LA REGRESSIONE LINEARE

Sommario

* Scopo dell'analisi della regressione
* Regressione bivariata: Modello di base
* Regressione multipla: Modello di base
* Stima e interpretazione dei parametri
* Adeguatezza della soluzione
* Misure dell'associazione lineare tra Variabili Indipendenti (VI) e Variabile Dipendente (VD)
* Assunzioni
* Approcci analitici alla regressione
* Limiti

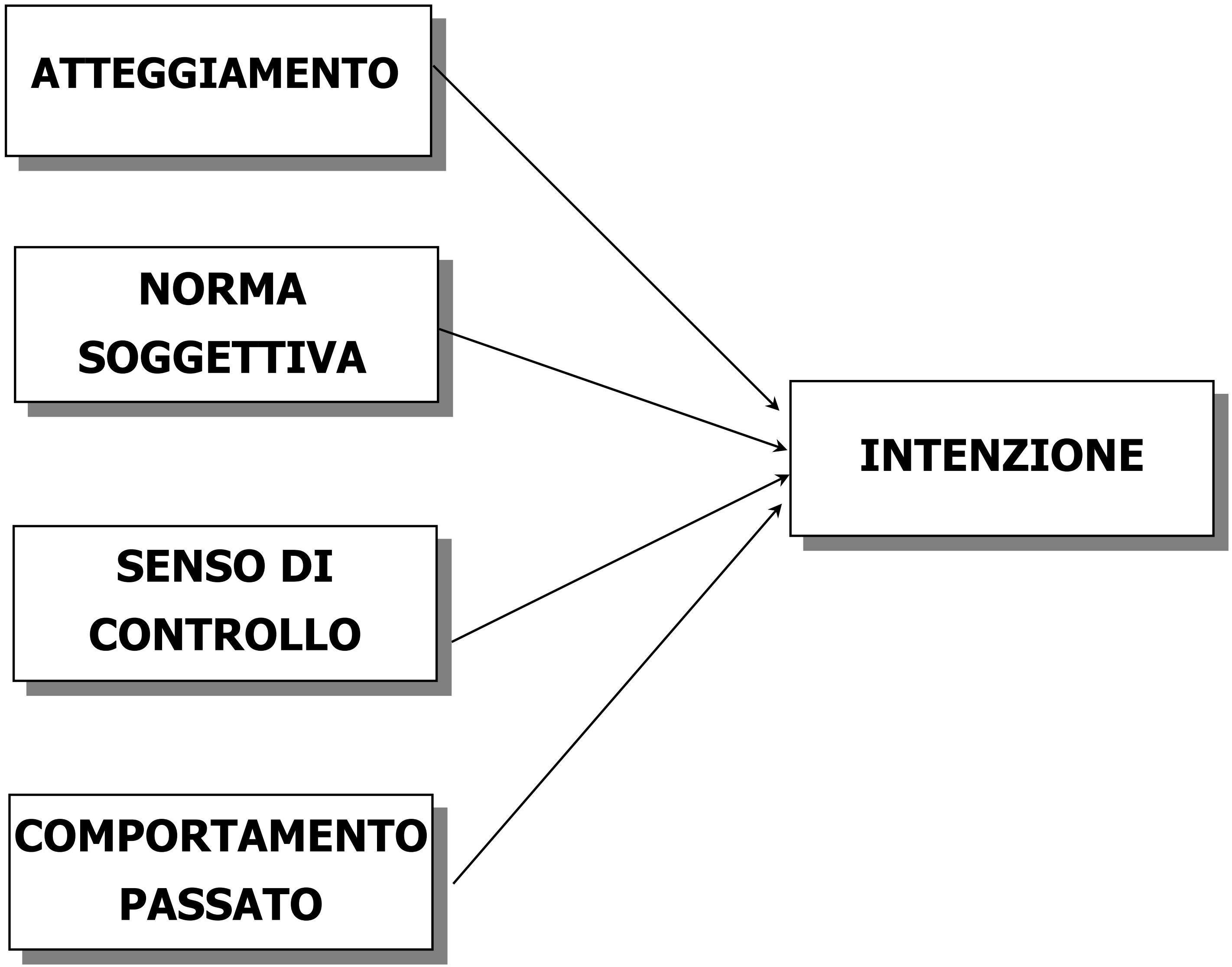
La Regressione esamina la relazione lineare tra una o più variabili esplicative (o indipendenti, VI, o "predittori") e una variabile criterio (o dipendente, VD).

Duplice scopo:

a) esplicativo: studiare e valutare gli effetti delle VI sulla VD in funzione di un determinato modello teorico

b) predittivo: individuare una combinazione lineare di VI per predire in modo ottimale il valore assunto dalla VD

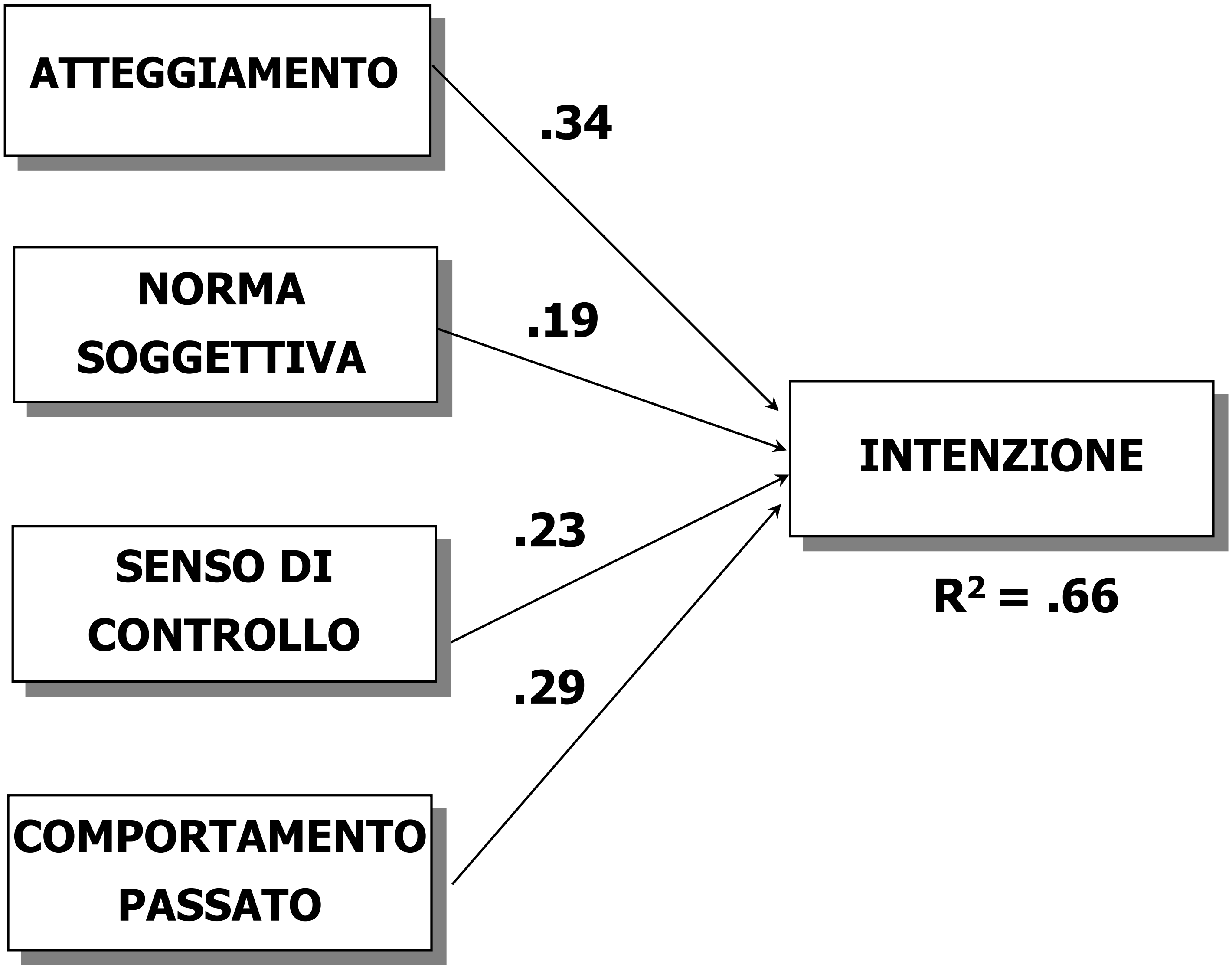
Da dove si parte: Modello concettuale



Da dove si parte: Matrice delle covarianze

	1	2	3	4	5
1.INT	6.438				
2.ATT	12.491	53.186			
3.NS	2.657	6.791	3.228		
4.CONTCO	2.650	5.534	1.149	3.453	
5.COMPAS	3.235	7.114	1.637	1.596	3.858

Dove si arriva: Modello empirico



Dove si arriva: Risultati del modello empirico

Variabile	B	Beta	T	p
Atteggiamento	.12	.34	6.38	.001
Norma Soggettiva	.28	.19	3.83	.001
Senso di Controllo	.32	.23	4.82	.001
Comport. Passato	.38	.29	5.65	.001

R^2 = .66; t= 16.74. p<.0001

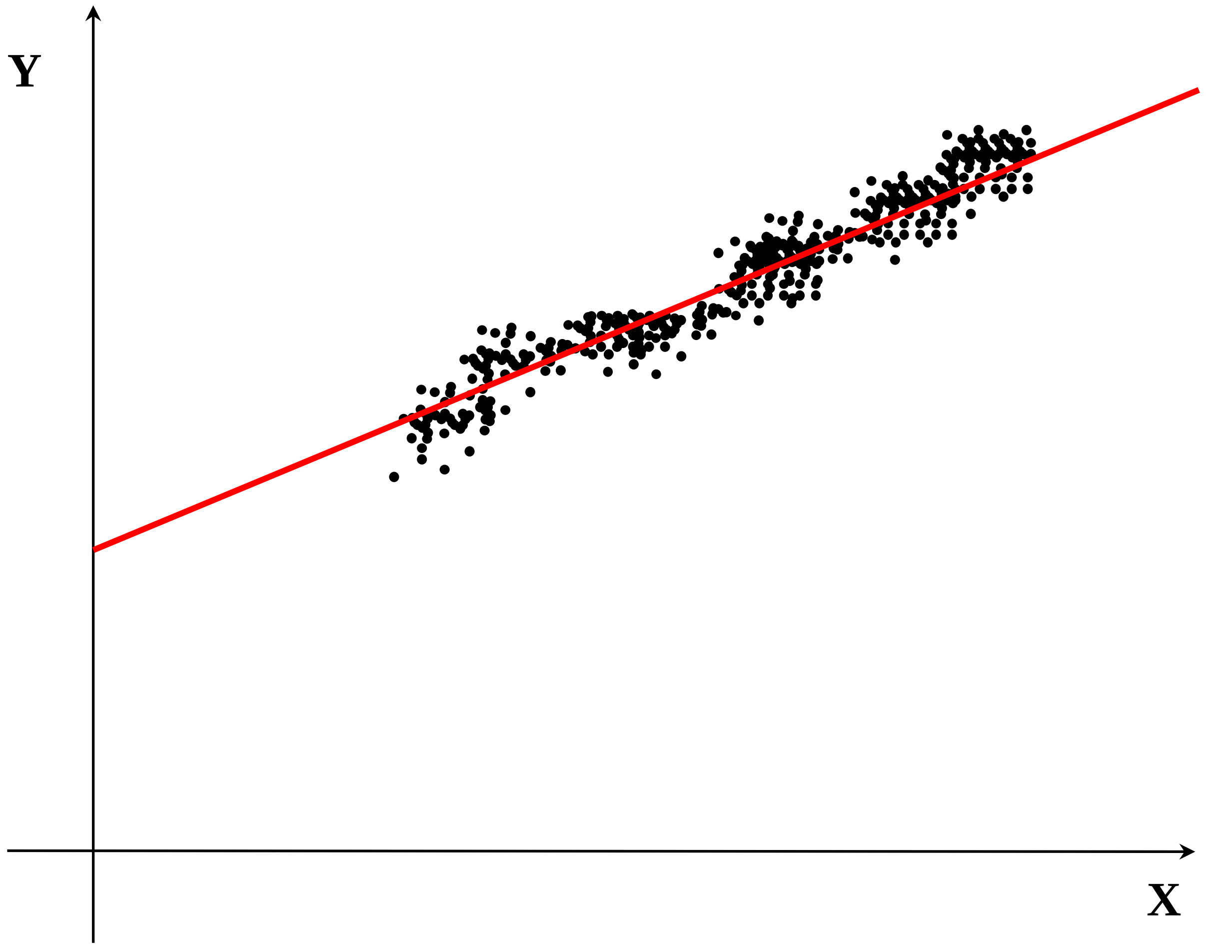
Regressione bivariata (o semplice)

Una sola variabile indipendente (VI) sulla quale "regredisce" la variabile dipendente (VD). Si ipotizza che la VI "determini" o "influenzi" o "predica" la VD.

Individuare quella retta che consente di prevedere al meglio i punteggi nella VD a partire da quelli nella VI.

Individuare la retta che "interpola" meglio la nuvola di punti (o "scatterplot") della distribuzione congiunta delle due variabili.

La retta di regressione (regressione bivariata)



Regressione bivariata (o semplice)

La relazione lineare è quella più parsimoniosa ed è quella più realistica in moltissimi casi.
L'equazione che lega Y a X è la seguente:

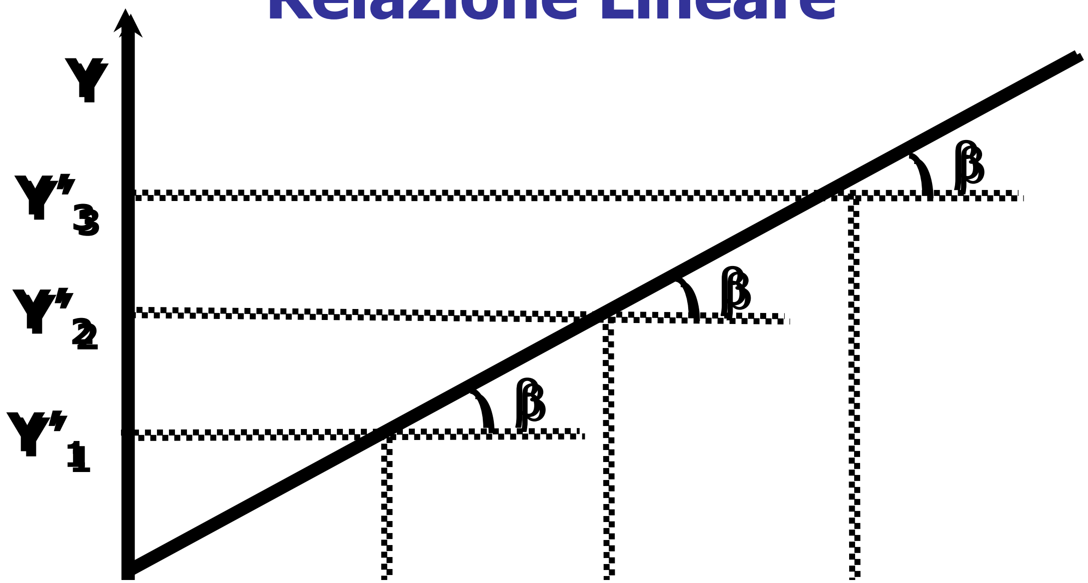
$$Y = \alpha + \beta X$$

Parametri dell'equazione:

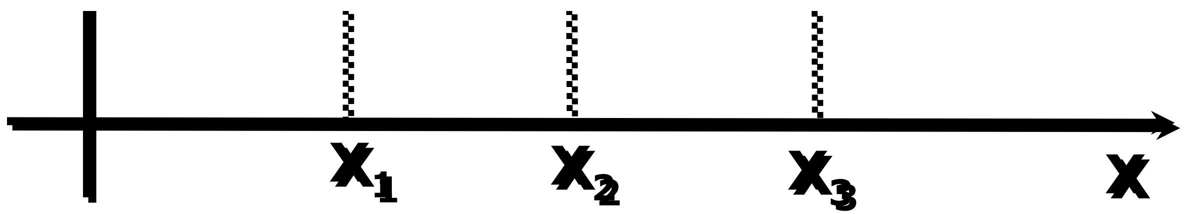
<u>Intercetta</u>: α, punto in cui la retta incrocia l'asse delle ordinate (altezza della linea).

