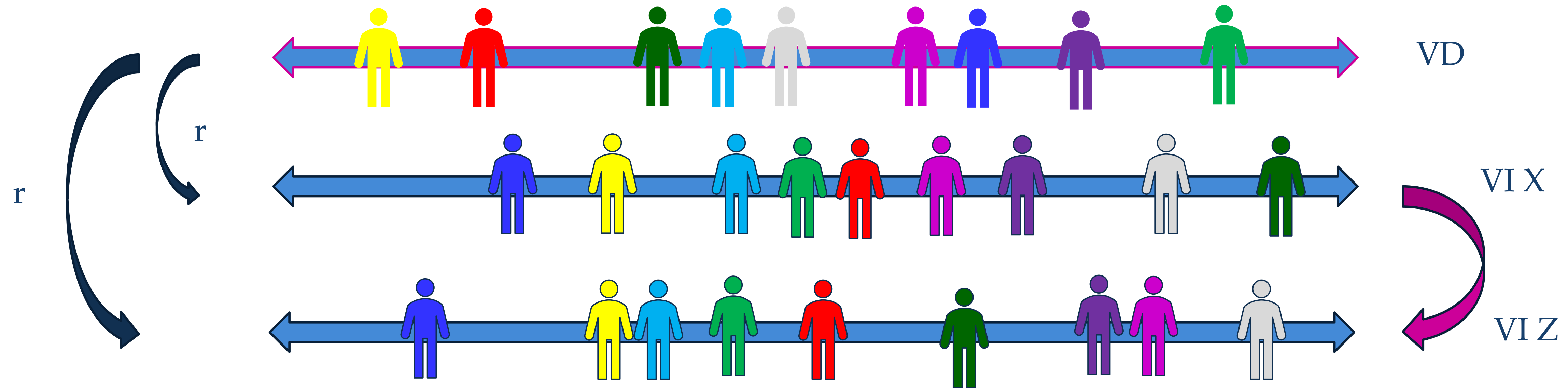
Analisi della regressione multipla: la parzializzazione

Depurare per controllare

Possiamo controllare l'impatto di una «terza» variabile tenendola costante attraverso i livelli dell'altra variabile fin dalla progettazione del disegno di ricerca bilanciato oppure possiamo controllare la «terza» variabile statisticamente

Ma perché controllare e depurare una relazione tra VD e VI dall'interferenza di una «terza» variabile?

Between People: 3 o più variabili quantitative



Se X e Z NON sono correlati, allora ogni VI spiega una parte di VD unica
Se X e Z sono correlati, allora X spiega di VD parte di varianza che anche Z può spiegare e vice versa

Descriviamo una relazione unica tra due variabili (intensità, direzione e forma) e prevediamo l'andamento di una variabile VD a partire dalle altre VI (coefficienti di associazione «parziale»)

