IBM SPSS Categories 19

Jacqueline J. Meulman

Willem J. Heiser

SPSS Inc.

Note: Before using this information and the product it supports, read the general information under Notices a pag. 310.

Prefazione

IBM® SPSS® Statistics è un sistema completo per l'analisi dei dati. Il modulo aggiuntivo opzionale Categories include le tecniche di analisi aggiuntive descritte nel presente manuale. Il modulo aggiuntivo Categories deve essere usato con il modulo Core SPSS Statistics in cui è completamente integrato.

Informazioni su SPSS Inc., una società del gruppo IBM

SPSS Inc., una società del gruppo IBM, è fornitore leader mondiale nel settore del software e delle soluzioni per l'analisi predittiva. L'offerta completa dei prodotti dell'azienda (raccolta di dati, statistica, modellazione e distribuzione) consente di acquisire i comportamenti e le opinioni delle persone, prevedere i risultati delle future interazioni con i clienti ed elaborare questi dati integrando le analitiche nelle procedure aziendali. Le soluzioni SPSS Inc. consentono la gestione di attività interconnesse all'interno dell'intera organizzazione, con particolare attenzione alla convergenza di analitiche, architettura IT e procedure aziendali. Clienti commerciali, istituzionali e accademici di tutto il mondo si affidano alla tecnologia SPSS Inc. ottenendo un vantaggio competitivo in termini di attrazione, mantenimento e ampliamento della base clienti, riducendo al contempo frodi e rischi. SPSS Inc. è stata acquisita da IBM nell'ottobre 2009. Per ulteriori informazioni, visitare il sito *http://www.spss.com*.

Supporto tecnico

Ai clienti che richiedono la manutenzione, viene messo a disposizione un servizio di supporto tecnico. I clienti possono contattare il supporto tecnico per richiedere assistenza per l'utilizzo dei prodotti SPSS Inc. o per l'installazione di uno degli ambienti hardware supportati. Per il supporto tecnico, visitare il sito Web di SPSS Inc. all'indirizzo *http://support.spss.com* o contattare la filiale del proprio paese indicata nel sito Web all'indirizzo *http://support.spss.com/default.asp?refpage=contactus.asp*. Ricordare che durante la richiesta di assistenza sarà necessario fornire i dati di identificazione personali, i dati relativi alla propria società e il numero del contratto di manutenzione.

Servizio clienti

Per informazioni sulla spedizione o sul proprio account, contattare la filiale nel proprio paese, indicata nel sito Web all'indirizzo *http://www.spss.com/worldwide*. Tenere presente che sarà necessario fornire il numero di serie.

Corsi di formazione

SPSS Inc. organizza corsi di formazione pubblici e onsite che includono esercitazioni pratiche. Tali corsi si terranno periodicamente nelle principali città. Per ulteriori informazioni sui corsi, contattare la filiale nel proprio paese, indicata nel sito Web all'indirizzo *http://www.spss.com/worldwide*.

Pubblicazioni aggiuntive

I documenti *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion* e *SPSS Statistics: Advanced Statistical Procedures Companion*, scritti da Marija Norušis e pubblicati da Prentice Hall sono disponibili come materiale supplementare consigliato. Queste pubblicazioni descrivono le procedure statistiche nei moduli SPSS Statistics Base, Advanced Statistics e Regression. Utili sia come guida iniziale all'analisi dei dati che per applicazioni avanzate, questi manuali consentono di ottimizzare l'utilizzo delle funzionalità presenti nell'offerta IBM® SPSS® Statistics. Per ulteriori informazioni, inclusi contenuti delle pubblicazioni e capitoli di esempio, visitare il sito Web dell'autrice: *http://www.norusis.com*

Ringraziamenti

Le ottime procedure di scaling e la loro implementazione in IBM® SPSS® Statistics sono state sviluppate da DTSS (Data Theory Scaling System Group), un gruppo costituito da membri dei dipartimenti di scienze dell'educazione e psicologia della Facoltà di Scienze Sociali e Comportamentali dell'Università di Leiden.

Willem Heiser, Jacqueline Meulman, Gerda van den Berg e Patrick Groenen hanno partecipato allo sviluppo delle procedure iniziali del 1990. Jacqueline Meulman e Peter Neufeglise hanno contribuito allo sviluppo delle procedure per la regressione categorica, l'analisi delle rispondenze, l'analisi delle componenti principali categoriale e lo scaling multimediale. Inoltre, Anita van der Kooij ha contribuito in particolare allo sviluppo delle analisi CATREG, CORRESPONDENCE e ATPCA. Willem Heiser, Jacques Commandeur, Frank Busing, Gerda van den Berg e Patrick Groenen hanno partecipato allo sviluppo della procedura PROXSCAL. Frank Busing, Willem Heiser, Patrick Groenen e Peter Neufeglise hanno partecipato allo sviluppo della procedura PREFSCAL.

Contenuto

Parte I: Manuale dell'utente

1 Introduzione alle procedure di scaling ottimale per i dati categoriali 1

2 Regressione categoriale (CATREG) 15

Parte II: Esempi

Appendici

Parte I:
Manuale dell'utente

Capitolo

1

Introduzione alle procedure di scaling ottimale per i dati categoriali

Le procedure di Categorie utilizzano lo scaling ottimale per analizzare i dati che risulta difficile o impossibile analizzare tramite le procedure statistiche standard. Il capitolo illustra le operazioni eseguite da ciascuna procedura, le situazioni in cui ogni procedura è più adatta, le relazioni tra le procedure e le relazioni tra queste procedure e le corrispondenti procedure statistiche standard.

Nota: Queste procedure e la relativa implementazione in IBM® SPSS® Statistics sono state sviluppate dal Data Theory Scaling System Group (DTSS), composto dai membri dei dipartimenti di Didattica e Psicologia dalla Facoltà di Scienze sociali e del comportamento della Leiden University.

Informazioni sullo scaling ottimale

Il concetto alla base dello scaling ottimale è l'assegnazione di quantificazioni numeriche alle categorie di ciascuna variabile, che rende possibile l'utilizzo delle procedure standard per ottenere una soluzione sulle variabili quantificate.

I valori di scala ottimali vengono assegnati alle categorie di ciascuna variabile in base al criterio di ottimizzazione della procedura in uso. Diversamente dalle etichette originali delle variabili nominali o ordinali nell'analisi, questi valori di scala hanno proprietà metriche.

Nella maggioranza delle procedure del modulo Categories, la quantificazione ottimale per ciascuna variabile scalata viene ottenuta tramite un metodo iterativo detto dei **minimi quadrati alternati** nel quale, dopo essere state utilizzate per trovare una soluzione, le quantificazioni correnti vengono aggiornate utilizzando la soluzione stessa. Le quantificazione aggiornate vengono quindi utilizzate per trovare una nuova soluzione, impiegata a sua volta per aggiornare le quantificazioni, e così via, fino a raggiungere un criterio che indichi al processo di arrestarsi.

Motivi di utilizzo dello scaling ottimale

I dati categoriali sono spesso presenti nelle ricerche di marketing, nelle indagini di mercato e nella ricerca nelle scienze sociali e del comportamento. In effetti, molti ricercatori hanno a che fare quasi esclusivamente con dati categoriali.

Sebbene gli adattamenti della maggior parte dei modelli standard siano finalizzati specificatamente all'analisi dei dati categoriali, spesso non funzionano altrettanto bene per insiemi di dati che includono:

- un numero troppo limitato di osservazioni

- un numero troppo limitato di variabili
- un numero troppo limitato di valori per variabile

Tramite la quantificazione delle categorie, le tecniche di scaling ottimale evitano i problemi relativi a queste situazioni. Inoltre, sono particolarmente utili quando è necessario utilizzare tecniche speciali.

Anziché sulle stime dei parametri, l'interpretazione dell'output dello scaling ottimale si basa spesso su rappresentazioni grafiche. Le tecniche di scaling ottimale offrono eccellenti funzioni di analisi esplorativa, che integrano bene altri modelli IBM® SPSS® Statistics. Limitando l'obiettivo principale dell'analisi, la visualizzazione dei dati tramite scaling ottimale può costituire la base di un'analisi basata sull'interpretazione dei parametri del modello.

Livello di scaling ottimale e livello di misurazione

Questo concetto può generare molta confusione al primo utilizzo delle procedure del modulo Categories. Il livello specificato non è il livello di *misurazione* delle variabili, ma quello di *scala*. Il concetto è che le variabili da quantificare possono includere relazioni non lineari indipendentemente dalla modalità di misurazione.

Per quanto concerne Categories, esistono tre livelli fondamentali di misurazione:

- Il livello**nominale** implica che i valori di una variabile rappresentano categorie non ordinate. Esempi di variabili che possono essere nominali sono la regione, il codice postale, la religione e le categorie a scelta multipla.
- Il livello**ordinale** implica che i valori di una variabile rappresentano categorie ordinate. Tra gli esempi, le scale di atteggiamento corrispondenti a gradi di soddisfazione o fiducia e i punteggi di preferenza.
- Il livello **numerico** implica che i valori di una variabile rappresentino categorie ordinate con una metrica significativa, tale che i confronti fra le categorie siano appropriati. Esempi di variabili sono l'età espressa in anni o il reddito espresso in migliaia di Euro.

Ad esempio, si supponga che le variabili *regione*, *lavoro* ed *età* siano codificate come illustrato nella tabella seguente.

Tabella 1-1
Schema di codifica per regione, lavoro ed età

Regione		Lavoro		Età	
1	Nord	1	stagista	20	venti anni
2	Sud	2	commerciale	22	ventidue anni
3	Est	3	manager	25	venticinque anni
4	Ovest			27	ventisette anni

I valori illustrati rappresentano le categorie di ciascuna variabile. *Regione* sarà una variabile nominale. Esistono quattro categorie di *regioni*, senza ordinamento intrinseco. I valori da 1 a 4 rappresentano semplicemente le quattro categorie; lo schema di codifica è totalmente arbitrario. D'altro canto, si presume che *Lavoro* sia una variabile ordinale. Le categorie originali formano una progressione da stagista a manager. Maggiore è il codice numerico, maggiore il livello della

posizione lavorativa all'interno della scala aziendale. Tuttavia, sono note solo informazioni sull'ordinamento, mentre non ci sono dati sulla distanza tra categorie adiacenti. Al contrario, si può presumere che *età* sia un valore numerico. Nel caso di *età*, le distanze tra i valori sono intrinsicamente significative. La distanza tra 20 e 22 è la stessa esistente tra 25 e 27, mentre la distanza tra 22 e 25 è maggiore di entrambe le precedenti.

Selezione del livello di scaling ottimale

È importante comprendere che nessuna proprietà intrinseca di una variabile predefinisce automaticamente il livello di scaling ottimale da specificare per la variabile. È possibile esplorare i dati in qualsiasi modo, purché sia appropriato e faciliti l'interpretazione. Analizzando ad esempio una variabile di livello numerico a livello ordinale, l'utilizzo di una trasformazione non lineare può consentire una soluzione in un numero minore di dimensioni.

I due esempi che seguono illustrano come il livello "ovvio" di misurazione possa non corrispondere al livello di scaling ottimale migliore. Si supponga che una variabile ordini gli oggetti in gruppi di età. Sebbene l'età possa essere scalata come variabile numerica, è potenzialmente vero che, per le persone con meno di 25 anni, la relazione tra sicurezza ed età è positiva, mentre è negativa per le persone con più di 60 anni. In questo caso, può essere preferibile considerare l'età come una variabile nominale.

Sempre a titolo di esempio, una variabile che ordina le persone in base alle preferenze politiche è essenzialmente nominale. Tuttavia, se si ordinano i partiti politici da sinistra a destra, è possibile che si voglia che la quantificazione dei partiti rispetti quest'ordine, utilizzando un livello di analisi ordinale.

Anche se non esistono proprietà predefinite di una variabile che le attribuiscano esclusivamente un livello specifico, l'utente inesperto può fare riferimento ad alcune linee guida generali. Quando si utilizza la quantificazione nominale singola, normalmente non si conosce l'ordine delle categorie, ma si desidera applicarne uno tramite l'analisi. Se l'ordine delle categorie è noto, utilizzare la quantificazione ordinale. Se le categorie non sono ordinabili, utilizzare la quantificazione nominale multipla.

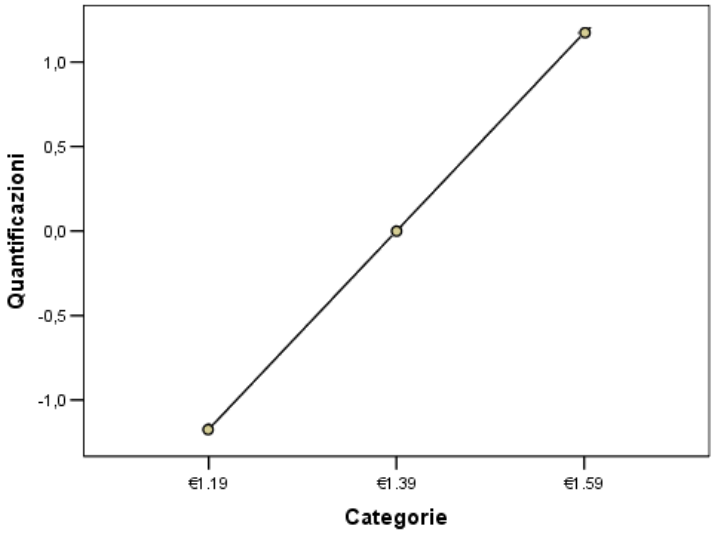
Grafici di trasformazione

I diversi livelli di scaling di ciascuna variabile applicano vincoli diversi alle quantificazioni. I grafici di trasformazione illustrano la relazione tra le quantificazioni e le categorie originali risultanti dal livello di scaling ottimale selezionato. Ad esempio, un grafico di trasformazione lineare viene generato quando una variabile viene considerata come numerica. Le variabili considerate come ordinali determinano la generazione di un grafico di trasformazione non decrescente. I grafici di trasformazione per le variabili considerate come nominali con forma a U (o l'inverso) visualizzano una relazione quadratica. Le variabili nominali possono inoltre generare grafici di trasformazione senza trend apparenti modificando completamente l'ordine delle categorie. La figura seguente mostra un grafico di trasformazione di esempio.

I grafici di trasformazione sono particolarmente adatti per determinare l'adeguatezza del livello di scaling ottimale selezionato. Se più categorie ricevono quantificazioni analoghe, la loro compressione in una categoria unica potrebbe essere giustificata. In alternativa, se una variabile considerata come nominale riceve quantificazioni che visualizzano un trend crescente, una trasformazione ordinale può generare un adattamento analogo. Se il trend è lineare, può essere

appropriato considerare la variabile come numerica. Tuttavia, se la compressione delle categorie o la modifica dei livelli di scaling è giustificata, l'analisi non si modificherà in modo significativo.

Figura 1-1
Grafico di trasformazione del prezzo (numerico)



Codici di categoria

Prestare attenzione nella codifica delle variabili categoriali, in quanto alcuni schemi di codifica possono generare output indesiderati o analisi incomplete. Gli schemi di codifica applicabili per la variabile *lavoro* sono visualizzati nella tabella seguente.

Tabella 1-2
Schemi di codifica alternativi per lavoro

	Schema			
Categoria	**T**	**P**	**O**	**D**
stagista	1	1	5	1
commerciale	2	2	6	5
manager	3	7	7	3

Alcune procedure di Categories richiedono la definizione dell'intervallo di ogni variabile utilizzata. Qualsiasi valore esterno all'interno viene considerato come mancante. Il valore di categoria minimo è sempre 1. Il valore di categoria massimo è specificato dall'utente. Questo valore non è il *numero* delle categorie per una variabile, ma il valore di categoria *massimo*. Ad esempio, nella tabella, il valore di categoria massimo per lo schema A è 3 e per lo schema B è 7, ma entrambi gli schemi codificano le stesse tre categorie.

L'intervallo delle variabili determina quali categorie saranno omesse dall'analisi. Qualsiasi categoria con codici esterni all'intervallo definito sarà omessa dall'analisi. Tuttavia, questo semplice metodo per escludere le categorie può determinare analisi indesiderate. Un errore nel determinare la categoria massima può determinare l'esclusione di categorie *valide* dall'analisi. Ad esempio definire per lo schema B il valore della categoria massima uguale a 3, significa che

a *lavoro* sono associate categorie codificate da 1 a 3; la categoria *manager* verrà considerata mancante. Poiché nessuna categoria è stata effettivamente codificata con il numero 3, la terza categoria nell'analisi non contiene nessun caso. Questa analisi sarebbe appropriata per omettere tutte le categorie manager. Tuttavia, per includere i manager, è necessario definire la categoria massima uguale a 7 e codificare i valori mancanti con valori superiori a 7 o inferiori a 1.

Per le variabili considerate come nominali o ordinali, l'intervallo delle categorie non influisce sui risultati. Per le variabili nominali, a essere significativa è solo l'etichetta, non il valore a essa associato. Per le variabili ordinali, l'ordine delle categorie viene mantenuto nelle quantificazioni; i valori di categoria non sono significativi. Tutti gli schemi di codifica risultanti nello stesso ordine di categoria avranno risultati identici. Ad esempio, i primi tre schemi nella tabella sono equivalenti,da un punto di vista funzionale, se *lavoro* viene analizzato a livello ordinale. L'ordine delle categorie è identico in questi schemi. Nello schema D, invece, la seconda e la terza categoria vengono invertite e i risultati generati sono diversi rispetto agli altri schemi.

Sebbene molti schemi di codifica per una variabile siano equivalenti da un punto di vista funzionale, schemi con piccole differenze tra i codici sono preferibili, perché i codici influiscono sulla quantità di output generata da una procedura. Tutte le categorie codificate con valori compresi tra 1 e il valore massimo definito dall'utente sono valide. Se una di tali categorie è vuota, le quantificazioni corrispondenti saranno mancanti di sistema o uguali a 0, in base alla procedura. Sebbene nessuna di queste assegnazioni influenzi le analisi, l'output viene generato per queste categorie. Di conseguenza, per lo schema B, *lavoro* ha quattro categorie che ricevono valori mancanti di sistema. Per lo schema C, sono inoltre presenti quattro categorie che ricevono indicatori di mancanti di sistema. Al contrario, per lo schema A non sono presenti quantificazioni mancanti di sistema. L'utilizzo di interi consecutivi come codici per le variabili considerate come nominali o ordinali determina una quantità molto minore di output senza influenzare i risultati.

Gli schemi di codifica per le variabili considerate come numeriche sono più limitati rispetto al caso di variabili considerate come nominali. Per tali variabili, le differenze tra categorie consecutive sono importanti. La tabella seguente mostra tre schemi di codifica per *età*.

Tabella 1-3
Schemi di codifica alternativi per età

	Schema		
Categoria	**T**	**P**	**O**
20	20	1	1
22	22	3	2
25	25	6	3
27	27	8	4

Qualsiasi ricodifica di variabili numeriche deve conservare le differenze tra le categorie. L'utilizzo dei valori originali è un metodo per assicurare la conservazione delle differenze. Tuttavia, può determinare la presenza di indicatori mancanti di sistema in molte categorie. Ad esempio, si supponga che lo schema A utilizzi i valori osservati originali. Per tutte le procedure di Categories, fatta eccezione per l'analisi delle corrispondenze, il valore di categoria massimo è 27 e il minimo 1. Le prime 19 categorie sono vuote e ricevono indicatori mancanti di sistema. L'output può diventare rapidamente piuttosto complesso se la categoria massima è molto superiore a 1 e ci sono molte categorie vuote tra 1 e il valore massimo.

Per ridurre la quantità di output è possibile eseguire la ricodifica. Tuttavia, nel caso di una variabile numerica, non utilizzare lo strumento Ricodifica automatica. La codifica in interi consecutivi determina differenze di 1 tra tutte le categorie consecutive; di conseguenza, tutte le quantificazioni saranno distanziate in modo uniforme. Le caratteristiche metriche ritenute importanti nella considerazione di una variabile come numerica vengono eliminate quando si esegue la ricodifica in interi consecutivi. Ad esempio, lo schema C nella tabella corrisponde alla ricodifica automatica di *età*. La differenza tra le categorie 22 e 25 è passata da tre a uno e le quantificazioni rifletteranno quest'ultimo valore.

Uno schema di ricodifica alternativo che conservi le differenze tra le categorie consiste nel sottrarre il valore di categoria minore da ciascuna categoria e nell'aggiungere uno a ciascuna differenza. Lo schema B è generato da questa trasformazione. Il valore di categoria minore, 20, è stato sottratto da ciascuna categoria e a ogni risultato è stato aggiunto 1. I codici trasformati hanno un valore minimo di 1 e tutte le differenze sono identiche ai dati originali. Il valore di categoria massimo è ora uguale a otto, e tutte le quantificazioni uguali a zero precedenti alla prima quantificazione diversa da zero vengono eliminate. Tuttavia, le quantificazioni diverse da zero corrispondenti a ciascuna categoria risultante dallo schema B sono identiche alle quantificazioni risultanti dallo schema A.

Procedura ottimale per l'applicazione

Le tecniche integrate in quattro di queste procedure (Analisi delle corrispondenze, Analisi delle corrispondenze multiple, Analisi delle componenti principali categoriale e Analisi della correlazione canonica non lineare) appartengono all'area generale dell'analisi dei dati multivariati nota come **riduzione dimensionale**. Le relazioni tra variabili vengono cioè rappresentate nel minor numero di dimensioni possibile —(due o tre)—. Questo consente di descrivere le strutture o i modelli delle relazioni che sarebbe troppo complesso comprendere appieno nella loro complessità e ricchezza originali. Nelle applicazioni per le ricerche di mercato, queste tecniche possono rappresentare una forma di **segmentazione percettiva**. Uno dei principali vantaggi di queste procedure consiste nel fatto che dispongono i dati con diversi livelli di scaling ottimale.

La regressione categoriale descrive la relazione tra una variabile di risposta categoriale e una combinazione di variabili predittore categoriali. L'influenza di ciascuna variabile predittore sulla variabile di risposta è descritta dal peso della regressione corrispondente. Come nelle altre procedure, è possibile analizzare i dati con diversi livelli di scaling ottimale.

Lo scaling e l'unfolding multidimensionale descrivono le relazioni tra gli oggetti in uno spazio dimensionale ridotto utilizzando le distanze tra gli oggetti.

Seguono alcune linee guida per ciascuna delle procedure:

- Utilizzare la regressione categoriale per la previsione dei valori di una variabile dipendente categoriale da una combinazione di variabili dello stesso tipo.
- Utilizzare l'analisi delle componenti principali categoriale per tenere conto dei modelli di variazione in un singolo insieme di variabili con livelli di scaling ottimale misti.
- Utilizzare l'analisi della correlazione canonica non lineare per valutare il grado di correlazione tra due o più insiemi di variabili con livelli di scaling ottimale misti.

- Utilizzare l'analisi delle corrispondenze per analizzare le tavole di contingenza a due vie o i dati che è possibile esprimere in una tavola a due vie, ad esempio dati relativi alla marca preferita o di scelta sociometrica.
- Utilizzare l'analisi delle corrispondenze multiple per analizzare una matrice di dati multivariati categoriali quando non si desiderano avanzare ipotesi più forti sul fatto che tutte le variabili siano analizzate a livello nominale.
- Utilizzare lo scaling multidimensionale per analizzare i dati di distanza per individuare una rappresentazione dei minimi quadrati di un insieme di oggetti in uno spazio dimensionale ridotto.
- Utilizzare l'unfolding multidimensionale per analizzare i dati di distanza per individuare una rappresentazione dei minimi quadrati di due insiemi di oggetti in uno spazio dimensionale ridotto.

Regressione categoriale

La regressione categoriale è la più adatta quando l'obiettivo dell'analisi è prevedere una variabile (di risposta) dipendente da un insieme di variabili (predittore) indipendenti. Come in tutte le procedure di scaling ottimale, i valori di scala vengono assegnati a ciascuna categoria di ogni variabile, in modo che i valori siano ottimali rispetto alla regressione. La soluzione di una regressione categoriale massimizza la correlazione quadratica tra la risposta trasformata e la combinazione ponderata dei predittori trasformati.

Relazione con altre procedure di Categories. La regressione categoriale con scaling ottimale è paragonabile all'analisi della correlazione canonica con scaling ottimale con due insiemi, di cui uno contiene solo la variabile dipendente. In quest'ultimo caso, la similarità degli insiemi deriva dal confronto di ciascun insieme con una variabile sconosciuta compresa in un qualsiasi punto all'interno di tutti gli insiemi. Nella regressione categoriale, la similarità della risposta trasformata e la combinazione lineare dei predittori trasformati vengono valutate direttamente.

Relazione con le tecniche standard. Nella regressione lineare standard, le variabili categoriali possono essere ricodificate come variabili indicatore oppure considerate come variabili a livello di intervallo. Nel primo caso, il modello include un'intercetta e una inclinazione separate per ciascuna combinazione di livelli delle variabili categoriali. Questo determina un numero elevato di parametri da interpretare. Nel secondo caso, per ciascuna variabile viene stimato solo un parametro. Tuttavia, la natura arbitraria delle codifiche di categoria rende impossibile generalizzare.

Se alcune delle variabili non sono continue, è possibile utilizzare analisi alternative. Se la risposta è continua e i predittori sono categoriali, viene spesso utilizzata l'analisi della varianza. Se la risposta è categoriale e i predittori sono continui, può essere adatta la regressione logistica o l'analisi discriminante. Se la risposta e i predittori sono entrambi categoriali, vengono spesso utilizzati il modelli loglineari.

La regressione con scaling ottimale offre tre livelli di scaling per ciascuna variabile. Le combinazioni di questi livelli possono tenere conto di un'ampia gamma di relazioni non lineari, per le quali un singolo metodo "standard" sia inadatto. Di conseguenza, lo scaling ottimale offre maggiore flessibilità rispetto agli approcci standard e una complessità aggiuntiva minima.

Inoltre, le trasformazioni non lineari dei predittori in genere riducono le dipendenze tra i predittori. Se si confrontano gli autovalori della matrice di correlazione per i predittori con gli autovalori della matrice di correlazione per i predittori con scaling ottimale, quest'ultimo insieme sarà generalmente meno variabile del primo. In altre parole, nella regressione categoria, lo scaling ottimale riduce gli autovalori maggiori della matrice di correlazione dei predittori e aumenta gli autovalori minori.

analisi Componenti principali categoriale

L'analisi delle componenti principali categoriale è la più adatta per tenere conto dei modelli di variazione in un singolo insieme di variabili con livelli di scaling ottimale misti. Questa tecnica tenta di ridurre la dimensione di un insieme di variabili tenendo conto al contempo della maggiore variazione possibile. I valori di scala vengono assegnati a ciascuna categoria di ogni variabile, in modo che i valori siano ottimali rispetto alla soluzione delle componenti principali. Gli oggetti nell'analisi ricevono i punteggi delle componenti in base ai dati quantificati. I grafici dei punteggi delle componenti rivelano modelli tra gli oggetti nell'analisi e possono segnalare oggetti anomali nei dati. La soluzione di un'analisi delle componenti principali categoriale massimizza le correlazioni dei punteggi degli oggetti con ciascuna delle variabili quantificate, per il numero delle componenti (dimensioni) specificate.

Un'applicazione importante di questa analisi è l'esame dei dati relativi alle preferenze, in cui i rispondenti classificano o valutano un numero di item in ordine di preferenza. Nella normale configurazione dei dati IBM® SPSS® Statistics, le righe sono valori individuali, le colonne misure per gli item e i punteggi tra le righe i punteggi di preferenza (ad esempio su una scala da 0 a 10); di conseguenza, i dati sono condizionali per le righe. Per i dati di preferenza, è possibile considerare i valori individuali come variabili. Utilizzando la procedura Trasponi è possibile trasporre i dati. I predittori diventano le variabili e tutte le variabili sono dichiarate ordinali. Non esistono controindicazioni all'utilizzo di più variabili che oggetti in CATPCA.

