STRUMENTI E METODI
PER
LE SCIENZE SOCIALI

Claudio Barbaranelli

ANALISI DEI DATI CON SPSS

II.

LE ANALISI MULTIVARIATE

LED
Edizioni Universitarie di Lettere Economia Diritto

ISBN 978-88-7916-315-9

LED Edizioni Universitarie di Lettere Economia Diritto
Via Cervignano 4 - 20137 Milano
Catalogo: www.lededizioni.com - E-mail: led@lededizioni.com

Videoimpaginazione e redazione grafica: Studio Venturini
Stampa: Bianca & Volta

INDICE

INTRODUZIONE

Questo testo fa parte di una "trilogia" dedicata all'analisi dei dati tramite SPSS. Il primo volume introduce al programma e alle sue funzionalità e affronta le analisi statistiche di base (descrizione e screening dei dati, analisi delle differenze delle medie, statistiche non parametriche). Il secondo volume affronta le analisi statistiche multivariate (regressione lineare multipla, analisi fattoriale, analisi della varianza, analisi discriminante, analisi multivariata della varianza, cluster analysis). Il terzo volume infine affronta l'analisi dei dati categoriali (regressione logistica, modelli loglineari, modelli logit, analisi delle corrispondenze semplici e multiple).

La divisione in tre volumi, oltre ad essere funzionale all'usabilità dei testi nei diversi percorsi formativi, universitari e non, vuole marcare una gradualità nell'approccio all'analisi dei dati con SPSS [1]. Anche se i tre volumi hanno una loro identità autonoma, nel complesso possono delineare una serie di percorsi nei quali il primo volume rappresenta una base comune per chi si affaccia per la prima volta al programma, mentre il secondo e il terzo volume affrontano tipologie diversificate di modelli di analisi *avanzati*. Da un lato si tratta di tecniche applicabili in contesti usualmente definiti *quantitativi*, dall'altro di tecniche che solitamente vengono associate (riteniamo in modo riduttivo e semplicistico) al mondo del *qualitativo*. Si tratta comunque di tecniche di analisi dei dati che si basano su modelli statistico-matematici sofisticati e complessi.

La diffusione del pacchetto SPSS e la sua estrema semplificazione (rispetto alle vecchie versioni "non" windows) possono aver creato, soprattutto negli utenti che si affacciano per la prima volta all'analisi dei dati, l'illusione di poter effettuare con un semplice clic del mouse le analisi

[1] Per la preparazione dei tre volumi è stata utilizzata la versione di SPSS 11.5.

più complesse. Niente di più errato! L'apparente semplicità con cui è possibile "comandare" il programma presuppone una approfondita conoscenza dei presupposti e dei modelli statistici alla base delle diverse tecniche. Come abbiamo già sostenuto altrove, spesso è la mancanza di conoscenza delle assunzioni e dei principi alla base delle diverse tecniche utilizzate ad essere la causa di gravi errori. A questo riguardo, la disponibilità di calcolatori super-veloci non solo nei centri di calcolo delle università ma direttamente a casa nostra, e di programmi sempre più *user friendly* per effettuare complesse analisi statistiche, rischia di produrre più guasti che altro. In definitiva, la facilità e semplicità del *software* non possono rappresentare un alibi per la mancata conoscenza dell'*hardware* rappresentato dai modelli statistici di riferimento.

Per queste ragioni la lettura dei volumi dedicati al programma SPSS presuppone la conoscenza teorica degli argomenti affrontati. Anche per queste ragioni, nessuno dei tre volumi affronta i modelli statistico-matematici alla base delle diverse tecniche trattate, per la conoscenza dei quali si rimanda ai diversi manuali (in italiano o in inglese) dedicati all'analisi dei dati. Gli argomenti discussi, oltre a riflettere quelle che sono le tecniche maggiormente utilizzate nella ricerca psicologico-sociale, risentono degli interessi degli autori, e della loro preferenza e inclinazione verso approcci più di tipo "confermativo" o di "test di ipotesi", rispetto ad approcci puramente esplorativi.

Diverse pagine del sito *www.lededizioni.com* completano il materiale contenuto nei tre volumi, con esempi concreti di dati da analizzare, esercizi di autoverifica, e descrizioni di opzioni di analisi che, pur non essendo utilizzate correntemente per effettuare le diverse tecniche discusse, consentono di completare lo spettro delle possibilità di analisi offerte da SPSS.

Il presente volume (dedicato alle analisi multivariate) affronta dalla prospettiva di SPSS gli argomenti che ho trattato nel mio precedente libro edito da LED, *Analisi dei dati*. Spesso gli esempi di analisi discussi in questo testo sono esattamente gli stessi che ho descritto nel testo precedente. La novità di questa nuova trattazione riguarda la descrizione delle procedure necessarie per un'implementazione dei diversi esempi tramite il programma SPSS, e dei risultati ottenibili dal programma. Una volta completata la lettura degli esempi, il lettore è invitato a scaricare i file di dati presenti sul sito e ripetere per conto proprio quanto presentato nel testo. In questo modo, la comprensione degli argomenti, e la sedimentazione di quanto appreso, possono risultare facilitate.

Voglio infine ringraziare Francesca D'Olimpio, Roberta Fida, Elena Natali e Michele Vecchione che, in diverse fasi della preparazione di questo volume, mi hanno aiutato a definirne meglio i contenuti e a chiarire quello che poteva risultare ambiguo.

Roma, ottobre 2006

1.

LA REGRESSIONE LINEARE

PREMESSA

L'analisi della regressione è una tecnica che esamina la relazione tra una o più variabili esplicative (o indipendenti) e una variabile criterio (o dipendente) con un duplice scopo:

a) *esplicativo*: comprendere e ponderare gli effetti delle Variabili Indipendenti (VI) sulla Variabile Dipendente (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.

In questo testo ci interessiamo solamente della relazione di tipo lineare e rimandiamo ad un testo dedicato all'analisi dei dati categoriali con SPSS per la trattazione dei modelli di regressione non lineari.

Il punto di partenza della regressione è rappresentato da una matrice che riassume le relazioni lineari tra la VD (misurata per lo meno su una scala ad intervalli equivalenti) e le VI (che invece possono essere quantitative oppure dicotomiche), e tra le VI stesse.

Il punto di arrivo è rappresentato da:

a) un insieme di parametri che riassumono la relazione tra VD e VI, sotto le ipotesi che la prima sia effetto o *determinata* dalle seconde, e che nell'esame dell'influenza di ogni VI sulla VD, il valore delle altre VI sia mantenuto costante;
b) una statistica per l'esame della significatività dei parametri, e un valore di probabilità associato ad ognuno di questi parametri;
c) un valore che riassume la proporzione di varianza della VD che complessivamente è spiegata dalle VI.

L'analisi di regressione si articola in alcuni passi fondamentali:

1. valutazione dell'adeguatezza delle variabili (livello di misura, distribuzione, collinearità, ecc.);
2. scelta della strategia analitica per inserire le VI in equazione;
3. interpretazione della soluzione;
4. validazione della soluzione.

Nelle pagine che seguono presenteremo la procedura per la regressione con SPSS e vedremo come tramite tale procedura è possibile affrontare i diversi passi indicati sopra.

1.1. Una descrizione della procedura per l'Analisi di Regressione Multipla Lineare

Per effettuare l'analisi di regressione multipla lineare bisogna scegliere dal menù Analizza la procedura "Regressione" e quindi l'opzione "Lineare" (*Figura 1.1.*). Si aprirà, dunque, la finestra di dialogo principale per specificare le opzioni di esecuzione dell'Analisi di Regressione (*Figura 1.2.*). Innanzi tutto è necessario scegliere dalla lista delle variabili attive una Variabile Dipendente (VD), che verrà inserita nel box rettangolare che ha l'etichetta "Dipendente", e una o più Variabili Indipendenti (VI), che verranno inserite nel box che ha l'etichetta "Indipendenti". Queste selezioni verranno effettuate con un clic del mouse sul corrispondente pulsante con il triangolino posto al centro della finestra di dialogo. Le variabili selezionate si sposteranno in questo modo nei due box per le variabili nella finestra di dialogo iniziale, come riportato nella *Figura 1.2.*

Le variabili indipendenti possono essere inserite nell'analisi in diversi blocchi. L'utente inoltre può scegliere il metodo attraverso il quale le variabili in ciascun blocco vengono elaborate, tramite le opzioni presenti nel menù a tendina "Metodo". Utilizzando metodi differenti, dunque, si possono implementare modelli diversi di regressione, a partire dallo stesso set di variabili.

I metodi disponibili sono i seguenti.

– *Per blocchi*: si usa per la regressione standard e per la regressione gerarchica.
– *Per passi* (Stepwise), *Rimozione* (Remove), *Indietro* (Backward) e *Avanti* (Forward): si usano per la regressione statistica.

Descriveremo brevemente queste diverse strategie analitiche per la regressione nelle pagine successive (per una trattazione più approfondita si vedano Barbaranelli, 2003, Pedhazur, 1997, e Tabachnick e Fidell, 1989).

