Correlazione tra due variabili

Federico Plazzi

26 Novembre 2015

Correlazione tra due variabili

Correlazione tra due variabili

Variabili dipendenti e variabili indipendenti

- La **variabile indipendente** è quella che, secondo le nostre aspettative, dovrebbe essere in grado di influenzare l'altra; di solito viene indicata sull'asse delle ascisse.

Correlazione tra due variabili

Variabili dipendenti e variabili indipendenti

- La **variabile indipendente** è quella che, secondo le nostre aspettative, dovrebbe essere in grado di influenzare l'altra; di solito viene indicata sull'asse delle ascisse.
- La **variabile dipendente** è quella che, secondo le nostre aspettative, viene invece influenzata dall'altra; di solito viene indicata sull'asse delle ordinate.

Correlazione tra due variabili

Variabili dipendenti e variabili indipendenti

- La **variabile indipendente** è quella che, secondo le nostre aspettative, dovrebbe essere in grado di influenzare l'altra; di solito viene indicata sull'asse delle ascisse.
- La **variabile dipendente** è quella che, secondo le nostre aspettative, viene invece influenzata dall'altra; di solito viene indicata sull'asse delle ordinate.
- È una semplice *convenzione*: se non abbiamo idea di quale variabile influenzi l'altra, o se ci sembra che si influenzino a vicenda, la scelta dell'asse su cui sistemare le nostre variabili è indifferente.

Correlazione tra due variabili

Variabili dipendenti e variabili indipendenti

- La **variabile indipendente** è quella che, secondo le nostre aspettative, dovrebbe essere in grado di influenzare l'altra; di solito viene indicata sull'asse delle ascisse.
- La **variabile dipendente** è quella che, secondo le nostre aspettative, viene invece influenzata dall'altra; di solito viene indicata sull'asse delle ordinate.
- È una semplice *convenzione*: se non abbiamo idea di quale variabile influenzi l'altra, o se ci sembra che si influenzino a vicenda, la scelta dell'asse su cui sistemare le nostre variabili è indifferente.
- **La correlazione è sempre un concetto simmetrico**.

Pearson product-moment correlation coefficient

Pearson product-moment correlation coefficient

La codevianza

- **Correlazione significa co-variabilità**: le due variabili *tendono a variare insieme*, positivamente o negativamente.

Pearson product-moment correlation coefficient

La codevianza

- **Correlazione significa co-variabilità**: le due variabili *tendono a variare insieme*, positivamente o negativamente.
- La "co-variabilità" viene stimata attraverso il concetto di **covarianza**.

Pearson product-moment correlation coefficient

La codevianza

- **Correlazione significa co-variabilità**: le due variabili *tendono a variare insieme*, positivamente o negativamente.
- La "co-variabilità" viene stimata attraverso il concetto di **covarianza**.
- Partiamo dal concetto di *devianza*:

$$D_X = \sum_{i=1}^{N} (X_i - \overline{X})^2 = \sum_{i=1}^{N} (X_i - \overline{X})(X_i - \overline{X}) \quad (1)$$

$$D_Y = \sum_{i=1}^{N} (Y_i - \overline{Y})^2 = \sum_{i=1}^{N} (Y_i - \overline{Y})(Y_i - \overline{Y}) \quad (2)$$

Pearson product-moment correlation coefficient

La codevianza

- **Correlazione significa co-variabilità**: le due variabili *tendono a variare insieme*, positivamente o negativamente.
- La "co-variabilità" viene stimata attraverso il concetto di **covarianza**.
- Partiamo dal concetto di *devianza*:

$$D_X = \sum_{i=1}^{N} (X_i - \overline{X})^2 = \sum_{i=1}^{N} (X_i - \overline{X})(X_i - \overline{X}) \tag{1}$$

$$D_Y = \sum_{i=1}^{N} (Y_i - \overline{Y})^2 = \sum_{i=1}^{N} (Y_i - \overline{Y})(Y_i - \overline{Y}) \tag{2}$$

