# Psicometria con Laboratorio di SPSS 2

## Regressione lineare multipla

(v. 1.7a, 7 novembre 2022)

Germano Rossi[1]
germano.rossi@unimib.it

[1]Dipartimento di Psicologia, Università di Milano-Bicocca

2017-18

# Sommario

1 Regressione lineare multipla

2 In SPSS

3 Correlazioni multiple e parziali

**ATTENZIONE:** Tutte le slide con una lettera nell'angolo destro del titolo, indicano: A = delle informazioni Avanzate da leggere successivamente dopo aver compreso le altre; M = una parte che utilizza l'algebra Matriciale che è possibile saltare.

# Sommario locale


1. Regressione lineare multipla
   - Introduzione generale
   - Influenze fra variabili
   - Varianza spiegata o ampiezza dell'effetto
2. In SPSS
3. Correlazioni multiple e parziali

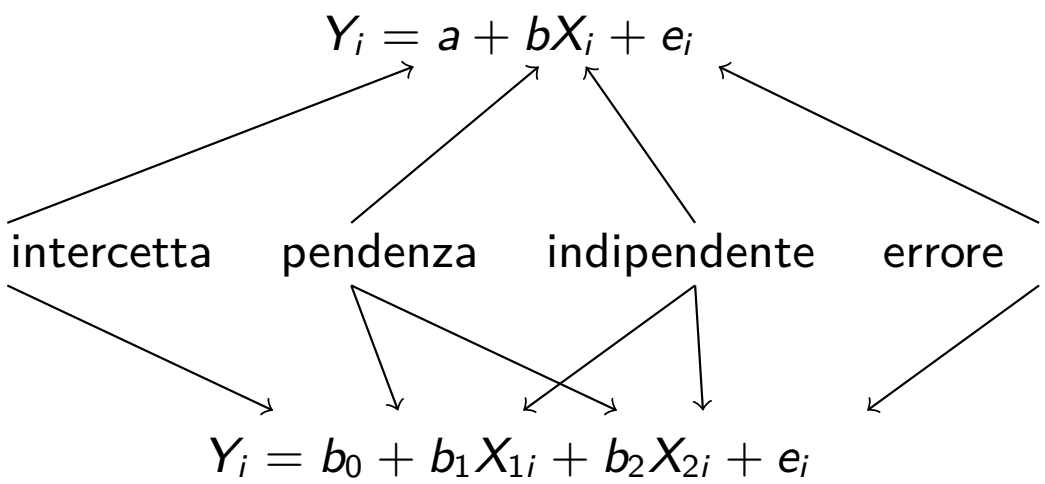
# Regressione lineare multipla

- Analoga a quella semplice

- Una sola variabile dipendente (Y) da spiegare

- Due o più variabili indipendenti (X) predittive, esplicative

Regressione lineare semplice (1 dip, 1 indip)




Regressione lineare multipla (2 indip, 1 dip)

- a, $b_0$, intercetta, costante
- b, $b_1$ $b_2$ ..., pendenza, coefficienti/parametri di regressione

# Regressione come modello generale

All'equazione della retta

$$Y_i = \beta_0 + \beta_1 X_i \quad \text{oppure} \quad Y_i = a + bX_i \quad \text{se 1 VI}$$

possiamo aggiungere più variabili esplicative quantitative

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \quad \text{se 2 VI}$$
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \quad \text{se 3 VI}$$

oppure variabili categoriali dicotomiche (0,1) [uso D per evidenziare]

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 D_{2i} \quad \text{se 1 Qt, 1 dicot}$$

e con un'interazione (con I=XD]

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 D_{2i} + \beta_3 I_{3i} \quad \text{se 1 Qt, 1 dicot, 1 inter}$$

# Regressione multipla: caso per caso

Ipotizziamo di avere una Y e due X $(X_1, X_2)$ l'equazione sarà:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

| $Y$ | $X_1$ | $X_2$ | $Y = b_0 + X_{1i} b_1 + X_{2i} b_2 + e$ |
|---|---|---|---|
| 3 | 2 | 1 | $3 = 1b_0 + 2b_1 + 1b_2 + e_1$ |
| 2 | 3 | 5 | $2 = 1b_0 + 3b_1 + 5b_2 + e_2$ |
| 4 | 5 | 3 | $4 = 1b_0 + 5b_1 + 3b_2 + e_3$ |
| 5 | 7 | 6 | $5 = 1b_0 + 7b_1 + 6b_2 + e_4$ |
| 8 | 8 | 7 | $8 = 1b_0 + 8b_1 + 7b_2 + e_5$ |

- Per ogni caso statistico, sostituiamo i valori nell'equazione generale
- La costante $(b_0)$ viene idealmente moltiplicata per 1
- Per ogni caso statistico, sostituiamo i valori nell'equazione generale

# Regressione multipla: caso per caso

Ipotizziamo di avere una Y e due X ($X_1, X_2$) l'equazione sarà:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

| $Y$ | $X_1$ | $X_2$ | $Y = b_0 + X_{1i} b_1 + X_{2i} b_2 + e$ |
|---|---|---|---|
| 3 | 2 | 1 | $3 = 1b_0 + 2b_1 + 1b_2 + e_1$ |
| 2 | 3 | 5 | $2 = 1b_0 + 3b_1 + 5b_2 + e_2$ |
| 4 | 5 | 3 | $4 = 1b_0 + 5b_1 + 3b_2 + e_3$ |
| 5 | 7 | 6 | $5 = 1b_0 + 7b_1 + 6b_2 + e_4$ |
| 8 | 8 | 7 | $8 = 1b_0 + 8b_1 + 7b_2 + e_5$ |

- Per ogni caso statistico, sostituiamo i valori nell'equazione generale
- La costante ($b_0$) viene idealmente moltiplicata per 1
- Per ogni caso statistico, sostituiamo i valori nell'equazione generale

# Regressione multipla: caso per caso

Ipotizziamo di avere una Y e due X ($X_1, X_2$) l'equazione sarà:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

| $Y$ | $X_1$ | $X_2$ | $Y = b_0 + X_{1i} b_1 + X_{2i} b_2 + e$ |
|---|---|---|---|
| 3 | 2 | 1 | $3 = 1b_0 + 2b_1 + 1b_2 + e_1$ |
| 2 | 3 | 5 | $2 = 1b_0 + 3b_1 + 5b_2 + e_2$ |
| 4 | 5 | 3 | $4 = 1b_0 + 5b_1 + 3b_2 + e_3$ |
| 5 | 7 | 6 | $5 = 1b_0 + 7b_1 + 6b_2 + e_4$ |
| 8 | 8 | 7 | $8 = 1b_0 + 8b_1 + 7b_2 + e_5$ |

- Per ogni caso statistico, sostituiamo i valori nell'equazione generale
- La costante ($b_0$) viene idealmente moltiplicata per 1
- Per ogni caso statistico, sostituiamo i valori nell'equazione generale

Passando al matriciale, abbiamo vettori e matrici:

$$\begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 3 & 5 \\ 1 & 5 & 3 \\ 1 & 7 & 6 \\ 1 & 8 & 7 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$\begin{matrix} 5 \times 1 \\ \mathbf{y} \end{matrix} = \begin{matrix} 5 \times 3 \\ \mathbf{X} \end{matrix} \begin{matrix} 3 \times 1 \\ \mathbf{b} \end{matrix} + \begin{matrix} 5 \times 1 \\ \mathbf{e} \end{matrix}$$

- Ignoriamo gli errori (**e**)
- **b** è il vettore dei parametri incogniti (non standardizzati)
- $\mathbf{b}^*$ o $\beta$ è il vettore dei parametri (standardizzati)
- In algebra, per trovare **b** useremmo $\mathbf{y}/\mathbf{X}$
- ma in algebra matriciale la "divisione" diventa un"'inversa" (indicata con $^{-1}$)

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\left( \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \\ 1 & 5 & 3 & 6 & 7 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 3 & 5 \\ 1 & 5 & 3 \\ 1 & 7 & 6 \\ 1 & 8 & 7 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \\ 1 & 5 & 3 & 6 & 7 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix}$$

$$\begin{bmatrix} 5 & 25 & 22 \\ 25 & 151 & 130 \\ 22 & 130 & 120 \end{bmatrix}^{-1} \begin{bmatrix} 22 \\ 131 \\ 111 \end{bmatrix} = \begin{bmatrix} 0.50 \\ 1 \\ -0.25 \end{bmatrix} \begin{matrix} b_0 \\ b_1 \\ b_2 \end{matrix}$$

