

# Dispensa di *Statistical Modeling*

## Indice

<b>I</b>	<b>Modello lineare</b>	<b>2</b>
<b>1</b>	<b>Stima del modello.</b>	<b>2</b>
1.1	Significatività statistica dei parametri. . . . .	3
<b>2</b>	<b>Bontà di adattamento.</b>	<b>3</b>
<b>3</b>	<b>Modello lineare classico.</b>	<b>4</b>
3.1	Assunzioni del modello lineare classico. . . . .	5

## 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 useful*”. Infatti un modello lineare può rappresentare in modo esatto i dati tramite una funzione  $y = f(x_1, x_2, x_3 \dots x_k)$  senza errore  $\varepsilon$ . L'errore  $\varepsilon$  è dovuto a variazioni individuali, componenti sistematiche del modello, errori di misura o di campionamento. Un modello è così caratterizzato dalla formula:

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

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  $\beta$ ) basandosi sul campione (di numerosità  $n$ ), mentre la componente erratica  $\varepsilon$  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 lineari:

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

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.

Questa stima di  $b$  è *BLUE* (*Best Linear Unbiased Estimator*), dimostrabile grazie al Teorema di Gauss Markov: risulta essere lo stimatore con efficienza maggiore tra gli stimatori corretti. La sua varianza è pari a:

$$Var(b) = \sigma^2(X'X)^{-1}$$

### 1.1 Significatività statistica dei parametri.

Un parametro  $b_i$  è considerato *significativo* solamente se supera un test  $t$  (con  $n - 1$  gradi di libertà) con  $H_0 : b_i = 0, H_1 : b_i \neq 0$  con significatività  $\alpha$  arbitraria.

$$\text{p-value} = T_{n-1}\left(\frac{b_i}{\sigma_{b_i}^2}\right)$$

$H_0$  è accettato per  $\text{p-value} < \alpha$ , altrimenti è rigettata l'ipotesi nulla e accettata l'ipotesi alternativa  $H_1$ .

Analogamente, si può costruire un test  $F$  (con  $k$  gradi di libertà al numeratore e  $n - k - 1$  al denominatore) per verificare la significatività dell'intero vettore  $b$ :  $H_0 : b_i = 0 \quad \forall i \leq k$ ,  $H_1 : \exists b_i \neq 0$ .

$$\text{p-value} = F_{k, n-k-1}\left(\frac{\frac{\sigma_{res}^2}{k}}{\frac{\sigma_{sp}^2}{n-k-1}}\right)$$

Grazie a questo ultimo test, è possibile confrontare due modelli annidati: posto  $q < k$  il numero di parametri del secondo modello, si verifica quanto sia significativo l'aumento di complessità del modello.

$$\text{p-value} = F_{k-q, n-k-1}\left(\frac{\frac{\sigma_{p.sp}^2 - \sigma_{k.sp}^2}{k-q}}{\frac{\sigma_{res}^2}{n-p-1}}\right)$$

## 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*:

$$y_i = \hat{y}_i + e_i$$

$$\sigma_{tot}^2 = \sigma_{sp}^2 + \sigma_{res}^2$$

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

$$R^2 = \frac{\sigma_{sp}^2}{\sigma_{tot}^2}$$

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).

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

### 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 perché 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  $\varepsilon$  rappresenta la variazione della variabile dipendente non spiegata dal modello e dovrebbe essere casuale. Il valore atteso dunque della variabile  $\varepsilon$  è nullo:  $E(\varepsilon|X) = 0$ .

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

$$\begin{aligned} Var(e_i) &= \sigma^2 \\ Cov(e_i, e_j) &= 0 \end{aligned}$$

**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  $\varepsilon$ .

**Non collinearità delle variabili esplicative.** Per far sì che la matrice del disegno abbia rango pieno, non possono esserci due variabili con correlazione perfetta ( $\rho = \pm 1$ ): 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  $\varepsilon$  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.