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.

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 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). La formula è lineare:

$$y = Xb + \varepsilon$$

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 non è imparziale):

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

e quindi segue che:

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

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 $n \times p+1$; 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 $Cov(X, \varepsilon) = 0$. La parte non fissa delle x_j finisce in ε.

Normalità degli errori.

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

Non collinearità.

Le variabili della matrice X sono linearmente indipendenti. X ha rango uguale al numero delle colonne (i caratteri, più una costante). Da ciò deriva che $X'X$ non è singolare: in caso contrario $X'X$ non è invertibile e 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.

1 Stima dei parametri.

Per stimare i parametri, la formula matriciale è:

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

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:

$$\lim_{n\to\infty} Var(\hat{\theta}) = 0$$

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, supponiamo $\beta_j \sim N(\mu, \sigma^2)$ (ha una distribuzione normale), si effettua un test (con probabilità dell'errore di primo grado α arbitrario) per verificare la significatività di un parametro β_j con varianza σ nota:

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

Solitamente nel modello lineare l'ipotesi nulla (H_0) è l'ipotesi che il parametro sia nullo $\beta_j \sim N(0, \sigma^2)$.

I test statistici per la significatività sono generalmente F e T (il primo è il quadrato del secondo) quando la varianza della popolazione σ^2 è ignota; nel caso invece di σ^2 nota allora è possibile sfruttare il test Z.

La distribuzione T di Student è anche esplicitabile come:

$$\frac{Z}{\chi^2 * gdl}$$

mentre la F di Snedecor è riassumibile come:

$$\frac{\frac{SSE}{k}}{\frac{SSR}{n-k-1}}$$

Nello specifico questa formula è sintetizzabile come il rapporto tra due distribuzioni χ^2 ed i rispettivi gradi di libertà.

Oltre al test d'ipotesi per verificare la significatività dei parametri è possibile anche calcolare l'intervallo di confidenza di un

parametro così da misurarne il possibile impatto.

La stima dei *minimi quadrati ordinari (OLS)* e la *stima di massima verosimiglianza* sono equivalenti poichè ricavano gli stessi valori per i parametri sfruttando due meccanismi differenti. Mentre la prima trova il valore dei parametri che minimizza la varianza, la seconda calcola i parametri che massimizzano la probabilità di osservare i valori da noi osservati.

2 Variabili nominali.

Il sistema più diffuso per la gestione di variabili nominali è convertirle in *dummys*: una serie di variabili fittizie con valore 0 o 1 in base al valore assunto dalla variabile. In questo modo è possibile interpretare facilmente il risultato ottenuto. Il numero di variabili dummy generate da un carattere qualitativo è pari a $v - 1$ (dove v è il numero di possibili valori), per evitare multicollinearità; una variabile non è mai inserita.

Parte III
Violazione delle ipotesi classiche

Dato che "tutti i modelli sono falsi", le ipotesi classiche non sono scorrette, tuttavia semplificano in modo eccessivo: in circostanze reali è difficile trovare fenomeni che le soddisfino tutte. Si possono però eliminare tutte le ipotesi classiche (a eccezione delle ultime due elencate precedentemente) in modo tale da ottenere un modello più veritiero.

1 Residui eteroschedastici.

La varianza dell'errore spesso dipende dalle variabili indipendenti: la distribuzione dell'errore è diversa per ogni osservazione. Continuano a valere correttezza, linearità e consistenza:

$$\begin{aligned} E(\hat{B}) &= E(Hy) \\ &= HE(y) \\ &= HE(Xb + e) \\ &= HXE(b) + E(e) \\ &= (X'X)^{-1}X'XE(b) + 0 = b \end{aligned}$$

La varianza tuttavia è maggiore:

$$\begin{aligned} Var(\hat{B}) &= E((Hy - b)(Hy - b)') \\ &= E((H(Xb + e) - b)(H(Xb + e) - b)') \end{aligned}$$

Ciò che non vale più, quindi, è l'efficienza: per questo motivo lo stimatore non è più considerabile BLUE. Questo avrà una conseguenza anche sui test d'ipotesi, per cui la regione di rifiuto diventerà molto più ampia, e per la stima degli intervalli di confidenza, che saranno più stretti e meno affidabili.