Analisi della Regressione multipla LINEARE

equazione di previsione di Y con

- 1 VI

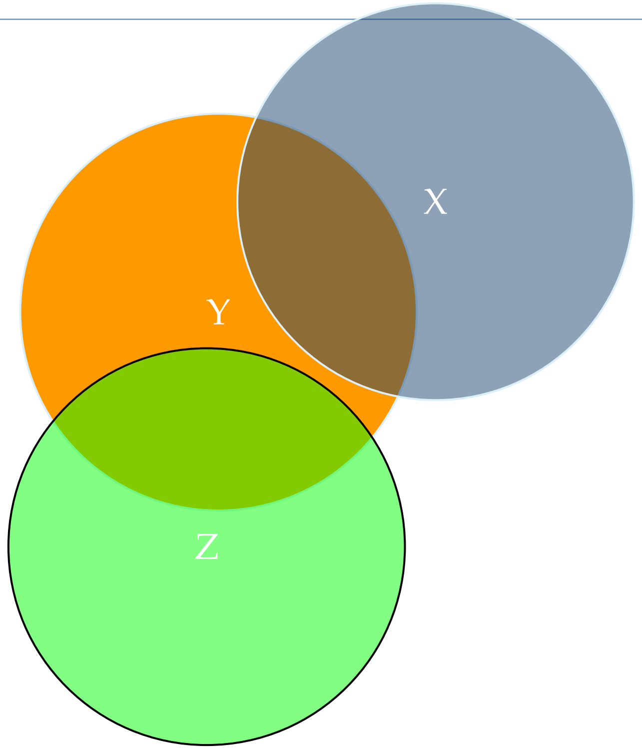
$$\hat{Y}_i = a + bX_i$$

- 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

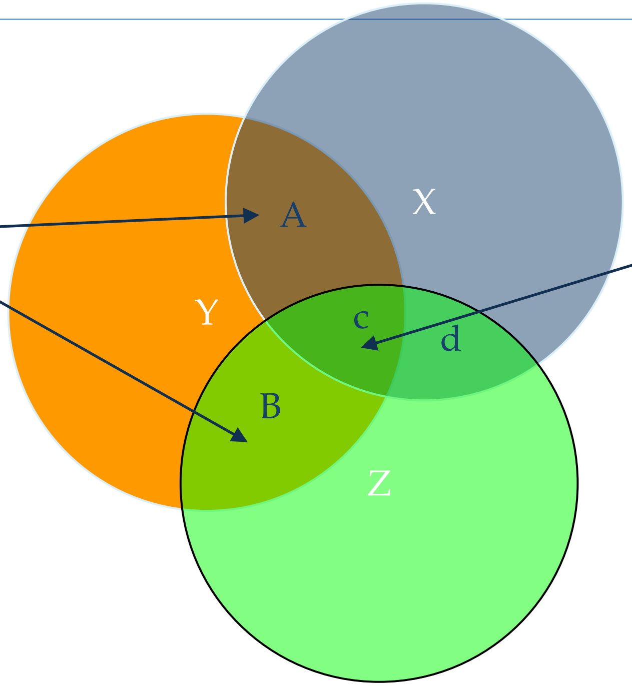


$$R_Y^2 = r_{YX}^2 + r_{YZ}^2$$

Cosa accade se le VI sono correlate tra loro?

Aree A e B rappresentano quota associazione unica tra Y e X e tra Y eZ, rispettivamente;

sono quantificate da coeff di associazione parziale



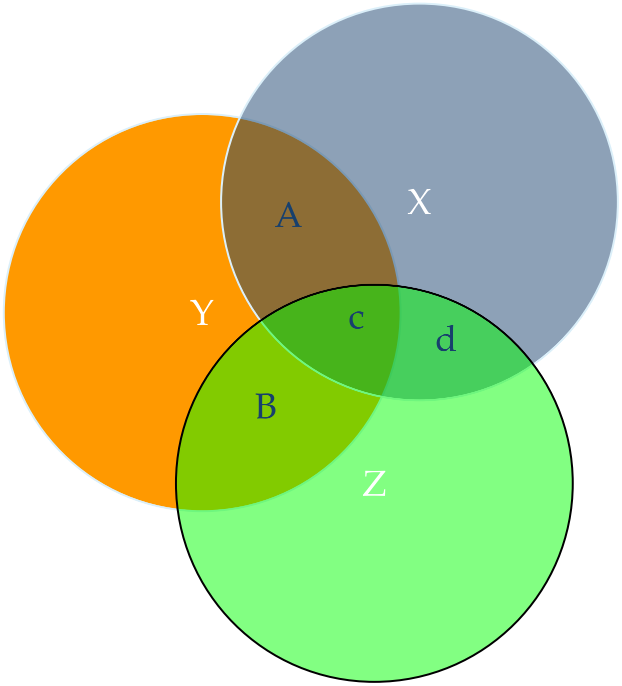
$$R_Y^2 = r_{YX}^2 + r_{YZ}^2$$

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

Serviamoci dei «punteggi residui» per definire la relazione «unica» tra ogni VI e la VD



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

Per «depurare» X da Z, possiamo regredire X su Z

$$X' = a + bZ$$

Se applichiamo l'eq. possiamo calcolare X' o atteso sulla base dei valori di Z

Se poi calcoliamo lo scarto X – X' otteniamo il residuo di X, quella parte del punteggio che NON si associa a Z ossia che Z non riesce a prevedere.

