Introduzione alla regressione multipla

Giovanni Battista Flebus
Lezioni di psicometria

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:

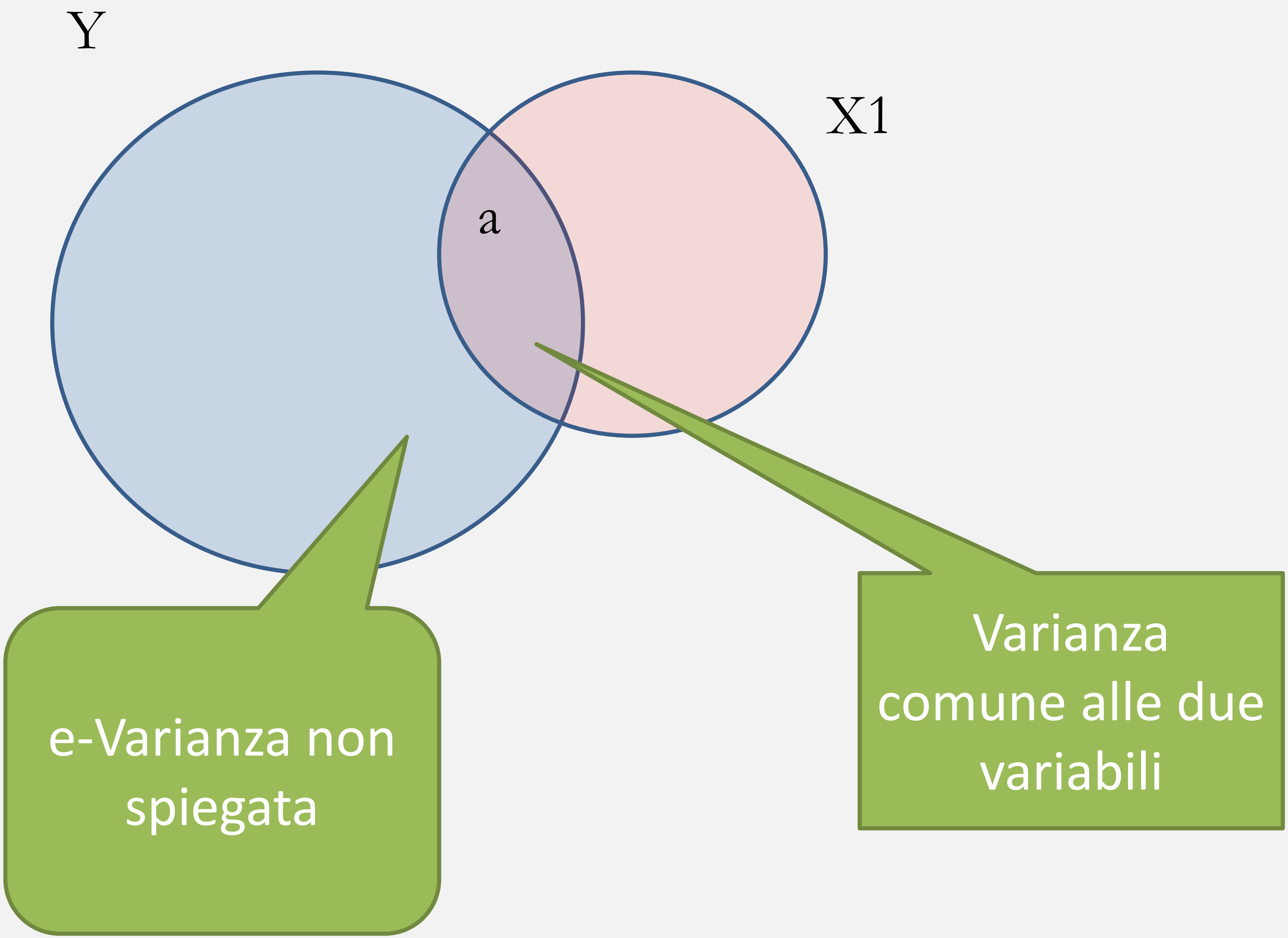
- a) Criterio dei minimi quadrati per gli errori
- b) Una sola intercetta
- c) Errori o residui calcolabili nello stesso modo
- d) Verifica della significatività per ciascuna VI