<u>Coefficiente angolare</u>: β inclinazione della retta di regressione di Y su X; indica di quante unità cambia Y per una variazione unitaria che si verifica nella X.

Relazione Lineare

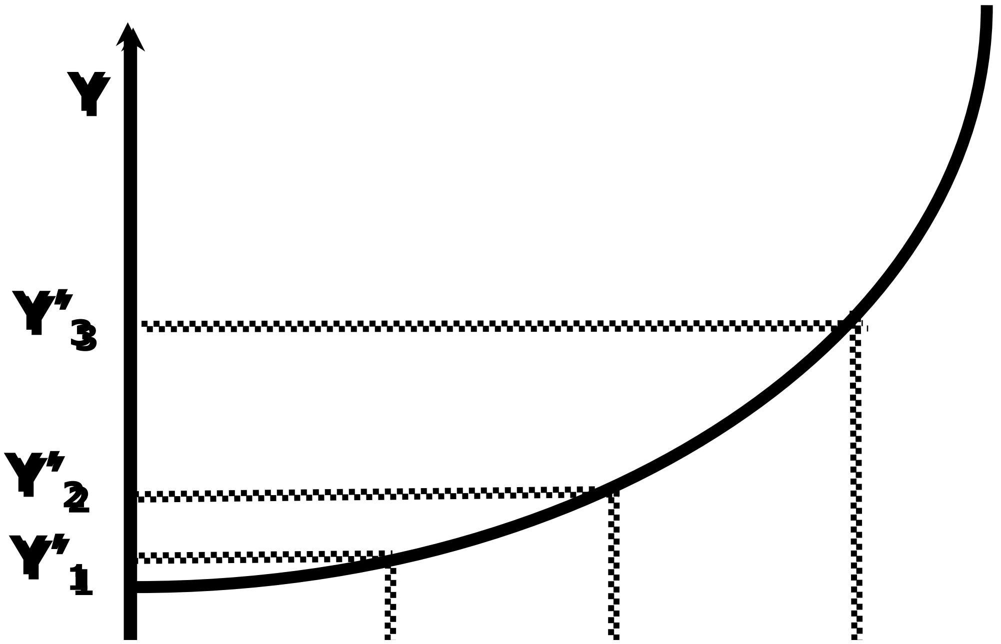


Per ogni variazione in X si determina sempre la stessa variazione in Y qualunque sia il valore di X sull'asse delle ascisse.

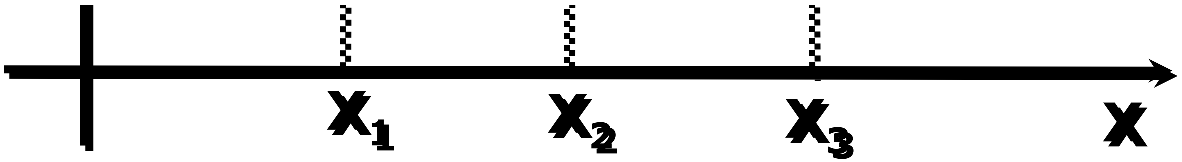


$$(X_3-X_2)=(X_2-X_1) \Rightarrow (Y'_3-Y'_2)=(Y'_2-Y'_1)$$

Relazione Non Lineare



La stessa variazione in X determina variazioni diverse in Y per diversi valori di X sull'asse delle ascisse.



$(X_3-X_2)=(X_2-X_1)$ **Ma** $(Y'_3-Y'_2)\neq(Y'_2-Y'_1)$

Errore o residuo

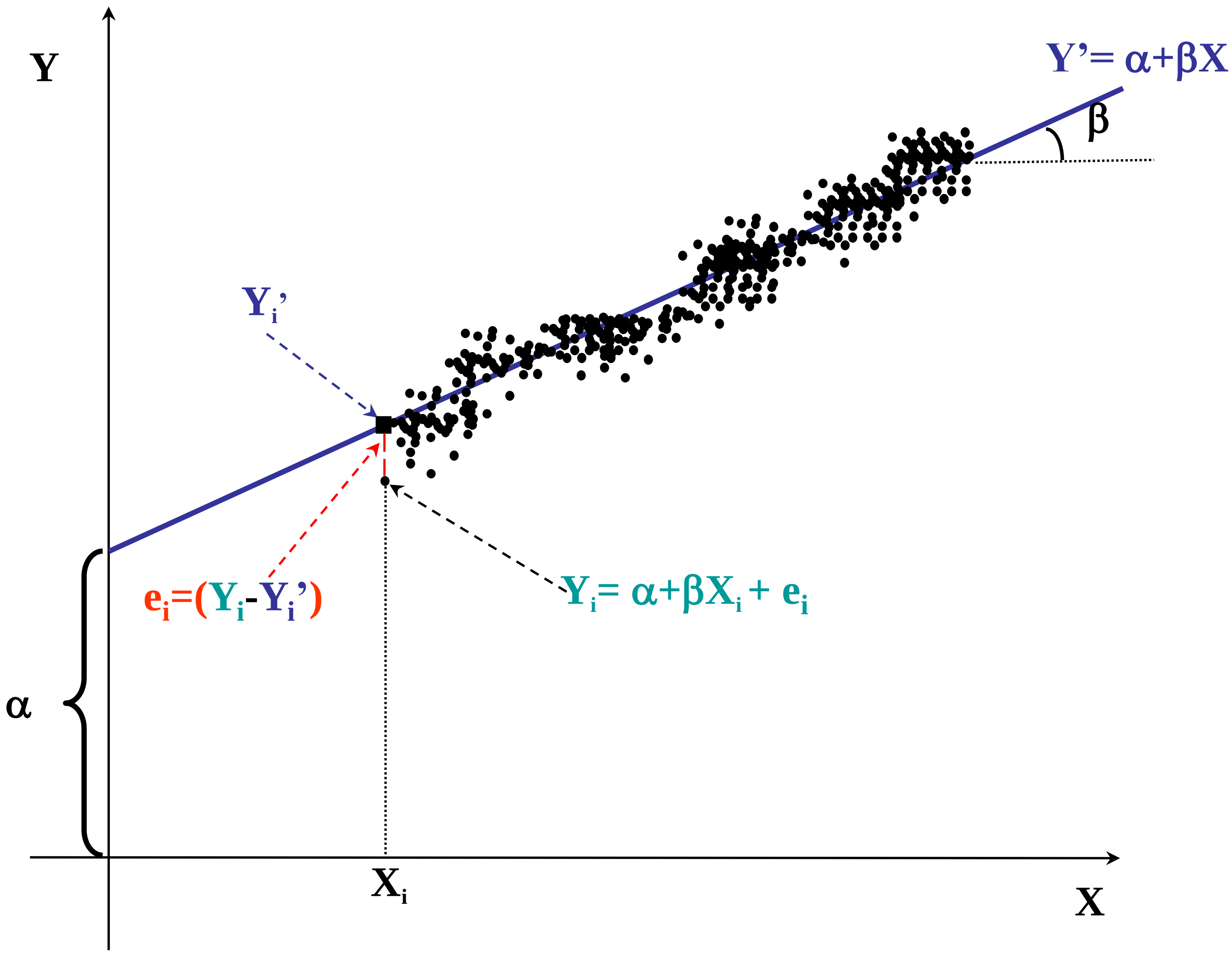
I punti sono dispersi intorno alla retta di regressione perché:
- le variabili sono misurate con errore;
- la relazione può non essere perfettamente lineare;
- predittori importanti possono essere omessi.

L'equazione quindi deve incorporare un termine di errore (o residuo) per ogni caso.

$$Y = \alpha + \beta X + e = Y' + e$$

$Y' = \alpha + \beta X$: valore "teorico" della Y, ottenuto dalla regressione.

"<u>e</u>": Residuo, deviazione del punteggio osservato Y dal punteggio teorico Y'.


Y
Y'= α+βX
β
Yi'
ei=(Yi-Yi')
Yi= α+βXi + ei
α
Xi
X

La Stima dei parametri

Bisogna identificare la retta che meglio si <u>adatta</u> ai punti che descrivono la distribuzione delle Y sulle X.

La retta che interpola meglio il diagramma di dispersione, cioè quella retta che passa più vicina possibile alla nuvola dei punti, è quella che rende minima la somma delle differenze al quadrato tra le Y osservate e le Y' teoriche.

I parametri α e β vengono stimati nel campione attraverso il metodo dei minimi quadrati, ovvero il metodo che rende minimo l'errore che si commette quando Y viene "stimato" dalla equazione di regressione.

Equazione dei minimi quadrati:

$$\Sigma(Y_i - Y_i')^2 = \Sigma(Y_i-(a+bx_i))^2 = \min$$

Identifica la retta che riduce al minimo l'errore che viene commesso nello stimare Y da X.

Formule dei minimi quadrati per il calcolo di a e b:

$$b = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)} \qquad a = \bar{Y} - b\bar{X}$$

Il coefficiente "a" rappresenta il valore atteso di Y quando X è uguale a 0.
Il coefficiente "b" rappresenta il cambiamento atteso in Y associato a un cambio di una unità in X.

Stime standardizzate

Il coefficiente di regressione esprime la relazione tra Y e X nell'unità di misura delle 2 variabili. Per esprimere questa relazione in una scala di misura comprensibile si deve standardizzarlo.

Il coefficiente standardizzato si ottiene moltiplicando il coefficiente "grezzo" (non standardizzato) per il rapporto delle deviazioni standard della VI e della VD:

$$\hat{\beta} = b\,(s_x/s_y)$$

Nella regressione semplice è uguale al coefficiente di correlazione "semplice", ovvero: $\hat{\beta} = r_{yx}$

La regressione multipla

Una variabile dipendente che regredisce su almeno due variabili indipendenti. Equazione di regressione:

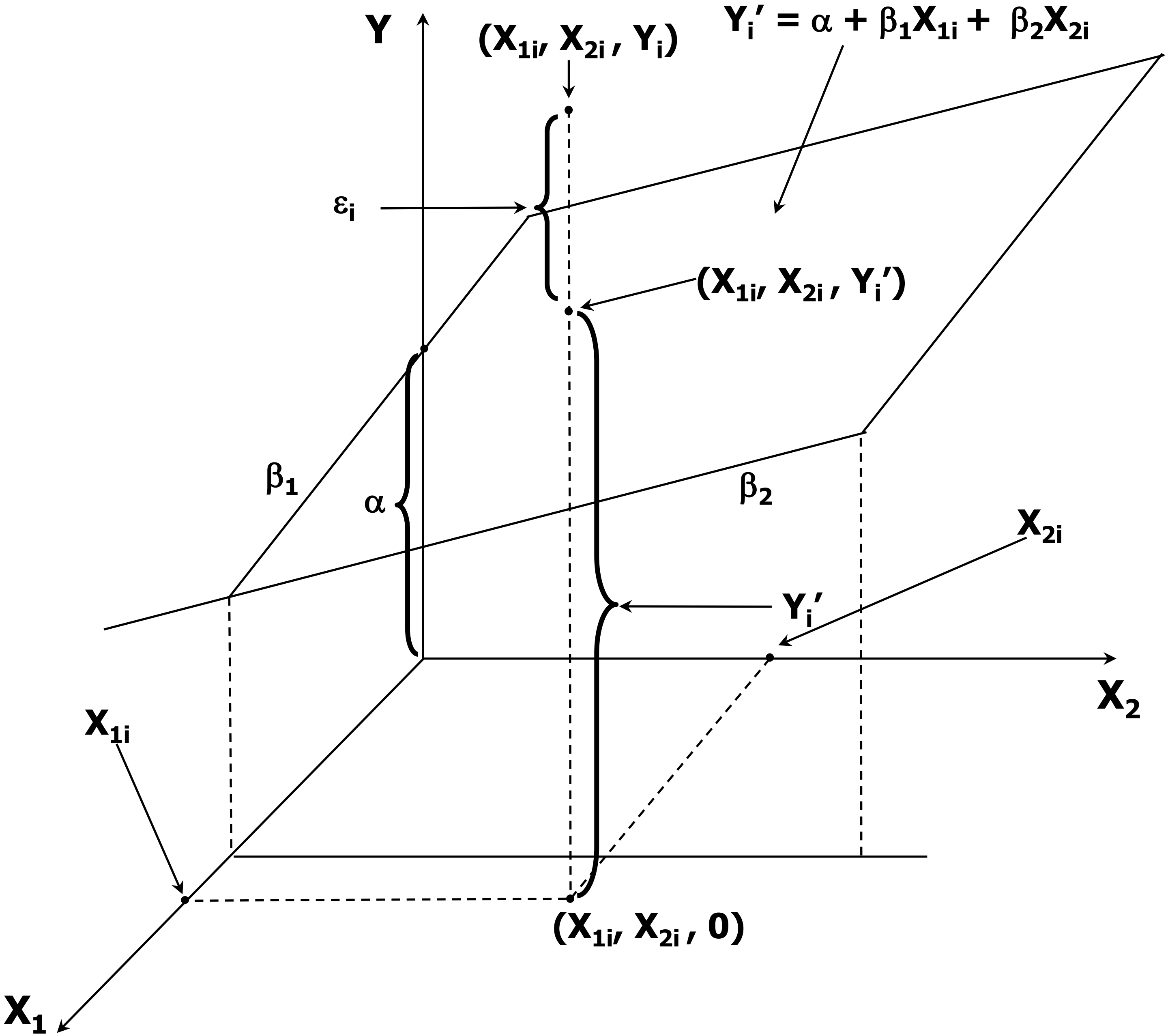
$$Y' = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon_i$$

Piano di regressione (due VI);
Iperpiano (più di 2 VI).
Equazione del piano di regressione:

$$Y' = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Rappresentazione grafica: il piano di regressione



Coefficienti di regressione della regressione multipla: coefficienti "parziali" o "netti" (partial slope o partial regression coefficient).

Dipendenza della variabile Y da ciascuna delle VI X_i, al netto delle altre VI nell'equazione.
Per ogni VI rappresentano l'inclinazione della retta di regressione della variabile dipendente, ottenuta mantenendo costanti i valori delle altre VI.

Nel piano:

β_1 è l'inclinazione della retta di regressione di Y su X_1 quando si mantiene costante X_2
β_2 è l'inclinazione della retta di regressione di Y su X_2, se si mantiene costante X_1.

Stime dei coefficienti: minimi quadrati.

Individuare un iperpiano di dimensioni k che si adatti meglio ai punti nello spazio di dim. k+1 (k VI e 1 VD).

$$\Sigma\,[Y - (\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)]^2 = \min$$

Espressioni matriciali delle equazioni:

$y = bX + e$	equazione di regressione	(1)
$b=(X'X)^{-1} X'Y$	coefficienti di regressione	(2)
$e=Y - (Xb+a)$	residui	(3)

X'X rappresenta la codevianza tra le VI, X'Y rappresenta la codevianza tra VI e VD.