- Definiamo la *codevianza*:

$$D_{XY} = \sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y}) \tag{3}$$

Pearson product-moment correlation coefficient

Pearson product-moment correlation coefficient

La covarianza

- Ricordiamo il concetto di *varianza*:

$$\sigma_X^2 = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N} = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(X_i - \overline{X})}{N} \tag{4}$$

$$\sigma_Y^2 = \frac{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}{N} = \frac{\sum_{i=1}^{N} (Y_i - \overline{Y})(Y_i - \overline{Y})}{N} \tag{5}$$

Pearson product-moment correlation coefficient

La covarianza

- Ricordiamo il concetto di *varianza*:

$$\sigma_X^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(X_i - \overline{X})}{N} \tag{4}$$

$$\sigma_Y^2 = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N} = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})(Y_i - \overline{Y})}{N} \tag{5}$$

- Definiamo la *covarianza*:

$$\sigma_{XY}^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{N} \tag{6}$$

Pearson product-moment correlation coefficient

Pearson product-moment correlation coefficient

La covarianza massima

- Qual è la covarianza massima?

Pearson product-moment correlation coefficient

La covarianza massima

- Qual è la covarianza massima?
- Deve dipendere dalle due varianze. In particolare, se una delle due varianze è 0, anche la covarianza deve risultare 0, perché non è possibile che l'altra co-vari con essa. Perciò, anzichè una media aritmetica, dobbiamo usare una *media geometrica*:

$$\max(\sigma^2_{XY}) = \sqrt{\sigma^2_X \cdot \sigma^2_Y} \tag{7}$$

Pearson product-moment correlation coefficient

Pearson product-moment correlation coefficient

Il coefficiente *r*

- Il coefficiente *r* stima il rapporto tra la covarianza osservata e la covarianza massima possibile:

$$r = \frac{\sigma^2_{XY}}{\max(\sigma^2_{XY})} \tag{8}$$

Pearson product-moment correlation coefficient

Il coefficiente r

- Il coefficiente r stima il rapporto tra la covarianza osservata e la covarianza massima possibile:

$$r = \frac{\sigma^2_{XY}}{\max(\sigma^2_{XY})} \tag{8}$$

- Sostituendo nella 8 usando la 6 e la 7:

$$r = \frac{\frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{N}}{\sqrt{\sigma^2_X \cdot \sigma^2_Y}} \tag{9}$$

Pearson product-moment correlation coefficient

Il coefficiente r

- Il coefficiente r stima il rapporto tra la covarianza osservata e la covarianza massima possibile:

$$r = \frac{\sigma^2_{XY}}{\max(\sigma^2_{XY})} \tag{8}$$

- Sostituendo nella 8 usando la 6 e la 7:

$$r = \frac{\frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{N}}{\sqrt{\sigma^2_X \cdot \sigma^2_Y}} \tag{9}$$

- Applicando la 4 e la 5:

$$r = \frac{\frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{N}}{\sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N} \cdot \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N}}} = \frac{D_{XY}}{\sqrt{D_X \cdot D_Y}} \tag{10}$$

Coefficient of determination

Coefficient of determination

Il coefficiente r

Il coefficiente r che abbiamo appena calcolato indica la direzione della correlazione, in particolare:

- un valore di -1 indica una perfetta correlazione negativa;

Coefficient of determination

Il coefficiente r

Il coefficiente r che abbiamo appena calcolato indica la direzione della correlazione, in particolare:

- un valore di -1 indica una perfetta correlazione negativa;
- un valore di 0 indica totale assenza di correlazione;

Coefficient of determination

Il coefficiente r

Il coefficiente r che abbiamo appena calcolato indica la direzione della correlazione, in particolare:

- un valore di -1 indica una perfetta correlazione negativa;
- un valore di 0 indica totale assenza di correlazione;
- un valore di $+1$ indica una perfetta correlazione positiva.

Coefficient of determination

Il coefficiente r

Il coefficiente r che abbiamo appena calcolato indica la direzione della correlazione, in particolare:

- un valore di -1 indica una perfetta correlazione negativa;
- un valore di 0 indica totale assenza di correlazione;
- un valore di $+1$ indica una perfetta correlazione positiva.

