Antonio Manno (mannoanto@libero.it, www.statistica.too.it). Palermo, Italia

Modellazione statistica a fini previsivi

1 Introduzione

Data una variabile casuale *y*, di qualsiasi natura (ossia sia di tipo qualitativo che quantitativo), uno dei principali obiettivi di un ricercatore è quello di poter prevedere i valori che essa potrebbe assumere in un arco temporale o in un contesto spaziale diverso da quello che è stato osservato attraverso l'estrazione di un campione.

Se è stata osservata una serie di *n* osservazioni, $(y_1, y_2, \ldots y_i, \ldots, y_{n,})$, y_i può indicare il valore che la variabile *y* assume al tempo *i* ed in questo caso si parla di analisi temporale o di serie storiche, oppure il valore assunto dall'*i*-esima unità osservata, ed in questo caso si parla di analisi di tipo cross-section; è pure possibile considerare un campione di osservazioni sulle stesse *n* unità campionarie per *T* periodi temporali, in tal caso si avranno le osservazioni $\{y_{ij}\}$, con *i=1,2,…,n* e *j=1,2,…T*, dove l'indice *i* indica l'unità campionaria di osservazione e l'indice *j* indica il periodo temporale di osservazioni: si parla in questo caso di analisi longitudinale.

Data la serie $\{y_i\}_{i=1}^{n}$, temporale o di tipo cross-section, il ricercatore vuole quindi essere in grado di poter valutare *y* in un'unità (temporale o spaziale) diversa da quelle osservate.

Per raggiungere questo obiettivo è necessario costruire un modello statistico – matematico, nel quale sia espressa la relazione che lega la variabile d'interesse *y* ad un insieme di altre variabili, dette esplicative o predittori. Un modello in cui una variabile è espressa in funzione di altre, ossia in cui un insieme di variabili esplicative influenzano la

determinazione della variabile d'interesse, per cui da un punto di vista logico-deduttivo tale variabile è "conseguente" alle altre, è detto modello di tipo regressivo; se invece nel modello non c'è un insieme di variabili che "spiega" un'altra variabile, ma tutte le variabili si influenzano fra di loro, allora il modello è detto correlativo. Per riuscire a fare previsione su una singola variabile bisogna considerare un modello di tipo regressivo.

In un tale modello, le variabili esplicative, indicate con il vettore *d*-dimensionale $\underline{x}$ ossia $\underline{x} = (x_1, x_2, \ldots, x_d)^t$, non determinano la variabile *y*, ma la influenzano; ossia ad un valore ben preciso del vettore $\underline{x}$ non è associato un unico valore della variabile *y*, in modo deterministico: si parla infatti di modellazione stocastica e non deterministica. La modellazione deterministica è una modellazione che va bene nelle scienze esatte, ma non è applicabile nelle scienze socio-economiche, caratterizzate da fenomeni altamente variabili ed intrinsecamente stocastici.

Un modello regressivo di tipo stocastico tenta di modellare in maniera "sistematica" il valore atteso della variabile *y* condizionatamente al vettore di esplicative $\underline{x}$, ossia si cerca di trovare una funzione *f* che soddisfi la seguente relazione:

$$E(y|\underline{x}) = f(\underline{x}).$$

La funzione *f* che lega il vettore $\underline{x}$ e la variabile *y*, non è però nota al ricercatore (tranne casi particolari), per cui ciò che il ricercatore deve fare è determinare una funzione $\hat{f}$ che approssimi bene la funzione ignota *f*.

Per fare ciò si possono due seguire due approcci diversi; un primo approccio è quello parametrico, nel quale il ricercatore definisce a priori una forma per la funzione *f*, che dipende da un vettore di parametri $\underline{\theta}$ (oltre che da $\underline{x}$) , per cui la relazione precedente viene riscritta nel seguente modo:

$$E[y|\underline{x}] = f(\underline{x};\underline{\theta})$$

con *f* completamente esplicitata e $\underline{\theta}$ vettore di parametri che dovrà essere stimato per mezzo di un campione di osservazioni $\{(\underline{x}_i, y_i)\}_{i=1}^{n}$; l'altro approccio, che viene definito non parametrico, non fa nessuna assunzione a priori sulla funzione *f* e non considera nessun tipo di parametri $\underline{\theta}$ da stimare, ma stima direttamente la funzione *f* attraverso una media ponderata locale delle *y* presenti in un intorno multidimensionale di $\underline{x}$; alcuni approcci tentano di combinare i due approcci considerando la funzione *f* come somma di due funzioni: una definita a priori e specificata da un vettore di parametri e l'altra non specificata, ma stimata attraverso medie ponderate locali.