Relazione con altre procedure di Categories. Se tutte le variabili vengono dichiarate nominali multiple, l'analisi dei componenti principali categoriale genera un'analisi equivalente a un'analisi delle corrispondenze multiple eseguita sulle stesse variabili. Di conseguenza, l'analisi delle componenti principali categoriale può essere considerata un tipo di analisi delle corrispondenze multiple, in cui alcune variabili vengono dichiarate ordinali o numeriche.

Relazione con le tecniche standard. Se tutte le variabili sono scalate a livello numerico, l'analisi delle componenti principali categoriale equivale all'analisi delle componenti principali standard.

Più in generale, l'analisi delle componenti principali categoriale è un'alternativa al calcolo delle correlazioni tra scale non numeriche e all'analisi di queste ultime attraverso un approccio di analisi fattoriale o delle componenti principali standard. Un utilizzo non attento della normale correlazione di Pearson come misura dell'associazione per i dati ordinali può portare a distorsioni significative nella stima delle correlazioni.

Analisi della correlazione canonica non lineare (OVERALS)

L'"analisi della correlazione canonica non lineare è una procedura estremamente generale con numerose applicazioni diverse. L'obiettivo è l'analisi delle relazioni tra due o più insiemi di variabili anzichè tra le variabili, come avviene nell'analisi delle componenti principali. Ad

esempio, si supponga di avere due insiemi di variabili, uno dei quali include item di background demografico in un insieme di rispondenti e il secondo le risposte a un insieme di item di atteggiamento. I livelli di scaling nell'analisi possono essere una qualsiasi combinazione dei livelli ordinale, numerico e nominale. L'analisi della correlazione canonica dello scaling ottimale determina la similarità tra gli insiemi confrontando contemporaneamente le variabili canoniche di ogni insieme con un insieme intermedio di punteggi assegnati agli oggetti.

Relazione con altre procedure di Categories. Se sono presenti due o più insiemi di variabili con una sola variabile per insieme, l'analisi della correlazione canonica dello scaling ottimale equivale all'analisi delle componenti principali dello scaling ottimale. Se tutte le variabili in un'analisi in cui ogni insieme include un'unica variabile sono nominali multiple, l'analisi della correlazione canonica dello scaling ottimale equivale all'analisi delle corrispondenze multiple. Se sono presenti due o più insiemi di variabili, uno dei quali include una sola variabile, l'analisi della correlazione canonica dello scaling ottimale equivale alla regressione categoriale con scaling ottimale.

Relazione con le tecniche standard. L'analisi della correlazione canonica standard è una tecnica statistica che individua una combinazione lineare di un insieme di variabili e una combinazione lineare di un secondo insieme di variabili con la massima correlazione. Dato questo insieme di correlazioni lineari, l'analisi della correlazione canonica è in grado di individuare insiemi indipendenti successivi di combinazioni lineari, detti variabili canoniche, fino a un numero massimo pari al numero delle variabili nell'insieme più piccolo.

Se sono presenti due o più insiemi di variabili nell'analisi e tutte le variabili sono definite come numeriche, l'analisi della correlazione canonica dello scaling ottimale equivale all'analisi della correlazione canonica standard. Sebbene IBM® SPSS® Statistics non includa una procedura di analisi della correlazione canonica, molte delle statistiche rilevanti possono essere ottenute tramite un'analisi della varianza multivariata.

L'analisi della correlazione canonica dello scaling ottimale ha svariate altre applicazioni. Se sono presenti due insiemi di variabili uno dei quali include una variabile nominale dichiarata come nominale singola, i risultati dell'analisi possono essere interpretati in modo analogo a quelli di un'analisi di regressione. Se si considera la variabile come nominale multipla, l'analisi rappresenta un'alternativa all'analisi discriminante. Raggruppando le variabili in più di due insiemi è possibile analizzare i dati in numerosi modi.

Analisi corrispondenze

L'obiettivo dell'analisi delle corrispondenze è generare biplot per le tabelle di corrispondenza. In una tabella di corrispondenza, si suppone che le variabili di riga e colonna rappresentino categorie non ordinate; di conseguenza, viene sempre utilizzato il livello nominale di scaling ottimale. Entrambe le variabili vengono esaminate solo per quanto riguarda le relative informazioni nominali. In altre parole, l'unica considerazione è il fatto che alcuni oggetti appartengono alla stessa categoria e altri no. Non viene fatta alcuna ipotesi circa la distanza o l'ordine tra le categorie della stessa variabile.

Un utilizzo specifico dell'analisi delle corrispondenze è l'analisi delle tavole di contingenza a due vie. Se una tabella include *r* righe attive e *c* colonne attive, il numero delle dimensioni nella soluzione dell'analisi delle corrispondenze è il valore minimo tra *r* meno 1 e *c* meno 1. In altre parole, è possibile rappresentare perfettamente le categorie delle righe o delle colonne di una tavola di contingenza in uno spazio dimensionale. Da un punto di vista pratico, tuttavia, è utile

rappresentare le categorie di righe e colonne in una tabella a due vie in uno spazio dimensionale ridotto, ad esempio con due dimensioni, in quanto i grafici bidimensionali sono più facilmente comprensibili delle rappresentazioni spaziali multidimensionali.

Quando viene utilizzato un numero di dimensioni possibili inferiore al massimo, le statistiche generate nell'analisi descrivono il grado di attendibilità della rappresentazione delle categorie di righe e colonne nella rappresentazione dimensionale ridotta. A condizione che la qualità della rappresentazione della soluzione a due dimensioni sia buona, è possibile esaminare i grafici dei punti di riga e di colonna per comprendere quali categorie della variabile di riga sono simili, quali categorie della variabile di colonna sono simili e quali categorie di riga e di colonna sono simili tra loro.

Relazione con altre procedure di Categories. L'analisi delle corrispondenze semplice è limitata a tabelle a due vie. Se le variabili di interesse sono più di due, è possibile combinarle per creare variabili di interazione. Ad esempio, per le variabili *regione*, *lavoro* ed *età*, è possibile combinare *regione* e *lavoro* per creare una nuova variabile *relav* inclusiva delle 12 categorie illustrate nella tabella seguente. La nuova variabile forma una tabella a due vie con *età* (12 righe, 4 colonne), che può essere analizzata tramite analisi delle corrispondenze.

Tabella 1-4
Combinazioni di regione e lavoro

Codice categoria	Definizione categoria	Codice categoria	Definizione categoria
1	Nord, stagista	7	Est, stagista
2	Nord, commerciale	8	Est, commerciale
3	Nord, manager	9	Est, manager
4	Sud, stagista	10	Ovest, stagista
5	Sud, commerciale	11	Ovest, commerciale
6	Sud, manager	12	Ovest, manager

Uno svantaggio di questo approccio è rappresentato dal fatto che ciascuna coppia di variabili può essere combinata. È possibile combinare *lavoro* ed *età*, generando un'altra variabile a 12 categorie. Oppure, è possibile combinare *regione* ed *età*, generando una nuova variabile a 16 categorie. Ciascuna di queste variabili di interazione forma una tabella a due vie con la variabile rimanente. Le analisi delle corrispondenze di queste tre tabelle non genereranno risultati identici, tuttavia ciascuna costituisce un approccio valido. Inoltre, in presenza di quattro o più variabili, è possibile creare tabelle a due vie per mettere a confronto due variabili di interazione. Il numero delle tabelle che è possibile analizzare può diventare ampio, anche in presenza di un numero limitato di variabili. È possibile selezionare una di queste tabelle da analizzare, oppure analizzarle tutte. In alternativa, la procedura Analisi delle corrispondenze multiple può essere utilizzata per esaminare tutte le variabili contemporaneamente senza necessità di creare variabili di interazione.

Relazione con le tecniche standard. La procedura Tavole di contingenza può essere utilizzata anche per analizzare tavole di contingenza, con l'indipendenza come elemento chiave comune delle analisi. Tuttavia, anche in tavole di piccole dimensioni, può essere difficile rilevare la causa degli scostamenti dall'indipendenza. L'utilità dell'analisi delle corrispondenze risiede nella visualizzazione di questi modelli per tabelle a due vie di qualsiasi dimensione. Se esiste un'associazione tra le variabili di riga e di colonna, ovvero se il valore chi-quadrato è significativo, l'analisi delle corrispondenze può essere utile per rivelare la natura della relazione.

Analisi corrispondenze multiple

L'analisi delle corrispondenze multiple tenta di generare una soluzione in cui gli oggetti della stessa categoria sono rappresentati in un grafico vicini tra loro, mentre quelli di categorie diverse sono inseriti in posizioni distanti. Ciascun oggetto si trova il più vicino possibile ai punti delle categorie a esso applicabili. In questo modo, le categorie dividono gli oggetti in sottogruppi omogenei. Le variabili sono considerate omogenee quando classificano gli oggetti nelle stesse categorie negli stessi sottogruppi.

Per una soluzione monodimensionale, l'analisi delle corrispondenze multiple assegna valori di scala ottimali (quantificazioni di categoria) a ciascuna categoria di ciascuna variabile, in modo che globalmente, in media, le categorie abbiano la massima variabilità. Per una soluzione bidimensionale, l'analisi delle corrispondenze multiple individua un secondo insieme di quantificazioni delle categorie per ciascuna categoria di ciascuna variabile non collegata al primo insieme, tentando nuovamente di massimizzare la variabilità, e così via. Poiché le categorie di una variabile ricevono tanti punteggi quante sono le dimensioni, si suppone che le variabili nell'analisi siano nominali multiple a livello di scaling ottimale.

L'analisi delle corrispondenze multiple assegna anch'essa punteggi agli oggetti nell'analisi, in modo che le quantificazioni di categoria siano le medie, o centroidi, dei punteggi degli oggetti inclusi in tale categoria.

Relazione con altre procedure di Categories. L'analisi delle corrispondenze multiple è conosciuta anche come analisi di omogeneità o scaling duale. In presenza di due sole variabili,essa fornisce risultati confrontabili, ma non identici, all'analisi delle corrispondenze. L'analisi delle corrispondenze genera un output univoco che riassume l'adattamento e la qualità della rappresentazione della soluzione, incluse informazioni sulla stabilità. Di conseguenza, l'analisi delle corrispondenze è generalmente preferibile all'analisi delle corrispondenze multiple in presenza di due variabili. Un'altra differenza tra le due procedure è rappresentata dal fatto che l'input per l'analisi delle corrispondenze multiple è una matrice di dati, in cui le righe sono oggetti e le colonne sono variabili, mentre l'input per l'analisi delle corrispondenze può essere la stessa matrice di dati, una matrice di distanza generale o una tavola di contingenza congiunta, vale a dire una matrice aggregata in cui sia le righe che le colonne rappresentano categorie di variabili. L'analisi delle corrispondenze multiple può essere considerata anche come un'analisi delle componenti principali dei dati scalati a livello nominale multiplo.

Relazione con le tecniche standard. L'analisi delle corrispondenze multiple può essere considerata come l'analisi di una tavola di contingenza a più vie. Le tavole di contingenza a più vie possono essere analizzate anche tramite la procedura Tavole di contingenza, che però fornisce statistiche riassuntive distinte per ciascuna categoria di ciascuna variabile di controllo. Con l'analisi delle corrispondenze multiple, è spesso possibile riassumere la relazione tra tutte le variabili in un unico grafico a due dimensioni. Un utilizzo avanzato dell'analisi delle corrispondenze multiple consiste nel sostituire i valori di categoria originali con i valori di scala ottimali della prima dimensione, eseguendo quindi un'analisi multivariata secondaria. Poiché l'analisi delle corrispondenze multiple sostituisce le etichette di categoria con valori di scala numerici, dopo l'analisi è possibile applicare molte procedure diverse che richiedono dati numerici. Ad esempio, la procedura Analisi fattoriale genera una prima componente principale equivalente alla prima dimensione dell'analisi delle corrispondenze multiple. I punteggi delle componenti nella prima dimensione sono uguali ai punteggi degli oggetti e i pesi quadrati delle componenti sono uguali

alle misure di discriminazione. La seconda dimensione dell'analisi delle corrispondenze multiple, tuttavia, non è uguale alla seconda dimensione dell'analisi fattoriale.

Scaling multidimensionale

L'utilizzo dello scaling multidimensionale è il più adatto quando l'obiettivo dell'analisi è individuare la struttura in un insieme di misure di distanza tra un insieme di oggetti o casi. Questa operazione viene compiuta assegnando le osservazioni a posizioni specifiche in uno spazio concettuale ridotto, in modo tale che le distanze tra i punti nello spazio corrispondano il più possibile alle dissimilarità specificate. In questo modo si ottiene una rappresentazione dei minimi quadrati degli oggetti all'interno dello spazio dimensionale ridotto, che nella maggior parte dei casi aiuta a comprendere meglio i dati.

Relazione con altre procedure di Categories. Quando sono presenti dati multivariati dai quali si creano distanze e che quindi si analizzano tramite scaling multidimensionale, i risultati sono simili a quelli dell'analisi dei dati tramite analisi delle componenti categoriali principali con normalizzazione principale degli oggetti. Questo tipo di PCA è nota anche come analisi delle coordinate principali.

Relazione con le tecniche standard. La procedura di scaling multidimensionale del modulo Categories (PROXSCAL) offre numerosi miglioramenti rispetto alla procedura di scaling disponibile nel modulo Statistics Base (ALSCAL). PROXSCAL offre un algoritmo più rapido per alcuni modelli e consente di assegnare vincoli sullo spazio comune. Inoltre, PROXSCAL tenta di ridurre al minimo il raw stress normalizzato, anzichè l's-stress (anche denominato **deformazione**). Il raw stress normalizzato è generalmente preferibile in quanto rappresenta una misura basata sugli scostamenti, mentre l's-stress si basa sui quadrati degli scostamenti.

Unfolding multidimensionale

L'unfolding multidimensionale è particolarmente indicato se lo scopo dell'analisi è quello di individuare la struttura di un insieme di misure di distanza tra due insiemi di oggetti (ovvero gli oggetti riga e colonna). Questa operazione viene compiuta assegnando le osservazioni a posizioni specifiche in uno spazio concettuale ridotto, in modo tale che le distanze tra i punti nello spazio corrispondano il più possibile alle dissimilarità specificate. In questo modo si ottiene una rappresentazione dei minimi quadrati degli oggetti riga e colonna all'interno dello spazio dimensionale ridotto, che nella maggior parte dei casi aiuta a comprendere meglio i dati.

Relazione con altre procedure di Categories. Se i dati si riferiscono a distanze di un unico insieme di oggetti (quadrato, matrice simmetrica), usare lo scaling multidimensionale.

Relazione con le tecniche standard. La procedura di unfolding multidimensionale del modulo Categories (PREFSCAL) offre numerosi miglioramenti rispetto alla procedura di unfolding disponibile nel modulo Statistics Base (tramite il comando ALSCAL). PREFSCAL permette di limitare lo spazio comune. Inoltre, tenta di minimizzare la misura dello stress penalizzata, evitando che venga generate soluzioni inadeguate (problema che si verifica con gli algoritmi più vecchi).

Proporzioni nei grafici di scaling ottimale

Le proporzioni nei grafici di scaling ottimale sono isotropiche. In un grafico a due dimensioni, la distanza che rappresenta un'unità nella dimensione 1 è uguale alla distanza che rappresenta un'unità nella dimensione 2. Se si modifica l'intervallo di una dimensione in un grafico a due dimensioni, il sistema cambia le dimensioni dell'altra dimensione per mantenere uguali le distanze fisiche. Le proporzioni isotropiche non possono essere ignorate per le procedure di scaling ottimale.

Letture consigliate

Per informazioni generali sulle tecniche di scaling ottimale, vedere i seguenti testi:

Barlow, R. E., D. J. Bartholomew, D. J. Bremner, e H. D. Brunk. 1972. *Statistical inference under order restrictions*. New York: John Wiley and Sons.

Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. In: *Methodologies of Pattern Recognition,* S. Watanabe, ed. New York: Academic Press.

Bishop, Y. M., S. E. Feinberg, e P. W. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass.: MIT Press.

De Leeuw, J. 1984. The Gifi system of nonlinear multivariate analysis. In: *Data Analysis and Informatics III,* E. Diday, et al., ed..

De Leeuw, J. 1990. Multivariate analysis with optimal scaling. In: *Progress in Multivariate Analysis,* S. Das Gupta, e J. Sethuraman, ed. Calcutta: Indian Statistical Institute.

De Leeuw, J., e J. Van Rijckevorsel. 1980. HOMALS and PRINCALS—Some generalizations of principal components analysis. In: *Data Analysis and Informatics,* E. Diday, et al., ed. Amsterdam: North-Holland.

De Leeuw, J., F. W. Young, e Y. Takane. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .

Gifi, A. 1990. *Nonlinear multivariate analysis*. Chichester: John Wiley and Sons.

Heiser, W. J., e J. J. Meulman. 1995. Nonlinear methods for the analysis of homogeneity and heterogeneity. In: *Recent Advances in Descriptive Multivariate Analysis,* W. J. Krzanowski, ed. Oxford: Oxford University Press.

Israëls, A. 1987. *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Krzanowski, W. J., e F. H. C. Marriott. 1994. *Multivariate analysis: Part I, distributions, ordination and inference*. London: Edward Arnold.

Lebart, L., A. Morineau, e K. M. Warwick. 1984. *Multivariate descriptive statistical analysis*. New York: John Wiley and Sons.

Max, J. 1960. Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, 6, .

Meulman, J. J. 1986. *A distance approach to nonlinear multivariate analysis*. Leiden: DSWO Press.

Meulman, J. J. 1992. The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57, .

Nishisato, S. 1980. *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S. 1994. *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Rao, C. R. 1973. *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley and Sons.

Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In: *Multivariate Analysis, Vol. 5,* P. R. Krishnaiah, ed. Amsterdam: North-Holland.

Roskam, E. E. 1968. *Metric analysis of ordinal data in psychology*. Voorschoten: VAM.

Shepard, R. N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, .

Wolter, K. M. 1985. *Introduction to variance estimation*. Berlin: Springer-Verlag.

Young, F. W. 1981. Quantitative analysis of qualitative data. *Psychometrika*, 46, .

Capitolo
2

Regressione categoriale (CATREG)

La procedura **Regressione categoriale** consente di quantificare i dati categoriali mediante l'assegnazione di valori numerici alle categorie e di ottenere quindi un'equazione della regressione lineare ottimale per le variabili trasformate. La regressione categoriale è nota anche con l'acronimo CATREG (*reg*ressione *cat*egoriale).

L'analisi della regressione lineare standard comporta la riduzione al minimo della somma dei quadrati delle differenze tra una variabile (dipendente) di risposta e una combinazione ponderata di variabili (indipendenti) predittore. Le variabili sono in genere quantitative e i dati categoriali (nominali) vengono ricodificati in variabili binarie o di contrasto. Di conseguenza, le variabili categoriali consentono di distinguere i gruppi di casi e le stime della tecnica consentono di distinguere gli insiemi di parametri per ciascun gruppo. I coefficienti stimati riflettono il modo in cui le modifiche dei predittori influiscono sulla risposta. È possibile stimare la risposta per qualsiasi combinazione di valori dei predittori.

Un approccio alternativo consiste nell'analisi della regressione della risposta rispetto ai valori stessi dei predittori categoriali. Per ciascuna variabile viene pertanto stimato un singolo coefficiente. I valori di categoria delle variabili categoriali sono tuttavia arbitrari. Se le categorie vengono codificate in modi diversi, anche i coefficienti saranno diversi e i confronti tra analisi delle stesse variabili risulteranno difficoltosi.

La procedura CATREG consente di ampliare l'approccio standard poiché applica lo scaling simultaneamente alle variabili nominali, ordinali e numeriche. Questa procedura quantifica le variabili categoriali in modo tale che le quantificazioni riflettano le caratteristiche delle categorie originali e considera le variabili categoriali quantificate allo stesso modo delle variabili numeriche. L'utilizzo delle trasformazioni non lineari consente di analizzare le variabili in una gamma di livelli diversi e di individuare pertanto il modello che meglio si adatta alle specifiche esigenze.

Esempio. La regressione categoriale consente di illustrare in quale modo il grado di soddisfazione dipende dalla categoria lavorativa, dall'area geografica e dalla quantità di spostamenti richiesti. Si potrebbe scoprire che un grado elevato di soddisfazione è correlato ai manager e a un numero ridotto di spostamenti. L'equazione di regressione risultante può essere utilizzata per prevedere il grado di soddisfazione relativo a qualsiasi combinazione delle tre variabili indipendenti.

Statistiche e grafici. Frequenze, coefficienti di regressione, tabella ANOVA, cronologia delle iterazioni, quantificazioni di categoria, correlazioni tra predittori non trasformati, correlazioni tra predittori trasformati, grafici dei residui e grafici di trasformazione.

Dati. La procedura CATREG opera sulle variabili indicatore di categoria, che dovrebbero essere rappresentate da interi positivi. Nella finestra di dialogo Discretizzazione è possibile convertire le variabili rappresentate da frazioni o da stringhe in interi positivi.

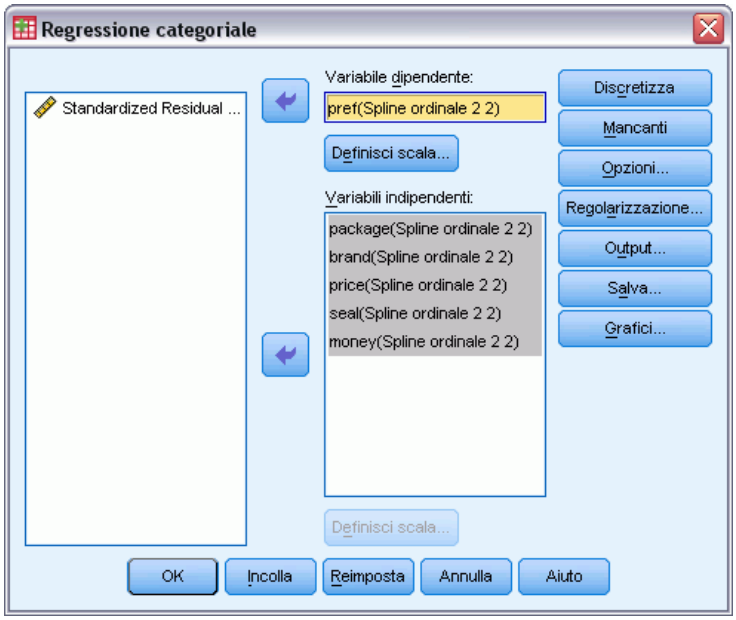
Assunzioni. È consentita una sola variabile di risposta, ma il numero massimo di variabili predittore è uguale a 200. I dati devono includere almeno tre casi validi e il numero di casi validi deve essere uguale al numero delle variabili predittore più uno.

Procedure correlate. La procedura CATREG equivale all'analisi della correlazione canonica categoriale con scaling ottimale (OVERALS) con due insiemi, di cui uno contiene solo una variabile. Lo scaling di tutte le variabili a livello numerico corrisponde all'analisi della regressione multipla standard.

Per ottenere una regressione categoriale

- Dai menu, scegliere:
  Analizza > Regressione > Scaling ottimale (CATREG)...

Figura 2-1
Finestra di dialogo Regressione categoriale



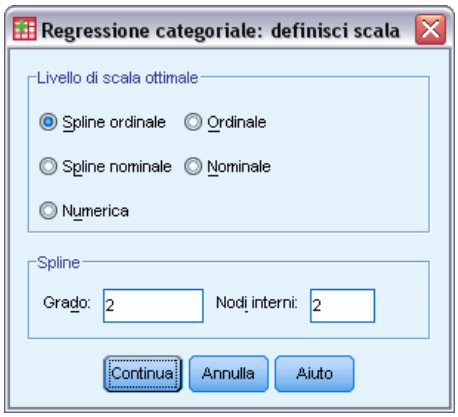
- Selezionare la variabile dipendente e le variabili indipendenti.
- Fare clic su OK.

È inoltre possibile modificare il livello di scaling per ciascuna variabile.

Definisci scala in regressione categoriale

È possibile impostare il livello di scaling ottimale per le variabili dipendenti e indipendenti, che vengono scalate per impostazione predefinita come spline (ordinali) monotoni di secondo grado con due nodi interni. È inoltre possibile impostare il peso delle variabili dell'analisi.

Figura 2-2
Finestra di dialogo Definisci scala



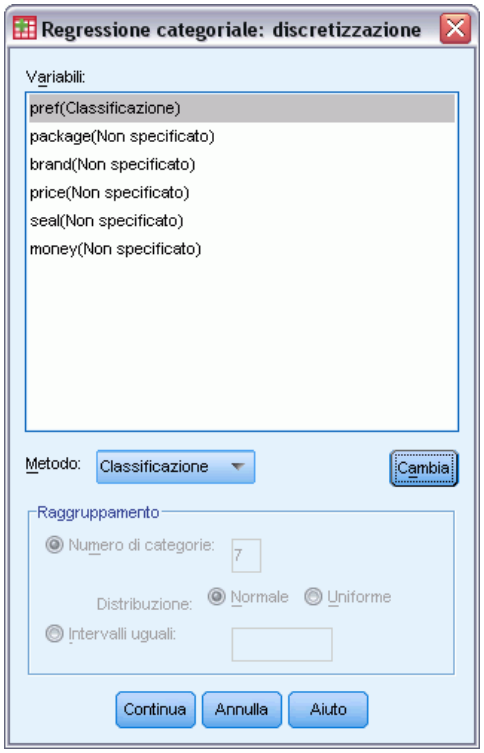
Livello di scaling ottimale. È inoltre possibile selezionare il livello di scaling per la quantificazione di ciascuna variabile.

- **Spline ordinale.** Nella variabile con scaling ottimale viene mantenuto l'ordine delle categorie della variabile osservata. I punti di categoria si troveranno su una linea retta (vettore) che passa per l'origine. La trasformazione ottenuta è un polinomio livellato monotono del grado specificato. Gli elementi vengono determinati dal numero di nodi interni definito dall'utente e dalla relativa posizione stabilita dalla procedura.
- **Spline nominale.** Le uniche informazioni della variabile osservata che verranno mantenute nella variabile con scaling ottimale sono quelle relative al raggruppamento degli oggetti in categorie. Non viene mantenuto l'ordine delle categorie della variabile osservata. I punti di categoria si troveranno su una linea retta (vettore) che passa per l'origine. La trasformazione ottenuta è un polinomio livellato possibilmente non monotono del grado specificato. Gli elementi vengono determinati dal numero di nodi interni definito dall'utente e dalla relativa posizione stabilita dalla procedura.
- **Ordinale.** Nella variabile con scaling ottimale viene mantenuto l'ordine delle categorie della variabile osservata. I punti di categoria si troveranno su una linea retta (vettore) che passa per l'origine. La trasformazione ottenuta ha un grado di adeguatezza maggiore di quello ottenuto con la trasformazione dello spline ordinale, ma è meno regolare.
- **Nominale.** Le uniche informazioni della variabile osservata che verranno mantenute nella variabile con scaling ottimale sono quelle relative al raggruppamento degli oggetti in categorie. Non viene mantenuto l'ordine delle categorie della variabile osservata. I punti di categoria si troveranno su una linea retta (vettore) che passa per l'origine. La trasformazione ottenuta ha un grado di adeguatezza maggiore di quello ottenuto con la trasformazione dello spline nominale, ma è meno regolare.
- **Numerica.** Le categorie vengono considerate come ordinate ed equamente distanziate (a livello di intervallo). L'ordine delle categorie e le distanze uguali tra i numeri delle categorie della variabile osservata vengono mantenuti nella variabile con scaling ottimale. I punti di categoria si troveranno su una linea retta (vettore) che passa per l'origine. Se tutte le variabili sono a livello numerico, l'analisi corrisponde all'analisi delle componenti principali standard.

Regressione categoriale: Discretizzazione

Nella finestra di dialogo Discretizzazione è possibile selezionare un metodo di ricodifica delle variabili. Le variabili con valori frazionari sono raggruppate in sette categorie (o nel numero di valori distinti della variabile se tale numero è inferiore a sette) con distribuzione approssimativamente normale, se non viene specificato diversamente. Le variabili stringa vengono sempre convertite in interi positivi tramite l'assegnazione di indicatori di categoria in base a un ordinamento alfanumerico crescente. La discretizzazione delle variabili stringa è valida per questi valori interi. Le altre variabili rimangono distinte per impostazione predefinita. Le variabili discretizzate vengono quindi utilizzate per l'analisi.