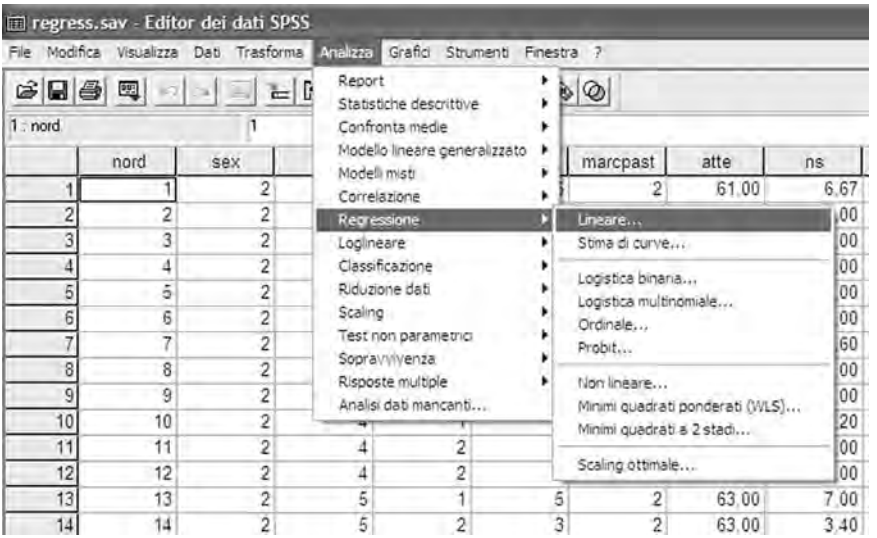


Figura 1.1. – Procedura Regressione del menù Analizza.

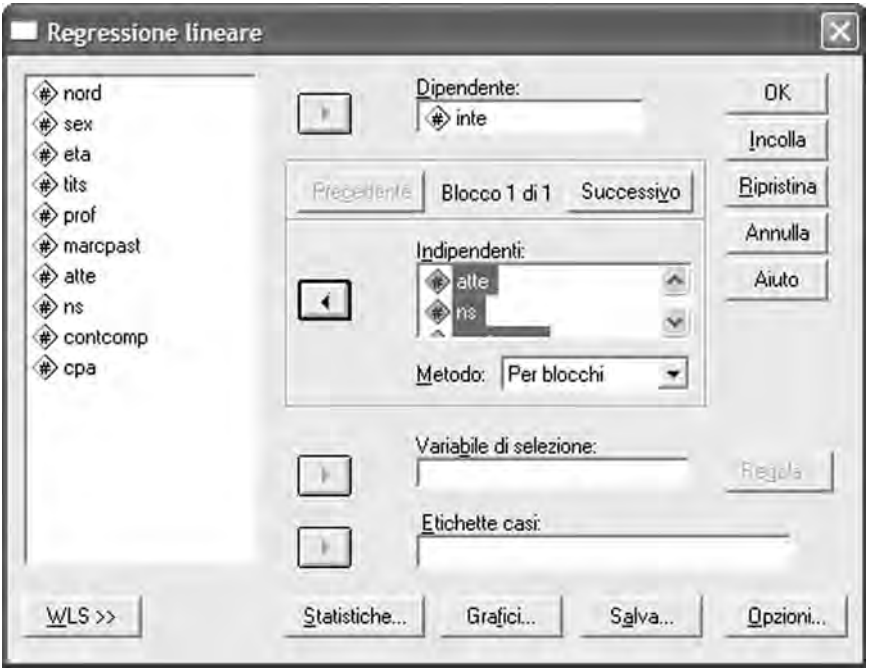
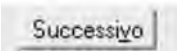


Figura 1.2. – Finestra di dialogo principale per la procedura di regressione lineare.

Le variabili indipendenti vengono inserite a partire dal primo blocco. Il metodo selezionato in questo blocco viene applicato a tutte le variabili indipendenti presenti nel blocco stesso. Una volta selezionati la variabile dipendente (criterio), le variabili indipendenti e il metodo, si procede, se necessario, al blocco successivo, con un clic del mouse sul pulsante



posto sopra la finestra delle variabili indipendenti.

Anche se tutte le variabili indipendenti selezionate in un blocco vengono trattate con lo stesso modello di regressione, è possibile specificare differenti metodi di entrata per differenti blocchi di variabili, a seconda delle necessità del ricercatore. Ad esempio, si può inserire un primo blocco di variabili nel modello di regressione utilizzando l'opzione "Per passi" ed un secondo blocco utilizzando l'opzione "Per blocchi". Vedremo nelle prossime pagine esempi di applicazione dei diversi metodi di Regressione.

Nella finestra di dialogo principale (*Figura 1.2.*) sono presenti diversi pulsanti che consentono di arricchire l'output di un'analisi di regressione. Innanzi tutto il box "*Variabile di selezione*" serve per specificare la variabile che eventualmente può essere utilizzata per selezionare i soggetti da inserire in analisi, se si vuole effettuare l'analisi solo su una parte

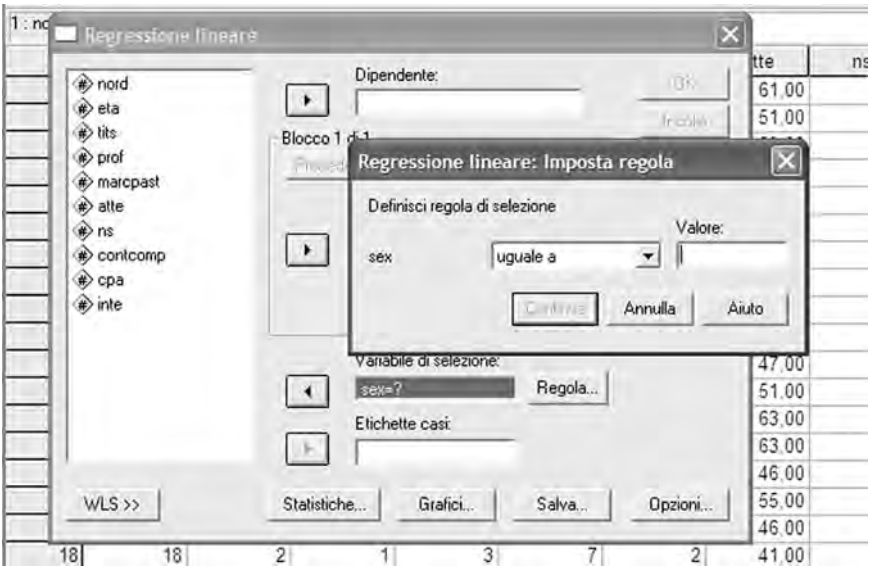


Figura 1.3. – Impostazione delle operazioni di selezione dei soggetti tramite la variabile di selezione.

del campione. Specificando, tramite il pulsante "*Regola…*", il valore corrispondente al gruppo da selezionare verranno selezionati solo i soggetti che presentano quel valore specificato (vedi *Figura 1.3.*). Ad esempio, nella *Figura 1.3.*, inserendo il numero 1 nel campo 1 della finestra "Imposta regola", l'analisi verrà effettuata soltanto sui soggetti il cui sesso è codificato con il numero 1 (nel nostro file saranno i maschi).

Con un clic del mouse sul pulsante Statistiche della finestra di dialogo principale, si aprirà una nuova finestra di dialogo (*Figura 1.4.*). Di default vengono fornite le Stime dei coefficienti di regressione, e il test di adattamento del modello (R^2). In particolare, l'opzione "*Stime*" produrrà un output con il coefficiente di regressione *B*, l'errore standard di *B*, il coefficiente beta standardizzato, il valore *t* per *B*, e il livello di significatività bidirezionale di *t*. L'opzione "Adattamento del modello" produrrà un output con le statistiche relative alla bontà dell'adattamento: *R* multiplo, R^2, e R^2 corretto (*adjusted*), errore standard della stima e tabella di analisi della varianza per la verifica della significatività dell'R^2 (per un quadro delle formule relative all'equazione di regressione multipla e ai principali coefficienti di regressione discussi in questo capitolo si veda Barbaranelli, 2003).

Nella sezione relativa ai coefficienti di regressione della finestra Statistiche è possibile selezionare due ulteriori opzioni. "*Intervalli di*

Figura 1.4. – Finestra di dialogo Statistiche.

confidenza" produrrà in output gli intervalli di confidenza al 95% per le stime non standardizzate dei parametri: se il valore 0 è compreso tra i limiti inferiore e superiore, allora la stima non risulta significativamente diversa da 0. Se invece i due valori sono entrambi negativi o entrambi positivi (quindi non comprendono il valore di zero all'interno dell'intervallo), allora la stima risulta significativa al livello di probabilità di .05 (ovvero, al 95%).

L'opzione "*Matrice di covarianza*" produrrà in output le matrici di correlazione e di covarianza tra tutte le variabili (indipendenti e dipendente) selezionate per l'analisi.

L'opzione "*Cambiamento di R quadrato*" produrrà in output i cambiamenti nell'R^2 attraverso i diversi passi di una regressione gerarchica o statistica, e il test di verifica delle ipotesi sul cambiamento dell'R^2 (*Variazione di F*) con la relativa significatività statistica. Questo test è particolarmente importante quando le variabili indipendenti sono inserite in blocchi diversi. Come vedremo negli esempi sulla regressione gerarchica e sulla regressione statistica, infatti, il test di verifica delle ipotesi sul cambiamento di R^2 valuta il valore aggiunto di una variabile indipendente (o di un insieme di variabili indipendenti all'interno di un blocco) nello spiegare la varianza della variabile dipendente, rispetto a quanto già spiegato dalle variabili indipendenti (o dai blocchi) inserite in precedenza.

L'opzione "*Descrittive*" produrrà in output due tabelle di informazioni, la prima con il numero di casi validi, la media, e la deviazione standard per ciascuna variabile nelle analisi, la seconda con i coefficienti di correlazione.

L'opzione "*Correlazioni di ordine zero e parziali*" produrrà in output i coefficienti di correlazioni semplici (o "di ordine zero"), quelli parziali e quelli semiparziali (o "parziali indipendenti") della variabile dipendente rispetto a ciascuna variabile indipendente selezionata.

L'opzione "*Test di collinearità*" produrrà in output alcuni importanti indici (come la Tolleranza e il VIF) che servono per analizzare la ridondanza tra le variabili indipendenti.