Considerando un approccio di tipo parametrico nel quale, come già affermato, occorre definire a priori una funzione ben specificata *f*, la metrica della variabile *y* gioca un ruolo chiave nella scelta del tipo di funzione *f* da considerare. Ciò perché il codominio della funzione *f*, relativamente all'insieme *D* ossia lo spazio dei predittori $\underline{x}$, deve coincidere con lo spazio a cui appartiene il valore atteso $E[y|\underline{x}]$; si consideri, ad esempio, una variabile *y* di tipo bernoulliano, per cui il suo valore atteso condizionato $E[y|\underline{x}]$, che indica la probabilità di successo condizionatamente a $\underline{x}$, giace nell'intervallo (*0,1*) ed un vettore di esplicative $\underline{x}$ tale che $\underline{x} \in D \subset \Re^d$, in questo caso bisogna scegliere una funzione *f* il cui codominio, relativamente all'insieme *D*, sia l'intervallo (*0,1*), per cui appare ovvio che una funzione di tipo lineare $f(\underline{x}) = \underline{x}^t \underline{\theta}$ non è ammissibile, in quanto il codominio di *f* relativamente a *D* è l'insieme dei numeri reali $\Re$, ossia $f(D) = \Re \neq (0,1)$; in una simile situazione si scelgono, ad esempio, funzioni di tipo logistico $f(\underline{x};\underline{\theta}) = \frac{e^{\underline{x}^t \underline{\theta}}}{1 + e^{\underline{x}^t \underline{\theta}}}$ che soddisfano la condizione richiesta.

In generale occorre distinguere modelli relativi a variabili di tipo quantitativo, da modelli per variabili qualitative: nel primo caso si parla di regressione, mentre nel secondo caso si parla di classificazione.

2 **Regressione**

Nei modelli di tipo regressivo, alla componente deterministica del modello, ossia $E[y|\underline{x}]= f(\underline{x})$, trattandosi di modellazione di tipo stocastico, si aggiunge una componente casuale, per cui il modello viene riscritto in termini di *y* e si considera:

$$y = f(\underline{x},\varepsilon);$$

impiegando un approccio di tipo parametrico, occorre aggiungere il vettore dei parametri $\underline{\theta}$ come argomento della funzione f ; usualmente si ipotizza che l'influenza della variabile casuale ε sulla variabile d'interesse *y* sia di tipo additivo, per cui, aggiungendo il vettore dei parametri come argomento, il modello viene scritto nel seguente modo:

$$y = f(\underline{x};\underline{\theta})+\varepsilon\,.$$

Si noti che ciò che è importante è la metrica della variabile *y*, mentre la metrica della variabile $\underline{x}$, che comunque ha una sua importanza, non entra direttamente nella scelta del tipo di funzione *f*. Infatti, è sempre possibile trasformare una variabile qualitativa in una variabile di tipo numerico, attraverso l'utilizzo di indicatori, senza modificare l'informazione presente nei dati. Si ipotizzi ad esempio che la variabile *x* (con *d=1*, per ipotesi) sia qualitativa e che abbia 3 possibili modalità, per trasformare tale variabile in numerica occorre considerare un vettore tridimensionale nel quale ogni singola componente è associata ad una modalità qualitativa, che avrà (relativamente ad ogni osservazione) un valore pari ad uno per la

componente associata alla modalità osservata e zero per le altre due modalità; per cui se un'unità campionaria presenta un valore pari alla seconda modalità qualitativa, allora il vettore numerico associato sarà (0,1,0).

Con tale operazione è conservata tutta l'informazione presente nei dati: la possibilità di poter trattare come numerica una variabile qualitativa, ha l'unico svantaggio di "complicare" la trattazione del modello, ma data la potenza degli attuali elaboratori ciò non costituisce più un problema.

La trasformazione di una variabile quantitativa in qualitativa è pure possibile, attraverso una categorizzazione in classi, ma in questo caso si ha una notevole perdita d'informazione, per cui non è consigliabile effettuare tale trasformazione, tranne nei casi in cui non interessa tutta l'informazione presente nella variabile, ma soltanto un'informazione più ridotta.

Seguendo un approccio di tipo parametrico, occorre definire la forma funzionale della funzione *f* e fare una serie di assunzioni plausibili, più o meno forti, sulla variabile casuale ε.

Definito completamente il modello e le ipotesi sottostanti, si passa alla stima del modello, che in un approccio parametrico corrisponde a stimare il vettore parametrico $\underline{\theta}$; i metodi di stima sono diversi e dipendono dalle informazioni disponibili e dalle ipotesi fatte sull'errore casuale ε. I principali metodi di stima sono:

- della massima verosimiglianza, qualora si ritenga di conoscere la distribuzione del campione di osservazioni $\underline{y} = (y_1, y_2, \ldots, y_n)^t$, ossia $p(\underline{y}; \underline{\theta})$;
- dei minimi quadrati ordinari, se si ipotizza che il vettore di errori $\underline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^t$ abbia una matrice di varianza e covarianze *V* del tipo $V = \sigma^2 \cdot I$ (dove *I* è la matrice identità), ossia gli errori siano

omoschedastici (cioè abbiano la stessa varianza $\sigma_i^2 = \sigma^2 \ \ \forall i = 1,2,\ldots,n$) e non correlati $\left(Cov(\varepsilon_i, \varepsilon_j) = 0 \ \ \forall i \neq j \right)$;

- dei minimi quadrati ponderati, se il vettore degli errori $\underline{\varepsilon}$ ha una matrice di varianze e covarianze *V* diagonale, del tipo:

$$V = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \sigma_n^2 \end{pmatrix}$$

ossia gli errori sono eteroschedastici (cioè con varianze diverse), ma incorrelati;

- dei minimi quadrati generalizzati, per *V* qualsiasi.

