

Studio della dipendenza: la regressione

Fino ad ora abbiamo studiato delle relazioni *simmetriche* fra due caratteri. Spesso invece siamo interessati a valutare se una variabile può essere **spiegata** e/o **prevista** da un'altra variabile (relazione *direzionale*).

Per esempio sono l'associazione agricoltori e sono interessato a valutare se la quantità di riso raccolto (variabile **DIPENDENTE**) può essere spiegata dalla quantità di pioggia (variabile **INDIPENDENTE** o **ESPLICATIVA**)

Oppure sono una banca voglio prevedere il rischio di credito (variabile **DIPENDENTE**) di un'impresa che mi chiede un prestito sulla base di indicatori di bilancio (più variabili **INDIPENDENTI** o **ESPLICATIVE**) o prevedere il rischio di insolvenza di un cliente (variabile **DIPENDENTE**) sulla base delle caratteristiche del cliente stesso (età , numero componenti famiglia, etc. ...) e di indici ricavati della sua dichiarazione dei redditi (variabili **INDIPENDENTI**)

Nell'ambito dello studio della regressione vedremo:

- la *funzione o spezzata di regressione*
(funzione non analitica ma disegnata per punti)
- la *retta di regressione*
(funzione analitica con un'espressione)

Notazione

Variabile **DIPENDENTE**:

Y

Variabile **INDIPENDENTE** o **ESPLICATIVA**:

X

vedremo solo il caso di una sola variabile esplicativa. Nel corso di previsione economica si considereranno anche più variabili esplicative

La funzione di regressione

si può costruire solo se **almeno uno** dei due caratteri oggetto di studio è **quantitativo**. Indichiamo con Y il carattere quantitativo che solitamente coincide anche con il carattere che vogliamo andare a spiegare (variabile dipendente).

La funzione di regressione si ottiene

1. costruendo tutte le h distribuzioni subordinate di $Y|X = x_i$ per $i = 1, \dots, h$
2. calcolando i valori attesi di queste distribuzioni subordinate $E(Y|X = x_i)$ per $i = 1, \dots, h$
3. mettendo su un grafico i punti con ascissa uguale ad x_i ed ordinata corrispondente uguale ad $m_Y(x_i) = E(Y|X = x_i)$ per $i = 1, \dots, h$

Quindi la **funzione di regressione di Y su X** è formata da h punti di coordinate

$$[x_i, E(Y|X = x_i)]$$

ovvero

$$[x_i, m_Y(x_i)]$$

e permette di fare previsioni per Y solo in corrispondenza dei valori di X già osservati.

Quindi la previsione di Y per un soggetto che ha $X = x_i$ è $E(Y|X = x_i)$

In modo simile si costruisce la **funzione di regressione di X su Y** (in questo caso X deve essere quantitativo) ed è formata da k punti di coordinate

$$[y_j, E(X|Y = y_j)]$$

ovvero

$$[y_j, m_X(y_j)]$$

e permette di fare previsioni per X solo in corrispondenza dei valori di Y già osservati.

Quindi la previsione di X per un soggetto che ha $Y = y_j$ è $E(X|Y = y_j)$.

La previsione è tanto migliore quando minore è la variabilità delle distribuzioni subordinate

Se le medie delle condizionate di Y data X variano molto al variare delle modalità di X allora la funzione di regressione di Y su X fornisce delle previsioni su Y migliori di quelle che sarebbero fornite considerando più semplicemente la media di Y

La funzione di regressione è la funzione di (X) che minimizza l'errore quadratico medio di previsione.