Relazioni tra una VD Y e due VI X1 e X2, espresse in termini della varianza che condividono:

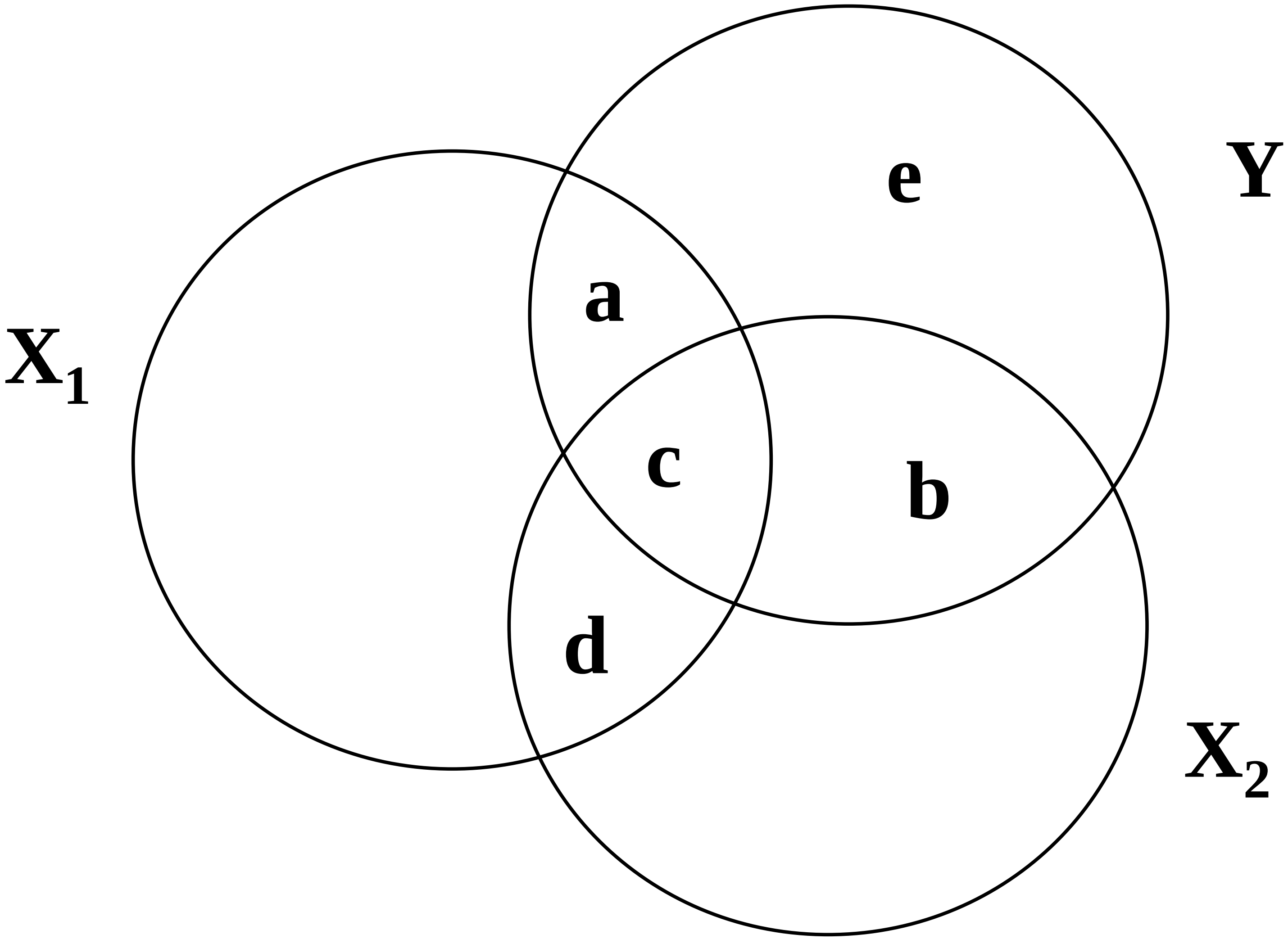
- "a+c": varianza in comune tra X1 e Y, e "a": varianza che Y condivide <u>solo</u> con X1;

- "c+b": varianza in comune tra X2 e Y, e "b": che Y condivide <u>solo</u> con X2;

- "c+d": varianza in comune tra X1 e X2;

- "e" var. che Y non condivide né con X1 né con X2;

Relazioni tra una VD Y e due VI X1 e X2



Coefficienti che misurano l'associazione tra VD e VI.

1. Coefficiente di Correlazione Semi-parziale: corr. tra X1 e Y, se X2 viene parzializzata solo da X1.

$$sr_{y1.2} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{1 - r_{12}^2}} \qquad sr^2_{y1.2...k} = R^2_{y.12..i...k} - R^2_{y.12..(i)...k}$$

$$sr^2_{y1.2} = a/(a+c+b+e)$$

Proporzione della varianza totale di Y spiegata unicamente da X1 al netto di X2,

$$F_i = \frac{sr^2_{y1.2...k}}{(1 - R^2)/df_{res}}, \; df = (1, N - k - 1)$$

Coefficienti che misurano l'associazione tra VD e VI.

2. Coefficiente di Correlazione Parziale: corr. tra X1 e Y, se X2 viene parzializzata da X1 e da Y.

$$pr_{y1.2} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{(1 - r_{y2}^{2})(1 - r_{12}^{2})}}$$

$$pr^2_{y1.2} = a/(a+e)$$

Proporzione della varianza di Y non spiegata da X2, spiegata unicamente da X1 al netto di X2.

Formula alternativa:

$$pr^2_{y1.2....k} = \frac{sr^2_{y1.2...k}}{1 - R^2_{y.12..(i)...k}}$$

Coefficienti che misurano l'associazione tra VD e VI.

3. Coefficiente di Regressione:
Inclinazione della retta di regressione di Y su X_1 per valori costanti di X_2, cambiamento atteso in Y in seguito ad un cambiamento di una unità (b) o di una deviazione standard (b^) in X_1 al netto di X_2.

$$b_{y1.2} = \frac{b_{y1} - b_{y2} b_{12}}{1 - r_{12}^2} \qquad \beta_{y1.2}\wedge = b_{y1.2} \frac{s_y}{s_1} = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2}$$

b_{y1}, b_{y2}, b_{12}: coefficienti delle regressioni bivariate rispettivamente di Y su X_1, di Y su X_2 e di X_1 su X_2.

Adeguatezza della equazione di regressione

1) $\Sigma(Y_i - \overline{Y})^2$ devianza totale delle Y_i dalla loro media.

2) $\Sigma(Y_i' - \overline{Y})^2$ devianza di Y_i spiegata dalla regressione. Scarto tra la retta dei minimi quadrati e la media: quanto migliora la previsione di Y per il fatto di conoscere X.

3) $\Sigma(Y_i - Y_i')^2$ è la devianza di Y_i non spiegata dalla regressione. Scarto di Y_i dalla retta dei minimi quadrati: quantità di errore che si commette per predire Y con Y'.

Adeguatezza della equazione di regressione

E' possibile dimostrare che:

$$r^2 = \frac{\sum (Y_i' - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\text{Devianza Spiegata}}{\text{Devianza Totale}}$$

Dividendo i due termini per n:

$$r^2 = \frac{\sum (Y_i' - \bar{Y})^2 / n}{\sum (Y_i - \bar{Y})^2 / n} = \frac{\text{Varianza Spiegata}}{\text{Varianza Totale}}$$

r^2 = <u>coefficiente di determinazione</u> = indice della proporzione della varianza totale di Y che viene spiegata dalla regressione lineare di Y su X.

Adeguatezza della equazione di regressione

$(1\text{-}r^2)$ = proporzione della varianza totale di Y che <u>non</u> è spiegata dalla regressione di Y su X.

E' possibile dimostrare infatti che:

$$(1-r^2) = \frac{\sum (Y_i - Y_i^{'})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\text{Devianza Residua}}{\text{Devianza Totale}}$$

$\sqrt{(1-r^2)} =$ **<u>coefficiente di alienazione</u> = parte di deviazione standard di Y non spiegata dalla regressione**

Adeguatezza della equazione di regressione

Da $\sqrt{(1-r^2)}$ è possibile ricavare il coefficiente che rappresenta la dispersione intorno alla retta dei minimi quadrati per ogni valore di X: "errore standard della stima" ed è un indice della precisione della retta di regressione

$$S_e = \sqrt{(1-r^2)}S_y = \sqrt{\frac{\sum (Y-Y')^2}{n-2}}$$

Se r = 0, $S_e = S_y$ e la varianza d'errore coincide con la varianza totale di Y;
Se r = 1 $S_e = 0$ tutti gli Y cadono sulla retta di regressione Y', quindi l'errore è uguale a 0.

Varianza spiegata nella regressione multipla

Coefficiente di determinazione multiplo (R^2): indica la proporzione di varianza della VD spiegata dalle VI prese nel loro complesso.

$$R^2_{y.12...k} = \sum r_{yi}\hat{\beta}_{yi}$$

Nel caso di due variabili indipendenti la formula è:

$$R^2_{y.12} = r_{y1}\hat{\beta}_{y1} + r_{y2}\hat{\beta}_{y2}$$

Somma dei prodotti delle correlazioni semplici (o "di ordine zero") e dei coefficienti β^ tra VD e ogni VI.

Varianza spiegata nella regressione multipla

R^2 non diminuisce mai se si aggiungono altre VI. Correzione per il numero di VI: coefficiente <u>corretto</u> (<u>Adjusted</u>, o Shrunken).

$$AR^2 = R^2 - (1-R^2)*(k/(N-k-1))$$

Può diminuire rispetto a R^2 se le VI aggiunte forniscono un contributo mediocre alla spiegazione della varianza della VD.

Coefficiente di correlazione multiplo (R o RM): associazione tra una VD e un insieme di VI.

Coefficiente di correlazione multiplo:

$$R_{y.12...k} = \sqrt{R^2_{y.12...k}}$$

R è sempre maggiore/uguale a 0, ed è maggiore dei singoli coefficienti di ordine zero.

VI molto correlate: R vicino al più elevato coefficiente di correlazione semplice tra le VI e la VD.

VI poco correlate: R più elevato del più grande dei coefficienti di correlazione di ordine zero.

Verifica delle ipotesi (test di significatività)

Significatività statistica di R^2
Ipotesi statistiche: H_0: r = 0; H_1: r > 0
(equivale a H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$)

Varianza	Somme dei quadrati	Gradi di Libertà	Stime della Varianza	F
Totale	Σy^2	N−1		
Spiegata	$R^2\Sigma y^2$	k	$\dfrac{R^2\Sigma y^2}{K}$	$\dfrac{(N-k-1)R^2}{k(1-R^2)}$
Non Spiegata	$(1-R^2)\Sigma y^2$	N−k−1	$\dfrac{(1-R^2)\Sigma y^2}{(N-k-1)}$	

dove $y=(Y - \bar{Y})$ e k è il numero di VI.

Verifica delle ipotesi (test di significatività)

Significatività statistica dei singoli b:

H_0: b = 0; H_1: b $\neq$ 0

t =(b - 0)/S_b, con N-k-1 gradi di libertà.

Stima dell'errore standard di β:

$$S_b = \frac{s_y}{s_i}\sqrt{\frac{1-R_Y^2}{N-k-1}}\sqrt{\frac{1}{1-R_i^2}} = \sqrt{\frac{S_e^2}{S_i^2(1-R_i^2)}}$$

Assunzioni alla base della regressione multipla

1. Assenza di errore di specificazione
   a. Relazione tra le X_i e Y lineare
   b. Non sono state omesse VI rilevanti
   c. Non sono state incluse VI irrilevanti
2. Assenza di errore di misurazione: variabili misurate senza errore
3. VI quantitative o dicotomiche, VD quantitativa
4. Varianza della VI è > 0
5. Campionamento casuale
6. Nessuna VI è combinazione lineare perfetta delle altre (assenza di perfetta multicollinearità)

Assunzioni alla base della regressione multipla

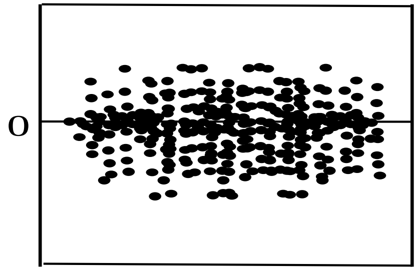
7. **Assunzioni sui residui (o termini di errore) ε_i**
   a. **Media uguale a zero: $E(\varepsilon_i)=0$**
   b. **Omoschedasticità, $VAR(\varepsilon_i)=s2$**
   c. **Assenza di autocorrelazione: $Cov(\varepsilon_i, \varepsilon_j)=0$**
   d. **VI non correlate con gli errori: $Cov(\varepsilon_i,X_i)=0$**
   e. **Normalità: Le distribuzioni dei valori di ε_i per ogni valore dato di X sono di forma normale**

Violazione delle assunzioni:

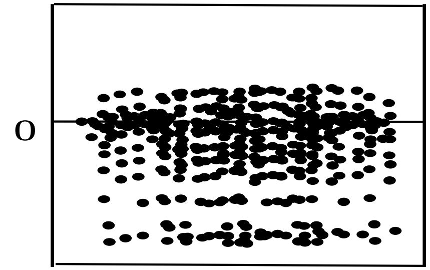
Esame della distribuzione dei residui e=(Y-Y’) rispetto ai punteggi teorici Y’.

Utile per rilevare:

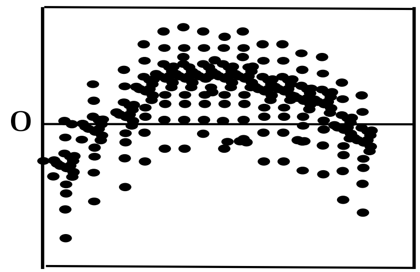
- **La non linearità della relazione tra VI e VD, e tra VI,**
- **La non omogeneità della varianza**
- **La non normalità dei residui**



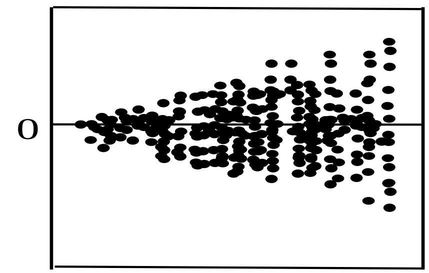
1.Assunzioni rispettate



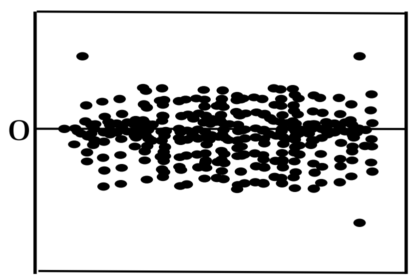
2.Non normalità



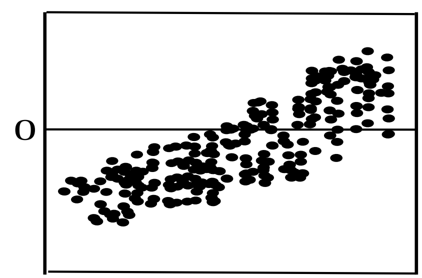
3.Non linearità



4.Eteroschedasticità



5. Casi estremi



6. Autocorrelazione

Nei riquadri 1-5: Punteggi predetti Y': in ascisse; Residui (Y-Y'): in ordinate.
Nel riquadro 6: Tempo o ordine di acquisizione: in ascisse; Residui (Y-Y'): in ordinate.

Rilevare la collinearità (correlazione elevata tra le VI):

- **Correlazioni tra le VI (se sono >.8);**
- **R^2 elevati e b bassi;**
- **Errori standard elevati;**
- **Indici di tolleranza e VIF.**

Tolleranza di una VI: quantità di varianza che *non* è spiegata dalle altre VI: $T_i = (1 - R_i^2)$
valori bassi di tolleranza indicano alta collinearità, valori alti bassa collinearità.

Variance Inflaction Factor (VIF): $VIF_i = 1/T_i = 1/(1 - R_i^2)$;

valori bassi del VIF indicano bassa collinearità, valori alti elevata collinearità.

Non indipendenza degli errori (Autocorrelazione):

Test di Durbin-Watson.
Ha un valore compreso tra 0 e 4: se i residui di osservazioni consecutive non sono correlati il test di Durbin-Watson ha un valore intorno a 2.