Stimati i parametri del modello ed ottenuto il vettore di stimatori $\underline{\hat{\theta}}$, occorre verificare se il modello stimato $f(\underline{x}; \underline{\hat{\theta}})$ fornisce risultati soddisfacenti; per tale ragione si implementano una serie di test statistici per verificare che le assunzioni fatte siano plausibili con i dati e quindi per verificare la plausibilità del modello stimato. Uno strumento indispensabile per verificare l'attendibilità di un modello è l'analisi dei residui, dove i residui $\{r_i\}_{i=1}^n$ sono definiti come la differenza fra valori *y* osservati e calcolati per mezzo del modello ossia $r_i = (y_i - f(\underline{x}_i; \underline{\hat{\theta}}))$.

Dopo aver verificato la validità del modello, lo stesso può essere utilizzato per vari scopi, i principali sono:

- descrittivi;
- previsivi;
- simulativi.

Il modello consente, quindi, di poter fornire una spiegazione logica ai fenomeni e quindi di descriverne la dinamica. Per tale scopo è però necessario implementare un modello di facile lettura, che consenta quindi

di capire con chiarezza il ruolo che ogni variabile esplicativa gioca nell'influenzare le determinazioni di *y*, ciò corrisponde a definire una funzione *f* non troppo complessa e tale che consenta di attribuire un significato specifico ai parametri $\underline{\theta}$ che la specificano.

Se invece l'utilizzo principale del modello è per scopi previsivi, allora non importa assolutamente il tipo di funzione *f* che lega $\underline{x}$ ed *y*, ma interessa costruire semplicemente un meccanismo che dato un insieme di input sia in grado di fornire una previsione accurata dell'output associato. Per tale motivo, in un contesto previsivo, stanno avendo una continua diffusione modelli definiti "black box", nei quali il ricercatore non conosce esattamente qual è la funzione matematica che lega input ed output; il modello definisce un meccanismo attraverso il quale tali variabili sono legate, attraverso una serie di osservazioni, in maniera adattiva. Per evitare che tali modelli si adattino completamente ai dati osservati, cioè che si crei una funzione per cui i residui di tale modello siano tutti nulli, vengono considerati termini di penalizzazione alla funzione d'errore da minimizzare per stimare il modello, la quale misura la distanza fra dati osservati e dati calcolati mediante il modello da adattare. Un esempio di modello a scatola nera è quello degli algoritmi a rete neurale, dove input ed output sono legati da una funzione complessa, che il ricercatore non conosce esattamente.

Lo scopo della previsione, come già accennato all'inizio del lavoro, è quello di conoscere il valore di *y* in unità (temporali o spaziali) non osservate; per cui se il modello adattato per l'*i*-esima unità è $\hat{y}_i = f\left(\underline{x}_i ; \underline{\hat{\theta}}\right) + \varepsilon_i$, il ricercatore potrà prevedere il valore di *y* in corrispondenza di una certa configurazione del vettore di variabili esplicative.

In un modello di serie storiche con *d* ritardi, cioè in cui le esplicative siano valori antecedenti della variabile d'interesse, ossia $\underline{x} = (y_{t-1}, y_{t-2}, \ldots, y_{t-d})^t$ e le *n* osservazioni sono riferite ad *n* osservazioni in periodi diversi, allora il ricercatore potrebbe voler prevedere il valore y_{n+1}, ossia il valore successivo all'ultima osservazione temporale, e impiegando il modello adattato prevedrà y_{n+1} nel seguente modo:

$$\hat{y}_{n+1} = f\left(y_n, y_{n-1}, \ldots, y_{n-d+1}; \underline{\hat{\theta}}\right).$$

Facendo assunzione più forti sugli errori $\{\varepsilon_i\}$, sarà possibile ricavare un intervallo di previsione per *y*, ma affinché ciò sia possibile è necessario conoscere la distribuzione di probabilità della variabile casuale $\hat{y}$, che dipenderà dalla distribuzione congiunta degli errori, dalla distribuzione congiunta degli stimatori e dalla funzione *f* che esprime la dipendenza.