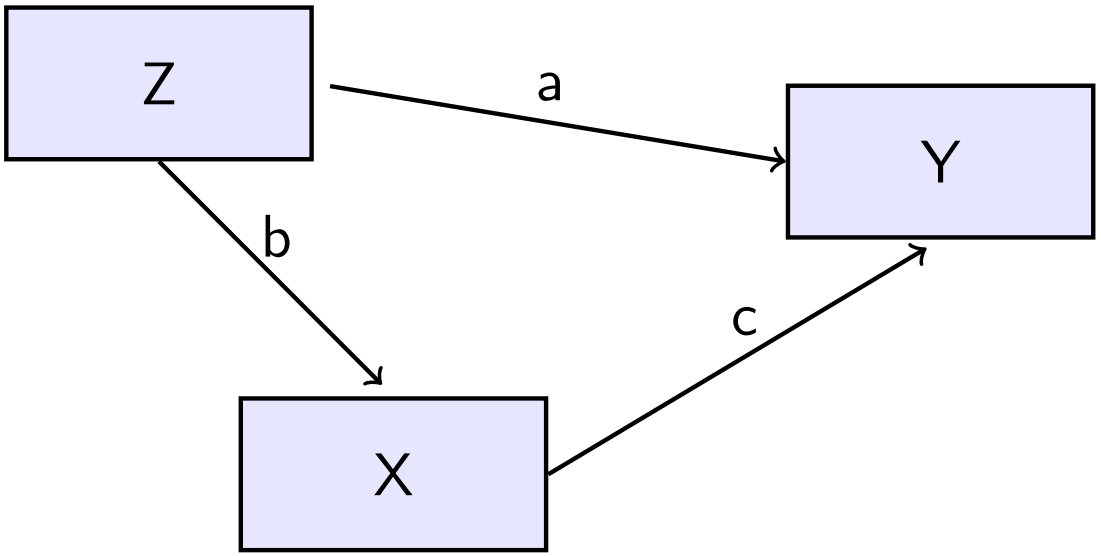
$$\hat{Y}_i = .50 + 1X_{1i} + (-.25)X_{2i}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} N & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 \end{bmatrix} \qquad \mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \end{bmatrix}$$

Più in generale, date $n$ variabili indipendenti, **X'X** e **X'y** diventeranno:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} N & \sum X_1 & \cdots & \sum X_n \\ \sum X_1 & \sum X_1^2 & \cdots & \sum X_1 X_n \\ \cdots & \cdots & \cdots & \cdots \\ \sum X_n & \sum X_1 X_n & \cdots & \sum X_n^2 \end{bmatrix} \qquad \mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \cdots \\ \sum X_n Y \end{bmatrix}$$

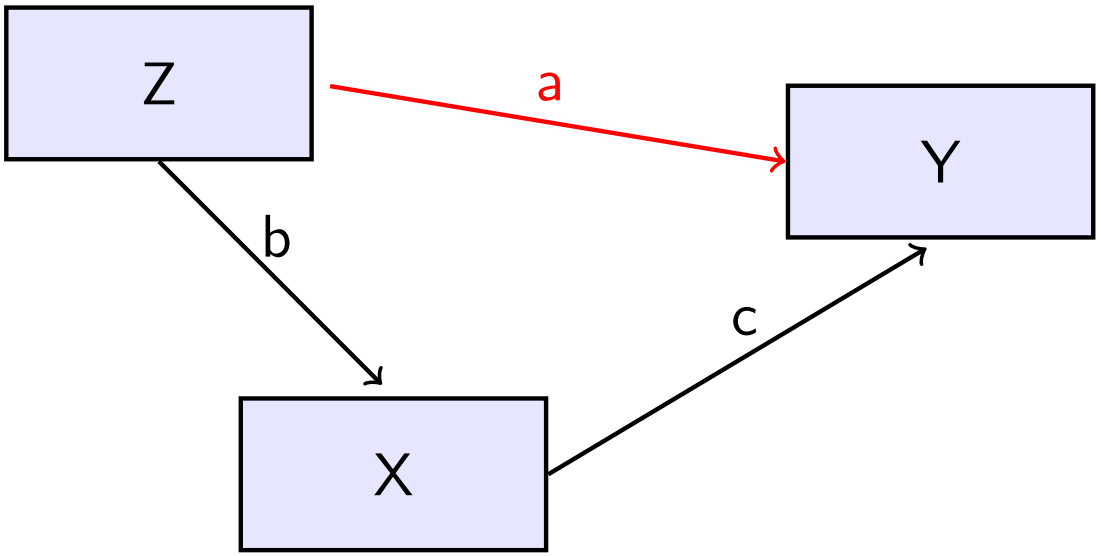
# Percorsi causali/relazionali: definizioni




$Z \rightarrow Y$
significa che $Z$ spiega $Y$
e $a$ è il valore dell'influenza

- **Influenza diretta** = percorso semplice ($Z \rightarrow Y = a$, $Z \rightarrow X = b$, $X \rightarrow Y = c$)
- **Influenza indiretta** = percorso composto ($Z \rightarrow X \rightarrow Y = bc$) con anche le eventuali covarianze
- Il valore di un'influenza indiretta è pari al prodotto delle influenze semplici ($b \cdot c$)

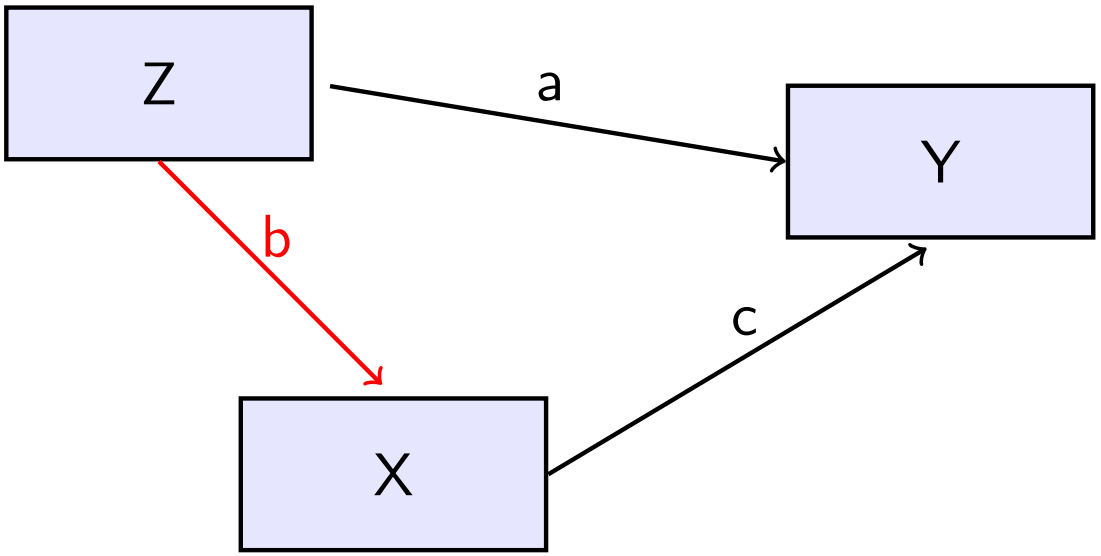
# Percorsi causali/relazionali: definizioni




$Z \rightarrow Y$
significa che $Z$ spiega $Y$
e $a$ è il valore dell'influenza

- **Influenza diretta** = percorso semplice ($Z \rightarrow Y = a$, $Z \rightarrow X = b$, $X \rightarrow Y = c$)
- **Influenza indiretta** = percorso composto ($Z \rightarrow X \rightarrow Y = bc$) con anche le eventuali covarianze
- Il valore di un'influenza indiretta è pari al prodotto delle influenze semplici ($b \cdot c$)