Se $n \geq 100$ e le VI almeno 2, valori compresi tra 1.5 e 2.2 possono essere considerati indicativi di assenza di autocorrelazione, quindi:
Valori inferiori a 1.5 = autocorrelazione positiva.
Valori superiori a 2.2 =autocorrelazione negativa.

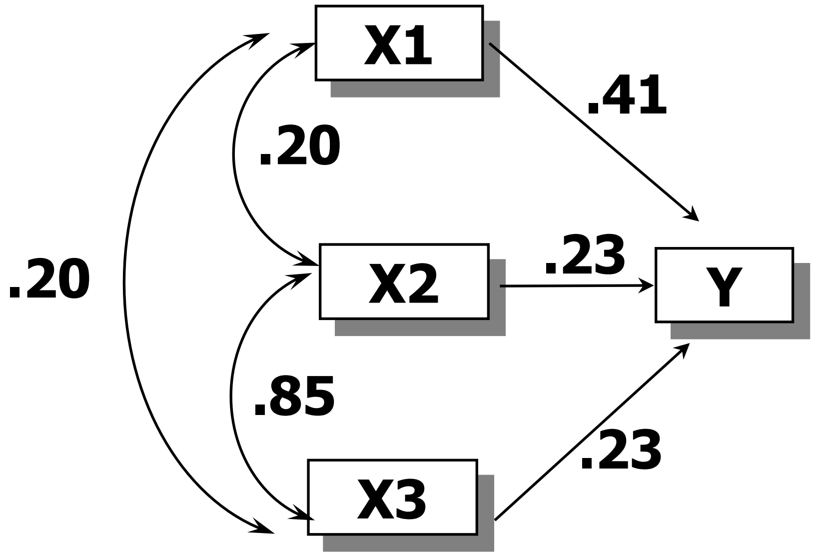
Rimedi per risolvere le violazioni: trasformazione delle variabili originali (logaritmo, reciproco, radice quad.).

Scomposizione degli effetti

La *ridondanza* riguarda il caso in cui i coefficienti di correlazione semiparziale (sr), parziale (pr) e di regressione standardizzato (β) sono inferiori (in valore assoluto) al coefficiente di correlazione semplice r e hanno il suo stesso segno.

Allora ogni variabile indipendente porta un'informazione sulla variabile dipendente che in parte si sovrappone con quella veicolata dalle altre variabili indipendenti.

Ridondanza



r(X1,Y)=r(X2,Y)=r(X3,Y)=.50

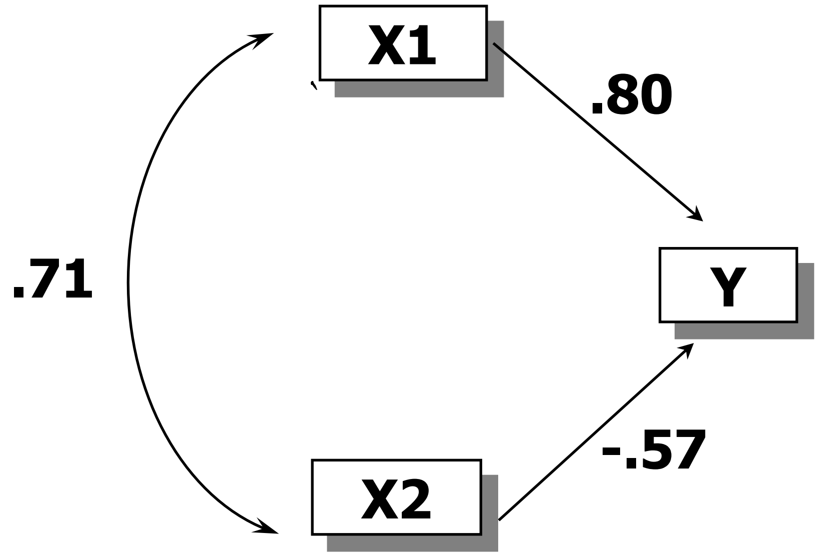
Scomposizione degli effetti

La *soppressione* riguarda il caso in cui i coefficienti sr, pr e β sono maggiori (in valore assoluto) del coefficiente di correlazione semplice r.

Il termine soppressione indica che la relazione tra le variabili indipendenti "maschera" o "sopprime" la loro reale relazione con la variabile dipendente, che potrebbe essere maggiore o addirittura di segno opposto se le variabili indipendenti *non* fossero correlate. Il soppressore è una VI la cui inclusione nella regressione aumenta l'effetto di un'altra VI sulla VD.

Un caso particolare di soppressione è il ribaltamento, dove il coefficiente parziale assume il segno opposto del coefficiente semplice.

Soppressione

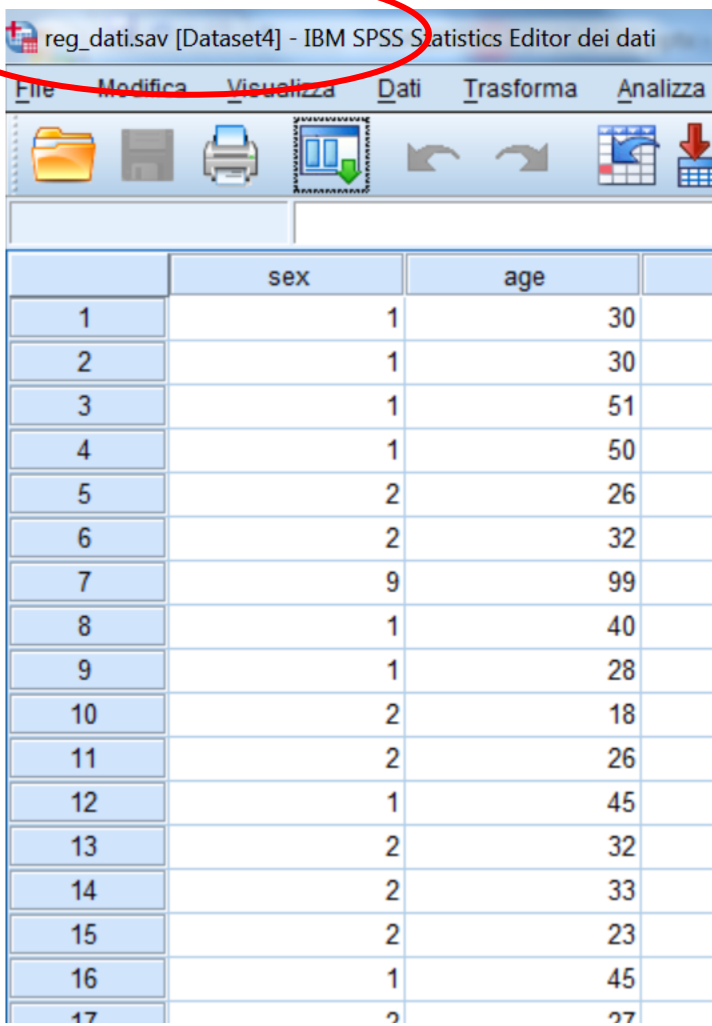


r(X1,Y)=.40; r(X2,Y)=0

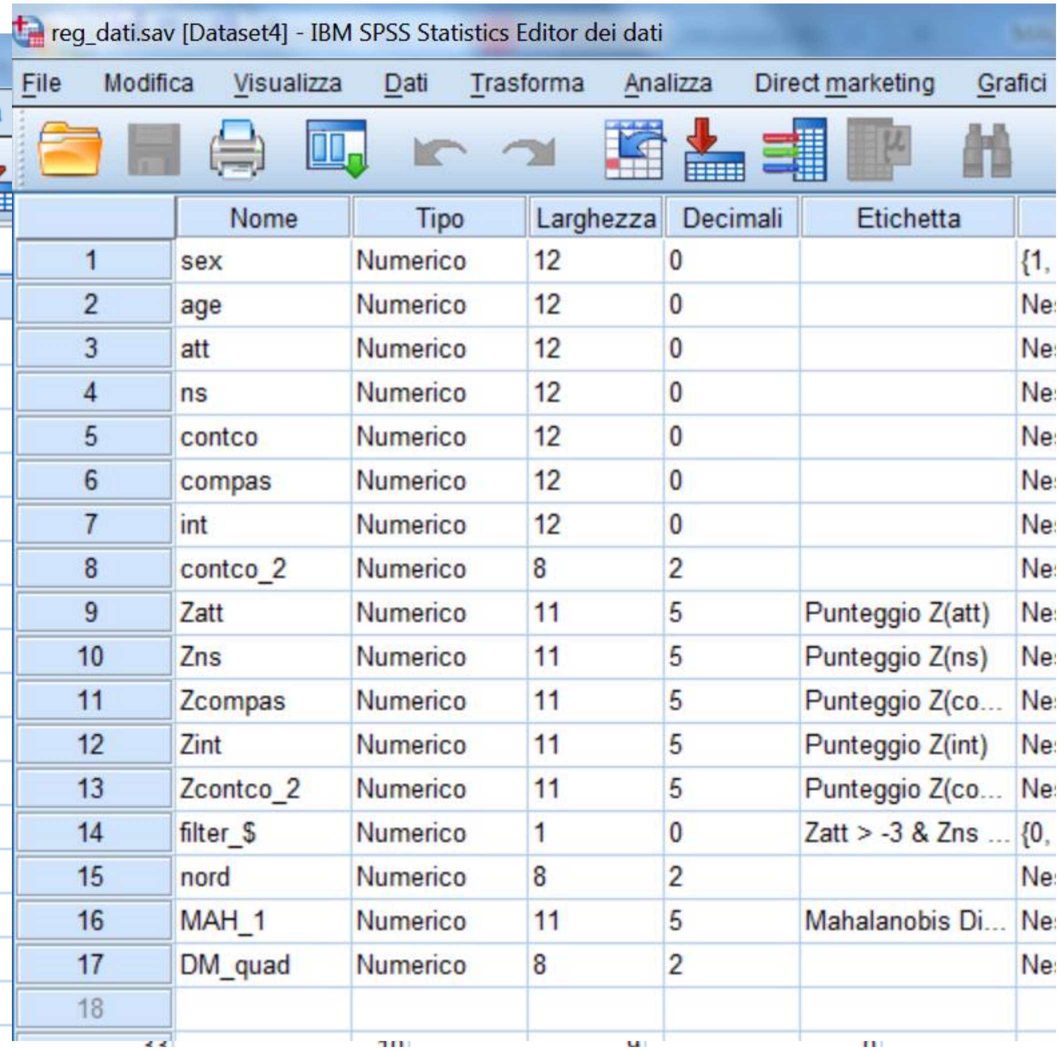
REGRESSIONE

CON SPSS

Carichiamo i dati utilizzati per l'esempio sul trattamento preliminari (rinominare il file come reg_dati.sav).

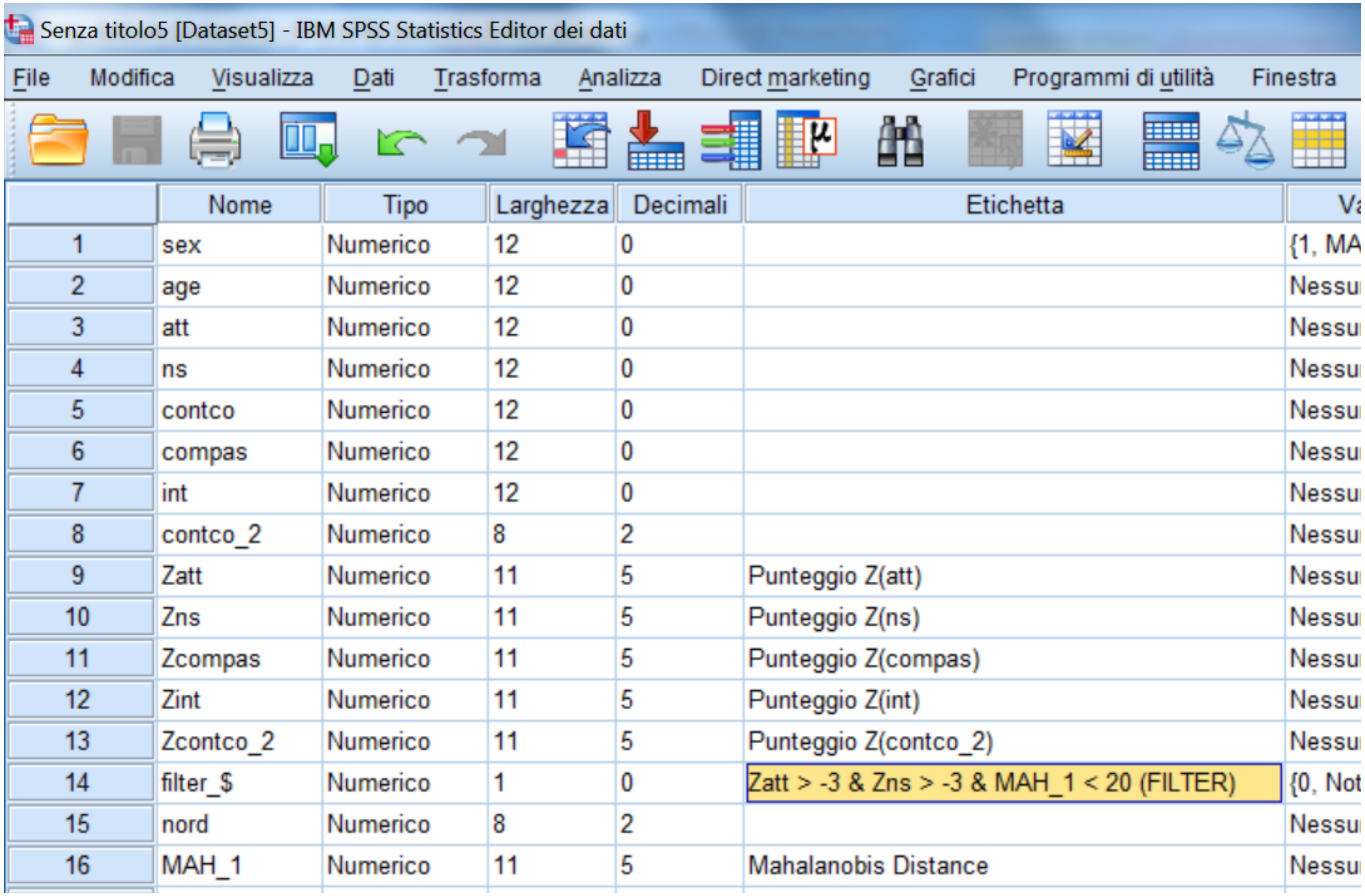


	sex	age
1	1	30
2	1	30
3	1	51
4	1	50
5	2	26
6	2	32
7	9	99
8	1	40
9	1	28
10	2	18
11	2	26
12	1	45
13	2	32
14	2	33
15	2	23
16	1	45



