Le variabili binarie (dummy variables) nell'analisi di regressione

Giovanni Battista Flebus

Lezioni di Psicometria

Le variabili dicotomiche hanno due soli valori

Si possono applicare anche a variabili categoriali (scale nominali)

La presenza di un solo intervallo le trasforma in una vera scala a intervalli

La codifica numerica ammette qualsiasi scelta, ma questa è quella più conveniente

0 -->assenza di caratteristica

1 --> presenza di caratteristica

(altri valori numeri sono possibili, ma questi sono i più pratici)

Le variabili binarie nella regressione o indicatori binari

- Sono utilizzate soprattutto per le variabili categoriali
- Ogni categoria viene usata per creare un indicatore binario.
- Se k sono le categorie, servono k-1 indicatori binari
- Ma si possono usare anche per le variabili continue, per rilevare situazioni particolari, per esempio, $x > k$

Piccolo esempio: 10 adulti hanno comunicato la loro età al momento del matrimonio

A	donna	21
B	donna	22
C	donna	23
D	donna	24
E	donna	25
F	donna	29
M	uomo	23
N	uomo	26
P	uomo	27
Q	uomo	28

Report

età età matrimonio

SESSO	Media	N	Deviazione std.
1 uomo	26,00	4	2,160
2 donna	24,00	6	2,828
Totale	24,80	10	2,658

Costruiamo una variabile dicotomica FEM

- vale 1 se il soggetto è una donna
- Vale 0 se è un uomo.
- La categoria di riferimento – quella che non compare nella codifica - è pertanto quella degli uomini.

- Usando l'equazione di regressione con la variabile dicotomica FEM (uguale a 1 per le donne e 0 per gli uomini), possiamo predire l'età del campione, secondo il sesso

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sign.
		T	Errore std	Beta		
1	(Costante)	26,000	1,299		20,015	,000
	FEM	-2,000	1,677	-,389	-1,193	,267

a. Variabile dipendente: età età matrimonio

L'equazione è sempre uguale: -2 x (genere) +26

Valore predetto per il gruppo di riferimento

Cambiamento attribuibile a FEM =1

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sign.
		T	Errore std	Beta		
1	(Costante)	26,000	1,299		20,015	,000
	FEM	-2,000	1,677	-,389	-1,193	,267

a. Variabile dipendente: età età matrimonio

L'equazione è sempre uguale:

età = -2 x (genere) +26

Per una donna, la predizione è uguale a

-2 x 1 +26 = 24 media del gruppo di donne

Per un uomo, la predizione è uguale a

-2 X 0 +26 =26 media del gruppo di uomini

Usando la variabile dicotomica MAS (1= uomo), otteniamo questi risultati:

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sign.
		T	Errore std	Beta		
1	(Costante)	24,000	1,061		22,627	,000
	MAS	2,000	1,677	,389	1,193	,267

a. Variabile dipendente: età età matrimonio

Valore predetto per il gruppo di riferimento (le donne)

Cambiamento attribuibile a MAS =1

La categoria di riferimento

- È quella che non compare nella regressione
- Può essere scelta secondo l'agio di interpretazione e utilizzazione.
- E' fondamentale però ricordare quale è stata scelta per diventare il riferimento

Esempio con la codifica della scuola

scuola

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	1 CFP	128	20,2	20,2	20,2
	2 IPSIA	105	16,5	16,5	36,7
	3 ipscom	62	9,8	9,8	46,5
	4 ITC	120	18,9	18,9	65,4
	5 ITI	72	11,3	11,3	76,7
	6 Classico	31	4,9	4,9	81,6
	7 Scientifico	80	12,6	12,6	94,2
	8 Magistrali	37	5,8	5,8	100,0
	Totale	635	100,0	100,0	

Esempi di codifiche

	Variabili dicotomiche nuove			
Scuola consigliata	D_CFP	D_IPSIA	D_IPSC	Ecc ecc
1 CFP	1	0	0	
2 IPSIA	0	1	0	
3 ipscom	0	0	1	
4 ITC	0	0	0	
5 ITI	0	0	0	
6 Classico	0	0	0	
7 Scientifico	0	0	0	
8 Magistrali	0	0	0	