A questo punto sorge un grosso problema: per scopi previsivi si utilizzano spesso funzioni molto complesse, che spieghino bene il comportamento di *y* in funzione di $\underline{x}$, ma se *f* è molto complessa allora è praticamente impossibile ricavare la distribuzione esatta o asintotica della variabile aleatoria $\hat{y}$, per cui sembrerebbe impossibile riuscire a fornire un intervallo di previsione (*a*,*b*) al 95%, cioè tale che sia:

$$\Pr\{a \le \hat{y} \le b\} = 0{,}95.$$

Per risolvere tale problema si possono impiegare tecniche simulative di tipo Monte Carlo, che ovviamente non ci danno una risposta esatta al problema, ma forniscono risposte molto soddisfacenti. Per implementare tali procedure è necessario disporre di un potente elaboratore e quindi simulare la generazione di pseudo-campioni dalle distribuzioni di probabilità considerate; più precisamente, la procedura da seguire può essere la seguente:

- si genera un elevato numero di campioni di errori accidentali $\{\varepsilon_{ij}\}$, provenienti dalla distribuzione teorica ipotizzata $p(\varepsilon)$, con l'indice *i* che indica l'unità "campionaria" e l'indice *j* che indica il campione estratto;
- si fissa il vettore parametrico $\underline{\theta}$ (la procedura dovrà essere effettuata per varie configurazioni di tale parametro);
- si generano campioni $\{y_{ij}\}$ dalla distribuzione teorica $p(\underline{y};\underline{\theta})$;
- si generano campioni per le variabili esplicative, oppure si fissano configurazioni a priori del tipo $\{\underline{x}_{ij}\}$;
- per ogni campione *j*, si stima il vettore dei parametri attraverso il metodo di stima scelto e si avrà $\underline{\hat{\theta}}_j$;
- si calcolano i valori teorici per ogni campione e per ogni unità, ossia:

$$\hat{y}_{ij} = f\left(\underline{x}_{ij};\underline{\hat{\theta}}_j\right) + \varepsilon_{ij}$$

- si determina la distribuzione di probabilità empirica di $\hat{y}$ che rappresenta una stima della vera distribuzione di probabilità; per cui è possibile stimare un intervallo di previsione sulla base di tale funzione.

È comprensibile che generando un numero elevato di osservazioni, si avrà una distribuzione empirica che sarà praticamente simile alla vera distribuzione teorica incognita, per cui il problema incontrato in precedenza è risolvibile ricorrendo a questo approccio.

Sebbene siano attualmente disponibili elaboratori molto potenti, non sempre è possibile seguire tale via; si consideri infatti il caso di modelli molto complessi, la cui stima richiede elevati tempi di elaborazione (come avviene ad esempio per i modelli a scatola nera), è impensabile in tale situazione generare un elevato numero di campioni e per ognuno di questi stimare il modello considerato, al fine di ricavare la distribuzione empirica

di $\hat{y}$: se per un certo modello, ad esempio, i tempi di elaborazione richiesti per la stima sono di 1 ora, generando 100.000 campioni, il tempo necessario per poter stimare 100.000 volte il modello è di 100.000 ore, ossia pari a circa 11 anni!

In alcuni casi è possibile collegare in parallelo alcuni computer e generare un certo numero di campioni e stimarne il relativo modello su ognuno di essi, quindi raccogliere le informazioni indipendenti fornite dai vari computer al fine di ricavare la distribuzione empirica di $\hat{y}$.

Un utilizzo dei modelli molto analogo a quello previsivo è quello simulativo: il ricercatore è interessato a capire quale dinamica avrebbe la *y* in funzione di una certa configurazione delle esplicative $\underline{x}$.

Posto in questi termini, il problema sembra analogo a quello della previsione; formalmente, ossia da un punto di vista algoritmico, le due procedure si equivalogono, poiché bisogna sempre verificare cosa accade alla *y* facendo certe ipotesi sulla $\underline{x}$. La differenza sostanziale fra i due approcci è di tipo concettuale: in ambito previsivo il ricercatore considera una configurazione di $\underline{x}$ effettiva, cioè, per esempio, considerando un'analisi di serie storiche con $\underline{x}_t = (y_t, y_{t-1},,\ldots, y_{t-d+1})^t$, il ricercatore conosce esattamente il valore che le esplicative $\underline{x}$ assumono al tempo *t*, in quanto tali dati sono presenti nel campione di osservazioni di cui dispone e che ha utilizzato per stimare il modello; in ambito simulativo, invece, il ricercatore fa un'ipotesi di lavoro, ossia si chiede "cosa accadrebbe alla *y* qualora capitasse che la $\underline{x}$ sia pari a un valore fissato $\underline{x}^0$", ma non è detto che il valore $\underline{x}^0$ si verifichi; ad esempio, nell'analisi delle serie storiche, il ricercatore potrebbe assumere che ad un certo periodo futuro ($t+j$) si possa verificare una certa configurazione delle *y* antecedenti tale periodo, ma conseguenti il periodo attuale *n* e chiedersi, quindi, che valore assumerebbe

y_{t+j}, data l'ipotesi di lavoro $\underline{x}^0_{t+j} = \left(y^0_{t+j-1}, y^0_{t+j-2}, \ldots, y^0_{t+j-d}\right)^t$ e ipotizzando che il modello stimato al tempo *n* sia valido anche per il periodo futuro oggetto di analisi (problema della stabilità).

