Stima dei parametri nei modelli non-lineari

Indice

Introduzione

Come sappiamo dallo studio di altre discipline, la degradazione di un erbicida nel terreno può essere modellizzata ricorrendo ad una funzione di 'decadimento esponenziale".

$$C(t) = C_0 \, exp(-kt) + \epsilon$$

dove t è il tempo, C è la concentrazione al temp t ed ϵ (l'errore del modello) segue una qualche distribuzione di probabilità.

La parametrizzazione consiste nell'individuare i valori più appropriati per $ C_0 $ e k in un determinato scenario. Studiando la funzione anzidetta è facile vedere che $ C_0 $ rappresenta la concentrazione iniziale, mentre k è il tasso relativo di decremento della concentrazione e misura la velocità del processo.

Nella lezione precedente abbiamo visto che la parametrizzazione può avvenire tramite:

- informazioni in letteratura;
- secondo l'esperienza locale;
- mediante esperimenti 'ad hoc'

Per parametrizzare il modello organizziamo un esperimento, nel quale prepariamo un campione di terreno a concentrazione nota, lo poniamo in cella climatica a temperatura ed umidità costante e, a tempi prestabiliti, preleviamo un'aliquota e misuriamo la concentrazione residua.

Otteniamo i seguenti risultati medi:

t	C(t)
0	100
10	50
20	25
30	15
40	7
50	3
60	2
70	1

Possiamo ora parametrizzare il modello di decadimento esponenziale in base ai dati osservati. Se siamo pronti ad assumere che la componente stocastica del modello è perfettamente normale con varianze omogenee ed indipendenti dalla X, il metodo di parametrizzazione più semplice e praticato è quello dei minimi quadrati, che consiste nel cercare i valori dei parametri che minimizzano lo scostamento tra i valori osservati e le stime del modello.

Nel caso di modelli lineari nei parametri (modelli polinomiali), il problema dei minimi quadrati ha una soluzione algebrica, che invece non necessariamente esiste per modelli intrinsecamente non lineari. Abbiamo tre soluzioni possibili, che illlustriamo brevemente.

Approssimazione locale

La prima cosa che possiamo fare, dopo aver osservato i dati, è approssimare la cinetica di decadimento esponenziale con una funzione curvilinea, ma lineare nei parametri, come una parabola.

Questa soluzione consente di eseguire facilmente i calcoli, ma, come possiamo osservare, non è scevra da inconvenienti, dato che il modello prevede che la concentrazione di erbicida inizi a crescere ad un certo punto della degradazione. Per questa mancanza di senso biologico, l'approssimazione locale deve essere sempre eseguita con molta prudenza.

