

Introduzione alla regressione multipla

Che cos'è la regressione multipla

- A differenza della regressione semplice, utilizza più variabili indipendenti
- Usa più coefficiente angolari, uno per ciascuna VI

Restano le caratteristiche dell'equazione di predizione:

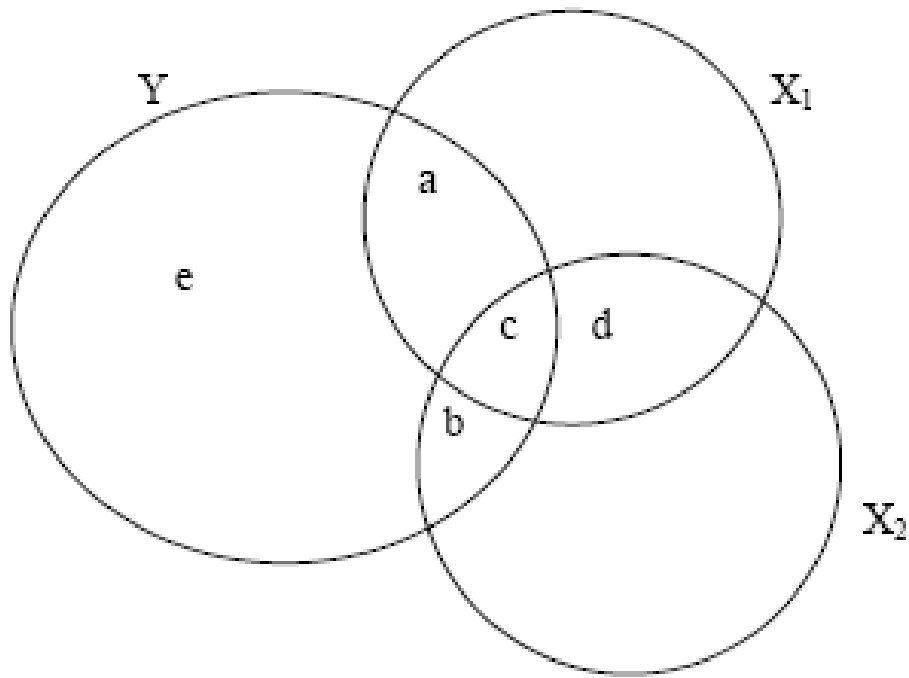
- a) Criterio dei minimi quadrati per gli errori
- b) Una sola intercetta
- c) Errori o residui calcolabili nello stesso modo
- d) Verifica della significatività per ciascuna VI

Diagramma di Venn per
la **varianza** di VI e VD

I coefficienti parziali e
semiparziali

Difficoltà del passaggio da una a più VI

- Correlazioni fra VI
- Calcolo dei coefficienti angolari
- Introduzione del concetto di **correlazione parziale**



