lezione n. 12 (a cura di Francesca Godioli)

Ad ogni categoria della variabile qualitativa si può assegnare un valore numerico che viene chiamato SCORE.

Passare dalla variabile qualitativa X2 a dei valori numerici:

(Per esempio X_2= titolo di studio) quindi il rispettivo score sarà Z_2 che è uguale a:

$Z_2 =$

1	diploma
2	scuola superiore
3	laurea triennale (laurea 3)
4	laurea di carattere superiore (laurea 5)

L'ipotesi che sta dietro questa assegnazione, che è arbitraria, è:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_{i2} + \varepsilon_i \qquad \text{dove } X_1 = \text{età}$$

Dipoma $\longrightarrow$ $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 + \varepsilon_i$ dove $Z_{i2} = 1$

Scuola superiore $\longrightarrow$ $Y_i = \beta_0 + \beta_1 X_{i1} + 2\beta_2 + \varepsilon_i$ dove $Z_{i2} = 2$

Laurea 3 $\longrightarrow$ $Y_i = \beta_0 + \beta_1 X_{i1} + 3\beta_2 + \varepsilon_i$ dove $Z_{i2} = 3$

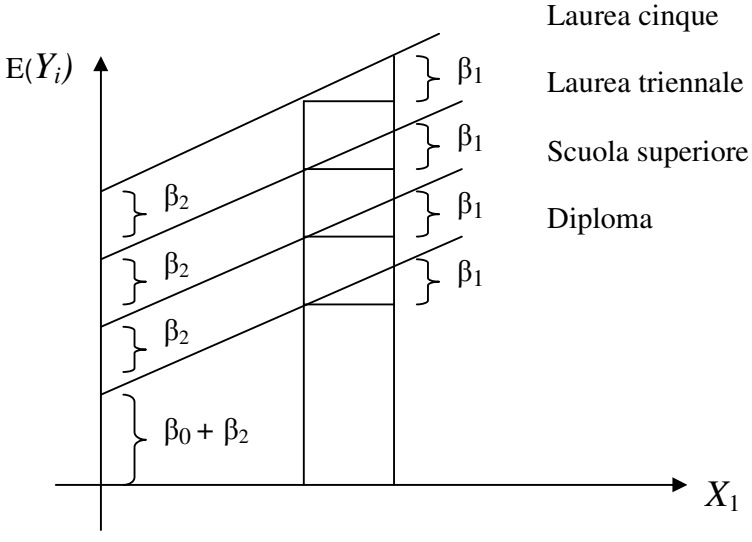
Laurea 5 $\longrightarrow$ $Y_i = \beta_0 + \beta_1 X_{i1} + 4\beta_2 + \varepsilon_i$ dove $Z_{i2} = 4$

$E(Y_i$ / Scuola superiore) - $E(Y_i$ / Dipoma) = $\beta_2 = (2\beta_2 - \beta_2)$

$E(Y_i$ / Laurea 3) - $E(Y_i$ / Scuola superiore) = $\beta_2 = (3\beta_2 - 2\beta_2)$

$E(Y_i$ / Laurea 5) - $E(Y_i$ / Laurea 3) = $\beta_2 = (4\beta_2 - 3\beta_2)$

L'incremento di reddito (Y_i) che c'è tra laurea triennale e scuola superiore è lo stesso che c'è tra scuola superiore e diploma.
Il vantaggio di questo assegnazione è che si può lavorare con un modello più parsimonioso, poiché utilizza un minor numero di parametri quindi giunge a delle conclusioni più semplici.
Lo svantaggio, invece, è di imporre delle ipotesi nel passaggio di grado che non è detto rispecchiare la realtà.



Le rette sono equidistanti (distanza pari a β_2)
Con le variabili dummy la distanza tra le rette può essere differente.

