LEZIONE N. 11 (a cura di MADDALENA BEI)

F- test
Assumiamo l’ipotesi nulla
$H_0 : \beta_1, \ldots, B_k = 0$
E’ diverso dal verificare che $H_0 : B_j = 0$
In realtà F - test è più generale
$H_0 : A\boldsymbol{\underline{\beta}} \boldsymbol{=} \underline{0}$
$H_1 : A\boldsymbol{\underline{\beta}} \neq \underline{0}$
A è una matrice di dimensioni opportune che si chiama matrice dei vincoli
Prendiamo $A = (0, \ldots, \underset{3}{1}, 0, \ldots 0)$

Se facciamo $A\boldsymbol{\beta} = \beta_2$ viene fuori un solo valore; quindi il caso più semplice è considerando A come vettore riga in cui tutti gli elementi sono uguali a zero tranne uno, uguale a 1. Se 1 è nella terza posizione otterremo β_2 perché prima vi è l’intercetta.

$\Downarrow$

$$(0, 0, 1, 0) = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \beta_2$$

Ovviamente possiamo esprimere anche forme più complesse. Potrei avere:
k=3

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\underset{2*(k+1)}{A\underline{\boldsymbol{\beta}}} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

$H_o : \beta_1, \beta_2 = 0$ $\Longrightarrow$ Non va verificata con t bensì con F - Fisher e col caso generale
Posso dire che alcuni coefficienti sono congiuntamente uguali a zero
Ipotizzo:
k=3 $\qquad A = (0 \;\; -1 \;\; 1 \;\; 0)$
$A \underline{\boldsymbol{\beta}} = (\beta_2 - \beta_1)$

$\Uparrow$
Qui in corrispondenza ho 1
$H_0 : \beta_1 = \beta_2$

Avremo che il coefficiente della prima covariata è uguale al coefficiente della seconda covariata
$H0 : A\underline{\boldsymbol{\beta}} = \underline{\mathbf{0}}$

1) Singolo parametro quindi (r (A) =1)
2) Blocco di parametri (r (A) > 1)
3) $\Longrightarrow$ Contrasto (r (A) = 1) nella stessa riga ho il vettore i cui valori sommano zero (tipicamente ho una riga)

Come si fa a verificare H_0?

1) Si stima il modello senza vincolo (MODELLO SVINCOLATO) con tutte le covariate e memorizzo la DR_1 (devianza residua) . Il modello è così espresso:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

2) Si stima il modello con il vincolo (MODELLO VINCOLATO) che è l' ipotesi nulla H_0.

$Yi = \beta_0 + \beta_3 X_{i3} + \varepsilon_i$ (se riguarda il secondo caso)

Una volta stimata ottengo DR_0 .

A questo punto posso calcolare la statistica F

$$Foss = \frac{(DR_0 - DR_1)r(A)}{\underbrace{DR_1 / (n-(k+1))}_{S^2 \text{ svincolato}}}$$

F oss = F osservato

Ci dice quanto è l'errore senza covariate.

Ipotizzo un modello pieno

Modello vincolato ha $DR_0 > DR_1$ dove DR_1 rappresenta il modello svincolato

Va ricordato che :

DR = misura l'errore di previsione del modello

r (A) = dimensione di A uguale al numero dei vincoli

Stimo il modello con e senza ipotesi e ricapitolando avremo che il modello svincolato usa tutte le covariate, il vincolato né toglie alcune con effetto sulla DR che sarà più piccola

Vediamo come la F ha distribuzione

$F \sim F(r(A), n-(k+1))$ sotto H_0

Se è vera avrà valori abbastanza piccoli

SE $F \geq F_\alpha \Longrightarrow$ Allora si rifiuta H_0 (RH_0) $\Longrightarrow$ Allora almeno la prima o seconda covariata è utile per vedere la reale influenza di y

Anche qui posso calcolare il p - value = $p(F \geq F_{oss})$

Allora posso verificare diverse teorie:

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

$H_0 : \beta_1 = \beta_= \ldots = B_k = 0$ che dice che nessuna covariata è utile per spiegare la y

Pongo

$$\underset{K+(k+1)}{A} = \begin{pmatrix} 0 & 1 & 0 \ldots 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \ldots 1 \end{pmatrix}$$

Per verificare le ipotesi

1)Stimare modello svincolato DR1

2)Stimare modello vincolato (H_0) che è $Yi = \beta_0 + \varepsilon_i \Longrightarrow \beta_0 = \bar{Y} \Longrightarrow DR_0 = \sum_i (Yi - \hat{Y}_i)^2 = \sum_i (Yi - \bar{Y})^2 =$ Dev (Y)

