Psicometria

6-Regressione lineare multipla
vers. 1.4

Germano Rossi[1]
germano.rossi@unimib.it
Giovanni Battista Flebus[1]
giovannibattista.flebus@unimib.it

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

2008-2008

Regressione lineare multipla

- Analoga a quella semplice
- Una sola variabile dipendente (Y) o da spiegare
- Due o più variabili indipendenti (X) o predittive, esplicative

Regressione lineare semplice (1 dip, 1 indip)

$$Y_i = a + bX_i + e_i$$

intercetta pendenza indipendente errore

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + e_i$$

Regressione lineare multipla (2 indip, 1 dip)

Regressione multipla matriciale

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

Y	X_1	X_2	$Y = b_0 + b_1 X_1 + b_2 X_2 + e$
3	2	1	$3 = 1b_0 + 2b_1 + 1ba_2 + e_1$
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	$8 = 1b_0 + 8b_1 + 7b_2 + e_5$

Regressione multipla matriciale

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

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

Regressione multipla matriciale

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

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

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

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

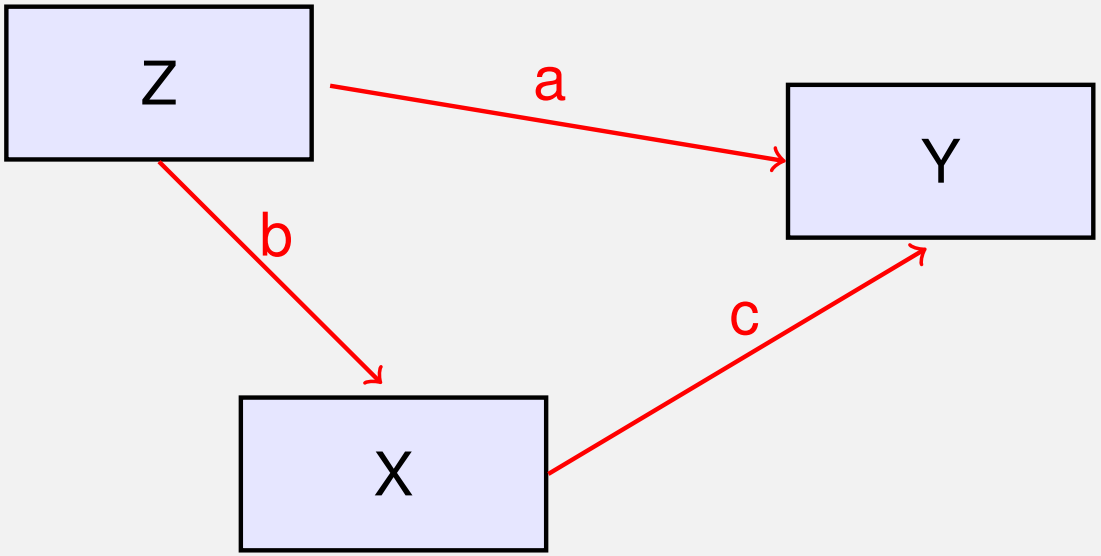
Regressione multipla matriciale

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

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

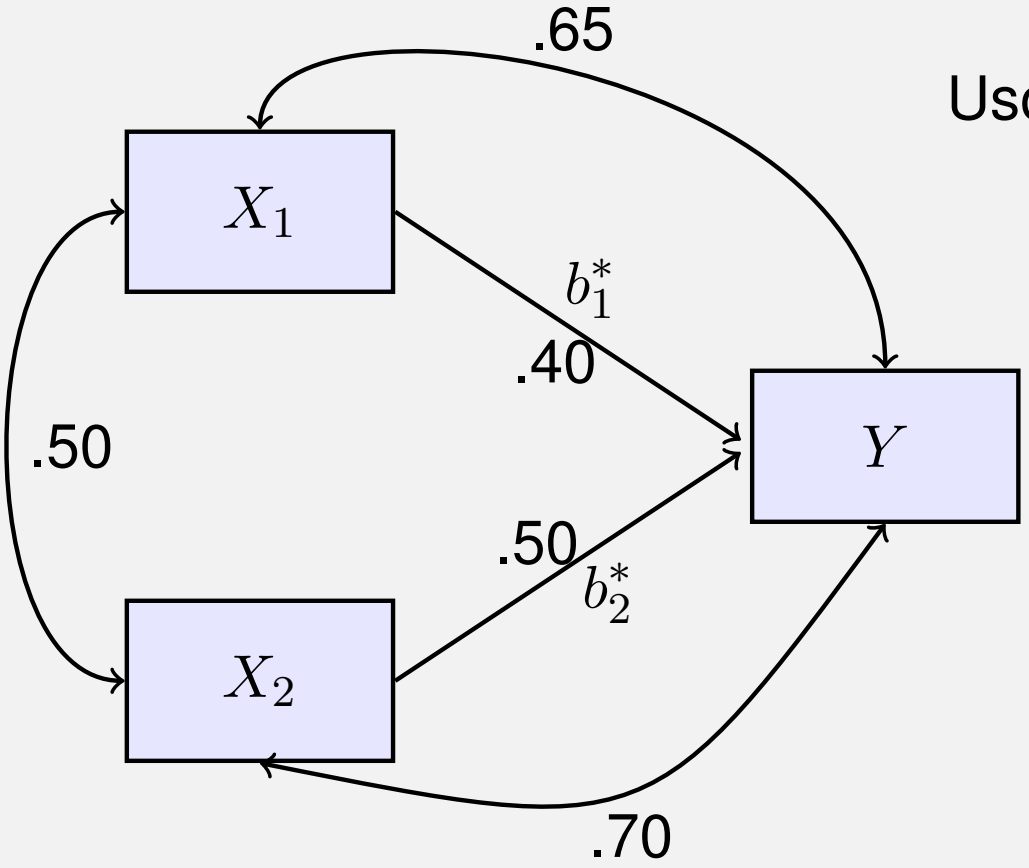
$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} N & \sum X_1 & \cdots & \sum X_n \\ \sum X_1 & \sum X_1^2 & \cdots & \sum X_1 X_n \\ \cdots & \cdots & \cdots & \cdots \\ \sum X_n & \sum X_1 X_n & \cdots & \sum X_n^2 \end{bmatrix} \qquad \mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \cdots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

Percorsi causali/relazionali



- **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 covarianze
- Il valore di un'influenza indiretta è pari al prodotto delle influenze semplici

Percorsi causali/relazionali



Uso 1 anziché X_1 e 2 anziché X_2

$$r_{1y} = b_1^* + b_2^* r_{12}$$

$$r_{2y} = b_2^* + b_1^* r_{12}$$

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

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

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

Percorsi causali/relazionali

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

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

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

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

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

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

Regressione matriciale

- Ci sono tre formule alternative

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

- C_{xx} è la matrice delle varianze/covarianze fra le X
- c_{yx} è il vettore delle covarianze fra la Y e le X
- R_{xx} è la matrice delle correlazioni fra le X
- r_{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_1Y	X_2Y	X_1X_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

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

Esempio con le covarianze

Varianze e covarianze calcolate con $N - 1$

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

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

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

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

Esempio con le correlazioni

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

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

$$b_0 = 0$$

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

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

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

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

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

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

Test di significatività

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

a) un test globale: che include tutte le variabili

Se il test globale è significativo

b) un test per ciascuna variabile indipendente (per ogni X)

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

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

Test globale

$$F = \frac{(R^2_f - R^2_r)/(d_f - d_r)}{(1 - R^2_f)/d_f}$$

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

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

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

k=numero di variabili indipendenti (X)

R^2 tende ad aumentare al numero delle X, quindi

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

Test globale

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

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

$$F = \frac{.824/6}{(1-.824)/(287-6-1)} = 218.484 \quad F = \frac{48461.740}{221.318} = 218.969$$