Se cioè voglio andare a prevedere la variabile dipendente Y sulla base di una funzione S della variabile indipendente o esplicativa X : $S(X)$ posso scegliere la funzione S in modo che sia minimo l'errore quadratico medio di previsione

$$\text{Min}_S E[Y - S(X)]^2$$

$$\begin{aligned} E[Y - S(X)]^2 &= \sum_i \sum_j [y_j - S(x_i)]^2 f(x_i, y_j) = \\ &= \sum_i \sum_j [y_j - S(x_i)]^2 f(y_j | X = x_i) f(x_i) \end{aligned}$$

per la proprietà di minimo della media aritmetica (la media minimizza la somma degli scarti

da un polo al quadrato) ricaviamo che l'errore quadratico medio è minimo quando

$$S(x_i) = E[Y|X = x_i]$$

Se la funzione di regressione di Y su X è COSTANTE cioè se

$$E(Y|X = x_i) = \text{costante per ogni valore di } i$$

significa che la previsione di Y non varia al variare di X quindi X non contribuisce in alcun modo alla spiegazione di Y . Si dice allora che Y è **REGRESSIVAMENTE INDIPENDENTE** o **INDIPENDENTE in MEDIA** da X

Si dimostra che il valore della costante è pari alla media di Y e che la media della funzione di regressione di Y su X , $E[E(Y|X = x_i)]$, è uguale alla media di Y , $E(Y)$ cioè

$$E[E(Y|X = x_i)] = E(Y)$$

Quindi se Y è regressivamente indipendente da X allora i punti della funzione di regressione di Y su X si trovano su una retta parallela all'asse delle ascisse ad altezza pari a $E(Y)$

In modo simile, se la funzione di regressione di X su Y è COSTANTE cioè se

$$E(X|Y = y_j) = \text{costante per ogni valore di } j$$

allora X è **REGRESSIVAMENTE INDIPENDENTE** o **INDIPENDENTE in MEDIA** da Y

Si dimostra che il valore della costante è pari alla media di X e che la media della funzione di regressione di X su Y , $E[E(X|Y = y_j)]$, è uguale alla media di X , $E(X)$ cioè

$$E[E(X|Y = y_j)] = E(X)$$

Quindi se X è regressivamente indipendente da Y allora i punti della funzione di regressione di X su Y si trovano su una retta parallela all'asse delle ascisse ad altezza pari a $E(X)$

L'indipendenza regressiva **NON è un concetto SIMMETRICO** cioè se Y è regressivamente indipendente da X non è detto che X sia regressivamente indipendente da Y .

L'indipendenza statistica fra X ed Y implica l'indipendenza regressiva sia di Y su X ed di X su Y .

Non è vero il contrario: l'indipendenza regressiva non implica l'indipendenza statistica

Se almeno una delle due variabili è regressivamente indipendente dall'altra allora le due variabili sono correlativamente indipendenti.

Non è vero il contrario: l'indipendenza correlativa non implica quella regressiva

ESERCIZIO

Per una popolazione costituita da donne coniugate, si considerino i due caratteri

X = numero di figli delle madri delle donne coniugate,

Y = numero di figli delle donne coniugate,

aventi distribuzione congiunta data dalla seguente tabella:

$X \backslash Y$	0	1	2
1	0.2	0.1	0.1
2	0.1	0.2	0.1
3	0.05	0.1	0.05

Determinate e rappresentate graficamente la funzione di regressione di Y su X .

RISPOSTA: devo calcolare i seguenti valori attesi:

$$E(Y|X = 1), \quad E(Y|X = 2), \quad E(Y|X = 3)$$

Per fare questo è bene prima determinare le 3 distribuzioni subordinate e poi calcolare la media delle stesse:

$$(Y|X = 1) = \begin{cases} 0 & 1 & 2 \\ 0.5 & 0.25 & 0.25 \end{cases}$$

quindi $E(Y|X = 1) = 0.75$

$$(Y|X = 2) = \begin{cases} 0 & 1 & 2 \\ 0.25 & 0.5 & 0.25 \end{cases}$$

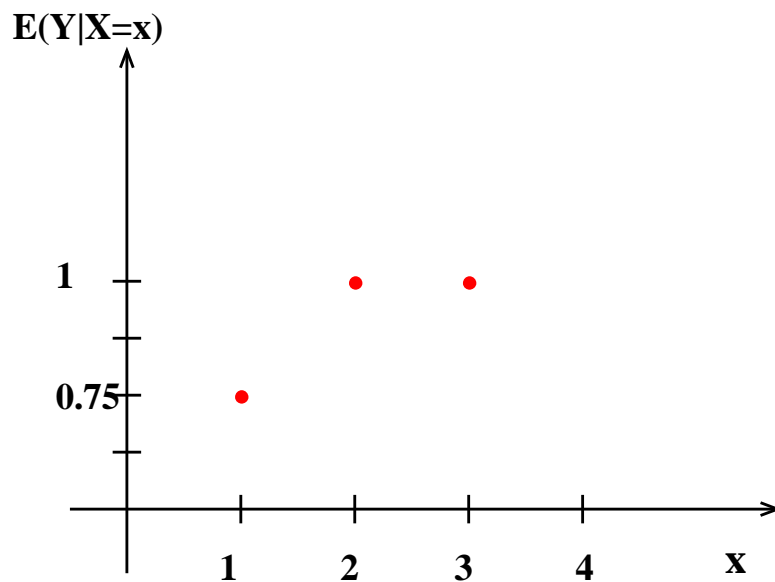
quindi $E(Y|X = 2) = 1$

$$Y|X = 3 = \begin{cases} 0 & 1 & 2 \\ 0.25 & 0.5 & 0.25 \end{cases}$$

quindi $E(Y|X = 3) = 1$.

La rappresentazione grafica della funzione di regressione è riportata in figura. Si tratta di una funzione disegnata per punti di coordinate:

$$[x_i, E(Y|X = x_i)]$$



La retta di regressione

la funzione di regressione, non essendo una funzione analitica permette di fare previsioni solo in corrispondenza dei valori di X osservati.

Cerco allora un livello di **sintesi** maggiore costruendo una funzione analitica $Y = f(X)$ che sintetizzi “al meglio” la nuvola di punti.

Fra tutte le possibili funzioni analitiche di X , $S(X)$, mi limito a considerare le **rette**, quindi cerco una funzione lineare delle osservazioni sulla variabile esplicativa del tipo:

$$S(X) = \beta_0 + \beta_1 X$$

e scelgo l'intercetta a e la pendenza β_1 in modo da minimizzare l'errore quadratico medio, applichiamo cioè il criterio dei **MINIMI QUADRATI**

La retta di regressione o interpolante lineare

Il criterio dei minimi quadrati consiste nel minimizzare, rispetto ad a (intercetta) e β_1 (pendenza), la media degli scostamenti al quadrato fra valori osservati y_j e valori previsti $\beta_0 + \beta_1 x_i$.

Si minimizza cioè la quantità

$$MIN_{(\beta_0, \beta_1)} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Per trovare il minimo di una funzione si prende la derivata prima e la si uguaglia a zero.

In questo caso ho una funzione a due variabili (β_0 e β_1) quindi devo fare due derivate, una rispetto ad β_0 ed una rispetto a β_1 ed eguaglio entrambe a zero.

Il risultato è :

$$\hat{\beta}_0 = E(Y) - \beta_1 E(X) \quad \text{e} \quad \hat{\beta}_1 = \frac{\text{Cov}(X, Y)}{V(X)} = \frac{C_{X,Y}}{D_X}$$

L'errore quadratico medio della **retta** di regressione sarà maggiore dell'errore quadratico medio della **funzione** di regressione

Immaginiamo di aver rilevato, per i 20 studenti della Facoltà di Economia il numero di **mesi** (Y) impiegato per laurearsi ed il **voto** (X) ottenuto all'esame di maturità (in sessantesimi):

$$Y = 48; 52; 60; 48; 72; 56; 76; 54;$$

$$58; 50; 84; 48; 50; 52; 56; 60; 54; 68; 52; 66$$

$$X = 60; 57; 46; 50; 40; 54; 40; 57; 50; 60; 38; 54$$

$$60; 57; 50; 40; 46; 38; 40; 46$$

abbiamo che:

$$E(Y) = 58.2, V(Y) = 97.16$$

$$E(X) = 49.15, V(X) = 60.03$$

$$Cov(X, Y) = -60.66$$

$$\hat{\beta}_1 = \frac{Cov(X, Y)}{V(X)} = -0.96 \text{ e}$$

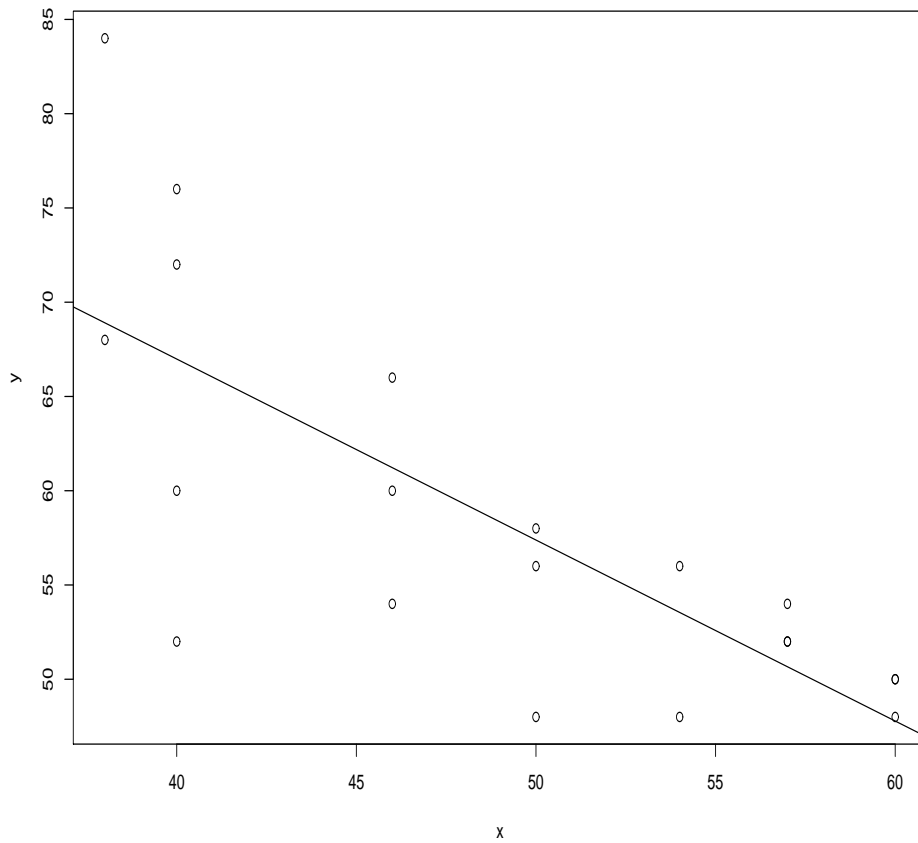
$$\hat{\beta}_0 = E(Y) - \hat{\beta}_1 E(X) = 105.39$$

La retta di regressione ha quindi equazione:

$$Y = 105.39 - 0.96X$$

Il segno di β_1 dipende dal segno della covarianza. Poichè β_1 è negativo deduciamo che la correlazione fra i due caratteri oggetto di studio è negativa cioè i due caratteri sono **discordanti**

Rappresentiamo su un grafico la nuvola di punti (diagramma di dispersione) e la retta di regressione:



Scomposizione della varianza

risulta che:

$$\sum_{i=1}^n [y_i - E(Y)]^2 = \sum_{i=1}^n [y_i - \hat{y}_i]^2 + \sum_{i=1}^n [\hat{y}_i - E(Y)]^2$$

$$D_Y = D_{RL} + D_{SL}$$

dove $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ e $\hat{\beta}_0, \hat{\beta}_1$ sono i valori dell'intercetta e della pendenza nella retta di regressione ottenuti sulla base delle osservazioni.

Coefficiente di determinazione

$$R^2 = \rho^2 = \left[\frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \right]^2$$

si dimostra che

$$R^2 = \frac{D_{SL}}{D_Y} = 1 - \frac{D_{RL}}{D_Y}$$

inoltre

$$0 \leq R^2 \leq 1$$

$R^2 = 1$ se c'è una perfetta relazione lineare fra X ed Y

$R^2 = 0$ se X ed Y sono correlativamente indipendenti e quindi la retta di regressione è una costante

R^2 rappresenta la quota della varianza totale spiegata dal modello. Tanto maggiore è il valore di R^2 tanto migliori (prive di errore) saranno le mie previsioni