REGRESSIONE LINEARE

In questo documento presentiamo alcune opzioni analitiche della procedura della regressione lineare di SPSS che non sono state incluse nel testo pubblicato. Si tratta di opzioni che, pur non essendo utilizzate correntemente per effettuare la regressione multipla lineare, consentono di completare lo spettro delle possibilità di analisi offerte da SPSS. In particolare, il documento completa la descrizione della procedura di analisi fattoriale con in seguenti paragrafi:

1. Altri metodi di selezione dei predittori
2. Selezione di blocchi di predittori
3. L'analisi di regressione con il linguaggio SINTASSI
4. Opzioni di analisi ottenibili solo attraverso SINTASSI
    - Statistiche di selezione
    - Errore standard per i coefficienti standardizzati
    - Opzioni relative alla Tolleranza
    - Metodo TEST per l'inserimento delle variabili indipendenti
    - Analisi diretta di matrici
    - Differenze tra i coefficienti di regressione

1. Altri metodi di selezione dei predittori

Ci sono tre ulteriori opzioni selezionabili nel box "Metodo" nella finestra di dialogo principale della procedura di regressione lineare di SPSS. Due sono metodi statistici di selezione delle variabili indipendenti: il Metodo "*Avanti*" (Forward) e quello "*Indietro*" (Backward). Anche in questi casi le variabili indipendenti vengono inserite in un solo blocco. Il metodo "*Avanti*" inserisce le variabili una alla volta, in base al criterio di entrata (probabilità della F minore di .05). Se una variabile entra in equazione, non vi esce più (anche se il suo contributo successivamente diventa non significativo

per l'entrata di altre variabili). Il metodo "*Indietro*" inserisce tutte le variabili nel blocco in un singolo passo e poi le rimuove una alla volta in base al criterio di rimozione (probabilità della F maggiore di .10). Se una variabile esce dall'equazione, non può più rientrarvi (anche se il suo contributo successivamente diventa significativo grazie all'uscita di altre variabili).

L'altro metodo di selezione delle variabili è "*Rimozione*" ed è di tipo gerarchico. Inizialmente bisogna inserire tutte le variabili nell'equazione, e scegliere come metodo di analisi "Per blocchi". Poi bisogna selezionare (con un clic del mouse su "Successivo") i blocchi che consentiranno di rimuovere le variabili. Lo scopo è di vedere quanto diminuisce il coefficiente R-quadrato avendo rimosso una data variabile: se la rimozione della variabile non diminuisce significativamente l'R-quadrato, il contributo della variabile è inutile. Consideriamo il modello gerarchico descritto nel testo da pag. 45 a pag. 53; l'ordine di entrata delle variabili in equazione era il seguente: COMPAS, NS, CONTCO, ATT. Per esaminare lo stesso modello con il metodo "Rimozione" dovremmo selezionare:

a) le quattro variabili nel primo blocco lasciando il metodo "Per blocchi";

b) la variabile ATT nel secondo blocco e scegliere il metodo "Rimozione";

c) la variabile CONTCO nel terzo blocco e scegliere il metodo "Rimozione";

d) la variabile NS nel quarto blocco e scegliere il metodo "Rimozione";

e) la variabile COMPAS nel quinto ed ultimo blocco e scegliere il metodo "Rimozione" (vedi figura 1).

La tabella 1 riassume la storia delle variabili inserite e rimosse nei diversi passi/modelli: nel modello 1 tutte le variabili sono in equazione, poi una per volta vengono eliminate fino al modello 5 nel quale nessuna variabile è in equazione. La tabella 2 presenta i risultati relativi al decremento dell'R-quadrato nei 5 modelli. E' facile verificare confrontando questa tabella con la 1.21 del testo che i coefficienti nelle due matrici sono esattamente gli stessi ma in ordine di comparsa inverso. Ad esempio, nella colonna R-quadrato .640 è il primo coefficiente nella tabella 2 ma nella 1.21 è l'ultimo, .mentre 389 è l'ultimo (prima dello 0) ma nella 1.21 è il primo. Ovviamente nella colonna

Variazione di R-quadrato della tabella 2 i valori dopo .640 sono tutti negativi, perché ora indicano una diminuzione dell'R-quadrato mentre nella 1.21 indicavano un aumento del coefficiente. Ad esempio rimuovendo la variabile ATT il coefficiente diminuisce significativamente di .085 (ovvero si perde l'8.5% di varianza spiegata), passando così da .640 a .555. Rimuovendo la variabile CONTCO il coefficiente diminuisce significativamente di .036, passando così da .555 a .519. Rimuovendo la variabile NS il coefficiente diminuisce significativamente di .130, passando così da .519 a .389. Infine, rimuovendo l'unica variabile rimasta, COMPAS, il coefficiente diminuisce significativamente di .389, diventando ovviamente uguale a 0 visto che nessuna variabile è più presente in equazione.

