1. Analisi della regressione multipla: un (relativam) rapido ed essenziale ripasso (prerequisiti per comprenedre i contenuti di questo insegnamento)

Elementi di metodologia
Disegni correlazionali

- ricerche che descrivono una relazione o una rete di relazioni tra variabili
- una relazione viene descritta in termini di

  intensità o *effect size*

  direzione
- si avvalgono di un'ampia serie di tecniche d'analisi statistica che rispondono a vari scopi
- non permettono, in generale, inferenze di natura causale

Disegni correlazionali

- ma permettono di attuare una serie di controlli sulla relazione tra variabili (moderazione, mediazione,…)
  - *scomponendo la relazione* in componente diretta o componente indiretta (mediazione)
  - *stratificando la relazione* attraverso i livelli di una terza variabile (moderazione)
- e di indagare le relazioni nel tempo a lungo termine (*studi longitudinali*) o a breve termine (*studi sulle fluttuazioni o dinamiche intrapersonali)*

Tipi di variabili e di relazioni tra 2 variabili

- co-occorrenza: indipendenza semantica e simmetria (es. peso e altezza)
- dipendenza: indipendenza semantica e asimmetria come priorità, il disegno non è sperimentale, ma quasi- o pre-sperimentale (cfr disegni longitudinali)
- causalità: indipendenza semantica e asimmetria, piena separabilità tra explanans ed explanandum, disegno sperimentale (randomizzazione/manipolazione)

Quale il ruolo di ogni variabile?

In generale,

la validità di una ricerca dipende dalla sostenibilità delle relazioni ipotizzate.

Si possono allora trovare spiegazioni diverse da quelle ipotizzate (cioè VI) per spiegare la variabilità di VD?

Altre variabili potrebbero intervenire

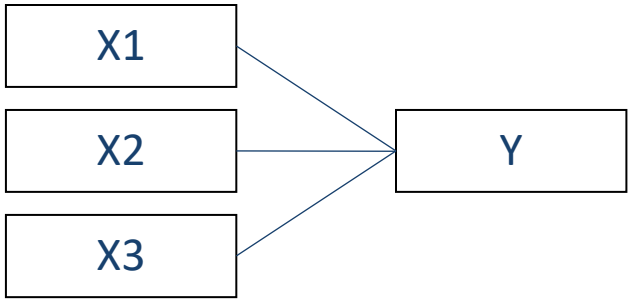
nella relazione tra VD e VI?

Si inserisce la terza variabile per controllare

quei disegni di ricerca non sperimentali o sperimentali ma non bilanciati

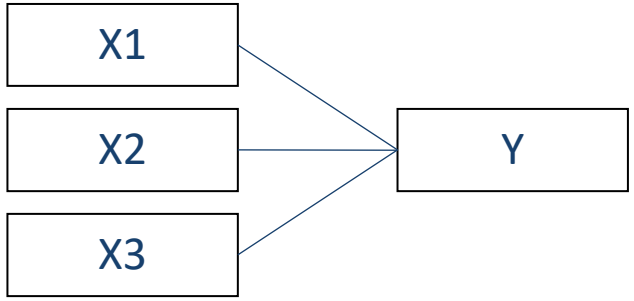
I disegni correlazionali: ruoli e relazioni tra 3 (o più) variabili

focus sulla variabile



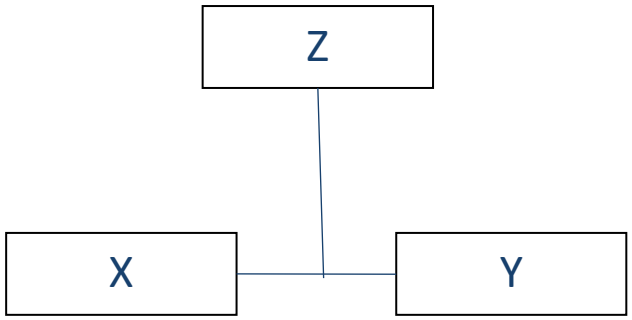
Spiegare una variabile:
Peso o impatto unico di ogni stimatore
(l'impatto unico del QI sulla prestazione accademica, al di là dell'ansia e della motivazione)

Prevedere una varabile:
la combinazione più efficiente
(QI, ansia e motivazione permettono un'ampia previsione della prestazione accademica)



I disegni correlazionali: ruoli e relazioni tra 3 (o più) variabili

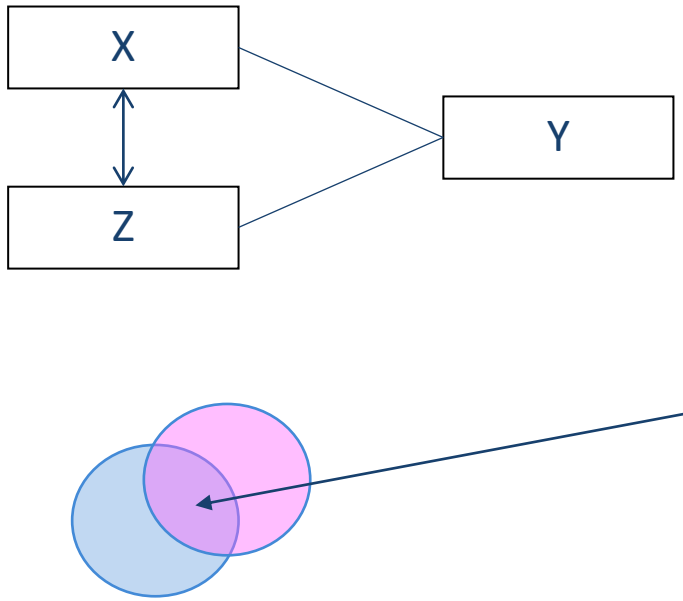
focus sulla relazione



Stratificare una relazione:
Come varia la relazione tra X e Y al variare dei livelli di Z?
Se la relazione tra XY varia dipendentem dai livelli di Z, allora la relazione tra XY è **MODERATA** da Z, X e Z interagiscono
(la prestazione accademica varia in funzione del QI soprattutto nelle persone introverse ovvero QI e introversione interagiscono nella previsione della prestazione accademica)