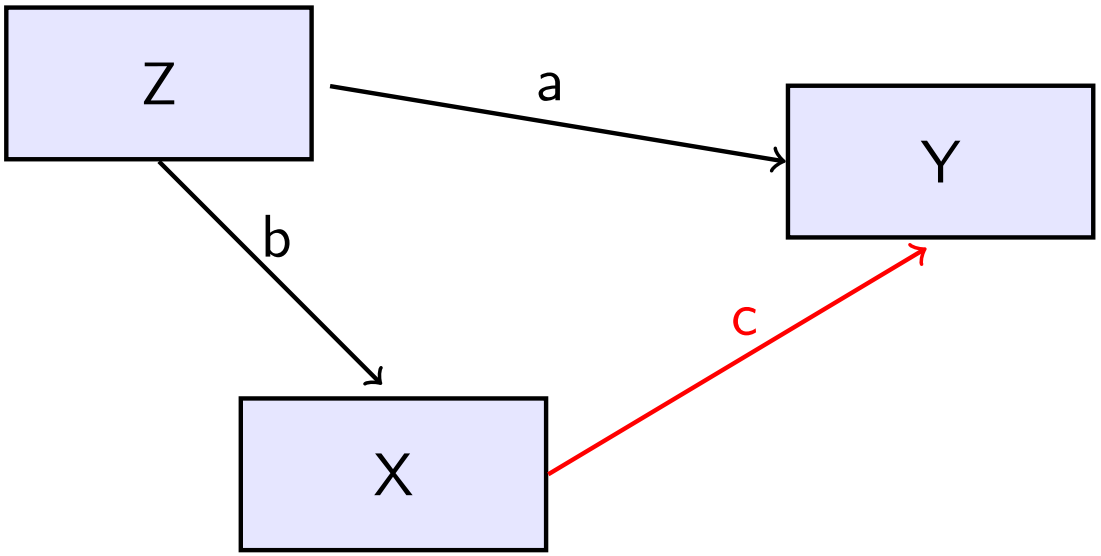
# Percorsi causali/relazionali: definizioni




$Z \to Y$
significa che $Z$ spiega $Y$
e $a$ è il valore dell'influenza

- **Influenza diretta** = percorso semplice ($Z \to Y = a$, $Z \to X = b$, $X \to Y = c$)
- **Influenza indiretta** = percorso composto ($Z \to X \to Y = bc$) con anche le eventuali covarianze
- Il valore di un'influenza indiretta è pari al prodotto delle influenze semplici ($b \cdot c$)

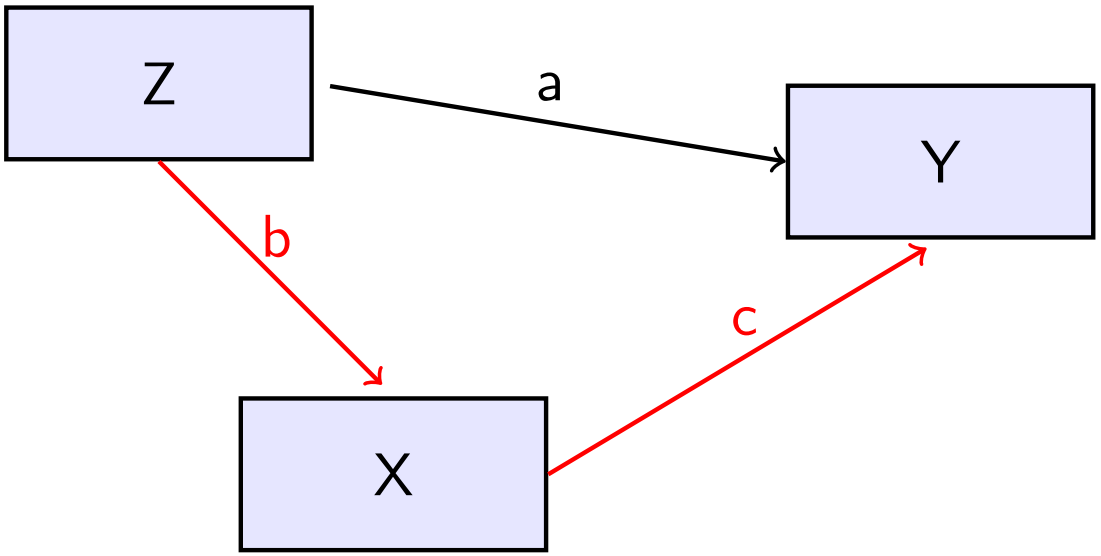
# Percorsi causali/relazionali: definizioni




$Z \to Y$
significa che $Z$ spiega $Y$
e $a$ è il valore dell'influenza

- **Influenza diretta** = percorso semplice ($Z \to Y = a$, $Z \to X = b$, $X \to Y = c$)
- **Influenza indiretta** = percorso composto ($Z \to X \to Y = bc$) con anche le eventuali covarianze
- Il valore di un'influenza indiretta è pari al prodotto delle influenze semplici ($b \cdot c$)

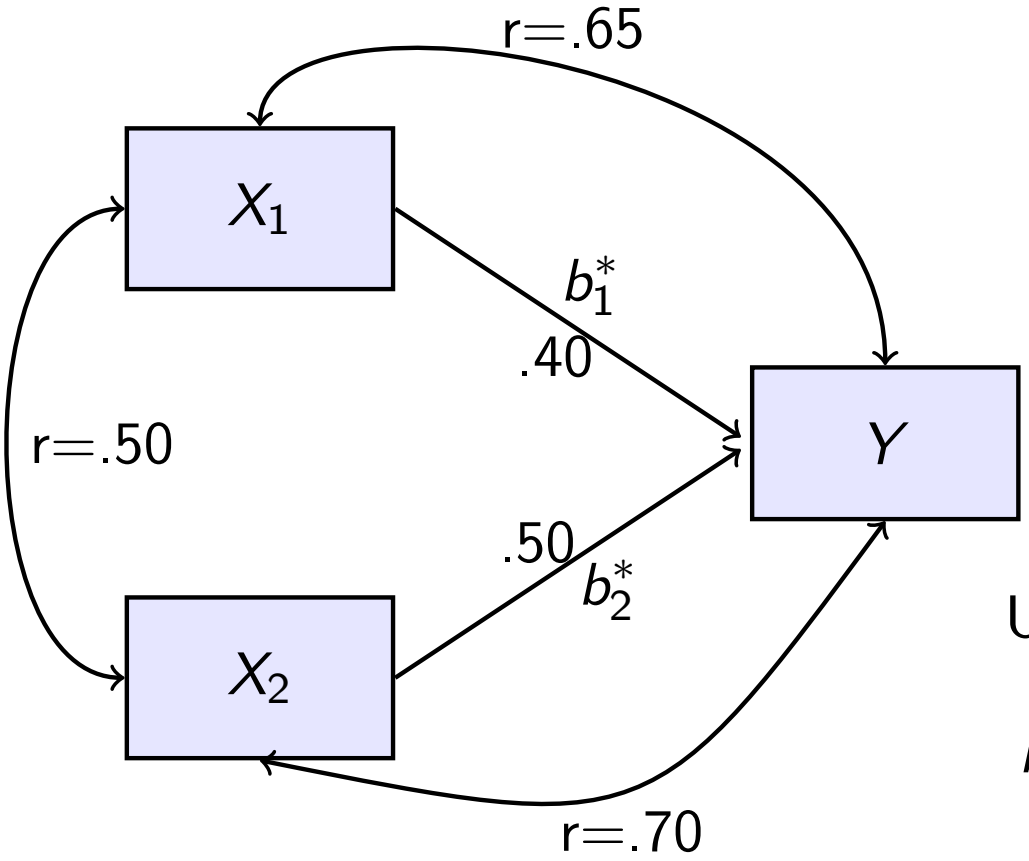
# Percorsi causali/relazionali: definizioni




$Z \rightarrow Y$
significa che $Z$ spiega $Y$
e $a$ è il valore dell'influenza

- **Influenza diretta** = percorso semplice ($Z \rightarrow Y = a$, $Z \rightarrow X = b$, $X \rightarrow Y = c$)
- **Influenza indiretta** = percorso composto ($Z \rightarrow X \rightarrow Y = bc$) con anche le eventuali covarianze
- Il valore di un'influenza indiretta è pari al prodotto delle influenze semplici ($b \cdot c$)

# Percorsi causali/relazionali




$$r_{X_1 Y} = b_1^* + b_2^* r_{X_1 X_2}$$

$$r_{X_2 Y} = b_2^* + b_1^* r_{X_1 X_2}$$

Possiamo ricavare il valore di $b_1^*$ e di $b_2^*$

Uso 1 anziché $X_1$ e 2 anziché $X_2$

$$b_1^* = r_{1Y} - r_{12} b_2^* = .65 - .50 b_2^*$$

$$b_2^* = r_{2Y} - r_{12} b_1^* = .70 - .50 b_1^*$$

La correlazione fra 2 variabili è la somma delle influenze dirette e indirette delle due variabili