Ecco, ora possiamo legare quei RESIDUI di X a Y e calcolare quanto X, depurato da Z, si associa a Y

Analisi della Regressione multipla LINEARE

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

$$b_{YX1\bullet X2} = \hat{\beta}_{YX1\bullet X2}\frac{s_Y}{s_{X1}} \qquad \hat{\beta}_{YX1\bullet X2} = \frac{r_{YX1} - r_{YX2}r_{X1X2}}{1 - r_{X1X2}^2}$$

Analisi della Regressione multipla LINEARE

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

Numeratore: a r tra Y e X1 sottraggo la quota che X1 condivide con X2, pesandola però per quella parte che X2 condivide anche con Y (se r tra Y e X2 = 0, allora al numeratore rimane la r semplice tra Y e X1), Così DEPURO r tra Y e X1 della quota «indiretta» dovuta al legame che X2 ha sia con X1 che con Y, lasciando solo legame «diretto» tra Y e X1

Denominatore: rapporto quel legame diretto tra Y e X1 alla varianza di X1 depurata da X2

Analisi della Regressione multipla LINEARE

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

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

Analisi della Regressione multipla LINEARE

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

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

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

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

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 Y
(o Y parzializzati da Z per pr)

Usiamo i nostri dati in Jamovi ed excel per capire meglio

VD = RSE

VI = HUMOR self-enhancing (X2) + HUMOR self-defeating (X1)

1. usiamo Jamovi per osservare le correlazioni semplici

Matrice di Correlazione

		Y_RSE	**X1_humor_self-def**	**x2_humor_self_enhanc**
Y_RSE	r di Pearson	—		
	gdl	—		
	valore p	—		
X1_humor_self-def	r di Pearson	-0.329	—	
	gdl	182	—	
	valore p	< .001	—	
x2_humor_self_enhanc	r di Pearson	0.204	0.221	—
	gdl	182	182	—
	valore p	0.006	0.003	—

Usiamo i nostri dati in Jamovi ed excel per capire meglio

VD = RSE

VI = HUMOR self-enhancing (X2) + HUMOR self-defeating (X1)

Proviamo a calcolare i coefficienti di associazione parziale sulla base delle correlazioni semplici

$$\hat{\beta}_{YX1\bullet X2} = \frac{r_{YX1} - r_{YX2}r_{X1X2}}{1 - r^2_{X1X2}} = \frac{-0.329 - (0.204 * 0.221)}{1 - (0.221^2)} = \frac{-0.374}{0.951} = -0.393$$

$$sr_{YX1\bullet X2} = \frac{r_{YX1} - r_{YX2}r_{X1X2}}{\sqrt{1 - r^2{}_{X1X2}}} = \frac{-0.329 - (0.204 * 0.221)}{\sqrt{1 - (0.221^2)}} = \frac{-0.374}{\sqrt{0.951}} = -0.383$$

$$pr_{YX1\bullet X2} = \frac{r_{YX1} - r_{YX2}r_{X1X2}}{\sqrt{1 - r^2{}_{YX2}}\sqrt{1 - r^2{}_{X1X2}}} \frac{-0.329 - (0.204 * 0.221)}{\sqrt{1 - (0.204^2)}\sqrt{1 - (0.221^2)}} = \frac{-0.374}{\sqrt{0.958}\sqrt{0.951}} = -0.392$$

Usiamo i nostri dati in Jamovi ed excel per capire meglio

VD = RSE

VI = HUMOR self-enhancing (X2) + HUMOR self-defeating (X1)

2. usiamo Jamovi per regredire Y su X1 e X2 e calcolare i punteggi previsti di Y e i suoi residui

Ogni coefficiente b e beta di regressione informa sul legame che uno stimatore ha con Y osservato, depurato lo stimatore dall'altro stimatore

Coefficienti del Modello - Y_RSE

						95% Intervallo di Fiducia	
Predittore	**Stima**	**SE**	**t**	**p**	**Stima Standard**	**Inferiore**	**Superiore**
Intercettare	17.901	1.8735	9.56	< .001			
x2_humor_self_enhanc	0.209	0.0495	4.23	< .001	0.290	0.155	0.426
X1_humor_self-def	-0.256	0.0448	-5.72	< .001	-0.393	-0.528	-0.257

Usiamo i nostri dati in Jamovi ed excel per capire meglio

VD = RSE

VI = HUMOR self-enhancing (X2) + HUMOR self-defeating (X1)

3. usiamo Jamovi per regredire X1 su X2 calcolare i punteggi previsti di X1 e i suoi residui