Figura 1. Finestra di dialogo della regressione lineare con il metodo di selezione dei predittori "Rimozione"

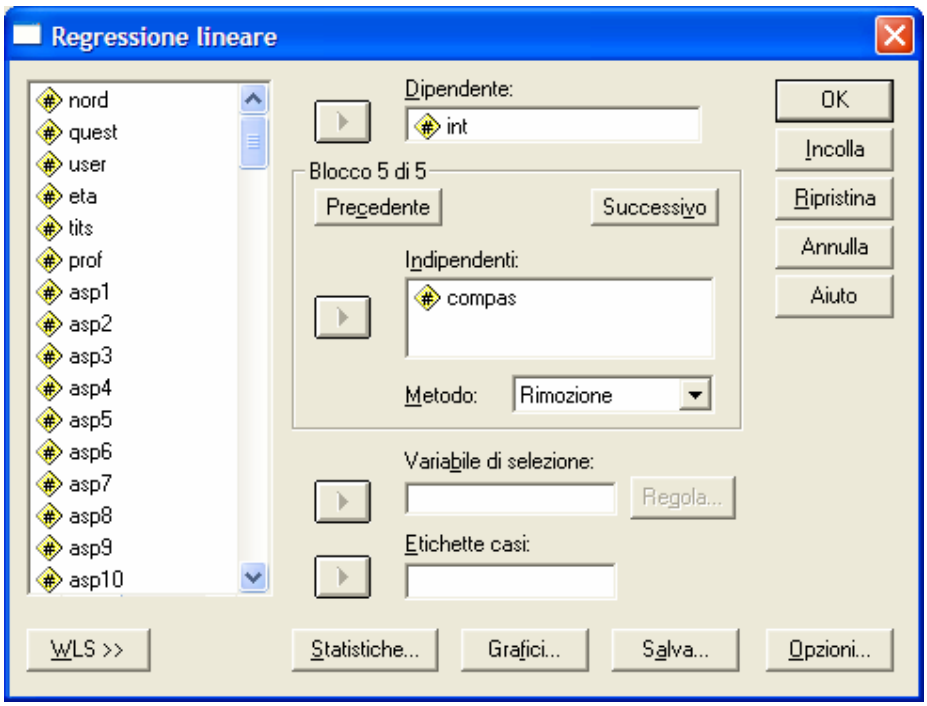


Tabella 1. Elenco delle variabili inserite e rimosse nella regressione lineare con il metodo di selezione dei predittori "Rimozione"

Variabili inserite/rimosse[c]

Modello	Variabili inserite	Variabili rimosse	Metodo
1	ATT, CONTCO, COMPAS, NS[a]	.	Per blocchi
2	.[a]	ATT[b]	Rimuovi
3	.[a]	CONTCO[b]	Rimuovi
4	.[a]	NS[b]	Rimuovi
5	.[a]	COMPAS[b]	Rimuovi

a. Tutte le variabili richieste sono state inserite

b. Tutte le variabili richieste sono state rimosse

c. Variabile dipendente: INT

Tabella 2. Riepilogo dei risultati nella regressione lineare con il metodo di selezione dei predittori "Rimozione"

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima	Variazione dell'adattamento				
					Variazione di R-quadrato	Variazione di F	df1	df2	Sig. variazione di F
1	,800[a]	,640	,632	1,41298	,640	79,417	4	179	,000
2	,745[b]	,555	,547	1,56649	-,085	42,237	1	179	,000
3	,720[c]	,519	,514	1,62344	-,036	14,398	1	180	,000
4	,624[d]	,389	,386	1,82458	-,130	48,893	1	181	,000
5	,000[e]	,000	,000	2,32779	-,389	115,861	1	182	,000

a. Stimatori: (Costante), ATT, CONTCO, COMPAS, NS

b. Stimatori: (Costante), CONTCO, COMPAS, NS

c. Stimatori: (Costante), COMPAS, NS

d. Stimatori: (Costante), COMPAS

e. Stimatore: (costante)

2. Selezione di blocchi di predittori

Utilizzando il pulsante "Successivo" nella finestra di dialogo principale è possibile selezionare non solo una variabile per blocco ma più variabili, e trattare le variabili inserire nei diversi blocchi con

uno qualsiasi dei metodi previsti da SPSS. Proviamo ad esempio ad esaminare un modello nel quale vogliano tenere sotto controllo il comportamento passato e vedere quale è il valore aggiunto complessivamente dalle altre variabili senza voler specificare però un ordine specifico per le altre variabili in analisi. La figura 2 presenta le opzioni nella finestra di dialogo principale, che possono essere utilizzate: nel primo blocco inseriamo la variabile COMPAS e nel secondo le altre 3 variabili lasciando l'opzione di default per il metodo. Il risultato di questa analisi è riportato nella tabella 3. Le tre variabili aggiungono alla varianza spiegata da COMPAS il 25.1% che risulta significativo. Poiché non abbiamo specificato un ordine di entrata delle variabili in equazione, questo 25.1% non viene diviso tra le 3 variabili inserite nel secondo blocco, quindi il loro contributo è considerato solo complessivamente. Per calcolare il contributo unico delle variabili nel secondo blocco è però possibile considerare la tabella relativa ai coefficienti che è uguale a quella ottenuta nella regressione standard (vedi tabella 1.18 del testo) e che per questo non riportiamo. Utilizzando i coefficienti di correlazione semiparziali al quadrato possiamo concludere che il 25.1% spiegato complessivamente da ATT, NS e CONTCO è divisibile in questo modo: 8.5% attribuibile solo alla variabile ATT, 3.6% attribuibile solo alla variabile NS, 1.7% attribuibile solo alla variabile CONTCO, e 11.3% (=25.1- 8.5 - 3.6 – 1.7) non attribuibile a nessuna delle 3 variabili in modo unico, ma che rappresenta la percentuale di varianza *comune* spiegata simultaneamente dalle 3 variabili prese in combinazione. La figura 3 fornisce una rappresentazione grafica di questo risultato utilizzando i diagrammi di Venn.