Considerando che la correlazione di una variabile con se stessa è 1

$$r_{y1} = b_1^* + b_2^* r_{12} = b_1^* r_{11} + b_2^* r_{12} = b_1^* r_{11} + b_2^* r_{12}$$

$$r_{y2} = b_2^* + b_1^* r_{12} = b_2^* r_{22} + b_1^* r_{12} = b_1^* r_{12} + b_2^* r_{22}$$

$$\begin{bmatrix} r_{y1} \\ r_{y2} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{bmatrix} \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix}$$

$$\mathbf{r}_{yx} = \mathbf{R}_{xx} \mathbf{b}_{yx}^*$$

$$\mathbf{b}_{yx}^* = \mathbf{R}_{xx}^{-1} \mathbf{r}_{yx}$$

- Ci sono tre formule alternative

| | | |
|---|---|---|
| Dati grezzi | $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ | parametri non standardizzati |
| Varianze/covarianze | $\mathbf{b} = \mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{c}_{\mathbf{yx}}$ | parametri non standardizzati |
| Correlazioni | $\mathbf{b}^* = \mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{r}_{\mathbf{yx}}$ | parametri standardizzati |

- $\mathbf{C}_{\mathbf{xx}}$ è la matrice delle varianze/covarianze fra le X
- $\mathbf{c}_{\mathbf{yx}}$ è il vettore delle covarianze fra la Y e le X
- $\mathbf{R}_{\mathbf{xx}}$ è la matrice delle correlazioni fra le X
- $\mathbf{r}_{\mathbf{yx}}$ è il vettore delle correlazioni fra la Y e le X

# Calcoliamo

| $Y$ | $X_1$ | $X_2$ | $Y^2$ | $X_1^2$ | $X_2^2$ | $X_1 Y$ | $X_2 Y$ | $X_1 X_2$ | |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 2 | 1 | 9 | 4 | 1 | 6 | 3 | 2 | |
| 2 | 3 | 5 | 4 | 9 | 25 | 6 | 10 | 15 | |
| 4 | 5 | 3 | 16 | 25 | 9 | 20 | 12 | 15 | |
| 5 | 7 | 6 | 25 | 49 | 36 | 35 | 30 | 42 | |
| 8 | 8 | 7 | 64 | 64 | 49 | 64 | 56 | 56 | |
| 22 | 25 | 22 | 118 | 151 | 120 | 131 | 111 | 130 | somme |
| 4,4 | 5 | 4,4 | | | | | | | medie |
| | | | 5,3 | 6,5 | 5,8 | 5,25 | 3,55 | 5,0 | var/cov |
| | | | | | | .894 | .640 | .814 | cor |

$$var = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N-1} \qquad cov = \frac{\sum xy - \frac{\sum x \sum y}{N}}{N-1} \qquad cor = \frac{cov(xy)}{\sqrt{var(x)var(y)}}$$

Calcolare a mano i parametri di una regr. multipla ($b_0$, $b_1$, $b_2$) è complesso, usiamo il calcolo matriciale

# Esempio con covarianze: parametri NON standardizzati **M**

Varianze e covarianze calcolate con $N - 1$

$$\mathbf{C_{xx}} = \begin{bmatrix} 6.5 & 5.0 \\ 5.0 & 5.8 \end{bmatrix} \qquad \mathbf{c_{yx}} = \begin{bmatrix} 5.25 \\ 3.55 \end{bmatrix}$$

$$\frac{1}{12.7} \begin{bmatrix} 5.8 & -5.0 \\ -5.0 & 6.5 \end{bmatrix} \begin{bmatrix} 5.25 \\ 3.55 \end{bmatrix} = \begin{bmatrix} 1.00 \\ -0.25 \end{bmatrix} \begin{matrix} \leftarrow b_1 \\ \leftarrow b_2 \end{matrix}$$

$$b_0 = \bar{Y} - \sum(b_i \bar{X}_i) = 4.4 - 1 \cdot 5 - (-.25 \cdot 4.4) = 0.5$$

L'equazione finale per calcolare le stime per ogni caso è:

$$\hat{Y}_i = .50 + 1X_{1i} + (-.25)X_{2i}$$

$$\mathbf{R_{xx}} = \begin{bmatrix} 1 & .814 \\ .814 & 1 \end{bmatrix} \qquad \mathbf{r_{yx}} = \begin{bmatrix} .894 \\ .640 \end{bmatrix}$$

$$\frac{1}{0.337} \begin{bmatrix} 1 & -.814 \\ -.814 & 1 \end{bmatrix} \begin{bmatrix} .894 \\ .640 \end{bmatrix} = \begin{bmatrix} 1.107 \\ -0.261 \end{bmatrix} \begin{matrix} \leftarrow b_1^* \\ \leftarrow b_2^* \end{matrix}$$

$$b_0 = 0$$

$$z_{\hat{Y}_i} = 1.107 z_{X_{1i}} + (-.261) z_{X_{2i}}$$

# Standardizzare/destandardizzare

Con i dati dell'esempio precedente

$$b_{yx_i} = b^*_{yx_i} \frac{s_y}{s_{x_i}} \qquad 1 = 1.107 \frac{\sqrt{5.3}}{\sqrt{6.5}} \qquad -0.25 = -.261 \frac{\sqrt{5.3}}{\sqrt{5.8}}$$

$$b^*_{yx_i} = b_{yx_i} \frac{s_{x_i}}{s_y} \qquad 1.107 = 1 \frac{\sqrt{6.5}}{\sqrt{5.3}} \qquad -0.261 = -.25 \frac{\sqrt{5.8}}{\sqrt{5.3}}$$

# Proporzione di varianza spiegata

$$R^2 = (r_{y\hat{y}})^2 = \frac{(spiegata)}{\text{totale}} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{SQ_m}{SQ_t} =$$

$$= \frac{\sum(Y - \bar{Y})^2 - \sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2} =$$

$$= \underbrace{r_{y1} b^*_{1.2} + r_{y2} b^*_{2.1}}_{\text{con 2 X}} = \underbrace{\sum(r_{yi} b^*_i)}_{\text{generico}}$$

$$\mathbf{r_{yx}} = \begin{bmatrix} .894 \\ .640 \end{bmatrix} \quad \mathbf{b}^* = \begin{bmatrix} 1.107 \\ -0.261 \end{bmatrix} \quad R^2 = (.894 \cdot 1.107) + (.640 \cdot -.261)$$

# Proporzione di varianza spiegata o ampiezza dell'effetto

Anche se $R^2$ viene spesso utilizzato per stimare la bontà di una regressione multipla

- $R^2$ è semplicemente una proporzione (varianza comune / varianza totale)
- $R^2 \times 100$ è la % di varianza spiegata dal modello
- $R^2$ non è un'ampiezza dell'effetto
- Ciò nonostante, una versione di $R^2$ (chiamata eta quadro, $\eta^2$) viene spesso interpretata come se fosse un'ampiezza dell'effetto, anche se è un po' distorta.
- altri tentativi di creare una vera misura di ampiezza dell'effetto sono: $\omega$ (omega) e $f$

- $R^2 = \eta^2 = \dfrac{SQ_m}{SQ_t}$
- $\omega = \dfrac{SQ_m + df_m \times MQ_e}{SQ_t + MQ_e}$
- $f = \dfrac{R^2}{1 - R^2}$ proposta da Cohen (1988)

dove $SQ_M$ = Somma quadrati dell'effetto; $SQ_t$ = Somma quadrati totale $df_m$ gradi di libertà dell'effetto; $MQ_e$ media dei quadrati dell'errore

# Test di significatività

Sono i test che facciamo per verificare i passaggi dell'analisi

a) un test globale: che include tutte le variabili (tramite Anova)

Se il test globale è significativo

b) un test per ciascuna variabile indipendente (tramite Anova o t-test)

- Anche se il modello globale è significativo, questo non significa che tutte le X siano significativamente associate a Y

# Test globale

**A**

Nel test globale si fa un confronto fra il modello studiato e il modello nullo

| | | |
|---|---|---|
| nullo (ristretto) | $Y = b_0 + e$ | gdl=N-1 |
| completo | $Y = b_0 + b_1 X_1 + b_2 X_2 + e$ | gdl=N-3 |

con N=numero di soggetti; 1 e 3 sono il numero di parametri

L'ipotesi nulla è $H_0$: $b_1 = b_2 = 0$

Usiamo la statistica F di Snedecor (rapporto di varianze) tramite un'analisi della varianza (Anova)

Se è significativa, c'è una relazione consistente fra le X e la Y; la regressione ha senso. **N.B.**: In genere, è significativa perché nel modello nullo, $b_0$ equivale alla media.