Il ricercatore, anche in questo caso, ricaverà il valore indagato allo stesso modo, cioè:

$$\hat{y}^0_{t+j} = f\left(\underline{x}^0_{t+j}; \underline{\hat{\theta}}\right)$$

ma mentre in ambito previsivo il ricercatore avrebbe effettuato (*j-1*) previsioni, impiegando il modello stimato, al fine di avere un vettore di previsioni $\underline{\hat{x}}_{t+j} = \left(\hat{y}_{t+j-1}, \hat{y}_{t+j-2}, \ldots, \hat{y}_{t+j-d}\right)^t$, in ambito simulativo il ricercatore fa "una forzatura" ponendo $\underline{x}_{t+j} = \underline{x}^0_{t+j}$ (con $\underline{x}^0_{t+j} \neq \underline{\hat{x}}_{t+j}$ in generale) e utilizza il modello.

3 **Classificazione**

Se la variabile *y* è di natura qualitativa, la tecnica regressiva che consente di modellare la dipendenza di *y* dal vettore esplicativo $\underline{x}$ viene definita classificazione ed il modello generale è del tipo:

$$E\left[y|\underline{x}\right] = f\left(\underline{x}\right).$$

Anche in questo caso è possibile seguire un approccio di tipo parametrico oppure non parametrico; se l'approccio è parametrico, allora la funzione *f* viene definita a priori e completamente specificata da un vettore di parametri $\underline{\theta}$, per cui il modello viene scritto nel seguente modo:

$$E\left[y|\underline{x}\right] = f\left(\underline{x}; \underline{\theta}\right).$$

Nell'approccio non parametrico, invece, la funzione *f* non viene definita a priori, ma viene stimata a livello locale, ossia si considerano vari intorni di

$\underline{x}$ e in ogni intorno si stima *f* mediante una media aritmetica ponderata delle osservazioni *y* associate all'intorno considerato.

Le varie tecniche non parametriche (kernel, spline, k-NN, regressogram, ecc.) differiscono fra loro per il modo in cui ponderano tali valori e per il modo in cui viene selezionato l'intorno $\underline{x}$ sul quale verrà operata la "mediazione".

Nel caso di variabili qualitative, molto spesso è nota la distribuzione di probabilità di *y* condizionatamente ad $\underline{x}$, ad esempio per una variabile binaria si considera una distribuzione bernoulliana con valore atteso $E(y|\underline{x}) = f(\underline{x}) = \pi_{\underline{x}}$, per cui la sua distribuzione sarà:

$$P(y|\underline{x}) = \pi_{\underline{x}}^{y} \cdot (1-\pi_{\underline{x}})^{(1-y)};$$

se invece supponiamo che la variabile *y* sia di tipo categoriale con *J* modalità, e che $C = \{1,2,...,J\}$ sia l'insieme delle possibili etichette che essa può assumere, allora la sua distribuzione condizionata, considerato $\{y = j\}$ l'evento che un'osservazione *y* sia pari alla j-esima modalità avrà una distribuzione multinomiale definita dal vettore parametrico $\underline{\pi}_{\underline{x}}$ con:

$$\underline{\pi}_{\underline{x}} = \begin{pmatrix} \pi_1(\underline{x}) \\ \pi_2(\underline{x}) \\ \vdots \\ \pi_J(\underline{x}) \end{pmatrix}$$

dove $P\{y = j|\underline{x}\} = \pi_i(\underline{x})$, con il vincolo che :

$$\sum_{j=1}^{J} \pi_j(\underline{x}) = 1.$$

Se invece *y* è una variabile con $C \equiv \aleph \bigcup \{0\}$, indicante il numero di eventi che si verificano in un prefissato dominio temporale o spaziale, allora la sua distribuzione condizionata ad $\underline{x}$ segue una distribuzione di Poisson di parametro $E(y|\underline{x}) = f(\underline{x}) = \lambda_{\underline{x}}$, per cui la sua distribuzione sarà:

$$P\left(y|\underline{x}\right)=e^{-\lambda_{\underline{x}}}\frac{\lambda_{\underline{x}}^{y}}{y!}.$$

Da quanto sopra affermato, si comprende che in tali situazioni è preferibile seguire un approccio di tipo parametrico: nota la distribuzione di *y*, assumendo che le osservazioni $\{y_i\}_{i=1}^{n}$ siano fra di loro indipendenti, allora la distribuzione congiunta del campione è pari al prodotto delle singole distribuzioni marginali, per cui si ha:

$$P\left(\underline{y}|\underline{x};\underline{\theta}\right)=P\left((y_1,y_2,...,y_n)^t|\underline{x};\underline{\theta}\right)=\prod_{i=1}^{n}P\left(y_i|\underline{x};\underline{\theta}\right)$$

nota questa distribuzione è allora possibile applicare il metodo di stima parametrico della massima verosimiglianza, attraverso il quale si cerca il vettore $\underline{\theta}$ che massimizza tale funzione, osservandola però come funzione dei parametri, date le osservazioni. Per implementare tale metodo, però, è spesso necessario ricorrere a procedure iterative, che determinano tale vettore per via numerica attraverso procedure a passi.