Figura 2. Finestra di dialogo per la regressione gerarchica a blocchi

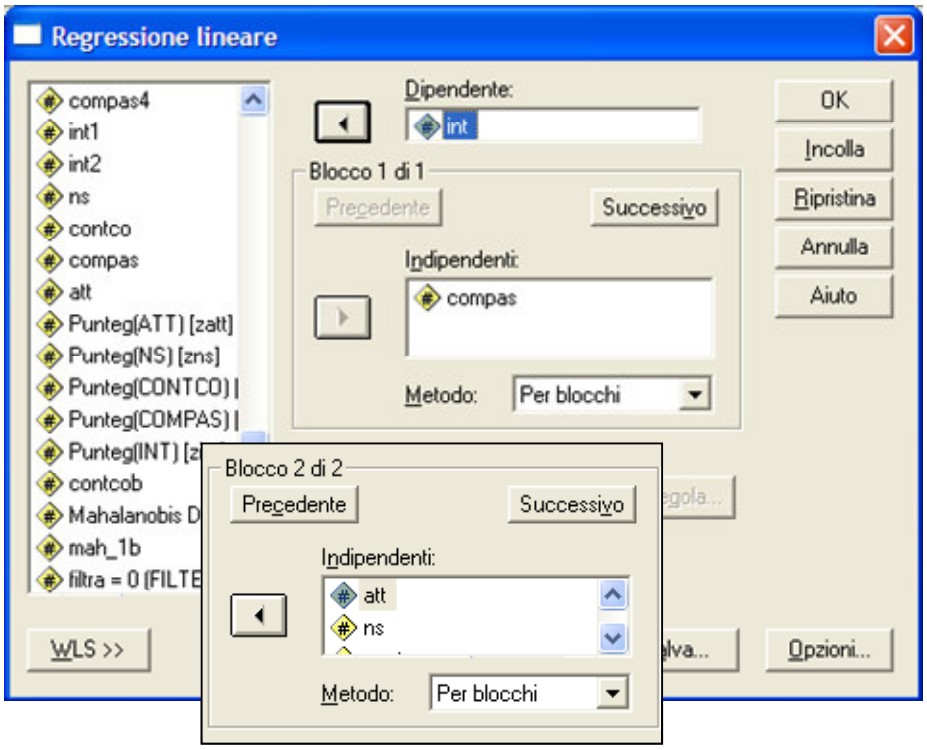


Tabella 3. Riepilogo del modello della regressione gerarchica a blocchi

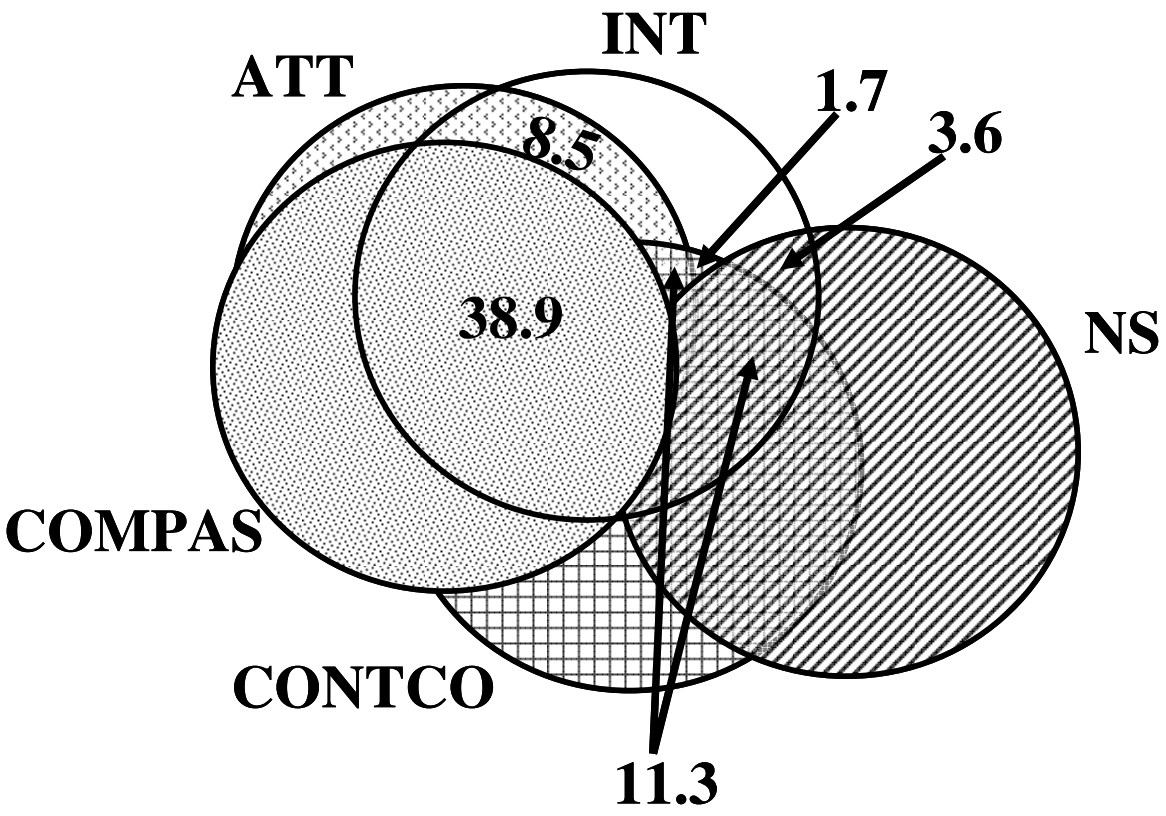
Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima	Variazione dell'adattamento				
					Variazione di R-quadrato	Variazione di F	df1	df2	Sig. variazione di F
1	,624[a]	,389	,386	1,82458	,389	115,861	1	182	,000
2	,800[b]	,640	,632	1,41298	,251	41,492	3	179	,000

