maratona: il grafico è poco leggibile.

Quale modello continuo può adattarsi a questi punti?

Osserviamo innanzitutto che al tendere a 0 di x (distanze) deve tendere a 0 anche y (tempi): cerchiamo una curva che passi dall'origine. Inoltre deve avere concavità verso l'alto; se scartiamo una funzione lineare per l'origine $x \to ax$ possiamo ipotizzare una funzione potenza

$$f(x) := ax^b,$$

con un esponente b lievemente maggiore di 1.

Dunque il modello f che vogliamo adottare non è lineare rispetto ai parametri a e b. Anche in questo caso definiamo funzione potenza di regressione quella che minimizza la somma dei quadrati degli scarti:

$$S(a,b) := \sum_{k=1}^{n} \left(ax_k^b - y_k \right)^2$$

Per esempio, la funzione potenza di regressione dei punti $X := \{1,2,3\}$, $Y := \{5,7,8\}$ è quella che minimizza la funzione

$$S(a,b) := (a-5)^2+(a2^b-7)^2+(a3^b-8)^2.$$

Ma, ahimé, minimizzare questa funzione è tutt'altro che banale. Il sistema delle derivate parziali uguagliate a 0 non è lineare nei parametri a e b:

$$\begin{cases} S_a'(a,b) = 2\sum_{i=1}^{n} \left(ax_i^b - y_i \right)^2 x_i^b \\ S_b'(a,b) = 2\sum_{i=1}^{n} \left(ax_i^b - y_i \right)^2 ax_i^b \ln(b) \end{cases}.$$

Tale sistema non solo non ammette una soluzione simbolica, ma anche la ricerca di un'approssimazione richiede algoritmi di livello superiore. Seguiamo un'altra strada, che si può percorrere in classe.

Se la relazione tra x e y fosse del tipo $y = ax^b$, passando ai logaritmi (per esempio in base 10) risulterebbe

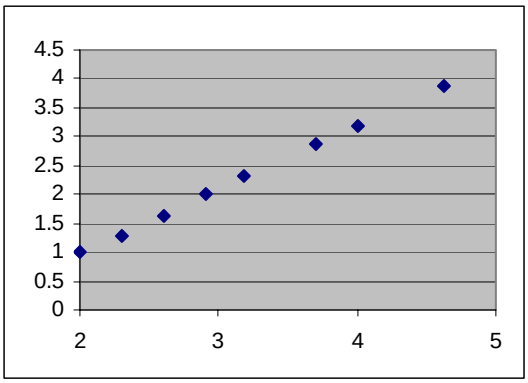
$$\log(y) = \log(a) + b\log(x)$$

cioè, ponendo $\log(y) = Y$, $\log(x) = X$:

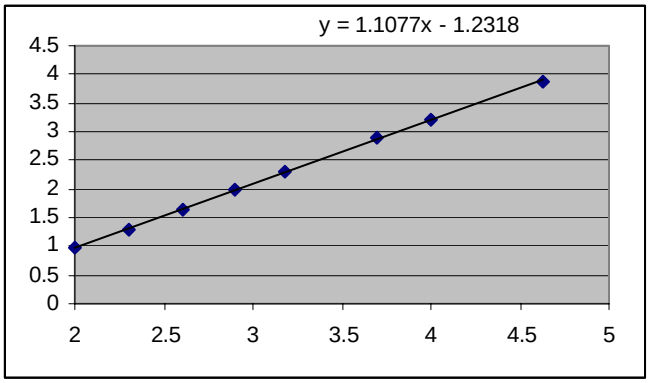
$$Y = A + BX,$$

il che significa: se y fosse una funzione potenza di x allora il logaritmo di y si comporterebbe linearmente rispetto al logaritmo di x, con pendenza $B = b$ e intercetta $A = \ln(a)$. Confrontiamo allora $\log(y)$ con $\log(x)$.

x	y	log(x)	log(y)
100	9.77	2.000	0.990
200	19.32	2.301	1.286
400	43.18	2.602	1.635
800	101.11	2.903	2.005
1500	206.00	3.176	2.314
5000	757.35	3.699	2.879
10000	1577.53	4.000	3.198
42195	7440.55	4.625	3.872



I punti sembrano allineati! Calcoliamo la retta di regressione di $\log(y)$ rispetto a $\log(x)$ direttamente con “Aggiungi linea di tendenza” di Excel.