Test per ciascuna X

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: $b_2 = 0$

Usiamo 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.

$$t = \frac{b_i}{s_{y.i}} \qquad s_{y.i} \text{ è l'err. della stima}$$

Test per ciascuna X

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

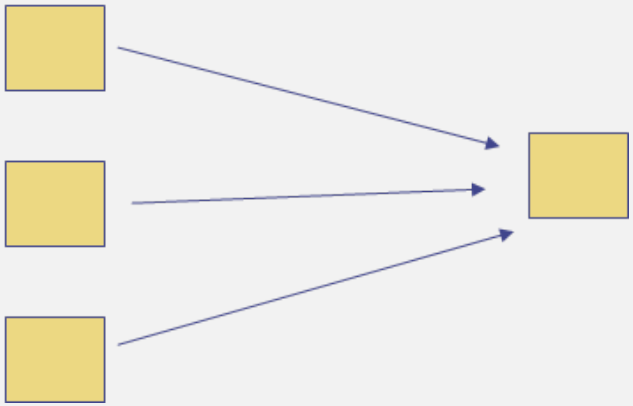
a. Variabile dipendente: Fondam

$$t = \frac{2.195}{0.177} = 12.385$$

Ci sono 2 var che non servono (Sig > .05): oEsPers e oEsSoc

Multicollinearità

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

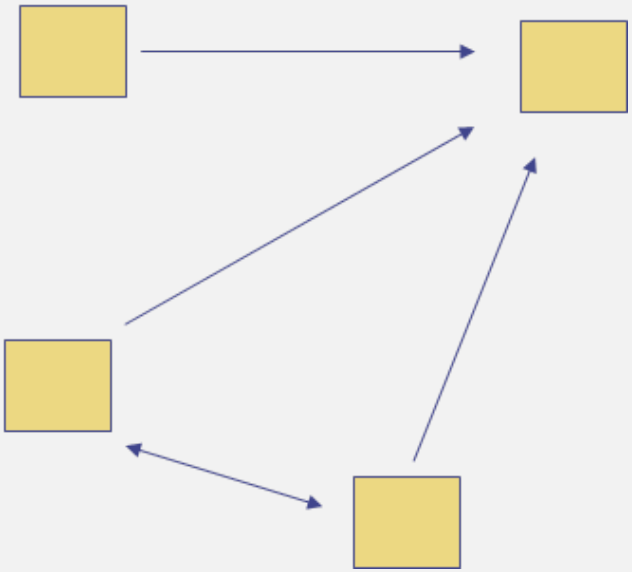


	X_1	X_2	X_3
Y	.60	.50	.70
X_1		**.20**	.30
X_2			.20

Idealmente, le correlazioni tra le X, dovrebbero essere 0; in questo modo b^* coinciderà con r e non con r parzializzato

Multicollinearità

Spesso però, 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*.

Multicollinearità

Problemi

- fa diminuire la R multipla
- l'effetto dei predittori si confonde
- aumenta la varianza e l'instabilità dell'equazione

Diminuire la multicollinearità

- combinare fra loro i predittori altamente correlati (ad esempio sommandoli)
- se ci sono molti predittori altamente correlati, usare un'analisi delle componenti principali per ridurre il numero delle X

Multicollinearità

	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 $\hat{Y} = b_0 + b_2 V$

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

Ipotizzando due regressioni semplici (separate):

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

Scegliere i predittori

- **Usare la teoria**: Si usano solo variabili teoricamente sensate; la sensatezza può essere ricavata, oltre che dalla logica, da una ricerca bibliografica
- Metodi semi-automatici sequenziali
- *Standard*: Tutte le variabili vengono inserite
- *Forward*: Le variabili X vengono inserite una alla volta
- *Backward*: Tutte le variabili vengono inserite e poi cancellate una alla volta
- *Stepwise*: Si inizia con un blocco di variabili e poi si inseriscono o si tolgono una alla volta