Varianza di Y spiegata dalla regressione = $a + b + c$

Varianza di Y non spiegata dal modello di regressione = e

$$R^2_{Y12} = \frac{(a + b + c)}{(a + b + c + e)}$$

Semiparziale per $X_1 = a / (a + b + c + e)$

Semiparziale per $X_2 = b / (a + b + c + e)$

Parziale per $X_1 = pr_{21} = a / (a + e)$

Parziale per $X_2 = pr_{22} = b / (b + e)$

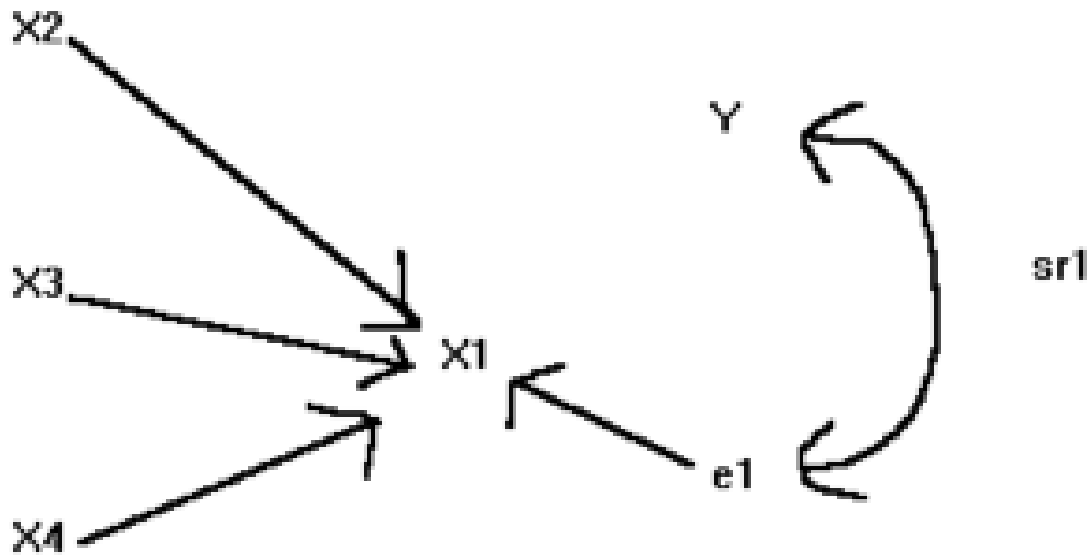
Definizione di correlazione parziale

- È la correlazione fra la variabile X e Y al netto di una terza K.
- Al netto significa: facendo in modo che la variabile K non abbia nessuno effetto su X e Y

Oppure

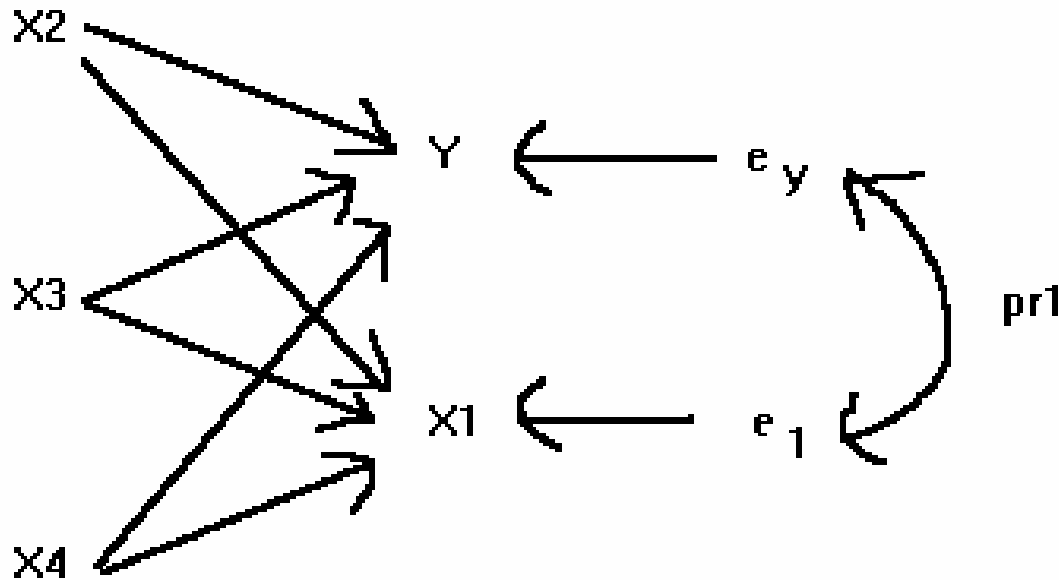
Che X e Y siano indipendenti da K

Correlazione semi-parziale o indipendente



- X2 x3 e x4 predicano x1. I residui correlati con Y (VD) sono la correlazione **semiparziale (sr1) o parziale indipendente**

Correlazione parziale



- x_2 , x_3 e x_4 predicano x_1 .
- X_2 , x_3 e x_4 predicano Y .
- I residui delle due predizioni, messi in correlazione, producono la correlazione **parziale ($pr1$)**

Intepretazione dei coefficienti parziale e semiparziali

- Il coefficiente parziale informa in che modo la VI x è correlata con la VD tenendo costante l'influsso delle altre VI
- Il coefficiente semiparziale è invece interpretabile più semplicemente in riferimento a R^2 : il quadrato del coefficiente semiparziale è uguale all'incremento di R attribuibile alla VI.

