

Statistical Modeling

Sommario

Lo scopo del corso è riuscire a muoversi con disinvoltura all'interno di un dataset: è privilegiata la teoria perché indipendente dalla piattaforma; inoltre sono presentati numerosi esercizi svolti e database di esempio. Durante il corso si generalizza il modello lineare classico andando oltre alle sue premesse, arrivando al modello lineare multi-livello (che costruisce una gerarchia nei dati).

L'esame è composto da una parte teorica di due domande (da un database di 15 note) e un esercizio da eseguire in R o SAS. Le slide sono sufficienti alla preparazione dell'esame, ma in più è offerta una dispensa ufficiale.

Parte I

Premesse all'analisi

Il modello stabilisce cosa fare coi dati: la finalità del modello stabilisce l'interpretazione da dare al risultato e i dati da raccogliere. I test sul modello devono essere fatti su un campione *significativo*: i risultati potrebbero non essere veritieri. Il campione è necessario anche coi *Big Data*, dato che aumentano l'eterogeneità dei dati.

Nella prima parte della costruzione del modello, lo statistico deve collaborare con l'esperto di dominio per individuare il fine del modello e i caratteri da osservare. In un secondo momento si procede con un'analisi descrittiva (o esplorativa) del dataset (tramite grafici come istogrammi o boxplot, oppure calcolando valori indice). Si individuano dunque gli *outliers* (ovvero i valori anomali), da eliminare prima dell'analisi vera e propria.

Si analizza poi la matrice di correlazione R , per eliminare casi di *multicollinearità* (ovvero i casi in cui la correlazione $\rho \rightarrow 1$).

Si deve sciogliere il conflitto tra adattamento dei dati e parsimonia in modo tale da rendere comprensibile il modello ma garantendo una certa qualità nell'adattamento.

Il modello reale tiene conto dunque dell'errore, che considera cosa non è conoscibile e di componenti sistematici (esclusione di variabili) o di errori di misurazione; dai campioni (poco significativi) inoltre si può avere un errore stocastico. Il modello statistico ha come scopo la comprensione e la minimizzazione dell'errore. Non si arriva a formulare regole di causa-effetto ma solamente a relazioni empiriche.

Si stabilisce il modello di regressione individuando le variabili (e il loro grado) e il valore dei parametri. Dunque si calcolano i valori sti-

mati sui valori noti (\hat{y}_i) e si calcola l'errore con la formula $\epsilon_i = y_i - \hat{y}_i$. Al variare del campione, i valori dei parametri cambiano: è così possibile costruire una distribuzione (col metodo Monte-carlo). L'errore dunque non è deterministico ma stocastico (cioè casuale), dato che varia in base al campione selezionato.

Parte II

Modello lineare classico

Il modello lineare classico ha delle premesse molto rigide a causa della sua natura. Si tratta però di *ipotesi semplificatrici* che saranno rimosse con modelli più avanzati (tranne per le ultime due).

Linearità.

Le variabili e i parametri del modello sono lineari:

$$E(y|X) = Xb$$

Non sistematicità degli errori.

Affinchè il modello sia imparziale, il valore medio del termine di errore deve essere uguale a zero. Quindi gli errori ϵ hanno media nulla. Formalmente:

$$E(\epsilon_i = 0)E(\epsilon|X) = 0$$

ne consegue che:

$$E(y|X) = Xb$$

Sfericità degli errori.

Gli errori sono omoschedastici (cioè la varianza è costante) e non correlati:

$$\begin{aligned} E(\epsilon) &= E(y - Xb) \\ &= E(Xb - Xb) = 0 \end{aligned}$$

e quindi:

$$\begin{aligned} var(\epsilon) &= E(\epsilon^2) = \sigma^2 \\ cov(\epsilon_i, \epsilon_j) &= E(\epsilon_i \epsilon_j) = 0 \end{aligned}$$

Di fatto molti casi reali sono correlati tra di loro (ovvero un'osservazione influenza le altre), come nella finanza, nelle serie temporali o nelle coordinate spaziali.

Non omoschedasticità.

È una mera convenzione: si presume che la matrice del disegno X sia fissa e nota; questa astrazione garantisce una semplificazione del problema. Le componenti non stocastiche sono riassunte dalla componente erratica.

Non stocasticità delle variabili esplicative

I valori delle x_j variabili esplicative non sono soggetti a fluttuazioni da campione a campione, perciò $E(X) = X$ e $Cov(X, \epsilon) = 0$. La parte non fissa delle x_j finisce in ϵ .

Non collinearità.

Le variabili della matrice X sono linearmente indipendenti. X ha rango uguale al numero delle variabili, costante inclusa. Da ciò deriva che $X'X$ non è singolare: in caso contrario $X'X$ non è invertibile ed il modello non risolvibile.

Numerosità della popolazione.

Il numero di osservazioni è sempre maggiore del numero di caratteri osservati: $n \geq p + 1$. Se questa proprietà non è soddisfatta, la matrice del disegno non è invertibile e dunque non è possibile la costruzione di alcun modello.

Normalità degli errori.

Il soddisfacimento dell'ipotesi che $\epsilon_i \sim N(0, \sigma^2)$ ovvero che la distribuzione dell'errore campionario sia normale provoca anche la normalità nella distribuzione di y e di $\hat{\beta}$.

1 Stima dei parametri.

Per stimare i parametri, la formula è:

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

Si dice che uno stimatore è *corretto* se il valore medio della sua distribuzione è pari a quello della popolazione ($E(\theta) = E(x)$). Invece per *consistenza* si intende che con una popolazione infinita

il valore stimato sia quello della popolazione. In sostanza:

$$\lim_{n \rightarrow \infty} Var(\hat{\beta}) = 0$$

Tra due (o più) stimatori è preferibile lo stimatore più *efficiente*, ovvero con una varianza minore nella sua distribuzione. Generalmente uno stimatore segue una distribuzione normale: si effettua un test per verificare la significatività del parametro:

$$\begin{aligned} P[-Z_{\frac{\alpha}{2}} < \frac{\beta_j}{\frac{\sigma}{\sqrt{n\sigma_{j,j}^{-1}}}} < +Z_{\frac{\alpha}{2}}] = \\ = P[-Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n\sigma_{j,j}^{-1}}} < \beta_j < Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n\sigma_{j,j}^{-1}}}] = 1 - \alpha \end{aligned}$$