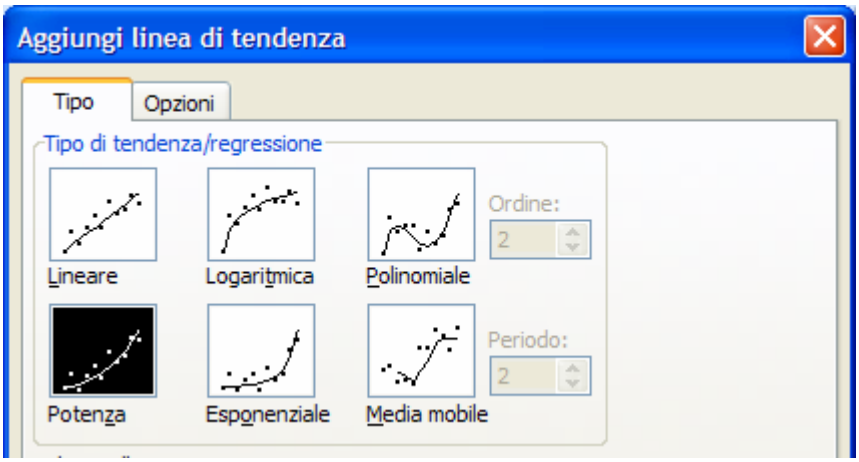
Otteniamo $b \approx 1.1077$ e $\log(a) \approx {}^{-}1.2318$, da cui $a \approx 10^{1.2318} \approx 0.05864$. Un modello per i record di corsa è dunque il seguente.

$$y = 0.05864 \cdot x^{1.1077}.$$

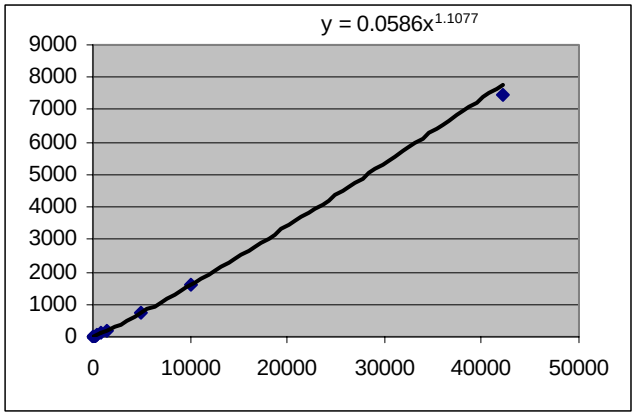
Siamo arrivati a questo risultato attraverso un’osservazione importante:

– y è una funzione potenza di x se e solo se $\log(y)$ è una funzione lineare di $\log(x)$.

Tale osservazione ci autorizza d'ora in avanti a chiedere direttamente, con il comando "Aggiungi linea di tendenza", il modello potenza sulle colonne *x* e *y*.



Otteniamo infatti la stessa funzione: per Excel il miglior modello potenza su *x* e *y* è la retta di regressione su log(*x*) e log(*y*).



Secondo tale modello il record sulla mezza maratona (20 km) dovrebbe essere circa 3407 s, cioè 56'48". In realtà è 58'55" (H. Gebrselassie, 15/1/2006).

Dobbiamo ricordare che la retta di regressione di log(*y*) rispetto a log(*x*) non fornisce la "migliore" funzione potenza di *y* rispetto a *x*, nel senso che la coppia (*a*, *b*) che minimizza la funzione

$$T(a,b) := \sum_{k=1}^{n}\left(\log(a)+b\log(x_k)-\log(y_k)\right)^2 .$$

(è questo il problema che abbiamo risolto) non è uguale alla coppia (a, b) che minimizza la funzione:

$$S(a,b) := \sum_{k=1}^{n}\left(ax_k^b - y_k\right)^2 .$$

Tuttavia in generale ne è una buona approssimazione, e per questo motivo viene implementata nei sistemi di calcolo; il fatto è che minimizzare $T(a, b)$ è facile. Infatti, posto $A := \log(a)$, $B := b$, $X := \log(x)$, $Y := \log(y)$ possiamo scrivere