Esempio

Per predire il Numero di esercizi a scuola, usiamo un primo modello (regressione semplice)

- Modello 1

Esercizi per casa → Esercizi a scuola

Aggiungiamo anche le Ore di studio

- Modello 2

Esercizi per casa →

Ore di studio → Esercizi a scuola

La varianza spiegata aumenta

Model Summary

Model	R	R Square
1	,663 ^a	,440
2	,889 ^b	,791

a. Predictors: (Constant), Esercizi_casa

b. Predictors: (Constant), Esercizi_casa, Ore_studio

Coefficients^a

Model		Unstandardized Coefficients		Stand Coeff	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-.49	1.94		-.25	,807			
	Esercizi_casa	.80	.30	.66	2.66	,026	,663	,663	,663
2	(Constant)	.35	1.28		.27	,791			
	Esercizi_casa	.23	.25	.19	.93	,378	,663	,313	.151
	Ore_studio	.53	.15	.76	3.67	,006	,876	,792	.593

a. Dependent Variable: Num_esercizi

Model Summary

Model	R	R Square
1	,663 ^a	,440
2	,889 ^b	,791

a. Predictors: (Constant), Esercizi_casa

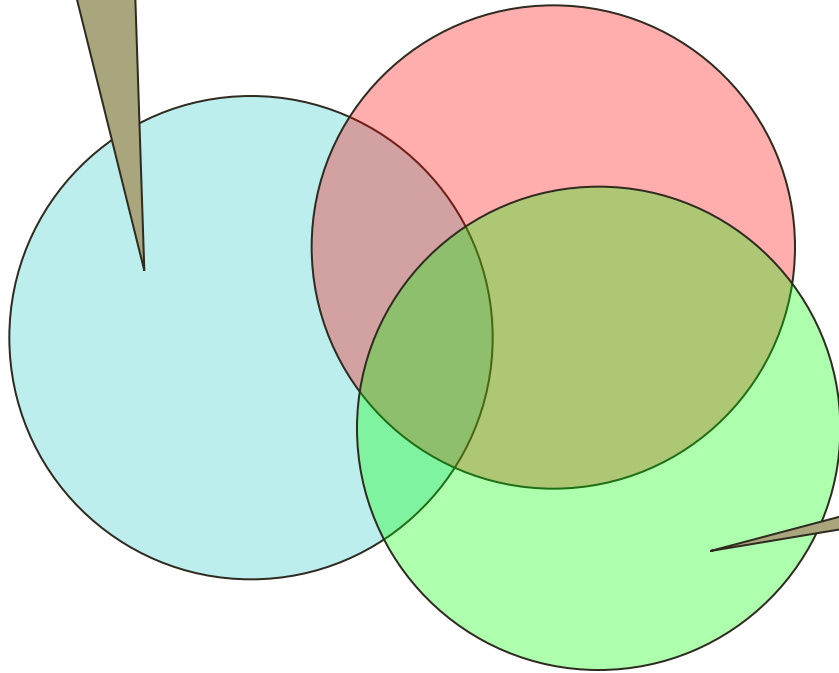
b. Predictors: (Constant), Esercizi_casa, Ore_studio

$$0,791 - 0,440 = 0,351$$

$$0,593^2 = 0,351$$

Aggiungendo le OreStudio si può spiegare 35% di varianza in più

Esercizi
svolti



Esercizi a
casa

Ore di
studio

Esempio terzo

- Aggiungiamo nel modello anche il numero di libri che uno studente possiede a casa

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,663 ^a	,440	,377	1,846	,440	7,060	1	9	,026
2	,889 ^b	,791	,739	1,196	,351	13,437	1	8	,006
3	,943 ^c	,889	,842	,930	,098	6,226	1	7	,041

a. Predictors: (Constant), Esercizi_casa

b. Predictors: (Constant), Esercizi_casa, Ore_studio

c. Predictors: (Constant), Esercizi_casa, Ore_studio, Libri_casa

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,663 ^a	,440	,377	1,846	,440	7,060	1	9	,026
2	,889 ^b	,791	,739	1,196	,351	13,437	1	8	,006
3	,943 ^c	,889	,842	,930	,098	6,226	1	7	,041

- a. Predictors: (Constant), Esercizi_casa
 b. Predictors: (Constant), Esercizi_casa, Ore_studio
 c. Predictors: (Constant), Esercizi_casa, Ore_studio, Libri_casa

L'aumento nella predizione è quasi uguale al 10%

Il livello di significatività è inferiore a 0,05, rifiutiamo l'ipotesi nulla di mancanza di effetto.

