LEZIONE n. 5 (a cura di Antonio Di Marco)

IL P-VALUE (α)

Data un'ipotesi nulla (H_0), questa la si può accettare o rifiutare in base al valore del p-value. In genere il suo valore è un numero molto piccolo, vicino allo zero. Volendo dare una definizione, si può dire che:

Definizione: è il livello di significatività assegnato, ossia una misura di evidenza contro l'ipotesi nulla.

Qual è la sua interpretazione?

Assegnato un valore soglia, in genere 0,05 si ha:

1) $\alpha < 0{,}05$ rifiuto H_0;
2) $\alpha \geq 0{,}05$ non rifiuto H_0.

Quindi più piccolo è il p-value, tanto maggiore è l'evidenza contro l'ipotesi nulla.

Come si calcola il p-value?

Esso si calcola utilizzando la normale standard.

$H_0 : \mu=\mu_0$ $\quad X \sim (\mu, \sigma^2)$ $\quad$ con σ^2 nota

$H_1 : \mu>\mu_0$ $\quad z = \dfrac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$ $\quad$ p-value = $P(Z \geq z)$ $\longrightarrow$ Probabilità di osservare un campione che contrasta ancora di più con H_0

$H_1 : \mu<\mu_0$ $\quad$ p-value = $P(Z \leq z)$

$H_1 : \mu\neq\mu_0$ $\quad$ p-value = $P(|Z| \geq |z|)$

Esempio di interpretazione del p-value

$H_0 : \mu=\mu_0$ $\quad H_1 : \mu\neq\mu_0$

p-value = 0,02 $\quad$ significa che è più plausibile l'ipotesi alternativa

VERIFICA DELLE IPOTESI (CONFRONTO TRA VARIANZE)

Immaginiamo di avere due popolazioni, indicheremo con X una variabile di interesse per la prima popolazione e con Y quella relativa alla seconda popolazione:

$P_1 : X \sim (\mu_1, \sigma_1^2)$ → ad esempio altezza di un umbro

$P_2 : X \sim (\mu_2, \sigma_2^2)$ → ad esempio altezza di un lombardo

$H_0 : \sigma_1^2 = \sigma_2^2$ → se nella verifica delle ipotesi riferita alla media, avevamo imposto l'uguaglianza tra le due varianze, ora invece, andiamo a verificare se esse sono uguali e se lo sono possiamo applicare il test sulla media.

$H_1 : \sigma_1^2 > \sigma_2^2$

$H_1 : \sigma_1^2 < \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

Uno stimatore tipico della varianza è S:

$$S_1^2 = \frac{1}{n1-1} \Sigma i (xi - \bar{x})2$$

$$S_2^2 = \frac{1}{n2-1} \Sigma i (yi - \bar{y})2$$

Un modo per fare il confronto tra le due varianze, oltre alla differenza è quello del rapporto tra le due, infatti se il valore è maggiore di uno significa che $S_1^2 > S_2^2$, se minore dell'unità è il contrario, infine se uguale ad uno, esse sono uguali:

$F = S_1^2 / S_2^2 \sim F(n_1 - 1, n_2 - 2)$ sotto ipotesi H_0

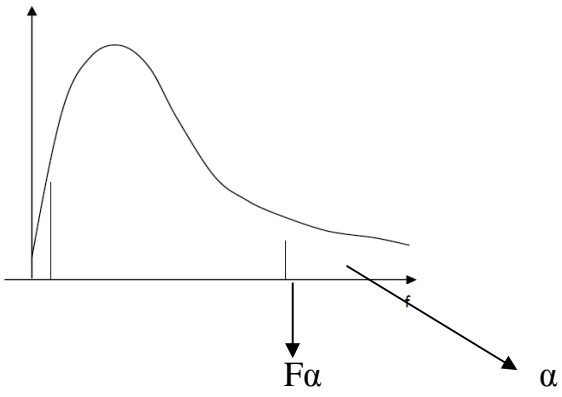
Se H_0 fosse vera tale statistica avrebbe distribuzione di Fisher.

Quando si rifiuta?