I disegni correlazionali: ruoli e relazioni tra variabili

focus sulla relazione

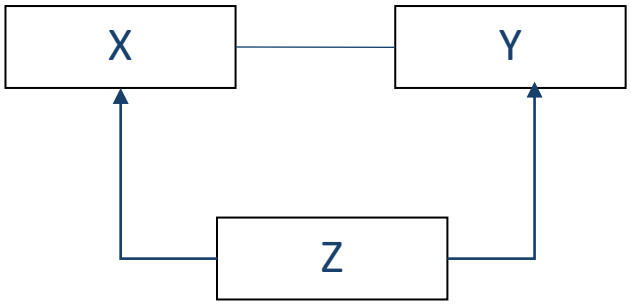


Depurare una relazione per controllare la relazione tra due variabili indipendenti per comprendere la relazione unica tra variabile dipendente e variabile indipendente

(Immagine corporea e stato affettivo correlano e l'impatto della immagine corporea su Y diviene nullo / si riduce se depurata dalla sua associazione con lo stato affettivo)

I disegni correlazionali: ruoli e relazioni tra variabili

focus sulla relazione



Depurare una relazione: se Z precede logicamente sia X sia Y
e la direzione delle relazione YZ e di quella XZ è la stessa e di simile intensità, si ha una correlazione XY **SPURIA**, vale a dire la componente diretta che lega X a Y è 0, mentre la correlazione semplice osservata è > 0

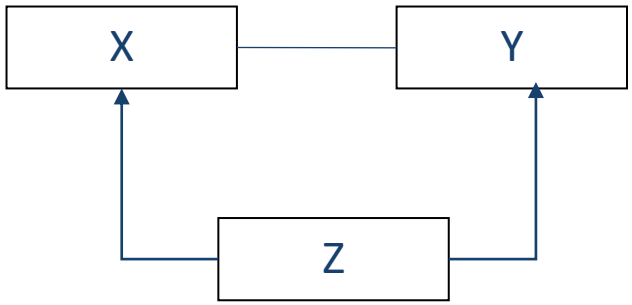
(all'aumentare dell'altezza aumentano le competenze matematiche: questa relazione svanisce se inserisco l'età e depuro l'altezza dalla sua relazione con l'età)

I disegni correlazionali: ruoli e relazioni tra variabili

focus sulla relazione

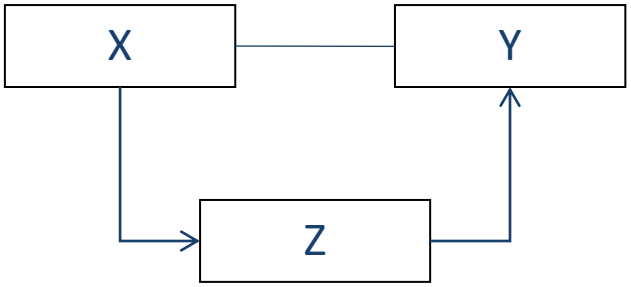
Depurare una relazione: se Z precede logicamente sia X sia Y e la direzione delle relazione YZ e di quella XZ è opposta e di simile intensità, si ha una correlazione XY **SOPPRESSA**, vale a dire la component diretta che lega X a Y è > 0, mentre la correlazione semplice osservata è 0

(la relazione tra NA e PA è nulla, ma se introduciamo e controlliamo los tile acquiescente di risposta diventa negativa)



I disegni correlazionali: ruoli e relazioni tra variabili

focus sulla relazione



Depurare (per spiegare) una relazione:
In che modo X ha un peso, un impatto su Y? Attraverso una terza variabile Z? Se X precede logicamente Z e Z precede Y, allora la relazione XY è **MEDIATA** da Z
(la perdita di un lavoro favorisce uno stato depressivo che a sua volta può favorire uno stile genitoriale inconsistente)

Regressione multipla:
controllare le relazioni tra VI per depurare la relazione tra una VI e VD

