Statistical Modeling

Sommario

Lo scopo del corso è riuscire a muoversi con dimestichezza 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 $\varrho \to 1$).

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.

In termini matriciali per la regressione multipla abbiamo:

$$\mathbf{y} = X\mathbf{b} + \varepsilon$$

dove il modello reale tiene conto dell'errore indicato con ε, che tiene conto di 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 nel vettore **b**. Dunque si calcolano i valori stimati sui valori noti ($\hat{y}_i$) e si calcola l'errore con la formula $\varepsilon_i = y_i - \hat{y}_i$. Al variare del campione nella popolazione, i valori dei parametri cambiano: è così possibile costruire una distribuzione (col metodo Montecarlo) di tali parametri. L'errore dunque non è deterministico ma stocastico (cioè casuale), dato che varia in base al campione selezionato.

Il criterio scelto per ricercare **b** è il criterio dei minimi quadrati ordinari che rende minima la norma del vettore di scarti ε

Parte II
Modello lineare classico

Il modello lineare classico ha delle premesse molto rigide a causa della sua natura. Si tratta perlopiù 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.

Non sistematicità degli errori

Il vettore casuale ha valore atteso nullo (altrimenti il nostro modello è inadeguato):

$$E(\varepsilon_i) = 0, i = 1, \dots, n$$

$$E(\varepsilon|X) = 0$$

e quindi segue

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

$$E(y) = E(xb + \varepsilon) = E(Xb) + E(\varepsilon) = Xb$$

Sfericità degli errori.

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

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

e quindi:

$$\begin{aligned} &var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma_i^2 \\ &cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0) \end{aligned}$$

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 di dimensione (n,p+1) ; questa astrazione garantisce una semplificazione del problema. Le componenti non stocastiche sono riassunte dalla componente erratica.

Non numerosità.

Il numero di osservazioni è sempre maggiore del numero di caratteri osservati: $n > p + 1$. Se questa proprietà non è soddisfatta, la matrice del disegno $(X'X)$ non è invertibile e dunque non è possibile la costruzione di alcun modello. Ciò presuppone inoltre che la matrice X abbia rango massimo $p+1$ (visto che $n > p+1$) che può non succedere nel caso di matrici con colonne alta mente correlate ($|\varrho| \to 1$).

Normalità degli errori

Questa non appartiene in senso proprio alle ipotesi ma permette agli stimatori di costruire test e intervalli di confidenza:

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

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.

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, supponiamo $\beta_j \sim N(\mu, \sigma^2)$(ha una distribuzione normale), si effettua un test per verificare la significatività del parametro:

$$P[-Z_{\frac{\alpha}{2}} < \frac{\beta_j}{\frac{\sigma}{\sqrt{n\sigma^{-1}_{j.j}}}} < +Z_{\frac{\alpha}{2}}] =$$
$$= P[-Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n\sigma^{-1}_{j.j}}} < \beta_j < Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n\sigma^{-1}_{j.j}}}] = 1 - \alpha$$

Solitamente nel modello lineare l'ipotesi nulla è l'ipotesi che il parametro sia nullo ($\beta_j \sim N(0, \sigma^2)$).