Si può verificare l'ipotesi attraverso la statistica F (test-F)

Confronto tra il modello con variabili dummy e il modello con gli score e trovare un vincolo affinché i due modelli coincidano:

DUMMY : $Y_i = \beta^*_0 + \beta^*_1 X^*_{i1} + \beta^*_2 Z^*_{i2} + \beta^*_3 Z^*_{i3} + \beta^*_4 Z^*_{i4} + \varepsilon_i$

SCORE : $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_{i2} + \varepsilon_i$

VINCOLI: *I° CASO* $(\beta^*_3 - \beta^*_2) = \beta^*_2$ *II° CASO* $(\beta^*_4 - \beta^*_3) = (\beta^*_3 - \beta^*_2)$

Nel *I° CASO* si è posto che la differenza tra laurea triennale e scuola superiore sia uguale alla differenza tra scuola superiore e diploma; nel *II° CASO* si è posto che la differenza tra laurea di livello superiore e laurea triennale sia uguale alla differenza tra laurea triennale e scuola superiore. Si esprimono i vincoli attraverso la matrice A:

$H_0 : A\underline{\boldsymbol{\beta}}^* = 0$

$\underline{\boldsymbol{\beta}}^*$(vettore con 5 elementi) = $\begin{pmatrix} \beta^*_0 \\ \beta^*_1 \\ \beta^*_2 \\ \beta^*_3 \\ \beta^*_4 \end{pmatrix}$

Risolvendo il sistema:
$$\begin{cases} (\beta^*_3 - \beta^*_2) = \beta^*_2 \longrightarrow (\beta^*_3 - 2\beta^*_2) = 0 \\ (\beta^*_4 - \beta^*_3) = (\beta^*_3 - \beta^*_2) \longrightarrow (\beta^*_4 - 2\beta^*_3 + \beta^*_2) = 0 \end{cases}$$

Quindi la corrispondente matrice A sarà:

$$A_{2x5} = \begin{pmatrix} 0 & 0 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{pmatrix}$$

$$F = \frac{(DR_{score} - DR_{dummy}) / r(A)}{DR_{dummy} / (n - \#\text{ parametri})}$$

$DR_{score} > DR_{dummy}$

Questo perché il modello con gli score utilizza meno parametri quindi ha un errore più elevato ed ha, quindi, una minore capacità predittiva.

$F \sim F(r(A), n - \#\text{ par}) \longrightarrow$ Nel nostro caso specifico F (2,5)

La scelta degli score è arbitraria, per esempio:

$Z_2 =$ 1 diploma
3 scuola superiore
4 laurea triennale (laurea 3)
4,5 laurea di carattere superiore (laurea 5)

In questo caso l'incremento più alto è tra scuola superiore e diploma (pari a $2\beta_2$) e si può inoltre notare che l'incremento è decrescente.

Modello analisi varianza

Quando un modello di regressione include solo variabili qualitative si chiama "Modello analisi varianza".

La notazione è differente se si tratta un modello di analisi ad un fattore, cioè con una sola variabile qualitativa oppure di un modello di analisi a due fattori, cioè con due variabili qualitative.

1 FATTORE $\longrightarrow$ $Y_{ij}=$ { i si riferisce al soggetto; j si riferisce alla categoria } $i= 1,2,\ldots n_j$

(dove n_j è il numero di soggetti nella categoria j)

$Y_{ij} = \mu_j + \varepsilon_{ij}$

μ_j= reddito medio, valor medio della variabile risposta, per i soggetti nella categoria j

$$\hat{\mu}_{j=} \frac{\sum_{i=1}^{n} Y_{ij}}{n_j}$$ Metodo dei minimi quadrati

2 FATTORI $\longrightarrow$ si hanno tre indici per Y dove gli ultimi due indicano le due categorie $Y_{ijk\ ;}$ dove j e k sono le due categorie.

Selezione delle variabili esplicative:

$Y \longrightarrow X_1, X_2, \ldots\ldots, X_k$

Nella realtà si trovano tantissime variabili esplicative. Il problema è selezionare le variabili veramente significative (importanti).
Attraverso un modello statistico cerco una spiegazione semplice della realtà. Si vuole spiegare la Y attraverso un modello *parsimonioso*, cioè che coinvolga poche covariate. Per la selezione vengono utilizzati vari metodi detti *stepwise.* Quest'ultimi procedono passo passo eliminando una variabile alla volta fino a giungere al modello ottimale.
I due metodi più utilizzati sono:

1) *STEPWISE-BACKWARD*: questo parte da un modello con tutte le covariate (completo) e le elimina via via, poi torna indietro ai punti precedenti.

