Dispensa di *Statistical Modeling*

Parte I
Modello lineare

I modelli esplicitano la relazione tra variabili trovando un compromesso tra adattamento ai dati e parsimonia (Popper): "*all models are wrong, some are usefull*". Infatti un modello lineare può rappresentare in modo esatto i dati tramite una funzione $y = f(x_1, x_2, x_3...x_k)$ senza errore ε. L'errore ε è dovuto a variazioni individuali, componenti sistematiche del modello, errori di misura o di campionamento. Un modello è così caratterizzato dalla formula:

$$\underline{y} = f(x_1, x_2, x_3, ...x_k) + \varepsilon$$

A livello teorico la scelta delle variabili indipendenti deve tenere conto di relazioni, anche ipotizzate; di fatto però sono comunque influenzate dalla metodologia di raccolta, dalla popolazione e dal campionamento effettuato. Il modello dunque è prima specificato, poi stimato e verificato e infine, se supera i test, utilizzato.

Specificazione del modello. Inizialmente si specifica quali sono le variabili dipendenti e quale la dipendente e la relazione tra queste (lineare o meno).

Stima del modello. Successivamente si procede col calcolo del vettore b (stima del vettore dei parametri β) basandosi sul campione (di numerosità n), mentre la componente erratica ε rimane per definizione ignota.

Verifica del modello. Si passa alla verifica della bontà del modello grazie a indicatori descrittivi e test statistici; inoltre anche l'analisi dei residui consente una valutazione della bontà di adattamento. La fase di verifica serve per rifiutare il modello se eccessivamente sbagliato (dunque è necessario specificare un nuovo modello), ma non fornisce indicazioni sulla correttezza del modello.

La formula del modello è riassumibile in forma matriciale con la formula

$$y = \beta X + \varepsilon$$

dove però la matrice del disegno X è composta da una colonna fittizia composta solamente da 1 e dai dati campionari, per aggiungere al vettore b la quota b_0. I coefficienti b_j si interpretano come la variazione unitaria della variabile x_j a parità di altre condizioni (*ceteris paribus*).

1 Stima del modello.

Uno degli approcci utilizzati per il calcolo del vettore b è il metodo dei minimi quadrati lineri:

$$\begin{aligned}&\min \sum (y_i - x_i b)^2 \\ &= \min \sum e_i^2 \\ &= \min\{(y - Xb)'(y - Xb)\}\end{aligned}$$

Si calcola quindi il gradiente della funzione obiettivo:

$$\frac{\partial e'e}{\partial b} = -2X'(y - Xb)$$

L'esistenza di un minimo impone la nullità del gradiente ($X'(y - Xb)$); e quindi, se la matrice X ha rango pieno, il sistema possiede un'unica soluzione per:

$$b = (X'X)^{-1}X'y$$

Lo stimatore rappresenta quindi il vettore dei parametri dei minimi quadrati; inoltre questa stima coincide con la stima lineare efficiente dei parametri.

2 Bontà di adattamento.

Per misurare quanto un modello si adatta ai dati, generalmente si usa l'indice di correlazione R^2: l'obiettivo è trovare una relazione che permette di spiegare la variabilità del fenomeno. Si divide dunque la variabilità in *spiegata* e *residua*:

$$\begin{aligned}y_i &= \hat{y}_i + e_i \\ \sigma^2_{tot} &= \sigma^2_{sp} + \sigma^2_{res}\end{aligned}$$

Si può quindi calcolare il coefficiente di correlazione multipla (o di determinazione) R^2 che rappresenta la quantità di varianza spiegata dal modello:

$$R^2 = \frac{\sigma^2_{sp}}{\sigma^2_{tot}}$$

Tuttavia questa misura aumenta con l'inserimento di ogni nuovo regressore (senza mai diminuire) andando contro al principio di parsimonia: si introduce quindi una penalità per la complessità del modello ($\tilde{R}^2$, o R^2 aggiustato).

$$\tilde{R}^2 = 1 - \frac{\sigma^2_{res}}{\sigma^2_{tot}} \cdot \frac{n-1}{n-k-1}$$

3 Modello lineare classico.

Il modello lineare classico si basa su alcune assunzioni (5 delle quali semplificatrici) per descrivere la realtà:

1. linearità;
2. numerosità della popolazione;
3. non sistematicità degli errori;
4. sfericità degli errori;
5. non stocasticità delle variabili esplicative;
6. non collinearità delle variabili esplicative;
7. normalità degli errori.

Linearità. Necessaria per la forma funzionale del modello (modello di regressione *lineare*): i parametri sono stimati in modo funzionale.

Numerosità della popolazione. È necessario che la matrice inversa di $X'X$ sia unica perchè si possa effettuare una stima dei parametri del modello. Dunque il numero di osservazioni deve essere maggiore del numero di variabili più la costante: $n > k + 1$.

Non sistematicità degli errori. L'errore ε rappresenta la variazione della variabile dipendente non spiegata dal modello e dovrebbe essere casuale. Il valore atteso dunque della variabile ε è nullo: $E(\varepsilon|X) = 0$.

Sfericità degli errori (omoschedasticità e incorrelazione). La varianza della variabile ε è costante (omoschedastica) e non correlata con le variabili esplicative (incorrelata). Di fatto però in molte situazioni si verifica correlazione tra le osservazioni. Formalmente:

$$Var(e_i) = \sigma^2$$
$$Cov(e_i, e_j) = 0$$

Non stocasticità delle variabili esplicative. I valori delle variabili esplicative non sono soggetti a fluttuazioni dipendenti dal campione, perciò $Cov(X, \varepsilon) = 0$. Eventuali componenti stocastiche sono riassunte dalla componente erratica ε.

Non collinearità delle variabili esplicative. Per far sì che la matrice del disegno abbia rango pieno, non possono esserci due variabili con correlazione perfetta ($\varrho = \pm 1$): si deve procedere eliminando una delle due variabili, che rappresentano lo stesso fenomeno registrato in modi diversi.

Normalità degli errori. Se gli errori seguono una distribuzione normale è possibile effettuare test statistici o costruire intervalli di confidenza e previsione. Infatti, sotto ipotesi di normalità si ha che:

$$\varepsilon \sim N(0, \sigma^2)$$

e quindi:

$$b \sim N(\beta, \sigma^2(X'X)^{-1})$$

3.1 Assunzioni del modello lineare classico.

Il modello lineare classico si basa su quattro assunzioni semplificatrici, che cadono nei modelli più complessi.

Prima assunzione. La distribuzione della componente erratica ε condizionata a X ha media nulla: $E(\varepsilon_i|X) = 0$

Seconda assunzione. Le osservazioni sono tra di loro indipendenti.

Terza assunzione. I valori anomali (*outliers*) sono improbabili e con momento quarto finito. Un valore anomalo è un'osservazione che da sola altera significativamente la distribuzione: è necessario rimuoverla dalla matrice del disegno per poter ottenere un modello veritiero.

Quarta assunzione. Non si verificano casi di collinearità perfetta (ovvero un regressore non è una funzione lineare esatta degli altri). Eventuali variabili che non rispettano questa assunzione sono semplicemente rimosse senza danno.