REGRESSIONE E CORRELAZIONE

Nella Statistica, per studio della connessione si intende la ricerca di eventuali relazioni, di dipendenza ed interdipendenza, intercorrenti tra due variabili statistiche[1]. Esso prende il nome di **regressione** se lo scopo è quello di ricercare legami di dipendenza, di **correlazione** se lo scopo è quello di evidenziare tra le variabili legami di interdipendenza. Se X e Y sono le due variabili statistiche ed (x_i, y_i) le coppie di valori per esse osservati, riportando i valori in una tabella semplice entrata (composta da due sole colonne) e tracciando il diagramma a dispersione, si può capire il tipo di funzione idonea a rappresentare la distribuzione iniziale.

Lo studio della regressione consiste, appunto, **nel determinare**, **con il metodo dei minimi quadrati**[2], **una funzione, detta funzione di regressione**, **che ci permetta di valutare le variazioni della Y al variare della X e viceversa. Se la funzione prescelta è la retta si parlerà di regressione lineare.** Eventuale assenza di regressione lineare non significa assenza di regressione perché le due variabili potrebbero dipendere secondo una parabola, una iperbole, ecc.

RETTE DI REGRESSIONE

La retta, specie se i dati sono numerosi, è la funzione matematica più utilizzata per esprimere la dipendenza tra X e Y. Dette:

$\boldsymbol{y = a_1 + b_1 x}$ e $\boldsymbol{x = a_2 + b_2 y}$ le due rette di regressione si determinano, con il metodo dei minimi quadrati, le incognite a_1, b_1, a_2, b_2 mediante le formule seguenti:

$$\boldsymbol{b_1 = \frac{\sum_{i=1}^{n} x_i' y_i'}{\sum_{i=1}^{n} (x_i')^2}}; \quad \boldsymbol{a_1 = \overline{y} - b_1 \overline{x}}; \quad \boldsymbol{b_2 = \frac{\sum_{i=1}^{n} x_i' y_i'}{\sum_{i=1}^{n} (y_i')^2}}; \quad \boldsymbol{a_2 = \overline{x} - b_2 \overline{y}};$$

[1] Si definisce **variabile statistica** X un insieme di valori, rilevati per un carattere di tipo quantitativo (altezza, peso, cilindrata di un'automobile, numero di vani di un'abitazione ecc.) associati ad un insieme di frequenze assolute o relative (si ottengono dividendo ogni frequenza assoluta per il totale delle frequenze). Se i valori sono espressi da singoli numeri, si parla di variabile statistica a carattere discreto, se sono espressi in classi si parla di variabile statistica a carattere continuo. Se al posto dei valori troviamo degli attributi (nomi, anni, ecc.) si parla di **mutabile statistica**.

[2] **Il metodo dei minimi quadrati** è un procedimento di calcolo approssimato che consente di determinare una funzione rappresentativa di un fenomeno di natura qualsiasi. La funzione si ricava trovando dei parametri $a, b, c, \cdots, k$ che rendono minima la somma dei quadrati delle differenze tra i valori y_i osservati e quelli $\hat{y}_i$ teorici.

dove $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ ed $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ sono le medie aritmetiche dei valori iniziali x_i ed y_i. Entrambe le rette devono passare per il punto $(\bar{x}, \bar{y})$, baricentro della distribuzione, e possono essere scritte anche nella forma:

$$y - \bar{y} = b_1(x - \bar{x}) \qquad x - \bar{x} = b_2(y - \bar{y})$$

Le differenze $x_i' = x_i - \bar{x}$ ed $y_i' = y_i - \bar{y}$ sono gli scarti da $\bar{x}$ ed $\bar{y}$ rispettivamente di x_i ed y_i; b_1 (coefficiente angolare della retta $y = a_1 + b_1 x$) è detto coefficiente di regressione di Y rispetto a X ed indica di quanto varia Y al variare di una unità di X, mentre b_2 (reciproco del coefficiente angolare della retta $x = a_2 + b_2 y$) è detto coefficiente di regressione di X rispetto a Y indica di quanto varia X al variare di una unità di Y. Il segno di b_1 e b_2 è lo stesso e dipende dal numeratore: se b_1 e b_2 sono entrambi positivi al crescere di X la Y cresce, se b_1 e b_2 sono entrambi negativi al crescere di X la Y decresce. Le rette $y = a_1 + b_1 x$ e $x = a_2 + b_2 y$ prendono il nome di rette di regressione di X rispetto a Y e viceversa. La bontà della retta, come funzione di regressione, si può dedurre dall'ampiezza dell'angolo da esse formate: più l'angolo è piccolo e più la funzione lineare rappresenta bene il legame di dipendenza. Casi limite: $\boldsymbol{b_2 = \frac{1}{b_1}}$ (cioè $\boldsymbol{b_1 \cdot b_2 = 1}$) $\Longrightarrow$ rette coincidenti e regressione lineare ottima; $\boldsymbol{b_1 = b_2 = 0} \Longrightarrow$ rette perpendicolari e regressione nulla.