equazione di previsione di Y con

- 1 VI

$\hat{Y}_i = a + bX_i$ $\quad$ $r_{xy} = \beta_{xy}$

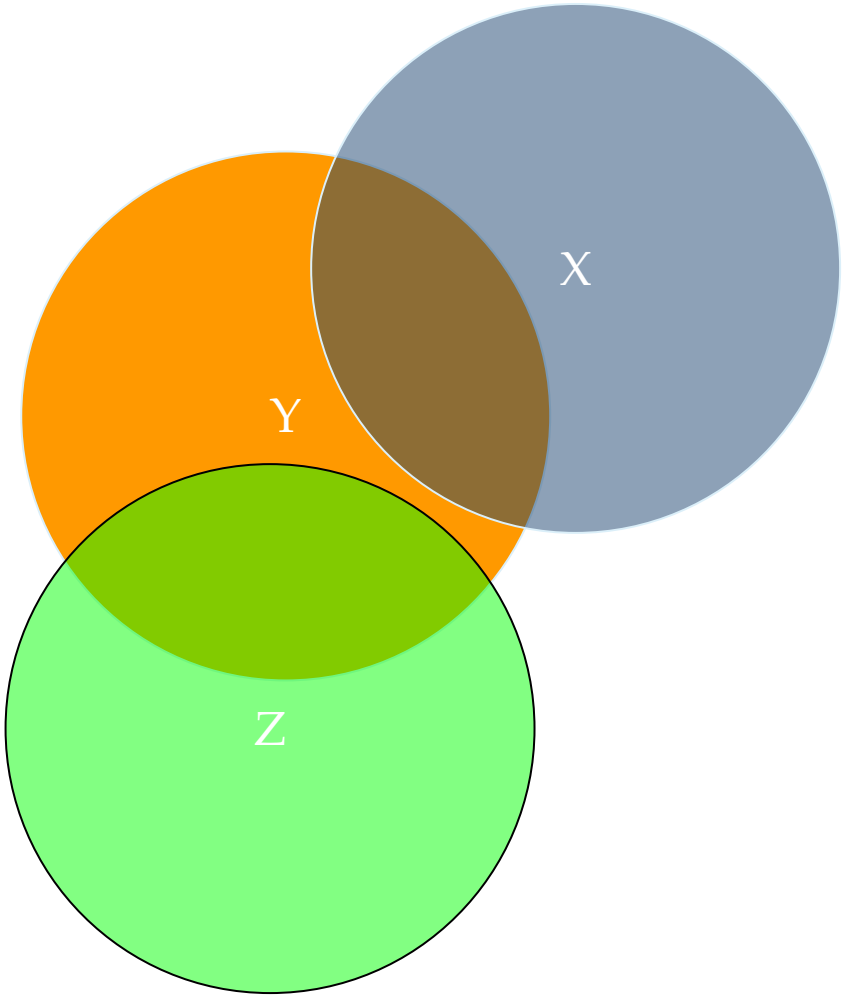
- 2 VI

$\hat{Y}_i = a + b_{YX1 \bullet X2} X_{1i} + b_{YX2 \bullet X1} X_{2i}$

Che cosa cambia quando da 1 si passa a 2 o più stimatori?

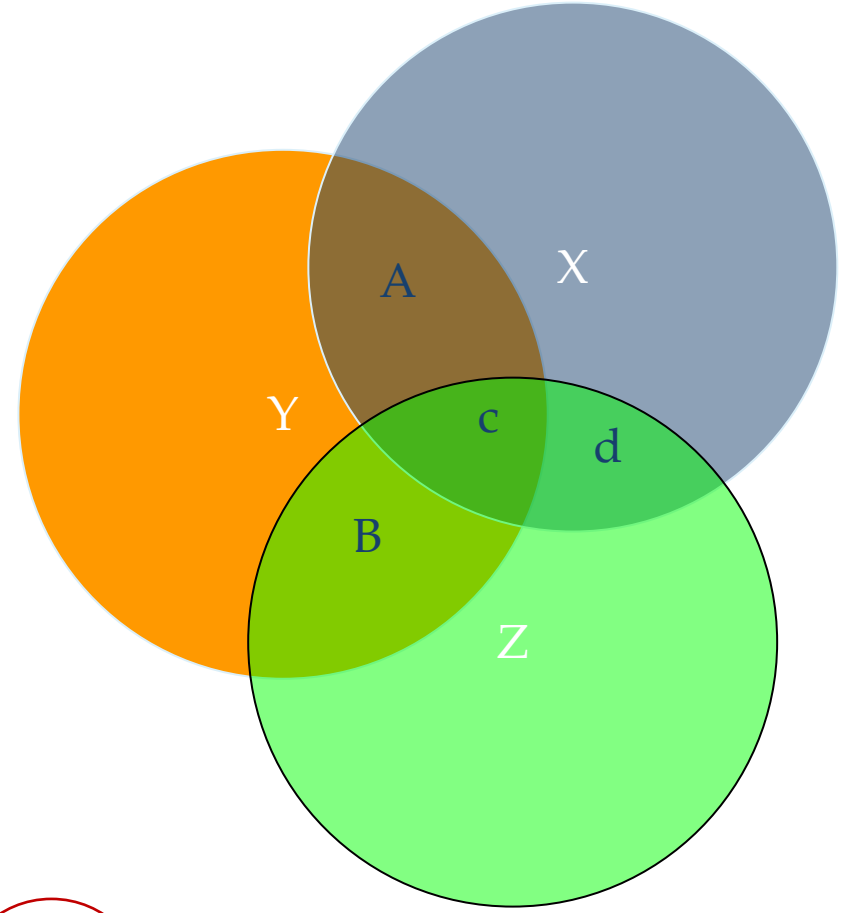
LE RELAZIONI TRA LE VARIABILI INDIPENDENTI VENGONO TENUTE SOTTO CONTROLLO: PARZIALIZZAZIONE

Se le due VI non sono correlate tra loro



$$R_Y^2 = r_{YX}^2 + r_{YZ}^2$$

X ripulita da c + d corrisponde ai residui di X regredito su Z (e viceversa per la varabile Z)



Se le due VI sono correlate tra loro

$R_Y^2 = r_{YX}^2 + r_{YZ}^2$

Aree A e B rappresentano quota associazione unica tra Y –X e Y-Z, rispettivamente; sono quantificate da coeff di associazione parziale che lega i residui di X (o di Z) a Y

Area C: quota di varianza di Y che spiegano sia X sia Z in virtù della loro associazione

Area B+C rappresenta la quota di variabilità condivisa da Y e Z e quantificata dal coeff di correlazione semplice tra le 2 var

Area A+C rappresenta la quota di variabilità condivisa da Y e X e quantificata dal coeff di correlazione semplice tra le 2 var