# Test globale: esempio

**A**

$$F = \frac{(R_f^2 - R_r^2)/(d_f - d_r)}{(1 - R_f^2)/d_f}$$

*vale anche per i test parziali*; f=full (completo); r=ristretto [$R^2 = 0$ per il modello nullo]

$$= \frac{\sum(Y - \bar{Y})^2 - \sum(Y - \hat{Y})^2/(d_r - d_f)}{\sum(Y - \hat{Y})^2/d_f}$$

$$= \frac{R_f^2/k}{(1 - R_f^2)/(N - k - 1)}$$

k=numero di variabili indipendenti (X)

$R^2$ tende ad aumentare al numero delle X, quindi viene "aggiustato"

$$R^2_{Adj} = R^2 - (1 - R^2)\frac{k}{N - k - 1}$$

# Test globale: varianza spiegata in Spss

**Riepilogo del modello**

| Modello | R | R-quad | R-quad corr | Err. std. stima |
|---|---|---|---|---|
| 1 | ,908(a) | 0,824 | 0,821 | 14,87675 |

a. Stimatori: (Costante), rwait, oEsSoc, oEsPers, Pacific, oIntrins, Ricerca
b. Variabile dipendente: Fondam

- La variabile `Fondam` viene spiegata da 6 variabili
- R è la correlazione fra tutte le variabili X e la Y (correlazione multipla)
- R-quad è la correlazione al quadrato ($R^2$)
- R-quad corr è la versione corretta dell'$R^2$

# Test globale: esempio in Spss

**ANOVA(b)**

| Modello | | Somma quad. | df | Media quad. | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regress. | 290.770,441 | 6 | 48.461,740 | 218,969 | ,000(a) |
| | Residuo | 61.968,960 | 280 | 221,318 | | |
| | Totale | 352.739,401 | 286 | | | |

a. Stimatori: (Costante), rwait, oEsSoc, oEsPers, Pacific, oIntrins, Ricerca
b. Variabile dipendente: Fondam

$$F = \frac{.824/6}{(1-.824)/(287-6-1)} = 218.484 \quad F = \frac{48.461,740}{221,318} = 218,969$$

$R^2 = .824$

| | | |
|---|---|---|
| ristretto | $Y = b_0 + b_1 X_1 + e$ | gdl=N-2 |
| completo | $Y = b_0 + b_1 X_1 + b_2 X_2 + e$ | gdl=N-3 |

con N=numero di soggetti; 2 e 3 sono il numero di parametri

L'ipotesi nulla è $H_0$: $b_2 = 0$

Potremmo usare ancora la statistica F di Snedecor (rapporto di varianze)

La maggior parte dei programmi utilizza un semplice t-test. Se il test è significativo, la $X_n$ considerata, può stare nel modello, altrimenti si dovrebbe togliere.

$$t = \frac{b_i}{s_{y.i}} = \frac{\text{parametro}}{\text{err. della stima}}$$

# Test per ciascuna X: esempio in Spss

**Coefficienti(a)**

| Modello | | Coef. non stand. B | Coef. non stand. Err. std. | Coef. stand. Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Costante) | 66,265 | 14,455 | | 4,584 | 0,000 |
| | Pacific | -0,383 | 0,106 | -0,104 | -3,619 | 0,000 |
| | oIntrins | 2,195 | 0,177 | 0,437 | 12,385 | 0,000 |
| | oEsPers | 0,351 | 0,365 | 0,028 | 0,962 | **0,337** |
| | oEsSoc | 0,179 | 0,366 | 0,013 | 0,489 | **0,625** |
| | Ricerca | -0,456 | 0,058 | -0,306 | -7,845 | 0,000 |
| | rwait | 0,702 | 0,102 | 0,247 | 6,880 | 0,000 |

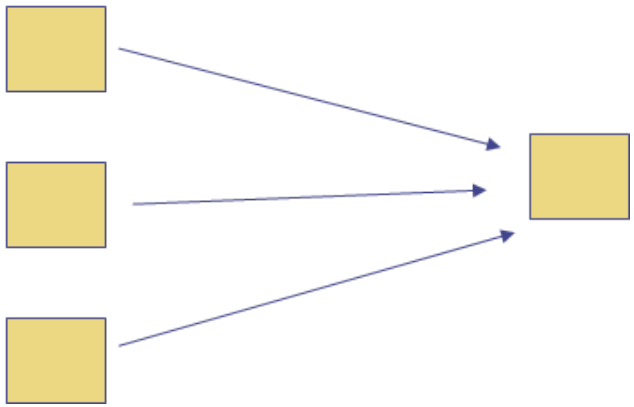
a. Variabile dipendente: Fondam

$$t = \frac{2,195}{0,177} = 12,385$$

Ci sono 2 variabili che non servono ($p > .05$): oEsPers e oEsSoc

Il Coef. stand. è la relazione della X con la Y depurata del contributo delle altre

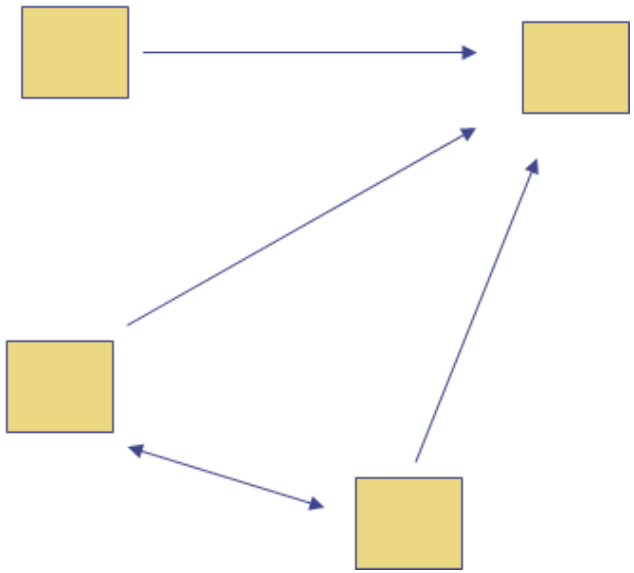
La situazione ideale per una regressione multipla dovrebbe essere: ogni X è altamente correlata con Y, ma le X non sono correlate fra loro



| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $Y$ | .60 | .50 | .70 |
| $X_1$ | | **.20** | **.30** |
| $X_2$ | | | **.20** |

- Idealmente, le correlazioni tra le X, dovrebbero essere 0
- in questo modo $b^*$ coinciderà con $r$ e non con $r$ parzializzato
- è una situazione che capita raramente

Spesso, due o più X sono correlate fra loro



| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $Y$ | .60 | .50 | .70 |
| $X_1$ | | **.70** | .30 |
| $X_2$ | | | .20 |

- Quando due variabili X o più, sono tra loro correlate (moderatamente o più), parliamo di *multicollinearità*
- che può essere più o meno grave

**Problemi**

- Fa aumentare l'errore standard della stima
- limita la R multipla: se A<–>B, ciascuna spiegherebbe una certa parte di Y, ma se A e B spiegano una stessa parte di varianza....
- l'effetto dei predittori si confonde: se A viene usato per prima, B non spiega più niente (e viceversa)
- aumenta la varianza e l'instabilità dell'equazione

**Riconoscere la collinearità**

- *Dopo aver deciso il modello da studiare*:
- Ispezione della matrice di correlazione
- Statistiche della collinearità

# Multicollinearità: statistiche

**A**

- **Variance inflation factor** (VIF)
- Indice di **Tolleranza**: (1/VIF)

**Interpretazione** (non ci sono valori "significativi")

- VIF $>$ 10 (c'è multicollinearità)
- VIF $>$ 1 (è possibile che ci sia)
- Tolleranza $<$ .1 (grossi problemi)
- Tolleranza $<$ .2 (potenzialmente problematico)
- VIF $<=$1 **OK**;
- Tolerance $>=$.2 **OK**