$$T(A,B) := \sum_{k=1}^{n}\left(A+BX_k - Y_k\right)^2 .$$

Così T è quadratica nelle variabili A e B, e, come prima, il sistema delle derivate parziali uguagliate a 0

$$\begin{cases} T_A'(A,B)=0 \\ T_B'(A,B)=0 \end{cases}$$

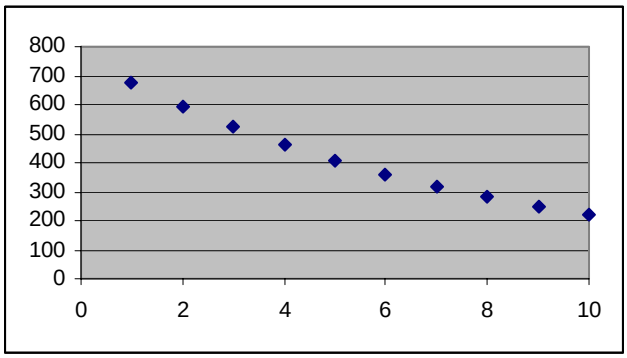
è un sistema lineare nelle variabili A e B. Una volta calcolati A e B si risale facilmente ad a e b.

Ancora un'osservazione sugli esempi sin qui svolti: la scelta di dati che non sono legati da una legge deterministica è voluto e non casuale. Non c'è scritto da nessuna parte che il PIL debba seguire un andamento lineare nel tempo, né che le migliori prestazioni sulle diverse distanze debbano seguire un modello potenza. La scelta del modello è nostra. Può essere una scelta ragionevole, sensata, convincente (e non mancano argomenti statistici in questa direzione), ma non *vera* o *falsa*. Come accade spesso nella vita, dobbiamo operare una scelta e ce ne assumiamo la responsabilità.

La regressione esponenziale

Vediamo ora un esempio governato da una legge fisica. I dati seguenti riportano la pressione atmosferica media (in millimetri di mercurio) in funzione dell'altezza sul livello del mare (in km).

h (km)	p (mmHg)
1	674
2	596
3	526
4	462
5	405
6	360
7	318
8	281
9	248
10	219



Il grafico e le informazioni sulla grandezza fisica in esame lasciano pensare ad un andamento esponenziale decrescente, cioè del tipo

$$y = ab^x$$

con $0 < b < 1$. La funzione esponenziale di regressione è, per definizione, quella che minimizza la funzione in a e b:

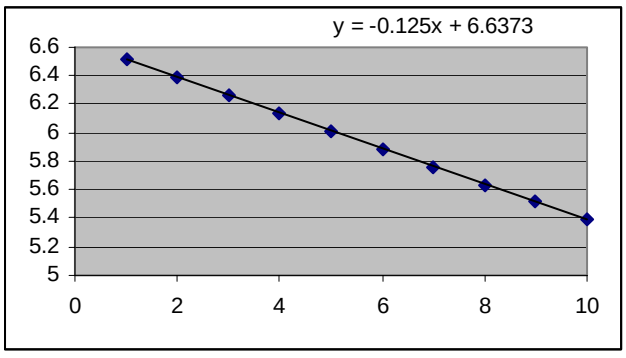
$$S(a,b) := \sum_{i=1}^{n}\left(ab^{x_i} - y_i\right)^2$$

Come prima, passiamo ai logaritmi (usiamo questa volta il logaritmo naturale). Se y fosse una funzione esponenziale di x, risulterebbe

$$\ln(y) = \ln(a) + x\ln(b)$$

cioè $\ln(y)$ sarebbe espresso da una funzione lineare di x, di pendenza $\ln(b)$ e intercetta $\ln(a)$. Proviamo allora a tracciare il grafico del logaritmo della pressione in funzione dell'altezza.

h (km)	ln(p)
1	6.51323
2	6.390241
3	6.265301
4	6.135565
5	6.003887
6	5.886104
7	5.762051
8	5.638355
9	5.513429
10	5.389072