Regressione multipla: parzializzare, controllare, depurare

Attraverso la parzializzazione, l'impatto di X su Y ossia la relazione osservata tra Y e X viene scomposta in relazione diretta e relazione indiretta e viene depurata dalla sua componente indiretta,

dovuta alla relazione tra gli stimatori

$$r_{YX} = d_{YX} + i_{YX}$$

$$i_{YX} = r_{YZ}r_{XZ}$$

$$d_{YX} = pr_{def}r_{YX \cdot Z} = r_{YX} - r_{XZ}r_{YZ}$$

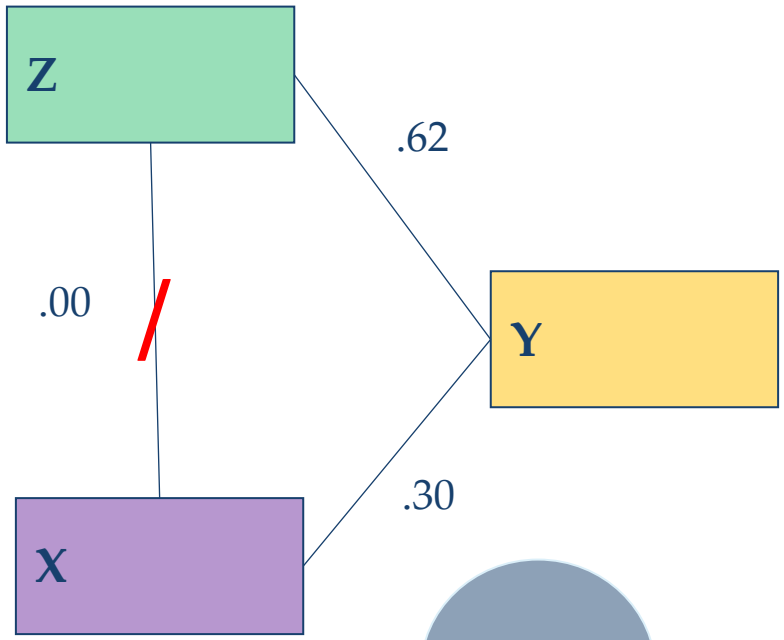
$$= r_{YX} - i_{YX}$$

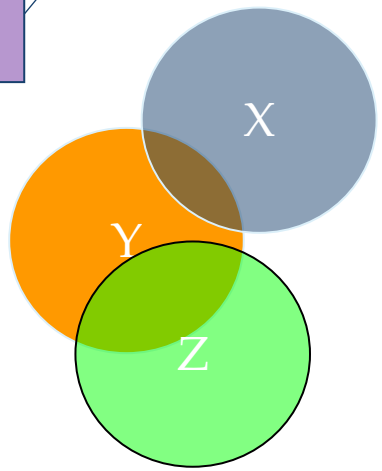
(coefficiente di correlazione parziale deflazionato)

Regressione multipla: per un ripasso essenziale

$$r_{YX} = d_{YX} + i_{YX}$$
$$i_{YX} = r_{YZ}r_{XZ}$$
$$d_{YX} = pr_{def}r_{YX\cdot Z} = r_{YX} - r_{XZ}r_{YZ} = r_{YX} - i_{YX}$$

$$i_{YX} = .00 \times .62 = .00$$
$$d_{YX} = .30 - .00 = .30$$
$$r_{YX} = .30 + .00 = .30$$

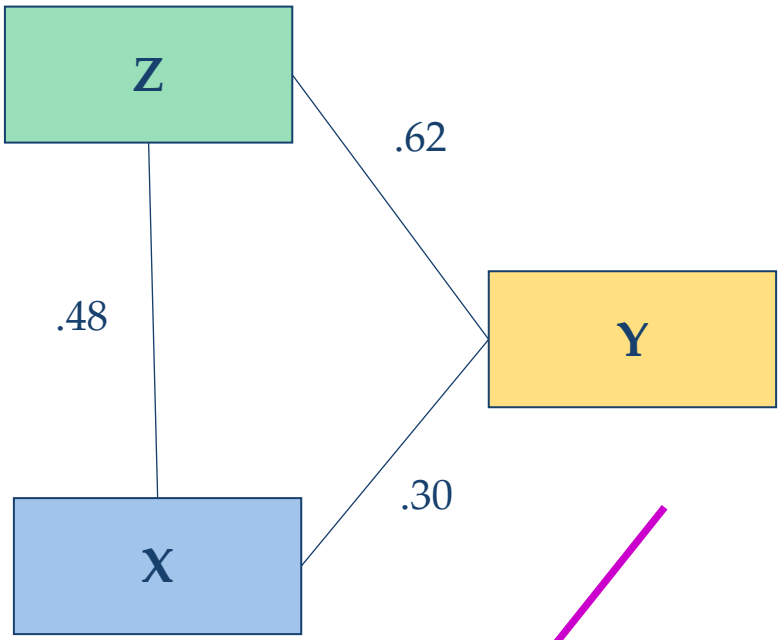




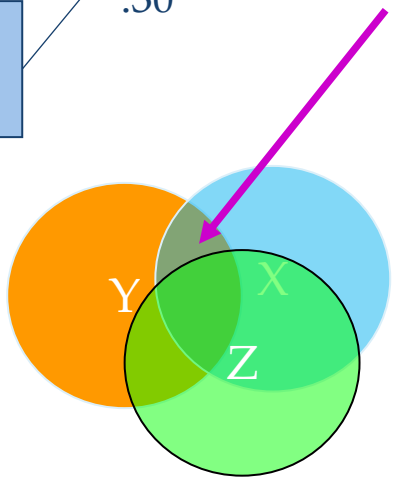
Regressione multipla: per un ripasso essenziale

$$r_{YX} = d_{YX} + i_{YX}$$
$$i_{YX} = r_{YZ}r_{XZ}$$
$$d_{YX} = p_{def}r_{YX \cdot Z} = r_{YX} - r_{XZ}r_{YZ}$$
$$= r_{YX} - i_{YX}$$

