CORRELATION COEFFICIENT

Utile per analizzare la correlazione fra due variabili numeriche.

$$r_{A,B} = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n \cdot \sigma_A \cdot \sigma_B} = \frac{\sum_{i=1}^{n} (a_i b_i) - n\,\bar{A}\bar{B}}{n \cdot \sigma_A \cdot \sigma_B}$$

Dove:

- a_i e b_i : valori attuali in A e B
- $\bar{A}$ e $\bar{B}$: media dei valori in A e B
- σ_A e σ_B : deviazione standard in A e B
- n : numero di tuple nel DataSet

L'intuizione della formula è la seguente:
Se assumiamo che A e B hanno lo stesso trend, per esempio entrambi crescono in un determinato punto, quello che succede nella formula è che il prodotto al numeratore mantiene il segno positivo. Se per esempio assumiamo che i valori siano ordinati in modo crescente, se partiamo da a_ϕ e b_ϕ, essi sono necessariamente $a_\phi < \bar{A}$ e $b_\phi < \bar{B}$ perché il trend è crescente. Sotto questa luce abbiamo:

$(a_i - \bar{A}) \Rightarrow$ NEGATIVO
$(b_i - \bar{B}) \Rightarrow$ NEGATIVO
$(a_i - \bar{A})(b_i - \bar{B}) \Rightarrow$ POSITIVO

Ma questo prodotto, sotto queste ipotesi è sempre positivo.

Le cose cambiano il trend, con A che cresce e B decresce. Ora otteniamo che il prodotto è sempre negativo:

$(a_1 - \bar{A}) \Rightarrow$ NEGATIVO ... $(a_N - \bar{A}) \Rightarrow$ POSITIVO
$(b_1 - \bar{B}) \Rightarrow$ POSITIVO ... $(b_N - \bar{B}) \Rightarrow$ NEGATIVO
$(a_1 - \bar{A})(b_1 - \bar{B}) \Rightarrow$ NEGATIVO ... $(a_N - \bar{A})(b_N - \bar{B}) \Rightarrow$ NEGATIVO

SE $\tau_{A,B} > 0$ ALLORA A e B POSITIVAMENTE CORRELATE
SE $\tau_{A,B} < 0$ ALLORA A e B NEGATIVAMENTE CORRELATE
SE $\tau_{A,B} \cong 0$ ALLORA A e B INDIPENDENTI

Quest'ultimo caso è motivato dal fatto che, essendo A e B indipendenti, il prodotto a volte è negativo altre e positivo (confrontiamo valori random), quindi nella somma a volte sottraggo, altre somma.

ATTENZIONE
Questo metodo rileva solo dipendenze lineari.

VALORI: $-1 < \tau_{A,B} < 1$

Standardizzare

Per valutare la correlazione, i valori vengono spesso standardizzati:

$$a'_k = \frac{(a_k - \bar{A})}{\sigma_A}, \quad b'_k = \frac{(b_k - \bar{B})}{\sigma_A}$$

COVARIANZA

$$Cov(A,B) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}$$

Da cui possiamo dire che $\rho_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

Valgono le stesse considerazioni fatte per $\rho_{A,B}$, tranne per $Cov(A,B) \cong \emptyset$.

Per $Cov(A,B) \cong \emptyset$, A e B sono indipendenti. Il contrario non è sempre vero, ovvero che:

$$A,B: \text{indipendenti} \Rightarrow Cov(A,B) \cong \emptyset$$