Sono possibili codifiche più generali

	Variabili dicotomiche nuove			
Scuola	Licei	Tecnici	Professionali	Altre
1 CFP	0	0	1	0
2 IPSIA	0	0	1	0
3 ipscom	0	0	1	0
4 ITC	0	1	0	0
5 ITI	0	1	0	0
6 Classico	1	0	0	0
7 Scientifico	1	0	0	0
8 Magistrali	0	0	0	1

Esempio con la variabile Età (file Aurisina)

età

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	13	21	3,3	3,3	3,3
	14	515	81,1	81,1	84,4
	15	87	13,7	13,7	98,1
	16	12	1,9	1,9	100,0
	Totale	635	100,0	100,0	

Variabile misurata

g4 vocabolario

età	Media	N	Deviazione std.
13	18,905	21	5,638
14	18,468	515	5,163
15	15,368	87	4,273
16	17,167	12	4,407
Totale	18,033	635	5,157

Categoria di riferimento

Usiamo la sintassi di SPSS per costruire le variabili binarie

Definiamo tre nuove variabili (i quattordicenni sono il gruppo di riferimento)

- compute tredici=0.
- *compute quindici=0.*
- *compute sedici=0.*

Istruzioni condizionali

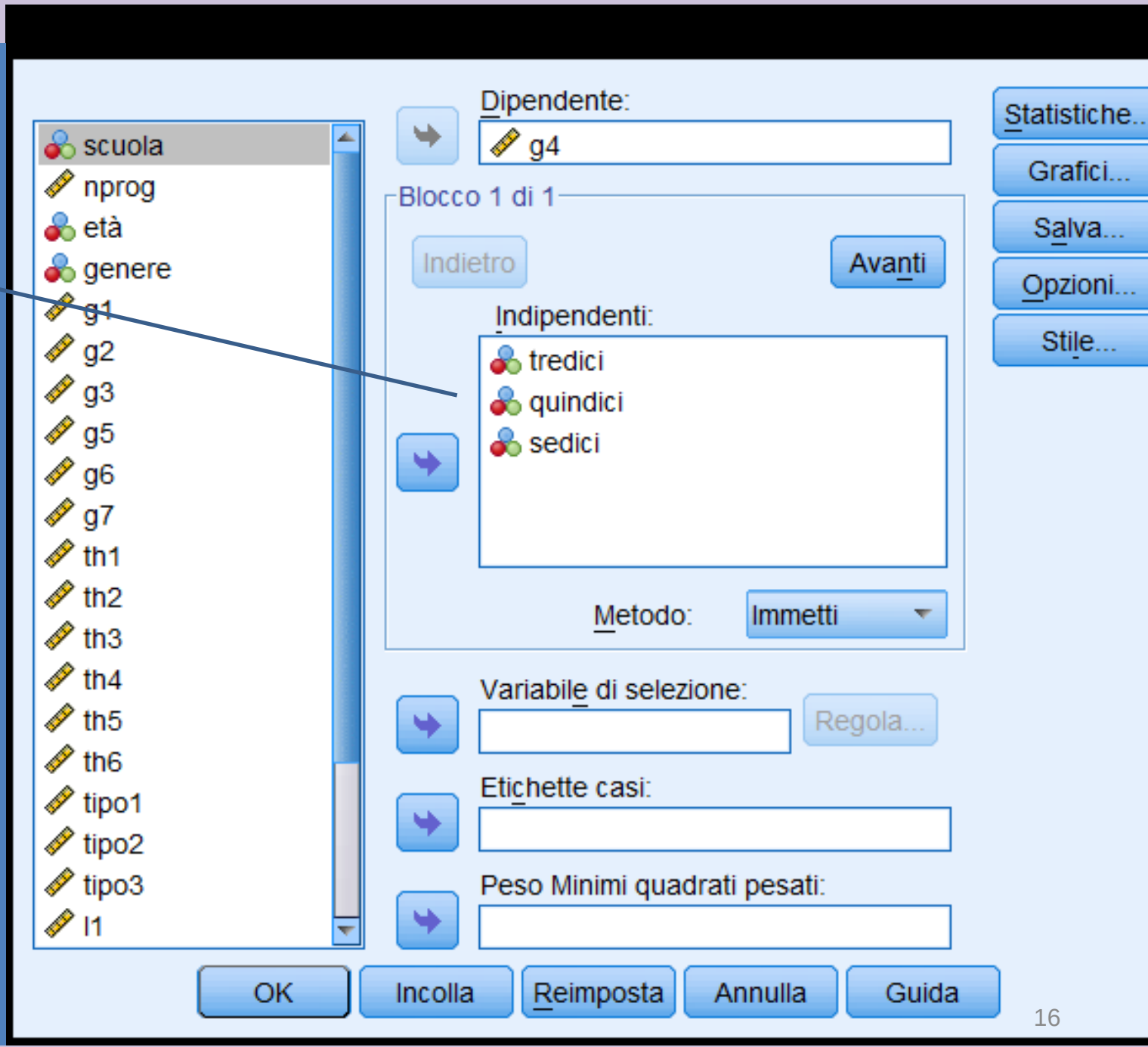
- if età eq 13 tredici = 1.
- if età eq 15 quindici = 1.
- if età eq 16 sedici = 1.