Il coefficiente r^2

- Il coefficiente r^2 indica la *forza* della correlazione ed è definito semplicemente come $r^2 = r \cdot r$.

Coefficient of determination

Il coefficiente r

Il coefficiente r che abbiamo appena calcolato indica la direzione della correlazione, in particolare:

- un valore di -1 indica una perfetta correlazione negativa;
- un valore di 0 indica totale assenza di correlazione;
- un valore di $+1$ indica una perfetta correlazione positiva.

Il coefficiente r^2

- Il coefficiente r^2 indica la *forza* della correlazione ed è definito semplicemente come $r^2 = r \cdot r$.
- Se, ad esempio, r^2 è pari a $0,75$, vuol dire che il 75% della varianza di X è spiegato dalla varianza di Y e viceversa.

Coefficient of determination

Il coefficiente r

Il coefficiente r che abbiamo appena calcolato indica la direzione della correlazione, in particolare:

- un valore di -1 indica una perfetta correlazione negativa;
- un valore di 0 indica totale assenza di correlazione;
- un valore di $+1$ indica una perfetta correlazione positiva.

Il coefficiente r^2

- Il coefficiente r^2 indica la *forza* della correlazione ed è definito semplicemente come $r^2 = r \cdot r$.
- Se, ad esempio, r^2 è pari a $0,75$, vuol dire che il 75% della varianza di X è spiegato dalla varianza di Y e viceversa.
- La *varianza residua* sarà pari a $1 - r^2$ per entrambe la variabili.

Linea di regressione

- La linea di regressione è la linea che meglio interpola i punti del grafico, secondo la strategia dei *minimi quadrati*.

Linea di regressione

- La linea di regressione è la linea che meglio interpola i punti del grafico, secondo la strategia dei *minimi quadrati*.
- Prendiamo l'equazione generica di una retta nel piano:

$$y = mx + q$$

Linea di regressione

- La linea di regressione è la linea che meglio interpola i punti del grafico, secondo la strategia dei *minimi quadrati*.
- Prendiamo l'equazione generica di una retta nel piano:

$$y = mx + q$$

- Il coefficiente angolare della retta di regressione è dato da

$$m = \frac{D_{XY}}{D_X} \tag{11}$$

Linea di regressione

- La linea di regressione è la linea che meglio interpola i punti del grafico, secondo la strategia dei *minimi quadrati*.
- Prendiamo l'equazione generica di una retta nel piano:

$$y = mx + q$$

- Il coefficiente angolare della retta di regressione è dato da

$$m = \frac{D_{XY}}{D_X} \tag{11}$$

- L'intercetta della retta di regressione è data da

$$q = \overline{Y} - m\overline{X} \tag{12}$$

Previsioni in base alla retta di regressione

Il concetto di residuo

- Un *residuo* è la differenza tra il valore di Y letto sulla retta di regressione e quello osservato.

Previsioni in base alla retta di regressione

Il concetto di residuo

- Un *residuo* è la differenza tra il valore di Y letto sulla retta di regressione e quello osservato.
- Possiamo calcolare la somma dei residui quadrati come

$$\Sigma R^2 = D_Y \cdot (1 - r^2) \tag{13}$$

dove R^2 è un residuo quadrato e $1 - r^2$, come ricorderete, è la varianza residua.

Previsioni in base alla retta di regressione

L'errore standard dei residui

- ΣR^2 ha la forma di una devianza; per ottenere una deviazione standard, che chiameremo *errore standard* (dei residui), possiamo dividere per N ed estrarre la radice quadrata.

$$\mathrm{SE}_R = \sqrt{\frac{\Sigma R^2}{N}} \tag{14}$$

Previsioni in base alla retta di regressione

L'errore standard dei residui

- ΣR^2 ha la forma di una devianza; per ottenere una deviazione standard, che chiameremo *errore standard* (dei residui), possiamo dividere per N ed estrarre la radice quadrata.