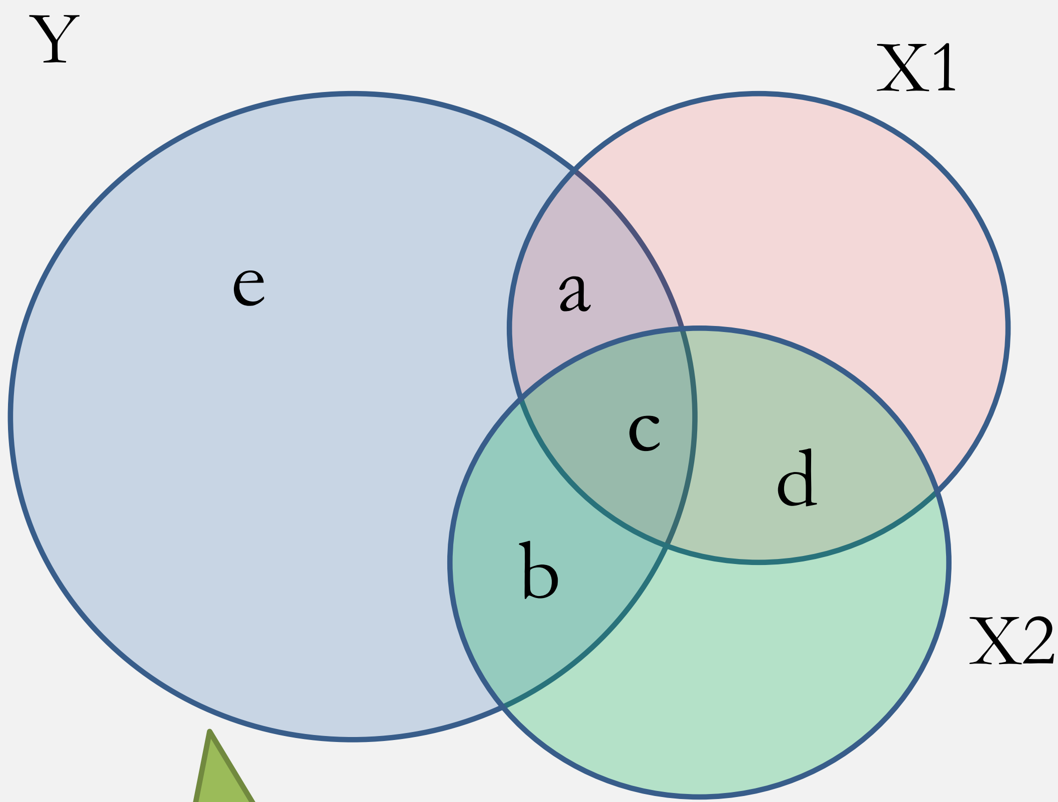
Difficoltà del passaggio da una a più VI