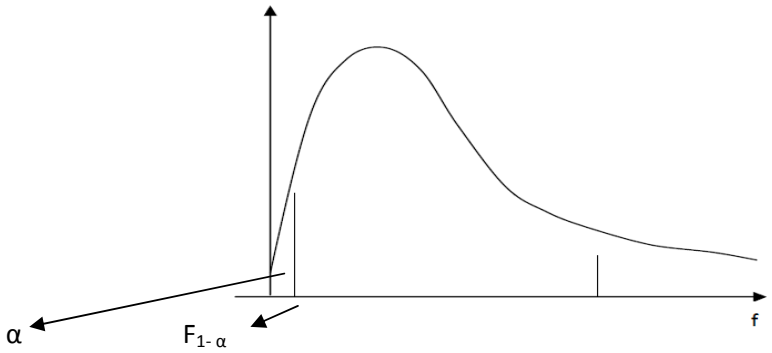
$H_1 : \sigma_1^2 > \sigma_2^2$ → se $F \geq F_\alpha$ si rifiuta H_0

↓

quantile della F – Fisher



$H_1 : \sigma_1^2 < \sigma_2^2$ → se $F \leq F_{1-\alpha}$ si rifiuta H_0



$H_1 : \sigma_1^2 \neq \sigma_2^2$ → se $F \leq F_{1-\alpha/2}$ oppure $F \geq F_{1-\alpha/2}$ si rifiuta H_0

REGRESSIONE LINEARE SEMPLICE

Perche si definisce lineare e semplice? Perché c'è una funzione lineare e una sola x.

Definizione del concetto:

immaginiamo di avere n soggetti e in corrispondenza dei quali osserviamo una variabile x ed una variabile y:

$x_i \rightarrow$ valore della variabile in corrispondenza del soggetto i (ad esempio anni di istruzione post scuola dell'obbligo

$y_i \rightarrow$ valore della variabile in corrispondenza del soggetto i (ad esempio reddito) con i= 1,2,3,……..,n

1) sulla base di tali osservazioni voglio capire come x influenza y;
2) voglio fare delle previsioni, ossia poter dire ad una persona, la quale intende studiare due anni in più, se il suo reddito incrementa o meno.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$\beta_0 + \beta_1 x_i$: parte deterministica

ε_i: parte casuale

Cos'è y_i? è una variabile casuale che è data dalla somma della parte deterministica del modello (la quale dice che la y è funzione lineare di x), più la parte casuale (stocastica), la quale rappresenta l'errore, cioè la differenza tra la parte casuale e quella deterministica.

La parte deterministica coinvolge 2 parametri:

$\beta_0 \rightarrow$ è l'intercetta

$\beta_1 \rightarrow$ è il coefficiente angolare

Qual è l'interpretazione di β_0 ? è il valore atteso di y se x=0: $\beta_0 = E(y_i|x_i=0)$

Qual è l'interpretazione di β_1 ? è l'incremento del valore atteso di y se x cresce di 1:

$\beta_1 = E(y_i \mid x_i = x+1) - E(y_i \mid x_i = x)$

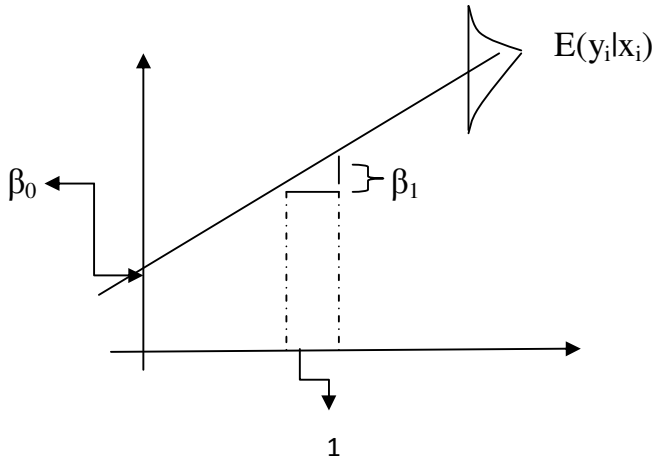
Esempio

x = numero anni di istruzione dopo la scuola dell'obbligo

y = reddito

Se il modello dice che $\beta_0 = 15$ significa che una persona, la quale ha completato solo la scuola dell'obbligo guadagna **intorno** a 15000€, il che è ben diverso dal dire, che tutti guadagno 15000 €, questo perché c'è la componente di errore.