Figura 2-3
Finestra di dialogo Discretizza



Metodo. Scegliere un metodo di raggruppamento, di classificazione o di moltiplicazione.

- **Raggruppamento.** Ricodifica in un numero specificato di categorie o ricodifica per intervallo.
- **Classificazione.** La variabile viene discretizzata tramite la classificazione dei casi.
- **Moltiplicazione.** I valori correnti della variabile vengono standardizzati, moltiplicati per 10, arrotondati e viene aggiunta una costante in modo tale che il valore discretizzato minore sia uguale a 1.

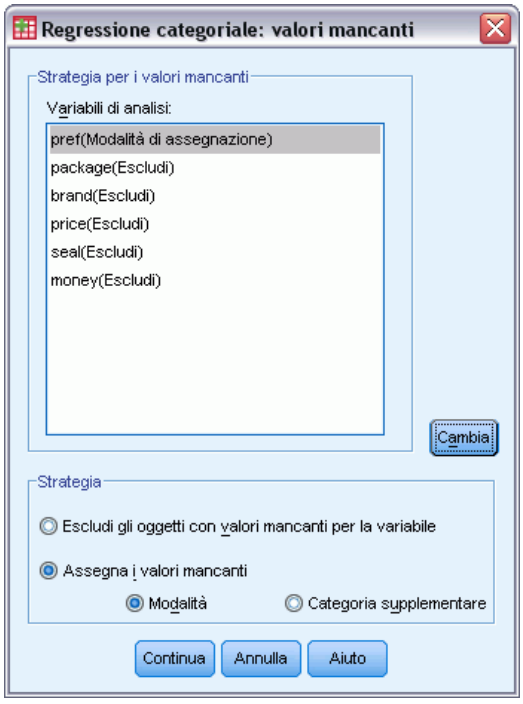
Raggruppamento. Per la discretizzazione delle variabili tramite raggruppamento sono disponibili le seguenti opzioni:

- **Numero di categorie.** Specificare un numero di categorie e se la distribuzione dei valori della variabile nelle categorie deve essere normale o uniforme.
- **Intervalli uguali.** Le variabili vengono ricodificate in categorie definite in base agli intervalli di dimensioni uguali specificati. È necessario specificare la lunghezza degli intervalli.

Regressione categoriale: Valori mancanti

Nella finestra di dialogo Valori mancanti è possibile scegliere la strategia di gestione dei valori mancanti delle variabili dell'analisi e delle variabili supplementari.

Figura 2-4
Finestra di dialogo Valori mancanti



Strategia. Specificare se si desidera escludere i valori mancanti (eliminazione listwise) o aggiungere gli oggetti con valori mancanti (trattamento attivo).

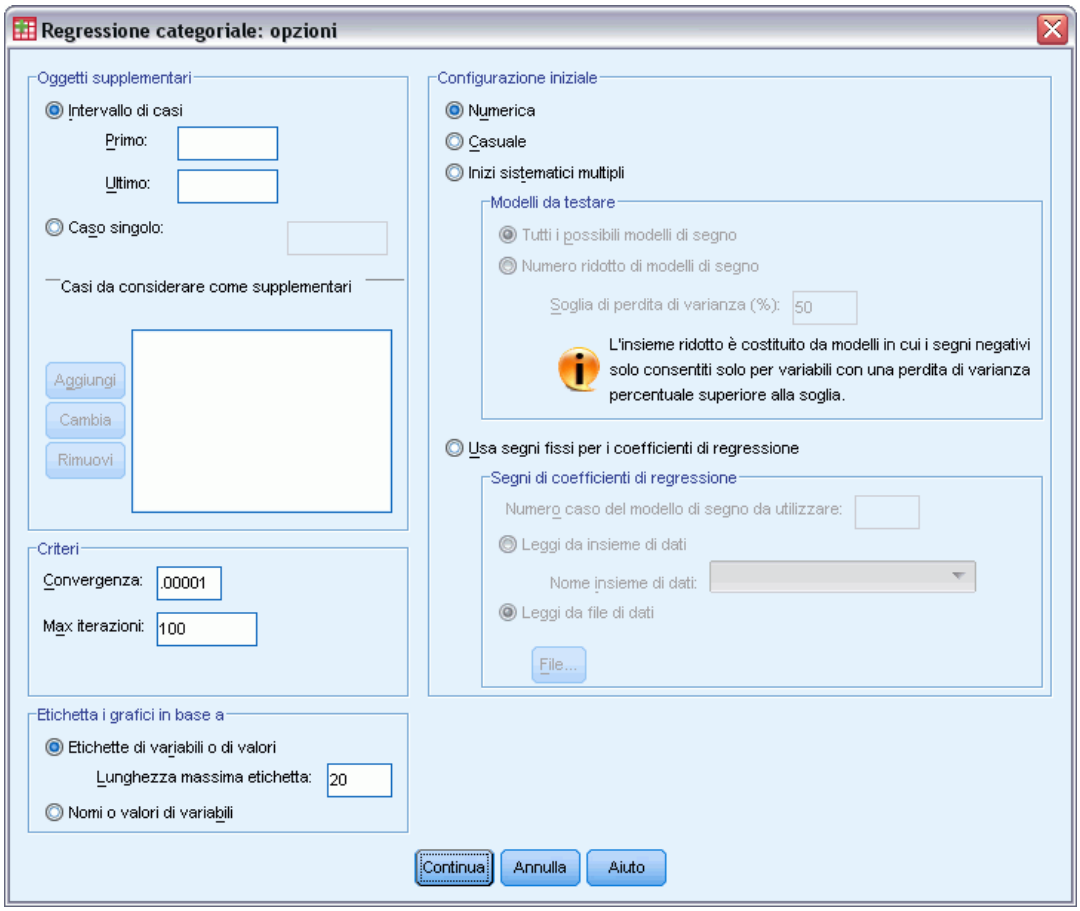
- **Escludi gli oggetti con valori mancanti per la variabile** Gli oggetti con valori mancanti della variabile selezionata sono esclusi dall'analisi. Questa strategia non è disponibile per le variabili supplementari.
- **Assegna i valori mancanti.** Agli oggetti con valori mancanti della variabile selezionata vengono assegnati i valori ed è possibile scegliere il metodo di assegnazione. Selezionare Moda per sostituire i valori mancanti con la categoria più frequente. Se sono disponibili più mode, verrà utilizzata quella con l'indicatore di categoria minore. Selezionare Categoria distinta per sostituire i valori mancanti con la stessa quantificazione di una categoria supplementare. Ciò

implica che gli oggetti con un valore mancante nella variabile specificata vengono considerati come appartenenti alla stessa categoria supplementare.

Regressione categoriale: Opzioni

Nella finestra di dialogo Opzioni è possibile selezionare lo stile di configurazione iniziale, specificare i criteri di iterazione e di convergenza, selezionare gli oggetti supplementari e impostare le etichette dei grafici.

Figura 2-5
Finestra di dialogo Opzioni



Oggetti supplementari. Consente di specificare gli oggetti che si desidera considerare come supplementari. Digitare il numero di un oggetto supplementare (o indicare un intervallo di casi) e fare clic su Aggiungi. Non è possibile ponderare oggetti supplementari (i pesi specificati vengono ignorati).

Configurazione iniziale. Se nessuna variabile viene considerata come nominale, selezionare la configurazione Numerica. Se almeno una variabile viene considerata come nominale, selezionare la configurazione Casuale.

In alternativa, se almeno una delle variabili ha un livello di scaling ordinale o spline ordinale, l'algoritmo adatto al modello utilizzato solitamente può rivelarsi una soluzione non proprio ottimale. La selezione di Inizi sistematici multipli con tutti i possibili modelli di segni da testare troverà sempre la soluzione ottimale, ma il tempo necessario per l'elaborazione aumenta rapidamente con l'aumento delle variabili ordinali e spline ordinali nell'insieme di dati. Per ridurre il numero di modelli di test, è possibile specificare una soglia percentuale di perdita di varianza, dove l'incremento della soglia comporta l'aumento dei modelli di segni che verranno esclusi. Questa opzione non garantisce la soluzione ottimale, ma diminuisce la possibilità di ottenere una soluzione non ottimale. Inoltre, se non viene trovata la soluzione ottimale, diminuiscono le possibilità che la soluzione non ottimale sia significativamente diversa dalla soluzione ottimale. Quando sono richiesti gli inizi sistematici multipli, i segni dei coefficienti di regressione per ogni avvio vengono scritti in un file di dati IBM® SPSS® Statistics esterno o in un insieme di dati della sessione corrente. Per ulteriori informazioni, vedere l'argomento Regressione categoriale: Salva a pag. 25.

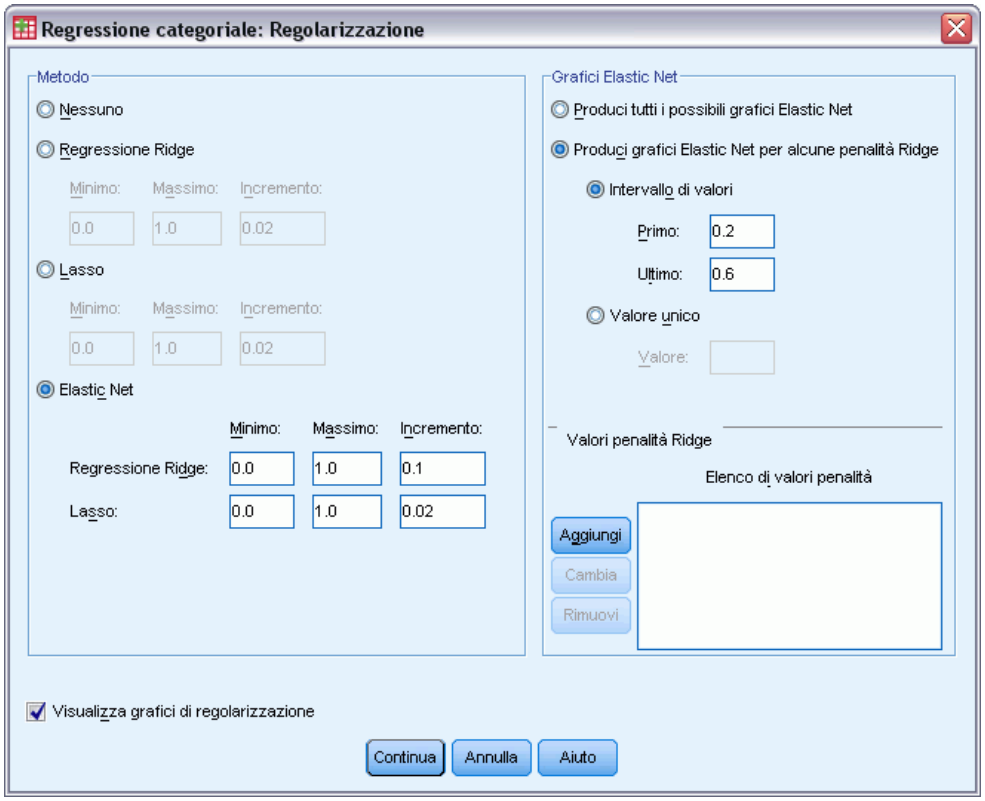
I risultati di un'esecuzione precedente con inizi sistematici multipli consentono di Usare segni fissi per i coefficienti di regressione. I segni (indicati da 1 e −1) devono essere in una riga dell'insieme di dati o file specificato. Il numero iniziale intero è il numero del caso della riga di questo file che contiene i segni da utilizzare.

Criteri. È possibile specificare il numero massimo di iterazioni che possono essere eseguite dalla regressione durante i calcoli e inoltre selezionare un valore per il criterio di convergenza. La procedura si interrompe se la differenza dell'adattamento totale delle due ultime iterazioni è inferiore al valore di convergenza o se viene raggiunto il numero massimo di iterazioni.

Etichetta i grafici in base a. Consente di specificare se nei grafici verranno utilizzati le variabili e le etichette dei valori o i nomi delle variabili e i valori. È inoltre possibile specificare una lunghezza massima per le etichette.

Regolarizzazione della regressione categoriale

Figura 2-6
Finestra di dialogo Regolarizzazione



Metodo. I metodi di regolarizzazione consentono di migliorare l'errore predittivo del modello diminuendo la variabilità delle stime del coefficiente di regressione riducendo le stime verso lo 0. Lasso ed Elastic Net ridurranno alcune stime dei coefficienti esattamente a 0, fornendo in questo modo una forma di selezione variabile. Quando è richiesto un metodo di regolarizzazione, il modello regolarizzato e i coefficienti per ogni valore di coefficiente di penalità vengono scritti in un file di dati IBM® SPSS® Statistics esterno o in un insieme di dati della sessione corrente. Per ulteriori informazioni, vedere l'argomento Regressione categoriale: Salva a pag. 25.

- **Regressione Ridge.** La regressione Ridge riduce i coefficienti introducendo un termine di penalità uguale alla somma dei coefficienti al quadrato moltiplicata per un **coefficiente di penalità**. Questo coefficiente può andare da 0 (nessuna penalità) a 1; la procedura cercherà il valore "migliore" della penalità se si indica un intervallo e un incremento.
- **Lasso.** Il termine di penalità Lasso si basa sulla somma dei coefficienti assoluti e la specifica di un coefficiente di penalità è simile a quella della regressione Ridge; tuttavia, Lasso prevede un numero maggiore di operazioni di calcolo.
- **Elastic net.** Elastic net è semplicemente una combinazione delle penalità Lasso e regressione Ridge ed esegue una ricerca nella griglia dei valori specificati al fine di trovare i coefficienti di penalità Lasso e regressione Ridge "migliori". Per una data coppia di penalità Lasso e regressione Ridge, Elastic Net non prevede un numero di calcoli particolarmente più alto rispetto a Lasso.

Visualizza grafici di regolarizzazione. Si tratta di grafici dei coefficienti di regressione rispetto alla penalità di regolarizzazione. Quando si cerca un intervallo di valori per trovare il coefficiente di penalità "migliore", offre una visualizzazione del modo in cui i coefficienti di regressione cambiano nell'arco dell'intervallo.

Grafici Elastic Net. Per il metodo Elastic Net, vengono prodotti dei grafici di regolarizzazione in base ai valori della penalità di regressione Ridge. Produci tutti i possibili grafici Elastic Net utilizza tutti i valori dell'intervallo determinato dai valori di penalità di regressione Ridge minimo e massimo specificati. Produci grafici Elastic Net per alcune penalità Ridge consente di specificare un sottoinsieme dei valori dell'intervallo determinato dal minimo e dal massimo. Digitare il numero di un valore di penalità (o indicare un intervallo di valori) e fare clic su Aggiungi.

Regressione categoriale: Output

Nella finestra di dialogo Output è possibile selezionare le statistiche che si desidera visualizzare nell'output.

Figura 2-7
Finestra di dialogo Output

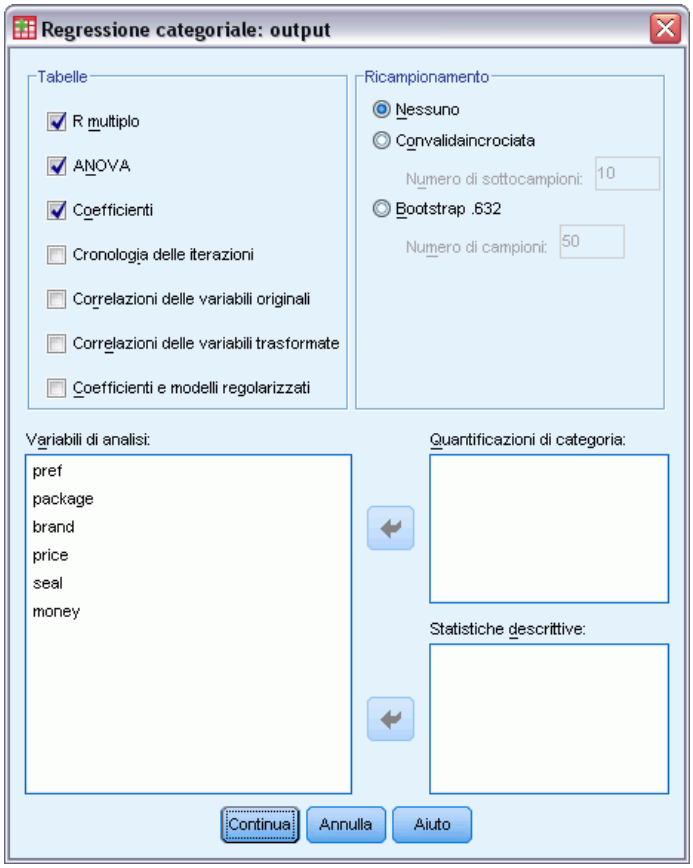


Tabelle. Consente di creare tabelle per:

- **R multiplo.** Comprende R^2, R^2 corretto ed R^2 corretto basati sulla scala ottimale.

- **ANOVA.** Questa opzione include le somme dei quadrati dei residui e della regressione, le medie dei quadrati e un test *F*. Vengono visualizzate due tabelle ANOVA: una con i gradi di libertà per la regressione equivalenti al numero delle variabili predittore e l'altra con i gradi di libertà per la regressione che tengono conto dello scaling ottimale.
- **Coefficienti.** Questa opzione rende disponibili tre tabelle: una tabella Coefficienti che include i coefficienti beta ed il relativo errore standard, i valori *t* e la significatività, una tabella dei coefficienti dello scaling ottimale con l'errore standard dei coefficienti beta che tiene conto dei gradi di libertà dello scaling ottimale e una tabella che include la correlazione di ordine zero e parziale, la misura di importanza relativa di Pratt per i predittori trasformati e la tolleranza precedente e successiva alla trasformazione.
- **Cronologia iterazioni.** Per ogni iterazione, inclusi i valori iniziali dell'algoritmo, vengono visualizzati gli errori relativi a *R* multiplo e alla regressione. Gli incrementi di *R* multiplo vengono visualizzati a partire dalla prima iterazione.
- **Correlazioni delle variabili originali.** Viene visualizzata una matrice con le correlazioni tra le variabili non trasformate.
- **Correlazioni delle variabili trasformate.** Viene visualizzata una matrice con le correlazioni tra le variabili trasformate.
- **Coefficienti e modelli regolarizzati.** Visualizza valori di penalità, R-quadrato e coefficienti di regressione per ogni modello regolarizzato. Se viene specificato un metodo di ricampionamento oppure vengono indicati degli oggetti supplementari (casi di test), viene visualizzato anche l'errore di previsione o l'errore MSE del test.

Ricampionamento. I metodi di ricampionamento forniscono una stima dell'errore di previsione del modello.

- **Convalida incrociata.** La convalida incrociata divide il campione in vari sottocampioni o campioni. A questo punto vengono generati i modelli di regressione categoriale, escludendo di volta in volta i dati da ciascun sottocampione. Il primo modello si basa su tutti i casi eccetto quelli contenuti nel primo sottocampione, il secondo modello si basa su tutti i casi eccetto quelli contenuti nel secondo sottocampione e così via. Il rischio di errore di previsione per ciascun modello viene stimato applicando il modello al sottocampione escluso al momento della generazione del modello stesso.
- **Bootstrap .632.** Mediante bootstrap, le osservazioni vengono derivate in modo casuale dai dati con sostituzione, ripetendo questo processo più volte in modo da ottenere una serie di campioni bootstrap. Per ogni campione bootstrap viene adattato un modello che esegue la stima dell'errore di previsione per ogni modello; l'errore di previsione viene quindi applicato ai casi che non fanno parte del campione bootstrap.

Quantificazioni di categoria. Vengono visualizzate le tabelle che includono i valori trasformati delle variabili selezionate.

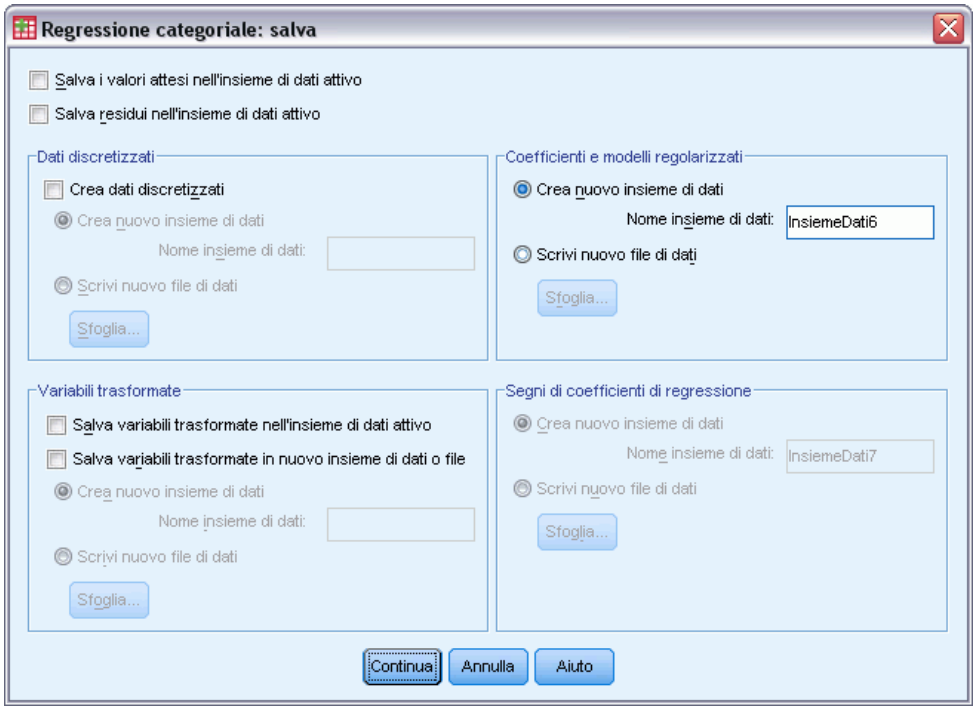
Statistiche descrittive. Vengono visualizzate le tabelle con le frequenze, i valori mancanti e le mode delle variabili selezionate.

Regressione categoriale: Salva

Dalla finestra di dialogo Salva è possibile salvare i valori previsti, i residui e i valori trasformati nel file di dati attivo e/o salvare i dati discretizzati, i valori trasformati, i coefficienti e i modelli regolarizzati nonché i segni dei coefficienti di regressione in un file di dati esterno di IBM® SPSS® Statistics o in un insieme di dati della sessione corrente.

- I file di dati sono disponibili durante la sessione corrente, ma non lo sono in quelle successive a meno che non li si salvi esplicitamente come file di dati. I nomi degli insiemi di dati devono rispettare le regole dei nomi delle variabili.
- I nomi dei file o i nomi dei file di dati devono essere diversi per ogni tipo di dati salvati.

Figura 2-8
Salva



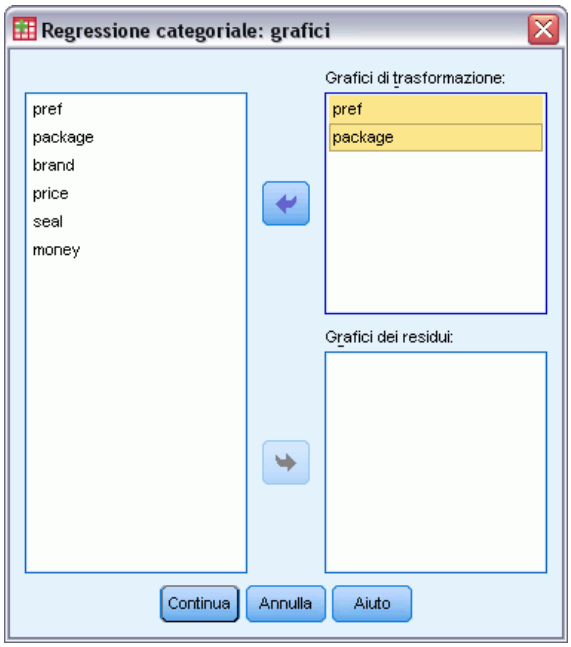
I coefficienti e i modelli regolarizzati vengono salvati ogni volta che viene selezionato un metodo di regolarizzazione nella finestra di dialogo Regolarizzazione. Per impostazione predefinita, questa procedura crea un nuovo insieme di dati con un nome univoco; è comunque possibile specificare un nome di propria scelta o salvare in un file esterno.

I segni dei coefficienti di regressione vengono salvati ogni volta che vengono utilizzati gli inizi sistematici multipli come configurazione iniziale nella finestra di dialogoOpzioni. Per impostazione predefinita, questa procedura crea un nuovo insieme di dati con un nome univoco; è comunque possibile specificare un nome di propria scelta o salvare in un file esterno.

Regressione categoriale: Grafici

Nella finestra di dialogo Grafici è possibile specificare le variabili in base alle quali verranno creati i grafici di trasformazione e dei residui.

Figura 2-9
Finestra di dialogo Grafici



Grafici di trasformazione. Per ciascuna variabile, le quantificazioni di categoria vengono inserite nel grafico mediante il confronto con i valori di categoria originali. Le categorie vuote vengono visualizzate sull'asse orizzontale, ma non influiscono sui calcoli. Queste categorie sono identificate da interruzioni sulla linea che collega le quantificazioni.

Grafici dei residui. Per ciascuna variabile, i residui (calcolati per la variabile dipendente attesa in base a tutte le variabili stimatore eccetto quella in questione) vengono inseriti nel grafico mediante il confronto con gli indicatori di categoria e le quantificazioni di categoria ottimali moltiplicate per i coefficienti beta e confrontate con gli indicatori di categoria.

Opzioni aggiuntive del comando CATREG

Per personalizzare la procedura Regressione categoriale è possibile incollare le selezioni in una finestra di sintassi e quindi modificare la sintassi dei comandi CATREG così ottenuta. Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare i nomi di radice per le variabili trasformate durante il salvataggio nel file di dati attivo (con il sottocomando SAVE).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
3

Analisi delle componenti principali categoriale (CATPCA)

Questa procedura consente di quantificare le variabili categoriali e contemporaneamente di ridurre la dimensione dei dati. L'analisi delle componenti principali categoriale è conosciuta anche con l'acronimo CATPCA *(Categorical Principal Component Analysis)*.

Lo scopo principale dell'analisi delle componenti principali categoriale è quello di ridurre un insieme originale di variabili in un insieme più limitato di componenti non correlate che rappresentano la maggior parte delle informazioni disponibili nelle variabili originali. Questa tecnica risulta particolarmente utile nel caso in cui non sia possibile interpretare in modo efficiente le relazioni tra gli oggetti (soggetti e unità) a causa della presenza di un numero troppo elevato di variabili. Se la dimensione viene ridotta, sarà possibile interpretare un numero ridotto di componenti, anziché un numero elevato di variabili.

L'analisi delle componenti principali standard presume l'esistenza di relazioni lineari tra le variabili numeriche. L'approccio di scaling ottimale consente d'altra parte di scalare le variabili a livelli diversi. Le variabili categoriali vengono quantificate in modo ottimale nella dimensione specificata ed è quindi possibile definire le relazioni non lineari tra variabili.

Esempio. L'analisi delle componenti principali categoriale consente di visualizzare graficamente la relazione esistente tra una categoria lavorativa, una divisione, una regione, la quantità di spostamenti richiesti (alta, media e bassa) e il grado di soddisfazione. A volte ci si può rendere conto che due dimensioni sono sufficienti per considerare un'entità notevole della varianza. La prima dimensione può distinguere la categoria lavorativa rispetto alla regione, mentre la seconda può distinguere la divisione dalla quantità di spostamenti. Può anche risultare che un grado di soddisfazione alto sia correlato a una quantità media di spostamenti.

Statistiche e grafici. Frequenze, valori mancanti, livello di scaling ottimale, moda, varianza spiegata in base alle coordinate del centroide, coordinate del vettore, totale per variabile e per dimensione, pesi di componente per le variabili quantificate in base al vettore, quantificazioni e coordinate di categoria, cronologia delle iterazioni, correlazioni delle variabili trasformate e autovalori della matrice di correlazione, correlazioni delle variabili originali e autovalori della matrice di correlazione, punteggi degli oggetti, grafici di categoria, grafici di categoria congiunti, grafici di trasformazione, grafici dei residui, grafici dei centroidi proiettati, grafici degli oggetti, biplot, triplot e grafici dei pesi di componente.