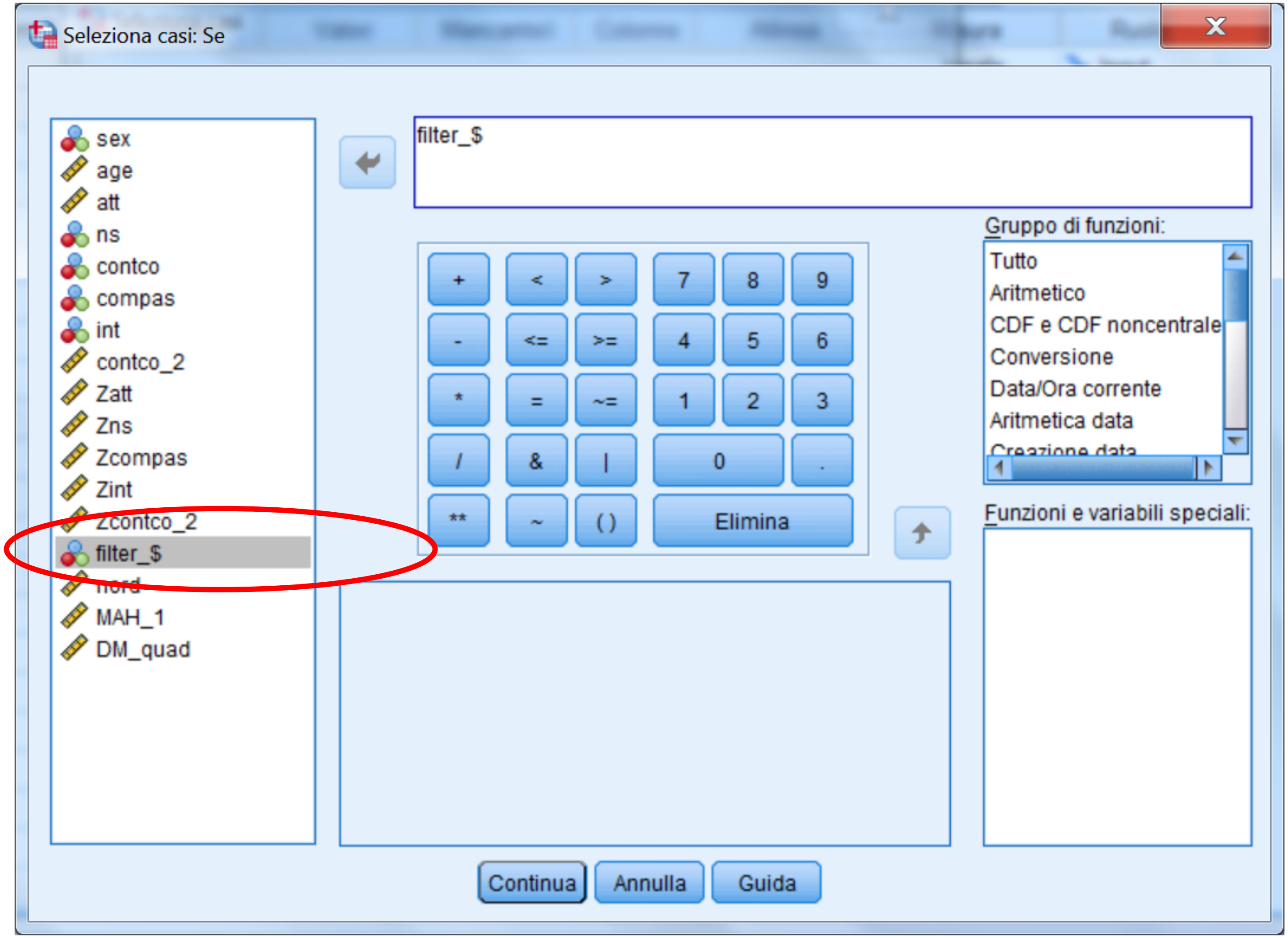
	Nome	Tipo	Larghezza	Decimali	Etichetta	
1	sex	Numerico	12	0		{1,
2	age	Numerico	12	0		Ne
3	att	Numerico	12	0		Ne
4	ns	Numerico	12	0		Ne
5	contco	Numerico	12	0		Ne
6	compas	Numerico	12	0		Ne
7	int	Numerico	12	0		Ne
8	contco_2	Numerico	8	2		Ne
9	Zatt	Numerico	11	5	Punteggio Z(att)	Ne
10	Zns	Numerico	11	5	Punteggio Z(ns)	Ne
11	Zcompas	Numerico	11	5	Punteggio Z(co...	Ne
12	Zint	Numerico	11	5	Punteggio Z(int)	Ne
13	Zcontco_2	Numerico	11	5	Punteggio Z(co...	Ne
14	filter_$	Numerico	1	0	Zatt > -3 & Zns ...	{0,
15	nord	Numerico	8	2		Ne
16	MAH_1	Numerico	11	5	Mahalanobis Di...	Ne
17	DM_quad	Numerico	8	2		Ne
18						

Riattiviamo il filtro per i 3 outliers



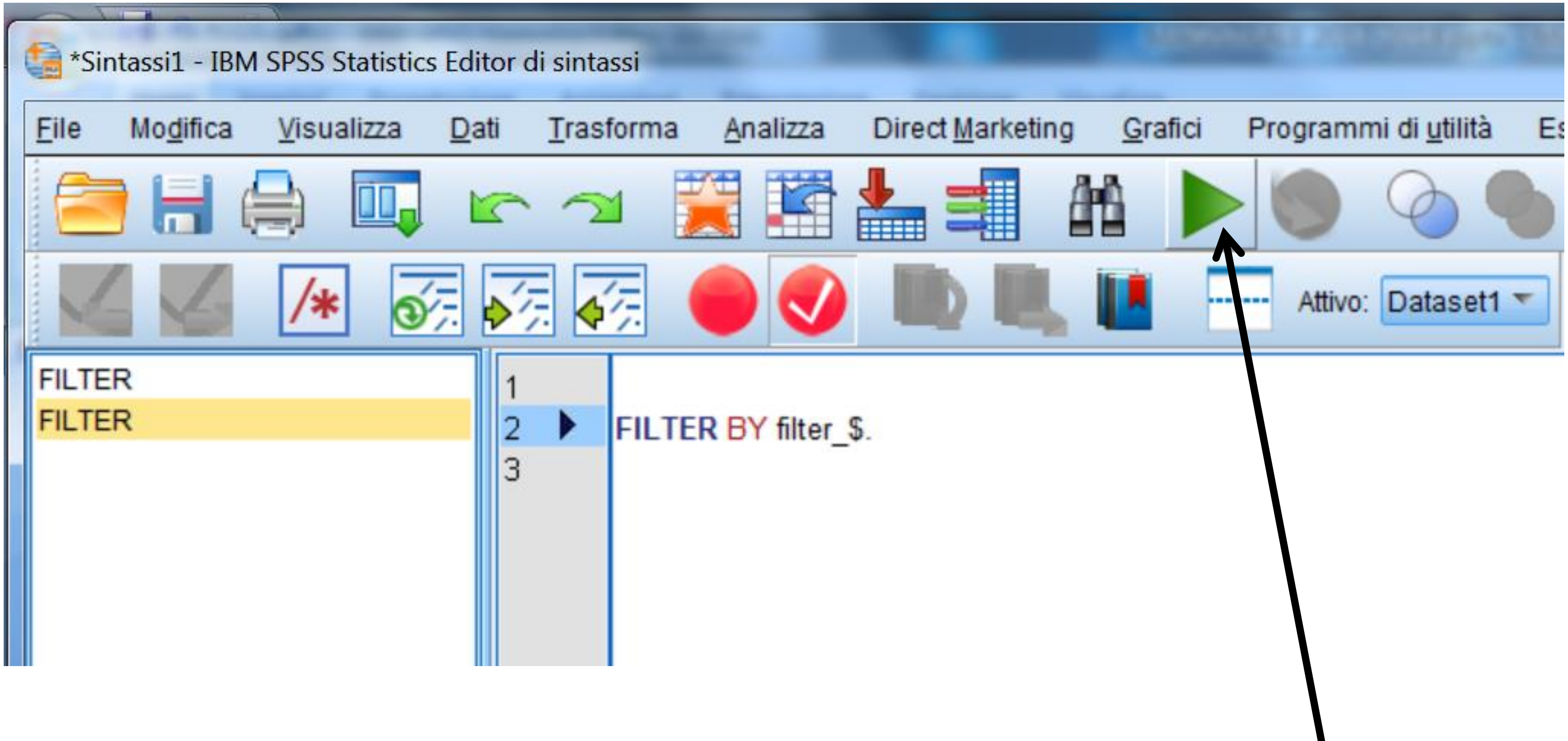
	Nome	Tipo	Larghezza	Decimali	Etichetta	Va
1	sex	Numerico	12	0		{1, MA
2	age	Numerico	12	0		Nessu
3	att	Numerico	12	0		Nessu
4	ns	Numerico	12	0		Nessu
5	contco	Numerico	12	0		Nessu
6	compas	Numerico	12	0		Nessu
7	int	Numerico	12	0		Nessu
8	contco_2	Numerico	8	2		Nessu
9	Zatt	Numerico	11	5	Punteggio Z(att)	Nessu
10	Zns	Numerico	11	5	Punteggio Z(ns)	Nessu
11	Zcompas	Numerico	11	5	Punteggio Z(compas)	Nessu
12	Zint	Numerico	11	5	Punteggio Z(int)	Nessu
13	Zcontco_2	Numerico	11	5	Punteggio Z(contco_2)	Nessu
14	filter_$	Numerico	1	0	Zatt > -3 & Zns > -3 & MAH_1 < 20 (FILTER)	{0, Not
15	nord	Numerico	8	2		Nessu
16	MAH_1	Numerico	11	5	Mahalanobis Distance	Nessu

Riattiviamo il filtro per i 3 outliers

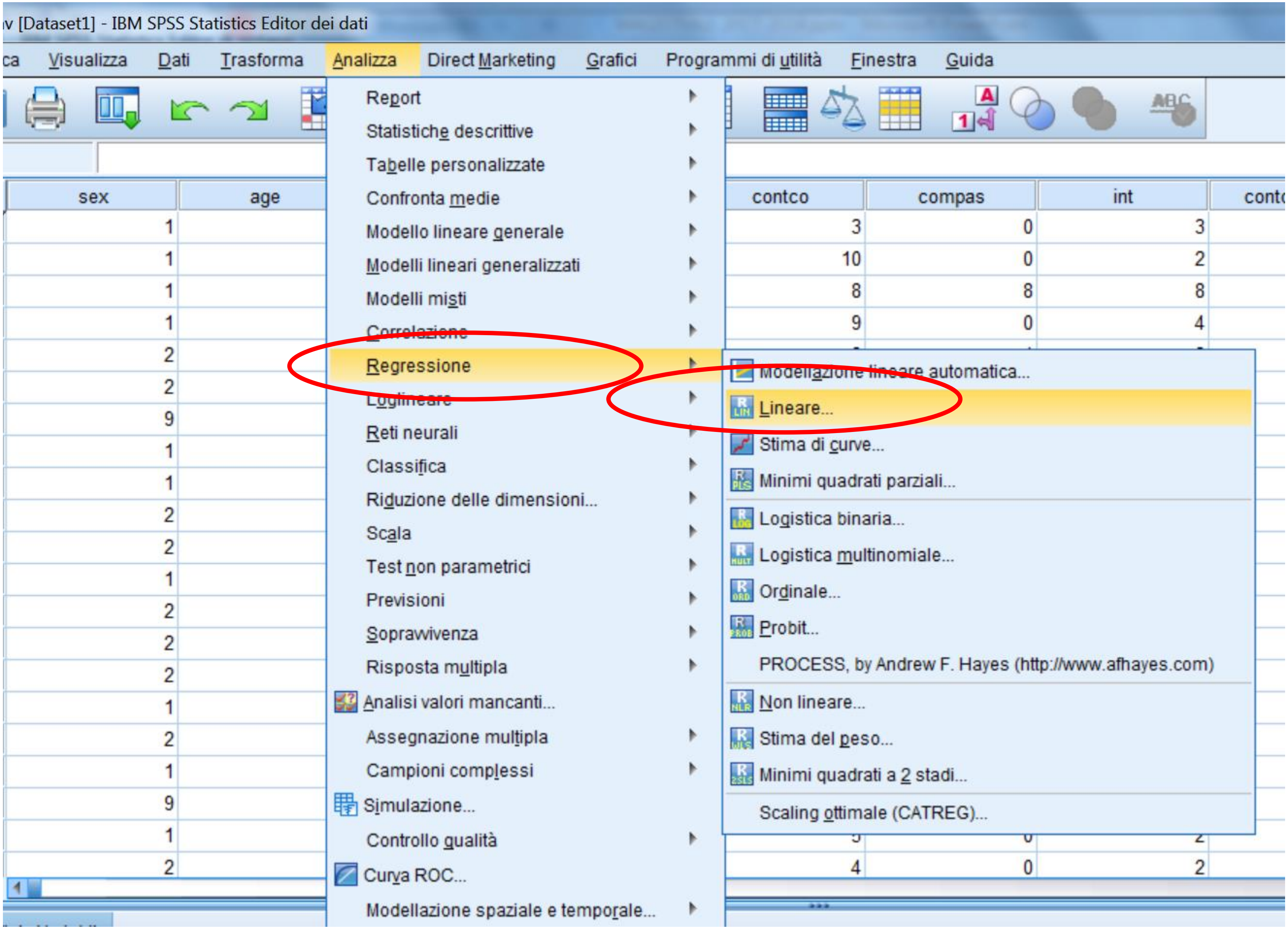


Questa procedura però cancella l'etichetta della variabile filter_$

Riattivare il filtro per i 3 outliers senza cancellare l'etichetta della variabile filter_$

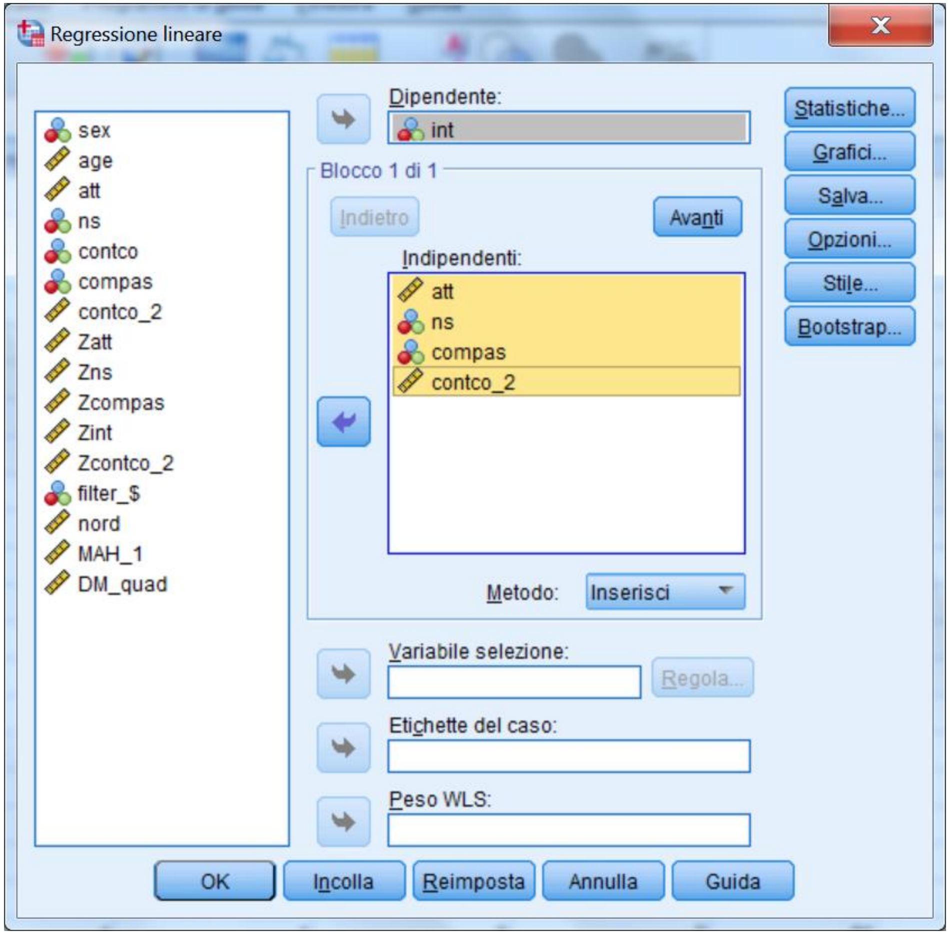


Dalla finestra Sintassi lanciare il comando FILTER BY filter_$ posizionando il cursore sulla linea del comando e cliccando sul triangolino verde.


v [Dataset1] - IBM SPSS Statistics Editor dei dati
ca Visualizza Dati Trasforma Analizza Direct Marketing Grafici Programmi di utilità Finestra Guida
Report
Statistiche descrittive
Tabelle personalizzate
Confronta medie
Modello lineare generale
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Reti neurali
Classifica
Riduzione delle dimensioni...
Scala
Test non parametrici
Previsioni
Sopravvivenza
Risposta multipla
Analisi valori mancanti...
Assegnazione multipla
Campioni complessi
Simulazione...
Controllo qualità
Curva ROC...
Modellazione spaziale e temporale...
Modellazione lineare automatica...
Lineare...
Stima di curve...
Minimi quadrati parziali...
Logistica binaria...
Logistica multinomiale...
Ordinale...
Probit...
PROCESS, by Andrew F. Hayes (http://www.afhayes.com)
Non lineare...
Stima del peso...
Minimi quadrati a 2 stadi...
Scaling ottimale (CATREG)...
sex
age
contco
compas
int

Regressione standard

Selezionare la variabile dipendente ("int") e poi tutte le variabili indipendenti ("att", "ns", "contco_2", "compas") che verranno inserite in un unico blocco. Lasciare nell'opzione "Metodo" il valore di default "Inserisci".