Le opzioni relative ai residui sono molto importanti perché consentono di esaminare la verifica delle assunzioni alla base della regressione e la presenza di outliers. In particolare, l'opzione "*Durbin-Watson*" produrrà in output l'indice omonimo che consente di esaminare la presenza di autocorrelazione dei residui. L'opzione "*Diagnostiche per casi*" produrrà una tabella che conterrà i valori dei residui (grezzi e standardizzati), il valore osservato e il valore atteso della variabile dipendente per

tutti i casi o per quei casi che presentano un residuo standardizzato maggiore della quantità specificata nel box "*Valori oltre*" (il valore di default è pari a 3). Selezionando una qualsiasi tra le due opzioni presenti in Residui verrà prodotta in output una tabella che contiene le statistiche descrittive (minimo, massimo, media, deviazione standard, numero di osservazioni) per il valore atteso grezzo e standardizzato e per il residuo grezzo e standardizzato.

Con un clic del mouse sul pulsante Grafici della finestra di dialogo principale, si aprirà una nuova finestra di dialogo (*Figura 1.5.*). I grafici rappresentano un valido supporto visivo non solo per l'esame delle assunzioni di normalità, linearità, e omogeneità della varianza, ma anche per rivelare valori errati, anomali (*outliers*) e casi influenti (*influential data points*). La finestra di dialogo riportata in *Figura 1.5.* consente di produrre un diagramma di dispersione (*scatterplot*) costruibile considerando due delle seguenti variabili:

– variabile dipendente originale (DEPENDNT);
– valori stimati standardizzati (*ZPRED);
– residui standardizzati (*ZRESID);
– residui cancellati (*DRESID) e valori stimati corretti (*ADJPRED): sono rispettivamente il residuo e il valore teorico di un caso quando il caso viene eliminato dalla regressione;

Figura 1.5. – Finestra di dialogo Grafici.

– residui studentizzati (*SRESID): si tratta di un particolare residuo standardizzato che tiene conto della distanza del caso cui si riferisce dal vettore delle medie delle variabili indipendenti, e quindi della peculiarità del caso in esame;
– residui cancellati studentizzati (*SDRESID): corrispondono al residuo cancellato per un caso diviso per il proprio errore standard. La differenza fra questo valore e il relativo residuo studentizzato indica quanto la cancellazione di un caso influenza la regressione.

Per i grafici richiesti sono visualizzate le statistiche riassuntive dei valori stimati standardizzati e dei residui standardizzati (*ZPRED e *ZRESID).

Per costruire un diagramma di dispersione, si seleziona una variabile per l'asse delle Y (verticale) e una variabile per l'asse delle X (orizzontale). Si possono richiedere più diagrammi di dispersione tramite un clic del mouse sul pulsante "Successivo".

Dalla finestra di dialogo per i grafici si possono ottenere anche i seguenti altri grafici.
– "*Istogramma*": mostra la distribuzione dei residui standardizzati.
– "*Grafico di probabilità normale*": confronta la distribuzione dei residui standardizzati con la distribuzione normale.

Sempre dalla finestra di dialogo per i grafici si possono ottenere tutti i "*grafici parziali*". Un grafico parziale è il diagramma di dispersione dei residui di una data variabile indipendente e dei residui della variabile dipendente quando entrambe sono regredite separatamente sul resto delle variabili indipendenti. Per produrre un grafico parziale è necessario che nell'equazione di regressione siano presenti almeno due variabili indipendenti. Questi grafici sono utili per evidenziare possibili violazioni di linearità, additività e omoschedasticità a carico di particolari variabili.

Con un clic del mouse sul pulsante Salva della finestra di dialogo principale, si aprirà la omonima finestra di dialogo (*Figura 1.6.*) che consente di salvare come nuove variabili una serie di punteggi particolarmente utili per l'esame dei casi anomali. I punteggi nelle sezioni "*Valori previsti*" e "*Residui*" sono gli stessi che possono essere utilizzati per effettuare i grafici (vedi sopra).

Vediamo brevemente le altre sezioni di questa finestra.

Nella sezione "*Distanze*" si trovano gli indici che consentono di identificare i casi con combinazioni inusuali di valori per le variabili indipendenti, e casi che possono avere un forte impatto sul modello di regressione. Si possono richiedere tre tipi di indici di distanza.
– *Distanza di Mahalanobis*: è una misura della distanza del caso specifi-

co dalla media di tutti i casi sulle variabili indipendenti. È utile per individuare quei casi che hanno valori estremi in una o più variabili indipendenti.
– *Distanza di Cook*: è una misura di quanto i residui di tutti i casi cambierebbero se un caso particolare fosse escluso dal calcolo dei coefficienti di regressione. Un coefficiente elevato indica che escludendo il caso dall'analisi i coefficienti verrebbero cambiati in modo sostanziale.
– *Valori di influenza (Leverage)*: è una misura dell'influenza di un caso sulla bontà dell'adattamento del modello di regressione.

Nella sezione "*Statistiche di influenza*" si possono scegliere alcuni indici che consentono di esaminare il cambiamento che avviene nei coefficienti di regressione (*DiffBeta*) e nel valore previsto della variabile

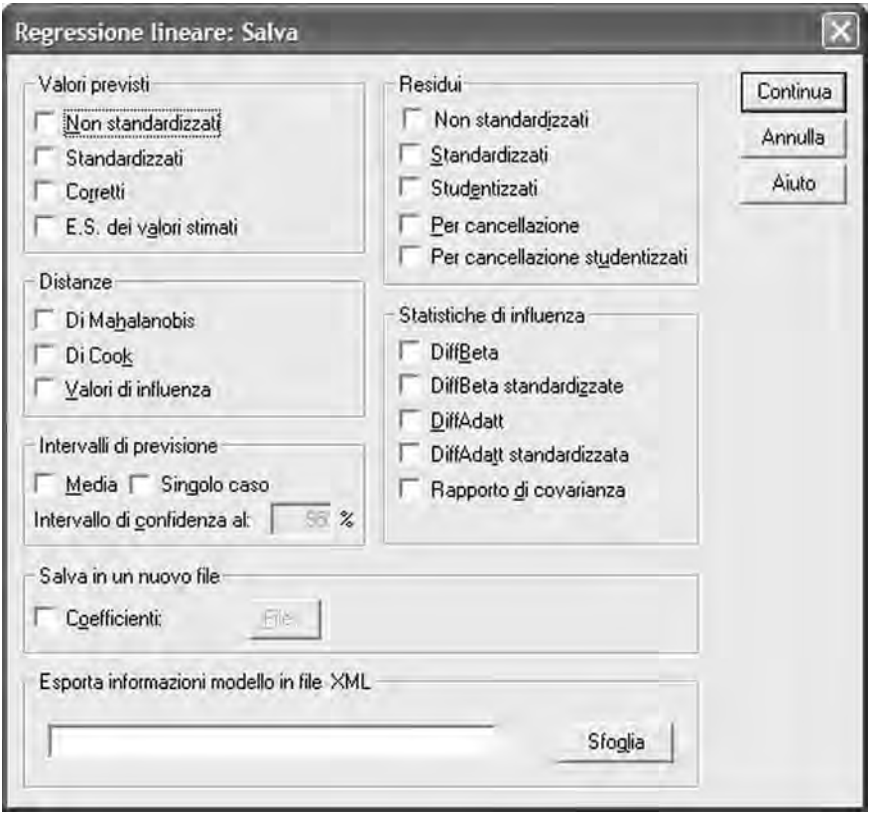


Figura 1.6. – Finestra di dialogo Salva.

dipendente (*DiffAdatt*) se il caso viene escluso dal calcolo della regressione. Inoltre, il Rapporto di covarianza (definito anche *Covratio*) indica il rapporto tra il determinante della matrice di covarianza, se il caso venisse escluso dal calcolo dei coefficienti di regressione, e il determinante con tutti i casi inclusi. Se il rapporto è vicino a 1, il caso non altera molto la matrice di covarianza.

Come già detto, è possibile salvare una serie di indici relativi ai punteggi presentati dal soggetto, e all'influenza del soggetto sulla soluzione. I valori della *distanza di Mahalanobis* e quelli di *influenza* (*Leverage*) consentono di identificare possibili outliers nelle variabili indipendenti. I due parametri danno la stessa informazione. Valori di *influenza* inferiori a .2 identificano casi che non sono outliers multivariati, valori compresi tra .2 e .5 identificano casi sospetti, mentre valori maggiori di .5 identificano casi che probabilmente rappresentano outliers multivariati. Per quanto riguarda la *distanza di Mahalanobis*, si considerano possibili outliers quelli il cui valore di distanza ha una probabilità <.001. La probabilità viene interpretata facendo riferimento alla distribuzione del χ^2, con gradi di libertà uguali al numero complessivo di variabili, dipendente e indipendenti.

Con l'opzione "*Intervalli di previsione*" è invece possibile salvare i limiti superiore ed inferiore degli intervalli relativi ai punteggi previsti nella variabile dipendente, sia medi sia individuali. Mentre il primo si ottiene considerando l'errore standard medio dei punteggi previsti, il secondo si ottiene considerando una stima individuale dell'errore standard (vedi Pedhazur, 1997, pp. 204-207): poiché il primo risulta più basso del secondo, l'intervallo medio sarà sempre più piccolo di quello individuale.