Dati. I valori delle variabili stringa vengono sempre convertiti in interi positivi disposti in ordine alfabetico crescente. I valori mancanti definiti dall'utente, i valori mancanti di sistema e i valori inferiori a 1 sono considerati valori mancanti. È possibile ricodificare o aggiungere una costante alle variabili con valori inferiori a 1 per fare in modo che siano considerate come non mancanti.

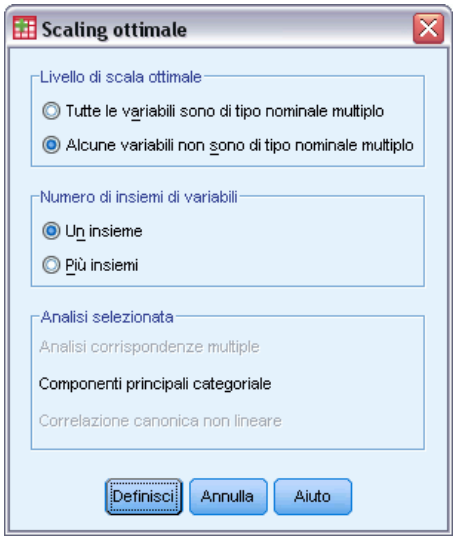
Assunzioni. I dati devono contenere almeno tre casi validi e l'analisi è basata su dati interi positivi. La funzione di discretizzazione classifica automaticamente una variabile con valore frazionario raggruppandone i valori in categorie con una distribuzione vicina a quella normale e converte automaticamente i valori delle variabili stringa in valori interi positivi. È possibile specificare altri schemi di discretizzazione.

Procedure correlate. Lo scaling di tutte le variabili a livello numerico corrisponde all'analisi delle componenti principali standard. Mediante l'utilizzo delle variabili trasformate in un'analisi delle componenti principali lineare standard, è possibile disporre di funzioni alternative per la creazione dei grafici. Se per tutte le variabili sono disponibili livelli di scaling nominale multipli, l'analisi delle componenti principali categoriale equivale all'analisi delle corrispondenze multiple. Se si desidera considerare insiemi di variabili, è consigliabile utilizzare l'analisi della correlazione canonica (non lineare) categoriale.

Per ottenere un'analisi delle componenti principali categoriale

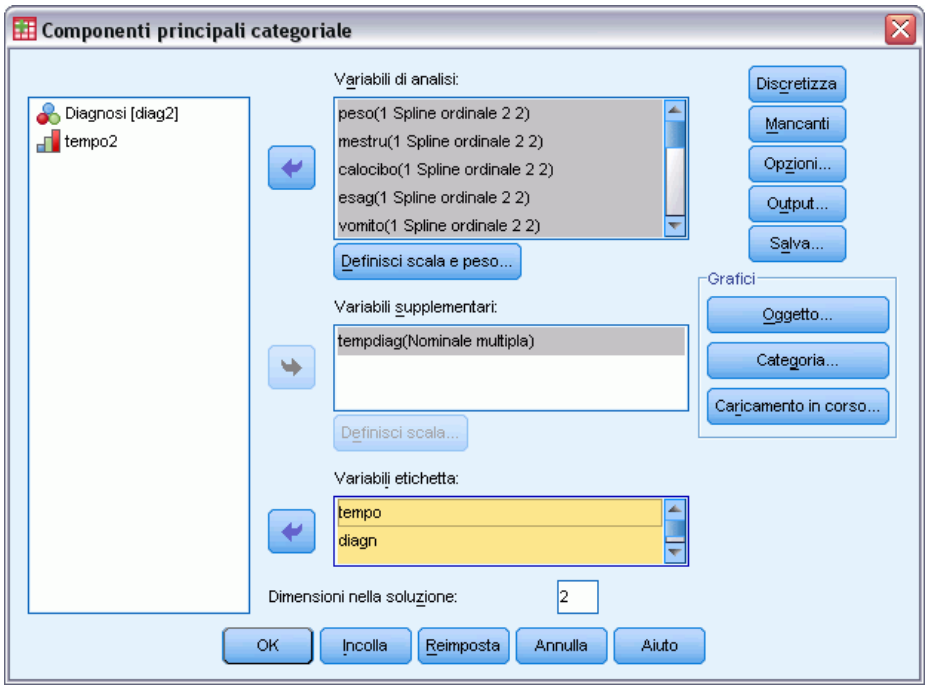
- Dai menu, scegliere:
  Analizza > Riduzioni dimensione > Scaling ottimale...

Figura 3-1
Finestra di dialogo Scaling ottimale



- Selezionare Una o più variabili non nominali multiple.
- Selezionare Un insieme.
- Fare clic su Definisci.

Figura 3-2
Finestra di dialogo Componenti principali categoriale



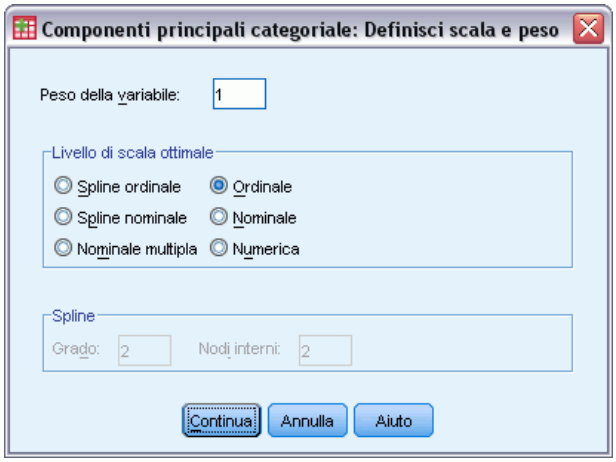
- Selezionare almeno due variabili dell'analisi e specificare il numero di dimensioni nella soluzione.
- Fare clic su OK.

  Se necessario, è possibile specificare variabili supplementari che verranno inserite nella soluzione oppure variabili di etichetta per i grafici.

Definisci scala e peso in CATPCA

È possibile impostare il livello di scaling ottimale per le variabili dell'analisi e le variabili supplementari, che vengono scalate per impostazione predefinita come spline (ordinali) monotoni di secondo grado con due nodi interni. È inoltre possibile impostare il peso delle variabili dell'analisi.

Figura 3-3
Finestra di dialogo Definisci scala e peso



Peso della variabile. Per ciascuna variabile può essere definito un peso, il cui valore deve essere un intero positivo. Il valore predefinito è 1.

Livello di scaling ottimale.È inoltre possibile selezionare il livello di scaling da utilizzare per quantificare ciascuna variabile.

- **Spline ordinale.** Nella variabile con scaling ottimale viene mantenuto l'ordine delle categorie della variabile osservata. I punti di categoria si troveranno su una linea retta (vettore) che passa per l'origine. La trasformazione ottenuta è un polinomio livellato monotono del grado specificato. Gli elementi vengono determinati dal numero di nodi interni definito dall'utente e dalla relativa posizione stabilita dalla procedura.
- **Spline nominale.** Le uniche informazioni della variabile osservata che verranno mantenute nella variabile con scaling ottimale sono quelle relative al raggruppamento degli oggetti in categorie. Non viene mantenuto l'ordine delle categorie della variabile osservata. I punti di categoria si troveranno su una linea retta (vettore) che passa per l'origine. La trasformazione ottenuta è un polinomio livellato possibilmente non monotono del grado specificato. Gli elementi vengono determinati dal numero di nodi interni definito dall'utente e dalla relativa posizione stabilita dalla procedura.
- **Nominale multipla.** Le uniche informazioni della variabile osservata che verranno mantenute nella variabile con scaling ottimale sono quelle relative al raggruppamento degli oggetti in categorie. Non viene mantenuto l'ordine delle categorie della variabile osservata. I punti di categoria saranno nel centroide degli oggetti delle categorie specifiche. Il termine *multipla* indica che per ciascuna dimensione si ottengono insiemi di quantificazioni diversi.
- **Ordinale.** Nella variabile con scaling ottimale viene mantenuto l'ordine delle categorie della variabile osservata. I punti di categoria si troveranno su una linea retta (vettore) che passa per l'origine. La trasformazione ottenuta ha un grado di adeguatezza maggiore di quello ottenuto con la trasformazione dello spline ordinale, ma è meno regolare.
- **Nominale.** Le uniche informazioni della variabile osservata che verranno mantenute nella variabile con scaling ottimale sono quelle relative al raggruppamento degli oggetti in categorie. Non viene mantenuto l'ordine delle categorie della variabile osservata. I punti di categoria si troveranno su una linea retta (vettore) che passa per l'origine. La trasformazione

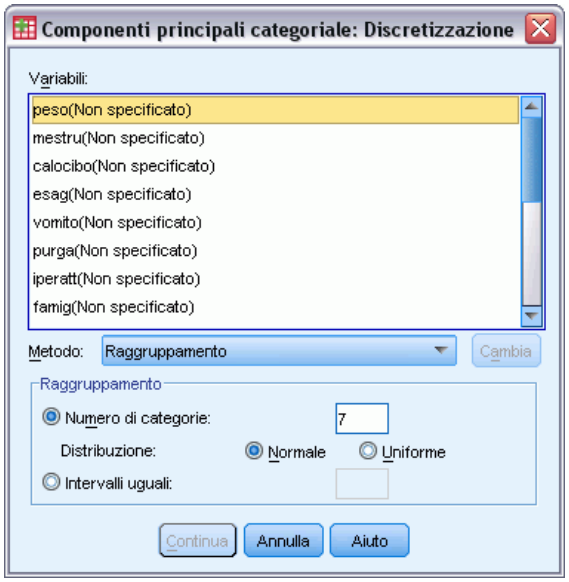
ottenuta ha un grado di adeguatezza maggiore di quello ottenuto con la trasformazione dello spline nominale, ma è meno regolare.

- **Numerica.** Le categorie vengono considerate come ordinate ed equamente distanziate (a livello di intervallo). L'ordine delle categorie e le distanze uguali tra i numeri delle categorie della variabile osservata vengono mantenuti nella variabile con scaling ottimale. I punti di categoria si troveranno su una linea retta (vettore) che passa per l'origine. Se tutte le variabili sono a livello numerico, l'analisi corrisponde all'analisi delle componenti principali standard.

Componenti principali categoriale: Discretizzazione

Nella finestra di dialogo Discretizzazione è possibile selezionare un metodo di ricodifica delle variabili. Le variabili con valori frazionari sono raggruppate in sette categorie (o nel numero di valori distinti della variabile se tale numero è inferiore a sette) con distribuzione approssimativamente normale, se non viene specificato diversamente. Le variabili stringa vengono sempre convertite in interi positivi tramite l'assegnazione di indicatori di categoria in base a un ordinamento alfanumerico crescente. La discretizzazione delle variabili stringa è valida per questi valori interi. Le altre variabili rimangono distinte per impostazione predefinita. Le variabili discretizzate vengono quindi utilizzate per l'analisi.

Figura 3-4
Finestra di dialogo Discretizza



Metodo. Scegliere un metodo di raggruppamento, di classificazione o di moltiplicazione.

- **Raggruppamento.** Ricodifica in un numero specificato di categorie o ricodifica per intervallo.
- **Classificazione.** La variabile viene discretizzata tramite la classificazione dei casi.
- **Moltiplicazione.** I valori correnti della variabile vengono standardizzati, moltiplicati per 10, arrotondati e viene aggiunta una costante in modo tale che il valore discretizzato minore sia uguale a 1.

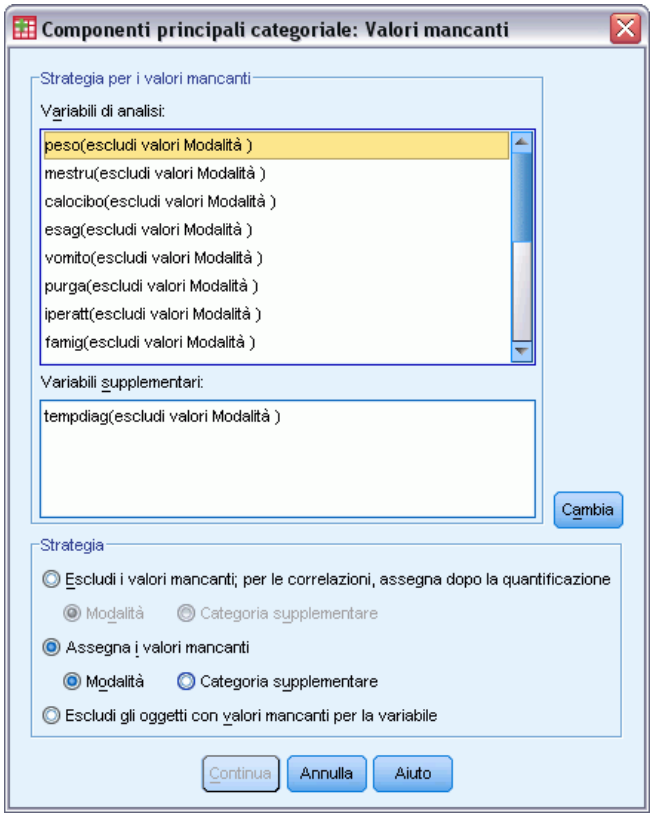
Raggruppamento. Per la discretizzazione delle variabili tramite raggruppamento sono disponibili le seguenti opzioni:

- **Numero di categorie.** Specificare un numero di categorie e se la distribuzione dei valori della variabile nelle categorie deve essere normale o uniforme.
- **Intervalli uguali.** Le variabili vengono ricodificate in categorie definite in base agli intervalli di dimensioni uguali specificati. È necessario specificare la lunghezza degli intervalli.

Componenti principali categoriale: Valori mancanti

Nella finestra di dialogo Valori mancanti è possibile scegliere la strategia di gestione dei valori mancanti delle variabili dell'analisi e delle variabili supplementari.

Figura 3-5
Finestra di dialogo Valori mancanti



Strategia. Specificare se si desidera escludere i valori mancanti (trattamento passivo), assegnare i valori mancanti (trattamento attivo) o escludere gli oggetti con valori mancanti (eliminazione listwise).

- **Escludi i valori mancanti; per le correlazioni, assegna dopo la quantificazione.** Gli oggetti con valori mancanti della variabile selezionata non vengono utilizzati nell'analisi di questa variabile. Se a tutte le variabili è applicato il trattamento passivo, gli oggetti con valori mancanti di tutte le variabili vengono considerati come supplementari. Se nella finestra di dialogo Output sono specificate correlazioni, dopo l'analisi ai valori mancanti viene assegnata

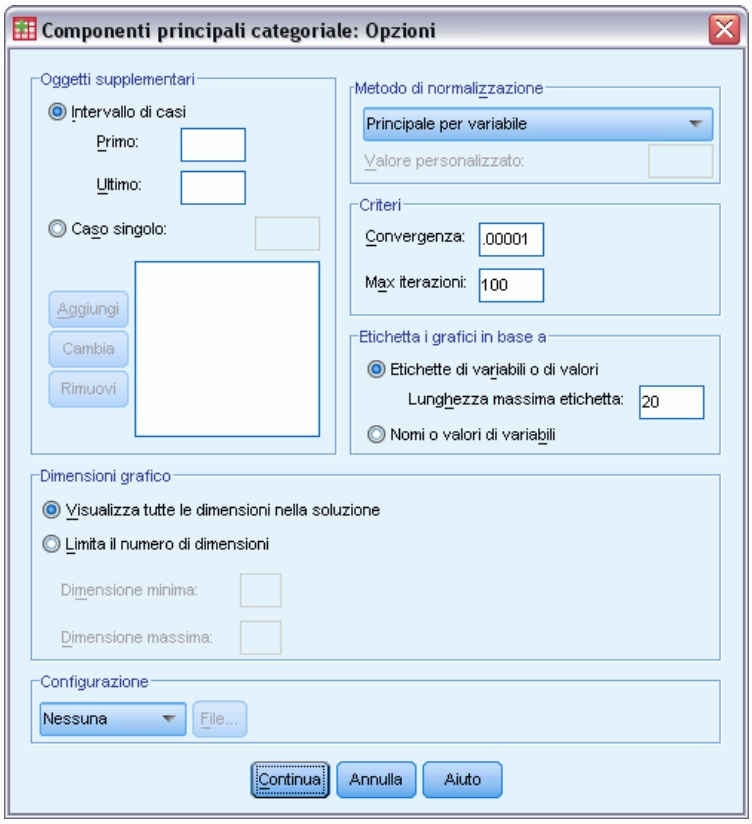
la categoria o moda più frequente della variabile per le correlazioni delle variabili originali. Per le correlazioni delle variabili scalate in modo ottimale è possibile scegliere il metodo di assegnazione. Selezionare Moda per sostituire i valori mancanti con la moda della variabile scalata in modo ottimale. Selezionare Categoria distinta per sostituire i valori mancanti con la quantificazione di una categoria supplementare. Ciò implica che gli oggetti con un valore mancante nella variabile specificata vengono considerati come appartenenti alla stessa categoria supplementare.

- **Assegna i valori mancanti.** Agli oggetti con valori mancanti della variabile selezionata vengono assegnati i valori ed è possibile scegliere il metodo di assegnazione. Selezionare Moda per sostituire i valori mancanti con la categoria più frequente. Se sono disponibili più mode, verrà utilizzata quella con l'indicatore di categoria minore. Selezionare Categoria distinta per sostituire i valori mancanti con la stessa quantificazione di una categoria supplementare. Ciò implica che gli oggetti con un valore mancante nella variabile specificata vengono considerati come appartenenti alla stessa categoria supplementare.
- **Escludi gli oggetti con valori mancanti per la variabile.** Gli oggetti con valori mancanti della variabile selezionata sono esclusi dall'analisi. Questa strategia non è disponibile per le variabili supplementari.

Componenti principali categoriale: Opzioni

Nella finestra di dialogo Opzioni è possibile specificare la configurazione iniziale, i criteri di iterazione e di convergenza, un metodo di normalizzazione, il metodo per etichettare i grafici e gli oggetti supplementari.

Figura 3-6
Finestra di dialogo Opzioni



Oggetti supplementari. Specificare il numero di caso dell'oggetto (o il primo e l'ultimo numero di caso per un intervallo di oggetti) che si desidera contrassegnare come supplementare e quindi fare clic su Aggiungi. Ripetere l'operazione fino ad aver specificato tutti gli oggetti supplementari. I pesi di caso di un oggetto definito come supplementare verranno ignorati.

Metodo di normalizzazione. Per normalizzare i punteggi degli oggetti e le variabili, è possibile specificare una delle cinque opzioni seguenti. In un'analisi può essere utilizzato un solo metodo di normalizzazione.

- **Principale per variabile.** Consente di ottimizzare l'associazione tra variabili. Le coordinate delle variabili nello spazio dell'oggetto sono i pesi di componente, ovvero le correlazioni con le componenti principali, quali le dimensioni e i punteggi degli oggetti. Questo metodo risulta utile se la correlazione tra variabili riveste un'importanza fondamentale.
- **Principale per oggetto.** Questo metodo consente di ottimizzare le distanze tra gli oggetti e risulta utile se le dissimilarità o le similarità tra gli oggetti sono di importanza fondamentale.
- **Simmetrico.** Utilizzare questo metodo di normalizzazione se la relazione tra oggetti e variabili è di importanza fondamentale.

- **Indipendente.** Utilizzare questo metodo se si desidera esaminare separatamente le distanze tra gli oggetti e le correlazioni tra le variabili.
- **Personalizzata.** È possibile specificare qualsiasi valore reale compreso nell'intervallo [–1, 1]. Il valore 1 equivale al metodo Principale per oggetto, il valore 0 equivale al metodo Simmetrico e il valore –1 equivale al metodo Principale per variabile. Se si specifica un valore maggiore di –1 e minore di 1, è possibile disperdere l'autovalore negli oggetti e nelle variabili. Questo metodo è utile per creare biplot o triplot adatti alle specifiche esigenze.

Criteri. È possibile specificare il numero massimo di iterazioni che possono essere eseguite dalla procedura durante i calcoli e inoltre selezionare un valore per il criterio di convergenza. L'algoritmo si interrompe se la differenza dell'adattamento totale delle due ultime iterazioni è inferiore al valore di convergenza o se viene raggiunto il numero massimo di iterazioni.

Etichetta i grafici in base a. Consente di specificare se nei grafici verranno utilizzati le variabili e le etichette dei valori o i nomi delle variabili e i valori. È inoltre possibile specificare una lunghezza massima per le etichette.

Dimensioni del grafico. Consente di controllare le dimensioni visualizzate nell'output.

- **Visualizza tutte le dimensioni nella soluzione.** Tutte le dimensioni nella soluzione sono visualizzate in una matrice di grafici a dispersione.
- **Limita il numero di dimensioni.** Le dimensioni visualizzate sono limitate alle coppie inserite nel grafico. Se le dimensioni vengono limitate è necessario selezionare la dimensione maggiore e minore da inserire nel grafico. La dimensione minore può variare da 1 al numero delle dimensioni nella soluzione meno 1 e viene inserita nel grafico a confronto con le dimensioni maggiori. La dimensione maggiore può variare da 2 al numero delle dimensioni nella soluzione e indica la dimensione massima da utilizzare nell'inserimento nel grafico delle coppie di dimensioni. Questa specifica si applica a tutti i grafici multidimensionali richiesti.

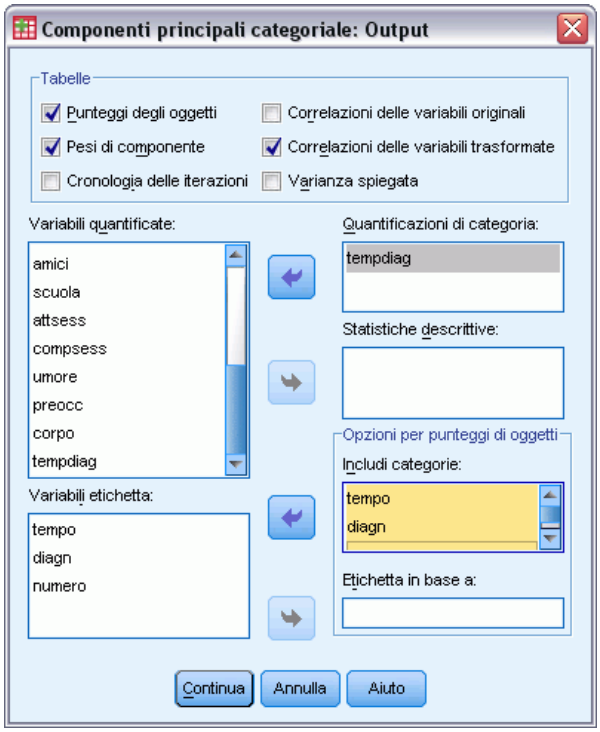
Configurazione. È possibile leggere i dati da un file che contiene le coordinate di una configurazione. La prima variabile del file deve contenere le coordinate della prima dimensione, la seconda variabile le coordinate della seconda dimensione e così via.

- **Iniziale.** La configurazione del file specificato verrà utilizzata come punto di partenza dell'analisi.
- **Fissa.** La configurazione del file specificato verrà utilizzata per inserire le variabili. Le variabili inserite devono essere selezionate come variabili dell'analisi, ma poiché la configurazione è fissa, vengono considerate come variabili supplementari e pertanto non è necessario selezionarle come variabili supplementari.

Componenti principali categoriale: Output

La finestra di dialogo Output consente di creare tabelle per i punteggi degli oggetti, i pesi di componente, la cronologia delle iterazioni, le correlazioni delle variabili originali e trasformate, le varianze spiegate per variabile e per dimensione, le quantificazioni di categoria delle variabili selezionate e le statistiche descrittive delle variabili selezionate.

Figura 3-7
Finestra di dialogo Output



Punteggi degli oggetti. Visualizza i punteggi degli oggetti e include le seguenti opzioni:

- **Includi categorie.** Visualizza gli indicatori di categoria per le variabili dell'analisi selezionate.
- **Etichetta in base a.** Per etichettare gli oggetti è possibile selezionare una variabile dall'elenco di variabili etichetta.

Pesi di componente. Visualizza pesi di componente per tutte le variabili a cui non sono stati assegnati livelli di scaling nominale multipli.

Cronologia iterazioni. Visualizza la varianza spiegata, la perdita e l'aumento della varianza spiegata per ciascuna iterazione.

Correlazioni delle variabili originali. Visualizza la matrice di correlazione delle variabili originali e gli autovalori di tale matrice.

Correlazioni delle variabili trasformate. Visualizza la matrice di correlazione delle variabili trasformate (con scaling ottimale) e gli autovalori di tale matrice.

Varianza spiegata. Visualizza l'entità della varianza spiegata in base alle coordinate del centroide, alle coordinate del vettore e al totale (combinazione delle coordinate del centroide e del vettore) per variabile e per dimensione.

Quantificazioni di categoria. Fornisce le quantificazioni di categoria e le coordinate per ciascuna dimensione delle variabili selezionate

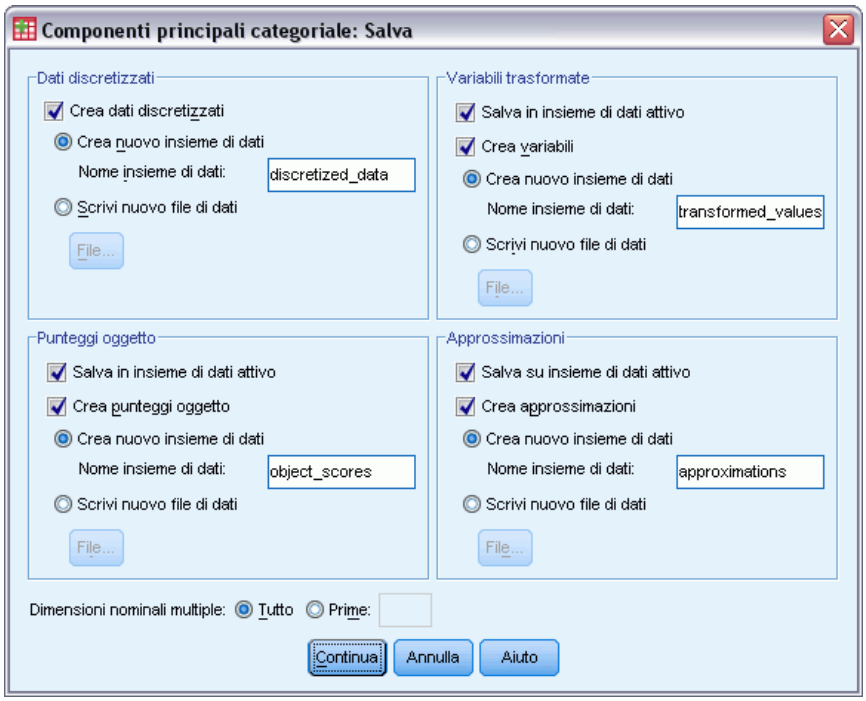
Statistiche descrittive. Visualizza le frequenze, il numero di valori mancanti e la moda delle variabili selezionate.

Componenti principali categoriale: Salva

Dalla finestra di dialogo Salva è possibile salvare i dati discretizzati, i punteggi degli oggetti, i valori trasformati e le approssimazioni in un file di dati esterno di IBM® SPSS® Statistics o un insieme di dati nella sessione corrente. Nel file di dati attivo è inoltre possibile salvare i valori trasformati, i punteggi degli oggetti e le approssimazioni.

- I file di dati sono disponibili durante la sessione corrente, ma non lo sono in quelle successive a meno che non li si salvi esplicitamente come file di dati. I nomi degli insiemi di dati devono rispettare le regole dei nomi delle variabili.
- I nomi dei file o i nomi dei file di dati devono essere diversi per ogni tipo di dati salvati.
- Se si salvano i punteggi degli oggetti o i valori trasformati nel file di dati attivo, è possibile specificare il numero delle dimensioni nominali multiple.