Strategie Analitiche per la regressione

Regressione standard:

- Quale è l'entità della relazione globale tra VD e VI?
- Quale è il contributo unico di ciascuna VI nel determinare questa relazione ?

Regressione gerarchica:

- Se la VI X1 è inserita dopo la VI X2, quale contributo aggiuntivo dà alla spiegazione della VD ?

Regressione statistica:

- Quale è la migliore combinazione lineare di VI per predire la VD in un determinato campione ?

La regressione standard

Tutte le VI vengono inserite nell'equazione simultaneamente.

Ogni VI è trattata come se fosse inserita nell'equazione dopo aver preso in considerazione tutte le altre VI.

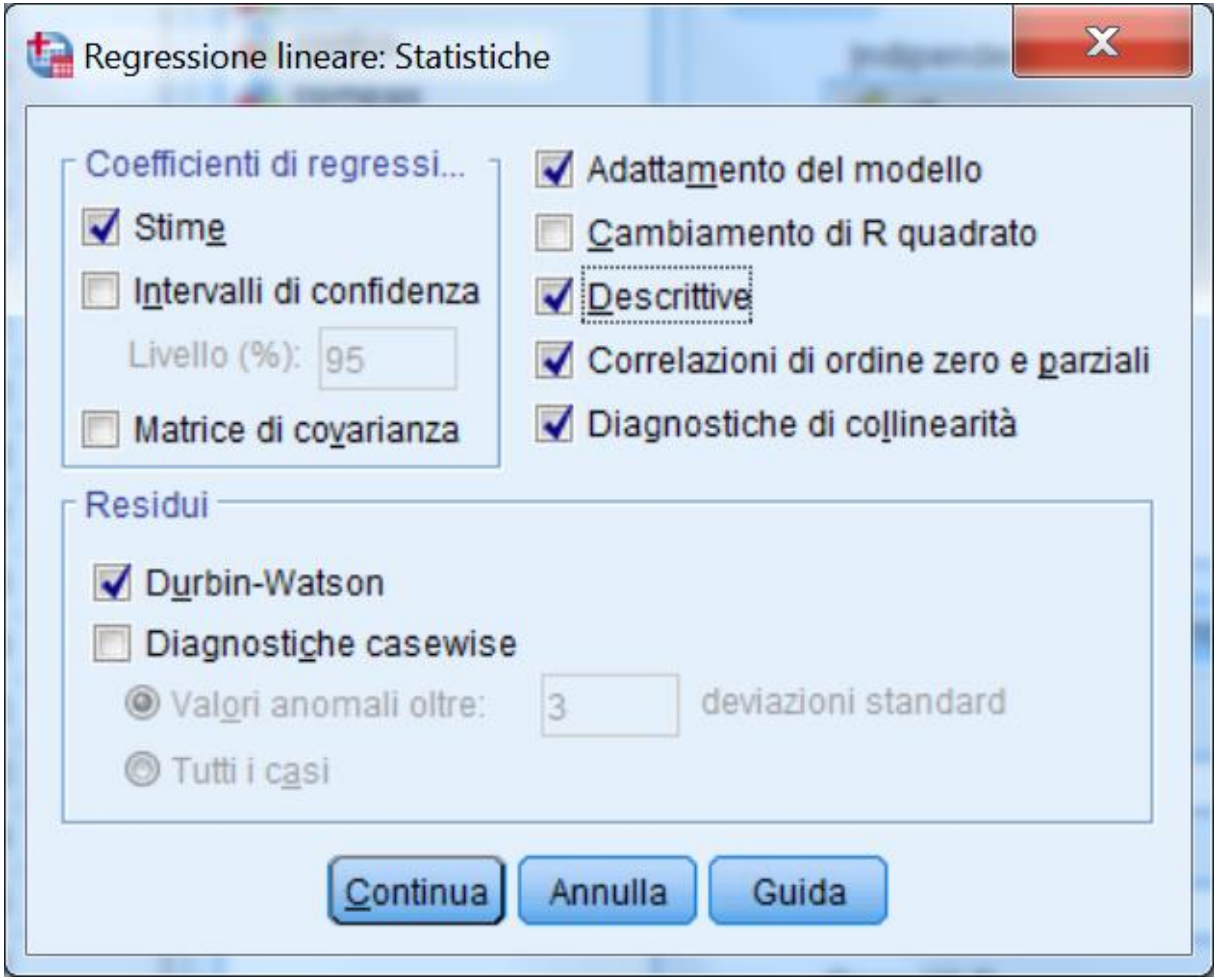
Ogni VI è valutata per quanto aggiunge, nello spiegare la VD, a quanto viene spiegato da tutte le altre VI.

Ogni VI spiega solo quella parte di varianza della VD che condivide unicamente con la VD, al netto delle VI.

La variabilità che la VD condivide simultaneamente con più VI viene ad aggiungersi all'R^2 ma non è assegnata individualmente a nessuna delle VI.

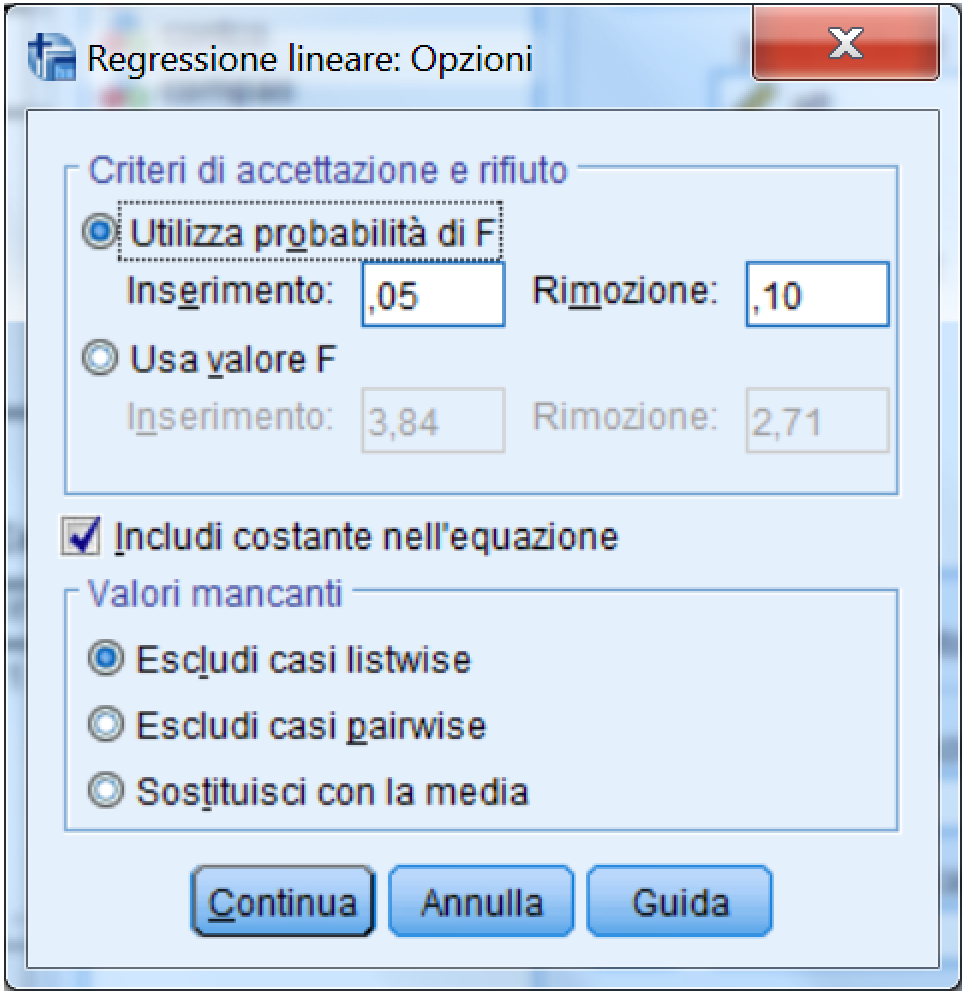
Regressione standard

Nella finestra di dialogo "Statistiche" bisogna selezionare determinati parametri per ottenere nell'output le informazioni necessarie per interpretare e valutare la soluzione.



Regressione standard

Nella finestra di dialogo "Opzioni" vengono presentate le opzioni relative al trattamento dei valori mancanti.



Pairwise

Vengono utilizzati tutti i valori disponibili
Le analisi vengono effettuate considerando tutti i soggetti che hanno valori validi sulle variabili di volta in volta considerate

Listwise

Vengono utilizzati solo quei soggetti che NON hanno alcun valore mancante. È sufficiente che un soggetto presenti un valore mancante in una sola variabile per essere escluso dalle analisi
Per molte procedure è il metodo di *default* di SPSS

Sostituzione con la media

Sostituisce i valori mancanti con la media della variabile nel campione

Statistiche descrittive

Statistica descrittiva

	Media	Deviazione std.	N
int	7,32	2,543	196
att	42,93	7,024	196
ns	7,92	1,758	196
compas	2,67	1,965	196
contco_2	,2545	,28864	196

Correlazioni

		int	att	ns	compas	contco_2
Correlazione di Pearson	int	1,000	,721	,589	,645	-,544
	att	,721	1,000	,556	,520	-,449
	ns	,589	,556	1,000	,454	-,315
	compas	,645	,520	,454	1,000	-,445
	contco_2	-,544	-,449	-,315	-,445	1,000
Sign. (a una coda)	int	.	,000	,000	,000	,000
	att	,000	.	,000	,000	,000
	ns	,000	,000	.	,000	,000
	compas	,000	,000	,000	.	,000
	contco_2	,000	,000	,000	,000	.
N	int	196	196	196	196	196
	att	196	196	196	196	196
	ns	196	196	196	196	196
	compas	196	196	196	196	196
	contco_2	196	196	196	196	196

Regressione standard

Il pannello iniziale evidenzia come che tutte le variabili siano state inserite in un unico passo

Variabili immesse/rimosse[a]

Modello	Variabili immesse	Variabili rimosse	Metodo
1	contco_2, ns, compas, att[b]	.	Inserisci

a. Variabile dipendente: int

b. Sono state immesse tutte le variabili richieste.

La varianza spiegata si trova in questa tabella

Riepilogo del modello[b]

Modello	R	R-quadrato	R-quadrato adattato	Errore std. della stima	Durbin-Watson
1	,819[a]	,671	,664	1,474	1,709

a. Predittori: (costante), contco_2, ns, compas, att

b. Variabile dipendente: int

ANOVA[a]

Modello		Somma dei quadrati	gl	Media quadratica	F	Sign.
1	Regressione	845,599	4	211,400	97,259	,000[b]
	Residuo	415,151	191	2,174		
	Totale	1260,750	195			

a. Variabile dipendente: int

b. Predittori: (costante), contco_2, ns, compas, att

Regressione standard

Per interpretare gli effetti delle VI guardare questa tabella

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sign.
		B	Errore std.	Beta		
1	(Costante)	-1,422	,816		-1,742	,083
	att	,141	,020	,390	7,045	,000
	ns	,273	,074	,189	3,676	,000
	compas	,354	,067	,273	5,274	,000
	contco_2	-1,656	,426	-,188	-3,885	,000

a. Variabile dipendente: int

Correlazioni			Statistiche di collinearità	
Ordine zero	Parziale	Parte	Tolleranza	VIF
,721	,454	,293	,563	1,775
,589	,257	,153	,653	1,530
,645	,357	,219	,642	1,557
-,544	-,271	-,161	,737	1,358

Risultati della regressione standard

sr^2 = contributo unico della VI all'R^2 nell'insieme di VI.

Somma degli sr^2: può non raggiungere il valore di R^2.

Differenza tra somma degli sr^2 e R^2: proporzione di varianza della VD spiegata simultaneamente da più VI, ma non attribuita a nessuna VI in particolare.

Dati dell'esempio:

$\Sigma sr^2 = (.29)^2+(.15)^2+(.22)^2+(-.16)^2=.183$; $R^2 = .671$;

$R^2 - \Sigma sr^2 = .67 - .183 = .488$

E' la varianza spiegata simultaneamente dalle VI

Regressione standard

Varianza unica e varianza comune spiegata dalla VI

	varianza unica	
	sr	sr^2
att	,293	0,086
ns	,153	0,023
compas	,219	0,048
contco_2	-,161	0,026
Varianza totale spiegata		0,671
Varianza unica spiegata		0,183
Varianza comune spiegata		0,488

La regressione gerarchica

Le VI vengono inserite nell'equazione secondo un ordine specificato dal ricercatore.

L'ordine di "entrata" viene assegnato dal ricercatore secondo considerazioni teoriche o logiche.

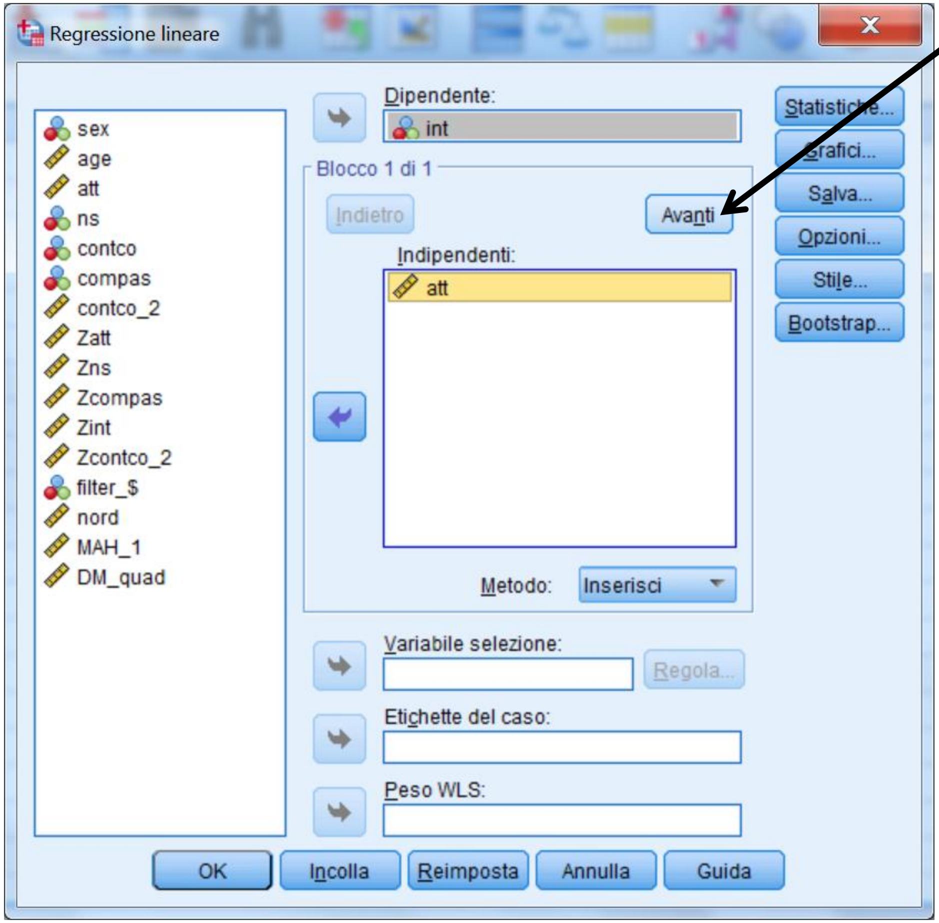
L'analisi procede attraverso "passi" sequenziali. Ogni VI è valutata per quanto aggiunge, nello spiegare la VD, rispetto a quanto è stato spiegato dalle VI inserite precedentemente. Partizione ordinata della varianza di VD spiegata dalle VI.