**Coefficienti**[a]

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | | | Statistiche di collinearità | |
|---|---|---|---|---|---|---|---|---|
| | | B | Errore standard | Beta | t | Sign. | Tolleranza | VIF |
| 1 | (Costante) | 27,841 | 2,500 | | 11,138 | ,000 | | |
| | SWLS | ,207 | ,066 | ,259 | 3,125 | ,002 | ,647 | 1,546 |
| | RIFAccet | ,533 | ,137 | ,357 | 3,888 | ,000 | ,530 | 1,888 |
| | DERSTOT | -,082 | ,018 | -,338 | -4,445 | ,000 | ,771 | 1,297 |

a. Variabile dipendente: RSE

**Come diminuire la multicollinearità**

- combinare fra loro i predittori altamente correlati (ad esempio sommandoli)
- eliminare un predittore molto correlato con un altro
- fra più predittori altamente correlati, eliminare quello meno correlato con la Y
- se ci sono molti predittori altamente correlati, usare un'analisi delle componenti principali per ridurre il numero delle X

Matrice di covarianze

| | Y | G | H | V | W |
|---|---|---|---|---|---|
| Y | 4.41 | | | | |
| G | 1.89 | 1.44 | | | |
| H | 1.26 | 0.96 | 2.25 | | |
| V | 1.96 | 0 | 0 | 4.00 | |
| W | 0.63 | 0 | 0 | 0.80 | 1 |

- Notiamo che G e H non correlano con V e W
- Quindi se calcolo $\hat{Y} = b_0 + b_1 H + b_2 V$ troverò esattamente gli stessi parametri di $\hat{Y} = b_0 + b_1 H$ e di $\hat{Y} = b_0 + b_2 V$

usando il calcolo matriciale

$$\begin{bmatrix} 2.25 & 0 \\ 0 & 4.0 \end{bmatrix}^{-1} \begin{bmatrix} 1.26 \\ 1.96 \end{bmatrix} = \begin{bmatrix} 0.56 \\ 0.49 \end{bmatrix}$$

Ipotizzando due regressioni semplici (separate):

$$b_H = \frac{cov(YH)}{var(H)} = \frac{1.26}{2.25} = 0.56 \qquad b_V = \frac{1.96}{4} = 0.49$$

# Scelta dei predittori

- **Usare la teoria**: Si usano solo variabili teoricamente sensate; la sensatezza può essere ricavata, oltre che dalla logica, da una ricerca bibliografica

**In SPSS**, possiamo usare diverse modalità di selezione:

- *Standard* (Inserisci): Tutte le variabili vengono inserite insieme (e poi andranno eliminate manualmente)
- (*Gerarchica*) Seguendo la teoria, si decidono i blocchi da utilizzare e vanno inserimenti per blocchi separati
- Metodi semi-automatici sequenziali (avanti, indietro, per passi)

# Scelta dei predittori

**Metodi semi-automatici sequenziali**

- Il software seleziona le variabili in base a determinati criteri
- *Forward* (Avanti): Le variabili X vengono inserite una alla volta dal software, iniziando con la correlazione fra $X_n$ e Y più alta, poi si prosegue con la seconda più alta e avanti così fino a che le variabili "aggiungono" un contributo significativo ($p<.05$)
- *Backward* (Indietro): Vengono inserite tutte e poi cancellate una alla volta se non significative ($p>.05$)
- *Stepwise* (a passi): Inizia con un blocco di variabili e poi vengono inserite o tolte una alla volta

# Inserimento dei predittori: Inserisci

- metodo **standard** o *Inserisci*: Tutte le X vengono inserite assieme e tutti i coefficienti di regressione (B o $\beta$) stimati contemporaneamente (come negli esempi che abbiamo calcolato con 2 X)

*Coefficienti*[a]

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | | | Statistiche di collinearità | |
|---|---|---|---|---|---|---|---|---|
| | | B | Errore standard | Beta | t | Sign. | Tolleranza | VIF |
| 1 | (Costante) | 7,846 | 2,432 | | 3,226 | ,002 | | |
| | INTRINS | ,311 | ,065 | ,350 | 4,766 | ,000 | ,807 | 1,239 |
| | ESTRPERS | 1,113 | ,194 | ,424 | 5,735 | ,000 | ,796 | 1,256 |
| | QUEST | -,043 | ,023 | -,150 | -1,881 | ,062 | ,680 | 1,471 |

a. Variabile dipendente: PosCOP

# Inserimento dei predittori: Avanti

- **Forward** o *In avanti*: Le variabili X vengono inserite una alla volta (in genere la X con la correlazione XY più alta) e vengono poi calcolate le correlazioni parziali e i test di significatività di tutte le altre.
- Una nuova variabile viene inserita se risulta statisticamente associata al modello
- Ci si ferma quando non ci sono più variabili significative

*Coefficienti*[a]

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | | | Statistiche di collinearità | |
|---|---|---|---|---|---|---|---|---|
| | | B | Errore standard | Beta | t | Sign. | Tolleranza | VIF |
| 1 | (Costante) | 9,216 | 1,504 | | 6,127 | ,000 | | |
| | ESTRPERS | 1,501 | ,198 | ,572 | 7,566 | ,000 | 1,000 | 1,000 |
| 2 | (Costante) | 4,320 | 1,566 | | 2,758 | ,007 | | |
| | ESTRPERS | 1,259 | ,180 | ,479 | 7,000 | ,000 | ,947 | 1,055 |
| | INTRINS | ,358 | ,061 | ,403 | 5,883 | ,000 | ,947 | 1,055 |

a. Variabile dipendente: PosCOP

# Inserimento dei predittori: Indietro

- **Backword** o *Indietro*: Le X vengono inserite tutte assieme e poi pian piano tolte se non risultano significative al t-test
- Ci si ferma quando tutte le non significative sono state tolte

*Coefficienti*[a]

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | | | Statistiche di collinearità | |
|---|---|---|---|---|---|---|---|---|
| | | B | Errore standard | Beta | t | Sign. | Tolleranza | VIF |
| 1 | (Costante) | 7,846 | 2,432 | | 3,226 | ,002 | | |
| | INTRINS | ,311 | ,065 | ,350 | 4,766 | ,000 | ,807 | 1,239 |
| | ESTRPERS | 1,113 | ,194 | ,424 | 5,735 | ,000 | ,796 | 1,256 |
| | QUEST | -,043 | ,023 | -,150 | -1,881 | ,062 | ,680 | 1,471 |
| 2 | (Costante) | 4,320 | 1,566 | | 2,758 | ,007 | | |
| | INTRINS | ,358 | ,061 | ,403 | 5,883 | ,000 | ,947 | 1,055 |
| | ESTRPERS | 1,259 | ,180 | ,479 | 7,000 | ,000 | ,947 | 1,055 |

a. Variabile dipendente: PosCOP

# Inserimento dei predittori: a passi

- **Stepwise** o *per passi*: Si parte con "alcune" variabili X o con la X maggiormente associata alla Y
- Le altre X vengono inserite, tolte o ignorate a seconda della loro importanza e significatività statistica aggiuntiva
- Il modello finale identificato "dovrebbe" essere il migliore (ma è molto legato alle caratteristiche del campione...)