Oppure ricorriamo alla finestra del menu per eseguire le stesse operazioni.

E' possibile anche ricodificare la variabile Età in tre nuovo variabili

esempio: recode età (13=1)(14 15 16 =0) into TREDICI.

Abbiamo già imparato a selezionare la variabile dipendente e indipendente, ora aggiungiamo le tre variabili dicotomiche che abbiamo appena creato



Ecco l'output di SPSS

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	18,468	,223		82,919	,000
	tredici	,437	1,125	,015	,388	,698
	quindici	-3,100	,586	-,207	-5,292	,000
	sedici	-1,301	1,476	-,034	-,882	,378

a. Variabile dipendente: g4 vocabolario

Risultati con la regressione

Coefficienti[a]

Modello		Coefficienti non standardizzati	
		B	Errore std.
1	(Costante)	18,468	,223
	tredici	,437	1,125
	quindici	-3,100	,586
	sedici	-1,301	1,476

a. Variabile dipendente: g4 vocabolario

Repo

g4 vocabolario

età	Media
13	18,905
14	18,468
15	15,368
16	17,167
Totale	18,033

La costante delle regressione è uguale alla media della categoria di riferimento, ossia quando i tre indicatori sono tutti uguali a zero

Risultati con la regressione

18,468 +0,437 =18,905

Modello		Coefficienti non standardizzati B	Errore std.
1	(Costante)	18,468	,223
	tredici	,437	1,125
	quindici	-3,100	,586
	sedici	-1,301	1,476

a. Variabile dipendente: g4 vocabolario

Repo

g4 vocabolario

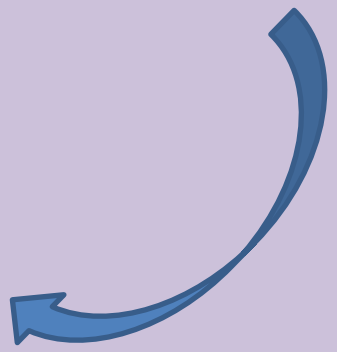
età	Media
13	18,905
14	18,468
15	15,368
16	17,167
Totale	18,033

Le medie degli altri gruppi sono il risultato della somma della costante e di ciascun coefficiente moltiplicativo

MEDIA DEL GRUPPO	costante moltiplicativa	valore della variabile	costante additiva	risultato
13 anni	,437	1,000	18,468	18,905
14 anni	0	nessuna	18,468	18,468
15 anni	-3,100	1,000	18,468	15,368
16 anni	-1,301	1,000	18,468	17,167

g4 vocabolario

età	Media
13	18,905
14	18,468
15	15,368
16	17,167
Totale	18,033



Uso degli indicatori nella regressione

Il ricorso agli indicatori dicotomici nella regressione soddisfa diverse esigenze:

- (1) predizione con una variabile categoriale o realmente dicotomica, come il genere
- (2) controllo o eliminazione di alcuni effetti privi di interesse (o che si vogliono controllare) in un'equazione di regressione multipla
- (3) esame delle interazioni fra indicatori diversi