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

Diagramma di Venn per la varianza di VI e VD

Rappresentazione grafica dei coefficienti parziali e semiparziali


Y
X1
a
e-Varianza non spiegata
Varianza comune alle due variabili

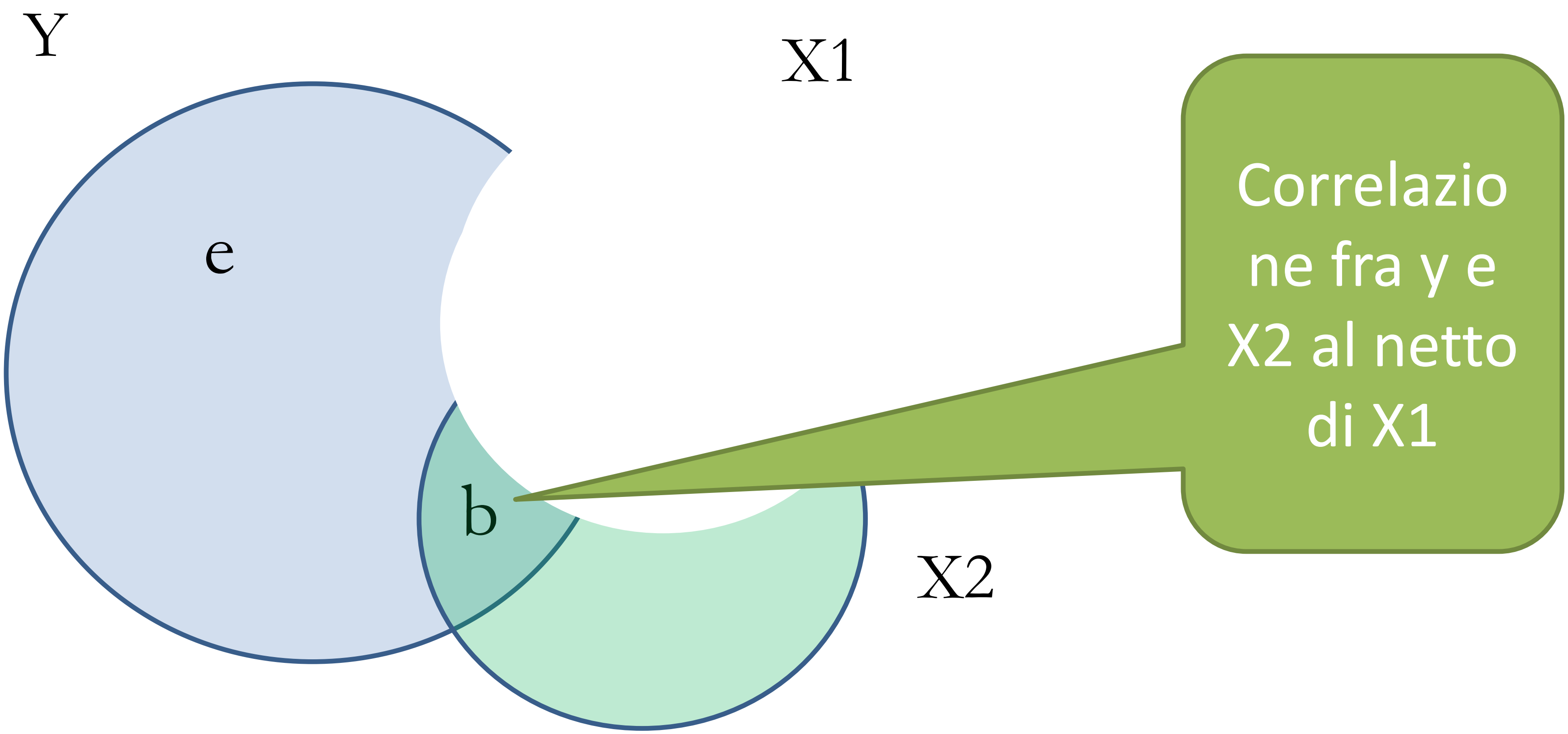


Varianza di Y spiegata dalla regressione
= a + b + c

Varianza di Y non spiegata dal modello di regressione
= e

e-Varianza non spiegata

$$R^2_{Y.12} = \frac{a+b+c}{a+b+c+e}$$



X1 predice x2, e si ottengono i residui X2- res X1
i residui di x2 non hanno nessuna correlazione con X1,

X1 predice Y e si ottengono i residui y-res X1
questi residui di Y non hanno nessuna correlazione con X1

Correlazione parziale (pr1)

$$X1 \xrightarrow{\text{Predice}} Y \longrightarrow \text{Residuo}$$

$$X1 \xrightarrow{\text{Predice}} X2 \longrightarrow \text{Residuo}$$

•I residui delle due predizioni, messi in correlazione, producono la correlazione parziale (pr1)