Contributo di una VI: può variare se la sua posizione nella gerarchia viene cambiata

Regressione gerarchica

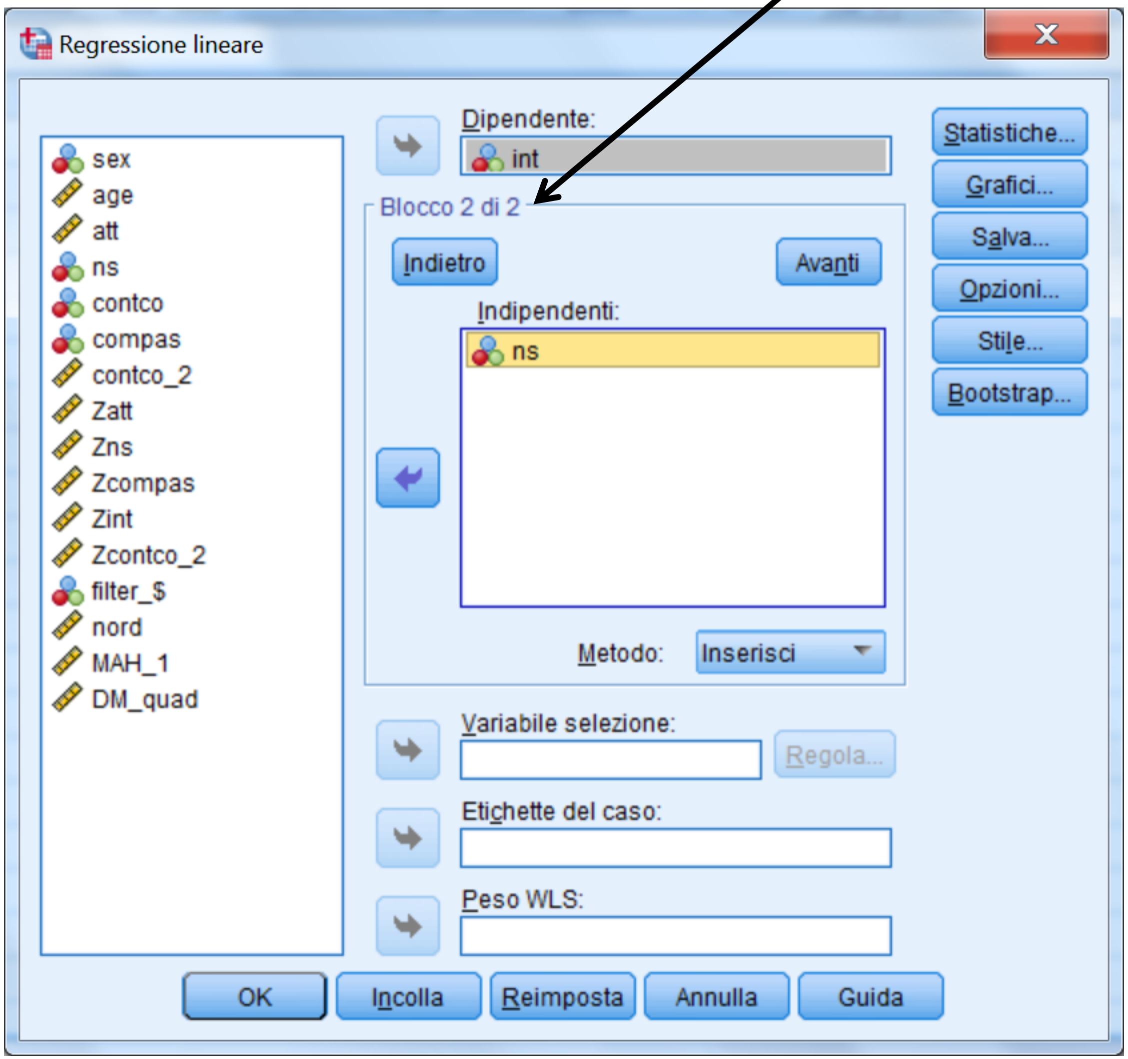
Selezionare la variabile dipendente ("int"). Quindi tutte le variabili indipendenti verranno inserite in blocchi separati, secondo un ordine consistente con il modello teorico che il ricercatore vuole esaminare.

Inserita la prima variabile ("att") cliccare sul pulsante "Avanti"



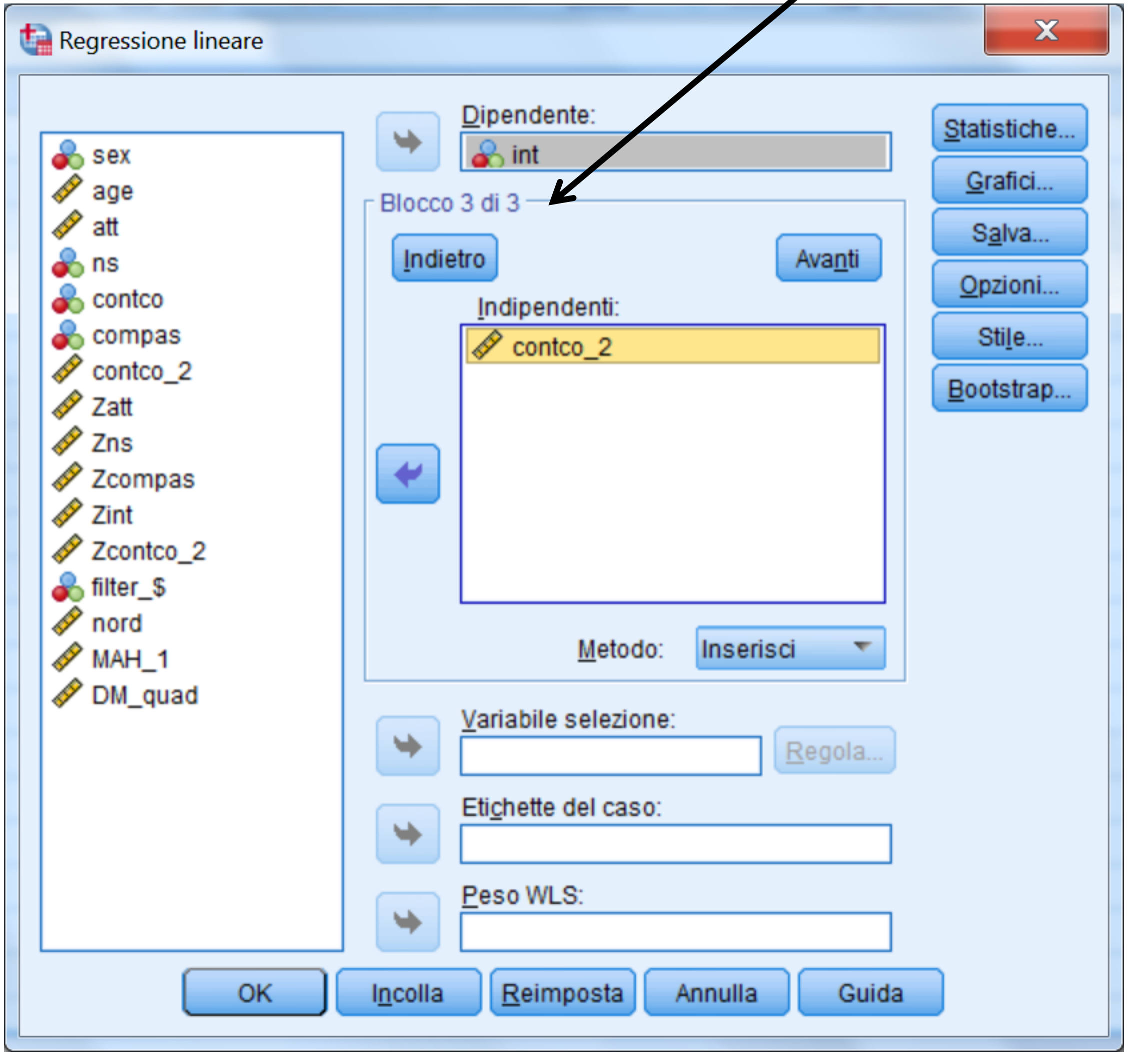
Regressione gerarchica

Inserire la seconda variabile nel "Blocco 2 di 2" ("ns") e di nuovo cliccare sul pulsante "Avanti"



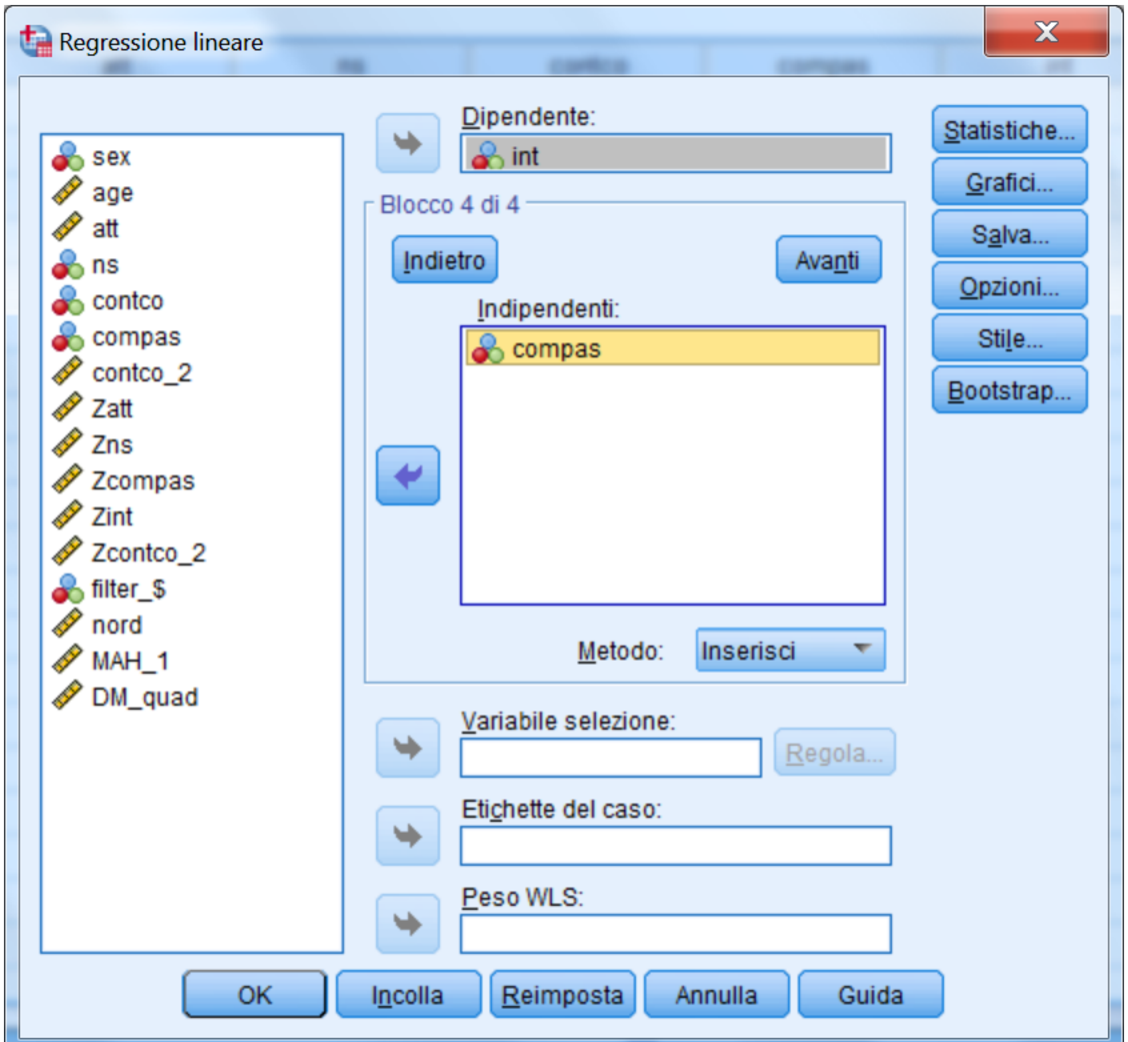
Regressione gerarchica

Inserire la terza variabile nel "Blocco 3 di 3" ("contco_2") e di nuovo cliccare sul pulsante "Avanti"



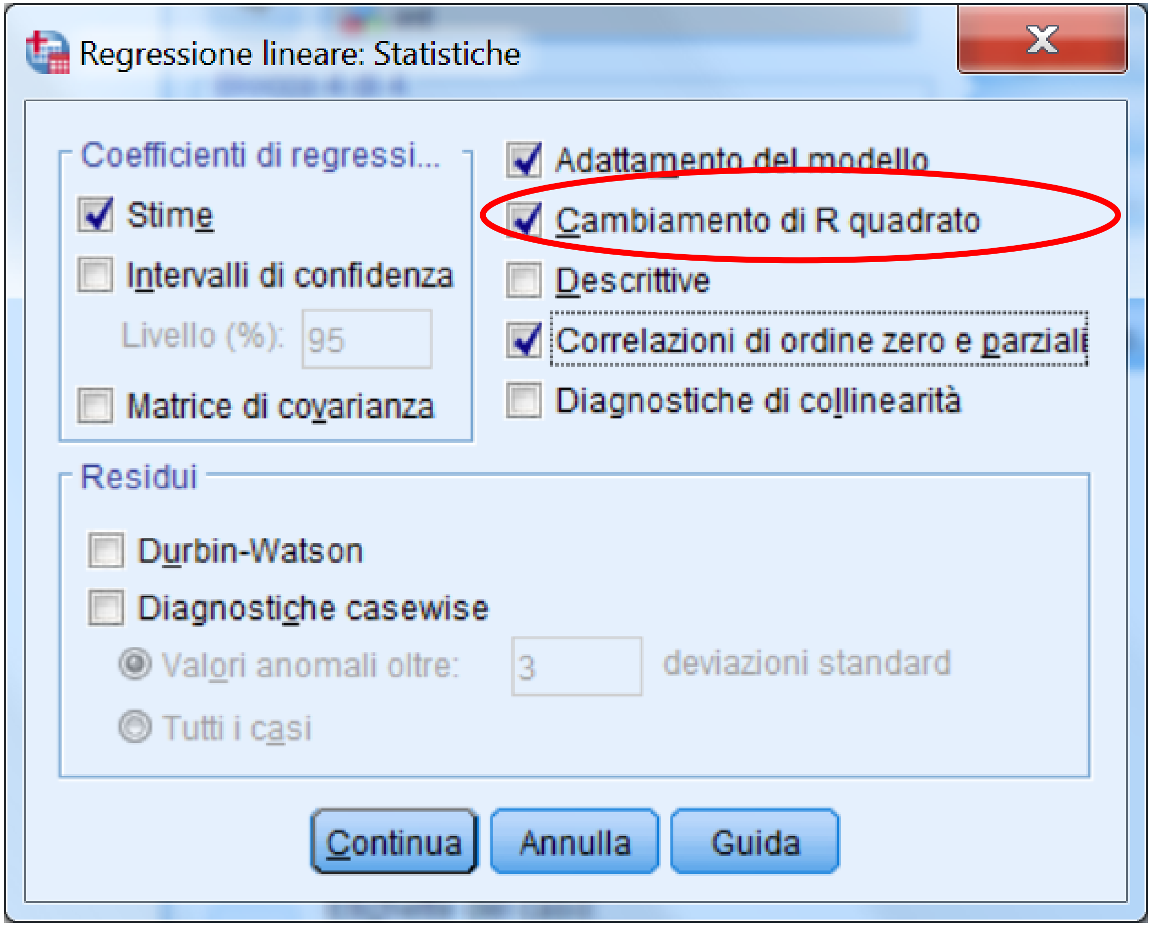
Regressione gerarchica

Inserire la quarta e ultima variabile nel "Blocco 4 di 4" ("compas"). In questi passaggi non cambiare mai il tipo di Metodo !!!