Rimane solo β_0 gli sono tutti uguali a zero

Il modello diventa semplice. DR è UGUALE ALLA Dev (Y) cioè il valore previsto per ogni soggetto è uguale alla Y

$$Foss = \frac{(DR_0 - DR_1) r(A)}{DR_1 / (n-(k+1))}$$

$DR_0 = Dev(Y)$

$$Foss = \frac{(Dev(Y) - DR_1)/k}{DR_1 / (n-(k+1))} = \frac{DS/k}{DR_1 / (n-(k+1))}$$

Questo è come verificare $H_0 : A\underline{\beta}=0$

Quindi abbiamo tre diverse ipotesi:

- un singolo coefficiente uguale a zero ;
- un blocco di coefficienti uguali a zero;
- due coefficienti uguali tra loro

Fin' ora abbiamo scritto il modello con le covariate in x e ci aspettiamo che siano variabili quantitative. Nella pratica ci interessa inserire il modello con covariate qualitative nel modello della regressione multipla

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots\ldots + \beta_k x_{ik} + \varepsilon_i$$

Y= reddito (variabile risposta)
X_1 = anni di istruzione (prima covariata)
X_2= sesso (seconda covariata)

Inserisco variabili dummy (sono sempre uguale a 0 o ad 1 e permettono di introdurre una variabile qualitativa). Creo:

$$Z_2 = \begin{cases} 0 \text{ se } X_2 = M \\ 1 \text{se } X_2 = F \end{cases}$$

Posso riscrivere il modello come

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

Ottengo tre coefficienti:

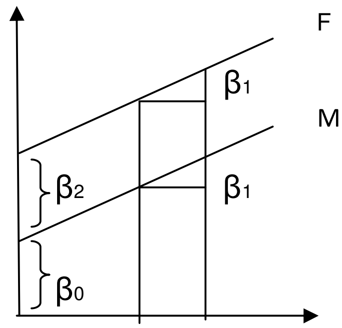
- $\beta_0 = E(Y_i)$ se $x_{i1} = 0$ $z_{i2} = 0$
  β_0 è il reddito di un maschio con $x_1 = 0$, è il valore atteso se tutte le covariate sono uguali a 0. In particolare qui se $\beta_0 = 20$ significa che un maschio a 18 anni ha un reddito atteso di 20.000 euro.

- $\beta_1 = E(Y_i)|\ x_{i1} = x+1) – E(Y_i)|\ x_{i1} = x)$
  β_1 è l'incremento del reddito atteso se la x cresce di 1. All'aumentare di x , β_1 cresce sia per i maschi che per le femmine. $\beta_1 = 2$

- $\beta_2 = E(A_i \mid F) – E(Y_i|M)$
  Se calcolo il reddito per una femmina avremo :
  $\cancel{\beta_0} + \cancel{\beta_1} x_{i1} + \beta_2 – (\cancel{\beta_0} + \cancel{\beta_1} x_{i1})$
  β_2 è la differenza nel reddito atteso fra femmine e maschi a parità di età. Se $\beta_2 = 2$ a parità di condizioni una donna guadagna più di un uomo.

M: $\beta_0 + \beta_1 x_{i1} + \varepsilon_i$
F: $\beta_0 + \beta_1 x_{i1} + \beta_2 + \varepsilon_i = (\beta_0 + \beta_2) + \beta_1 x_{i1} + \beta_2 + \varepsilon_i$

Cambia solo l'intercetta.



β_0 è il reddito di un maschio ; β_1 comprende sia i maschi che le femmine ; β_2 è la differenza tra le due rette.
L'introduzione di dummy permette di prendere in considerazione:

1) la variabile qualitativa
2) β_2
3) Due rette parallele

Come si verifica dando al modello $\beta_2 = 0$, cioè non consideriamo l'effetto sesso.
H0: $\beta_2 = 0$ lo possiamo fare con t-stat

Quando abbiamo le rette parallele significa che l'incremento della covariata ha lo stesso effetto su Y (reddito).
L'ipotesi più attendibile è che il reddito cresca in modo diverso.
Possiamo scrivere:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 z_{i2} + \underbrace{\beta_3\, x_{i1}\, z_{i2}}_{} + \varepsilon_i$$

TERMINE DI INTERAZIONE (inferenza fra le due covariate)

Diventa:
M: $Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$ $\qquad z_{i2} = 0$