I residui di X2 rappresentano
quella parte di variabilità di X2
che X1 non riesce a catturare, pertanto
i residui di X2 NON correlano con X1,
mentre i punteggi previsti di X2
Correlano perfettamente con X1

Coefficienti del Modello - X1_humor_self-def

						95% Intervallo di Fiducia	
Predittore	**Stima**	**SE**	**t**	**p**	**Stima Standard**	**Inferiore**	**Superiore**
Intercettare	20.558	2.7024	7.61	< .001			
x2_humor_self_enhanc	0.244	0.0799	3.05	0.003	0.221	0.0780	0.363

Usiamo i nostri dati in Jamovi ed excel per capire meglio

VD = RSE

VI = HUMOR self-enhancing (X1) + HUMOR self-defeating (X2)

4. ora possiamo calcolare la correlazione semi parziale che lega Y ai res di X1

5. e regredire Y su X2 per calcolarne punteggi previsti e residui, parzializzando così Y da X2

Coefficienti del Modello - Y_RSE

						95% Intervallo di Fiducia	
Predittore	**Stima**	**SE**	**t**	**p**	**Stima Standard**	**Inferiore**	**Superiore**
Intercettare	12.634	1.7686	7.14	< .001			
x2_humor_self_enhanc	0.147	0.0523	2.81	0.006	0.204	0.0605	0.347

Usiamo i nostri dati in Jamovi ed excel per capire meglio

VD = RSE

VI = HUMOR self-enhancing (X1) + HUMOR self-defeating (X2)

4. ora possiamo calcolare la correlazione semi parziale che lega Y ai res di X1

Correlazione Semiparziale

		Y_RSE	X1_humor_self-def
Y_RSE	r di Pearson	—	-0.383
	valore p	—	< .001
X1_humor_self-def	r di Pearson	-0.382	—
	valore p	< .001	—

Nota. controllo per 'x2_humor_self_enhanc'

Nota. La variazione dalle variabili di controllo viene rimossa solo dalle variabili nelle colonne

I valori sr e pr
ottenuti con Jamovi coincidono con quelli
calcolati in Excel

6. e calcolare la correlazione parziale tra i res di Y e i res di X1

Correlazione Parziale

		Y_RSE	X1_humor_self-def
Y_RSE	r di Pearson	—	
	valore p	—	
X1_humor_self-def	r di Pearson	-0.391	—
	valore p	< .001	—

Nota. controllo per 'x2_humor_self_enhanc'

Usiamo i nostri dati in Jamovi ed excel per capire meglio

VD = RSE

VI = HUMOR self-enhancing (X2) + HUMOR self-defeating (X1)