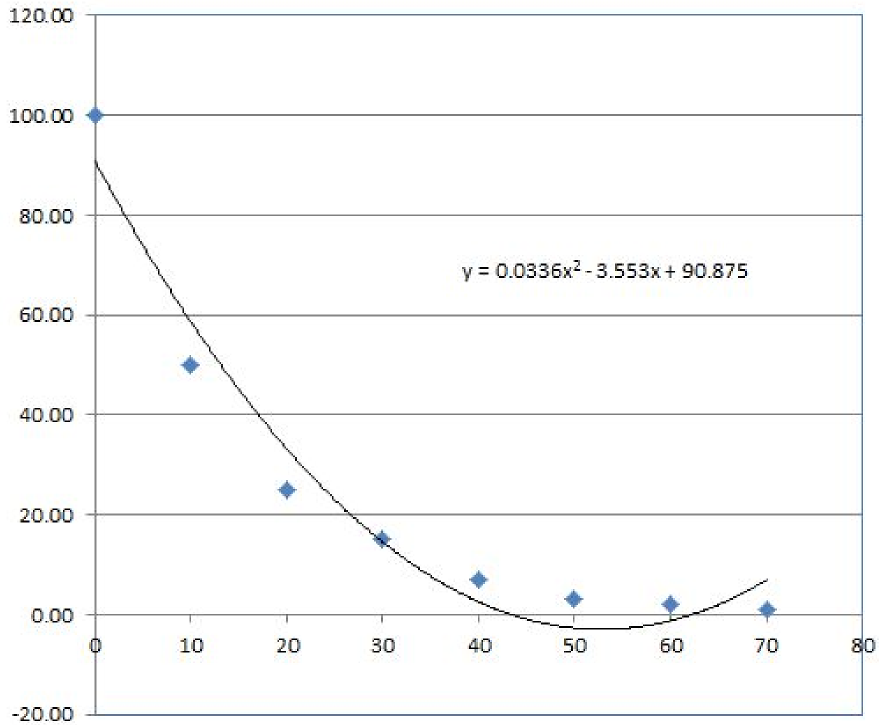


Figura 1: Approssimazione locale di una cinetica di decadimento esponenziale

Linearizzazione della funzione

La funzione di decadimento esponenziale può essere linearizzata prendendo il logaritmo della concentrazione. Infatti:

$$log[C(t)] = log(C_0) - kt + \epsilon$$

Questa soluzione è molto valida, e porta a facile soluzione anche con EXCEL.

Si tratta di una soluzione molto seguita in questo caso, bisogna però fare attenzione che, se gli errori sono normalmente distribuiti e omoscedastici sulla scala originale, la trasformazione logaritmica della Y può portare a deviazioni rispetto alle assunzioni di base per i minimi quadrati. In questo caso non vengono tanto invalidate le stime, quanto piuttosto le inferenze basate sugli anzidetti assunti parametrici.

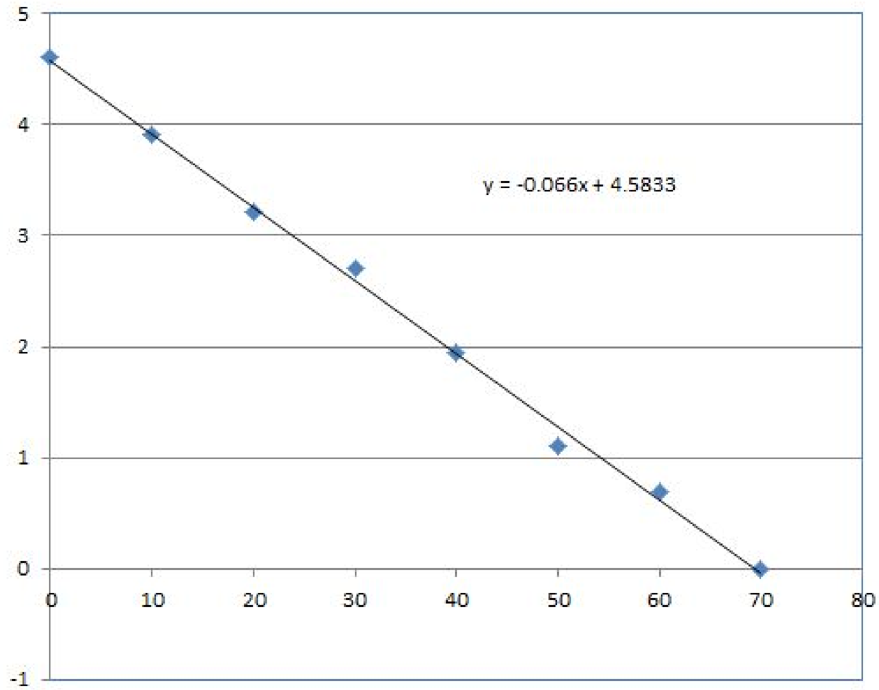


Figura 2: Linearizzazione di una cinetica di decadimento esponenziale, mediante trasformazione logaritmica della concentrazione.

Minimi quadrati 'non-lineari': uso del SOLVER in Excel

L'ampia disponibilità di potenza di calcolo permette di utilizzare metodiche di adattamento diretto di funzioni nonlineari, basate su algoritmi iterativi (Gauss-Newton, Steepest Descent, Marquardt, Simplex, si rinvia per ulteriori informazioni alla bibliografia riportata in calce) .