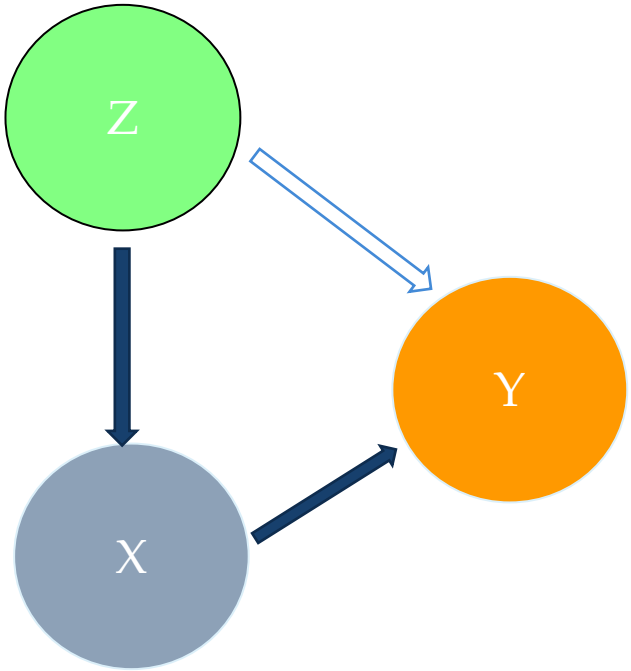
$i_{YX} = .48 \times .62 = .30$
$d_{YX} = .30 - .30 = .00$
$r_{YX} = .30 + .00 = .30$
$p_{def}r_{YX \cdot Z} = .30 - (.48 * .62) = .00$

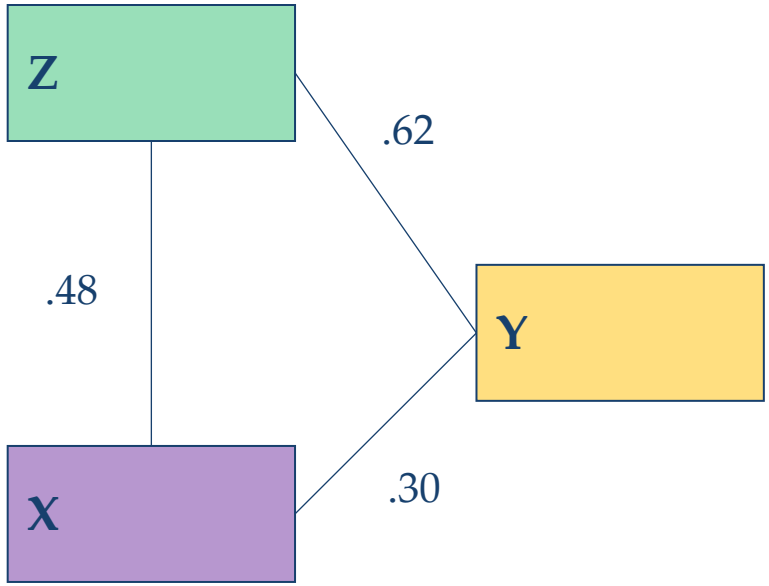


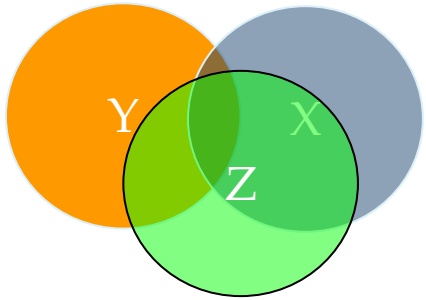
$i_{YZ} = .48 \times .30 = .14$
$d_{YZ} = .62 - .14 = .48$
$r_{YZ} = .48 + .14 = .62$
$p_{def}r_{YZ \cdot X} = .62 - (.48 * .30) = .48$



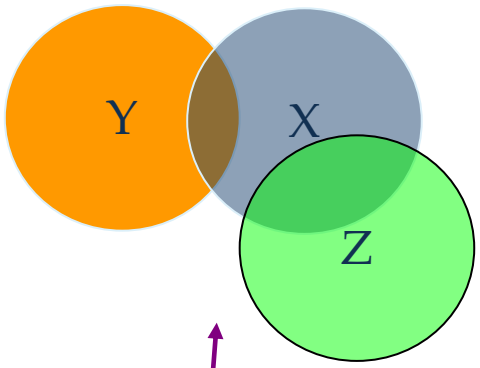
Regressione multipla: per un ripasso essenziale

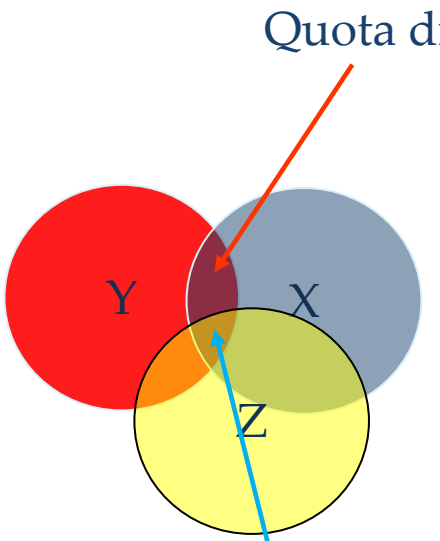






Regressione multipla: per un ripasso essenziale





Z variando porta ad una variazione in X che a sua volta comporta una variazione in Y: Quanto del peso di X su Y dipende dalla variazione di Z su X? La quota del peso di X su Y che dipende dal legame tra X e Z è proporzionale al legame tra Z e Y.
Se Y e Z non sono correlati, allora Z pesa su X ma non per quella parte che di X si lega a Y.
Altrimenti, il peso è proporzionale ($r_{XZ}r_{YZ}$) e corrisponde alla quota indiretta del legame tra X e Y dovuta al legame tra X e Z