Scegliere i predittori

- Regressione **standard**: Tutte le variabili X vengono considerate assieme e tutti i coefficienti di regressione (B o beta) stimati contemporaneamente (come negli esempi che noi abbiamo calcolato con 2 X)
- **Forward**: 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 variabili significative

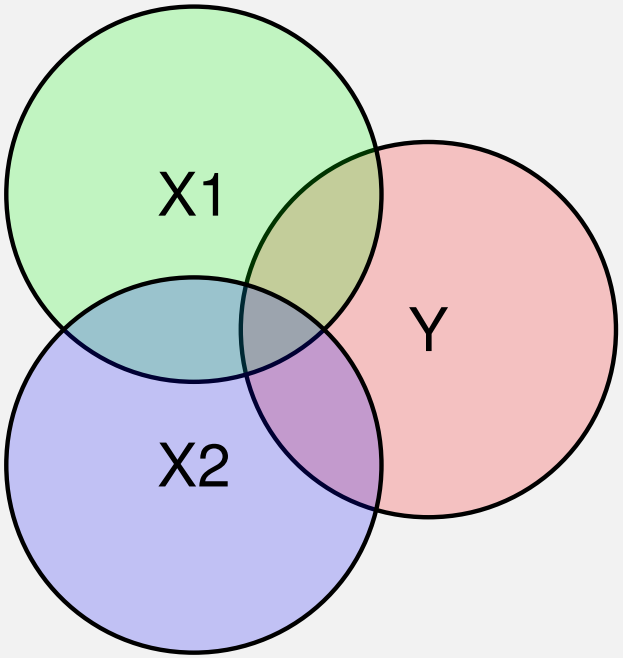
Scegliere i predittori

- **Backword**: 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
- **Stepwise**: 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 empirico...)

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)
- Per una **regressione standard**: verificare che `Metodo` sia impostato su `Per blocchi`
- Per una **forward**: impostare `Metodo` su `Avanti`
- Per una **backward**: impostare `Metodo` su `Indietro`
- Per una **stepwise**: impostare `Metodo` su `Per passi`
- Date l'`OK`

Correlazioni multiple e parziali



- Le intersezioni rappresentano le co-varianze in comune fra le variabili
- L'area in comune fra due variabili abbiamo la correlazione r (di ordine 0)
- L'area in comune fra tre variabili è la correlazione multipla
- le correlazioni parziali e semiparziali sono le aree in comune a due variabili ma non alla terza

Correlazione multipla

- E' la correlazione di una variabile con 2 o più variabili contemporaneamente
- Oscilla fra -1 e $+1$ come la correlazione di Pearson (come tutti gli indici di correlazione)

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

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

Correlazione parziale

- E' 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). Se l'età è correlata con una delle due, 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

Correlazione semi-parziale

- E' 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.

$$sr_{12.3} = r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^2}}$$

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

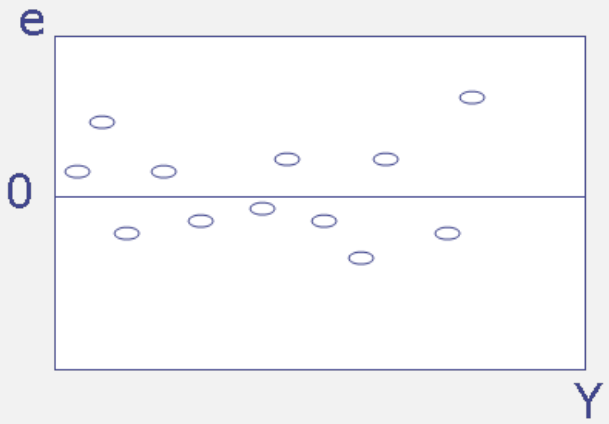
Correlazioni parziali in SPSS

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])

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