

Valutazione dei modelli matematici

Andrea Onofri

30 aprile 2013

Indice

1	Introduzione	2
2	Metodi grafici di valutazione	2
3	Metodi numerici	3
3.1	Il coefficiente di determinazione	5
4	Valore predittivo di un modello	6
5	Referenze bibliografiche per approfondimenti	7

1 Introduzione

Valutare un modello significa verificare il suo valore descrittivo e predittivo. Nel primo caso assumiamo che i parametri di un certo modello siano stati stimati adattando il modello stesso (ad esempio con il metodo dei minimi quadrati) ad un certo dataset. Di conseguenza vogliamo verificare che il modello così ottenuto garantisca effettivamente una buona descrizione dei dati di partenza (**bontà di adattamento**), altrimenti le stime dei parametri non hanno alcun valore scientifico. Insomma, il modello non descrive bene il fenomeno di partenza e non ha quindi alcun senso utilizzarlo a scopo previsionale.

Se invece abbiamo un modello già parametrizzato (con esperimenti, dati di letteratura o esperienza locale), possiamo preoccuparci di verificare la sua capacità predittiva, confrontando il suo output con altri dati sperimentali indipendenti da quelli che abbiamo utilizzato per parametrizzarlo.

In ogni caso, dovrebbe essere chiaro che la più semplice valutazione di un modello viene effettuata per confronto tra previsioni e dati osservati. A titolo di esempio, continueremo ad utilizzare il modello relativo alla cinetica di degradazione del primo ordine (decadimento esponenziale), già utilizzato nella lezione precedente. In questo caso, gli 'ingredienti' per la valutazione sono nella tabella seguente.

Time	Expected	Observed	Residuals
0	99.680487	100	0.319512841
10	50.724018	50	-0.72401819
20	25.811715	25	-0.8117146
30	13.134697	15	1.865302976
40	6.6837972	7	0.316202796
50	3.4011554	3	-0.40115535
60	1.7307314	2	0.269268596
70	0.8807099	1	0.119290098

2 Metodi grafici di valutazione

Immaginiamo di avere dei dati relativi alla crescita della biomassa e del LAI di una coltura nel tempo. Plottare questi dati vicino alle previsioni di un modello di simulazione consente di aver un ottimo colpo d'occhio sulla capacità descrittiva e predittiva (in questo caso solo se i dati sono 'indipendenti') dal modello.

Il grafico in figura 1 si può utilizzare solo per modelli con una sola variabile esplicativa. Per modelli più complessi si consiglia un grafico che riporti i valori osservati sull'asse delle ascisse e i valori attesi sull'asse delle ordinate. In questo grafico un modello perfetto giace esattamente sulla diagonale.

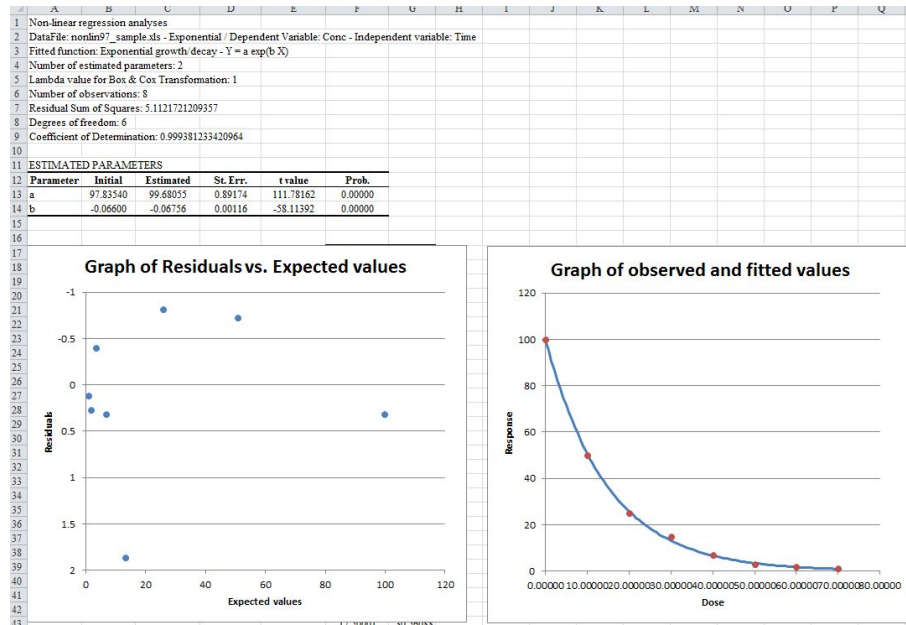


Figura 1: Valutazione di un modello di crescita di una coltura

3 Metodi numerici

I metodi numerici sono realizzati attraverso indicatori che quantificano lo scostamento tra i dati osservati e le simulazioni del modello. Tra questi indichiamo:

$$Bias = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)$$

ove Y e \hat{Y} sono gli N valori osservati e attesi. Il Bias è 0 per modelli perfetti, è positivo per modelli che sottostimano e negativo per modelli che sovrastimano. ATTENZIONE! il bias può anche essere 0 se gli errori sono molto alti ma di segno opposto e tali da compensarsi a vicenda. Considerando i dati in tabella, il bias è pari a 0.119 ed è quindi piuttosto basso.

Un altro indicatore molto utilizzato è:

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

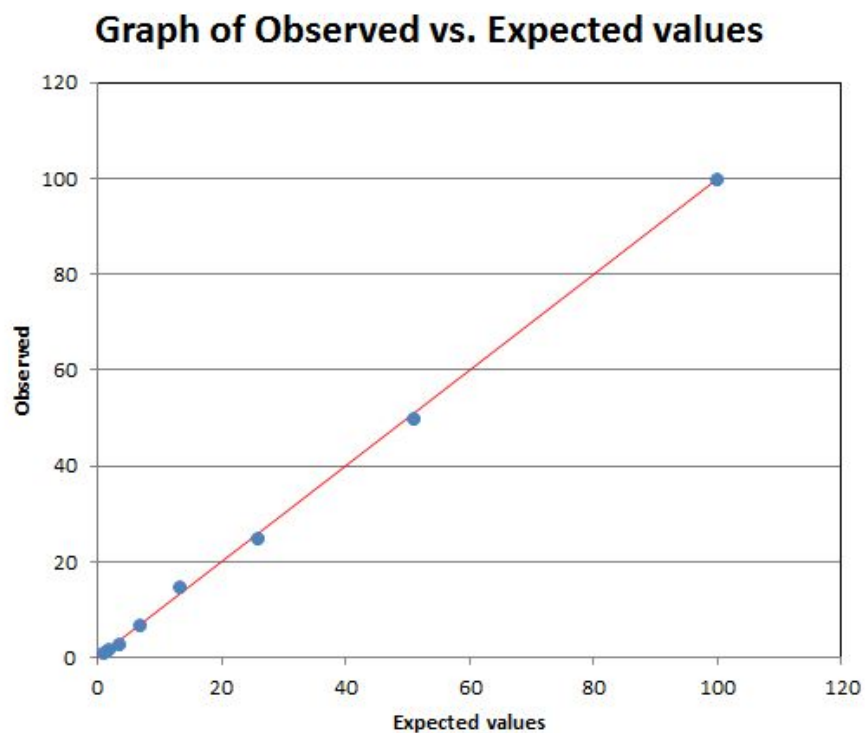


Figura 2: Altro esempio di valutazione grafica di un modello

A differenza del bias, MSE (Mean Square Error) è insensibile alla compensazione degli errori di segno opposto. Nel caso in esempio, il MSE è pari a 0.639.

Il MSE è spesso difficile da valutare perchè, essendo una somma di quadrati, la sua unità di misura non è quella dei dati. Per questo motivo, spesso si utilizza la sua radice quadrata:

$$RMSE = \sqrt{MSE}$$

che ha la stessa unità di misura delle osservazioni. In questo caso, il RRMSE (Root Mean Square Error) è uguale a 0.799, che, considerando i dati rappresenta un valore decisamente basso.

Una variante molto utilizzata è il Relative Root Mean Squared Error (RRMSE):

$$RRMSE = \frac{\sqrt{MSE}}{\bar{Y}} \times 100$$

dove \bar{Y} è la media dei dati. Si tratta di un indicatore analogo al coefficiente di variabilità, nel quale la bontà del modello viene espressa relativamente alla media delle previsioni. Nel nostro caso è pari al 3.16 %.

Un altro indicatore 'relativo' è:

$$EF = 1 - \frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2}$$

che pone a confronto il MSE e la devianza delle osservazioni. In pratica si confronta lo scostamento tra modello e osservazioni con lo scostamento tra la media e le stesse osservazioni, il che significa che al denominatore l'output del modello è sostituito con un valore costante e pari alla media delle osservazioni. Se il modello è perfetto EI è uguale ad 1 (osservato e simulato coincidono), mentre se EI è uguale a 0 il modello ha un valore predittivo molto scarso, pari alla semplice media delle osservazioni. EI può anche assumere valore negativo, se il modello è peggiore del modello della media. In questo caso, l'EF è uguale a 0.999, il che conferma che si tratta di un'ottimo modello.