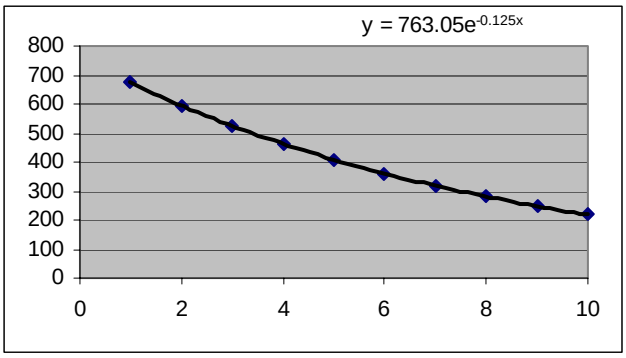
La retta di regressione tra x e $\ln(y)$ fornisce la relazione

$$\ln(y) = 6.6373 - 0.125x$$

e dunque il modello esponenziale tra x e y è

$$y = \exp(6.6373) \cdot \exp(^{-}0.125)^{x} \approx 763 \cdot 0.88^{x}.$$

Usando il comando "Aggiungi linea di tendenza" di tipo esponenziale otteniamo direttamente grafico ed equazione della funzione esponenziale cercata, espressa da Excel nella forma ae^{kx} anziché nella forma ab^{x}.



Sulla base dei dati analizzati la pressione atmosferica vale circa 763 mmHg al livello del mare e diminuisce circa del 12% per ogni aumento di 1 km dell'altezza sul livello del mare.

Anche per la regressione esponenziale è stato risolutivo "passare ai logaritmi" e risolvere poi un problema di regressione lineare. Ricordiamo che anche in questo caso la coppia (a, b) che minimizza la funzione

$$T(a, b) := \sum_{i=1}^{n} \left(\ln(a) + x_i \ln(b) - y_i \right)^2$$

non è uguale alla coppia (a, b) che minimizza la funzione

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

Ponendo $A := \ln(a)$, $B := \ln(b)$ risulta

$$T(a, b) = \sum_{i=1}^{n} \left(A + x_i B - y_i \right)^2 ,$$

cioè, ancora una volta, T è quadratica in A e B e il sistema delle derivate parziali uguagliato a 0 è lineare in A e B. I valori di a e b trovati sono delle approssimazioni dei valori a e b che minimizzano $S(a, b)$.

Il problema generale della regressione

Il problema generale della ricerca di una curva di regressione potrebbe essere formulato in questo modo: siano dati n punti di ascisse $X := \{x_1, \ldots, x_n\}$ e ordinate $Y := \{y_1, \ldots, y_n\}$, e la famiglia di funzioni

$$f(a_1, \ldots, a_p, x)$$

che dipendono da p parametri $a_1, \ldots, a_p$ e dalla variabile x. La curva di regressione della famiglia $f(a_1, \ldots, a_p, x)$ è quella che minimizza la somma dei quadrati degli scarti, cioè quella che minimizza la funzione

$$S(a_1, \ldots, a_p) := \sum_{i=1}^{n} \left(f\left(a_1, \ldots, a_p, x_i\right) - y_i \right)^2 .$$

Se il sistema

$$\begin{cases} S_{a_1}'\left(a_1, \ldots, a_p\right) = 0 \\ \ldots \\ S_{a_p}'\left(a_1, \ldots, a_p\right) = 0 \end{cases}$$

nelle incognite $a_1, \ldots, a_p$ non è lineare allora può essere impossibile esprimere la soluzione in forma simbolica in funzione di $x_1, \ldots, x_n$ e $y_1, \ldots, y_n$.

Si apre un problema arduo: risolvere un sistema non lineare.

Qui non resisto alla tentazione di osservare che nelle nostre scuole si insegna a ottimizzare (localmente) una funzione $f(x)$ attraverso lo strumento-chiave dell'equazione $f'(x) = 0$ ma non si dice in modo chiaro che in generale un'equazione non si risolve con metodi "onesti" e che occorre approssimare la soluzione per mezzo di opportuni algoritmi. Se f è una funzione in una variabile l'algoritmo di Newton è uno strumento molto efficiente: data un'equazione $g(x) = 0$ e un valore iniziale x_0, la successione

$$x_{n+1} := x_n - \frac{g(x_n)}{g'(x_n)},$$

sotto opportune ipotesi, converge ad una soluzione di $g(x) = 0$.