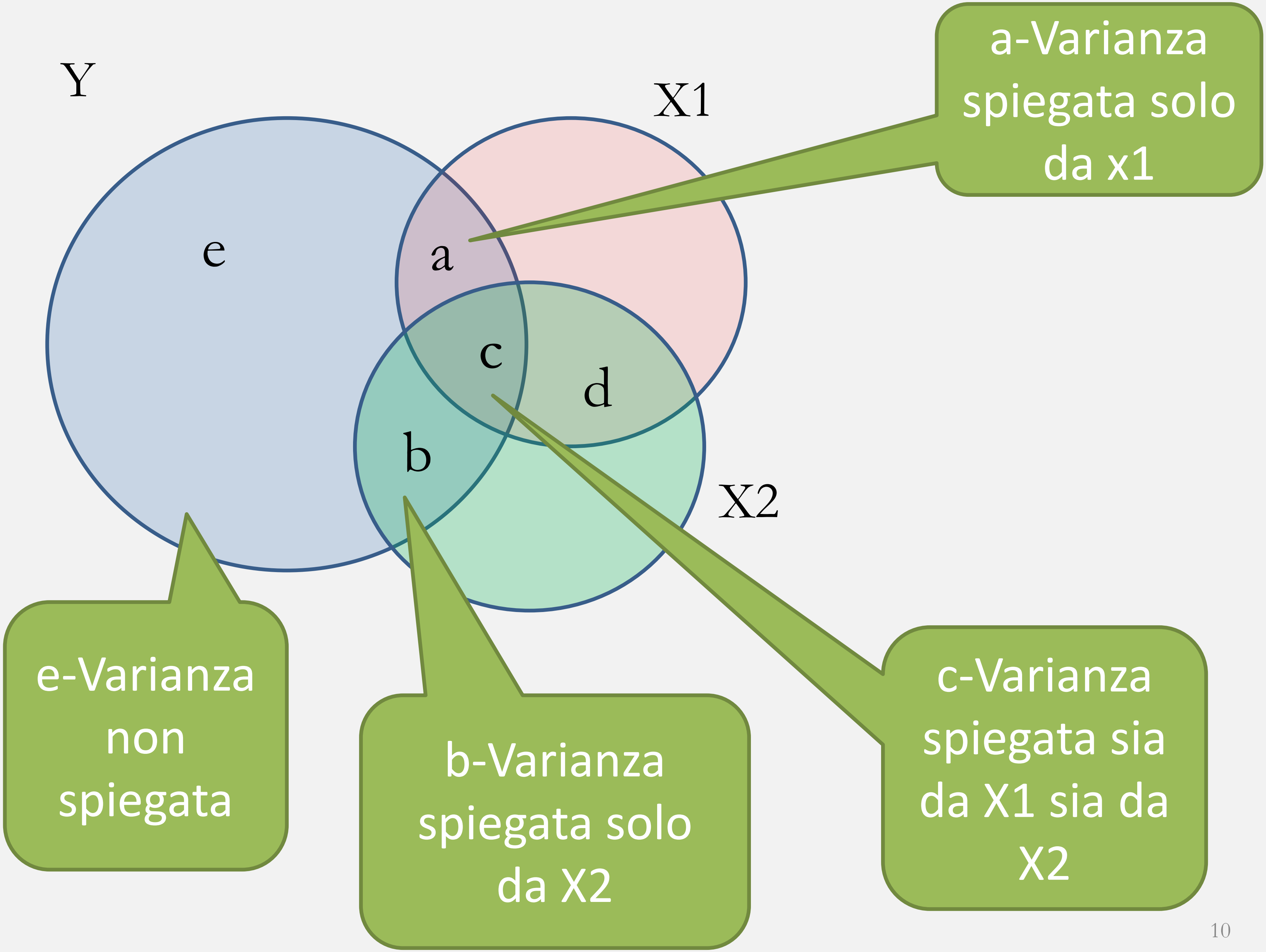
$$\text{pr} = \frac{b}{b+e}$$

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



Y
X1
X2
e
a
c
d
b
a-Varianza spiegata solo da x1
e-Varianza non spiegata
b-Varianza spiegata solo da X2
c-Varianza spiegata sia da X1 sia da X2

Correlazione semi-parziale (indipendente) Sr1

I residui correlati con Y (VD) sono la correlazione semiparziale (sr1) o parziale indipendente

La correlazione semiparziale ci informa su come la variabile X2 contribuisce **da sola** al coefficiente di correlazione multiplo
Il suo quadrato può essere tolto direttamente da R2 per indicare quanto contribuisce alla regressione

$$\mathbf{Sr} = \frac{b}{a+b+c+e}$$

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 numerico

Undici studenti ci forniscono dei dati sulla loro competenza e sul metodo di studio. Rileviamo il il numero di esercizi eseguiti correttamente in un compito in classe assieme alle ore di studio , e il numero di esercizi che hanno eseguito come preparazione a casa. Per finire aggiungere il numero di libri che hanno a casa.