E' evidente che i metodi iterativi di calcolo sono possibili solo con l'uso di un PC e pertanto assumiamo che lo studente abbia a disposizione un software come EXCEL. Per risolvere questo problema di minimizzazione, dobbiamo prima impostarlo ed è quindi necessario fornire delle stime iniziali dei parametri, che vengono poi corrette dall'algoritmo in ogni iterazione successiva fino ad ottenere la convergenza sui valori che minimizzano la funzione dei minimi quadrati.

Dobbiamo quindi:

1. Calcolare i valori attesi, utilizzando i valori iniziali dei parametri
2. Calcolare i residui ed elevarli al quadrato
3. Sommare i residui al quadrato ed ottenere il valore della devianza del residuo, che è la quantità da minimizzare (obiettivo)

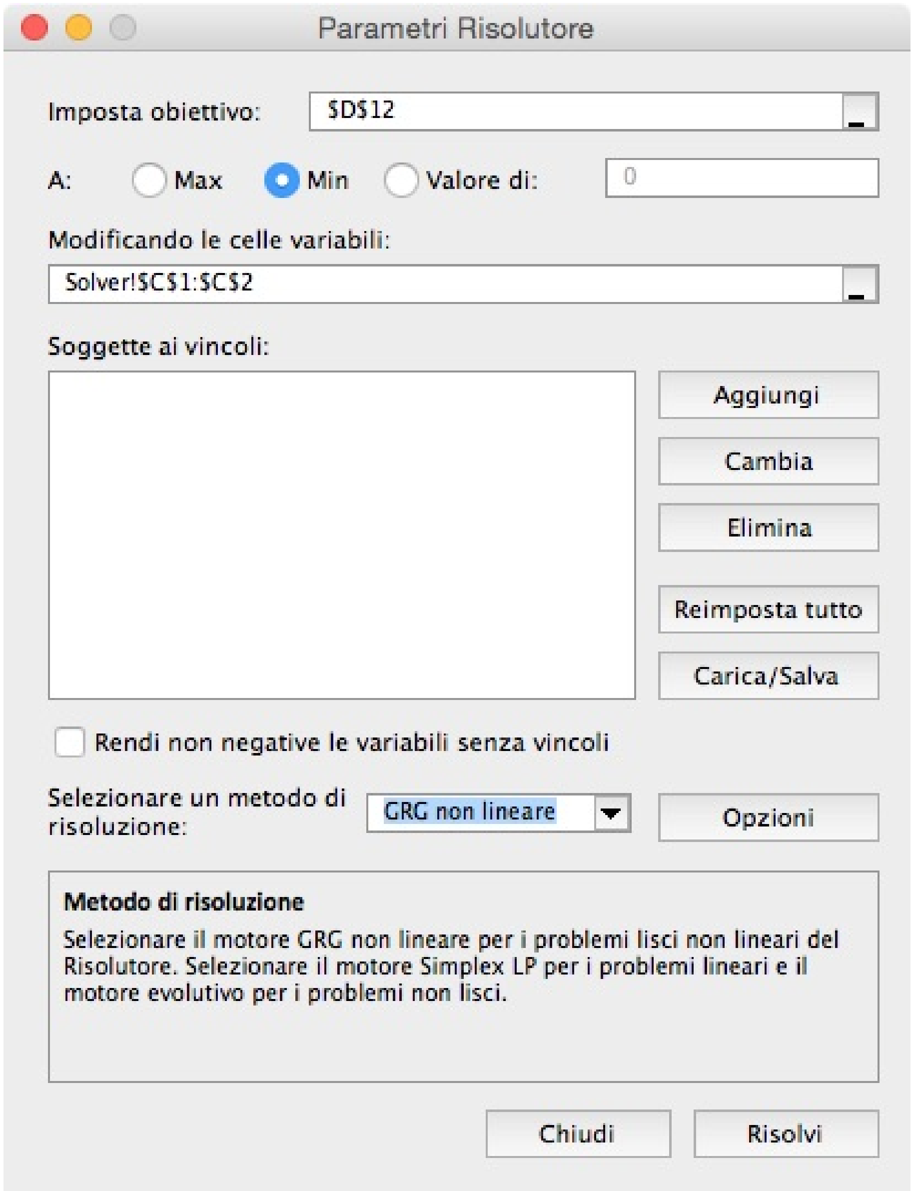
Il problema è impostato nell'immagine sottostante.