Come individuare l'eteroschedasticità.

La violazione dell'ipotesi di eteroschedasticità è rilevabile attraverso due modalità: una di ispezione grafica ed una orientata verso dei test su

misura. Per quanto riguarda l'ispezione grafica, preferibile come primo approccio, è necessario ricorrere a visualizzazioni quale lo **Scatterplot**: con quest'ultimo è possibile mettere a confronto diverse coppie di elementi sugli assi, tutte egualmente valide.
Tra queste coppie abbiamo:

- Y vs X
- *Residui stimati* vs *Ypredetta*
- *Residui al quadrato* vs *Regressori*
- *Valori osservati* vs *Valori predetti*
- *Residui* vs *Regressori*

Per quanto riguarda invece i test di misura, il primo utile allo scopo è il **Test di White**. Questo test si basa sull'ipotesi H_0 che vi sia effettivamente *omoschedasticità* tra i residui e che quindi:

$$Var(e|X) = \sigma^2 I_n$$

Il meccanismo del test è piuttosto semplice:

- Prima di tutto si procede a regredire Y rispetto alle variabili esplicative X_j e a ricavare l'errore ε_i.
- Successivamente si procede a regredire ε_i rispetto alle variabili esplicative x_j e al loro quadrato x_j^2, oltre che alle loro interazioni.
- Si determina il coefficiente di determinazione R^2 della regressione e si procede alla costruzione del valore

$$LM = nR^2$$
$$LM \sim \chi^2$$

  con n numero di regressori.
- Se LM cade all'interno della regione di rifiuto allora vorrà dire che ε_i^2 varia al variare delle x_j e sarà quindi da confermare la presenza di *eteroschedasticità*.

Molto simile al Test di White è il **Test di Breusch-Pagan** che invece di regredire su ε_i^2 effettua la regressione su

$$S^2 = \frac{\sum_i \varepsilon_i^2}{n}$$

Questo meccanismo punta quasi a normalizzare i residui ε_i^2 per cui se S^2 ed ε^2 divergono significa che c'è *eteroschedasticità*.

Modello WLS per soluzioni con errori eteroschedastici.

Una volta appurata la presenza di residui eteroschedastici ed incorrelati $Cov(\varepsilon_i^*; \varepsilon_j^*) = 0$ è necessario costruire un modello per correggere questa violazione dell'ipotesi di *sfericità degli errori*. Prima di tutto dal classico modello OLS

$$y = Xb + \varepsilon^*$$

ricaviamo la varianza campionaria dei residui $S_i^2 = h_i$ e poi procediamo a dividere ogni componente per $\sqrt{h_i}$

$$y^* = \frac{y}{\sqrt{h_i}}$$
$$X^* = \frac{X}{\sqrt{h_i}}$$
$$\varepsilon = \frac{\varepsilon^*}{\sqrt{h_i}}$$

ottenendo infine un modello trasformato con

$$Var(\varepsilon) = \frac{h_i}{h_i} = 1$$

Per questa ragione il modello è chiamato **WLS** ovvero **Weighted Least Squares** ed è definito nella sua forma funzionale come:

$$y^* = X^* b + \varepsilon$$

2 Residui correlati.

Per quanto riguarda l'ipotesi di *sfericità degli errori* oltre all'*omoschedasticità* abbiamo anche l'assunzione di *incorrelazione* dei ε_i. In caso quindi di residui **correlati** abbiamo che:

$$Cov(\varepsilon_i; \varepsilon_j) \neq 0$$
$$y_i = x_i b + \varepsilon_i^\sharp$$