Con un clic del mouse sul pulsante Opzioni della finestra di dialogo principale, si aprirà l'omonima finestra di dialogo (*Figura 1.7.*) nella quale i "*Criteri di accettazione e rifiuto*" permettono di cambiare i criteri secondo i quali le variabili possono essere inserite o rimosse dal modello: si tratta di un'opzione applicabile solo quando viene utilizzata una strategia analitica di tipo statistico (specificando i metodi "Per passi", "Rimozione", "Avanti" e "Indietro" nella finestra di dialogo principale). Se selezioniamo "*Usa probabilità di F*", una variabile è inserita in equazione se il livello di significatività della F è minore del valore specificato in "Inserimento", ed è rimossa se esso è maggiore del valore specificato in "Rimozione". Se selezioniamo "*Usa valore di F*", una variabile viene inserita se il suo valore di F è maggiore del valore specificato in "Inserimento", e rimossa se il valore di F è minore del valore specificato in "Rimozione".

L'opzione "*Includi termine costante nell'equazione*" è selezionata di default: se lasciamo questa selezione il termine costante (ovvero, l'intercetta dell'equazione) viene incluso nel modello di regressione. Se invece questa opzione viene deselezionata, l'intercetta viene eguagliata a 0 e la retta di regressione sarà forzata a passare per l'origine. In questo caso, però, i risultati potrebbero non essere confrontabili con quelli ottenuti da regressioni in cui il termine costante viene inserito.

Nelle opzioni relative ai "*Valori mancanti*" possiamo decidere tra tre differenti strategie per il trattamento dei valori mancanti.

Esclusione listwise: vengono inclusi solo i casi con valori validi per tutte le variabili;

Esclusione pairwise: ogni coefficiente di regressione è calcolato usando tutti i casi con valori validi per le due variabili che sono correlate. I gradi di libertà sono basati sul minimo N tra tutte le coppie di variabili esaminate.

Sostituisci con la media: selezionando questa opzione vengono utilizzati tutti i casi. Laddove manchino osservazioni, i valori vengono sostituiti con la media della variabile. Lo svantaggio di questo metodo è nel

Figura 1.7. – Finestra di dialogo Opzioni.

rischio di creare soggetti *fittizi* caratterizzati da valori nelle variabili che sono il risultato della sostituzione con la media. L'utente quindi deve filtrare i casi a seconda del numero di valori mancanti presenti, ed eliminare dall'analisi quei casi che presentano una percentuale eccessiva di valori mancanti. Solo dopo questa operazione di filtraggio è possibile applicare la modalità di sostituzione con la media senza incorrere nel rischio di creare soggetti fittizi.

Ora che abbiamo introdotto tutte le finestre di dialogo della procedura per la regressione multipla siamo pronti per discutere un esempio completo di analisi basato su dati reali.

1.2. Un esempio completo di analisi di regressione

Vediamo un esempio di analisi della regressione facendo riferimento alle opzioni di analisi illustrate nelle pagine precedenti. Si tratta di un'analisi effettuata sulle 5 variabili nel file "**regress.sav** ". I dati dell'esempio provengono da una nostra ricerca nella quale la teoria del comportamento pianificato (Ajzen, 1988) è stata applicata al comportamento di acquisto (vedi Caprara e Barbaranelli, 2000). La variabile dipendente era l'intenzione di acquistare una confezione di bagno-schiuma di una nota marca leader di quel settore di mercato. Le variabili indipendenti erano l'atteggiamento, la norma soggettiva, il controllo comportamentale percepito e il comportamento passato di acquisto. Il campione era costituito da 100 consumatrici della marca in esame, e 99 consumatrici di altre marche.

Per applicare correttamente le procedure di analisi di regressione lineare ed effettuare la verifica delle ipotesi sui parametri, è necessario che siano soddisfatti alcuni assunti essenziali che ricordiamo di seguito.

1. La variabile indipendente deve essere quantitativa o dicotomica, e la variabile dipendente deve essere quantitativa (ovvero, misurata almeno al livello degli intervalli equivalenti).
2. La varianza di ogni variabile indipendente deve essere maggiore di 0.
3. Il campionamento deve essere casuale semplice.
4. Non deve esserci errore di specificazione, ovvero:
   a. la forma della relazione tra le VI e la VD deve essere lineare;
   b. non devono essere state omesse VI rilevanti;
   c. non devono essere state incluse VI irrilevanti.

5. Tutte le variabili devono essere misurate senza errore (assenza di errore di misurazione).
6. Se vi è più di una VI, nessuna di esse deve essere una combinazione lineare perfetta delle altre. Questa condizione viene definita come assenza di *perfetta multicollinearità*. In generale è bene che le VI non siano troppo correlate, ovvero è bene che le loro correlazioni siano moderate. Se le VI sono fortemente, anche se non perfettamente, correlate si parla di *quasi-multicollinearità*, o più generalmente di *multicollinearità* (vedi Sadocchi, 1987). Mentre la perfetta multicollinearità rappresenta un evento abbastanza raro, la quasi-multicollinearità è invece un evento piuttosto frequente.
7. Completano il quadro le seguenti assunzioni sui residui (o termini di errore) ε_i:
   a. *media uguale a zero*: $E(\varepsilon_i) = 0$; per ogni combinazione di valori delle VI, il valore atteso dei residui deve essere uguale a 0;
   b. *omoschedasticità*: $\mathrm{VAR}(\varepsilon_i) = \sigma^2$, per ogni i; la varianza dei residui deve essere costante per tutte le combinazioni dei valori delle VI;
   c. *normalità*: le distribuzioni dei valori di ε_i per ogni combinazione di valori delle VI devono essere di forma normale;
   d. *assenza di autocorrelazione*: $\mathrm{Cov}(\varepsilon_i, \varepsilon_i) = 0$, per ogni i e j, con $i \neq j$; i residui associati ad osservazioni diverse non devono essere correlati;
   e. le VI *non devono essere correlate* con i residui: $\mathrm{Cov}(\varepsilon_i, X_i) = 0$.

1.2.1. Analisi delle assunzioni

Le variabili sono misurate tutte a livello degli intervalli equivalenti, quindi possiamo considerarle come quantitative. La prima assunzione dunque risulta rispettata. La *Tabella 1.1.* riassume le principali statistiche descrittive delle variabili in analisi. È facile verificare che tutte le variabili indipendenti (ma anche la variabile dipendente INT) hanno deviazione standard maggiore di 0. Il campionamento dei soggetti non è casuale semplice: tuttavia la violazione di questa assunzione non introduce distorsioni apprezzabili nei risultati, quindi possiamo ignorarla.

Per quanto riguarda la normalità delle relazioni tra le variabili, dalla *Tabella 1.1.* è facile verificare che solo la variabile CONTCO presenta problemi di normalità (tutte le altre variabili hanno valori di asimmetria e curtosi sotto la soglia di $|1|$). Anche se la normalità delle distribuzioni delle variabili non è un'assunzione della regressione, la presenza di

Tabella 1.1. – Statistiche descrittive delle variabili in analisi.

Statistiche descrittive

	N	Minimo	Massimo	Media	Deviazione	Asimmetria		Curtosi	
	Statistica	Statistica	Statistica	Statistica	Statistica	Statistica	Errore std	Statistica	Errore std
ATT	199	16,00	54,00	42,7990	7,31125	-,869	,172	,209	,343
NS	199	2,00	10,00	7,8844	1,80115	-,554	,172	-,339	,343
CONTCO	199	2,00	10,00	8,6834	1,86279	-1,850	,172	3,444	,343
COMPAS	199	,00	8,00	2,6432	1,96904	,391	,172	-,446	,343
INT	199	2,00	10,00	7,2915	2,54365	-,680	,172	-,630	,343
Validi (listwise)	199								

Nota. ATT = Atteggiamento, NS = Norme Soggettive, CONTCO = Controllo Comportamentale Percepito, COMPAS = Comportamento Passato, INT = Intenzione.

Tabella 1.2. – Distribuzioni di frequenze relative alle variabili che presentano punteggi estremi.

ZATT Punteg(ATT)

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	-3,66545	1	,5	,5	,5
	-2,84479	1	,5	,5	1,0
	-2,57124	1	,5	,5	1,5
	-2,29769	2	1,0	1,0	2,5
	-2,16092	1	,5	,5	3,0
	-2,02414	3	1,5	1,5	4,5

ZNS Punteg(NS)

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	-3,26703	1	,5	,5	,5
	-2,15663	9	4,5	4,5	5,0

ZCONTCO Punteg(CONTCO)

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	-3,58786	5	2,5	2,5	2,5
	-3,05103	2	1,0	1,0	3,5
	-2,51420	3	1,5	1,5	5,0
	-1,97737	4	2,0	2,0	7,0
	-1,44054	7	3,5	3,5	10,6
	-,90371	12	6,0	6,0	16,6
	-,36688	40	20,1	20,1	36,7
	,16995	26	13,1	13,1	49,7
	,70678	100	50,3	50,3	100,0
	Totale	199	100,0	100,0	

Tabella 1.3. – Statistiche descrittive relative alle variabili in analisi (tutti i soggetti).

Statistiche descrittive

	N	Minimo	Massimo	Media	Deviazione	Asimmetria		Curtosi	
	Statistica	Statistica	Statistica	Statistica	Statistica	Statistica	Errore std	Statistica	Errore std
ATT	190	22,00	50,00	43,2789	6,83680	-,858	,176	-,006	,351
NS	190	4,00	10,00	7,9895	1,72129	-,487	,176	-,607	,351
CONTCO	190	4,00	10,00	8,9158	1,43031	-1,386	,176	1,551	,351
COMPAS	190	,00	8,00	2,7474	1,94892	,348	,176	-,416	,351
INT	190	2,00	10,00	7,4579	2,45506	-,770	,176	-,401	,351
Validi (listwise)	190								

Tabella 1.4. – Statistiche descrittive relative alle variabili in analisi (solo i soggetti selezionati).