FASI DEL METODO:

a) stimare il modello completo con k covariate e k+1 parametri β_j
b) osservare tutti i p-value per β_j (H_0: $\beta_j = 0$), j=1,2,…..k. Lo si fa per tutti i coefficienti legati alle covariate senza considerare β_0.
c) trovare il p-value maggiore tra questi, che viene indicato con p*
d) controllare che p*> 0.05, se è maggiore si può eliminare la covarita corrispondente altrimenti non si procede, ci si ferma. Se si deve procedere all'eliminazione della variabile esplicativa si torna al passo a con un k differente; questo procedimento fino a quando trovo p*< 0.05 e ci si ferma, poiché non posso eliminare più alcuna covariata essendo tutte significative per il modello come sottolineato anche dal corrispondente p-value.

Esempio

Y $\longrightarrow$ X_1, X_2, X_3

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$ (4 stime di β)

p-value β_1= 0.01 p-value β_2= 0.07 p-value β_3= 0.25

p*= 0.25 quindi si elimina la terza covariata poiché non è più importante quindi si torna a stimare il modello.

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

p-value β_1= 0.04 p-value β_2= 0.06

p*= 0.06 quindi si elimina la seconda covariata, così si giunge al modello semplice con una sola variabile esplicativa, cioè:

$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$

p-value β_1= 0.03=p* quindi ci si ferma, non si può eliminare questa covariata.

A volte eliminare una covariata corrisponde ad eliminare più parametri; questo succede con le variabili qualitative.

DUMMY: $Y_i = \beta_0 + \underbrace{\beta_1 X_{i1}}_{\text{p-value con la t-test}} + \underbrace{\beta_2 Z_{i2} + \beta_3 Z_{i3} + \beta_4 Z_{i4}}_{\text{p-value con F-test}} + \varepsilon_i$

Per introdurre la covariata Z_{i4} ho bisogno di più parametri, quindi se si vuole valutare la possibilità di eliminare la covariata devo calcolare il p-value attraverso la statistica F(test-F)

Esempio

p-value β_1= 0.03 (statistica t-test)

p-value $\beta_2 = \beta_3 = \beta_4 = 0.08$ (statistica F-test) posso eliminare l'intero blocco di covariate

2) *STEPWISE-FORWARD:*questo parte da un modello senza covariate e ne aggiunge una alla volta.
FASI DEL MODELLO:

a) stimare k modelli del tipo : $Y_i = \beta_0 + \beta_j X_{ij} + \varepsilon_i$ j=1,2,….k

b) per ogni modello osservare il p-value per β_j (H_0: $\beta_j = 0$)

c) trovare il p-value minore (p^+)

d) se $p^+ < 0.05$ significa che la covariata corrispondente a questo livello di p-value è significativa, quindi la si introduce nel modello. Si torna al passo a) andando a stimare i modelli con più covariate. Nel caso in cui $p^+ > 0.05$ ci si ferma.

Esempio

Y $\longrightarrow$ X_1, X_2, X_3

$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \longrightarrow$ p-value= 0.05

$Y_i = \beta_0 + \beta_2 X_{i2} + \varepsilon_i \longrightarrow$ p-value= 0.01 = p^+

$Y_i = \beta_0 + \beta_3 X_{i3} + \varepsilon_i \longrightarrow$ p-value= 0.03

Si inserisce la seconda covariata e si procede quindi:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \longrightarrow$ p-value= 0.06= $p^+ > 0.05$

$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \longrightarrow$ p-value= 0.08

Nessuna delle due covariate è importante quindi si ferma l'algoritmo.
Si giunge al modello ottimale che è:

$$Y_i = \beta_0 + \beta_2 X_{i2} + \varepsilon_i$$

Nel caso delle variabili dummy il p-value, anche in questo caso, è sempre quello basato sulla statistica F.

Spesso i due metodi backward e forward conducono alla stessa conclusione, selezionando le stesse covariate, ma non essendo assicurato matematicamente, se la conclusione risulta differente è preferibile che si utilizzi il metodo backward.
Nelle versioni pratiche si possono utilizzare altri metodi più complessi; per esempio un metodo è lo *STEPWISE BACK-FORWARD*: questo modello ad ogni passo aggiunge una nuova covariata ed elimina una covariata già presente nel modello.
L'obiettivo comune di tutti questi metodi, comunque, è giungere ad un modello parsimonioso.