Regressione gerarchica

Nella finestra di dialogo "Statistiche" bisogna selezionare determinati parametri per ottenere nell'output le informazioni necessarie per interpretare e valutare la soluzione.



Regressione gerarchica

Il pannello iniziale riporta un riepilogo delle variabili inserite nel modello nei 4 passi della regressione: è diverso dal pannello analogo della regressione standard poiché ora non c'è più un unico blocco

Variabili immesse/rimosse[a]

Modello	Variabili immesse	Variabili rimosse	Metodo
1	att[b]	.	Inserisci
2	ns[b]	.	Inserisci
3	contco_2[b]	.	Inserisci
4	compas[b]	.	Inserisci

a. Variabile dipendente: int

b. Sono state immesse tutte le variabili richieste.

Regressione gerarchica

La varianza spiegata attraverso i diversi passi e il contributo unico delle variabili aggiunte ad ogni blocco si trova in questa tabella

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato adattato	Errore std. della stima	Statistiche delle modifiche: Modifica R-quadrato	Modifica F	gl1	gl2	Sign. Modifica F
1	,721[a]	,520	,517	1,766	,520	210,108	1	194	,000
2	,756[b]	,571	,567	1,674	,051	23,011	1	193	,000
3	,789[c]	,623	,617	1,574	,052	26,310	1	192	,000
4	,819[d]	,671	,664	1,474	,048	27,812	1	191	,000

a. Predittori: (costante), att

b. Predittori: (costante), att, ns

c. Predittori: (costante), att, ns, contco_2

d. Predittori: (costante), att, ns, contco_2, compas

Regressione gerarchica

La tabella dei coefficienti cambia a seconda del numero di predittori inseriti: l'ultima sezione (Modello 4) presenta risultati identici a quelli della regressione standard.

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sign.	Correlazioni		
		B	Errore std.	Beta			Ordine zero	Parziale	Parte
1	(Costante)	-3,886	,783		-4,960	,000			
	att	,261	,018	,721	14,495	,000	,721	,721	,721
2	(Costante)	-4,652	,759		-6,126	,000			
	att	,206	,021	,570	10,051	,000	,721	,586	,474
	ns	,393	,082	,272	4,797	,000	,589	,326	,226
3	(Costante)	-2,227	,856		-2,601	,010			
	att	,170	,021	,469	8,245	,000	,721	,511	,365
	ns	,358	,077	,248	4,627	,000	,589	,317	,205
	contco_2	-2,250	,439	-,255	-5,129	,000	-,544	-,347	-,227
4	(Costante)	-1,422	,816		-1,742	,083			
	att	,141	,020	,390	7,045	,000	,721	,454	,293
	ns	,273	,074	,189	3,676	,000	,589	,257	,153
	contco_2	-1,656	,426	-,188	-3,885	,000	-,544	-,271	-,161
	compas	,354	,067	,273	5,274	,000	,645	,357	,219

a. Variabile dipendente: int

Risultati della regressione gerarchica

Cambiamento di R e R^2 attraverso i riversi passi

Step	Variabile	R	R^2	R^2C	F	p
1	Atteggiamento	.72	.52	.52	210	.00
2	Norma Soggettiva	.76	.57	.05	23	.00
3	Senso di Controllo	.79	.62	.05	26	.00
4	Comport. Passato	.82	.67	.05	28	.00

sr^2: quantità di varianza aggiunta all' R^2 da ciascuna VI nel punto in cui la VI entra nell'equazione ("incremental sr^2" o cambiamento in R^2).

La somma degli sr^2 è uguale al valore di R^2.

Test statistico per valutare l'incremento nell'R^2

(Tabachnik & Fidell, 2007, p. 149)

$$F_{inc} = \frac{(R^2_{wi} - R^2_{wo})/m}{(1 - R^2)/df_{res}}$$

$R^2_{wi} = R^2$ ottenuto dall'inserimento della nuova variabile

$R^2_{wo} = R^2$ senza la nuova variabile

$m =$ numero di variabili nel nuovo blocco

$df_{res} = (N - k - 1)$

La regressione statistica

L'ordine di ingresso delle VI nell'equazione, e la decisione su quali VI vengono incluse o escluse dall'equazione di regressione sono determinati da criteri statistici

Limite: Differenze marginali rispetto a questi criteri possono influenzare in modo sostanziale l'importanza attribuita alle diverse VI

Tipi di regressione statistica

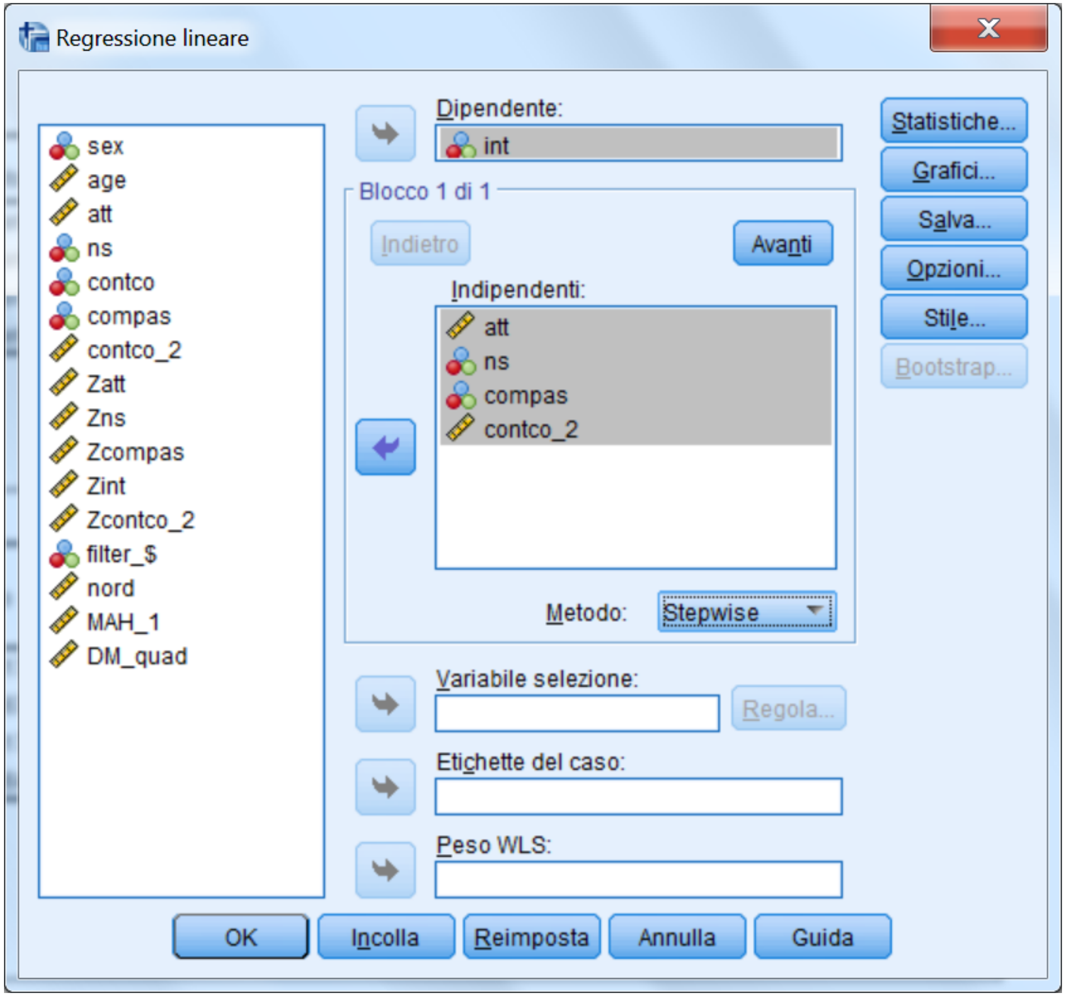
Regressione forward (in avanati): equazione inizialmente "vuota"; ad ogni step viene aggiunta la VI che presenta la correlazione più elevata con la VD. Se una VI entra in equazione, vi rimane

Regressione backward (all'indietro): l'equazione inizialmente comprende tutte le VI; ad ogni step viene eliminata la VI che non correla significativamente con la VD. Se una VI esce dall'equazione, non può più rientrarvi

Regressione stepwise: equazione inizialmente "vuota"; ad ogni step viene aggiunta la VI che correla di più con la VD. Le variabili che non forniscono più un contributo significativo vengono eliminate

Regressione Stepwise

Effettuare le stesse selezioni fatte per la regressione standard ma specificare "Stepwise" nel Metodo. Selezionare nelle Statistiche l'opzione per ottenere l'incremento dell'R^2.



Regressione Stepwise

Variabili immesse/rimosse[a]

Modello	Variabili immesse	Variabili rimosse	Metodo
1	att	.	Stepwise (criteri: Probabilità-di-F-da-inserire <= ,050, Probabilità-di-F-da-rimuovere >= ,100).
2	compas	.	Stepwise (criteri: Probabilità-di-F-da-inserire <= ,050, Probabilità-di-F-da-rimuovere >= ,100).
3	contco_2	.	Stepwise (criteri: Probabilità-di-F-da-inserire <= ,050, Probabilità-di-F-da-rimuovere >= ,100).
4	ns	.	Stepwise (criteri: Probabilità-di-F-da-inserire <= ,050, Probabilità-di-F-da-rimuovere >= ,100).

a. Variabile dipendente: int

Il pannello iniziale segnala quali variabili sono state inserite o rimosse durante la procedura Stepwise. Nella colonna metodo viene specificato quale è il metodo di inserimento/rimozione nell'equazione, e quali criteri determinano inserimento e rimozione

Regressione Stepwise

La varianza spiegata attraverso i diversi passi e il contributo unico delle variabili aggiunte ad ogni blocco si trova in questa tabella

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato adattato	Errore std. della stima	Statistiche delle modifiche				
					Modifica R-quadrato	Modifica F	gl1	gl2	Sign. Modifica F
1	,721[a]	,520	,517	1,766	,520	210,108	1	194	,000
2	,787[b]	,620	,616	1,575	,100	50,872	1	193	,000
3	,805[c]	,647	,642	1,522	,027	14,887	1	192	,000
4	,819[d]	,671	,664	1,474	,023	13,515	1	191	,000

a. Predittori: (costante), att

b. Predittori: (costante), att, compas

c. Predittori: (costante), att, compas, contco_2

d. Predittori: (costante), att, compas, contco_2, ns

La partizione della varianza è molto diversa da quella ottenibile nelle regressioni standard e gerarchica. L'ordine di importanza delle VI è quello dell'ultimo "modello"(ovvero passo): Atteggiamento, Comportamento Passato, Controllo, Norme Soggettive

Regressione Stepwise

La tabella dei coefficienti cambia a seconda dei predittori inseriti o rimossi: l'ultima sezione (Modello 4) presenta risultati identici a quelli della regressione standard e della gerarchica.

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sign.	Correlazioni		
		B	Errore std.	Beta			Ordine zero	Parziale	Parte
1	(Costante)	-3,886	,783		-4,960	,000			
	att	,261	,018	,721	14,495	,000	,721	,721	,721
2	(Costante)	-2,175	,739		-2,945	,004			
	att	,191	,019	,529	10,179	,000	,721	,591	,452
	compas	,479	,067	,370	7,132	,000	,645	,457	,316
3	(Costante)	-,657	,815		-,806	,421			
	att	,171	,019	,471	9,003	,000	,721	,545	,386
	compas	,407	,068	,315	6,027	,000	,645	,399	,258
	contco_2	-1,697	,440	-,193	-3,858	,000	-,544	-,268	-,165
4	(Costante)	-1,422	,816		-1,742	,083			
	att	,141	,020	,390	7,045	,000	,721	,454	,293
	compas	,354	,067	,273	5,274	,000	,645	,357	,219
	contco_2	-1,656	,426	-,188	-3,885	,000	-,544	-,271	-,161
	ns	,273	,074	,189	3,676	,000	,589	,257	,153

a. Variabile dipendente: int

Regressione Stepwise

Questa tabella è utile per capire quale variabile verrà inclusa nel prossimo passo. In questo caso è chiaro che tutte le variabili verranno incluse nell'analisi.

Variabili escluse[a]

Modello		Beta in	t	Sign.	Correlazione parziale	Statistiche di collinearità
						Tolleranza
1	ns	,272[b]	4,797	,000	,326	,691
	compas	,370[b]	7,132	,000	,457	,730
	contco_2	-,276[b]	-5,289	,000	-,356	,798
2	ns	,194[c]	3,647	,000	,255	,654
	contco_2	-,193[c]	-3,858	,000	-,268	,737
3	ns	,189[d]	3,676	,000	,257	,653

a. Variabile dipendente: int

b. Predittori nel modello: (costante), att

c. Predittori nel modello: (costante), att, compas

d. Predittori nel modello: (costante), att, compas, contco_2

Differenti metodi ⇒ Differenti risultati

Standard ⇒ 48% di varianza non attribuibile a nessuna variabile.

Gerarchica ⇒ Norma Soggettiva spiega più varianza del comportamento passato

Stepwise ⇒ Comportamento passato variabile più importante dopo l'atteggiamento

Regressione standard: strategia analitica migliore per studi esplorativi.
Regressione gerarchica: controllo maggiore sul processo della regressione; subordinata alla formulazione di ipotesi; studi confermativo.

Conclusioni

Tecnica flessibile per studiare la relazione di dipendenza tra variabili soprattutto nelle fasi esplorative di una ricerca.

Possibilità di definire modelli a priori (nel caso della regressione *gerarchica*): estensione anche a contesti di tipo confermativo.

Lo scopo è comunque quello di spiegare al meglio una variabile dipendente (y). E' una tecnica poco adatta a rendere ragione di modelli teorici complessi, in cui ci sono diverse variabili dipendente.

Conclusioni

Limiti legati alle assunzioni statistiche:

*** Assenza di errore nelle variabili: assai irrealistica.**

*** Problema della *multicollineratià:* spesso risolvibile all'interno del modello della regressione.**

*** Impossibile considerare <u>simultaneamente</u> più di una variabile dipendente alla volta nello stesso modello. Modelli complessi sono esaminabili solo scindendoli in tanti pezzi separati.**

*** Risultati soggetti ad interpretazioni assai differenti a seconda del metodo di regressione scelto (standard, gerarchica, statistica).**

Accertare le condizioni di applicabilità

Scegliere l'approccio più adeguato per gli scopi del ricercatore

ESERCIZIO 2:
REALIZZAZIONE DI UN MODELLO DI REGRESSIONE CON SPSS

Utilizzare i dati in formato testo nel file ES1.SAV, risultato dell'esercizio 1.

VARIABILI:
ATTEGGIAMENTO, NORME SOGGETTIVE, SENSO DI CONTROLLO, COMPORTAMENTO PASSATO, INTENZIONE.
LA VARIABILE DIPENDENTE E' "INTENZIONE"

1) Effettuare una regressione standard, calcolando la varianza unica spiegata da ogni variabile e la varianza comune

2) Effettuare una regressione gerarchica nella quale l'ordine di entrata della VI è il seguente: comportamento passato, norme soggettive, senso di controllo, atteggiamento