Regressione multipla: per un ripasso essenziale

equazione di previsione di Y con 2 VI:

$$\hat{Y}_i = a + b_{YX1 \bullet X2} X_{1i} + b_{YX2 \bullet X1} X_{2i}$$

$$Y_i = a + b_{YX1 \bullet X2} X_{1i} + b_{YX2 \bullet X1} X_{2i} + e_i$$

coefficienti di regressione parziale, b e $\hat{\beta}$: rappresentano il peso o impatto unico di ciascuna VI nell'equazione di previsione di Y, depurata dalla relazione che intercorre tra gli stimatori

$$b_{YX1 \bullet X2} = \hat{\beta}_{YX1 \bullet X2} \frac{s_Y}{s_{X1}}$$

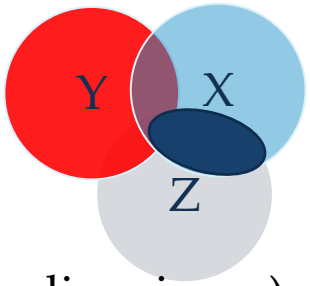
$$\hat{\beta}_{YX1 \bullet X2} = \frac{r_{YX1} - r_{YX2} r_{X1X2}}{1 - r_{X1X2}^2}$$

Regressione multipla: per un ripasso essenziale

Coefficiente di correlazione semi-parziale:

sr rappresenta la quota di variabilità (sr^2 = quota di varianza) che ogni VI, parzializzata dalle altre VI, spiega della varianza totale di Y

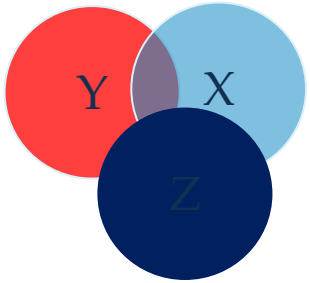
$$r_{YX1 \bullet X2(s)} = \frac{r_{YX1} - r_{YX2} r_{X1X2}}{\sqrt{1 - r^2{}_{X1X2}}}$$



Coefficiente di correlazione parziale:

pr rappresenta la quota di variabilità (pr^2 = quota di varianza) che ogni VI, parzializzata dalle altre VI, spiega della varianza di Y, parzializzata dalle altre VI

$$pr_{YX1 \bullet X2} = \frac{r_{YX1} - r_{YX2} r_{X1X2}}{\sqrt{1 - r^2{}_{YX2}}\sqrt{1 - r^2{}_{X1X2}}}$$



Regressione multipla: per un ripasso essenziale

$$b_{YX1 \bullet X2} = \hat{\beta}_{YX1 \bullet X2} \frac{s_Y}{s_{X1}}$$

$$\hat{\beta}_{YX1 \bullet X2} = \frac{r_{YX1} - r_{YX2} r_{X1X2}}{1 - r^2_{X1X2}}$$

$$sr_{YX1 \bullet X2} = \frac{r_{YX1} - r_{YX2} r_{X1X2}}{\sqrt{1 - r^2{}_{X1X2}}}$$

$$pr_{YX1 \bullet X2} = \frac{r_{YX1} - r_{YX2} r_{X1X2}}{\sqrt{1 - r^2{}_{YX2}}\sqrt{1 - r^2{}_{X1X2}}}$$

4 indicatori quantitativi dell'intensità e direzione della relazione che VD e VI condividono in modo unico. Ciascuno si presta a essere usato per scopi differenti, seppure forniscono la medesima informazione sostanziale.

Un legame unico può essere anche Pensato come l'associazione tra X1 e Y, avendo depurato X della sua relazione con X2, salvati i punteggi residui
E questi ultimi poi messi in relazione con (o Y parzializzati da Z per pr)

Per raccontare le associazioni parziali

usiamo i file dati

- Jamovi «A»
- Excel «A»

- E lavoriamo sui punteggi residui di X (depurato da Z o X2) e
- sui residui di Y (depurato da Z o X2)

Correlation Matrix

		InsoddisfazioneCorpo	UmoreNegativo	CiboFuoriPasto
InsoddisfazioneCorpo	Pearson's r	—		
	p-value	—		
UmoreNegativo	Pearson's r	0.489	—	
	p-value	< .001	—	
CiboFuoriPasto	Pearson's r	0.302	0.331	—
	p-value	0.005	0.002	—

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.368	0.135	0.114	6.42	2	82	0.003

Equazione di previsione

Model Coefficients - CiboFuoriPasto

Predictor	Estimate	SE	t	p	Stand. Estimate	pr	sr
Intercept	1.237	0.2018	6.13	< .001			
InsoddisfazioneCorpo	0.141	0.0905	1.56	0.122	0.184	.170	.160
UmoreNegativo	0.205	0.0999	2.05	0.044	0.241	.221	.211

Beta parziali std

Dal file in excel

- Calcolo punteggi stimati o attesi di Y in base all'eq di previsione (evidenziata in rosso in una slide precedente)
- Calcolo dell'eq di previsione di X2 a partire da X1 (X2-UN regredita su X1-IC) e relativi residui

Model Coefficients - UmoreNegativo

Predictor	Estimate	SE	t	p
Intercept	0.775	0.2048	3.78	< .001
InsoddisfazioneCorpo	0.443	0.0868	5.11	< .001