Se il modello dice che $\beta_1 = 3$ significa che ogni anno di istruzione in più, permette di far guadagnare 3000€ in più.



consiste nei valori che possono oscillare in questa campana, attorno al valore atteso e si distribuisce secondo una certa distribuzione dell'errore.

Qual è l'interpretazione della parte stocastica ε_i***?*** è una variabile casuale che possiede le seguenti proprietà:

1) $E(\varepsilon_i) = 0$, $\forall$ i $\rightarrow$ rispetto al valore atteso posso osservare variazioni in più o in meno, che vanno a compensarsi
2) $V(\varepsilon_i) = \sigma^2$, $\forall$ i $\rightarrow$ tale assunzione è una semplificazione perché nella realtà non è così
3) ε_i indipendente da ε_i , $\forall$ i $\rightarrow$ l'errore, rispetto al modello coinvolto nell'osservazione di quello di una persona non influenza quello di un'altra persona

Come si calcola $E(y_i)$?

$E(y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i) = \beta_0 + \beta_1 x_i$

Come si calcola $V(y_i)$?

$V(y_i) = V(\beta_0 + \beta_1 x_i + \varepsilon_i) = V((\varepsilon_i) = \sigma^2$

Come si applica tale modello?

Nella realtà si ha una serie di osservazioni:

x	y
x_1	y_1
x_2	y_2
.	.
.	.
x_n	y_n

→ sulla base di tali dati devo stimare β_0 e β_1 (assegnare loro dei valori plausibili)

Per fare ciò il metodo più semplice è quello dei **MINIMI QUADRATI.**

METODO DEI MINIMI QUADRATI

Se ho un certo valore di xi posso calcolare la previsione di yi:

$\hat{y}_i = \beta_0 + \beta_1 x_i$ → se conoscessi β_0 e β_1 e mi viene detto il valore della x, potrei fornire un valore plausibile della y

ad esempio una persona ha studiato per 10 anni dopo la scuola dell'obbligo:

$\hat{y_i}$= 15+ 3*10=45 ossia 45000€ di reddito

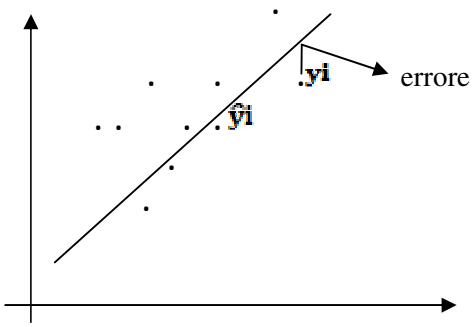
la previsione è soggetta ad un certo margine di errore che è quantificabile in $y_i - \hat{y_i}$

se volessi una misura di errore complessiva, dovrei fare:

$S = \sum (y_i - \hat{y_i})^2$ → misura complessiva di errore del modello, la quale è funzione di β_0 e β_1

Per stimare i parametri dovrò calcolare il minimo, questo perché voglio il modello col più basso valore di errore:

$\min_{\beta 0\, \beta 1} S(\beta_0, \beta_1) \rightarrow \hat{\beta}_0, \hat{\beta}_1$



errore di previsione (distanza di ogni punto dalla retta)

Se S è piccolo, implica che la retta passa quasi per tutti i punti, quindi è stata fatta una buona approssimazione.

$\hat{\beta}_0 = \frac{codevianza\ (x,y)}{devianza\ (x)}$

codevianza (x,y) = $\sum (xi - \bar{x})(yi - \bar{y})$

devianza (x) = $\sum xi - \bar{x}$ → è sempre positiva

$\hat{\beta}_1 = \bar{y} - \hat{\beta}_{1*}\bar{x}$

$\frac{\Sigma xi}{n}$ $\frac{\Sigma yi}{n}$