a. Stimatori: (Costante), COMPAS

b. Stimatori: (Costante), COMPAS, CONTCO, NS, ATT

Figura 3. Scomposizione della varianza nella regressione gerarchica a blocchi



3. L'analisi di regressione con il linguaggio SINTASSI

Di seguito presentiamo i programmi SINTASSI per le analisi descritte nel testo.

```
**** Regressione Standard.

REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT inte
  /METHOD=ENTER atte ns contcomp cpa
  /RESIDUALS DURBIN .
```

Illustriamo brevemente i comandi specificati nella procedura REGRESSION.

/DESCRIPTIVES MEAN STDDEV CORR SIG N consente di ottenere in output alcune statistiche descrittive univariate e bivariate calcolate sul campione.

/MISSING definisce il metodo utilizzato per il trattamento dei valori mancanti (in questo caso è lasciata l'opzione di default LISTWISE).

/STATISTIC specifica quali elementi verranno inseriti nell'output: in questo caso abbiamo lasciato le oltre alle opzioni di default (COEFF = coefficienti di regressione, OUTS = statistiche per le variabili escluse dall'equazione, R = indici di bontà dell'adattamento, ANOVA = significatività degli indici di bontà) abbiamo chiesto anche gli indici di collinearità (COLLIN) e di tolleranza (TOL), e i coefficienti di ordine zero, parziali e semiparziali (ZPP).

/CRITERIA serve a specificare i criteri che verranno utilizzati per inserire o escludere le variabili nella regressione statistica.

/NOORIGIN indica che verrà inclusa anche l'intercetta tra i parametri della regressione.

/DEPENDENT indica che la variabile dipendente sarà "int".

/METHOD=ENTER indica che verranno inserite in un unico blocco le variabili indipendenti che seguono.

/RESIDUALS DURBIN serve ad ottenere in output il test di Durbin-Watson per l'autocorrelazione dei residui.

```
**** Grafici per la verifica delle assunzioni e calcolo
della distanza di Mahalanobis.

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT int
  /METHOD=ENTER att ns contco compas
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZPRED ,*ZRESID )
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) OUTLIERS(MAHAL)
  /SAVE ZRESID MAHAL.
```

In grassetto sono indicati i comandi necessari per richiedere i risultati relativi alla verifica delle assunzioni. Il risultato di queste opzioni è descritto nel testo.

```
**** tabella del riepilogo dei casi rispetto alla distanza
di Mahalanobis.

SUMMARIZE
  /TABLES=nord mah_1
  /FORMAT=VALIDLIST NOCASENUM TOTAL LIMIT=20
  /TITLE='Riepiloghi dei casi'
  /MISSING=VARIABLE
  /CELLS=COUNT .


**** Plot dei quantili rispetto alla distanza di Mahalanobis
per la normalità multivariata.

PPLOT
  /VARIABLES=mah_1
  /NOLOG
  /NOSTANDARDIZE
  /TYPE=Q-Q
  /FRACTION=BLOM
  /TIES=MEAN
  /DIST=CHI(5).
```

```
***** Regressione Gerarchica.

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT int
  /METHOD=ENTER compas
  /METHOD=ENTER ns
  /METHOD=ENTER contco
  /METHOD=ENTER att  .
```

Rispetto alla regressione standard viene richiesto il calcolo del cambiamento del coefficiente R-quadrato e la significatività di tale cambiamento (CHANGE). Inoltre, il comando

/METHOD=ENTER viene ripetuto tante volte quante sono le variabili da inserire nell'analisi gerarchica. Ogni comando infatti corrisponde ad un blocco differente nella gerarchia, e l'ordine di entrata dei predittori nell'equazione di regressione è quello definito dalla sequenza dei comandi **/METHOD**.

```
***** Regressione Stepwise (per passi).

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT int
  /METHOD=STEPWISE att ns contco compas  .
```

Rispetto alla regressione standard, nel comando **/METHOD** viene specificata l'opzione **STEPWISE**, seguita dall'elenco delle variabili che si vogliono selezionare come predittori tramite il metodo statistico.

4. Opzioni di analisi ottenibili solo attraverso SINTASSI

Alcune funzionalità e opzioni di analisi della procedura regressione lineare di SPSS non sono accessibili tramite il menù, ma richiedono una programmazione nel linguaggio SINTASSI. Illustreremo qui alcune applicazioni di questa programmazione che riteniamo particolarmente interessanti. Esse rappresentano un'integrazione rispetto a quanto ottenibile utilizzando i menù e le finestre di dialogo fornite dal programma.