Per predire il numero di esercizi eseguiti nel compito in classe, usiamo un primo modello (regressione semplice)

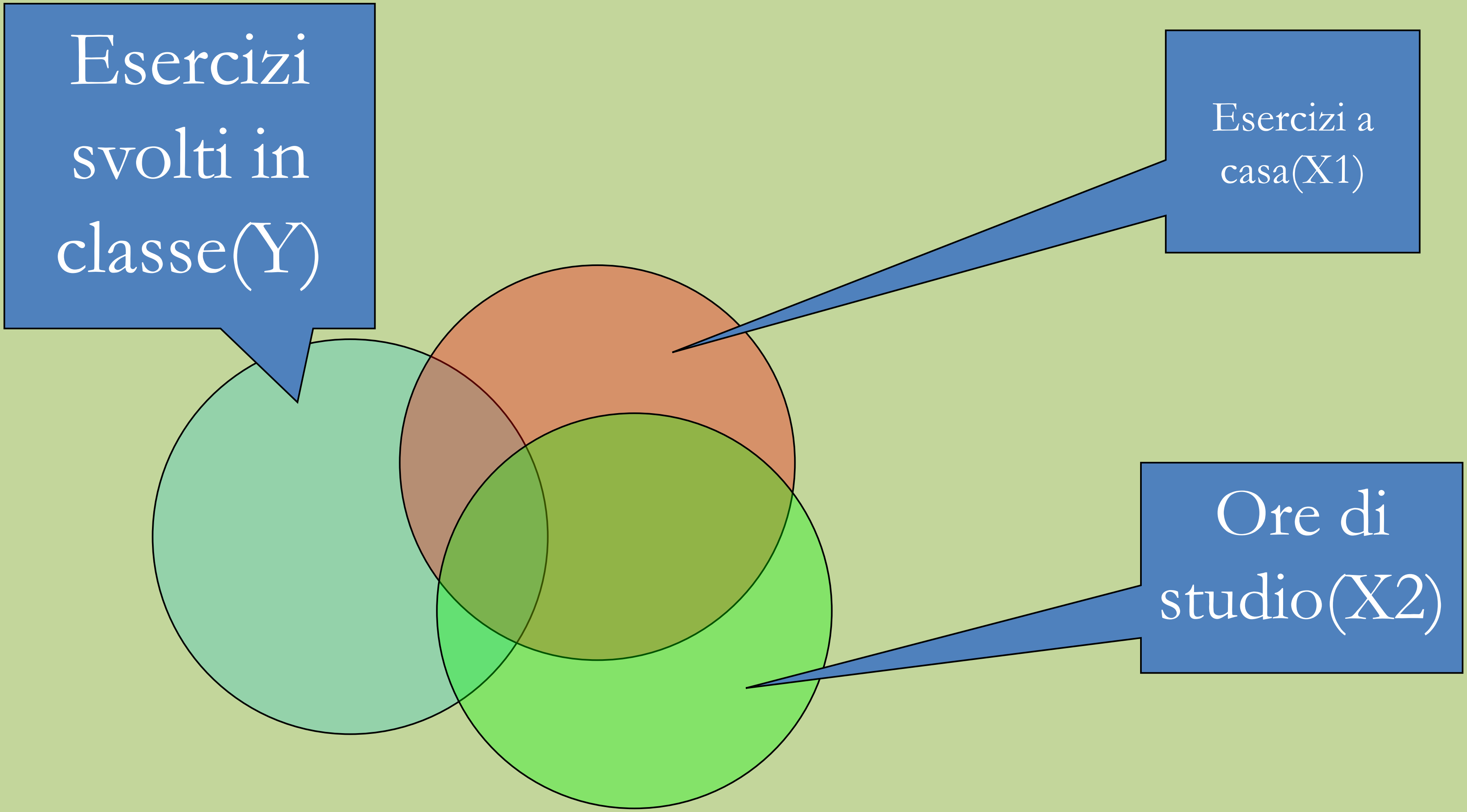
- Modello 1

Esercizi a casa→ Esercizi a scuola

Aggiungiamo anche le Ore di studio

- Modello 2

Esercizi per casa →
Ore di studio → Esercizi a scuola


Esercizi svolti in classe(Y)
Esercizi a casa(X1)
Ore di studio(X2)