CORRELAZIONE

Lo studio della correlazione ha lo scopo di evidenziare tra le variabili statistiche X e Y un legame di interdipendenza, cioè di vedere se esse si influenzano reciprocamente. Si premette alla regressione, perché se X e Y son prive di legame dovranno essere forzatamente indipendenti. La correlazione lineare si misura mediante un numero puro[3] $\boldsymbol{r}$ tale che sia $-1 \leq r \leq 1$. Si chiama **indice di correlazione lineare di Bravais-Pearson** ed è dato dal rapporto della covarianza (o varianza congiunta) di X e Y, σ_{xy}, ed il prodotto degli scarti quadratici medi di X e Y.

[3] Un numero che misura una grandezza adimensionale e, quindi, non seguito dall'unità di misura.

$$\boldsymbol{\sigma_{xy}} = \frac{\sum_{i=1}^{n} x_i' y_i'}{n} \quad \boldsymbol{\sigma_x} = \sqrt{\frac{\sum_{i=1}^{n} (x_i')^2}{n}} \quad \boldsymbol{\sigma_y} = \sqrt{\frac{\sum_{i=1}^{n} (y_i')^2}{n}} \Rightarrow \boldsymbol{r} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \Rightarrow$$

$$\boldsymbol{r} = \frac{\sum_{i=1}^{n} x_i' y_i'}{\sqrt{\sum_{i=1}^{n} (x_i')^2 \sum_{i=1}^{n} (y_i')^2}}$$

$r > 0 \Rightarrow$ correlazione diretta o positiva, $r < 0 \Rightarrow$ correlazione inversa o negativa, $r = 1 \Rightarrow$ correlazione perfetta positiva, $r = -1 \Rightarrow$ correlazione perfetta negativa, $r = 0 \Rightarrow$ assenza di correlazione, $-1 < r < 0 \Rightarrow$ rette decrescenti che formano un angolo acuto, mentre $0 < r < 1 \Rightarrow$ rette crescenti formanti un angolo acuto. Elevando entrambi i membri al quadrato abbiamo che:

$\boldsymbol{r^2 = b_1 b_2}$ da cui segue: $\boldsymbol{r = \pm\sqrt{b_1 b_2}}$ media geometrica di b_1 e b_2.

Essa va presa col segno più se b_1 e $b_2 > 0$, col segno meno se b_1 e $b_2 < 0$; $\boldsymbol{r^2}$ prende il nome di **coefficiente di determinazione**, ovviamente è : $\boldsymbol{r^2 \leq 1}$, ed indica la bontà del modello lineare: **più esso è vicino ad 1 più la retta rappresenta bene il legame di dipendenza tra le variabili**.

Moltiplicando $\boldsymbol{r^2}$ per 100 si ha la percentuale di varianza[4] spiegata, $\sigma_{\hat{Y}}^2$, dalla relazione di dipendenza, mentre il suo complemento a 100 ci dà la percentuale di varianza non spiegata, σ_d^2, da detta relazione ma da altre cause. Più alta è la prima e più il modello lineare va bene. La loro somma fornisce la varianza totale σ_Y^2. Risulta:

$$\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} + \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n} \Rightarrow \boldsymbol{\sigma_Y^2 = \sigma_d^2 + \sigma_{\hat{Y}}^2}$$

Dove $\hat{y}_i$ rappresenta il valore teorico ottenuto sostituendo le x_i iniziali nella retta di regressione $y = a_1 + b_1 x$. Se la funzione non è una retta, per determinare r^2 si devono calcolare $\sigma_{\hat{Y}}^2$ e σ_Y^2 il cui rapporto vale: $\boldsymbol{\frac{\sigma_{\hat{Y}}^2}{\sigma_Y^2} = b_1 b_2 = r^2}$.

[4] **La varianza** è un indice di dispersione (come lo è lo scarto quadratico medio o errore standard) che ci dice se la media aritmetica trovata è rappresentativa o no della distribuzione di dati iniziale. Trattasi di una media quadratica, cioè che conserva la somma dei quadrati. Essa è la più grande delle medie. Tra le medie aritmetica M , geometrica G , armonica A e quadratica Q vale la disuguaglianza seguente: $A \leq G \leq M \leq Q$ valendo l'uguale se i numeri sono uguali.