Dopo aver definito e stimato il modello, anche per variabili qualitative occorre verificare la validità dello stesso: si procede, dunque, all'implementazione di una serie di test statistici, che verificano la validità delle assunzione fatte, e in un contesto parametrico tendono a saggiare la significatività dei parametri inclusi nel modello stesso.

Una volta ottenuto la versione finale del modello è possibile utilizzarlo a fini previsivi; anche in questo caso è molto utile la conoscenza della distribuzione di *y* condizionatamente ai valori di $\underline{x}$.

Per cui assunto che in un certo arco temporale o dominio spaziale sia $\underline{x}=\underline{x}^0$, il valore y^0 previsto dal modello è quel valore di *y* che massimizza la funzione $P\left(y\left|\underline{x}^0\right.\right)$, ossia è:

$$y^0=\arg\max_j P\left(j\left|\underline{x}^0\right.\right).$$

La conoscenza della distribuzione $P\left(y\middle|\underline{x}\right)$ non è indispensabile, in generale basta disporre di una regola di decisione che consente di attribuire un'etichetta *j* relativamente alla variabile *y* in corrispondenza di una specificata configurazione del vettore di esplicative $\underline{x}$.

Ovviamente tale regola è in realtà basata su tale distribuzione di probabilità e comunque la conoscenza di essa ci consente di sapere con quale probabilità stiamo prevedendo un certo valore, in quanto se si ha che

$$P\left(y = j\middle|\underline{x}^0\right) = p$$

ciò vuol dire che il ricercatore può sostenere che qualora sia $\underline{x} = \underline{x}^0$, allora con una probabilità pari a *p* ci si deve aspettare che *y* sia pari alla *j*-esima modalità.

Uno strumento utile al fine di giudicare la validità del modello è la matrice di confusione. Stimato il modello parametrico $f\left(\underline{x};\hat{\underline{\theta}}\right)$ è possibile valutare il vettore dei dati teorici $\{\hat{y}_i\}_{i=1}^n$ dove il generico elemento $\hat{y}_i$ è dato da:

$$\hat{y}_i = \arg \max_j P\left(j\middle|\underline{x}_i;\hat{\underline{\theta}}\right)$$

ossia rappresenta il valore della variabile *y*, determinato dal modello stimato, in corrispondenza della configurazione del vettore $\underline{x}$ nell'i-esima unità campionaria.

È quindi utile confrontare i valori osservati $\{y_i\}_{i=1}^n$ con quelli previsti dal modello, ossia $\{\hat{y}_i\}_{i=1}^n$. Entra in gioco, dunque, la matrice di confusione *A*, ossia una matrice di dimensione $(J \times J)$, se *J* sono le possibili modalità di *y*, il cui generico elemento A_{hk} indica il numero di osservazioni, fra le *n*, per le quali il valore osservato è pari all'*h*-esima modalità, mentre il valore calcolato è pari alla *k*-esima modalità, ossia:

$$A_{hk} = \#\left\{i \in \{1,2,...,n\}:\ (y_i = h) \bigcap (\hat{y}_i = k)\right\}.$$

Dove la funzione # esprime il conteggio delle unità che soddisfano i requisiti dell'argomento posto fra parentesi.

I valori $\{A_{hh}\}$ della matrice, ossia i valori posti sulla diagonale principale di *A*, indicano il numero di unità in cui i dati osservati e quelli teorici calcolati per mezzo del modello sono uguali, relativamente all'*h*-esima modalità di *y*. Sommando tali valori si ottiene, quindi, il numero di osservazioni classificate in maniera corretta, ossia:

$$\text{unità classificate correttamente} = \sum_{h=1}^{J} A_{hh} \ .$$

Considerando diversi modelli si sceglierà quindi quello che classificherà correttamente il maggior numero di unità.

Operando in tal modo, si ipotizza implicitamente, che la "perdita" ottenuta classificando in modo non corretto le unità delle varie modalità è la medesima, a prescindere dagli indici considerati; in alcune situazioni, però, classificare nella generica *h*-esima modalità un'unità della *j*-esima classe, può avere un peso diverso in base alle modalità considerate, si rende pertanto necessario l'introduzione di un sistema di pesi $\{C_{hk}\}$, dove il generico elemento esprime la "perdita" subita nel classificare un'unità dell'*h*-esima classe nella *k*-esima modalità.

Introdotti tali pesi, si considera una funzione, associata al modello,da massimizzare, precisamente:

$$E = \sum_{h=1}^{J} A_{hh} C_{hh} - \sum_{h \neq k} \sum A_{hk} C_{hk}$$

4 Principali modelli per variabili qualitative binarie

Si consideri il caso in cui la variabile *y* possa assumere esclusivamente due valori, precisamente zero ed uno, dove un valore pari ad uno indica la presenza di un certo evento, detto anche "successo", mentre il valore zero indica il non presentarsi di un evento, spesso detto "insuccesso".

Se si assume che la probabilità che l'evento in questione si verifichi è influenzata da un certo insieme di variabili esplicative $\underline{x}$, allora è lecito modellare tale fenomeno in funzione di queste variabili.