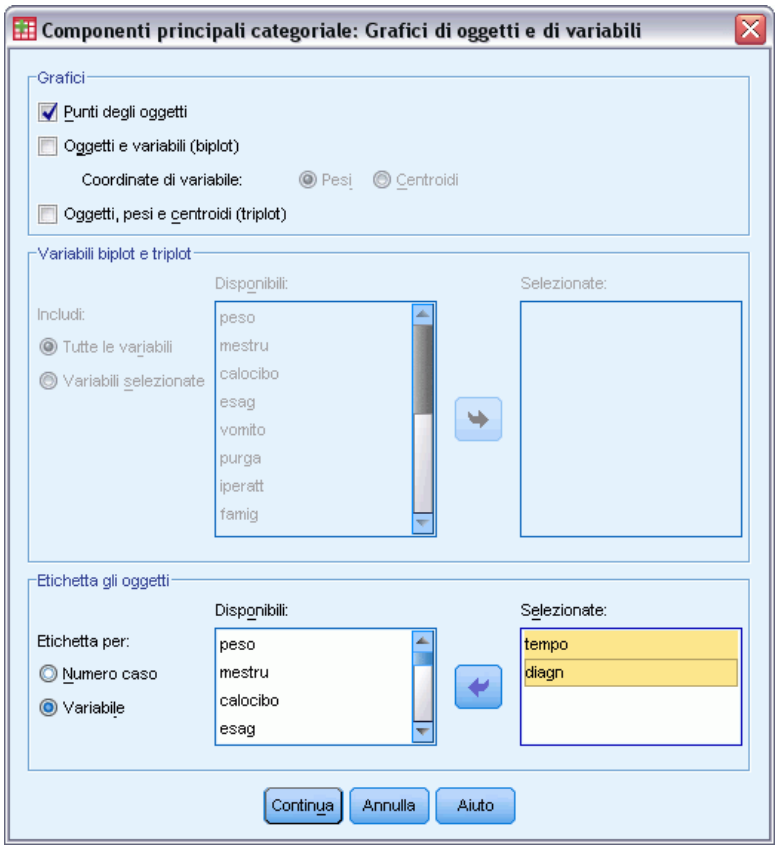
Figura 3-8
Salva



Componenti principali categoriale: Grafici di oggetti e di variabili

Nella finestra di dialogo Grafici di oggetti e di variabili è possibile specificare i tipi di grafici desiderati e le variabili per cui essi verranno creati.

Figura 3-9
Finestra di dialogo Grafici di oggetti e di variabili



Punti degli oggetti. Viene visualizzato un grafico dei punti degli oggetti.

Oggetti e variabili (biplot). I punti degli oggetti vengono tracciati nel grafico in base alle coordinate di variabile specificate, ovvero i pesi di componente e i centroidi di variabili.

Oggetti, pesi e centroidi (triplot). I punti degli oggetti vengono tracciati nel grafico in base ai centroidi di variabili con livelli di scaling nominale multipli e ai pesi di componente di altre variabili.

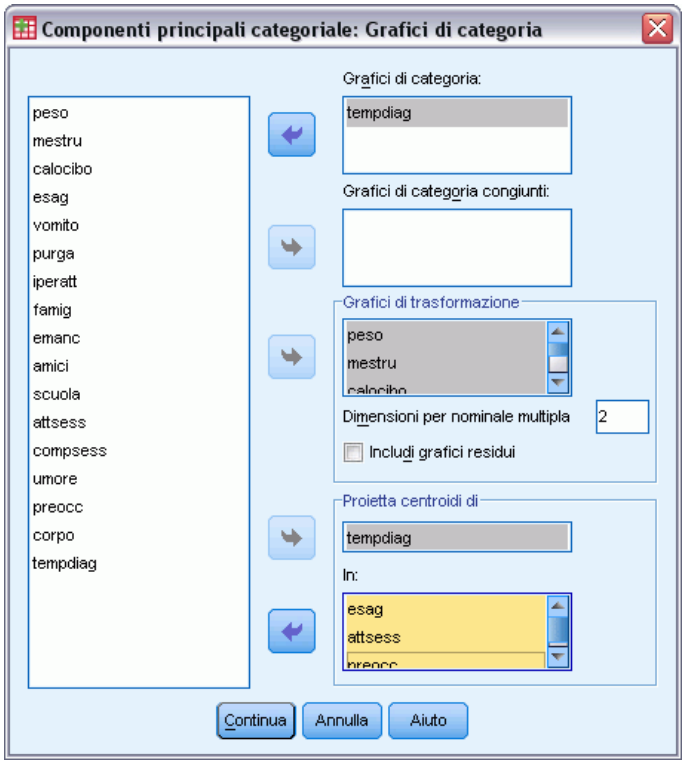
Variabili biplot e triplot. Per i biplot e i triplot è possibile utilizzare tutte le variabili o selezionarne un sottoinsieme.

Etichetta gli oggetti. È possibile etichettare gli oggetti con le categorie delle variabili selezionate (scegliendo i valori degli indicatori di categoria o le etichette dei valori nella finestra di dialogo Opzioni) oppure con i relativi numeri di caso. Se è selezionata l'opzione Variabile, viene creato un grafico per ogni variabile.

Componenti principali categoriale: Grafici di categoria

Nella finestra di dialogo Grafici di categoria è possibile specificare i tipi di grafici desiderati e le variabili per cui tali grafici verranno creati.

Figura 3-10
Finestra di dialogo Grafici di categoria



Grafici di categoria. Per ciascuna variabile selezionata viene creato un grafico delle coordinate del centroide e del vettore. Per le variabili con livelli di scaling nominale multipli, le categorie si trovano nei centroidi degli oggetti delle specifiche categorie. Per tutti gli altri livelli di scaling, le categorie sono su un vettore che passa per l'origine.

Grafici di categoria congiunti. Si tratta di un singolo grafico delle coordinate del centroide e del vettore relativo a ciascuna variabile selezionata.

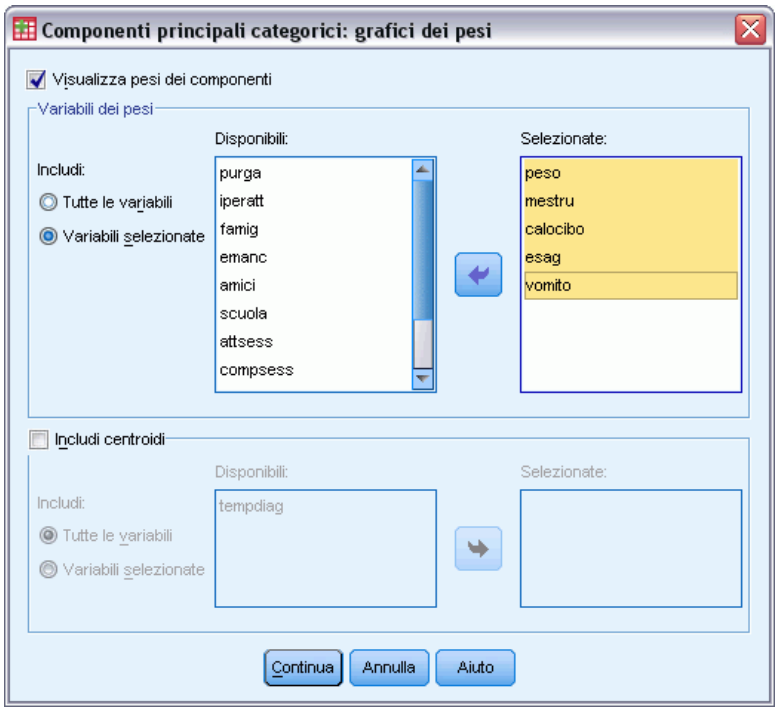
Grafici di trasformazione. Visualizza un grafico delle quantificazioni di categoria ottimali rispetto agli indicatori di categoria. È possibile specificare il numero di dimensioni desiderato per le variabili con livelli di scaling nominale multipli. Verrà generato un grafico per ciascuna dimensione. È inoltre possibile visualizzare i grafici dei residui per ciascuna variabile selezionata.

Proietta centroidi di. È possibile scegliere una variabile e proiettarne i relativi centroidi nelle variabili selezionate, che non possono tuttavia essere variabili con livelli di scaling nominale multipli. Insieme a questo grafico, viene visualizzata anche una tabella con le coordinate dei centroidi proiettati.

Componenti principali categoriale: Grafici dei pesi

Nella finestra di dialogo Grafici dei pesi è possibile specificare le variabili che verranno incluse nel grafico e se inserirvi o meno i centroidi.

Figura 3-11
Finestra di dialogo Grafici dei pesi fattoriali



Visualizza pesi di componente. Se selezionata, viene visualizzato un grafico dei pesi di componente.

Variabili dei pesi. Per il grafico dei pesi di componente è possibile utilizzare tutte le variabili o selezionarne un sottoinsieme.

Includi i centroidi. Le variabili con livelli di scaling nominale multipli non hanno pesi di componente, ma è possibile scegliere di includere nel grafico i centroidi di tali variabili. È possibile scegliere di utilizzare tutte le variabili nominali multiple o selezionarne un sottoinsieme.

Opzioni aggiuntive del comando CATPCA

Per personalizzare l'analisi delle componenti principali categoriale è possibile incollare le impostazioni selezionate in una finestra di sintassi e modificare la sintassi di comando CATPCA risultante. Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare i nomi di radice per le variabili trasformate, i punteggi degli oggetti e le approssimazioni quando vengono salvati nel file dati attivo (con il sottocomando SAVE).
- Specificare una lunghezza massima per le singole etichette di ciascun grafico (con il sottocomando PLOT).
- Specificare un elenco di variabili distinto per i grafici dei residui (con il sottocomando PLOT).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
4

Analisi della correlazione canonica non lineare (OVERALS)

L'analisi della correlazione canonica non lineare corrisponde all'analisi della correlazione canonica categoriale con scaling ottimale. Questa procedura consente di determinare la correlazione tra insiemi simili di variabili categoriali. È conosciuta anche con l'acronimo OVERALS.

L'analisi della correlazione canonica standard è un'estensione della regressione multipla, in cui il secondo insieme non contiene una sola variabile di risposta ma più variabili di risposta. L'obiettivo è quello di spiegare la maggior parte dei valori di varianza osservati nelle relazioni tra due insiemi di variabili numeriche in uno spazio dimensionale ridotto. Le variabili di ciascun insieme vengono inizialmente combinate linearmente in modo che la correlazione tra le combinazioni lineari sia massima. In base a tali combinazioni vengono determinate le combinazioni lineari successive non correlate con le precedenti e con la maggiore correlazione possibile.

L'approccio di scaling ottimale consente di estendere l'analisi standard in tre modi fondamentali. Innanzitutto, OVERALS consente di utilizzare più di due insiemi di variabili. In secondo luogo, le variabili possono essere scalate come nominali, ordinali o numeriche ed è quindi possibile analizzare le relazioni non lineari tra le variabili. Infine, anziché massimizzare le correlazioni tra gli insiemi di variabili, è possibile confrontare gli insiemi con un insieme intermedio non conosciuto definito dai punteggi degli oggetti.

Esempio. L'analisi della correlazione canonica categoriale con scaling ottimale consente di visualizzare graficamente la relazione esistente tra un insieme di variabili che include la categoria lavorativa e il livello di istruzione e un altro insieme di variabili che include l'area di residenza e il genere. Ci si può rendere conto che il livello di istruzione e l'area di residenza comportano una discriminazione maggiore rispetto alle altre variabili, oppure che il livello di istruzione comporta una maggiore discriminazione nella prima dimensione.

Statistiche e grafici. Frequenze, centroidi, cronologia delle iterazioni, punteggi degli oggetti, quantificazioni di categoria, pesi, pesi di componente, adattamento singolo e multiplo, grafici dei punteggi degli oggetti, grafici delle coordinate di categoria, grafici dei pesi di componente, grafici dei centroidi di categoria, grafici di trasformazione.

Dati. Utilizzare valori interi per la codifica delle variabili categoriali (livello di scaling nominale o ordinale). Per ridurre al minimo l'output, utilizzare interi consecutivi che iniziano con 1 per la codifica di ciascuna variabile. Le variabili scalate a livello numerico non devono essere ricodificate in interi consecutivi. Per ridurre al minimo l'output, per ogni variabile scalata a livello numerico sottrarre il valore osservato più piccolo da ogni valore e aggiungere 1. I valori frazionari vengono troncati dopo i decimali.

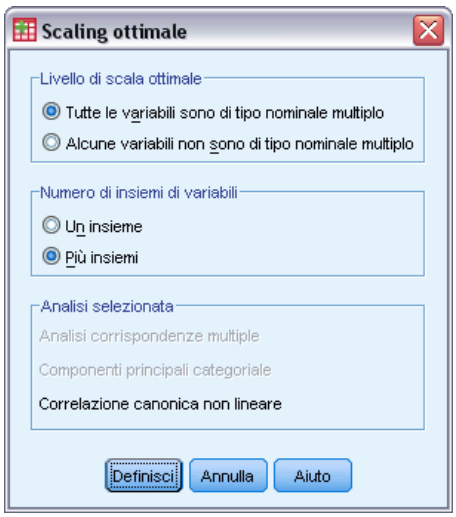
Assunzioni. Le variabili possono essere classificate in due o più insiemi. Le variabili dell'analisi vengono scalate come nominali multiple, nominali singole, ordinali o numeriche. Il numero massimo di dimensioni utilizzate nella procedura dipende dal livello di scaling ottimale delle variabili. Se tutte le variabili sono specificate come ordinali, nominali singole o numeriche, il numero massimo di dimensioni è inferiore ai due valori seguenti: Il numero di osservazioni meno 1 o il numero totale delle variabili. Se invece vengono definiti solo due insiemi di variabili, il numero massimo di dimensioni è uguale al numero delle variabili dell'insieme più piccolo. Se alcune variabili sono nominali multiple, il numero massimo di dimensioni è uguale al numero totale delle categorie nominali multiple più il numero di variabili nominali non multiple meno il numero di variabili nominali multiple. Ad esempio, se l'analisi implica cinque variabili, una delle quali è nominale multipla e ha quattro categorie, il numero massimo di categorie è uguale a 7 (4 + 4 – 1). Se si specifica un numero maggiore del massimo, verrà utilizzato il valore massimo.

Procedure correlate. Se ciascun insieme contiene una variabile, l'analisi della correlazione canonica non lineare equivale all'analisi delle componenti principali con scaling ottimale. Se ciascuna di tali variabili è nominale multipla, l'analisi corrisponde all'analisi delle corrispondenze multiple. Se sono implicati due insiemi di variabili e uno di essi contiene solo una variabile, l'analisi è identica alla regressione categoriale con scaling ottimale.

Per ottenere un'analisi della correlazione canonica non lineare

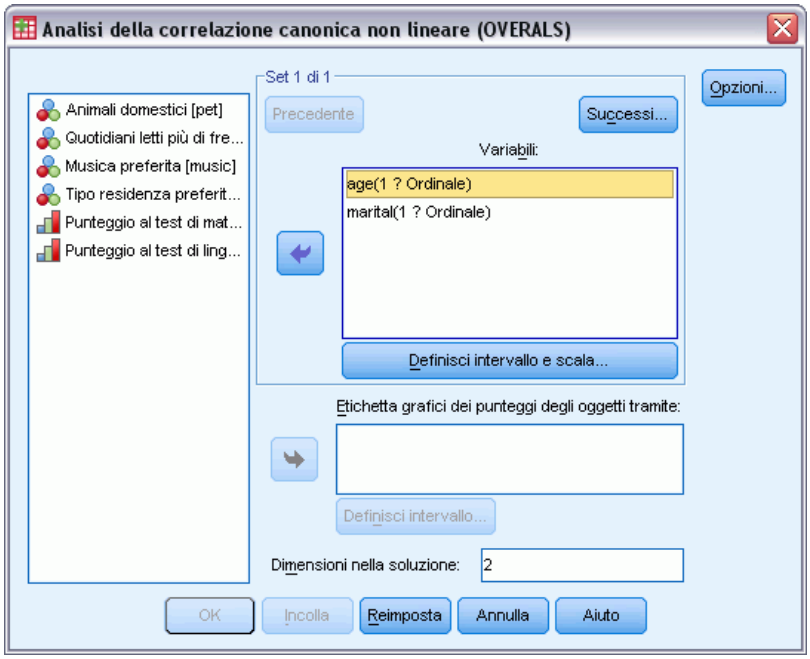
- Dai menu, scegliere:
  Analizza > Riduzioni dimensione > Scaling ottimale...

Figura 4-1
Finestra di dialogo Scaling ottimale



- Selezionare Tutte le variabili non nominali multiple o Una o più variabili nominali multiple.
- Selezionare Più insiemi.
- Fare clic su Definisci.

Figura 4-2
Finestra di dialogo Analisi della correlazione canonica non lineare (OVERALS).



- Specificare almeno due insiemi di variabili. Selezionare le variabili che si desidera includere nel primo insieme. Per passare all'insieme successivo, fare clic su Successivo e quindi selezionare le variabili da inserire nel secondo insieme. È possibile aggiungere il numero di insiemi desiderato. Fare clic su Precedente per tornare all'insieme di variabili definito in precedenza.

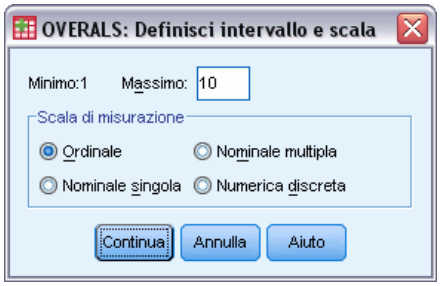
- Specificare l'intervallo di valori e la scala di misurazione (livello di scaling ottimale) per ciascuna variabile selezionata.

- Fare clic su OK.

- Oppure:
  - Selezionare una o più variabili per definire le etichette dei punti per i grafici dei punteggi degli oggetti. Per ciascuna variabile viene creato un grafico distinto, nel quale i punti sono etichettati in base ai valori della specifica variabile. È necessario definire un intervallo per ciascuna di tali variabili etichetta dei grafici. Nella finestra di dialogo, non è possibile definire contemporaneamente la stessa variabile come variabile dell'analisi e come variabile etichetta. Per etichettare il grafico dei punteggi degli oggetti mediante una variabile utilizzata nell'analisi, creare una copia della variabile scegliendo Calcola dal menu Trasforma e quindi utilizzare la nuova variabile per etichettare il grafico. In alternativa, utilizzare la sintassi di comando.
  - Specificare il numero di dimensioni desiderato per la soluzione. In genere si sceglie un numero di dimensioni sufficiente a spiegare la maggior parte della variazione. Se l'analisi implica più di due dimensioni, verranno creati grafici tridimensionali delle prime tre dimensioni. Le altre dimensioni possono essere visualizzate modificando il grafico.

Definisci intervallo e scala

Figura 4-3
Finestra di dialogo Definisci intervallo e scala



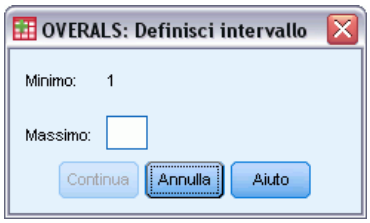
È necessario definire un intervallo per ciascuna variabile. Il valore massimo specificato deve essere un intero. I valori frazionari vengono troncati nell'analisi e i valori di categoria al di fuori dell'intervallo specificato vengono ignorati. Per ridurre al minimo l'output, utilizzare il comando Ricodifica automatica del menu Trasforma per creare categorie consecutive che iniziano con 1 per le variabili considerate come nominali o ordinali. La ricodifica in interi consecutivi è sconsigliabile per le variabili scalate a livello numerico. Per ridurre al minimo l'output per le variabili considerate come numeriche, sottrarre il valore minimo da ogni valore e aggiungere 1.

È inoltre necessario selezionare il livello di scaling da utilizzare per quantificare ciascuna variabile.

- **Ordinale.** Nella variabile quantificata viene mantenuto l'ordine delle categorie della variabile osservata.
- **Nominale singola.** Nella variabile quantificata, agli oggetti della stessa categoria è assegnato un punteggio uguale.
- **Nominale multipla.** Le quantificazioni possono variare a seconda della dimensione.
- **Numerica discreta.** Le categorie vengono considerate come ordinate ed equamente distanziate. Le differenze tra i numeri delle categorie e l'ordine delle categorie della variabile osservata vengono mantenute nella variabile quantificata.

Definisci intervallo

Figura 4-4
Finestra di dialogo Definisci intervallo



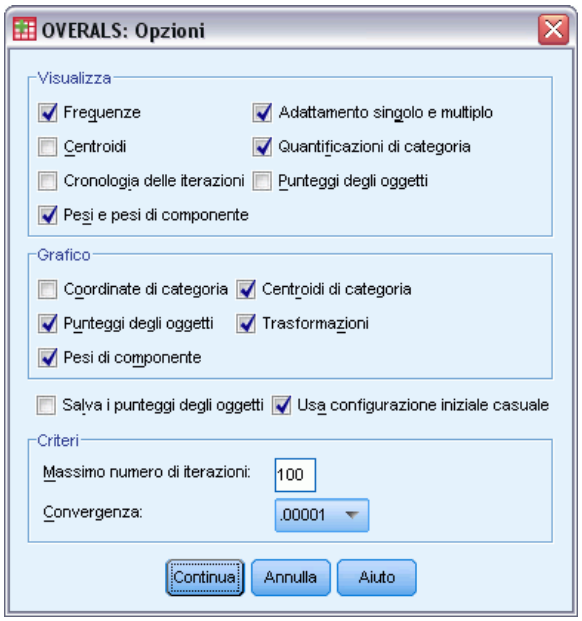
È necessario definire un intervallo per ciascuna variabile. Il valore massimo specificato deve essere un intero. I valori frazionari vengono troncati nell'analisi e i valori di categoria al di fuori dell'intervallo specificato vengono ignorati. Per ridurre al minimo l'output, utilizzare il comando Ricodifica automatica del menu Trasforma per creare categorie consecutive che iniziano con 1.

È inoltre necessario definire un intervallo per ciascuna variabile utilizzata per etichettare i grafici dei punteggi degli oggetti. Le etichette per le categorie con valori al di fuori dell'intervallo definito per la variabile non verranno tuttavia visualizzate nei grafici.

Analisi della correlazione canonica non lineare (OVERALS): Opzioni

Nella finestra di dialogo Opzioni è possibile selezionare le statistiche e i grafici opzionali, salvare i punteggi degli oggetti come nuove variabili nel file di dati attivo, specificare i criteri di iterazione e di convergenza nonché la configurazione iniziale dell'analisi.

Figura 4-5
Finestra di dialogo Opzioni



Visualizzazione. Le statistiche disponibili sono: frequenze marginali (conteggi), centroidi, cronologia delle iterazioni, pesi e pesi di componente, quantificazioni di categoria, punteggi degli oggetti e statistiche dell'adattamento singolo e multiplo.

- **Centroidi.** Quantificazioni di categoria e medie proiettate ed effettive dei punteggi per oggetti (casi) inclusi in ogni insieme e appartenenti alle medesime categorie di una variabile.
- **Pesi e pesi di componente.** I coefficienti di regressione in ciascuna dimensione per ogni variabile quantificata di un insieme, dove i punteggi di oggetto sono regressi sulle variabili quantificate e sulla proiezione delle variabili quantificate nello spazio. Fornisce un'indicazione del contributo di ciascuna variabile alla dimensione, all'interno di ogni insieme.
- **Adattamento singolo e multiplo.** Misure della bontà di adattamento delle coordinate di una o più categorie o delle quantificazioni di categoria rispetto agli oggetti.
- **Quantificazioni di categoria.** Punteggio ottimale assegnato alle categoria di ciascuna variabile.
- **Punteggi degli oggetti.** Quantificazione ottimale assegnata ad un oggetto (caso) in una delle dimensioni.

Grafico. È possibile creare grafici delle coordinate di categoria, dei punteggi degli oggetti, dei pesi di componente, dei centroidi delle categorie e delle trasformazioni.

Salva i punteggi degli oggetti. I punteggi degli oggetti possono essere salvati come nuove variabili nel file di dati attivo. Verranno salvati i punteggi relativi al numero di dimensioni specificato nella finestra di dialogo principale.

Usa configurazione iniziale casuale. Se alcune o tutte le variabili sono nominali singole, è consigliabile selezionare questa opzione. Se l'opzione non è selezionata, verrà utilizzata una configurazione iniziale nidificata.

Criteri. È possibile specificare il numero massimo di iterazioni che l'analisi della correlazione canonica non lineare può eseguire durante i calcoli e inoltre selezionare un valore per il criterio di convergenza. L'analisi si interrompe se la differenza dell'adattamento totale delle due ultime iterazioni è inferiore al valore di convergenza o se viene raggiunto il numero massimo di iterazioni.

Opzioni aggiuntive del comando OVERALS

Per personalizzare l'analisi della correlazione canonica non lineare, è possibile incollare le impostazioni selezionate in una finestra di sintassi e quindi modificare la sintassi del comando `OVERALS` così ottenuta. Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare le coppie di dimensioni da inserire nel grafico, evitando così di inserire tutte le dimensioni estratte (mediante la parola chiave `NDIM` del sottocomando `PLOT`).
- Specificare il numero dei caratteri delle etichette dei valori utilizzate per etichettare i punti dei grafici (mediante il sottocomando `PLOT`).
- Designare più di cinque variabili come variabili etichetta per i grafici dei punteggi degli oggetti (mediante il sottocomando `PLOT`).
- Selezionare le variabili utilizzate nell'analisi come variabili etichetta per i grafici dei punteggi degli oggetti (mediante il sottocomando `PLOT`).
- Selezionare le variabili che definiscono le etichette dei punti per il grafico del punteggio della quantificazione (mediante il sottocomando `PLOT`).
- Specificare il numero dei casi da includere nell'analisi se non si desidera utilizzare tutti i casi disponibili nel file di dati attivo (mediante il sottocomando `NOBSERVATIONS`).
- Specificare i nomi di radice per le variabili create salvando i punteggi degli oggetti (mediante il sottocomando `SAVE`).
- Specificare il numero di dimensioni da salvare, evitando quindi di salvare tutte le dimensioni estratte (mediante il sottocomando `SAVE`).
- Scrivere le quantificazioni di categoria in un file matrice (mediante il sottocomando `MATRIX`).
- Creare grafici a bassa risoluzione e più facili da leggere rispetto ai grafici ad alta risoluzione (mediante il comando `SET`).
- Creare grafici dei centroidi e delle trasformazioni unicamente per le variabili selezionate (con il sottocomando `PLOT`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
5

Analisi corrispondenze

Uno degli obiettivi dell'analisi delle corrispondenze è descrivere le relazioni esistenti tra due variabili nominali di una tabella di corrispondenza in uno spazio dimensionale ridotto e al tempo stesso descrivere le relazioni tra le categorie di ciascuna variabile.. Per ciascuna variabile, le distanze tra i punti delle categorie riportati in un grafico riflettono le relazioni tra le categorie e le categorie simili vengono inserite nel grafico una accanto all'altra. La relazione tra le variabili viene descritta dalla proiezione dei punti di una variabile sul vettore dall'origine a un punto di categoria dell'altra variabile.

L'analisi delle tavole di contingenza spesso include lo studio dei profili di riga e di colonna e i test di indipendenza mediante la statistica Chi-quadrato. Il numero dei profili può tuttavia essere piuttosto elevato e il test Chi-quadrato non è in grado rilevare la struttura della dipendenza. La procedura Tavole di contingenza rende disponibili numerosi test e misure di associazione, ma non può rappresentare graficamente le relazioni tra le variabili.

L'analisi fattoriale è una tecnica standard per la descrizione delle relazioni tra le variabili all'interno di uno spazio dimensionale ridotto, ma richiede tuttavia dati per intervallo e il numero di osservazioni deve essere pari a cinque volte il numero delle variabili. L'analisi delle corrispondenze, invece, assume variabili nominali ed è in grado di descrivere le relazioni tra le categorie di ciascuna variabile nonché la relazione tra le variabili. Può inoltre essere utilizzata per l'analisi di qualsiasi tabella di misure di corrispondenza positive.