# Ordine dei predittori

- Se le variabili X sono pochissimo correlate fra loro, l'ordine ha poca importanza
- Se sono correlate, usare le variabili maggiormente associate ad Y in letteratura
- Inserendole nell'ordine decrescente di efficacia indicata in letteratura
- **Preferibilmente**, usare l'approccio gerarchico
- Si inseriscono nel primo blocco le variabili maggiormente associate in letteratura/teoria
- successivamente si inseriscono altre variabili (in blocchi successivi)

# Sommario locale

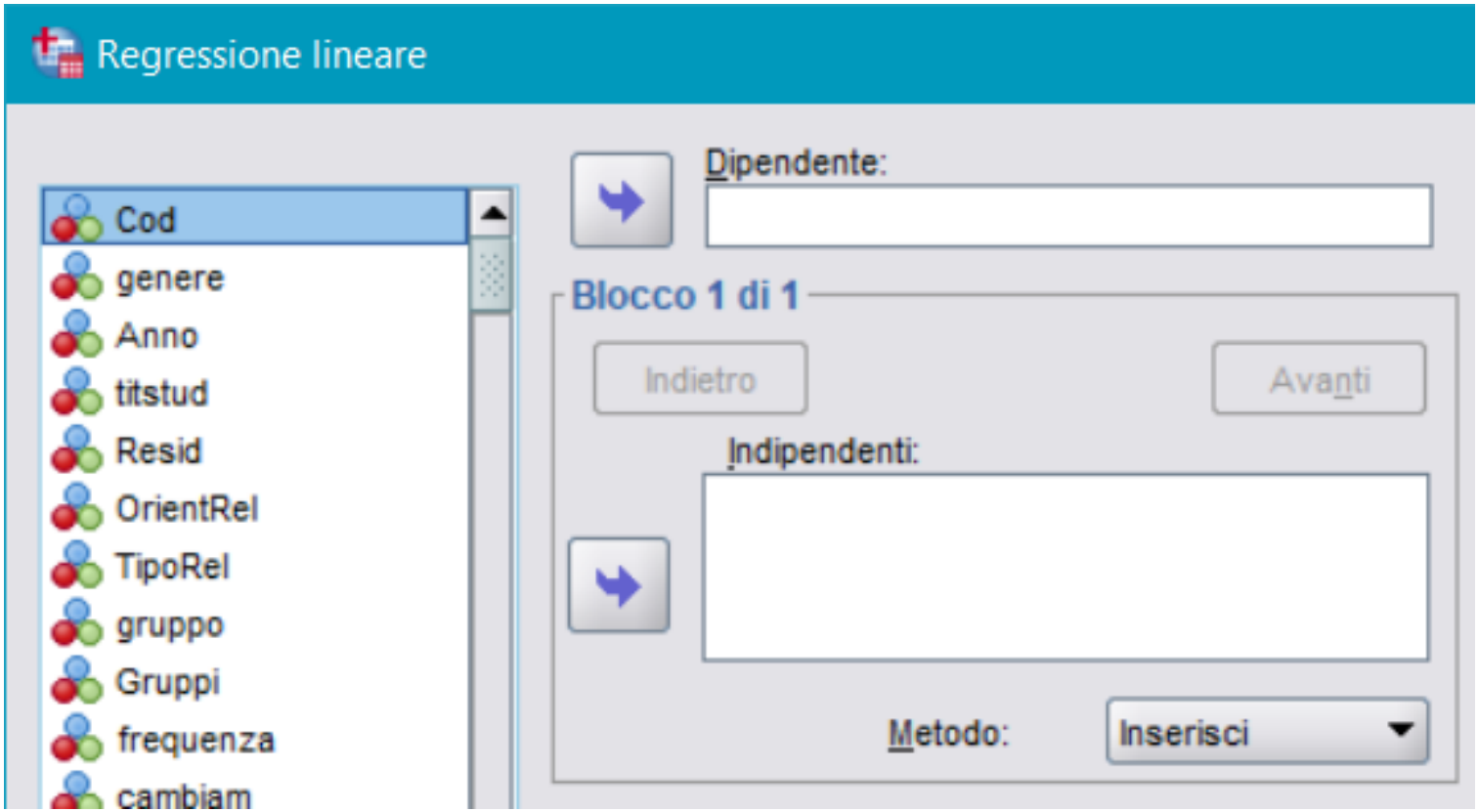
1 Regressione lineare multipla
- Introduzione generale
- Influenze fra variabili
- Varianza spiegata o ampiezza dell'effetto

2 In SPSS

3 Correlazioni multiple e parziali

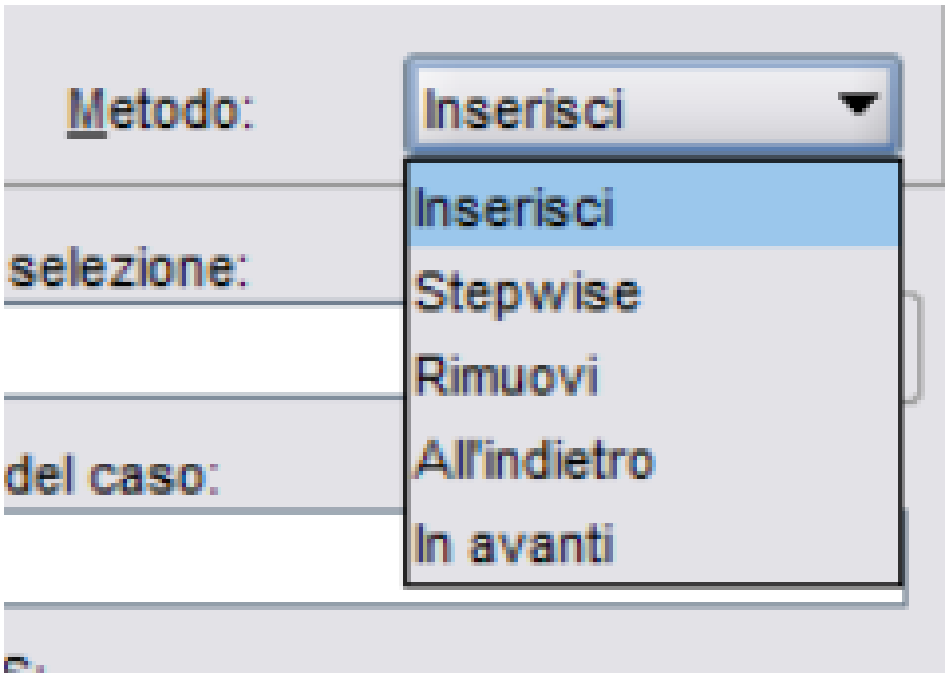
# In SPSS

- `Analizza | Regressione | Lineare...`
- Trasferite in `Dipendente` la variabile che volete spiegare (la Y)
- Trasferite in `Indipendenti` tutte le variabili che volete usare per spiegare (le diverse X)



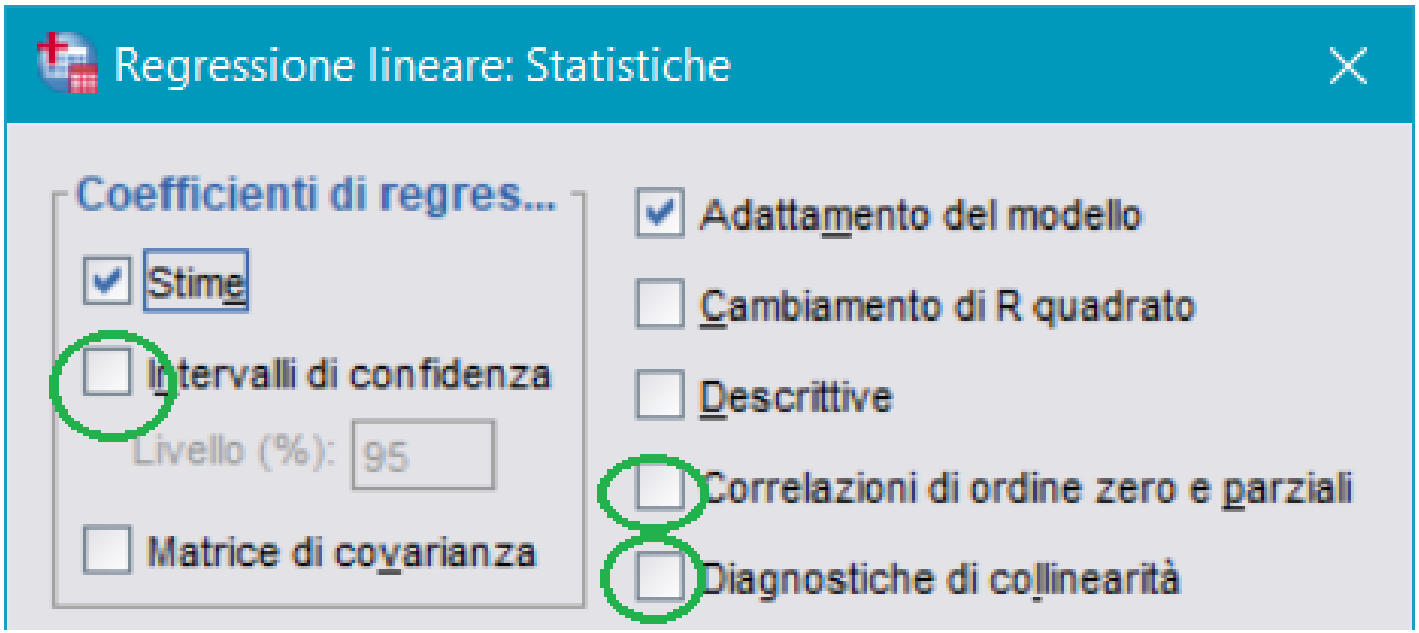
# In SPSS

- Per una **regressione standard**: verificare che `Metodo` sia impostato su `Inserisci`
- Per una **forward**: impostare `Metodo` su `In avanti`
- Per una **backward**: impostare `Metodo` su `All'indietro`
- Per una **stepwise**: impostare `Metodo` su `Per passi` (v.24)