Statistiche descrittive

	N	Minimo	Massimo	Media	Deviazione	Asimmetria		Curtosi	
	Statistica	Statistica	Statistica	Statistica	Statistica	Statistica	Errore std	Statistica	Errore std
ATT	187	22,00	50,00	43,4599	6,68027	-,852	,178	-,015	,354
NS	187	4,00	10,00	8,0107	1,69753	-,483	,178	-,592	,354
CONTCO	187	5,00	10,00	8,9947	1,29722	-1,185	,178	,698	,354
COMPAS	187	,00	8,00	2,7754	1,94904	,329	,178	-,420	,354
INT	187	2,00	10,00	7,5241	2,41483	-,811	,178	-,279	,354
Validi (listwise)	187								

distribuzioni che si discostano dalla normale può mettere a rischio il rispetto di assunzioni come la linearità delle relazioni, la normalità e la omoschedasticità dei residui. Poiché la normalità può essere fortemente influenzata dalla presenza di *outliers univariati*, esaminiamo se ci sono casi che possono essere considerati come tali: questo può essere importante soprattutto per la variabile CONTCO, che presenta la più forte deviazione dalla normale. Calcoliamo i punteggi standardizzati (con la procedura STATISTICHE DESCRITTIVE ⇒ DESCRITTIVE) e chiediamo la distribuzione delle frequenze per le 5 variabili standardizzate. La *Tabella 1.2.* riassume alcuni dei risultati ottenuti dalle distribuzioni di frequenze, limitatamente alle variabili che presentano casi con valori superiori a $|3|$. Eliminiamo dalle analisi successive tutti i soggetti che presentano tali punteggi nelle distribuzioni delle variabili in *Tabella 1.2.* Per le distribuzioni delle altre variabili non sembrano esserci casi estremi e proviamo a vedere come cambiano di conseguenza i valori delle statistiche descrittive delle variabili. La *Tabella 1.3.* evidenzia che i valori delle asimmetrie e delle curtosi sono diminuiti: tuttavia quelli di CONTCO risultano ancora elevati. Eliminiamo anche i 3 soggetti che nella *Tabella 1.2.* per la variabile CONTCO presentano un valore pari a circa – 2.5: la *Tabella 1.4.* evidenzia il risultato di questa selezione, mostrando valori di curtosi nei limiti per tutte le variabili, e di asimmetria fuori dai limiti per

la variabile CONTCO, ma comunque molto più bassi rispetto alla situazione iniziale di *Tabella 1.2*.

Sfortunatamente nessuna delle trasformazioni usuali (logaritmo, radice quadrata, reciproco) migliora ulteriormente le caratteristiche della distribuzione della variabile CONTCO. Dobbiamo mantenere tale variabile nella forma originale ricordandoci che comunque essa presenta una violazione significativa rispetto alla distribuzione normale.

Esaminiamo ora la presenza di outliers multivariati. Per condurre questa analisi dobbiamo utilizzare la procedura dell'analisi di regressione multipla. Descriveremo in maggiore dettaglio nell'appendice di questo capitolo i presupposti alla base della normalità multivariata e la procedura per calcolarla con SPSS. Nel nostro esempio, nella finestra di dialogo iniziale dobbiamo inserire una variabile qualsiasi (possiamo scegliere il numero d'ordine del soggetto, NORD) come dipendente e tutte le variabili sulle quali stiamo lavorando (quindi le 4 indipendenti ATT = Atteggiamento, NS = Norme Soggettive, CONTCO = Controllo Comportamentale Percepito, COMPAS = Comportamento Passato, e la dipendente INT = Intenzione) nel box delle indipendenti. Quindi, nella finestra di dialogo SALVA selezioniamo "Distanze di Mahalanobis". Nel file attivo verrà salvata la nuova variabile *mah_1* che quantifica la distanza ponderata di ogni soggetto dal centroide del campione (ovvero, dal vettore delle medie sulle 5 variabili calcolato sul nostro campione). Per ogni soggetto, il valore della distanza di Mahalanobis viene interpretato considerando la distribuzione del chi-quadrato e considerando un livello di probabilità critico pari a .001 e gradi di libertà uguali al numero di variabili che stiamo considerando (Tabachnik e Fidell, 1989): con 5 gradi di libertà il valore critico del chi-quadrato per $p < .001$ è pari a 20.51. Richiedendo una distribuzione delle frequenze di *mah_1* che viene mostrata nella *Tabella 1.5*. (limitatamente ai valori più bassi e più alti), è possibile vedere che nessun soggetto presenta un valore superiore al valore critico. Non sembrano dunque esserci outliers multivariati.

Per vedere se la distribuzione è normale multivariata possiamo utilizzare i valori della distanza di Mahalanobis e calcolare l'indice di curtosi multivariata di Mardia (vedi appendice). L'indice si può ricavare tramite SPSS calcolando la variabile mah_1 elevata al quadrato e richiedendo la media di questa variabile, che risulta pari a 26.7. Il valore critico si calcola considerando il numero di variabili (p) moltiplicato per $p + 2$, nel nostro caso 5 * 7 = 35. Quindi, la nostra distribuzione non devia dalla normale multivariata. È possibile effettuare anche un test grafico (vedi appendice) utilizzando il plot dei quantili rispetto alla distribuzione del

Tabella 1.5. – Distribuzione delle frequenze della distanza di Mahalanobis.

MAH_1 Mahalanobis Distance

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	,51422	1	,5	,5	,5
	,52835	1	,5	,5	1,1
	.....	.....	.....	.....	.....
	11,91927	1	,5	,5	97,9
	13,29379	1	,5	,5	98,4
	13,40615	1	,5	,5	98,9
	13,95142	1	,5	,5	99,5
	14,48132	1	,5	,5	100,0
	Totale	187	100,0	100,0	

chi-quadrato. La *Figura 1.8.* evidenzia la distribuzione che si ottiene e che risulta compatibile con la normale multivariata. In definitiva anche se la distribuzione univariata di una variabile non si conforma pienamente alla normale, la distribuzione multivariata può essere considerata sostanzialmente normale. Questo è importante poiché la normale multivariata garantisce che le relazioni tra le variabili siano sostanzialmente lineari.

Per quanto riguarda l'assenza di errore di specificazione, solo dopo l'esame dei risultati dell'analisi si può verificare che non siano state omesse variabili indipendenti rilevanti, e che non siano state incluse

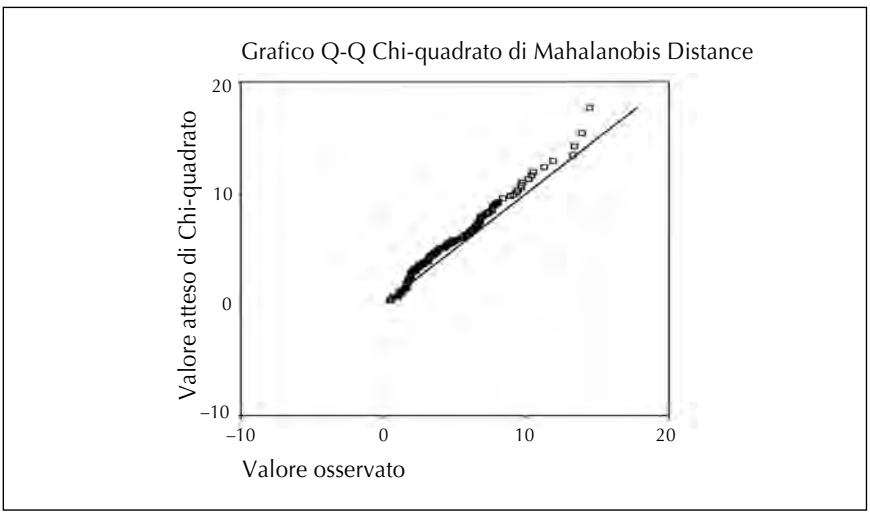


Figura 1.8. – Plot dei quantili della distribuzione della distanza di Mahalanobis rispetto alla distribuzione chi-quadrato.

variabili indipendenti irrilevanti. Da un punto di vista teorico si può garantire che il modello esaminato presenti tutte le variabili importanti individuate dalla teoria di riferimento, mentre è una questione empirica stabilire se alcune di esse possano non risultare significative per i dati considerati nell'esempio.

Per quanto riguarda l'assenza di errore di misurazione, i coefficienti di attendibilità calcolati tramite il coefficiente alfa di Cronbach sono risultati pari a .92 per l'Atteggiamento (che è misurato da 10 item), .84 per le Norme Soggettive (misurate da 2 item), .56 per il Controllo Comportamentale Percepito (misurato da 2 item), .46 per il Comportamento Passato (misurato da 2 item), e .99 per l'Intenzione (misurata da 2 item). Solamente per le variabili Intenzione e Atteggiamento è possibile concludere che l'errore di misurazione è assente o irrilevante. Le Norme Soggettive presentano un'incidenza modesta dell'errore (circa il 16%) mentre l'errore è assai più rilevante per il Controllo e soprattutto per il Comportamento Passato. È possibile correggere le stime dei coefficienti di regressione ricavati dall'analisi tramite la procedura di correzione per l'attenuazione (vedi Barbaranelli e Natali, 2005).

Per analizzare la presenza di collinearità occorre selezionare "Test di collinearità" nella finestra di dialogo Statistiche (vedi *Figura 1.4.*). La selezione di questo parametro produrrà in output i due importanti indici della Tolleranza e del VIF: descriviamoli.