4.1. Statistiche di selezione

Si tratta di una opzione che permette di calcolare il criterio di informazione di Akaike (Akaike Information Kriterion, AIK), il criterio di stima di Ameniya (PC), il criterio di stima di Mallow (C*p*), e il criterio bayesiano di Schwartz (Judge et al., 1980). Queste statistiche sono visualizzate nella tabella riassuntiva del modello. Si tratta di alcune statistiche supplementari che consentono di orientare il ricercatore nella scelta del modello. Le linee di programma SINTASSI per ottenere in output questi indici sono le seguenti:

```
**** Statistiche di selezione.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA SELECTION
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT int
  /METHOD=ENTER ns contco compas att  .
```

In queste linee di programma, oltre alle opzioni di default abbiamo chiesto anche le statistiche di selezione (**/STATISTICS .. SELECTION**). Se si confrontano modelli alternativi, il modello migliore è quello che presenta criterio di informazione, criterio di stima di Ameniya e criterio bayesiano più bassi, e criterio di stima di Mallow più elevato. La tabella 4 presenta i riepiloghi del modello con i risultati relativi alle statistiche di selezione (nelle colonne "Criteri di Selezione") con e senza la variabile "ns" nell'equazione di regressione. E' facile verificare che quando la variabile "ns" viene esclusa si assiste ad un generale peggioramento dei 4 Criteri di selezione considerati.

Tabella 4. Riepilogo del modello con i risultati relativi ai Criteri di Selezione

Soluzione con la variabile "ns"

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima	Criteri di selezione			
					Criterio di informazione di Akaike	Criterio di stima di Amemiya	Criterio di stima di Mallow	Criterio bayesiano di Schwartz
1	,812[a]	,659	,651	1,5016	166,743	,359	5,000	183,209

a. Stimatori: (Costante), ATT, CONTCO, NS, COMPAS

Soluzione senza la variabile "ns"

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima	Criteri di selezione			
					Criterio di informazione di Akaike	Criterio di stima di Amemiya	Criterio di stima di Mallow	Criterio bayesiano di Schwartz
1	,796[a]	,633	,628	1,5522	178,944	,382	4,000	192,117

a. Stimatori: (Costante), ATT, CONTCO, COMPAS

4.2. Errore standard per i coefficienti standardizzati

E' possibile calcolare l'errore standard per i parametri standardizzati. Si tratta dell'*Approximate standard error of the standardized regression coefficients.* (Meyer & Younger, 1976). Questa statistica viene riportata nella tabella dei coefficienti (vedi tabella 5) e si ottiene con le seguenti linee di programma SINTASSI:

```
**** Errore standard per i coefficienti standardizzati.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA SES
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT int
  /METHOD=ENTER ns contco compas att  .
```

Tabella 5. Errori standard per i coefficienti standardizzati

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati		t	Sig.
		B	Errore std.	Beta	Errore std.		
1	(Costante)	-3,999	,962			-4,157	,000
	NS	,313	,075	,230	,055	4,201	,000
	CONTCO	,267	,091	,146	,050	2,930	,004
	COMPAS	,336	,066	,280	,055	5,123	,000
	ATT	,132	,020	,378	,058	6,499	,000

a. Variabile dipendente: INT

Di fatto i risultati che è possibile ottenere dividendo i coefficienti standardizzati per il loro errore standard sono sovrapponibili a quelli riportati nella tabella 5 sotto la colonna "t".

4.3. Opzioni relative alla Tolleranza

E' possibile cambiare i valori di default di Tolleranza: ricordiamo che tutte le variabili che presentano un valore inferiore a quello stabilito di default (che è pari a 0.0001) o a quello eventualmente stabilito dall'utente vengono escluse dall'equazione di regressione. Ad esempio se l'utente vuole che siano considerate solo quelle variabili indipendenti che abbiano una tolleranza maggiore di .05 (ovvero, quelle variabili che presentano almeno il 5% di varianza unica non comune con le altre variabili indipendenti) le linee di programma SINTASSI per cambiare il valore di default da .0001 a .05 sono le seguenti:

```
**** Cambiare i valori della Tolleranza.
  /CRITERIA= TOLERANCE(0.05)
```

4.4. Metodo TEST per l'inserimento delle variabili indipendenti

Nel metodo TEST viene esaminato il cambiamento nell'R-quadrato e la sua significatività per insiemi (*sets*) differenti di variabili indipendenti. Questo metodo inizialmente esamina un modello con tutte le variabili specificate in METHOD tramite l'opzione TEST. Quindi rimuove a turno ciascuno degli insiemi specificati e mostra le statistiche che vengono richieste. Supponiamo ad esempio di confrontare la bontà dei modelli definibili considerando soltanto tre delle quattro variabili considerate nel nostro esempio. Le linee di programma SINTASSI che implementano questa analisi sono le seguenti:

```
**** Metodo TEST.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
   /NOORIGIN
  /DEPENDENT int
  /METHOD= TEST (ns) (att) (compas) (contco ).
```

Il programma calcolerà un R-quadrato considerando prima tutti i regressori. Quindi esaminerà la riduzione dell'R-quadrato dovuta alla rimozione della variabile "ns" (primo insieme), poi la riduzione dovuta alla rimozione della variabile "att" (secondo insieme), e quindi la riduzione dovuta alla rimozione della variabile "compas" (terzo insieme), e infine la riduzione dovuta alla rimozione della variabile "contoc" (quarto insieme). La tabella 6 presenta il risultato di questa analisi. Il primo insieme di variabili porta ad una riduzione dell'R-quadrato di .025, il secondo insieme porta ad una riduzione di .07, il terzo i insieme ad una riduzione di .055, il quarto infine ad una riduzione di .04. Sebbene tutte le riduzioni risultino statisticamente significative, è possibile sottolineare che l'insieme in cui non risulta compresa la variabile "ns" è quello cui è associata una riduzione minore dell'R-quadrato, mentre l'insieme in cui è la variabile "atte" ad essere esclusa porta alla riduzione più elevata.

