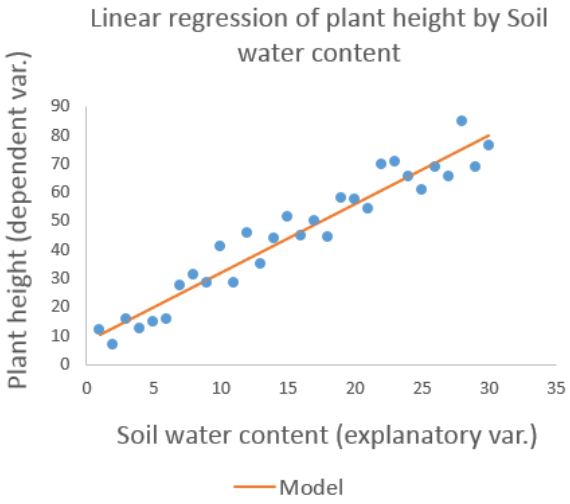
GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

Cos'è la modellazione statistica?

http://help.xlstat.com/customer/en/portal/articles/2062460-what-is-statistical-modeling-?b_id=9283
Versione italiana a cura di Luca Biasibetti.

La ***modellazione statistica*** è una metodologia semplificata e matematicamente formalizzata per approssimare la realtà e fare previsioni a partire da queste approssimazioni. Il modello statistico è sempre caratterizzato da una o più equazioni matematiche che rappresentano il modello stesso.
Per meglio comprendere quanto appena descritto, partiamo da qualche esempio.
Supponiamo di voler produrre un report relativo al peso di una varietà di patate considerando due diverse metodologie, una più difficile e un'altra più semplice. La più complessa consiste nello spendere anni a misurare il peso di ogni singola patata di questa varietà nel mondo e riportare i dati in uno sconfinato foglio di calcolo Excel. La più semplice invece consiste nel selezionare un campione rappresentativo di 30 patate di questa varietà, calcolandone il peso medio, la deviazione standard e riportando solo questi due "numeri" come descrizione approssimativa del campione. Rappresentare la caratteristica di un campione in termini di media e deviazione standard è una forma molto semplice di modellizzazione statistica.
Un altro esempio può essere l'analisi dell'altezza di alcune piante in base al contenuto idrico del suolo. Tale analisi può essere formalizzata e rappresentata mediante una linea retta ricavata dopo un esperimento su un campione di piante sottoposte ad un umidità crescente del suolo. Questo particolare modello è chiamato **regressione lineare semplice**.



Variabili dipendenti e variabili indipendenti (o esplicative)

In quasi tutti i casi, i modelli statistici possiedono variabili indipendenti (o esplicative) e dipendenti. La ***variabile dipendente*** è la variabile che vogliamo descrivere, spiegare e prevedere. In generale, la variabile dipendente è quella che rappresentiamo sull'asse delle ordinate (Y) nei grafici cartesiani che caratterizzano l'equazione del modello. Nell'ultimo esempio riportato, la variabile dipendente è l'altezza della pianta.

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

Le ***variabili indipendenti***, dette anche variabili ***esplicative***, sono quelle utilizzate per spiegare, descrivere o prevedere le variabili dipendenti. Le variabili esplicative sono spesso rappresentate sull'asse delle ascisse (X). L'esempio sull'altezza delle piante possiede una sola variabile esplicativa (ed anche quantitativa): il contenuto idrico del suolo.
Sia le variabili dipendenti che quelle esplicative possono essere singole o multiple, quantitative o qualitative; esistono modelli adattati per ogni diversa situazione.

Parametri del modello

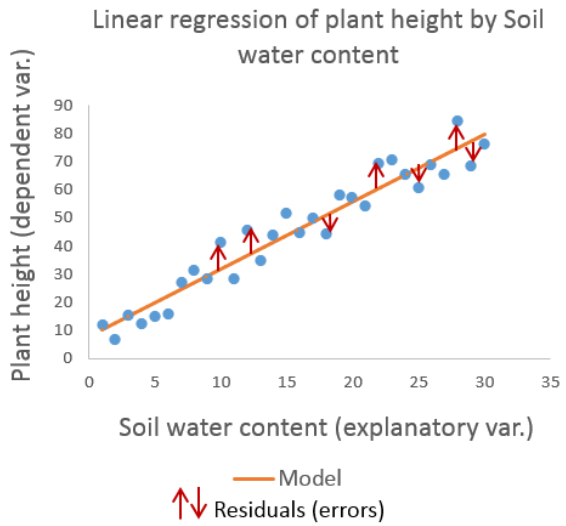
Nei modelli parametrici classici, le variabili dipendenti sono messe in relazione a quelle esplicative attraverso un'equazione matematica che rappresenta e contiene i ***parametri del modello***. Nell'esempio di regressione lineare semplice analizzato, i parametri sono l'intercetta e la pendenza della retta. L'equazione può essere scritta nel modo seguente:

*Altezza della pianta = intercetta + pendenza * contenuto idrico del suolo*

I calcoli sottostanti alla modellazione statistica non consentono solo la stima dei parametri del modello ma anche ulteriori previsioni della variabile dipendente.
Per completezza è corretto ricordare che la regressione lineare semplice comporta anche un terzo parametro: la varianza dei residui.