L'*indice di Tolleranza (Tolerance* T_i*)* viene utilizzato per stimare quanto una variabile indipendente è linearmente correlata alle altre variabili indipendenti. Questo parametro varia tra 0 e 1, indica la quantità di varianza di una variabile indipendente che non è spiegata dalle altre variabili indipendenti ed è uguale a: $T_i = (1 - R_i^2)$, dove R_i^2 è il coefficiente di determinazione ottenuto dalla regressione della variabile indipendente *i* sulle altre variabili indipendenti. Maggiore è l'indice di tolleranza, minore è la varianza che quella variabile indipendente condivide con le altre VI, maggiore è il contributo che essa può fornire nella spiegazione della variabile dipendente. Una variabile con un basso livello di tolleranza, invece, condivide molta varianza con le altre VI, quindi il contributo che può fornire nella spiegazione della variabile dipendente è solitamente più limitato. Un valore particolarmente basso di tolleranza (ad esempio, inferiore a .01) può risultare indicativo di variabili che rischiano di causare problemi computazionali nella stima dei coefficienti di regressione.

Il *Variance Inflaction Factor (VIF)* rappresenta il reciproco della tolleranza, ovvero, $VIF_i = 1/T_i = 1/(1 - R_i^2)$. Valori bassi del VIF indicano

Tabella 1.6. – Coefficienti di regressione e statistiche di collinearità (Tolleranza e VIF).

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.	Statistiche di collinearità	
		B	Errore std.	Beta			Tolleranza	VIF
1	(Costante)	-4,450	,991		-4,492	,000		
	ATT	,141	,021	,389	6,689	,000	,606	1,650
	NS	,284	,078	,199	3,621	,000	,675	1,482
	CONTCO	,291	,093	,156	3,120	,002	,814	1,229
	COMPAS	,351	,068	,284	5,177	,000	,683	1,465

a. Variabile dipendente: INT

Tabella 1.7. – Diagnostiche di collinearità.

Diagnostiche di collinearità[a]

Modello	Dimensione	Autovalore	Indice di collinearità	Variabilità spiegata				
				(Costante)	ATT	NS	CONTCO	COMPAS
1	1	4,713	1,000	,00	,00	,00	,00	,01
	2	,241	4,424	,00	,00	,00	,00	,76
	3	,026	13,438	,03	,00	,74	,16	,04
	4	,012	20,073	,01	,71	,26	,40	,00
	5	,008	24,379	,95	,28	,00	,43	,19

a. Variabile dipendente: INT

bassa collinearità, valori alti elevata collinearità. Valori tra 5 e 10 sono indicativi di forte collinearità.

La *Tabella 1.6.* presenta l'output (tabella dei coefficienti di regressione) relativo al nostro esempio in cui sono stati richiesti gli indici di collinearità. Per ottenere questo output, nella finestra di dialogo principale (*Figura 1.2.*) è stata specificata la variabile INT come dipendente, e le variabili ATT, NS, CONTCO, e COMPAS come indipendenti inserite in un unico blocco. Dagli indici di tolleranza e VIF possiamo osservare che nessuna delle variabili considerate nel nostro esempio presenta problemi di collinearità: infatti più del 60% della varianza di ogni variabile non risulta in comune con le altre variabili indipendenti in analisi.

Nella tabella "Diagnostiche di Collinearità" (*Tabella 1.7.*) vengono presentati indici aggiuntivi di collinearità. In particolare, gli autovalori sono ottenuti effettuando l'analisi delle componenti principali (vedi capitolo 2) della matrice dei prodotti scalari tra le variabili indipendenti, e danno un'indicazione della correlazione tra le variabili indipendenti. Se molti autovalori sono prossimi a 0 le variabili sono fortemente correlate. L'indice di collinearità (*condition index*) deriva dagli autovalori: se è compreso tra 15 e 30 indica possibili problemi di collinearità, se è mag-

giore di 30 la collinearità è grave (questo indice, comunque, è meno importante rispetto al VIF e alla tolleranza, Pedhazur, 1997). L'ultimo indice è la proporzione di variabilità spiegata dalle componenti principali ("Dimensioni") associate ad ogni autovalore.

La collinearità è un problema se una componente ("Dimensione") con un elevato indice di collinearità contribuisce in maniera sostanziale alla varianza di due o più variabili. Nei risultati della *Tabella 1.7.* nessuna delle componenti con indice di collinearità maggiore di 15 contribuisce in modo sostanziale a più di 2 variabili, non sono dunque presenti problemi di collinearità.

Effettuiamo ora la verifica delle assunzioni sui residui. Consideriamo il modello di regressione utilizzato per richiedere le statistiche di collinerarità e nella finestra di dialogo "Grafici" (*Figura 1.5.*) selezioniamo i residui standardizzati (*ZRESID) come variabile per l'asse delle X e i punteggi teorici standardizzati (*ZPRED) come variabile per l'asse delle Y, quindi selezioniamo le opzioni per i grafici: "Grafici dei residui standardizzati" (Istogramma, Grafico di probabilità normale) e "Produci tutti i grafici parziali".

Dalla *Tabella 1.8.* è facile verificare che la media dei residui (standardizzati o grezzi) è uguale a 0, quindi questa assunzione è verificata. Le *Figure 1.9.* e *1.10.* presentano i grafici delle distribuzioni dei residui: sia l'istogramma, sia il grafico dei percentili (P-P plot) sono compatibili con la normale. Quindi anche questa assunzione è verificata.

Il grafico riportato nella *Figura 1.11.* rappresenta il diagramma di dispersione dei residui rispetto ai valori teorici. In questo caso non sembrano presenti violazioni della linearità, mentre l'omoschedasticità potrebbe non essere garantita. La regressione però è molto robusta alla violazione di questa assunzione, quindi non dobbiamo preoccuparci. I plot parziali (*Figura 1.12.*) non sembrano evidenziare particolari problemi. Per un quadro esemplificativo di scatterplot residui/attesi che possono rappresentare schemi di riferimento per l'interpretazione dei risultati si vedano Barbaranelli (2003) e Tabachnick e Fidell (1989).

Per approfondire ulteriormente l'analisi dei residui dalla finestra di dialogo Statistiche (*Figura 1.4.*) selezioniamo le opzioni "*Durbin-Watson*" e "*Diagnostiche per casi*".

5.

L'ANALISI DEI CLUSTER

Premessa

Lo scopo dell'analisi dei cluster, o *analisi dei raggruppamenti* o dei *gruppi*, è quello di classificare casi sui quali sono state misurate differenti variabili in un numero inferiore e relativamente contenuto di classi, dette appunto *cluster* o *gruppi*. Il risultato è rappresentato dalla individuazione di una "tipologia" tramite la quale il ricercatore è in grado di classificare i casi (o meno spesso le variabili) oggetto di analisi.

Il *punto di partenza* è rappresentato da una matrice di indici di distanze o di similarità (solo nei cosiddetti metodi *gerarchici*) o da un file di dati grezzi Casi × Variabili (per i cosiddetti metodi *non gerarchici*). Il *punto di arrivo* è rappresentato da una gerarchia di partizioni (solo nei metodi gerarchici), dall'informazione sul gruppo di appartenenza per ogni caso, e da una serie di indici e rappresentazioni grafiche che consentono di valutare la bontà della soluzione.

I gruppi *non* sono determinati a priori. Gli individui vengono assegnati ai gruppi in modo tale che i casi all'interno di un cluster siano caratterizzati da un elevato grado di "*similarità*", mentre i cluster devono essere "*relativamente distinti*" l'uno dall'altro.

In generale è possibile dividere le procedure di cluster analysis in due grandi categorie.

a) Metodi *gerarchici*: sono procedimenti che generano un insieme di partizioni ordinate gerarchicamente, ovvero nelle quali ogni cluster ad un qualunque livello fa parte di un cluster più ampio ad un livello successivo, il quale a sua volta appartiene ad un cluster ancora più ampio ad un livello ulteriore. L'applicazione di un metodo gerarchico per classificare n unità darà origine a $n - 1$ partizioni, ordinate gerarchicamente.

b) Metodi *non gerarchici* (o "a partizioni ripetute" o "per suddivisioni iterative"): sono procedure che generano un'unica partizione. I metodi non gerarchici di clustering sono basati sull'attribuzione di un insieme di oggetti ad un *numero prefissato* di gruppi. L'appartenenza dei casi ai gruppi però *non* è nota a priori: il ricercatore deve decidere inizialmente il numero di gruppi presenti nella soluzione, l'algoritmo utilizzato assegnerà i casi ai diversi gruppi in base ad un criterio statistico. Si tratta di metodi che vengono applicati a matrici di dati Casi × Variabili e non a matrici di distanze. Possono consentire al ricercatore una maggiore flessibilità dei metodi gerarchici. Tali metodi, riservati a dati di tipo quantitativo, sono basati fondamentalmente sulla classificazione degli oggetti in base alla loro *distanza euclidea* dai centroidi dei gruppi.

Ogni procedura di cluster analysis *gerarchica* procede attraverso i seguenti passaggi:

1) identificazione delle variabili da utilizzare per la classificazione;
2) selezione di una misura di distanza tra unità;
3) selezione di una tecnica di raggruppamento delle unità;
4) identificazione del numero di gruppi entro i quali ripartire le unità;
5) valutazione ed interpretazione della soluzione.

Ogni procedura di cluster analysis *non-gerarchica* procede attraverso i seguenti passaggi:

1) identificazione delle variabili da utilizzare per la classificazione: il ricercatore sceglie le variabili sulle quali si basa il calcolo della distanza euclidea;
2) identificazione del numero di gruppi entro i quali ripartire le unità;
3) individuazione dei "semi" della partizione, ovvero dei casi che costituiscono inizialmente i gruppi, rendendo così possibile l'avvio del processo di classificazione;
4) valutazione ed interpretazione della soluzione.