$$\mathrm{SE}_R = \sqrt{\frac{\Sigma R^2}{N}} \tag{14}$$

- Se però vogliamo stimare l'errore standard dei residui di tutta la popolazione, dobbiamo dividere per $N - 2$:

$$\hat{\mathrm{SE}}_R = \sqrt{\frac{\Sigma R^2}{N - 2}} \tag{15}$$

Previsioni in base alla retta di regressione

L'errore standard dei residui

- ΣR^2 ha la forma di una devianza; per ottenere una deviazione standard, che chiameremo *errore standard* (dei residui), possiamo dividere per N ed estrarre la radice quadrata.

$$\mathrm{SE}_R = \sqrt{\frac{\Sigma R^2}{N}} \tag{14}$$

- Se però vogliamo stimare l'errore standard dei residui di tutta la popolazione, dobbiamo dividere per $N - 2$:

$$\hat{\mathrm{SE}}_R = \sqrt{\frac{\Sigma R^2}{N - 2}} \tag{15}$$

- Questo errore standard (dei residui) si distribuisce in modo normale!

Previsioni in base alla retta di regressione

Previsioni in base alla retta di regressione

Data la nostra retta di regressione

$$y = mx + q$$

Previsioni in base alla retta di regressione

Data la nostra retta di regressione

$$y = mx + q$$

possiamo quindi inserire il nostro errore standard come

$$y = mx + q \pm 1,96\hat{\mathrm{SE}}_R \tag{16}$$

ed otterremo un *intervallo di confidenza* di valori in cui abbiamo una probabilità di circa il 95% di ottenere una previsione corretta.

Significatività della correlazione

Significatività della correlazione

La distribuzione di r

- Come si distribuisce r? Se riusciamo a capirlo, possiamo stimarne la significatività.

Significatività della correlazione

La distribuzione di r

- Come si distribuisce r? Se riusciamo a capirlo, possiamo stimarne la significatività.
- Effettuiamo 10 lanci di un paio di dadi (diversi) per 10.000 volte: ogni serie di 10 lanci calcoliamo la correlazione tra il valore ottenuto sul primo dado ed il valore ottenuto sul secondo.

Significatività della correlazione

La distribuzione di r

- Come si distribuisce r? Se riusciamo a capirlo, possiamo stimarne la significatività.
- Effettuiamo 10 lanci di un paio di dadi (diversi) per 10.000 volte: ogni serie di 10 lanci calcoliamo la correlazione tra il valore ottenuto sul primo dado ed il valore ottenuto sul secondo.
- Ripetiamo il nostro esperimento per numeri di lanci diversi.

Significatività della correlazione

La distribuzione di r

- Come si distribuisce r? Se riusciamo a capirlo, possiamo stimarne la significatività.
- Effettuiamo 10 lanci di un paio di dadi (diversi) per 10.000 volte: ogni serie di 10 lanci calcoliamo la correlazione tra il valore ottenuto sul primo dado ed il valore ottenuto sul secondo.
- Ripetiamo il nostro esperimento per numeri di lanci diversi.
- La distribuzione di r è normale!

Significatività della correlazione

La distribuzione di r

- Come si distribuisce r? Se riusciamo a capirlo, possiamo stimarne la significatività.
- Effettuiamo 10 lanci di un paio di dadi (diversi) per 10.000 volte: ogni serie di 10 lanci calcoliamo la correlazione tra il valore ottenuto sul primo dado ed il valore ottenuto sul secondo.
- Ripetiamo il nostro esperimento per numeri di lanci diversi.
- La distribuzione di r è normale!
- Allora possiamo usare il solito trucco: quanta parte della curva è sottesa da $-\infty$ ad un valore pari al mio r?

Significatività della correlazione

La distribuzione di r

- Come si distribuisce r? Se riusciamo a capirlo, possiamo stimarne la significatività.
- Effettuiamo 10 lanci di un paio di dadi (diversi) per 10.000 volte: ogni serie di 10 lanci calcoliamo la correlazione tra il valore ottenuto sul primo dado ed il valore ottenuto sul secondo.
- Ripetiamo il nostro esperimento per numeri di lanci diversi.
- La distribuzione di r è normale!
- Allora possiamo usare il solito trucco: quanta parte della curva è sottesa da $-\infty$ ad un valore pari al mio r?
- Al solito, possiamo fare un test ad una od a due code.