Tabella 6. Risultato dell'analisi effettuata con il metodo TEST per l'inserimento delle variabili indipendenti

ANOVA[c]

Modello			Somma dei quadrati	df	Media dei quadrati	F	Sig.	Variazione di R-quadrato
1	Test sottoinsiemi	NS	32,358	1	32,358	14,350	,000[a]	,025
		ATT	89,697	1	89,697	39,779	,000[a]	,070
		COMPAS	70,426	1	70,426	31,233	,000[a]	,055
		CONTCO	51,426	1	51,426	22,807	,000[a]	,040
	Regressione		843,647	4	210,912	93,535	,000[b]	
	Residuo		437,448	194	2,255			
	Totale		1281,095	198				

a. Test effettuato sul modello completo.

b. Stimatori nel modello completo: (Costante), CONTCO, NS, COMPAS, ATT.

c. Variabile dipendente: INT

4.5. Analisi diretta di matrici

E' possibile leggere e scrivere file in formato SPSS che contengono una matrice tramite il comando **MATRIX**. Questo tipo di file vengono chiamati "*matrix data files*". Le linee di comandi SINTASSI che seguono consentono di effettuare l'analisi di regressione direttamente su una matrice di correlazioni opportunamente salvata come matrix data file SPSS.

```
**** Analisi diretta di una matrice di correlazioni.

REGRESSION MATRIX=IN(*)
   /VARIABLES=inte atte ns contcomp cpa
  /DEPENDENT=INTE
 /METHOD= ENTER atte ns contcomp cpa  .
```

Il comando MATRIX deve essere sempre specificato per primo E' possibile specificate solo una opzione IN e una opzione OUT nell'ambito di un unico comando MATRIX. Mentre l'opzione **IN** consente di leggere matrici, l'opzione **OUT (nome del file)** consente di scriverle su un file esterno.

Le opzioni IN e OUT nel comando MATRIX possono essere specificare in qualsiasi ordine. Nella procedura REGRESSION può essere utilizzata una matrice di correlazioni come quella presentata nella figura 25. **MATRIX=IN** non può essere utilizzato se non è stato già caricato un file SPSS. Per leggere una matrice in un file esterno all'inizio di una sessione SPSS, bisogna prima caricare il file, e poi specificare IN(*) nel comando MATRIX.

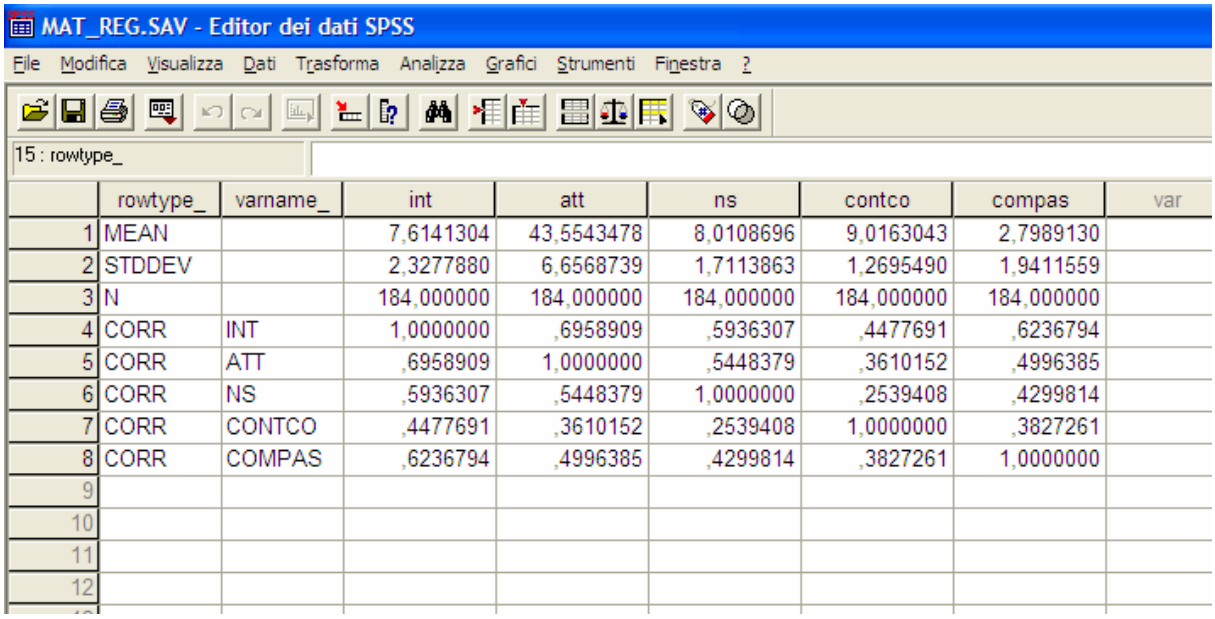
Nel nostro esempio abbiamo chiesto al programma di analizzare una matrice che è stata già caricata nel programma (quindi rappresenta il nostro file attivo) quindi abbiamo semplicemente scritto un asterisco nello spazio riservato per indicare il nome e il percorso del file. In questo caso ovviamente il file attivo è costituito da una matrice e non da dati nel solito formato Casi X Variabili. Se si legge una matrice da un file esterno questa non rimpiazza il file attivo in quel momento.