3.1 Il coefficiente di determinazione

Una quantità frequentemente utilizzata per valutare la bontà d'adattamento di un modello di regressione (valutazione puramente descrittiva, quindi) è il coefficiente di determinazione (R^2), dato dal rapporto tra la devianza della regressione e la devianza totale.

$$R^2 = \frac{SS_{reg}}{SS_{tot}}$$

Pur essendo utilizzato in modo pressochè ubiquitario, il coefficiente di determinazione fornisce solo un'indicazione abbastanza grezza sulla bontà del modello. Infatti:

- valori bassi possono essere ottenuti non solo perchè la devianza della regressione è bassa, ma anche perchè la devianza totale delle osservazioni è alta (molti dati e molto variabili).
- L' R^2 dipende dall'intervallo di variazione della variabile indipendente; di conseguenza, se nella regressione aggiungiamo uno o più livelli della X otteniamo un innalzamento del valore di R^2 , che tuttavia non necessariamente produce un miglior modello. Un esempio tipico è l'aggiunta del livello di fertilizzazione zero in una prova di concimazione: ciò aiuta ad alzare il valore di R^2 , ma non necessariamente implica una miglior stima del modello di regressione, soprattutto in vicinanza dei livelli ottimali.

Ciò è particolarmente importante con le regressioni non-lineari, laddove vi sono molti parametri di forma della curva. Ad esempio, se il valore di N ottimale è compreso tra 5 e 40 kg/ha, un esperimento con 0, 20, 40 e 80 kg/ha di N darà sicuramente valori di R^2 più alti che non un esperimento con 5, 10, 20 e 40 kg/ha anche se quest'ultimo fornirà più probabilmente una miglior stima dei livelli di N ottimale.

- Il coefficiente di determinazione è sensibile al numero di variabili esplicative presenti nel modello, e quindi non premia i modelli più semplici (rasoio di Occam).
- Il coefficiente di determinazione è sensibile al numero di parametri presenti nel modello. Modelli con molti parametri danno sempre valori di R^2 alti, ma non rispettano spesso le caratteristiche di semplicità e senso biologico che sono invece richieste ad un buon modello.

Per evitare almeno gli ultimi due problemi, viene proposto il coefficiente di determinazione corretto, dato dalla proporzione di varianza (MS) spiegata dalla regressione:

$$R_a^2 = 1 - \frac{MS_{residuo}}{MS_{tot}}$$

Dato che dalle devianze si passa alle varianze, il valore di R^2 corretto è detto anche ' R^2 corretto per i gradi di libertà'. Il suo rapporto con il coefficiente di determinazione tradizionale è:

$$R_a^2 = 1 - \frac{(1 - R^2)(n - 1)}{(n - k - 1)}$$

dove n è il numero di osservazioni e k il numero dei regressori. L' R^2 corretto è sempre più basso dell' R^2 e diminuisce con l'aggiunta al modello di un nuovo regressore se l'incremento di devianza totale è meno che quello della devianza residua. Può assumere valori negativi se la varianza del residuo è maggiore della varianza della variabile dipendente.

4 Valore predittivo di un modello

Nella sezione precedente abbiamo illustrato metodiche che confrontano gli output del modello con osservazioni sperimentali passate. Il nostro vero interesse è però capire quanto bene il modello prevede il futuro.

A questo fine deve realizzarsi la piena indipendenza tra modello e dati, che possono anche essere rilevati nel passato, ma non debbono essere stati utilizzati per costruire o parametrizzare il modello stesso.

Oltre all'indipendenza, i dati debbono anche essere rappresentativi dell'intero range di situazioni di interesse. In altre parole, un modello che permette di prevedere bene un gruppo di osservazioni sperimentali indipendenti è validato solo per il set di condizioni ambientali nelle quali il dataset è stato ottenuto.

Un esempio relativo alla valutazione del potere descrittivo e predittivo di un modello viene dato nel file 'tomato.xls' disponibile nel sito del docente (www.casaonofri.it/Biometry/).

5 Referenze bibliografiche per approfondimen- ti

Loague, K. & Green, R. Statistical and graphical methods for evaluating solute transport models: overview and applications. *Journal of contaminant hydrology*, 1991, 7, 51-73