Per modellare questo fenomeno, occorre quindi specificare una funzione *f* che lega il valore atteso di *y* condizionatamente a $\underline{x}$, ai valori $\underline{x}$ medesimi e che dipenda da un set di parametri ignoti $\underline{\theta}$.

In una situazione come questa un modello parametrico molto impiegato è quello che va sotto il nome di regressione logistica, dove la funzione *f* impiegata è una funzione a forma di S allungata, con il codominio pari all'intervallo (*0,1*), corrispondente allo spazio in cui varia il valore atteso condizionato di *y*, essendo una probabilità.

Precisamente il modello logistico è del tipo:

$$P\left(y=1\middle|\underline{x};\underline{\theta}\right)= f\left(\underline{x};\underline{\theta}\right)=\frac{e^{\underline{x}^t\underline{\theta}}}{1+e^{\underline{x}^t\underline{\theta}}}=\frac{1}{1+e^{-\underline{x}^t\underline{\theta}}}.$$

Assumendo che $\underline{x}\in D\subset\Re^d$ e che $\underline{\theta}\in\Theta\subset\Re^d$, segue che il prodotto interno fra i due vettori $\underline{x}^t\underline{\theta}$ è una variabile reale, per cui posto $\eta=\underline{x}^t\underline{\theta}$, si ha che :

$$f(\eta)=\frac{e^{\eta}}{1+e^{\eta}}$$

ed inoltre:

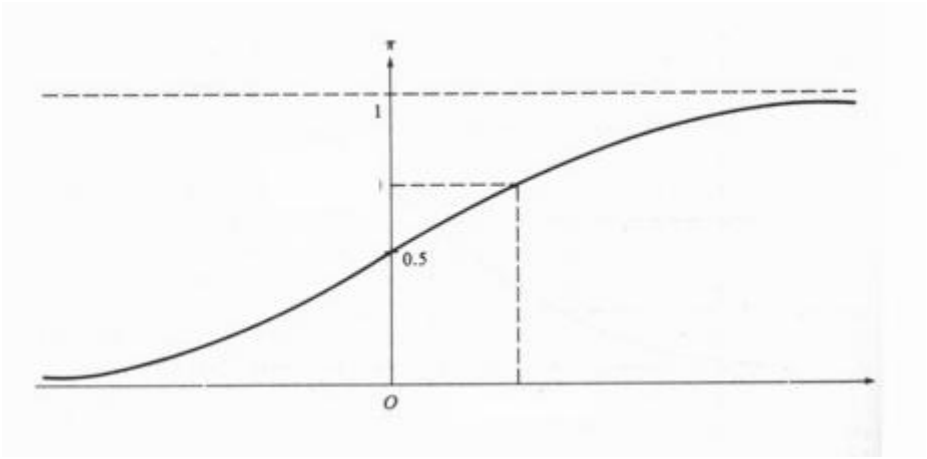
- $\lim_{\eta\to-\infty} f(\eta)=0$;

- $\lim_{\eta \to +\infty} f(\eta) = 1$;
- $f(0) = \frac{1}{2}$ punto di flesso della funzione.

La funzione *f* è una funzione simmetrica, ed ha una forma spiccatamente non lineare alle code, ossia per valori elevati di η in valore assoluto, mentre ha un andamento molto lineare intorno a valori nulli di η.

La figura seguente mostra l'andamento della funzione logistica *f*, in funzione della variabile η.

Fig. 1 Funzione logistica



Definito il modello logistico, si procede come specificato nella sezione precedente, per cui la distribuzione del campione $\underline{y} = (y_1, \ldots, y_n)^t$, è il prodotto di *n* distribuzioni bernoulliane indipendenti, ossia:

$$P(\underline{y}|\underline{x};\underline{\theta}) = \prod_{i=1}^{n} f(\underline{x}_i;\underline{\theta})^{y_i} (1 - f(\underline{x}_i;\underline{\theta}))^{(1-y_i)}$$

per semplicità si passa al logaritmo della funzione precedente e si ottiene la log-verosimiglianza:

$$\log L(\underline{\theta}) = \sum_{i=1}^{n} [y_i \log(f(\underline{x}_i;\underline{\theta})) + (1 - y_i)\log(1 - f(\underline{x}_i;\underline{\theta}))]$$

si determinano, quindi, le stime di massima verosimiglianza dei parametri, massimizzando tale funzione, tali stimatori non si ottengono in forma chiusa, ma attraverso un procedimento iterativo e sono i valori $\underline{\hat{\theta}}$ che soddisfano il sistema di equazioni:

$$\frac{\partial \log L}{\partial \underline{\theta}} = \underline{0}.$$

Stimati i parametri che definiscono il modello, si eseguono diversi test statistici per saggiare la significatività dei parametri.