Coefficients^a

	Unstandardized Coefficients		standard coeff	t	Sig.	Correlations		
	B	Std. Error	Beta			Zero-order	Partial	Part
1 (Constant)	-.49	1.94		-.25	,807			
Esercizi_casa	.80	.30	.66	2.66	,026	,663	,663	,663
2 (Constant)	.35	1.28		.27	,791			
Esercizi_casa	.23	.25	.19	.93	,378	,663	,313	,151
Ore_studio	.53	.15	.76	3.67	,006	,876	,792	,593
3 (Constant)	-11.30	4.77		-2.37	,050			
Esercizi_casa	.66	.26	.55	2.55	,038	,663	,694	,321
Ore_studio	.42	.12	.59	3.39	,012	,876	,788	,426
Libri_casa	.24	.10	.42	2.50	,041	,023	,686	,314

a. Dependent Variable: Num_esercizi

La correlazione di ordine zero è nulla, i libri a casa hanno un effetto **nascosto** sugli esercizi svolti nel compito, che appare solo in presenza di altre VI

Formula del coefficiente di Correlazione Parziale fra X e Y, al netto dell'effetto di Z

$$r_{xy.z} = \frac{r_{xy} - (r_{xz} \times r_{yz})}{\sqrt{1 - r_{xz}^2} \times \sqrt{1 - r_{yz}^2}}$$

Correlazione parziale fra x e y

correl fra x e y

correl fra x e z

correl fra y e z

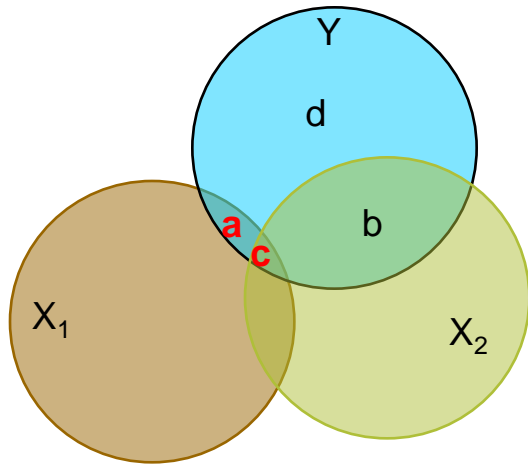
Quadrato della correl fra x e z

Quadrato della correl fra y e z

Rapporto fra la correlazione parziale e la correlazione semplice

Parziale può essere uguale a semplice

Quando "a" and "c" Sono piccoli:
Semplice = Parziale

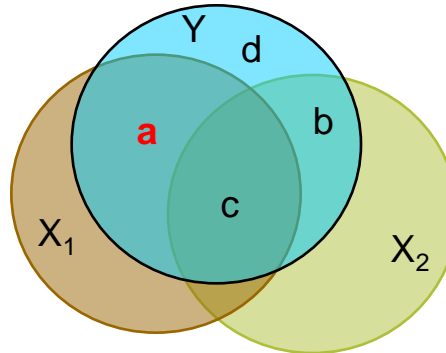


$$\frac{b}{b+d} \approx \frac{b+c}{a+b+c+d}$$

Il motivo più frequente:
X₁ è scarsamente correlato con Y

Parziale può essere maggiore di semplice

Quando "a" è grande (e "c" è grande o piccolo):
Parziale > Semplice

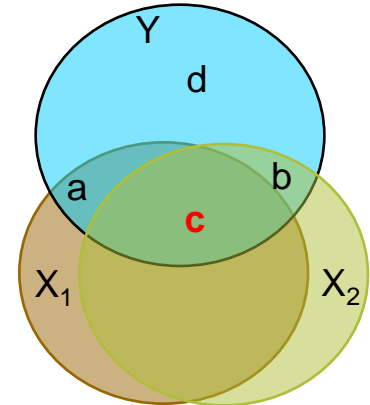


$$\frac{b}{b+d} > \frac{b+c}{a+b+c+d}$$

Il motivo più frequente :
X₁ è altamente correlato con Y

Parziale può essere più piccolo di semplice

Quando "c" è grande (e "a" non è molto grande):
Parziale < Semplice



$$\frac{b}{b+d} < \frac{b+c}{a+b+c+d}$$

Il motivo più frequente :
X₁ è altamente correlato con X₂