F : $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 + \beta_3\, x_{i1} + \varepsilon_i$

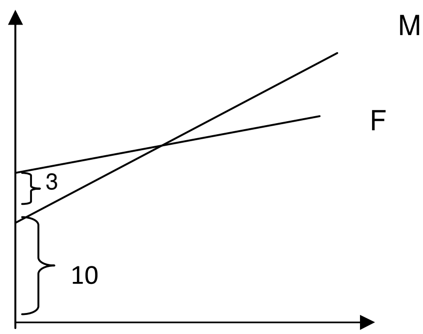
$$(\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x_{i1} + \varepsilon_i$$

(Ho creato una sola intercetta)

E' indispensabile scriverlo così per fare una stima.

- $\beta_0 = 10$ è il reddito atteso per un maschio a 18 anni
- $\beta_1 = 2$ è l' incremento di reddito di un maschio quando l'età aumenta di uno
- $\beta_2 = 3$ è la differenza fra reddito delle femmine e dei maschi a 18 anni ; è l'intercetta

- $\beta_3 = -1$ è la differenza fra femmine e maschi nell'incremento del reddito quando aumenta un anno di età. E' la differenza nel ritmo di crescita e coincide col coefficiente angolare



Se vogliamo verificare che non c'è differenza fra i maschi e le femmine dobbiamo porre
$H_0 = \beta_2 = \beta_3 = 0$
Dobbiamo usare la F-test (ipotesi su più parametri)

r(A)=2

Caso della variabile qualitativa con più livelli

Prendiamo X_2 con c categorie allora devo introdurre $\longrightarrow$ c-1 variabili dummy

Precisiamo che il sesso ha due categorie

$$X_2 = \text{titolo di studio} \begin{cases} \text{Licenza media} \\ \text{Superiore} \\ \text{Laurea triennale} \\ \text{Laurea quinquennale} \end{cases}$$

In questo caso ho c=4 quindi
3 variabili dummy

$$z_2 = \begin{cases} 0 & \text{altrimenti} \\ 1 & \text{superiore} \end{cases}$$

$$z_3 = \begin{cases} 0 & \text{altrimenti} \\ 1 & \text{laurea triennale} \end{cases}$$

$$z_4 = \begin{cases} 0 & \text{altrimenti} \\ 1 & \text{laurea quinquennale} \end{cases}$$

Dobbiamo escludere la prima (la licenza media in quanto non c'è bisogno di inserirla).
Per ogni titolo di studio ho una certa configurazione di dummy

	z_2	z_3	z_4
Media	0	0	0

Superiore	1	0	0
Laurea 3	0	1	0
Laurea 5	0	0	1

Ora si tratta di scrivere il modello di regressione lineare in una equazione che sia stimabile.

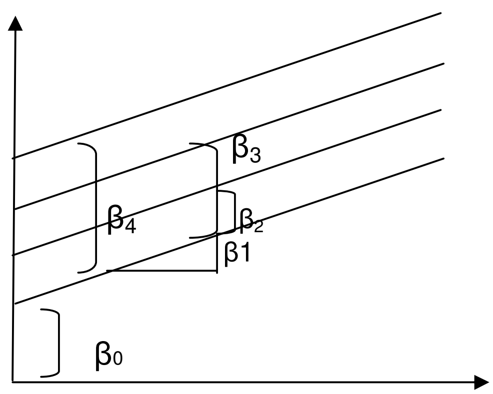
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4} + \varepsilon_i$$

Bisogna capire come diventa l'equazione:

media	$Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$
Superiore	$Y_i = (\beta_0 + \beta_2) + \beta_1 x_{i1} + \varepsilon_i$
Laurea triennale	$Y_i = (\beta_0 + \beta_3) + \beta_1 x_{i1} + \varepsilon_i$
laurea quinquennale	$Y_i = (\beta_0 + \beta_4) + \beta_1 x_{i1} + \varepsilon_i$

Le medie sono la categoria di riferimento; tutte le dummy sono uguali a zero; tutti i coefficienti che aggiungo li confronto con le dummy

Caso in cui le rette sono parallele:



Se ipotizzo che non c'è effetto sul titolo di studio

H_0: $\beta_2 = \beta_3 = \beta_4 = 0$ (TEST F con la matrice di contrasto)

H_0: $\beta_3 = \beta_4$ implica che non c'è differenza fra laurea triennale e quinquennale e quindi le due rette coincidono
Lo posso formulare con A = (0 0 0 -1 1)