Le distribuzioni delle statistiche test impiegate non sono note per qualsiasi ampiezza campionaria, ma considerando che le stime sono quelle di massima verosimiglianza è possibile effettuare test basati sul rapporto delle verosimiglianze, la cui distribuzione asintotica, come è noto, è di tipo χ^2 con gli opportuni gradi di libertà; più precisamente se $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)^t$ sono i parametri del modello, è possibile saggiare l'ipotesi che un sottoinsieme di questi siano nulli, per cui, ad esempio, si potrebbe saggiare l'ipotesi:

$$H_0: \quad \theta_1 = \cdots = \theta_k = 0.$$

La statistica test impiegata, basata sul rapporto delle verosimiglianze, è :

$$T = 2 \cdot \left(\log L\left(\underline{\hat{\theta}}\right) - \log L\left(\underline{\hat{\theta}}_0\right) \right)$$

dove $\log L\left(\underline{\hat{\theta}}\right)$ è la logverosimiglianza calcolata sotto l'ipotesi del modello con *d* parametri, ossia il modello completo di partenza, mentre $\log L\left(\underline{\hat{\theta}}_0\right)$ è la logverosimiglianza calcolata in relazione agli stimatori di massima verosimiglianza del modello ridotto, ossia quello in cui i primi *k* parametri sono nulli, cioè il modello valido sotto l'ipotesi nulla.

Tale statistica test segue una distribuzione asintotica di tipo χ^2, sotto l'ipotesi nulla, con *k* gradi di libertà (in questo esempio).

Determinata la versione "finale" del modello, è quindi possibile fare previsione attraverso il procedimento descritto nella sezione precedente.

Un modello parametrico alternativo a quello logistico è quello che va sotto il nome di probit. In tale modello anziché considerare la funzione *f* logistica, si utilizza la funzione Φ , ossia la funzione di ripartizione di una distribuzione normale di media zero e varianza uno. Dove tale funzione di ripartizione è definita come:

$$\Phi(\eta) = \int_{-\infty}^{\eta} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du .$$

Precisamente il modello è definito nel modo seguente:

$$E[y|\underline{x}] = f(\underline{x};\underline{\theta}) = \Phi(\underline{x}^t \underline{\theta}).$$

Anche in questo caso si seguirà la procedura eseguita in precedenza, poiché è nota la distribuzione di $\underline{y}$ condizionatamente ad $\underline{x}$, ma rispetto il caso logistico cambia il parametro $\pi_{\underline{x}}$, ossia la probabilità di successo.

Il modello logistico ed il modello probit, sono casi particolari di una vasta famiglia di modelli, applicabili anche a variabili risposta continue, che prende il nome di modelli lineari generalizzati, espressi con un acronimo GLM.

Tali modelli considerano una combinazione lineare dei parametri, detto predittore lineare, precisamente:

$$\eta = \underline{x}^t \underline{\theta}$$

quindi modellano il valore atteso condizionato di *y*, attraverso tale predittore lineare, precisamente considerano una funzione g invertibile che lega il predittore lineare con il valore atteso condizionato, ossia:

$$g(E[y|\underline{x}]) = \eta$$

per cui, essendo invertibile la funzione *g*, si può scrivere:

$$E[y|\underline{x}] = g^{-1}(\eta) = f(\eta).$$

Sostanzialmente, quindi, i modelli GLM sono modelli dove la funzione *f* che lega $\underline{x}$ ed il valore atteso di *y* è una funzione di $\underline{x}$ e $\underline{\theta}$, attraverso il loro prodotto interno, che per comodità si indica con η.

Variando la funzione *g* varia il modello adattato, ma se si fa questa ipotesi (di linearità generalizzata) il modello rientra comunque in questa vasta famiglia.

Bibliografia

- Chiodi M. (1997), *Tecniche elementari di simulazione in statistica.* Appunti del corso di statistica computazionale, Palermo.
- Davidian M., Giltinan D.M, *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall
- Ryan T. P., *Modern Regression Methods*, Wiley
- Granger C.W., Terasvirta T. (1992), *Modelling nonlinear economic relationships*. Oxford University Press
- Simonoff J.S. *Smoothing Methods in Statistics.* Springer
- Bowman, Azzalini *Applied smoothing techniques for data analysis.* Oxford Science Publications
- Hart P.E., *Non parametric smoothing and lack of fit.* Springer
- Green P.J., Silverman B.W. (1994), *Non parametric regression and generalized linear models.* Chapman & Hall
- Stone C.J. (1977), *Consistent nonparametric regression*. The Annals of Statistics, Vol.5, No.4, 595-645

- Silverman B.W.(1985), *Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting*. Royal Statistical Society B 47, No. 1, 1-52
- Hardle W. (1989), *Applied nonparametric regression*. Cambridge University Press
- Agresti A. (1990), *Categorical Data Analysis.* John Wiley & Sons.
- Bishop C.M. (1995), *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.
- Ripley B.D. (1996), *Pattern Recognition and Neural Networks.* Cambridge, University Press.
- Johnston J. (1996), *Econometrica.* Franco Angeli
- Corbetta P. (1992), *Metodi di analisi multivariata per le scienze sociali.* Il Mulino.
- Piccolo D. (1994), *Introduzione all'analisi delle serie storiche.* La Nuova Italia Scientifica.