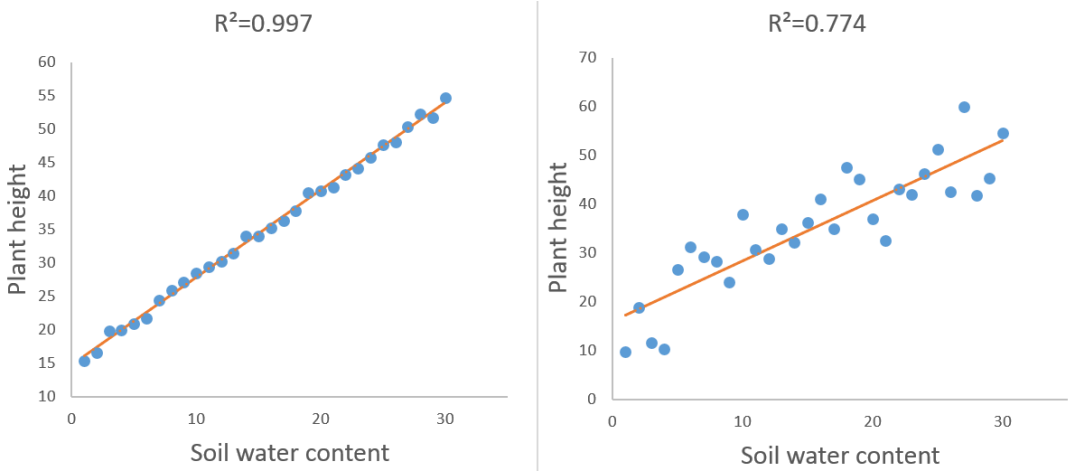
Residui

I ***residui*** del modello (o errori) sono le distanze tra i punti che rappresentano i dati e il modello stesso (rappresentato, nell'esempio di regressione lineare dell'altezza delle piante, dalla linea retta).



I residui rappresentano la parte di variabilità dei dati che il modello non è stato in grado di descrivere. La statistica R^2 invece è la parte di variabilità spiegata dal modello pertanto più bassi sono i residui, maggiore è la statistica R^2.

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it



Quale modello statistico dovremmo scegliere?

La scelta del modello statistico opportuno per ogni situazione non è un tema banale. Non è infatti del tutto corretto pensare che ogni set di dati abbia un proprio modello che si adatti alla perfezione.
In base alle singole esigenze, ogni strumento di modellizzazione risponde a domande specifiche. Ad esempio, la correlazione tra la glicemia e un determinato tipo di diabete può essere spiegata da una variabile qualitativa (il sesso per esempio); in questa situazione, è possibile utilizzare il modello ANOVA. Utilizzando lo stesso set, possiamo anche considerare i dati sull'età (variabile quantitativa) per verificare se è presente un trend lineare crescente o decrescente della glicemia in base all'età dei pazienti; in questa situazione useremmo la regressione lineare.
La scelta di un modello statistico può anche essere guidata dalla "forma" del grafico che rappresenta le relazioni tra le variabili dipendenti ed esplicative. Ad esempio, nel caso in cui queste "forme" siano delle curve, i modelli polinomiali o non lineari sono più indicati rispetto a quelli lineari.
Se lo scopo dello studio consiste unicamente nel fare previsioni partendo da un ampio insieme di variabili, allora si possono prendere in considerazione soluzioni diverse dai modelli parametrici. La regressione ai minimi quadrati parziali, ad esempio, è uno strumento specifico adottato per prevedere una variabile dipendente da un numero illimitato di variabili esplicative (eventualmente correlate). L'uso della regressione ai minimi quadrati parziali è molto utilizzata ad esempio nella chemiometria, in cui gli output sono spesso predetti da un ampio spettro di lunghezze d'onda.

Quanti parametri dovrebbero essere inclusi nel modello?

Una volta scelto il modello appropriato, in molte situazioni, ci si pone la domanda in testa al paragrafo: "Quanti parametri dovrei includere nel modello?"
Maggiore è il numero di parametri inclusi, migliore è l'adattamento del modello ai dati (ossia residui più bassi che implicano una più elevata statistica R^2). Quindi il numero di parametri nel modello dovrebbe essere massimizzato in modo tale che i residui siano estremamente ridotti al minimo? Non proprio. Un modello che si adatta troppo ai dati sarà troppo rappresentativo del

particolare campione utilizzato e la generalizzazione dell'intera popolazione sarà quindi meno accurata.
La qualità del modello, misurata come equilibrio tra un buon adattamento dei dati e un numero minimo di parametri, può essere valutata utilizzando alcuni indici come il Criterio di Informazione di Akaike (AIC) o il Criterio di Informazione Bayesiano (BIC o SBC). Confrontando l'uno con l'altro diversi modelli parametrici, il modello con indice più basso è caratterizzato da una migliore qualità. E' importante notare però che l'interpretazione di questi indici non risulta essere sensata in un contesto assoluto (ossia quando viene preso in considerazione un solo modello).

Scelta del modello

Al link riportato in calce è collegata una griglia utile per la scelta dei modelli. Tale griglia è suddivisa in base al tipo e al numero di variabili dipendenti e indipendenti. Vengono anche proposte soluzioni diverse dai modelli parametrici. I modelli visualizzati sono quelli comunemente utilizzati nelle statistiche e sono tutti disponibili in XLSTAT.
Prima di passare alla griglia, ricordiamo le condizioni di validità per i modelli parametrici:

1. dati indipendenti;
2. varianza omogenea;
3. residui normalmente distribuiti;
4. set di dati composto da almeno 20 elementi (consigliato);
5. assenza di multicollinearità (essenziale per la stima dei parametri del modello);
6. numero di dati superiore al numero delle variabili esplicative;
7. normalità multivariata dei residui;
8. varianza omogenea per ogni variabile dipendente; correlazioni tra variabili dipendenti omogenee.

Griglia - scelta del modello