Esempio. È possibile utilizzare l'analisi delle corrispondenze per visualizzare graficamente le relazioni tra la categoria lavorativa e le abitudini correlate al fumo. Si può scoprire che, per quanto riguarda il fumo, il comportamento dei manager di livello inferiore si differenzia da quello delle segretarie e che invece quello delle segretarie non si differenzia dal comportamento dei manager di livello superiore, oppure che i manager di livello inferiore fumano molto, mentre le segretarie fumano poco.

Statistiche e grafici. Misure di corrispondenza, profili di riga e di colonna, valori singolari, punteggi di riga e di colonna, inerzia, massa, statistiche del punteggio di riga e di colonna, statistiche di confidenza di valori singolari, grafici di trasformazione, grafici a punti di colonna e di riga e biplot.

Dati. Le variabili categoriali da analizzare vengono scalate in modo nominale. Per i dati aggregati o per una misura di corrispondenza diversa dalle frequenze, utilizzare una variabile peso con valori di similarità positivi. In alternativa, utilizzare la sintassi per leggere i dati della tabella.

Assunzioni. Il numero massimo di dimensioni utilizzate nella procedura dipende dal numero di righe attive e di categorie di colonna e dal numero dei vincoli di uguaglianza. Se non esistono vincoli e tutte le categorie sono attive, il numero massimo è inferiore di uno rispetto al numero delle categorie della variabile con il numero minimo di categorie. Se, ad esempio, una variabile ha cinque categorie e l'altra variabile ne ha quattro, il numero massimo di dimensioni è pari a tre. Le

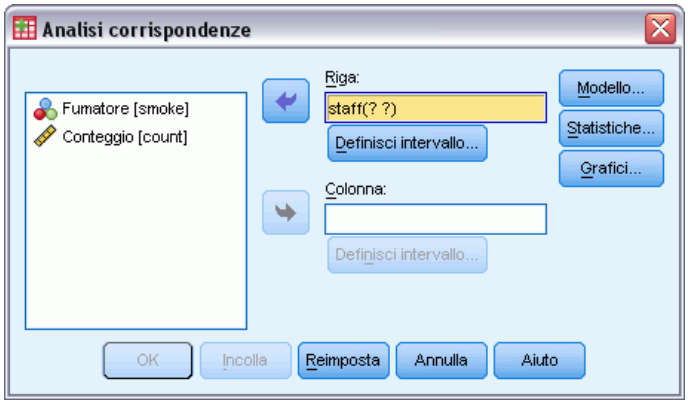
categorie supplementari non sono attive. Se, ad esempio, una variabile ha cinque categorie di cui due sono supplementari e l'altra variabile ha quattro categorie, il numero massimo di dimensioni è pari a due. Tutti gli insiemi di categorie vincolati devono essere considerati come un'unica categoria. Se, ad esempio, una variabile ha cinque categorie e per tre di esse è valido il vincolo di uguaglianza, per stabilire il numero massimo di dimensioni sarà necessario considerare che la variabile abbia tre categorie. Due delle categorie non sono vincolate, mentre la terza categoria corrisponde alle tre categorie vincolate. Se si specifica un numero di dimensioni maggiore del massimo, verrà utilizzato il valore massimo.

Procedure correlate. Se sono implicate più di due variabili, utilizzare l'analisi delle corrispondenze multiple. Se le variabili devono essere scalate in modo ordinale, utilizzare l'analisi delle componenti principali categoriale.

Per ottenere un'analisi delle corrispondenze

- Dai menu, scegliere:
  Analizza > Riduzioni dimensione > Analisi corrispondenze...

Figura 5-1
Finestra di dialogo Analisi della corrispondenze

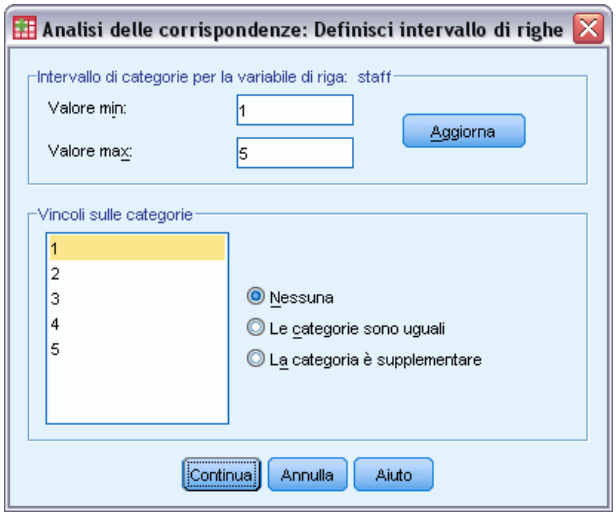


- Selezionare una variabile di riga.
- Selezionare una variabile di colonna.
- Definire gli intervalli per le variabili.
- Fare clic su OK.

Definire l'intervallo di righe nell'analisi delle corrispondenze

È necessario specificare un intervallo per la variabile di riga. I valori massimo e minimo specificati devono essere interi. I valori frazionari vengono troncati nell'analisi e i valori di categoria al di fuori dell'intervallo specificato vengono ignorati.

Figura 5-2
Finestra di dialogo Definisci intervallo di righe



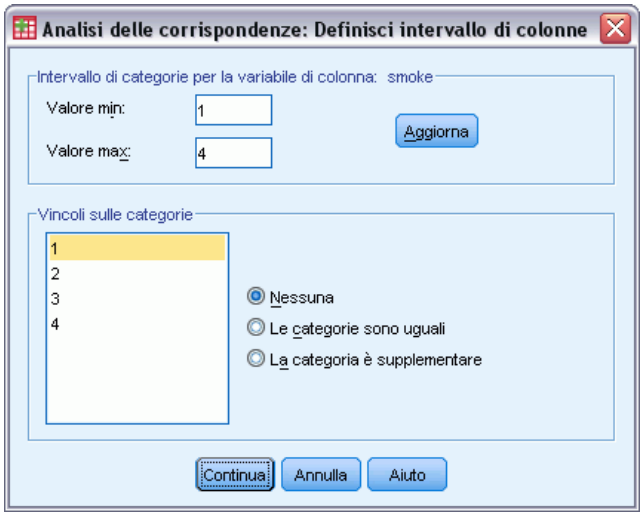
Tutte le categorie sono inizialmente libere da vincoli e attive. È possibile applicare alle categorie di riga il vincolo di essere uguali ad altre categorie di riga oppure definire una categoria di riga come supplementare.

- **Le categorie sono uguali.** I punteggi delle categorie devono essere uguali. Se l'ordine delle categorie non corrisponde alle aspettative o è di difficile comprensione, utilizzare il vincolo di uguaglianza. Il numero massimo di categorie di riga alle quali può essere applicato il vincolo di uguaglianza è pari al numero totale delle categorie di riga attive meno 1. Per applicare vincoli di uguaglianza diversi agli insiemi di categorie, utilizzare la sintassi. Ad esempio, mediante la sintassi è possibile vincolare l'uguaglianza delle categorie 1 e 2 e vincolare l'uguaglianza delle categorie 3 e 4.
- **La categoria è supplementare.** Le categorie supplementari non influiscono sull'analisi, ma devono essere rappresentate nello spazio definito dalle categorie attive. Non influiscono in alcun modo sulla definizione delle dimensioni. Il numero massimo di categorie di riga supplementari è uguale al numero totale delle categorie di riga meno 2.

Definire l'intervallo di colonne nell'analisi delle corrispondenze

È necessario specificare un intervallo per la variabile di colonna. I valori massimo e minimo specificati devono essere interi. I valori frazionari vengono troncati nell'analisi e i valori di categoria al di fuori dell'intervallo specificato vengono ignorati.

Figura 5-3
Finestra di dialogo Definisci intervallo di colonne



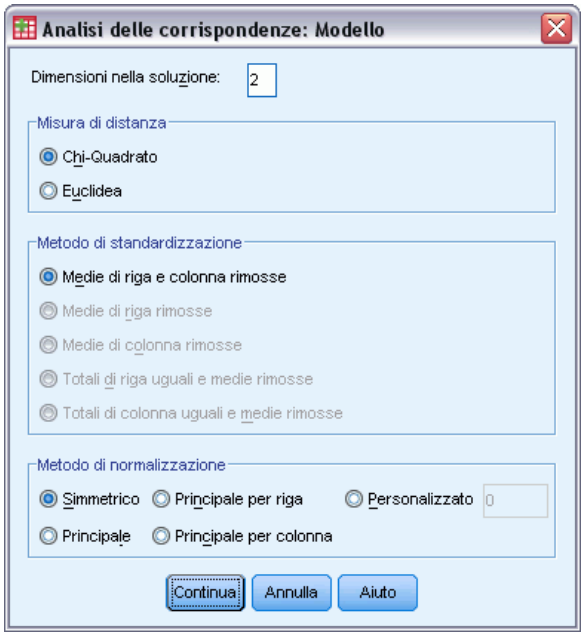
Tutte le categorie sono inizialmente libere da vincoli e attive. È possibile applicare alle categorie di colonna il vincolo di essere uguali ad altre categorie di colonna oppure definire una categoria di colonna come supplementare.

- **Le categorie sono uguali.** I punteggi delle categorie devono essere uguali. Se l'ordine delle categorie non corrisponde alle aspettative o è di difficile comprensione, utilizzare il vincolo di uguaglianza. Il numero massimo di categorie di colonna alle quali può essere applicato il vincolo di uguaglianza è pari al numero totale delle categorie di colonna attive meno 1. Per applicare vincoli di uguaglianza diversi agli insiemi di categorie, utilizzare la sintassi Ad esempio, mediante la sintassi è possibile vincolare l'uguaglianza delle categorie 1 e 2 e vincolare l'uguaglianza delle categorie 3 e 4.
- **La categoria è supplementare.** Le categorie supplementari non influiscono sull'analisi, ma devono essere rappresentate nello spazio definito dalle categorie attive. Non influiscono in alcun modo sulla definizione delle dimensioni. Il numero massimo di categorie di colonna supplementari è uguale al numero totale delle categorie di colonna meno 2.

Analisi delle corrispondenze: Modello

Nella finestra di dialogo Modello è possibile specificare il numero di dimensioni, la misura della distanza, il metodo di standardizzazione e il metodo di normalizzazione.

Figura 5-4
Finestra di dialogo Modello



Dimensioni nella soluzione. Specificare il numero di dimensioni. In genere si sceglie un numero di dimensioni sufficiente a spiegare la maggior parte della variazione. Il numero massimo di dimensioni dipende dal numero delle categorie attive utilizzate nell'analisi e dai vincoli di uguaglianza. Il numero massimo è minore di:

- il numero delle categorie di riga attive meno il numero delle categorie di riga vincolate all'uguaglianza, più il numero degli insiemi di categorie di riga vincolati.
- Il numero delle categorie di colonna attive meno il numero delle categorie di colonna vincolate all'uguaglianza, più il numero degli insiemi di categorie di colonna vincolati.

Misura di distanza. È possibile selezionare la misura della distanza tra le righe e le colonne della tabella delle corrispondenze. Sono disponibili i seguenti metodi:

- **Chi-quadrato.** Utilizza una distanza di profilo ponderata, dove il peso corrisponde alla massa delle righe e delle colonne. È richiesta per l'analisi delle corrispondenze standard.
- **Euclidea.** Utilizza la radice quadrata della somma delle differenze quadratiche tra le coppie di righe e le coppie di colonne.

Metodo di standardizzazione. Selezionare una delle alternative seguenti:

- **Medie di riga e colonna rimosse.** Vengono centrate sia le righe che le colonne. Questo metodo è richiesto per l'analisi delle corrispondenze standard .
- **Medie di riga rimosse.** Vengono centrate solo le righe.
- **Medie di colonna rimosse.** Vengono centrate solo le colonne.
- **Totali di riga uguali e medie rimosse.** Prima di centrare le righe vengono equalizzati i relativi margini.
- **Totali di colonna uguali e medie rimosse.** Prima di centrare le colonne vengono equalizzati i relativi margini.

Metodo di normalizzazione. Selezionare una delle alternative seguenti:

- **Simmetrico.** Per ciascuna dimensione, i punteggi di riga sono uguali alla media ponderata dei punteggi di colonna divisa per il valore singolare corrispondente e i punteggi di colonna sono uguali alla media ponderata dei punteggi di riga divisa per il valore singolare corrispondente. Utilizzare questo metodo se si desidera analizzare le differenze o le similarità tra le categorie delle due variabili.
- **Principale**. Le distanze tra i punti di riga e i punti di colonna sono approssimazioni delle distanze riportate nella tabella delle corrispondenze in base alla misura della distanza selezionata. Usare questo metodo per analizzare le differenze tra le categorie di una o entrambe le variabili anziché le differenze tra le due variabili.
- **Principale per riga.** Le distanze tra i punti di riga sono approssimazioni delle distanze riportate nella tabella delle corrispondenze in base alla misura della distanza selezionata. I punteggi di riga sono la media ponderata dei punteggi di colonna. ed è adatto per esaminare le differenze fra le categorie della variabile di riga.
- **Principale per colonna.** Le distanze tra i punti di colonna sono approssimazioni delle distanze riportate nella tabella delle corrispondenze in base alla misura della distanza selezionata. I punteggi di colonna sono la media ponderata dei punteggi di riga. Utilizzare questo metodo se si desidera analizzare le differenze o le similarità tra le categorie della variabile di colonna.
- **Personalizzata.** È necessario specificare un valore compreso tra –1 e 1. Il valore –1 corrisponde al valore principale per colonna, il valore 1 al valore principale per riga e il valore 0 al valore simmetrico. Tutti gli altri valori distribuiscono vari livelli di inerzia nei punteggi sia di riga che di colonna. Questo metodo è utile per creare biplot personalizzati.

Analisi delle corrispondenze: Statistiche

Nella finestra di dialogo Statistiche è possibile specificare l'output numerico.

Figura 5-5
Finestra di dialogo Statistiche

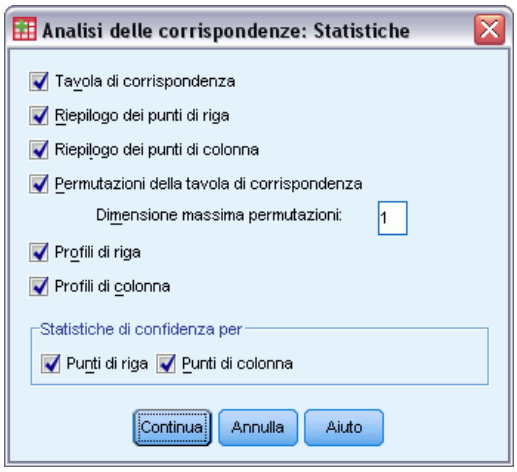


Tabella di corrispondenza. Tavola di contingenza delle variabili di input con i totali marginali di riga e di colonna.

Riassunto dei punti di riga. Per ciascuna categoria di riga, i punteggi, la massa, l'inerzia, il contributo all'inerzia della dimensione e il contributo della dimensione all'inerzia del punto.

Riassunto dei punti di colonna. Per ciascuna categoria di colonna, i punteggi, la massa, l'inerzia, il contributo all'inerzia della dimensione e il contributo della dimensione all'inerzia del punto.

Profili di riga. Per ciascuna categoria di riga, la distribuzione della variabile di colonna nelle categorie.

Profili di colonna. Per ciascuna categoria di colonna, la distribuzione della variabile di riga nelle categorie.

Permutazioni della tavola di corrispondenza. Tavola di corrispondenza riorganizzata in modo tale che le righe e le colonne sono disposte in ordine crescente in base ai punteggi nella prima dimensione. È inoltre possibile specificare il numero massimo di dimensioni per cui verranno create tabelle permutate. Viene creata una tabella permutata per ciascuna dimensione a partire da 1 fino al numero specificato.

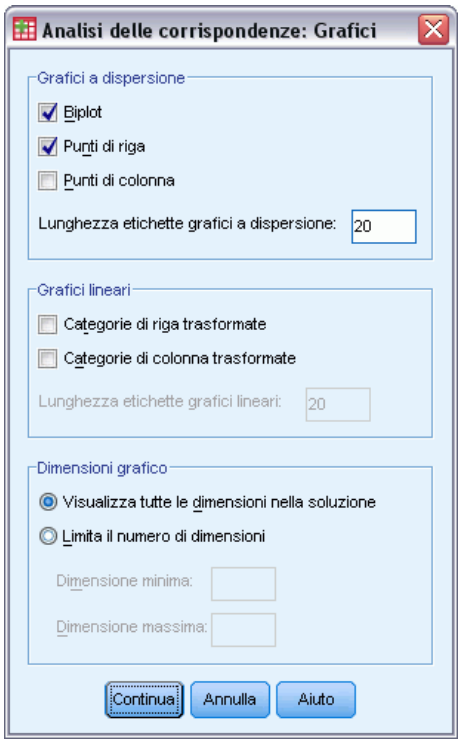
Statistiche di confidenza per Punti di riga. Include la deviazione standard e le correlazioni per tutti i punti di riga non supplementari.

Statistiche di confidenza per Punti di colonna. Include la deviazione standard e le correlazioni per tutti i punti di colonna non supplementari.

Analisi delle corrispondenze: Grafici

Nella finestra di dialogo Grafici è possibile specificare i grafici che si desidera creare.

Figura 5-6
Finestra di dialogo Grafici



Grafici a dispersione. Crea una matrice di tutti i grafici pairwise delle dimensioni. Sono disponibili i seguenti grafici a dispersione:

- **Biplot.** Crea una matrice dei grafici congiunti dei punti di riga e di colonna. Se è selezionata la normalizzazione principale, l'opzione non è disponibile.
- **Punti di riga.** Crea una matrice di grafici dei punti di riga.
- **Punti di colonna.** Crea una matrice di grafici dei punti di colonna.

È inoltre possibile specificare il numero di caratteri delle etichette dei valori da utilizzare per etichettare i punti, che deve essere un intero non negativo inferiore o uguale a 20.

Grafici lineari. Crea un grafico per ciascuna dimensione della variabile selezionata. Sono disponibili i seguenti grafici lineari:

- **Categorie di riga trasformate.** Crea un grafico dei valori delle categorie di riga originali confrontati con i punteggi di riga corrispondenti.
- **Categorie di colonna trasformate.** Crea un grafico dei valori delle categorie di colonna originali confrontati con i punteggi di colonna corrispondenti.

È inoltre possibile specificare il numero di caratteri delle etichette dei valori da utilizzare per etichettare l'asse delle categorie, che deve essere un intero non negativo inferiore o uguale a 20.

Dimensioni del grafico. Consente di controllare le dimensioni visualizzate nell'output.

- **Visualizza tutte le dimensioni nella soluzione.** Tutte le dimensioni nella soluzione sono visualizzate in una matrice di grafici a dispersione.
- **Limita il numero di dimensioni.** Le dimensioni visualizzate sono limitate alle coppie inserite nel grafico. Se le dimensioni vengono limitate è necessario selezionare la dimensione maggiore e minore da inserire nel grafico. La dimensione minore può variare da 1 al numero delle dimensioni nella soluzione meno 1 e viene inserita nel grafico a confronto con le dimensioni maggiori. La dimensione maggiore può variare da 2 al numero delle dimensioni nella soluzione e indica la dimensione massima da utilizzare nell'inserimento nel grafico delle coppie di dimensioni. Questa specifica si applica a tutti i grafici multidimensionali richiesti.

Opzioni aggiuntive del comando CORRESPONDENCE

Per personalizzare l'analisi delle corrispondenze è possibile incollare le impostazioni selezionate in una finestra di sintassi e quindi modificare la sintassi del comando `CORRESPONDENCE` così ottenuta. Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare i dati della tabella come input anziché utilizzare i dati per casi (con il sottocomando `TABLE = ALL`).
- Specificare il numero di caratteri delle etichette dei valori utilizzate per etichettare i punti di ciascun tipo di matrice dei grafici a dispersione o dei biplot (con il sottocomando `PLOT`).
- Specificare il numero di caratteri delle etichette dei valori utilizzate per etichettare i punti di ciascun tipo di grafico lineare (con il sottocomando `PLOT`).
- Scrivere una matrice di punteggi di riga e di colonna in un file dati della matrice (con il sottocomando `OUTFILE`).
- Scrivere una matrice delle statistiche di confidenza (varianze e covarianze) per i valori singolari e i punteggi in un file dati della matrice (con il sottocomando `OUTFILE`).
- Specificare più insiemi di categorie che devono essere uguali (con il sottocomando `EQUAL`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
6

Analisi corrispondenze multiple

L'analisi delle corrispondenze multiple quantifica i dati (categoriali) nominali assegnando valori numerici ai casi (oggetti) e alle categorie, in modo che gli oggetti all'interno della stessa categoria siano vicini tra loro e gli oggetti in diverse categorie siano distanti. Ciascun oggetto si trova il più vicino possibile ai punti delle categorie a esso applicabili. In questo modo, le categorie dividono gli oggetti in sottogruppi omogenei. Le variabili sono considerate omogenee quando classificano gli oggetti nelle stesse categorie negli stessi sottogruppi.

Esempio. È possibile utilizzare questa analisi per visualizzare graficamente la relazione tra la categoria lavorativa, la classificazione per minoranza e il genere. Può risultare che la classificazione per minoranza e il genere creino discriminazioni tra le persone, mentre ciò non accade per la categoria lavorativa. È inoltre possibile che le categorie Latino e Afro-americano siano simili tra loro.

Statistiche e grafici. Punteggi degli oggetti, misure di discriminazione, cronologia delle iterazioni, correlazioni delle variabili originali e trasformate, quantificazioni di categoria, statistiche descrittive, grafici a punti degli oggetti, biplot, grafici di categoria, grafici di categoria congiunti, grafici di trasformazione e grafici delle misure di discriminazione.

Dati. I valori delle variabili stringa vengono sempre convertiti in interi positivi disposti in ordine alfabetico crescente. I valori mancanti definiti dall'utente, i valori mancanti di sistema e i valori inferiori a 1 sono considerati valori mancanti. È possibile ricodificare o aggiungere una costante alle variabili con valori inferiori a 1 per fare in modo che siano considerate come non mancanti.

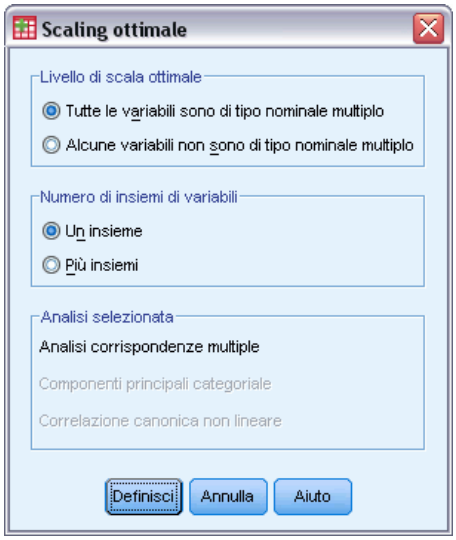
Assunzioni. Tutte le variabili hanno un livello di scaling nominale multiplo. I dati devono contenere almeno tre casi validi e l'analisi è basata su dati interi positivi. La funzione di discretizzazione classifica automaticamente una variabile con valore frazionario raggruppandone i valori in categorie con una distribuzione vicina a quella normale e converte automaticamente i valori delle variabili stringa in valori interi positivi. È possibile specificare altri schemi di discretizzazione.

Procedure correlate. Nel caso di due variabili, l'analisi delle corrispondenze multiple equivale all'analisi delle corrispondenze. Se si ritiene che le variabili abbiano proprietà ordinali o numeriche, è consigliabile utilizzare l'analisi delle componenti principali categoriale. Se si desidera considerare insiemi di variabili, è consigliabile utilizzare l'analisi della correlazione canonica (non lineare) categoriale.

Per ottenere un'analisi delle corrispondenze multiple

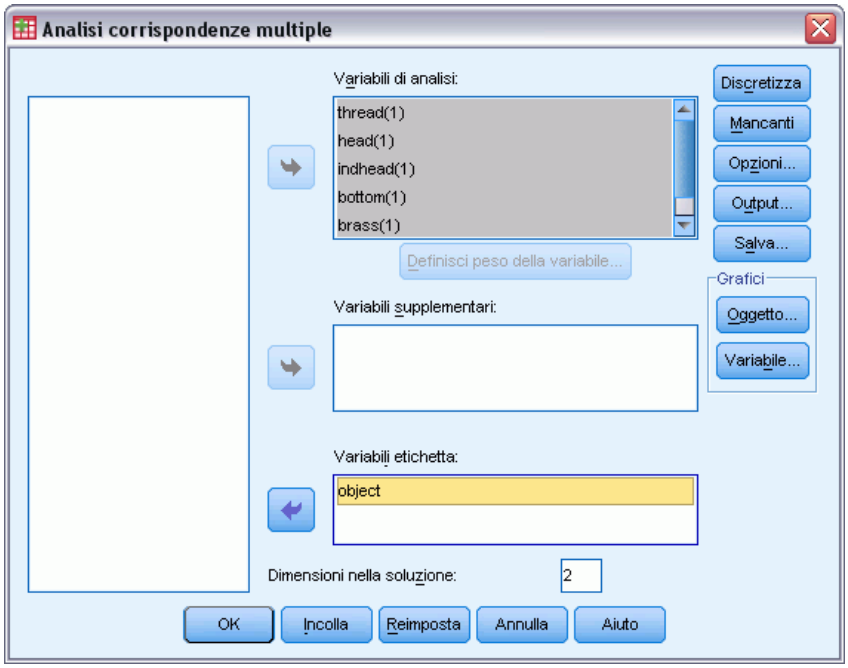
- Dai menu, scegliere:
  Analizza > Riduzioni dimensione > Scaling ottimale...

Figura 6-1
Finestra di dialogo Scaling ottimale



- Selezionare Tutte le variabili nominali multiple.
- Selezionare Un insieme.
- Fare clic su Definisci.

Figura 6-2
Finestra di dialogo Analisi delle corrispondenze multiple



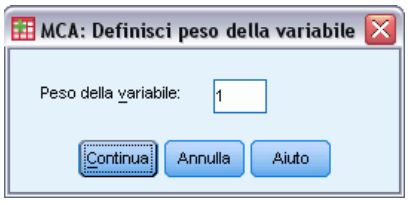
- Selezionare almeno due variabili dell'analisi e specificare il numero di dimensioni nella soluzione.
- Fare clic su OK.

Se necessario, è possibile specificare variabili supplementari che verranno inserite nella soluzione oppure variabili di etichetta per i grafici.

Definire il peso della variabile nell'analisi delle corrispondenze multiple

È possibile impostare il peso delle variabili dell'analisi.

Figura 6-3
Finestra di dialogo Definisci peso della variabile.

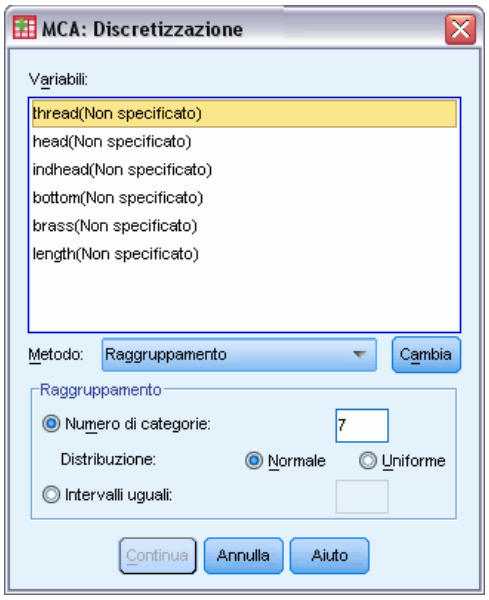


Peso della variabile Per ciascuna variabile può essere definito un peso, il cui valore deve essere un intero positivo. Il valore predefinito è 1.