# In SPSS

- Click-ate su `Statistiche...`
- Selezionate almeno: *Diagnostiche di collinearità* e *Correlazioni di ordine zero e parziali*
- A piacere selezionate *Intervalli di confidenza*
- Poi `Continua` e `OK`



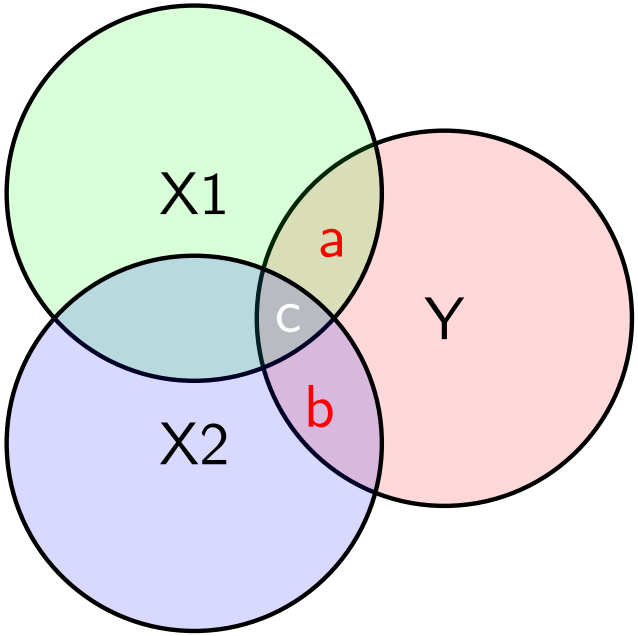
# Sommario locale

1 Regressione lineare multipla
- Introduzione generale
- Influenze fra variabili
- Varianza spiegata o ampiezza dell'effetto

2 In SPSS

3 Correlazioni multiple e parziali

# Correlazioni multiple e parziali




- Le intersezioni dei cerchi rappresentano le covarianze in comune fra le diverse variabili
- Con tre variabili, $R^2$ corrisponde a $(a + b + c)$ cioè la varianza che $Y$ ha in comune con $X_1$ e $X_2$
- la correlazione multipla è quindi $\sqrt{R^2} = \sqrt{a + b + c}$; tutto il resto dell'area di $Y$ è la varianza non spiegata $(1 - R^2)$

- Correlazione **parziale**: sovrapposizione fra $Y$ e $X_1$ $(\sqrt{a + c})$ o fra $Y$ e $X_2$ $(\sqrt{b + c})$
- Correlazione **semiparziale**: solo la parte unica di sovrapposizione fra $Y$ e $X_1$ $(\sqrt{a})$ o fra $Y$ e $X_2$ $(\sqrt{b})$

# Correlazione multipla

- È la correlazione di una variabile (Y) con 2 o più variabili ($X_1$, $X_2$, ...) contemporaneamente (oscilla fra $-1$ e $+1$)

$$R = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2 r_{12} r_{13} r_{23}}{1 - r_{23}^2}}$$

| r | 1-RSE | 2-SWLS |
|---|---|---|
| 2-SWLS | .520 | - |
| 3-RifAccet | .634 | .593 |

dove
$r_{12}$ è la correlazione fra le variabili 1 e 2;
$r_{13}$ fra la 1 e la 3... e
$r_{1.23}$ è la correlazione multipla (R)

$$R = \sqrt{\frac{.520^2 + .634^2 - 2 \times .520 \times .634 \times .593}{1 - .593^2}} = .659$$

In SPSS si ottiene solo come sottoprodotto della regressione lineare multipla

# Correlazione parziale

- È la correlazione di due variabili a cui viene "tolta" (contemporaneamente) l'influenza di una terza variabile.
- Es. correlazione fra "numero di parole conosciute" e "intelligenza" parzializzata in base all'età (tolto il contributo dell'età). Più l'età è correlata con una delle due, più la correlazione diminuirà.

$$pr_{12.3} = r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

dove
$r_{12}$ è la correlazione fra le variabili 1 e 2;
$r_{13}$ fra la 1 e la 3... e
$r_{12.3}$ è la correlazione fra 1 e 2 parzializzata sulla 3

$$R_p = \frac{.520 - .634 \times .593}{\sqrt{(1 - .634^2)(1 - .593^2)}} = .231$$

In SPSS si ottiene in *Correlazione parziale* o in *Regressione lineare multipla*

# Correlazione semi-parziale

- È la correlazione fra due variabili, ma solo ad una delle due è stato tolto il contributo di una terza.
- Es. correlazione fra "numero di parole conosciute" e "intelligenza". La parzializzazione in base all'età viene attuata solo con il numero di parole.

$$sr_{12.3} = r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^2}} \qquad R_{sp} = \frac{.520 - .634 \times .593}{\sqrt{1 - .593^2}} = .178$$

dove $r_{12}$ è la correlazione fra le variabile 1 e 2, $r_{13}$ fra la prima e la terza e così via

In SPSS si ottiene solo in *Regressione lineare multipla*

# Correlazioni parziali in SPSS

1. `Analizza | Correlazione | Parziale...`
2. `Analizza | Regressione | Lineare...`, bottone `Statistiche`, segnare `Correlazioni di ordine zero e parziale`

1. Calcola la correlazione di più variabili, parzializzate su un'altra variabile (identica per tutte)
2. Mentre effettua una regressione multipla, calcola la correlazione (ordine 0) fra ogni singola indipendente con la Y e quindi parzializza questa correlazione con tutte le altre indipendenti (parziale e semiparziale [parziale indipendenti])

# Correlazioni parziali in SPSS

*Riepilogo del modello*

| Modello | R | R-quadrato | R-quadrato adattato | Errore std. della stima |
|---|---|---|---|---|
| 1 | ,659[a] | ,434 | ,425 | 3,306 |

a. Predittori: (costante), RIFAccet, SWLS

*Coefficienti*[a]

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | | |
|---|---|---|---|---|---|---|
| | | B | Errore standard | Beta | t | Sign. |
| 1 | (Costante) | 19,549 | 1,551 | | 12,607 | ,000 |
| | SWLS | ,178 | ,070 | ,223 | 2,559 | ,012 |
| | RIFAccet | ,744 | ,129 | ,502 | 5,767 | ,000 |

a. Variabile dipendente: RSE

| 95,0% Intervallo di confidenza per B | | Correlazioni | | | Statistiche di collinearità | |
|---|---|---|---|---|---|---|
| Limite inferiore | Limite superiore | Ordine zero | Parziale | Parte | Tolleranza | VIF |
| 16,477 | 22,620 | | | | | |
| ,040 | ,316 | ,520 | ,232 | ,179 | ,649 | 1,542 |
| ,489 | 1,000 | ,634 | ,474 | ,404 | ,649 | 1,542 |

*Correlazioni*

| | | RSE | SWLS | RIFAccet |
|---|---|---|---|---|
| Correlazione di Pearson | RSE | 1,000 | ,520 | ,634 |
| | SWLS | ,520 | 1,000 | ,593 |
| | RIFAccet | ,634 | ,593 | 1,000 |

$$R = \sqrt{\frac{.520^2 + .634^2 - 2 \times .520 \times .634 \times .593}{1 - .593^2}} = .659$$

$$R_p = \frac{.520 - .634 \times .593}{\sqrt{(1 - .634^2)(1 - .593^2)}} = .231$$

$$R_{sp} = \frac{.520 - .634 \times .593}{\sqrt{1 - .593^2}} = .178$$

# Residui

- I residui ($e = Y - \hat{Y}$) dovrebbero essere dispersi casualmente attorno a Y
- Se **non** sono dispersi casualmente, esiste un'altra variabile X che può spiegarne una parte, oppure la relazione non è lineare