Caratteristiche di un "matrix data file"

La figura 4 presenta la matrice di correlazione analizzata tramite i comandi specificati nel quadro precedente. Una matrice di correlazioni del tipo leggibile da SPSS deve contenere due variabili "speciali" che sono create da SPSS nel momento in cui il file viene salvato in formato matix. **ROWTYPE_** è una variabile di tipo stringa che serve per codificare il tipo di parametri presenti nelle righe e nelle colonne della matrice: MEAN indica che la riga corrispondente conterrà le medie delle variabili, STDDEV indica che la riga corrispondente conterrà le deviazioni standard delle variabili, CORR indica che la riga corrispondente conterrà dei coefficienti di correlazione di Pearson, N indica che la riga corrispondente conterrà il numero di soggetti sui quali sono stati calcolati i coefficienti di correlazione. **VARNAME_** è una variabile di tipo stringa che viene utilizzata per indicare il nome delle variabili incluse nella matrice. Per poter effettuare correttamente un'analisi basta che nella matrice ci siano i coefficienti di correlazione: l'analisi infatti viene effettuata anche se nella matrice non sono presenti le medie, le deviazioni standard e il numero di casi. Ulteriori opzioni relative ai matrix data file vengono introdotte nel documento

sull'analisi fattoriale presente su questo sito, nel quale viene illustrato anche come costruire un file matriciale utilizzabile da SPSS.

Figura 4. Matrix data file per la regressione tramite analisi diretta di una matrice



	rowtype_	varname_	int	att	ns	contco	compas	var
1	MEAN		7,6141304	43,5543478	8,0108696	9,0163043	2,7989130	
2	STDDEV		2,3277880	6,6568739	1,7113863	1,2695490	1,9411559	
3	N		184,000000	184,000000	184,000000	184,000000	184,000000	
4	CORR	INT	1,0000000	,6958909	,5936307	,4477691	,6236794	
5	CORR	ATT	,6958909	1,0000000	,5448379	,3610152	,4996385	
6	CORR	NS	,5936307	,5448379	1,0000000	,2539408	,4299814	
7	CORR	CONTCO	,4477691	,3610152	,2539408	1,0000000	,3827261	
8	CORR	COMPAS	,6236794	,4996385	,4299814	,3827261	1,0000000	
9								
10								
11								
12								

4.6. Differenze tra i coefficienti di regressione

Differenze tra due coefficienti di regressione in due campioni indipendenti

Spesso un ricercatore è interessato a confrontare i risultati ottenuti considerando uno stesso modello esaminato su campioni diversi. Il problema è soprattutto quello di confrontare se i coefficienti di regressione ottenuti su più campioni possono essere considerati uguali (ovvero, non significativamente diversi). Il confronto può essere effettuato sia per quanto riguarda i coefficienti non standardizzati sia per i coefficienti standardizzati: il primo confronto è preferibile rispetto al secondo, poiché i coefficienti non standardizzati mantengono l'informazione relativa alla varianza

delle variabili nei diversi campioni. Laddove non ci sia differenza tra le varianze delle variabili nei due campioni, è sostanzialmente indifferente confrontare i coefficienti standardizzati ovvero quelli non-standardizzati. Per confrontare i coefficienti non-standardizzati si utilizza la seguente formula (vedi Cohen e Cohen, 1983. p. 111):

$$z = \frac{B_{i1} - B_{i2}}{\sqrt{SE_{B_i 1}^2 + SE_{B_i 2}^2}} \qquad (1)$$

Nella formula (1) il numeratore è ottenuto sottraendo i due coefficienti relativi all'impatto della variabile indipendente x_i sulla dipendente y nei due campioni 1 e 2; il denominatore si ottiene considerando gli errori standard al quadrato dei due parametri non standardizzati nei due campioni. Questo rapporto segue la distribuzione normale standardizzata alla quale dunque si può fare riferimento per interpretarne la significatività. Per effettuare la verifica delle ipotesi sulla differenza tra i coefficienti standardizzati basta sostituire gli elementi opportuni al numeratore e al denominatore della (1) e utilizzare lo stesso approccio. Consideriamo come esempio i risultati riportati in tabella 7, relativi all'analisi di regressione standard effettuata separatamente sulle consumatrici abituali del prodotto in esame (user) e consumatrici abituali di altri prodotti (non-user), nell'esempio discusso sul testo. La percentuale di varianza spiegata nei due sottocampioni sembra molto simile (.56 user, .54 non user), ma l'equazione di regressione sembra piuttosto diversa almeno per alcuni parametri. Mentre nelle user il comportamento passato non risulta significativo, nelle non-user è il controllo percepito a non essere significativo. Applicando la formula (1) ai coefficienti di regressione della tabella 7 (i calcoli sono stati effettuati tramite il programma Excell), solamente quelli relativi al controllo comportamentale percepito risultano significativamente diversi nei due campioni (z=3.61), mentre quelli relativi alla norma soggettiva (z=-.77), al comportamento passato (z=-1.84) e all'atteggiamento (z=-1.73) non risultano significativamente differenti. Risultati assolutamente analoghi si ottengono considerando i coefficienti standardizzati con relativi errori standard.