	A	B	C	D	E
1		a	100		
2		k	0.07		
3	Time	Conc	Atteso	Residui^2	
4	0	100	100	0	
5	10	50	49.65853	0.1166015	
6	20	25	24.6597	0.1158065	
7	30	15	12.24564	7.5864834	
8	40	7	6.081006	0.8445495	
9	50	3.5	3.019738	0.2306513	
10	60	2	1.499558	0.2504425	
11	70	1	0.744658	0.0651994	
12				9.2097341	**RSS**
13					

Come valori iniziali dei parametri abbiamo fornito 100 (cioè la concentrazione al tempo 0) e 0.07, che può essere ottenuto con il metodo della linearizzazione sopra esposto.

A questo punto possiamo utilizzare il risolutore di Excel (verificando che sia stato prima installato ed attivato tra i componenti aggiuntivi), selezionando la voce corrispondente dal menù strumenti. Il form che si evidenzia ci permette di:

1. Selezionare la cella obiettivo (D12) che deve essere minimizzata;
2. Selezionare le celle che debbono essere variate, per procedere alla minimizzazione (cioè i parametri nell'intervallo C1:C2)
3. Non è necessario in questo caso fare altre scelte, quindi lasciamo tutto il resto inalterato è clickiamo OK.



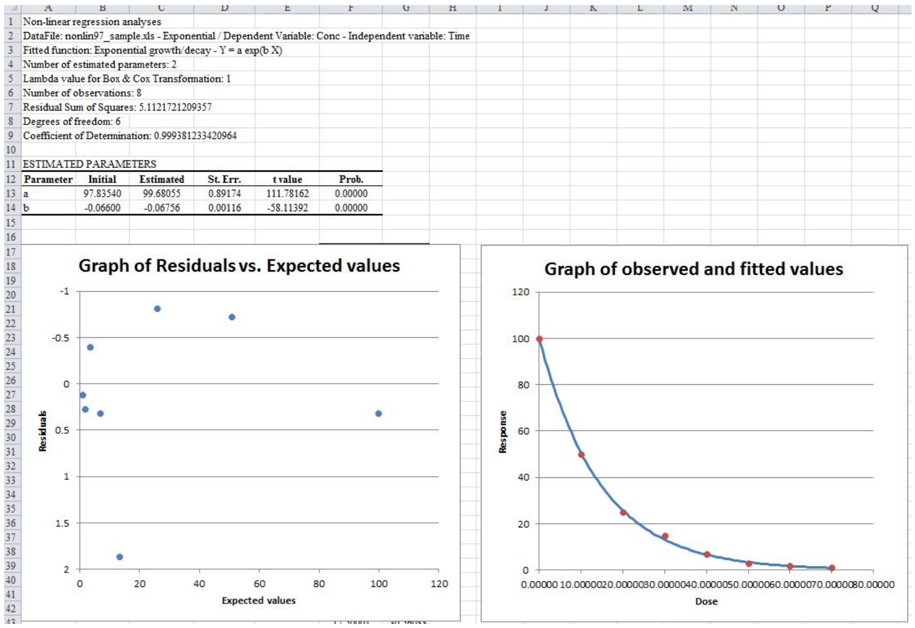
Il problema risolto si trova nella figura sottostante

	A	B	C	D	E
1		a	99.6478		
2		k	0.067428		
3	Time	Conc	Atteso	Residui^2	
4	0	100	99.6478	0.1240466	
5	10	50	50.77292	0.5974073	
6	20	25	25.87001	0.7569176	
7	30	15	13.18139	3.3073604	
8	40	7	6.716229	0.080526	
9	50	3.5	3.422078	0.0060718	
10	60	2	1.74363	0.0657255	
11	70	1	0.888421	0.0124499	
12				4.950505	**RSS**
13					

Uso di NONLIN97

Il solver è uno strumento importante di Excel, che merita di essere conosciuto perché consente di risolvere numerosi problemi di ottimizzazione, oltre alle regressioni non-lineari. Tuttavia il risultato che fornisce non contiene gli errori standard dei parametri, che sono invece molto utili per gli scopi metodologico-sperimentali. Per questo motivo abbiamo implementato una macro Excel (NONLIN97) che è basata sul risolutore e consente di calcolare gli errori standard dei parametri automaticamente. Inoltre, utilizzando questa macro non è necessario preoccuparsi delle stime iniziali, in quanto esse sono ottenute grazie a degli algoritmi di self-starting.