Nelle pagine che seguono presenteremo la procedura per l'analisi dei cluster gerarchica e non gerarchica con SPSS, e vedremo come tramite tale procedura è possibile affrontare i diversi passi in cui esse si articolano.

5.1. Una descrizione della procedura per l'analisi dei cluster gerarchica

Per eseguire analisi dei cluster gerarchica occorre scegliere dal menù Analizza ⇒ Classificazione ⇒ Cluster gerarchica (*Figura 5.1.*).

La procedura "Cluster gerarchica" effettua l'analisi dei gruppi gerarchica "per agglomerazione" (vedi Barbaranelli, 2003). Si utilizza un algoritmo che inizialmente prevede tanti gruppi quanti sono i casi (o le variabili), e unisce i gruppi tra di loro fino ad ottenere un unico gruppo. Le misure di distanze o di similarità sono generate dalla procedura *Prossimità* che viene richiamata automaticamente dalla procedura Cluster gerarchica.

Se si raggruppano casi (ovvero i soggetti) bisogna selezionare almeno una variabile numerica. Se si raggruppano variabili si devono selezionare almeno tre variabili numeriche. Le variabili sulle quali si basa l'analisi vengono selezionate e inserite nel box "*Variabili*" (vedi *Figura 5.2.*).

È possibile identificare i soggetti tramite una variabile alfanumerica, nel box "*Etichetta i casi in base a*". Le alternative dell'opzione "*Raggruppa*" consentono di formare sia gruppi di casi, sia gruppi di variabili. Nel primo caso si ha la classica applicazione dell'analisi dei cluster che porta all'identificazione di tipologie di soggetti. Nel secondo caso si ha un'applicazione simile, per i risultati che si possono ottenere, all'analisi fattoriale. Può essere utile effettuare l'analisi dei cluster sulle variabili in alternativa all'analisi fattoriale quando il livello di misurazione delle variabili è

Figura 5.1. – Procedura Classificazione (Analisi dei cluster).

basso (es., variabili nominali con più di 2 categorie) o quando ci sono severe violazioni della normalità o quando il numero dei soggetti è particolarmente scarso. Le alternative dell'opzione "*Visualizza*" consentono di avere in output statistiche e/o grafici.

La *Figura 5.3.* evidenzia le diverse opzioni statistiche che possono essere richieste. In particolare le seguenti.

Programma di agglomerazione: mostra nel dettaglio le aggregazioni tra i diversi oggetti nella gerarchia delle partizioni, specificando ad ogni stadio quali cluster si accorpano, qual è il valore del coefficiente di agglomerazione (o di fusione) in quello stadio, qual è lo stadio in cui i due cluster che ora si fondono erano stati generati, qual è lo stadio successivo in cui il cluster appena formato si fonderà con un altro cluster. Tramite la tabella che contiene il programma di agglomerazione è possibile ripercorrere il processo di fusione tra i cluster, dal punto di origine in cui ogni caso rappresenta un cluster distinto, al punto di arrivo in cui tutti i casi sono inclusi in un unico cluster.

Matrice delle distanze: è la matrice che contiene gli indici di distanza o di similarità tra le unità statistiche che vengono raggruppate (casi o variabili).

Cluster di appartenenza: mostra il cluster in cui ciascun caso viene assegnato ad uno o più stadi nelle diverse partizioni scelte dal ricercatore. Le opzioni disponibili sono le seguenti.

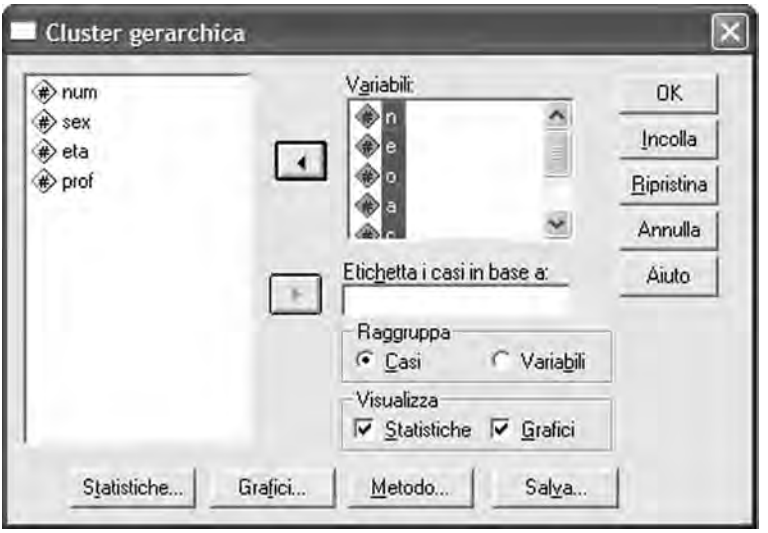


Figura 5.2. – Finestra di dialogo principale per la procedura Cluster gerarchica.

Assente: non fornisce nessuna statistica.

Soluzione unica: richiede l'appartenenza di ogni caso ad una singola partizione (ovvero, ad un singolo stadio).

Intervallo di soluzioni: richiede l'appartenenza di ciascun caso ad ogni partizione compresa nell'intervallo "Da" "a", e quindi agli stadi corrispondenti all'interno della gamma prescelta di soluzioni.

La *Figura 5.4*. evidenzia le diverse opzioni di visualizzazione grafica dei risultati che possono essere richieste. Esse sono le seguenti.

Dendrogramma: mostra il dendrogramma della gerarchia delle soluzioni. Si tratta come noto di un diagramma ad albero che dà un'immagine delle relazioni specificate fra gli oggetti che vengono raggruppati facendo riferimento ad una scala di distanza. Si costruisce partendo dai *rami*, corrispondenti ai casi iniziali, ognuno dei quali rappresenta un gruppo distinto, sino ad arrivare, per fusioni successive, ad un unico ramo finale (*radice*) che rappresenta appunto lo stadio finale in cui tutti i casi sono racchiusi in un unico cluster. Le "fusioni" sono rappresentate dai punti in cui due rami si congiungono. Come noto, i dendrogrammi possono essere utilizzati per avere informazioni circa il numero appropriato di cluster da mantenere: questo, in particolare, si ricava sezionando l'albero all'altezza del massimo salto tra livelli di distanza ai quali sono avvenute le fusioni, in modo da ottenere il minor numero di gruppi con omogeneità interna massima (ovvero, composti da individui simili).

Stalattite: mostra un grafico a stalattite, che include tutti i cluster o una gamma specificata di cluster. Questi grafici forniscono informazioni sul modo in cui i casi sono riuniti in cluster ad ogni stadio dell'analisi

Figura 5.3. – Finestra di dialogo Cluster gerarchica - Statistiche.

gerarchica e consentono di visualizzare esplicitamente quanti e quali casi si trovano nello stesso cluster ad un determinato stadio del processo aggregativo.

Le opzioni analitiche fondamentali che il ricercatore deve scegliere vengono selezionate dalla finestra di dialogo che si apre con un clic del mouse sul pulsante "*Metodo*" nella finestra di dialogo principale (vedi *Figura 5.5.*). Le voci selezionabili da questa finestra di dialogo sono molteplici.

Metodo di raggruppamento. Serve per specificare il metodo (ovvero il "criterio di fusione") da utilizzare per formare i cluster. Le alternative disponibili sono le seguenti.

Legame medio fra i gruppi (*Bavarage*): la distanza tra due gruppi è uguale alla media delle distanze tra ogni coppia di elementi appartenenti a gruppi differenti.

Legame medio entro gruppi (*Wavarage*): la distanza tra due gruppi è uguale alla media delle distanze tra ogni coppia di elementi, incluse le coppie di elementi che appartengono allo stesso gruppo.

Del vicino più vicino, o *legame singolo*: la distanza tra due gruppi è uguale alla distanza tra gli elementi dei due gruppi che sono più vicini.

Del vicino più lontano, o *legame completo*: la distanza tra due gruppi è uguale alla distanza tra gli elementi dei due gruppi che sono più lontani.

Centroide: la distanza tra due gruppi è uguale alla distanza tra i centroidi dei gruppi.

Figura 5.4. – Finestra di dialogo Cluster gerarchica - Grafici.

Mediana: la distanza tra due gruppi è uguale alla distanza tra i vettori dei valori mediani dei gruppi.
Ward o *Metodo di Ward*: la distanza tra due gruppi è una funzione delle devianze tra i due gruppi.

Per una definizione formale dei criteri di fusione considerati si vedano Barbaranelli (2003), Sadocchi (1987) e il documento sull'analisi dei cluster nel sito *www.lededizioni.com*.

Misura. Consente di specificare la misura di similarità o distanza da usare nelle analisi. È opportuno utilizzare misure di distanza appropriate al livello di misurazione delle variabili utilizzate. SPSS utilizza molte misure di distanza per una definizione delle quali si rimanda al testo di Barbaranelli (2003), alle funzioni di aiuto *on line* del programma, e al documento sull'analisi dei cluster nel sito *www.lededizioni.com*.

Trasforma valori. Consente di trasformare i valori dei dati sia per i casi sia per le variabili, prima di effettuare il calcolo delle prossimità. L'unità di misura delle variabili è importante: differenze in questo aspetto possono influenzare la soluzione. Se le variabili utilizzate mostrano grandi

Figura 5.5. – Finestra di dialogo Cluster gerarchica - Metodo.

differenze nelle unità di misura (ad esempio, una variabile varia da 1 a 5 e l'altra varia da 20 a 40) è conveniente che siano trasformate in modo tale che la loro unità di misura sia omogenea.