- Calcolo dell'eq di previsione di Y a partire da X1 (Y-CFP regredito su X1-IC) e relativi residui

Model Coefficients - CiboFuoriPasto

Predictor	Estimate	SE	t	p
Intercept	1.396	0.1899	7.35	< .001
InsoddisfazioneCorpo	0.232	0.0805	2.89	0.005

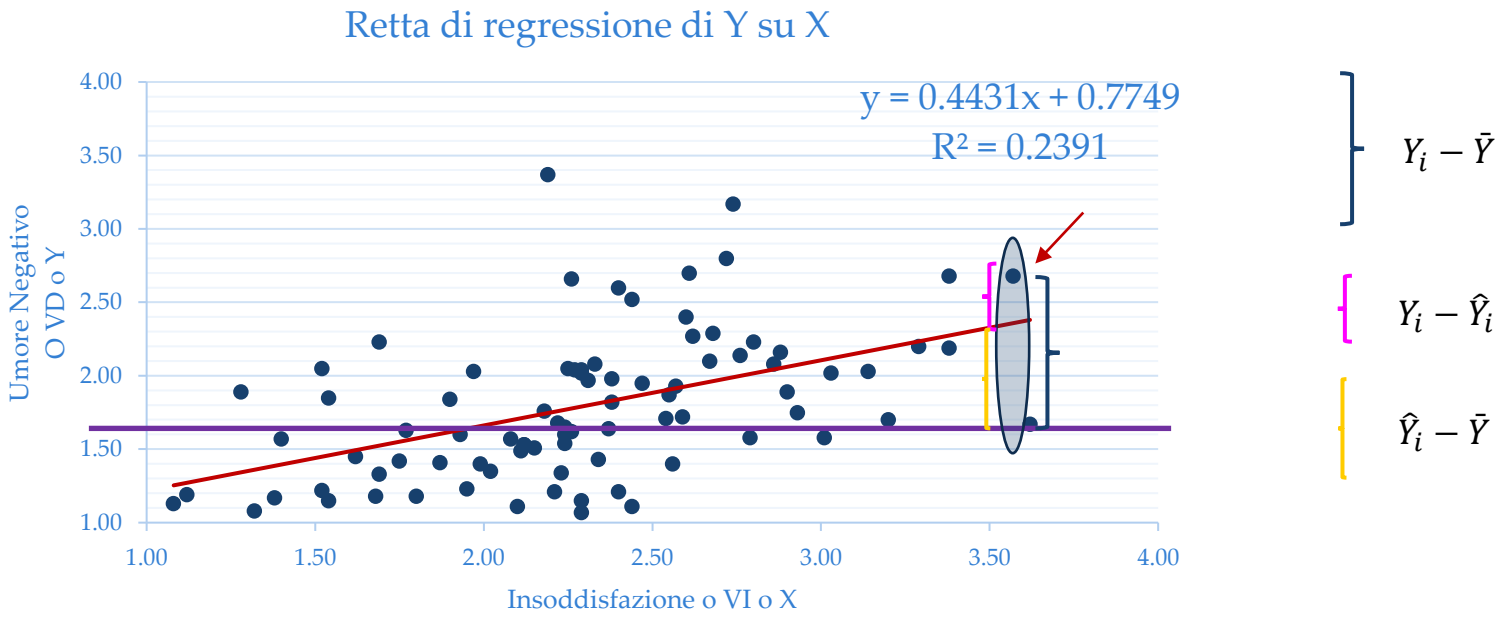
•Calcolo dei residui di Y o errore di stima in base alla regressione di Y su X1 2 X2

Dal file in excel «A»

Si può verificare come

- residui di Y regredita su X1 e X2 non siano correlati né a X1 né a X2
- La correlazione tra residui di X2, parzializzato dunque da X1, corrisponde a sr = 0.211 riportata nella tabella output (cerchiati in verde)
- La correlazione tra i residui di di X2, parzializzato da X1, e di Y, parzializzato da X1, corrisponde a pr = 0.221 riportata nella tabella output (cerchiati in lilla)

ARM: IL coefficiente di determinazione (MULTIPLO)



$$r^2_{YX} = \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = r^2_{Y\hat{Y}}$$

ARM: Il coefficiente di determinazione multiplo

Il coefficiente di correlazione elevato al quadrato R^2 rappresenta un indice quantitativo di RPE o riduzione proporzionale dell'errore di stima

$$RPE = \frac{e_{SENZA\ REGOLA\ DI\ DECISONE} - e_{CON\ REGOLA\ DI\ DECISIONE}}{e_{SENZA\ REGOLA\ DI\ DECISONE}}$$

Eta quadro = DEV tra / DEV tot

$$RPE = \frac{\sum(Y-\bar{Y})^2 - \sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2} = \frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2} = \frac{\sum(\hat{Y}-\bar{Y})^2/N-1}{\sum(Y-\bar{Y})^2/N-1} = \frac{s^2_{\hat{Y}}}{s^2_Y} = r^2_{Y'Y}$$

R^2 come rapporto tra varianza spiegata dalla regressione (dall'insieme di stimatori) e varianza totale di Y, in altre parole corrisponde alla correlazione al quadrato tra
punteggi Y previsti e punteggi Y osservati

ARM: IL coefficiente di determinazione MULTIPLO

Test F per verificare la significatività di R^2(Ipotesi nulla $R^2 = 0$)

$$F = \frac{dev.regress/k}{dev.residua/N-k-1}$$

multicollinearità tra VI

(P value for an F test: Calculator)