Le stime a cui perveniamo utilizzando NONLIN97 sono le seguenti.



Insomma, il modello parametrizzato è:

$$C(t) = 99.68\, exp(-0.0676t) + \epsilon$$

ed è in grado di descrivere l'andamento della concentrazione nel tempo.

Bontà delle stime

Come in tutti i metodi iterativi, le stime ottenute sono solo approssimate, ma dobbiamo comunque preoccuparci che questa approssimazione sia accettabile.

L'analisi dei residui è evidenziata nella figura precedente e non mostra problemi particolari:

1. il grafico di sinistra mostra una distribuzione random dei residui, senza pattern evidenti
2. il grafico di destra mostra che i valori osservati giacciono vicini al modello, che li rappresenta quindi adeguatamente.

Oltre ai metodi grafici di verifica, dobbiamo tener presenti gli errori standard dei parametri, che non debbono mai essere superiori alla metà del valore del parametro stimato, cosa che in questo caso è vera.

Cosa fare in caso di problemi di stima?

Se ci troviamo in una condizione patologica, dobbiamo prenderne atto ed adottare azioni correttive.

Può captare che:

1. il modello non funziona. DIAGNOSI: il grafico delle stime (quello a destra in alto) mostra che i dati osservati sono lontani dalla curva di regressione. AZIONE CORRETTIVA. Nessuna. Il nostro lavoro è finito e dobbiamo cambiare approccio modellistico.
2. il modello funziona, ma è troppo complesso, rispetto all'informazione contenuta nei dati. DIAGNOSI: le stime dei parametri hanno errori standard molto alti. AZIONI CORRETTIVA: semplificare il modello. Ripetere l'eseperimento. Utilizzare un altro dataset per la parametrizzazione. Utilizzare le stime dei parametri con molta cautela, perchè potrebbero non essere affidabili.
3. Il modello funziona, i parametri sono ben stimati, ma abbiamo problemi con le assunzioni di base dei minimi quadrati. DIAGNOSI: il grafico dei residui (a sinistra in alto) mostra problemi di non-normalità o eteroscedasticità). AZIONE CORRETTIVA. Utilizzare il metodo di Box-Cox.

Per quanto riguarda quest'ultimo aspetto (si veda la lezione relativa alla verifica dell'ANOVA), bisogna tener conto che la trasformazione dei dati comporta anche la modifica della scala sulla quale vengono stimati i parametri, che quindi non conservano il loro valore biologico. Per esempio, se invece che modellizzare il peso fresco, modellizziamo il logaritmo del peso fresco, le stime ottenute per i parametri cambiano la loro unità di misura e quindi potrebbero non essere più valide per i nostri scopi modellistici.

Per questo motivo, nelle regressioni non-lineari si preferisce adottare la cosiddetta tecnica della "trasformazione di entrambe le parti", o metodo TBS ("Transform Both Sides") e trasformare quindi sia i dati osservati per la variabile dipendente, sia il modello:

$$Y^\lambda = f(X)^\lambda$$

In questo modo si ottengono i parametri della funzione sulla scala originale come se la trasformazione non fosse stata eseguita per niente.

Per approfondire

Bates D. M. e Watts D. G., 1988. Nonlinear regression analysis and its applications. John Wiley and Sons, Inc., New York, pp.365\ Bliss C. I., 1967. Statistics in biology. McGraw-Hill Inc., Carrol R. J. e Ruppert D., 1988. Transformation and weighting in regression. Chapman and Hall, London\ Draper N. R. e

Smith H., 1981. Applied regression. John Wiley and Sons, Inc., New York, 2nd ed.\