Ma se f è una funzione in più variabili allora il problema è molto più complesso. Esiste una letteratura assai vasta sull'argomento, a cavallo tra ricerca matematica e statistica, (basta cercare *non linear regression* con Google; tra gli algoritmi più interessanti della ricerca recente segnalo gli "algoritmi genetici" di ottimizzazione) ma non esistono "vie regie"; esistono diversi algoritmi, che tengono conto anche del livello di regolarità di f (differenziabile, convessa, continua, definita a tratti, …).

Lo stesso algoritmo di Newton è generalizzabile in più variabili: il sistema non lineare

$$\begin{cases} S_{a_1}'(a_1,\ldots,a_p) = 0 \\ \ldots \\ S_{a_p}'(a_1,\ldots,a_p) = 0 \end{cases}$$

si può scrivere in forma vettoriale

$$\boldsymbol{S}'(\mathbf{x}) = \mathbf{0}.$$

Dato un punto iniziale $\mathbf{x}_0$, sotto opportune ipotesi, la successione

$$\mathbf{x}_{n+1} := \mathbf{x}_n - \boldsymbol{S}''(\mathbf{x}_n)^{-1}\boldsymbol{S}'(\mathbf{x}_n),$$

(dove $\boldsymbol{S}''(\mathbf{x}_n)$ è la matrice hessiana di $\boldsymbol{S}$, $\boldsymbol{S}''(\mathbf{x}_n)^{-1}$ è la sua matrice inversa e $\boldsymbol{S}'(\mathbf{x}_n)$ è il vettore gradiente) converge alla soluzione del sistema. Il fatto è che in più variabili l'algoritmo di Newton è instabile, dipende fortemente dalla condizione iniziale e se $\mathbf{x}_0$ non è "abbastanza vicino" alla soluzione allora la successione $\mathbf{x}_n$ può essere irregolare o addirittura divergere.

Vogliamo concludere mostrando con un esempio le potenzialità di Excel sulla regressione non lineare.

Esempio. Si vuole stimare la miglior funzione esponenziale decrescente

$$f(a, b, x) := a \exp(^{-}bx)$$

che approssima i punti $(1, 10)$, $(2, 5)$, $(3, 2)$; occorre dunque minimizzare la funzione

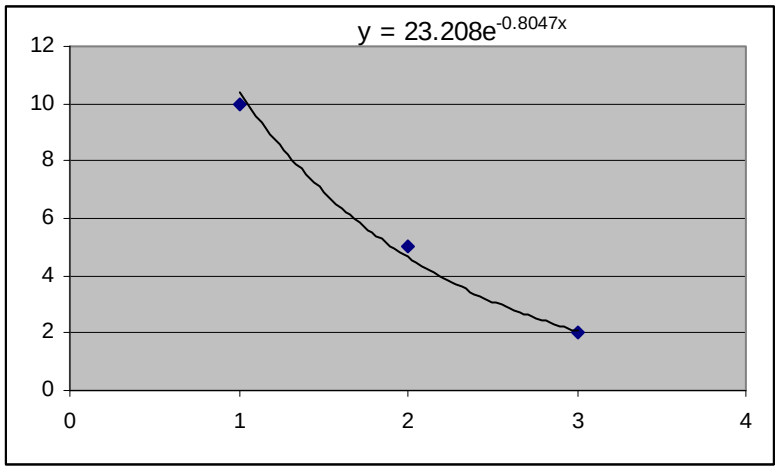
$$S(a, b) = \sum_{i=1}^{3} \left(f\left(a, b, x_i\right) - y_i \right)^2 = (ae^{-b} - 10)^2 + (ae^{-2b} - 5)^2 + (ae^{-3b} - 2)^2.$$