Coefficiente di determinazione multiplo secondo un modello di scomposizione gerarchico

$$R^2_{Y\bullet X1X2X3} = r^2_{YX1} + r^2_{YX2\bullet X1} + r^2_{YX3\bullet X1X2}$$

ARM: sulla significatività dei coefficienti di regressione parziale non std

Rapporto tra variabilità come DS

Multicollinearità:
coeff alienazione di *Xi*

Varianza residua (coefficiente di alienazione di Y)

$$SE_{b_{Xi}} = \frac{s_Y}{s_{X_i}} \sqrt{\frac{1}{1 - R^2_{X_i}}} \sqrt{\frac{1 - R^2_Y}{N - k - 1}}$$

H_0: $b = 0$ $(=\beta)$
attraverso t test (gl = N-k-1)

$$t_{Y_{Xi}} = sr_{YXi} \sqrt{\frac{N - k - 1}{1 - R^2_Y}}$$

$$SE_{\beta_{Xi}} = \sqrt{\frac{1}{1 - R^2_{X_i}}} \sqrt{\frac{1 - R^2_Y}{N - k - 1}}$$

Per tutte queste ragioni così illustrate
il numero di stimatori e la numerosità del campione vanno pensati e scelti accuratam
Insieme al livello p critico, così come gli stimatori, per spiegare quanta più VAR possi
col modello più parsimonioso possibile, contenendo la multicollinearità,
per poter esaminare statisticamente un modello valido con adeguata potenza del test

ARM: sulla significatività dei coefficienti di regressione parziale non std Excel (foglio 2)

t test di b_{YX2}	2,054383	se sr =.21 gl = 82 p < 0,05
t test di b_{YX2}	3,184253	Se sr =.21 N = 200 e gl = 197 p < 0,01
t test di b_{YX2}	2,920923	se sr = 0,3 e gl = 82 p < 0.01

SE di b_{YX2}	0,099901	
SE di b_{YX2}	0,087581	se diminuisce la correlazione tra X1 e X2
SE di b_{YX2}	0,064453	se N = 200
SE di b_{YX2}	0,157779	se aumenta SD di Y in rapporto a X2
SE di b_{YX2}	0,070647	se aumenta SD di X2 in rapporto a Y

Analisi della regressione multipla: Strategie analitiche

- Regressione simultanea o standard (enter)
  - tutte le VI sono inserite contemporaneam
  - per ogni VI si tiene sotto controllo la relazione con tutte le altre VI
- Regressione gerarchica
  - 1 o più VI vengono inserite secondo una successione predefinita, in base a obiettivi specifici
- Regressione statistica
  - *Forward*: 1 VI alla volta, incominciando da quella con corr semplice più alta con VD; poi di volta in volta VI con part corr maggiore con VD; una volta immessa una VI non viene più tolta
  - *Backward*: tutte le VI inserite simultaneam e poi tolte una alla volta, ogni volta quella che spiega minore quota di varianza di VD non significativa;
  - *Stepwise*: procede come forward, ma di volta in volta viene valutata ogni VI inserita nel modello e può essere tolta come in backward

un esempio con regressione gerarchica (model Builder in jamovi)

Model Fit Measures

			Overall Model Test			
Model	**R**	**R^2**	**F**	**df1**	**df2**	**p**
1	0.495	0.245	17.0	2	105	< .001
2	0.503	0.253	11.7	3	104	< .001

Model Comparisons

Comparison							
Model		**Model**	**ΔR^2**	**F**	**df1**	**df2**	**p**
1	-	2	0.00796	1.11	1	104	0.295

Omnibus ANOVA Test

	Sum of Squares	**df**	**Mean Square**	**F**	**p**
RSE	19.2	1	19.2	0.370	0.544
PHD_DEPRESSION	1447.2	1	1447.2	27.806	< .001
Residuals	5464.9	105	52.0		

Note. Type 3 sum of squares

(continua)

Model Coefficients - UCLA

			95% Confidence Interval				
Predictor	**Estimate**	**SE**	**Lower**	**Upper**	**t**	**p**	**Stand. Estimate**
Intercept	15.339	5.957	3.527	27.150	2.575	0.011	
RSE	0.125	0.205	-0.282	0.532	0.608	0.544	0.0604
PHD_DEPRESSION	0.981	0.186	0.612	1.350	5.273	< .001	0.5239

Model Coefficients - UCLA

			95% Confidence Interval				
Predictor	**Estimate**	**SE**	**Lower**	**Upper**	**t**	**p**	**Stand. Estimate**
Intercept	13.030	6.345	0.447	25.612	2.054	0.043	
RSE	0.123	0.205	-0.284	0.530	0.599	0.551	0.0594
PHD_DEPRESSION	0.869	0.214	0.445	1.294	4.064	< .001	0.4644
STRESS	0.379	0.360	-0.334	1.092	1.053	0.295	0.1070

Analisi della regressione: Alcune assunzioni

- VD quantitative, almeno scala a intervalli equivalenti
- VI quantitativa oppure qualitativa (es., dummy coding)
- ridotta multicollinearità
- ridotto errore di misurazione (affidabilità)
- controllo sui casi outlier univariati (es. z score estremi) e multivariati (test di Mahalanobis)

Analisi della regressione: Alcune assunzioni

- assenza di errore di specificazione
  - inclusione VI irrilaventi e omissione VI rilevanti
  - non linearità della relazione tra VI e VD
    → rimedio: si rende la relazione lineare, es., $Y_i' = a + b_1 X_1 + b_2 X_1^2$
  - non additività della relazione tra VD e VI (i.e., interazione tra VI)
    → si rende la relazione additive, i.e., $Y_i' = a + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2$
- controllo sui residui (es., assunzione dell'indipendenza e MLM)