Per la trasformazione dei valori sono disponibili le seguenti alternative:

Punteggi z: è la normale procedura di standardizzazione, che genera punteggi standard *z*, ovvero: $z = (x\text{-Media})/\text{Deviazione Standard}$;

Ampiezza massima di 1: la procedura divide il valore della variabile per il valore massimo, in modo tale che il valore massimo ottenibile sia 1, ovvero: $y = x/\text{Max}$;

Media di 1: la procedura standardizza i valori in modo che la loro media sia uguale a 1. I valori di una variabile sono divisi per la media della variabile, ovvero: $y = x/\text{Media}$;

Deviazione standard di 1: i valori di una variabile sono divisi per la deviazione standard, ovvero: $y = x/\text{Deviazione Standard}$;

Intervallo da –1 a 1: la procedura trasforma i valori in modo che l'intervallo di variazione sia compreso tra –1 e 1. I valori di una variabile sono divisi per l'intervallo di variazione della variabile, ovvero: $y = x/(\text{Max} - \text{Min})$;

Intervallo da 0 a 1: sottrae dal valore di una variabile il valore minimo e divide il risultato per l'intervallo, ovvero: $y = (x - \text{Min})/(\text{Max} - \text{Min})$.

Trasforma misure. Consente di trasformare i valori generati dalle misure di distanza. Si applica dunque dopo che è stato effettuato il calcolo delle distanze. Sono disponibili le seguenti opzioni.

– *Valori assoluti:* utilizza i valori assoluti delle distanze. Viene utilizzata quando interessa solo l'ampiezza della relazione, e non il segno.
– *Cambia segno:* trasforma i valori di similarità in dissimilarità, o viceversa. Viene utilizzata per invertire l'ordinamento delle distanze.

Figura 5.6. – Finestra di dialogo Cluster gerarchica - Salva nuove variabili.

– *Riscala all'intervallo tra 0 e 1:* trasforma le distanze prima sottraendo il valore più piccolo e poi dividendo per l'intervallo di variazione.

È possibile salvare alcuni dei risultati ottenuti dall'analisi dei cluster tramite la finestra di dialogo nella *Figura 5.6.* che si apre con un clic del mouse sul pulsante "*Salva*" nella finestra di dialogo principale. In questa finestra alla voce "*Cluster di appartenenza*" vengono specificate le opzioni per salvare l'appartenenza dei cluster nel caso di singole soluzioni o di una gamma di soluzioni. Le variabili salvate possono poi essere usate in analisi successive, ad esempio per esplorare le differenze tra i gruppi su qualsiasi variabile. Le opzioni previste sono:

Assente: non crea nuove variabili;

Soluzione unica: crea una nuova variabile che mostra l'appartenenza al cluster di ogni caso in una determinata soluzione;

Intervallo di soluzioni: crea nuove variabili che mostrano l'appartenenza al cluster di ogni caso ad ogni stadio tra quelli indicati nelle caselle "Da" e "a", che devono contenere numeri interi.

5.2. Un esempio completo di cluster analysis gerarchica

Presentiamo ora un esempio di cluster analysis gerarchica facendo riferimento alle opzioni illustrate nelle pagine precedenti. Si tratta di un'analisi effettuata sulle 5 variabili nel file "**cluster.sav** ". Le 5 variabili indipendenti analizzate rappresentano i punteggi ottenuti dai soggetti sulle 5 scale principali del questionario NEO-PI. Le 5 scale misurano altrettanti fattori di personalità: il Nevroticismo (N), l'Estroversione (E), l'Apertura all'Esperienza (O), la Coscienziosità (A), e l'Amicalità o Amabilità (A). Il file analizzato contiene i punteggi relativi a 50 soggetti: si tratta di un sotto-campione estrapolato da un campione più grande utilizzato in uno studio pubblicato sull'*European Journal of Personality* (Barbaranelli, 2002). Su queste variabili è possibile effettuare i passi delineati all'inizio di questo capitolo, attraverso i quali si dipana il processo dell'analisi dei cluster gerarchica.

Identificazione delle variabili da utilizzare per la classificazione. È un passo piuttosto semplice: il ricercatore decide quali saranno le variabili che il programma utilizzerà per calcolare le distanze tra i casi (o tra le variabili, nel caso in cui si raggruppino le variabili) che consentiranno di

definire la matrice di distanze o di prossimità, punto di partenza dell'analisi gerarchica. Nel nostro caso le variabili indipendenti sulle quali si baserà la classificazione sono i 5 punteggi nei 5 tratti di personalità misurati dal NEO-PI. Nella finestra di dialogo iniziale basterà dunque selezionare le variabili chiamate *n*, *e*, *o*, *a*, *c*, nel box delle variabili attive.

Selezione di una misura di distanza tra unità. Si tratta di selezionare un indice di distanza o di prossimità per calcolare appunto le distanze tra i casi che vengono raggruppati e quindi costruire così la matrice di distanze che verrà sottoposta ad analisi. Sono numerosi gli indici che SPSS consente di scegliere, in funzione del livello di misura delle variabili. Nel nostro caso abbiamo variabili misurate al livello degli intervalli equivalenti, quindi possiamo utilizzare senza problemi l'opzione di default del programma, ovvero la distanza euclidea al quadrato. Si ricorda che l'indice di distanza va specificato nella finestra di dialogo "Metodo" (vedi *Figura 5.5.*).

Selezione di una tecnica di raggruppamento delle unità. Si tratta di scegliere quale sarà l'algoritmo che verrà usato per calcolare la matrice delle distanze dopo ogni aggregazione. I metodi differiscono in merito al modo in cui viene calcolata la distanza tra due gruppi ad ogni passo del processo gerarchico. Alcuni metodi (come quelli *di legame*) possono essere utilizzati indipendentemente dal livello di misurazione delle variabili, altri metodi (come il metodo di Ward e il metodo del centroide) possono essere applicati solo a variabili misurate per lo meno al livello degli intervalli equivalenti. Nel nostro caso possiamo scegliere qualsiasi metodo, poiché le nostre variabili sono misurate al livello degli intervalli. Tra i diversi metodi scegliamo quello di *Ward*, poiché risulta particolarmente efficace nella classificazione di variabili che risultano misurate ai livelli più elevati. Anche la tecnica di raggruppamento viene scelta, come l'indice di distanza, dalla finestra di dialogo "Metodo". Nel nostro caso selezioniamo *Ward* come *metodo di raggruppamento* nell'apposito menù a tendina (vedi *Figura 5.5.*).

Identificazione del numero di gruppi entro i quali ripartire le unità. Si tratta della fase più importante della cluster analysis gerarchica: dalla gerarchia delle partizioni ne va scelta una che rappresenta la soluzione la quale consente di ripartire i soggetti nei gruppi individuati. Come sottolineato altrove (Barbaranelli, 2003), il ricercatore può utilizzare diversi metodi per arrivare a tale soluzione. Sono state, infatti, proposte diverse

soluzioni al problema di stabilire il numero di gruppi nella soluzione, al quale tuttavia non è stata data una risposta definitiva. Le principali soluzioni si basano sul calcolo di particolari indici o sull'ispezione visiva di grafici e tabelle. Purtroppo SPSS non consente il calcolo di alcuni degli indici più utilizzati (ad esempio, l'indice C, l'indice Gamma, la correlazione Punto-Biseriale, l'indice G+, la Somma degli errori *within clusters*, la statistica C). È inoltre piuttosto complicato calcolare a mano tali indici, pertanto si rimanda il lettore interessato a software che ne consentono il calcolo (ad esempio, SLEIPNER, di Bergman e El-Khouri, 1998). Per scegliere il numero di gruppi presenti nella soluzione di SPSS è possibile utilizzare il programma di agglomerazione e il dendrogramma. Descriviamo brevemente come si interpreta un programma di agglomerazione, e come è possibile utilizzarlo al fine di decidere la partizione ottimale. La *Tabella 5.1.* presenta il programma di agglomerazione relativo all'analisi dei 50 casi nel file **cluster.sav** tramite le opzioni descritte sopra.

La colonna "*Stadio*" indica i passi attraverso i quali procede l'algoritmo di cluster per aggregazione, dallo stadio in cui si fondono i primi 2 gruppi (stadio 1), allo stadio in cui abbiamo un solo gruppo costituito da tutti i soggetti (stadio 49). Lo stadio 0 è quello in cui ci sono tanti gruppi quanti sono i casi da raggruppare (nel nostro esempio, 50).

Le colonne "*Cluster accorpati*" indicano quali sono i cluster che si aggregano ad ogni stadio. Ad esempio, allo stadio 1 i cluster 19 e 26 si combinano in unico cluster che verrà identificato dal numero 19, ovvero il numero relativo al primo caso che si fonde nel nuovo gruppo.

La colonna "*Coefficienti*" riporta il valore della distanza tra i cluster che si aggregano. Ad esempio, i cluster 19 e 26 si combinano nello stadio 1 al livello di distanza .026. Il tipo di distanze è quello specificato dall'utente nella finestra di dialogo "Metodo".

Le colonne "*Stadio di formazione del cluster*" indicano il passo precedente in cui si sono formati i cluster che si aggregano in quel determinato stadio. Lo 0 viene utilizzato per indicare cluster che contengono un solo elemento, e che quindi si sono formati nello stadio 0, ovvero lo stadio in cui ogni cluster è composto da un solo caso. Ad esempio, nello stadio 25 si aggregano il cluster *29* e il cluster *32*: il primo si è formato allo stadio 6 dalla fusione dei casi 29 e 39, il secondo si è formato nello stadio 19 dalla fusione dei casi 32 e 33. Nello stadio 15, invece, si aggregano il cluster 13 e il cluster 14, entrambi composti da un solo caso, come si evince dalle colonne "*Stadio di formazione del cluster*" che presentano valore 0 per entrambi i casi.