Discretizzazione dell'analisi delle corrispondenze multiple

Nella finestra di dialogo Discretizzazione è possibile selezionare un metodo di ricodifica delle variabili. Le variabili con valori frazionari sono raggruppate in sette categorie (o nel numero di valori distinti della variabile se tale numero è inferiore a sette) con distribuzione approssimativamente normale, se non viene specificato diversamente. Le variabili stringa vengono sempre convertite in interi positivi tramite l'assegnazione di indicatori di categoria in base a un ordinamento alfanumerico crescente. La discretizzazione delle variabili stringa è valida per questi valori interi. Le altre variabili rimangono distinte per impostazione predefinita. Le variabili discretizzate vengono quindi utilizzate per l'analisi.

Figura 6-4
Finestra di dialogo Discretizza



Metodo. Scegliere un metodo di raggruppamento, di classificazione o di moltiplicazione.

- **Raggruppamento.** Ricodifica in un numero specificato di categorie o ricodifica per intervallo.
- **Classificazione.** La variabile viene discretizzata tramite la classificazione dei casi.
- **Moltiplicazione.** I valori correnti della variabile vengono standardizzati, moltiplicati per 10, arrotondati e viene aggiunta una costante in modo tale che il valore discretizzato minore sia uguale a 1.

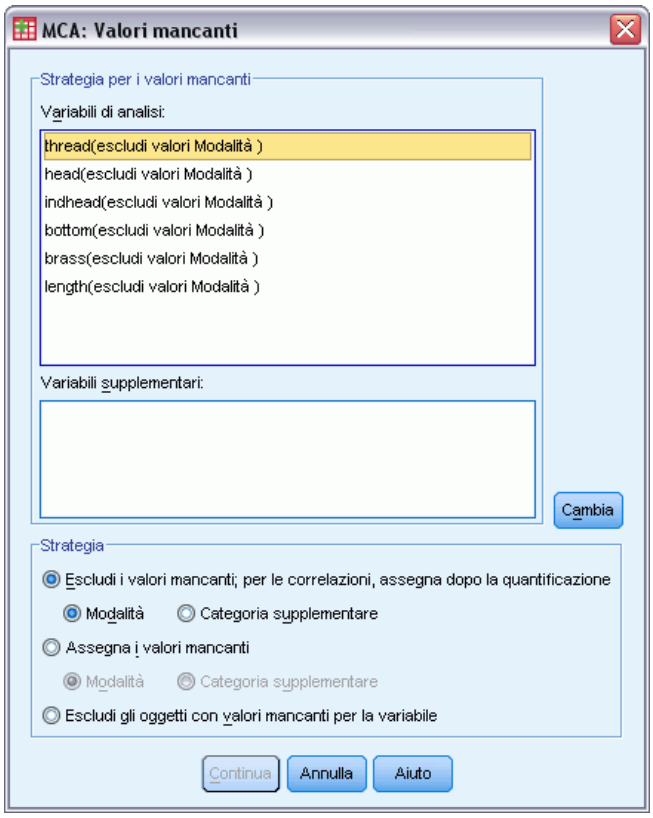
Raggruppamento. Per la discretizzazione delle variabili tramite raggruppamento sono disponibili le seguenti opzioni:

- **Numero di categorie.** Specificare un numero di categorie e se la distribuzione dei valori della variabile nelle categorie deve essere normale o uniforme.
- **Intervalli uguali.** Le variabili vengono ricodificate in categorie definite in base agli intervalli di dimensioni uguali specificati. È necessario specificare la lunghezza degli intervalli.

Valori mancanti nell'analisi delle corrispondenze multiple

Nella finestra di dialogo Valori mancanti è possibile scegliere la strategia di gestione dei valori mancanti delle variabili dell'analisi e delle variabili supplementari.

Figura 6-5
Finestra di dialogo Valori mancanti



Strategia per la gestione dei valori mancanti. Specificare se si desidera escludere i valori mancanti (trattamento passivo), assegnare i valori mancanti (trattamento attivo) o escludere gli oggetti con valori mancanti (eliminazione listwise).

- **Escludi i valori mancanti; per le correlazioni, assegna dopo la quantificazione** Gli oggetti con valori mancanti della variabile selezionata non vengono utilizzati nell'analisi di questa variabile. Se a tutte le variabili è applicato il trattamento passivo, gli oggetti con valori mancanti di tutte le variabili vengono considerati come supplementari. Se nella finestra di dialogo Output sono specificate correlazioni, dopo l'analisi ai valori mancanti viene assegnata la categoria o moda più frequente della variabile per le correlazioni delle variabili originali. Per le correlazioni delle variabili scalate in modo ottimale è possibile scegliere il metodo di assegnazione. Selezionare Moda per sostituire i valori mancanti con la moda della variabile scalata in modo ottimale. Selezionare Categoria distinta per sostituire i valori mancanti con la quantificazione di una categoria supplementare. Ciò implica che gli oggetti con un valore mancante nella variabile specificata vengono considerati come appartenenti alla stessa categoria supplementare.
- **Assegna i valori mancanti** Agli oggetti con valori mancanti della variabile selezionata vengono assegnati i valori ed è possibile scegliere il metodo di assegnazione. Selezionare Moda per sostituire i valori mancanti con la categoria più frequente. Se sono disponibili più mode, verrà utilizzata quella con l'indicatore di categoria minore. Selezionare Categoria distinta per sostituire i valori mancanti con la stessa quantificazione di una categoria supplementare. Ciò

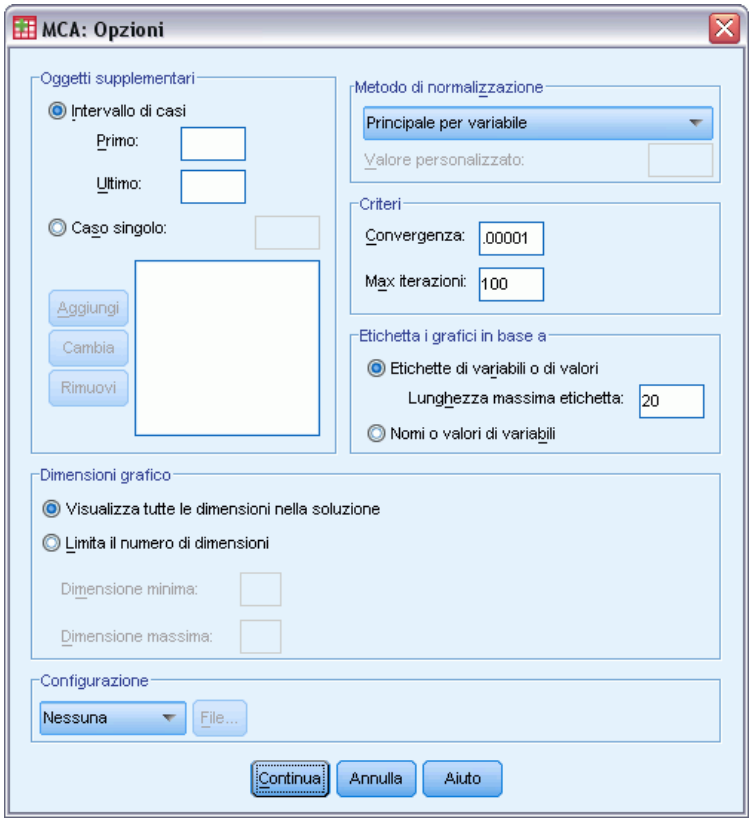
implica che gli oggetti con un valore mancante nella variabile specificata vengono considerati come appartenenti alla stessa categoria supplementare.

- **Escludi gli oggetti con valori mancanti per la variabile** Gli oggetti con valori mancanti della variabile selezionata sono esclusi dall'analisi. Questa strategia non è disponibile per le variabili supplementari.

Opzioni dell'analisi delle corrispondenze multiple

Nella finestra di dialogo Opzioni è possibile specificare la configurazione iniziale, i criteri di iterazione e di convergenza, un metodo di normalizzazione, il metodo per etichettare i grafici e gli oggetti supplementari.

Figura 6-6
Finestra di dialogo Opzioni



Oggetti supplementari. Specificare il numero di caso dell'oggetto (o il primo e l'ultimo numero di caso per un intervallo di oggetti) che si desidera contrassegnare come supplementare e quindi fare clic su Aggiungi. Ripetere l'operazione fino ad aver specificato tutti gli oggetti supplementari. I pesi di caso di un oggetto definito come supplementare verranno ignorati.

Metodo di normalizzazione. Per normalizzare i punteggi degli oggetti e le variabili, è possibile specificare una delle cinque opzioni seguenti. In un'analisi può essere utilizzato un solo metodo di normalizzazione.

- **Principale per variabile.** Consente di ottimizzare l'associazione tra variabili. Le coordinate delle variabili nello spazio dell'oggetto sono i pesi di componente, ovvero le correlazioni con le componenti principali, quali le dimensioni e i punteggi degli oggetti. Questo metodo risulta utile se la correlazione tra variabili riveste un'importanza fondamentale.
- **Principale per oggetto.** Questo metodo consente di ottimizzare le distanze tra gli oggetti e risulta utile se le dissimilarità o le similarità tra gli oggetti sono di importanza fondamentale.
- **Simmetrico.** Utilizzare questo metodo di normalizzazione se la relazione tra oggetti e variabili è di importanza fondamentale.
- **Indipendente.** Utilizzare questo metodo se si desidera esaminare separatamente le distanze tra gli oggetti e le correlazioni tra le variabili.
- **Personalizzata.** È possibile specificare qualsiasi valore reale compreso nell'intervallo [–1, 1]. Il valore 1 equivale al metodo Principale per oggetto, il valore 0 equivale al metodo Simmetrico e il valore –1 equivale al metodo Principale per variabile. Se si specifica un valore maggiore di –1 e minore di 1, è possibile disperdere l'autovalore negli oggetti e nelle variabili. Questo metodo è utile per creare biplot o triplot adatti alle specifiche esigenze.

Criteri. È possibile specificare il numero massimo di iterazioni che possono essere eseguite dalla procedura durante i calcoli e inoltre selezionare un valore per il criterio di convergenza. L'algoritmo si interrompe se la differenza dell'adattamento totale delle due ultime iterazioni è inferiore al valore di convergenza o se viene raggiunto il numero massimo di iterazioni.

Etichetta i grafici in base a. Consente di specificare se nei grafici verranno utilizzati le variabili e le etichette dei valori o i nomi delle variabili e i valori. È inoltre possibile specificare una lunghezza massima per le etichette.

Dimensioni del grafico Consente di controllare le dimensioni visualizzate nell'output.

- **Visualizza tutte le dimensioni nella soluzione.** Tutte le dimensioni nella soluzione sono visualizzate in una matrice di grafici a dispersione.
- **Limita il numero di dimensioni.** Le dimensioni visualizzate sono limitate alle coppie inserite nel grafico. Se le dimensioni vengono limitate è necessario selezionare la dimensione maggiore e minore da inserire nel grafico. La dimensione minore può variare da 1 al numero delle dimensioni nella soluzione meno 1 e viene inserita nel grafico a confronto con le dimensioni maggiori. La dimensione maggiore può variare da 2 al numero delle dimensioni nella soluzione e indica la dimensione massima da utilizzare nell'inserimento nel grafico delle coppie di dimensioni. Questa specifica si applica a tutti i grafici multidimensionali richiesti.

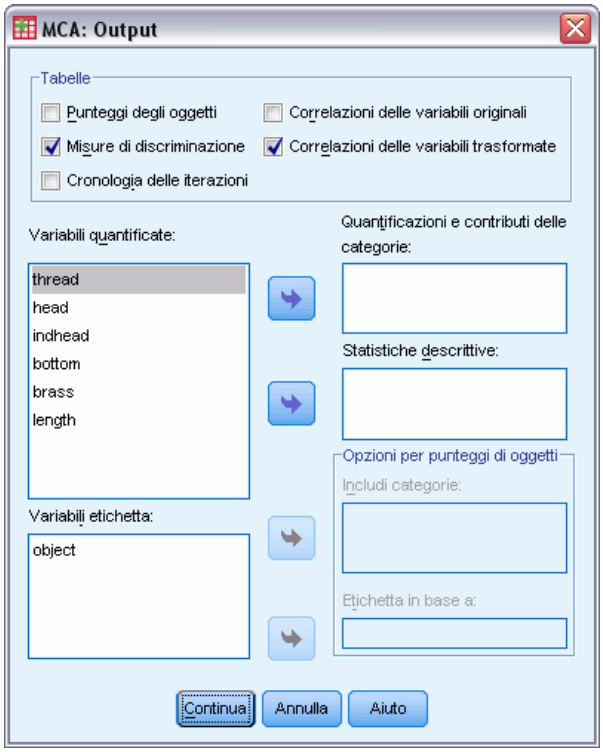
Configurazione. È possibile leggere i dati da un file che contiene le coordinate di una configurazione. La prima variabile del file deve contenere le coordinate della prima dimensione, la seconda variabile le coordinate della seconda dimensione e così via.

- **Iniziale.** La configurazione del file specificato verrà utilizzata come punto di partenza dell'analisi.
- **Fissa.** La configurazione del file specificato verrà utilizzata per inserire le variabili. Le variabili inserite devono essere selezionate come variabili dell'analisi, ma poiché la configurazione è fissa, vengono considerate come variabili supplementari e pertanto non è necessario selezionarle come variabili supplementari.

Output dell'analisi delle corrispondenze multiple

La finestra di dialogo Output consente di creare tabelle per i punteggi degli oggetti, le misure di discriminazione, la cronologia delle iterazioni, le correlazioni delle variabili originali e trasformate, le quantificazioni di categoria delle variabili selezionate e le statistiche descrittive delle variabili selezionate.

Figura 6-7
Finestra di dialogo Output



Punteggi degli oggetti. Visualizza i punteggi degli oggetti, compresi la massa, l'inerzie ed i contributi ed offre le seguenti opzioni:

- **Includi categorie.** Visualizza gli indicatori di categoria per le variabili dell'analisi selezionate.
- **Etichetta in base a.** Per etichettare gli oggetti è possibile selezionare una variabile dall'elenco di variabili etichetta.

Misure di discriminazione. Visualizza le misure di discriminazione per variabile e per dimensione.

Cronologia iterazioni. Visualizza la varianza spiegata, la perdita e l'aumento della varianza spiegata per ciascuna iterazione.

Correlazioni delle variabili originali. Visualizza la matrice di correlazione delle variabili originali e gli autovalori di tale matrice.

Correlazioni delle variabili trasformate. Visualizza la matrice di correlazione delle variabili trasformate (con scaling ottimale) e gli autovalori di tale matrice.

Quantificazioni e contributi delle categorie. Fornisce le quantificazioni di categoria e le coordinate compresi la massa, l'inerzia ed i contributi per ciascuna dimensione delle variabili selezionate

Nota: le coordinate e i contributi (comprese la massa e l'inerzia) sono visualizzati in strati diversi dell'output delle tabelle pivot, con le coordinate visualizzate per impostazione predefinita. Per visualizzare i contributi, attivare (mediante doppio clic) la tabella e selezionare Contributi dall'elenco a discesa Strato.

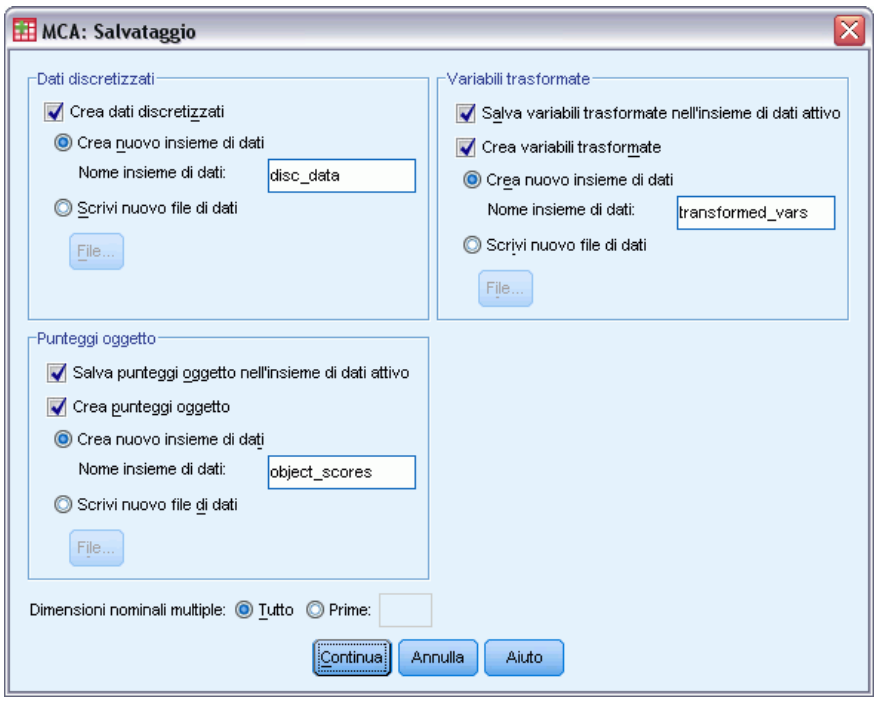
Statistiche descrittive. Visualizza le frequenze, il numero di valori mancanti e la moda delle variabili selezionate.

Analisi delle corrispondenze multiple: Salva

Dalla finestra di dialogo Salva è possibile salvare i dati discretizzati, i punteggi degli oggetti e i valori trasformati in un file di dati esterno di IBM® SPSS® Statistics o un insieme di dati nella sessione corrente. Nel file di dati attivo è inoltre possibile salvare i valori trasformati e i punteggi degli oggetti.

- I file di dati sono disponibili durante la sessione corrente, ma non lo sono in quelle successive a meno che non li si salvi esplicitamente come file di dati. I nomi degli insiemi di dati devono rispettare le regole dei nomi delle variabili.
- I nomi dei file o i nomi dei file di dati devono essere diversi per ogni tipo di dati salvati.
- Se si salvano i punteggi degli oggetti o i valori trasformati nel file di dati attivo, è possibile specificare il numero delle dimensioni nominali multiple.

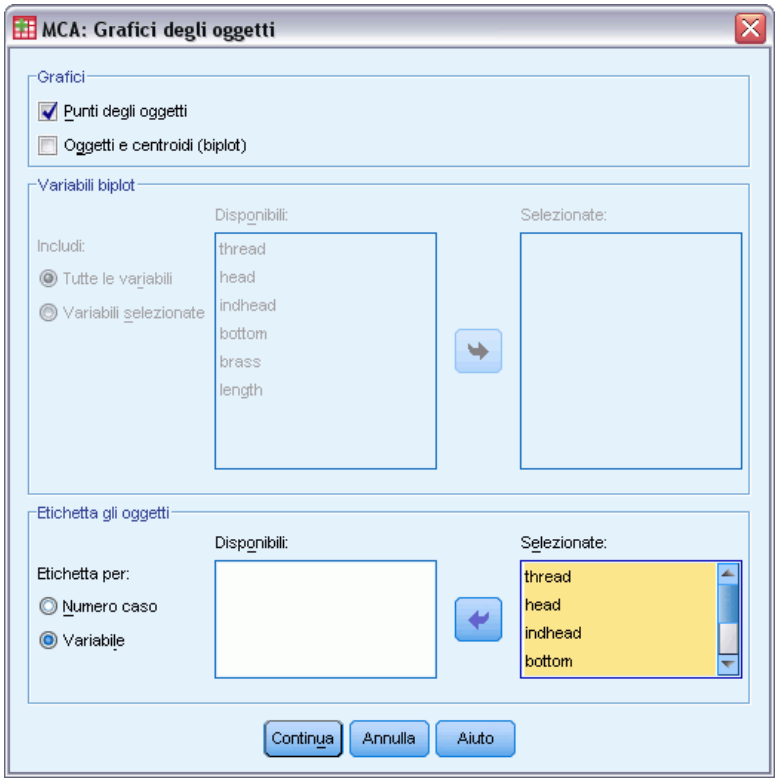
Figura 6-8
Salva



Grafici di oggetti dell'analisi delle corrispondenze multiple

Nella finestra di dialogo Grafici di oggetti è possibile specificare i tipi di grafici desiderati e le variabili da rappresentare graficamente.

Figura 6-9
Finestra di dialogo Grafici: Oggetto



Punti degli oggetti. Viene visualizzato un grafico dei punti degli oggetti.

Oggetti e centroidi (biplot). I punti degli oggetti vengono inseriti nel grafico insieme ai centroidi di variabili.

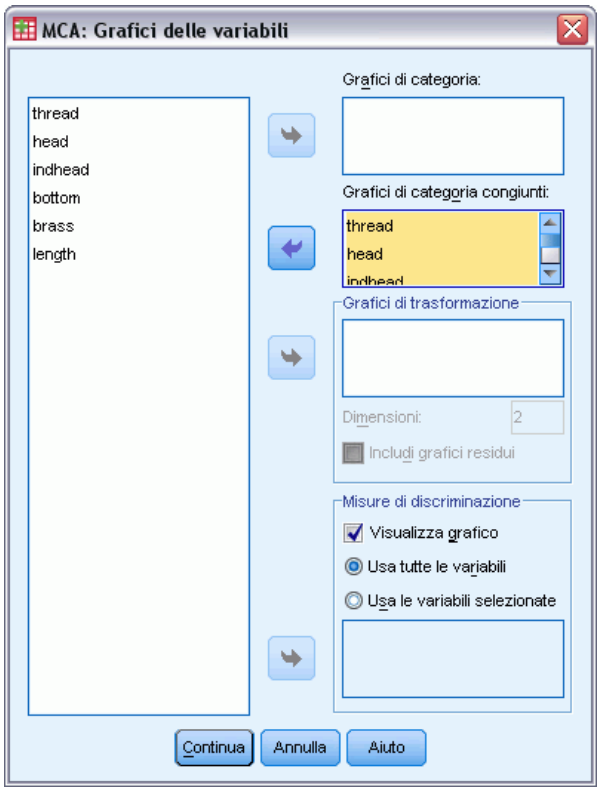
Variabili biplot. Per i biplot è possibile utilizzare tutte le variabili o selezionarne un sottoinsieme.

Etichetta gli oggetti. È possibile etichettare gli oggetti con le categorie delle variabili selezionate (scegliendo i valori degli indicatori di categoria o le etichette dei valori nella finestra di dialogo Opzioni) oppure con i relativi numeri di caso. Se è selezionata l'opzione Variabile, viene creato un grafico per ogni variabile.

Grafici di variabili dell'analisi delle corrispondenze multiple

Nella finestra di dialogo Grafici delle variabili è possibile specificare i tipi di grafici desiderati e le variabili da rappresentare graficamente.

Figura 6-10
Finestra di dialogo Grafici delle variabili



Grafici di categoria. Per ciascuna variabile selezionata viene creato un grafico delle coordinate del centroide. Le categorie saranno nei centroidi degli oggetti delle categorie specifiche.

Grafici di categoria congiunti. Si tratta di un singolo grafico delle coordinate del centroide di ciascuna variabile selezionata.

Grafici di trasformazione. Visualizza un grafico delle quantificazioni di categoria ottimali rispetto agli indicatori di categoria. È possibile specificare il numero di dimensioni desiderato. Verrà generato un grafico per ciascuna dimensione. È inoltre possibile visualizzare i grafici dei residui per ciascuna variabile selezionata.

Misure di discriminazione. Crea un singolo grafico delle misure di discriminazione per le variabili selezionate.

Opzioni aggiuntive del comando MULTIPLE CORRESPONDENCE

Per personalizzare l'analisi delle corrispondenze multiple è possibile incollare le impostazioni selezionate in una finestra di sintassi e quindi modificare la sintassi del comando `MULTIPLE CORRESPONDENCE` così ottenuta. Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare i nomi di radice per le variabili trasformate, i punteggi degli oggetti e le approssimazioni quando vengono salvati nel file dati attivo (con il sottocomando `SAVE`).

- Specificare una lunghezza massima per le singole etichette di ciascun grafico (con il sottocomando `PLOT`).
- Specificare un elenco di variabili distinto per i grafici dei residui (con il sottocomando `PLOT`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
7

Scaling multidimensionale (PROXSCAL)

La procedura Scaling multidimensionale consente di effettuare un tentativo per trovare la struttura in un insieme di misure di distanza tra oggetti. Questo processo viene effettuato assegnando le osservazioni a posizioni specifiche in uno spazio concettuale ridotto, in modo tale che le distanze tra i punti nello spazio corrispondano il più possibile alle dissimilarità specificate. In questo modo si ottiene una rappresentazione dei minimi quadrati degli oggetti all'interno dello spazio, che nella maggior parte dei casi aiuta a comprendere meglio i dati.

Esempio. La procedura Scaling multidimensionale può essere molto utile per definire le relazioni percettive. Se, ad esempio, si prende in considerazione l'immagine di un prodotto, si può svolgere un'analisi per definire un insieme di dati che descriva le similarità (o la distanza) percepibili del prodotto rispetto ai prodotti concorrenti. Tramite queste distanze e le relative variabili indipendenti, ad esempio il prezzo, si può stabilire quali variabili sono importanti in relazione al modo in cui le persone percepiscono tali prodotti ed è quindi possibile adattarne l'immagine di conseguenza.

Statistiche e grafici. Cronologia delle iterazioni, misure di stress, scomposizione di stress, coordinate dello spazio comune, distanze degli oggetti entro la configurazione finale, pesi dello spazio individuale, spazi individuali, distanze trasformate, variabili indipendenti trasformate, grafici di stress, grafici a dispersione dello spazio comune, grafici a dispersione dei pesi dello spazio individuale, grafici a dispersione dello spazio individuale, grafici di trasformazione, grafici dei residui di Shepard e grafici di trasformazione delle variabili indipendenti.

Dati. I dati possono essere forniti come matrici delle distanze o come variabili convertite in matrici delle distanze. Le matrici possono essere formattate in colonne o per colonne e per le distanze possono essere presi in considerazione i livelli di scaling di rapporto, intervallo, ordinale o spline.

Assunzioni. È necessario specificare almeno tre variabili e il numero di dimensioni non può essere maggiore del numero degli oggetti meno uno. La riduzione dimensionale non viene presa in considerazione se è combinata con inizi casuali multipli. Se viene specificata una sola origine, tutti i modelli sono equivalenti al modello di identità e pertanto l'analisi predefinita è il modello di identità.

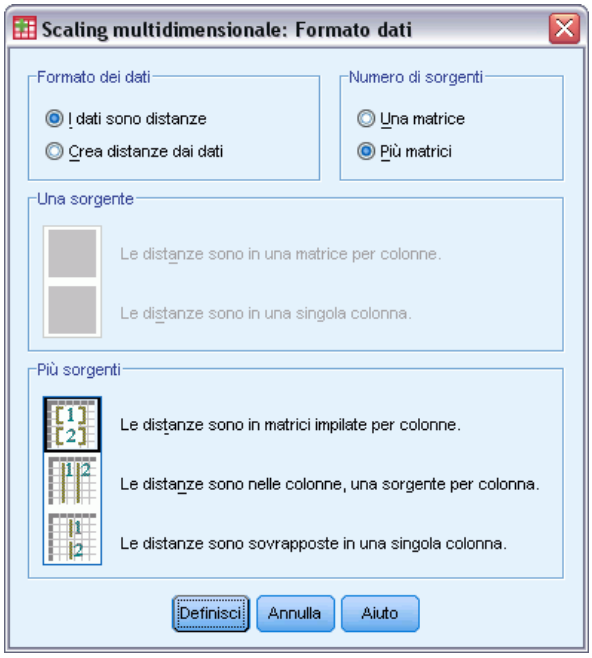
Procedure correlate. Lo scaling di tutte le variabili a livello numerico corrisponde all'analisi dello scaling multidimensionale standard.

Per ottenere un'analisi Scaling multidimensionale

▶ Dai menu, scegliere:
Analizza > Scala > Scaling multidimensionale (PROXSCAL)...

Verrà visualizzata la finestra di dialogo Formato dati.

Figura 7-1
Finestra di dialogo Formato dati



▶ Specificare il formato dei dati:

Formato dati. Specificare se i dati sono misure di distanza o se si desidera creare distanze dai dati.

Numero di sorgenti. Se i dati sono distanze, specificare se si dispone di una sorgente singola o di sorgenti multiple per le misure di similarità.

Una sorgente. Se esiste una sorgente delle distanze, specificare se il formato dell'insieme dei dati include le distanze in una matrice per colonne o in una singola colonna con due variabili distinte per l'identificazione della riga e della colonna di ciascuna distanza.