La varianza spiegata aumenta del 35 %

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

Model		Unstandardized Coefficients B	Std. Error
1	(Constant)	-.49	1.94
	Esercizi_casa	.80	.30
2	(Constant)	.35	1.28
	Esercizi_casa	.23	.25
	Ore_studio	.53	.15

a. Dependent Variable: Num_esercizi

Primo modello

Numero di esercizi = 0,80 x Esercizi a casa - 0,49

R2=0,44

Secondo modello

Numero di esercizi = 0,23 x Esercizi a casa + 0,53 x ore di studio + 0,35

R2=0,79

Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Stand Coeff Beta
1	(Constant)	-.49	1.94	
	Esercizi_casa	.80	.30	.66
2	(Constant)	.35	1.28	
	Esercizi_casa	.23	.25	.19
	Ore_studio	.53	.15	.76

a. Dependent Variable: Num_esercizi

Primo modello

S-Numero di esercizi-S = 0,66 x S-Esercizi a casa

R2=0,44

Secondo modello -

S-Numero di esercizi = 0,19 x Esercizi a casa + 0,76 x S-ore di studio

R2=0,79

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

Correlazioni semi -parziali

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Stand Coeff Beta	t	Sig.	Correlations 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	$\boxed{,593}$

a. Dependent Variable: Num_esercizi

Model Summary

Model	R	R Square
1	,663[a]	$\boxed{,440}$
2	,889[b]	$\boxed{,791}$

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

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

Correlazioni parziali

Modello 3

- 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		stand d coeff			Correlations		
		B	Std. Error	Beta	t	Sig.	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

correl fra x e y

correl fra x e z

correl fra y e z

Correlazione parziale fra x e y

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

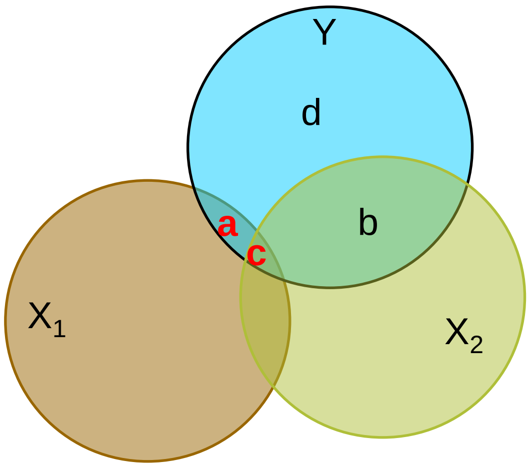
Quadrato della correl fra x e z

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

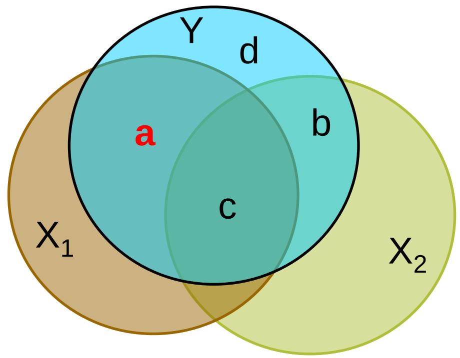


$$\frac{b}{b+d} \approx \frac{b+c}{a+b+c+d}$$

Il motivo più frequente:
X_1 è scarsamente correlato con Y

Parziale può essere **maggiore** di semplice

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

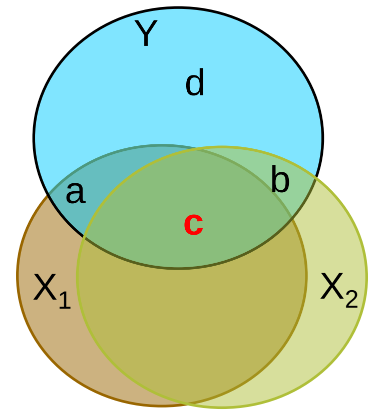


$$\frac{b}{b+d} > \frac{b+c}{a+b+c+d}$$

Il motivo più frequente :
X_1 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 :
X_1 è altamente correlato con X_2