LA CORRELAZIONE LINEARE

Supponiamo di essere interessati a valutare la relazione esistente tra importazioni ed esportazioni dei paesi dell'UE nel 2006.

Ci sono situazioni in cui sul piano logico non è possibile assegnare a uno dei due caratteri il ruolo di dipendente e all'altro quello di indipendente (perché i due caratteri si influenzano vicendevolmente o perché dipendono da un fattore latente).

Studieremo allora la dipendenza tra X e Y in maniera simmetrica, senza assegnare ruoli distinti ai caratteri. Si parla di studio della correlazione lineare tra X e Y.

Obiettivo: quantificare la dipendenza lineare reciproca tra X e Y attraverso un numero.

X e Y sono concordi? Discordi? Indifferenti?

Concordanza: alle intensità più elevate di X corrispondono le intensità più elevate di Y, e a quelle più basse di X si associano le più basse di Y.

Discordanza: alle intensità più elevate di X corrispondono le intensità più basse di Y, e a quelle più elevate di X si associano le più basse di Y.

Intensità più basse di *X*: quelle t.c. $(x_j - \bar{x}) < 0$.
Intensità più elevate di *X*: quelle t.c. $(x_j - \bar{x}) > 0$.
Analogamente per *Y*.

***X* e *Y* si dicono concordi nell'unità *j* se $(x_j - \bar{x})$ e $(y_j - \bar{y})$ sono entrambi positivi o entrambi negativi. Se il segno degli scarti è diverso, *X* e *Y* si dicono discordi nell'unità *j*.**

Se vi è prevalenza di scarti con segno concorde su quelli di segno discorde, *X* e *Y* si dicono concordi nell'insieme delle *n* unità statistiche.

Per capir bene, chiediamo l'aiuto del grafico.

Le coppie di scarti corrispondenti $(x_j - \bar{x})$ e $(y_j - \bar{y})$ ci dicono singolarmente per ogni unità *j*=1,…,*n* se *X* e *Y* sono concordi oppure no.

Problema: come misurare nel complesso delle *n* unità la concordanza o la discordanza?

Risposta: si sintetizza l'accordo-disaccordo tra scarti attraverso la covarianza

$$\mathrm{COVAR}(X,Y) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}).$$

La covarianza è la media del prodotto degli scarti corrispondenti dalle rispettive medie aritmetiche.

Covarianza nulla significa caratteri indipendenti? Consideriamo due situazioni (→ grafici).

Si può dimostrare che

$$\left|\mathrm{COVAR}(X,Y)\right| \le \sqrt{\mathrm{VAR}(X)\mathrm{VAR}(Y)}.$$

Da cui segue che un indice normalizzato di (inter)dipendenza lineare è

$$r = \frac{\mathrm{COVAR}(X,Y)}{\sqrt{\mathrm{VAR}(X)\mathrm{VAR}(Y)}} = \frac{\mathrm{CODEV}(X,Y)}{\sqrt{\mathrm{DEV}(X)\mathrm{DEV}(Y)}},$$

detto indice di Bravais-Pearson. $-1 \le r \le 1$.

$r \cdot 100$ quantifica il legame lineare (di concordanza o discordanza in base al segno) come percentuale del massimo legame lineare possibile.

Notiamo che

$$R^2 = \frac{\left[\mathrm{COVAR}(X,Y)\right]^2}{\mathrm{VAR}(X)\mathrm{VAR}(Y)} \text{ e che } r = b\frac{\sqrt{\mathrm{VAR}(X)}}{\sqrt{\mathrm{VAR}(Y)}}.$$

Esempio dei 30 appartamenti.

$$r = \frac{\mathrm{CODEV}(X,Y)}{\sqrt{\mathrm{DEV}(X)\mathrm{DEV}(Y)}} = \frac{25625}{\sqrt{25992{,}7 \cdot 28558{,}0}} = 0{,}942.$$

Come si interpreta 0,942?