Tabella 7. Risultati delle analisi di regressione standard effettuate separatamente sulle consumatrici abituali del prodotto in esame (user) e consumatrici abituali di altri prodotti (non-user)

Coefficienti[a]

USER	Modello		Coefficienti non standardizzati		Coefficienti standardizzati		t	Sig.
			B	Errore std.	Beta	Errore std.		
1 user	1	(Costante)	-4,136	1,256			-3,293	,001
		NS	,302	,082	,296	,081	3,667	,000
		CONTCO	,673	,113	,442	,074	5,979	,000
		COMPAS	,095	,096	,078	,079	,984	,328
		ATT	,078	,024	,249	,078	3,202	,002
2 non_user	1	(Costante)	-3,445	1,405			-2,453	,016
		NS	,399	,121	,295	,090	3,291	,001
		CONTCO	,040	,134	,024	,079	,297	,767
		COMPAS	,429	,154	,222	,080	2,790	,007
		ATT	,146	,031	,427	,092	4,669	,000

a. Variabile dipendente: INT

Riepilogo del modello

USER	Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1 user	1	,751[a]	,564	,546	1,0419
2 non_user	1	,734[b]	,539	,516	1,6577

a. Stimatori: (Costante), ATT, COMPAS, CONTCO, NS

b. Stimatori: (Costante), ATT, CONTCO, COMPAS, NS

Nota. Per una descrizione dell'esempio si vedano le pagine 22-57 del testo

Differenze tra due coefficienti di regressione nello stesso campione.

Il confronto dell'effetto di due diverse variabili indipendenti sulla variabile dipendenti nello stesso campione si può effettuare tramite una formula più complicata della (1). Questo confronto può essere effettuato considerando i coefficienti non-standardizzati *solo se* le variabili sono espresse nella stessa unità di misura. Invece i coefficienti standardizzati possono essere utilizzati sempre, indipendentemente dall'unità di misura in cui le variabili sono espresse.

Il confronto tra i coefficienti **non standardizzati** avviene tramite la formula (2) che utilizza la distribuzione della t di Student e presenta al denominatore un'espressione per l'errore standard calcolabile tramite la formula (3) (Cohen e Cohen, 1983, Appendice 2):

$$t = \frac{B_i - B_j}{SE_{B_i - B_j}} \tag{2}$$

$$SE_{B_i - B_j} = \sqrt{SE_{B_i}^2 + SE_{B_j}^2 - 2SE_{B_i}SE_{B_j}\left(\frac{r^{ij}}{r^{ii}r^{jj}}\right)} \tag{3}$$

Nella 3, oltre gli errori standard dei parametri B (indicati nel modo consueto) troviamo anche 3 nuovi termini, r^{ii}, r^{jj}, r^{ij}, che sono contenuti nelle posizioni *ii*, *jj*, e *ij* di $\mathbf{R^{-1}}$, l'inversa della matrice di correlazione tra le variabili indipendenti. Nel caso dei coefficienti standardizzati l'errore standard della differenza si calcola tramite la formula (4), mentre la t di Student si ottiene sostituendo nell'equazione (2) i termini appropriati (Cohen e Cohen, 1983).

$$SE_{\beta_i - \beta_j} = \sqrt{\frac{1 - R^2_{Y\bullet 12\ldots k}}{n - k - 1}(r^{ii} + r^{jj} - 2r^{ij})} \tag{4}$$

Nella (4) i termini si interpretano come nella (3), con l'aggiunta del coefficiente di determinazione $R^2_{Y\bullet 12\ldots k}$, del numero di variabili indipendenti *k* e del numero di soggetti *n*.

La matrice $\mathbf{R^{-1}}$ si ottiene tramite le seguenti linee di programma SINTASSI:

```
**** Inversa della matrice di correlazione tra le variabili
indipendenti R-1.
MATRIX.
READ A/FILE='C:\matriceR.txt'
  /SIZE={4;4}/FIELD = 1 TO 32  BY 8.
COMPUTE B=INV(A).
PRINT B.
end matrix.
```

La tabella 8 presenta la matrice **R** originale e la sua inversa calcolate considerando la soluzione sul gruppo totale presentata nella tabella 18. Poiché le unità di misura delle variabili sono diverse, il confronto tra i coefficienti va effettuato considerando i coefficienti standardizzati. Come esempio

consideriamo i coefficienti relativi all'influenza dell'Atteggiamento (.378) e del Controllo Comportamentale Percepito (.146). Applichiamo la formula (4) per calcolare l'errore standard della differenza:

$$SE_{\beta_i - \beta_j} = \sqrt{\frac{1-.640}{184-4-1}(1.680 + 1.227 - 2*-.266)} = .083$$

La t di Student sarà dunque uguale a:

$$t = \frac{.378 - .146}{.083} = 2.80$$

C'è dunque una differenza significativa tra i due coefficienti di regressione.

Tabella 8. Matrice di correlazione **R** e inversa $\mathbf{R}^{-1}$

Matrice di correlazione **R**

	Atteggiamento	Senso di Controllo	Comport. Passato	Norma Soggettiva
Atteggiamento	1.000	0.361	0.500	0.545
Senso di Controllo	0.361	1.000	0.383	0.254
Comport. Passato	0.500	0.383	1.000	0.430
Norma Soggettiva	0.545	0.254	0.430	1.000

Matrice inversa $\mathbf{R}^{-1}$

	Atteggiamento	Senso di Controllo	Comport. Passato	Norma Soggettiva
Atteggiamento	1.680	-0.266	-0.458	-0.651
Senso di Controllo	-0.266	1.227	-0.326	-0.027
Comport. Passato	-0.458	-0.326	1.486	-0.306
Norma Soggettiva	-0.651	-0.027	-0.306	1.493

Riferimenti bibliografici

Cohen, J., e Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T. C. Lee. 1980. *The theory and practice of econometrics.* 2nd ed. New York: John Wiley and Sons

Meyer, L. S., and M. S. Younger. 1976. Estimation of standardized coefficients. *Journal of the American Statistical Association*, 71: 154–57.