Ritorniamo al nostro output iniziale di ARM per rileggerlo e interpretarlo

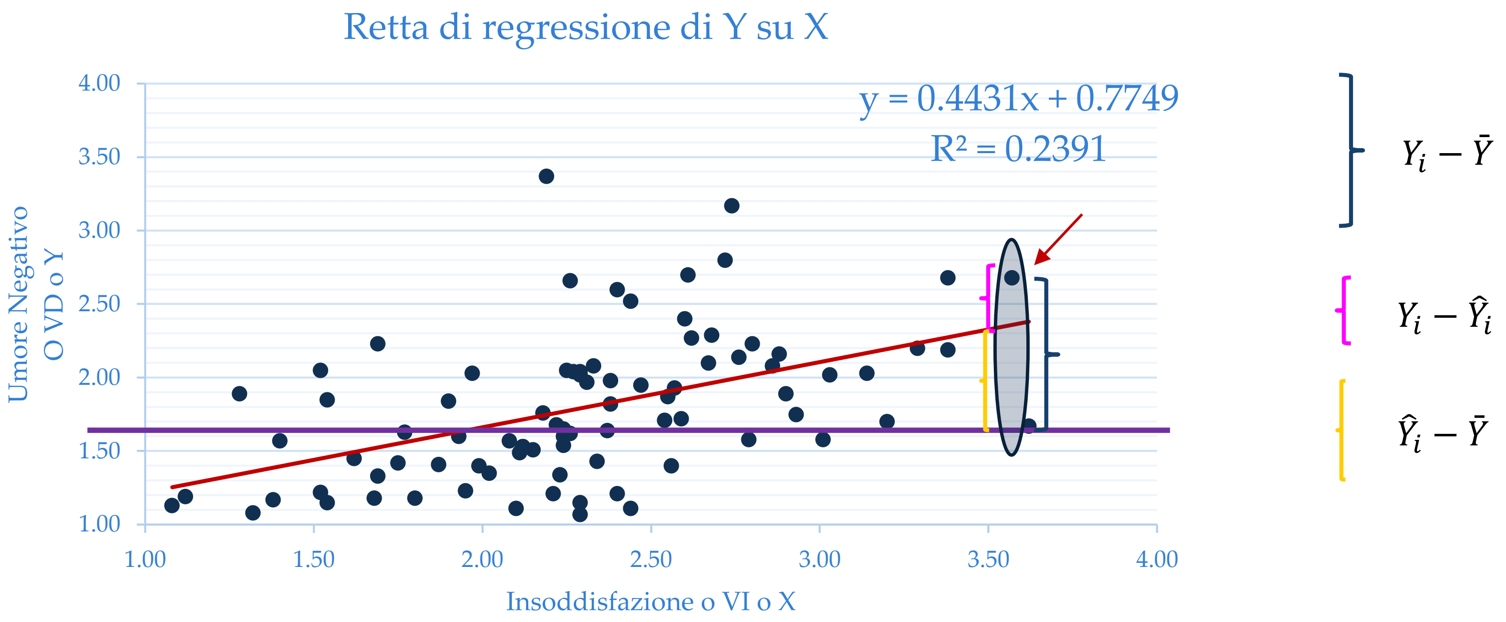
Coefficienti del Modello - Y_RSE

						95% Intervallo di Fiducia	
Predittore	**Stima**	**SE**	**t**	**p**	**Stima Standard**	**Inferiore**	**Superiore**
Intercettare	17.901	1.8735	9.56	< .001			
x2_humor_self_enhanc	0.209	0.0495	4.23	< .001	0.290	0.155	0.426
X1_humor_self-def	-0.256	0.0448	-5.72	< .001	-0.393	-0.528	-0.257

Mancano informazioni … quanta varianza di Y spiega il modello di previsione?

ARM: Il coefficiente di determinazione MULTIPLO

Ma riprendiamo quanto già sappiano dall'ARS



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

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

$$RPE = \frac{\sum(Y-\bar{Y})^2 - \sum\left(Y-\hat{Y}\right)^2}{\sum(Y-\bar{Y})^2} = \frac{\sum\left(\hat{Y}-\bar{Y}\right)^2}{\sum(Y-\bar{Y})^2} = \frac{\sum\left(\hat{Y}-\bar{Y}\right)^2/N-1}{\sum(Y-\bar{Y})^2/N-1} = \frac{s_{\hat{Y}}^2}{s_Y^2} = r_{Y'Y}^2$$

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: Il coefficiente di determinazione multiplo

Dai nostri dati i risultati mostrano un R quadro = 0.188, p < .001

R rappresenta la radice quadrata di R quadro, è sempre un valore positivo

Misure di Adattamento del Modello

			Test Globale del Modello			
Modello	**R**	**R^2**	**F**	**gdl1**	**gdl2**	**p**
1	0.434	0.188	21.0	2	181	< .001

ARM: Il coefficiente di determinazione multiplo

Dai nostri dati i risultati mostrano un R quadro = 0.188, p < .001

R rappresenta la radice quadrata di R quadro, è sempre un valore positivo

Misure di Adattamento del Modello

			Test Globale del Modello			
Modello	R	R^2	F	gdl1	gdl2	p
1	0.434	0.188	21.0	2	181	< .001

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

Strategia gerarchica e Delta di R quadro

Misure di Adattamento del Modello

			Test Globale del Modello			
Modello	R	R^2	F	gdl1	gdl2	p
1	0.329	0.108	22.1	1	182	< .001
2	0.434	0.188	21.0	2	181	< .001

Confronto dei Modelli

Confronto							
Modello		Modello	ΔR^2	F	gdl1	gdl2	p
1	-	2	0.0802	17.9	1	181	< .001

ARM: sulla significatività dei coefficienti di regressione parziale non std

Multicollinearità

Varianza residua

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

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

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

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

ARM: sulla significatività dei coefficienti di regressione parziale non std

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

rimando all'applicazione del calcolo di SE di b in excel