Utilizzando la regressione lineare di $\ln(y)$ rispetto a x, Excel trova la soluzione

$$a^* \approx 23.208,\ b^* \approx 0.8047,$$

per la quale risulta

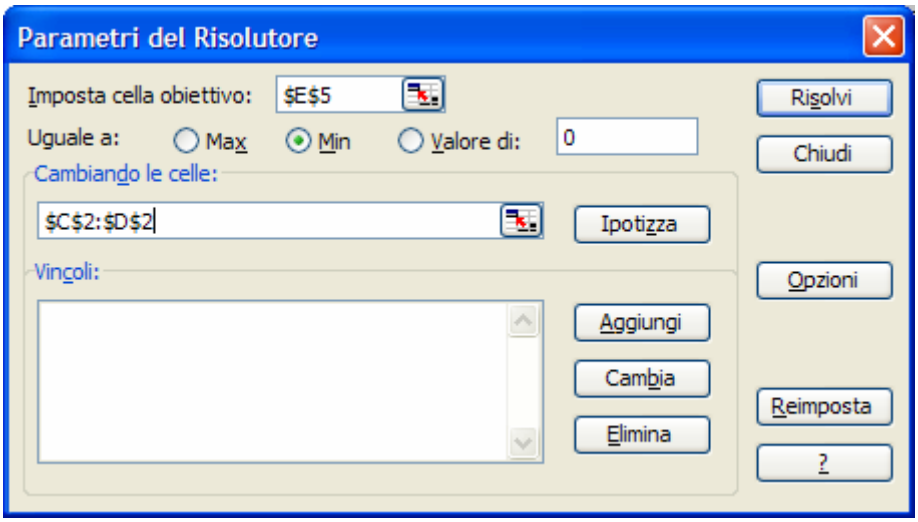
$$S(a^*, b^*) \approx 0.277825.$$



Si può far di meglio? E cioè: si può trovare una coppia (a, b) tale che $S(a, b)$ sia significativamente minore di 0.277825? Nelle colonne A e B scriviamo le ascisse e le ordinate dei punti. In C2 e D2 scriviamo i valori a^* e b^* trovati con la regressione lineare. Nella colonna E calcoliamo i quadrati degli scarti e sommiamoli in E5.

x	y	a	b	S(a,b)
1	10	23.208	0.8047	0.143739
2	5			0.128324
3	2			0.005762
				0.277825

Ora utilizziamo il Risolutore di Excel: impostiamo la cella E5 al valore minimo cambiando le celle C2 e D2.

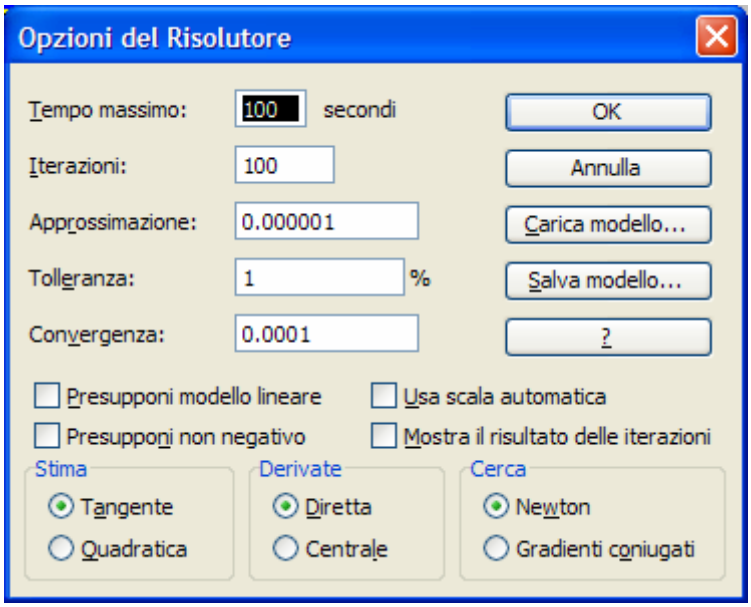


x	y	a	b	S(a,b)
1	10	21.24622	0.747936	0.003219
2	5			0.057464
3	2			0.064133
				0.124816

Come si vede, il Risolutore ha trovato i nuovi valori

$$a^* \approx 21.24622 \text{ e } b^* \approx 0.747936$$

per i quali $S(a^*, b^*) \approx 0.125$ è più che dimezzato! Nelle opzioni del Risolutore è possibile modificare alcuni parametri di ricerca per migliorare ulteriormente la soluzione.



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.