- **Le distanze sono in una matrice per colonne..** La matrice delle distanze è distribuita su un numero di colonne uguale al numero di oggetti. L'operazione porta alla finestra di dialogo Distanze in matrici per colonne.
- **Le distanze sono in una singola colonna..** La matrice delle distanze è riassunta in una sola colonna o variabile. Sono necessarie due variabili aggiuntive, che identificano riga e colonna per ciascuna cella. L'operazione porta alla finestra di dialogo Distanze in una sola colonna.

Più sorgenti. Se esistono più sorgenti delle distanze, specificare se il formato dell'insieme di dati include le distanze in matrici impilate per colonne, in più colonne con una sorgente per colonna o in una singola colonna.

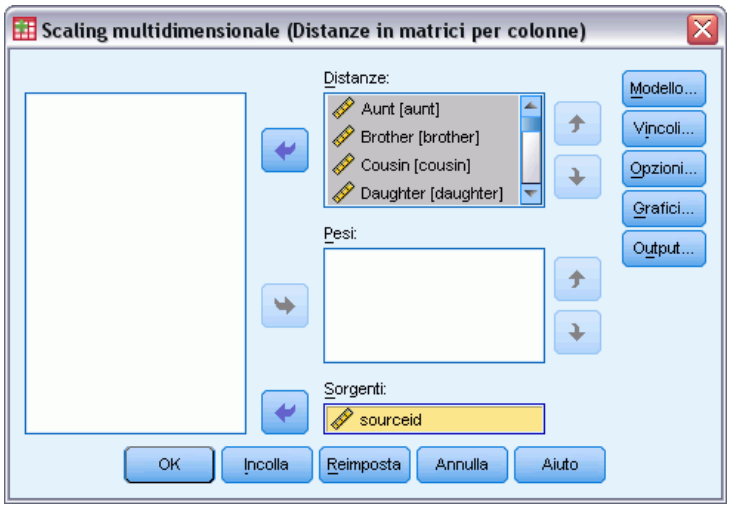
- **Le distanze sono in matrici impilate per colonne..** Le matrici di distanza sono distribuite tra un numero di colonne pari al numero di oggetti, e impilate una sopra l'altra per un numero di righe pari al numero di oggetti per il numero di sorgenti. L'operazione porta alla finestra di dialogo Distanze in matrici per colonne.
- **Le distanze sono nelle colonne, una sorgente per colonna..** Le matrici delle distanze sono riassunte in più colonne o variabili. Sono necessarie due variabili aggiuntive, che identificano riga e colonna per ciascuna cella. L'operazione porta alla finestra di dialogo Distanze in colonne.
- **Le distanze sono sovrapposte in una singola colonna..** Le matrici delle distanze sono riassunte in una sola colonna o variabile. Sono necessarie tre variabili aggiuntive, che identificano riga, colonna e sorgente per ciascuna cella. L'operazione porta alla finestra di dialogo Distanze in una sola colonna.

▶ Fare clic su Definisci.

Distanze in matrici per colonne

Se nella finestra di dialogo Formato dati si seleziona il modello di dati delle distanze nelle matrici per una sorgente o per più sorgenti, verrà visualizzata la seguente finestra principale:

Figura 7-2
Finestra di dialogo Distanze in matrici per colonne



▶ Selezionare tre o più variabili di distanza, accertandosi che l'ordine delle variabili nell'elenco corrisponda all'ordine delle colonne delle distanze.

▶ Se necessario, selezionare un numero di variabili di ponderazione uguale al numero delle variabili delle distanze. Accertarsi che l'ordine dei pesi corrisponda all'ordine delle distanze da essi rappresentate.

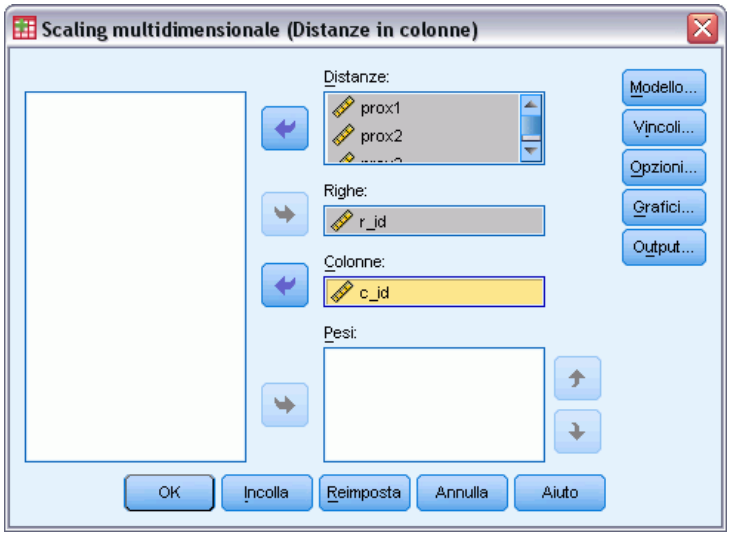
▶ In alternativa, se sono disponibili più sorgenti, selezionare una variabile di sorgenti. Il numero dei casi in ciascuna variabile di distanza deve essere uguale al numero delle variabili di distanza per il numero delle sorgenti.

È inoltre possibile definire un modello di scaling multidimensionale, assegnare vincoli sullo spazio comune, impostare criteri di convergenza, specificare la configurazione iniziale che dovrà essere utilizzata e scegliere i grafici e l'output.

Distanze in colonne

Se nella finestra di dialogo Formato dati si seleziona il modello in più colonne per più sorgenti, verrà visualizzata la seguente finestra principale:

Figura 7-3
Finestra di dialogo Distanze in colonne



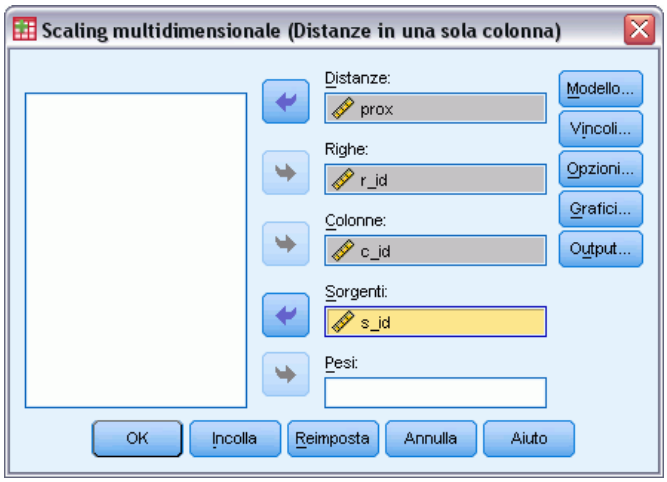
- Selezionare due o più variabili delle distanze. Si assume che ciascuna variabile sia una matrice delle distanze derivate da una sorgente distinta.
- Selezionare una variabile di riga per definire le posizioni delle righe per le distanze contenute in ciascuna variabile di distanza.
- Selezionare una variabile di colonna per definire le posizioni delle colonne per le distanze contenute in ciascuna variabile di distanza. Le celle della matrice delle distanze che non vengono designate come righe o colonne vengono considerate come mancanti.
- Se necessario, selezionare un numero di variabili di ponderazione uguale al numero delle variabili delle distanze.

  È inoltre possibile definire un modello di scaling multidimensionale, assegnare vincoli sullo spazio comune, impostare criteri di convergenza, specificare la configurazione iniziale che dovrà essere utilizzata e scegliere i grafici e l'output.

Distanze in una sola colonna

Se nella finestra di dialogo Formato dati si seleziona il modello a una colonna per una sorgente o per più sorgenti, verrà visualizzata la seguente finestra principale:

Figura 7-4
Finestra di dialogo Distanze in una sola colonna



- Selezionare una variabile delle distanze, che si assume sia costituita da una o più matrici delle distanze.
- Selezionare una variabile di riga per definire le posizioni delle righe per le distanze contenute nella variabile di distanza.
- Selezionare una variabile di colonna per definire le posizioni delle colonne per le distanze contenute nella variabile di distanza.
- Se sono disponibili più sorgenti, selezionare una variabile di sorgenti. Le celle della matrice delle distanze di ciascuna sorgente che non vengono designate come righe o colonne verranno considerate come mancanti.
- Se necessario, selezionare una variabile di ponderazione.

  È inoltre possibile definire un modello di scaling multidimensionale, assegnare vincoli sullo spazio comune, impostare criteri di convergenza, specificare la configurazione iniziale che dovrà essere utilizzata e scegliere i grafici e l'output.

Crea le distanze dai dati

Se nella finestra di dialogo Formato dati si sceglie di creare le distanze dai dati, verrà visualizzata la seguente finestra principale:

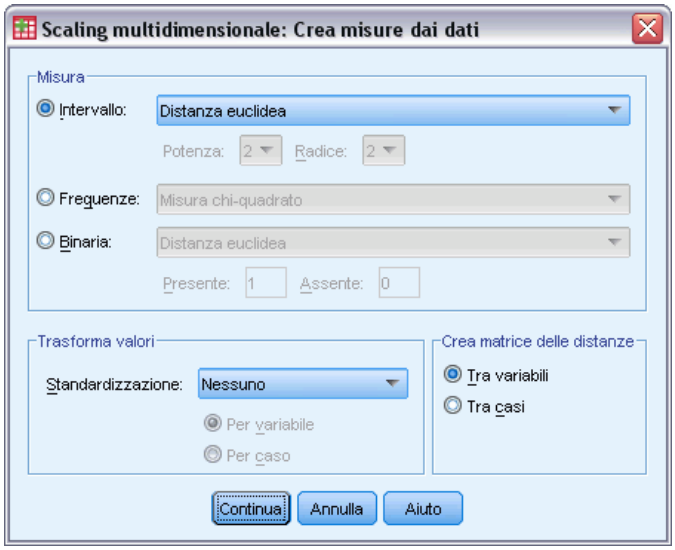
Figura 7-5
Finestra di dialogo Crea le distanze dai dati

- Se si creano distanze tra variabili (vedere la finestra di dialogo Crea misure dai dati), selezionare almeno tre variabili, che verranno utilizzate per creare la matrice delle distanze o le matrici delle distanze, se sono disponibili più sorgenti. Se si creano distanze tra casi, è sufficiente una sola variabile.
- Se sono disponibili più sorgenti, selezionare una variabile di sorgenti.
- Se necessario, scegliere una misura per la creazione delle distanze.

  È inoltre possibile definire un modello di scaling multidimensionale, assegnare vincoli sullo spazio comune, impostare criteri di convergenza, specificare la configurazione iniziale che dovrà essere utilizzata e scegliere i grafici e l'output.

Crea misure dai dati

Figura 7-6
Finestra di dialogo Crea misure dai dati



La procedura Scaling multidimensionale utilizza dati di dissimilarità per creare una soluzione di scaling. Se i dati disponibili sono dati multivariati (valori di variabili misurate), è necessario creare dati di dissimilarità in modo da calcolare una soluzione di scaling multidimensionale. È possibile specificare i dettagli della creazione delle misure di dissimilarità a partire dai dati disponibili.

Misura. Consente di specificare la misura di dissimilarità per l'analisi. Selezionare un'alternativa dal gruppo Misura corrispondente al tipo di dati desiderato e quindi selezionare una delle misure dall'elenco a discesa corrispondente a tale tipo di misura. Le alternative disponibili sono:

- **Intervallo**. Distanza euclidea, Distanza euclidea quadratica, Chebychev, City-Block, Minkowski o Personalizzato.
- **Conteggi**. Misura chi-quadrato e Misura phi-quadrato.
- **Binaria**. Distanza euclidea, Distanza euclidea quadratica, Differenza di dimensione, Differenza di modello, Varianza o Lance e Williams.

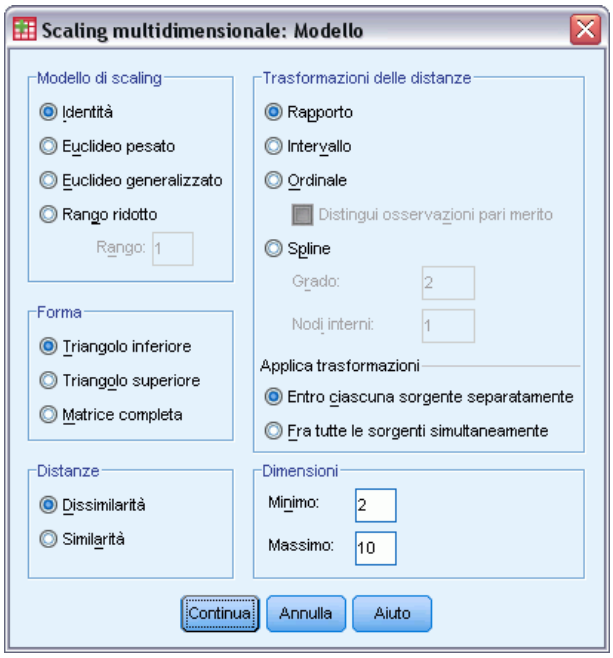
Crea matrice delle distanze. Consente di scegliere l'unità di analisi. Le alternative sono Fra variabili o Fra casi.

Trasforma valori. In alcuni casi, ad esempio quando le variabili sono misurate su scale molto diverse, è possibile standardizzarne i valori prima di calcolare le dissimilarità (non applicabile ai dati binari). Selezionare un metodo di standardizzazione dall'elenco a discesa Standardizzazione (se non è richiesta la standardizzazione, selezionare Nessuno).

Definire un modello di scaling multidimensionale

Nella finestra di dialogo Modello è possibile specificare un modello di scaling, il numero minimo e massimo delle dimensioni di tale modello, la struttura della matrice delle distanze, la trasformazione da utilizzare sulle distanze e se le distanze vengono trasformate in ciascuna sorgente separatamente o in modo non condizionale sulla sorgente.

Figura 7-7
Finestra di dialogo Modello



Modello di scaling. Scegliere una delle seguenti opzioni:

- **Identità.** Tutte le sorgenti hanno la stessa configurazione.
- **Euclideo pesato.** È un modello per differenze individuali. Ciascuna sorgente ha uno spazio individuale in cui ogni dimensione dello spazio comune viene pesata in modo differenziale.
- **Euclideo generalizzato.** È un modello per differenze individuali. Ciascuna sorgente ha uno spazio individuale uguale alla rotazione dello spazio comune, seguito da una pesatura differenziale delle dimensioni.
- **Rango ridotto.** Questo è un modello Euclideo generalizzato in cui è possibile specificare il rango dello spazio individuale. Il rango specificato deve essere maggiore o uguale a 1 e inferiore al numero massimo di dimensioni.

Forma. Specificare se le distanze devono essere prese dal triangolo inferiore o dal triangolo superiore della matrice delle distanze. È possibile specificare che deve essere utilizzata l'intera matrice e in questo caso verrà analizzata la somma ponderata del triangolo superiore e del triangolo inferiore. È tuttavia opportuno specificare la matrice completa, inclusa la diagonale, anche se verranno utilizzate solo le parti indicate.

Distanze. Specificare se la matrice di distanza contiene misure di similarità o di dissimilarità.

Trasformazioni delle distanze. Scegliere una delle seguenti opzioni:

- **Rapporto.** Le distanze trasformate sono proporzionali alle distanze originali. È consentita solo per le distanze con valore positivo.
- **Intervallo.** Le distanze trasformate sono proporzionali alle distanze originali, più il termine di un'intercetta. L'intercetta garantisce che tutte le distanze trasformate sono positive.
- **Ordinale.** L'ordine delle distanze trasformate è uguale all'ordine delle distanze originali. È possibile specificare se la distinzione delle distanze pari merito è consentita o meno.
- **Spline.** Le distanze trasformate sono una trasformazione polinomiale non decrescente livellata delle distanze originali. È possibile specificare il grado del polinomio e il numero dei nodi interni.

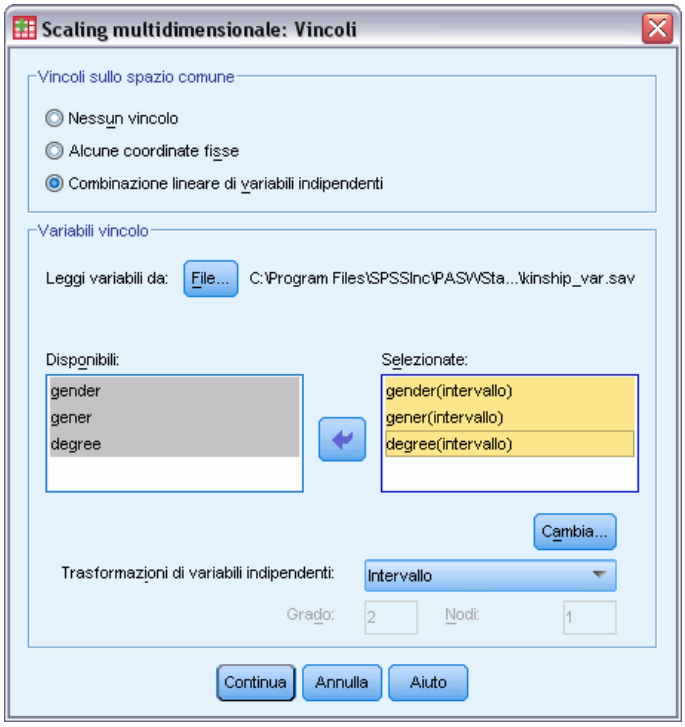
Applica trasformazioni. Specificare se il confronto avviene solo tra le distanze di ciascuna sorgente o se i confronti nella sorgente sono non condizionali.

Dimensioni. Per impostazione predefinita, una soluzione viene calcolata in due dimensioni (Minimo=2, Massimo=2). È possibile scegliere un minimo e un massimo compresi tra 1 e il numero degli oggetto meno 1, a patto che il minimo sia minore o uguale al massimo. La procedura consente di calcolare una soluzione nelle dimensioni massime e riduce quindi la dimensionalità in passaggi, fino al raggiungimento di quello inferiore.

Scaling multidimensionale: Vincoli

Nella finestra di dialogo Vincoli è possibile assegnare vincoli sullo spazio comune.

Figura 7-8
Finestra di dialogo Vincoli



Vincoli sullo spazio comune. Specificare il tipo di vincolo desiderato.

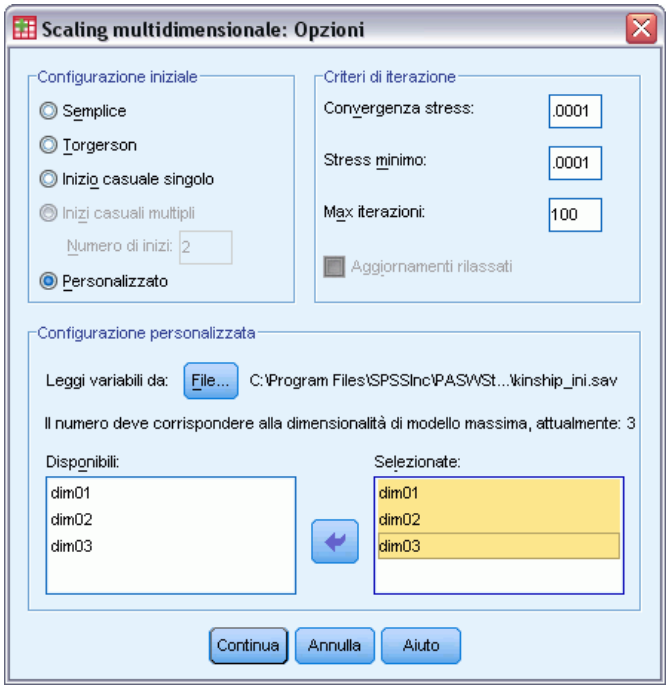
- **Nessun vincolo.** Non vengono assegnati vincoli sullo spazio comune.
- **Alcune coordinate fisse.** La prima variabile selezionata contiene le coordinate degli oggetti della prima dimensione, mentre la seconda corrisponde alle coordiante della seconda dimensione, e così via. Un valore mancante indica che una coordinata su una dimensione è libera. Il numero di variabili selezionate deve essere uguale al numero massimo di dimensioni richiesto.
- **Combinazione lineare delle variabili indipendenti.** Allo spazio comune è applicato il vincolo di essere una combinazione lineare delle variabili selezionate.

Variabili vincolo. Selezionare le variabili che definiscono i vincoli sullo spazio comune. Se è stata specificata una combinazione lineare, è possibile indicare una trasformazione di tipo intervallo, nominale, ordinale o spline per le variabili vincolo. Il numero dei casi di ciascuna variabile deve essere in ogni caso uguale al numero degli oggetti.

Scaling multidimensionale: Opzioni

Nella finestra di dialogo Opzioni è possibile selezionare lo stile di configurazione iniziale, specificare i criteri di iterazione e di convergenza e selezionare gli aggiornamenti standard o rilassati.

Figura 7-9
Finestra di dialogo Opzioni



Configurazione iniziale. Selezionare una delle alternative seguenti:

- **Semplice.** Gli oggetti sono posizionati alla stessa distanza l'uno dall'altro nella dimensione massima. Per ottenere una configurazione iniziale con il numero massimo di dimensioni specificato nella finestra di dialogo Modello, viene eseguita un'iterazione per migliorare questa configurazione di livello dimensionale elevato, seguita da una riduzione dimensionale.
- **Torgerson.** Come configurazione iniziale viene utilizzata una soluzione di scaling standard.
- **Inizio casuale singolo.** La scelta della configurazione è casuale.
- **Inizi casuali multipli.** Vengono scelte numerose configurazioni casuali e come configurazione iniziale viene utilizzata quella con il livello di Raw Stress normalizzato più basso.
- **Personalizzata.** È possibile selezionare le variabili che contengono le coordinate della configurazione iniziale specificata. Il numero delle variabili selezionate deve essere uguale al numero massimo di dimensioni specificato. La prima variabile corrisponde alle coordinate sulla dimensione 1, la seconda variabile alle coordinate sulla dimensione 2 e così via. Il numero dei casi di ciascuna variabile deve essere uguale al numero degli oggetti.

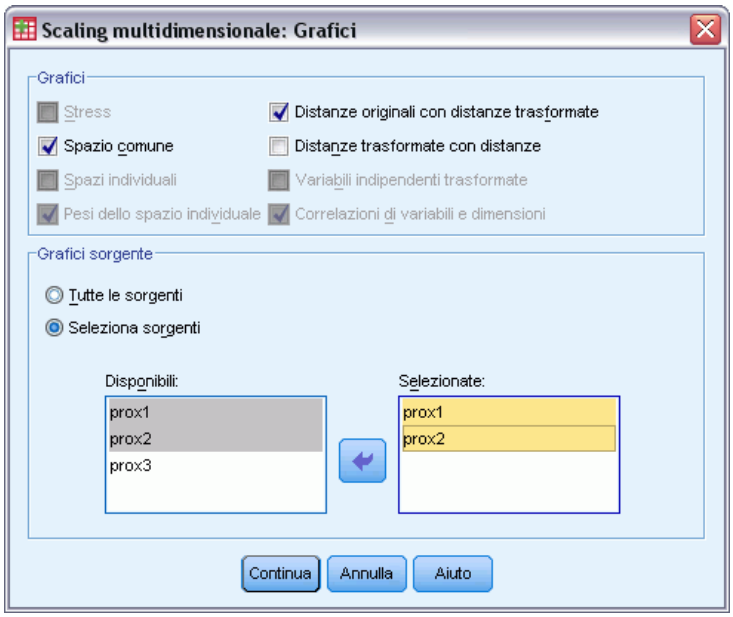
Criteri di iterazione. Specificare i valori dei criteri di iterazione.

- **Convergenza stress.** L'algoritmo di iterazione si interrompe quando la differenza tra i valori di Raw Stress normalizzato consecutivi è inferiore al numero specificato, che deve essere compreso tra 0.0 e 1.0.
- **Stress minimo.** L'algoritmo si interrompe quando il valore di Raw Stress normalizzato è inferiore al numero specificato, che deve essere compreso tra 0.0 e 1.0.
- **Max iterazioni.** L'algoritmo esegue il numero di iterazioni specificato, a meno che non sia stato soddisfatto in precedenza uno dei criteri sopra riportati.
- **Aggiornamenti rilassati.** Gli aggiornamenti rilassati consentono di velocizzare l'algoritmo. Questi aggiornamenti non possono essere utilizzati con modelli diversi dal modello di identità, né insieme a vincoli.

Scaling multidimensionale: Grafici, Versione 1

Nella finestra di dialogo Grafici è possibile specificare i grafici che si desidera creare. Se è stato impostato il formato dei dati Distanze in colonne, la finestra di dialogo Grafici includerà le opzioni riportate di seguito. Per i grafici Peso dello spazio individuale, Distanze originali con distanze trasformate e Distanze trasformate con distanze, è possibile specificare le sorgenti per cui devono essere creati i grafici. L'elenco delle sorgenti disponibili corrisponde all'elenco delle variabili delle distanze della finestra principale.

Figura 7-10
Finestra di dialogo Grafici, versione 1



Stress. Viene creato un grafico in cui sono rappresentati il valore di Raw Stress e le dimensioni. Il grafico viene creato solo se il numero massimo delle dimensioni è maggiore del numero minimo delle dimensioni.

Spazio comune. Viene visualizzata una matrice di grafici a dispersione delle coordinate dello spazio comune.

Spazi individuali. Per ciascuna sorgente, nelle matrici di grafici a dispersione vengono visualizzate le coordinate degli spazi individuali. Ciò avviene solo se nella finestra di dialogo Modello è specificato uno dei modelli delle differenze individuali.

Pesi dello spazio individuale. Viene creato un grafico a dispersione dei pesi dello spazio individuale. Ciò avviene solo se nella finestra di dialogo Modello è specificato uno dei modelli delle differenze individuali. Per il modello Euclideo pesato, nei grafici vengono rappresentati i pesi con una dimensione su ciascun asse. Per il modello Euclideo generalizzato, viene creato un grafico per dimensione, in cui sono indicate sia la rotazione che il peso della dimensione. Il modello Rango ridotto crea lo stesso grafico del modello Euclideo generalizzato, ma il numero delle dimensioni per gli spazi individuali viene ridotto.

Distanze originali con distanze trasformate. Vengono creati grafici delle distanze originali e delle distanze trasformate.

Distanze trasformate con distanze. Nel grafico vengono rappresentate le distanze trasformate e le distanze.

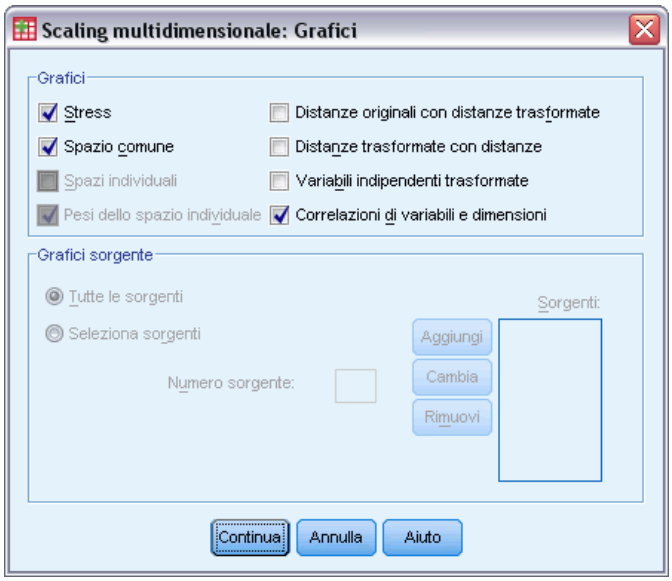
Variabili indipendenti trasformate. Vengono creati grafici di trasformazione per le variabili indipendenti.

Correlazioni di variabili e dimensioni. Viene visualizzato un grafico della correlazione tra le variabili indipendenti e le dimensioni dello spazio comune.

Scaling multidimensionale: Grafici, Versione 2

Nella finestra di dialogo Grafici è possibile specificare i grafici che si desidera creare. Se il formato dei dati specificato è diverso da Distanze in colonne, la finestra di dialogo Grafici includerà le opzioni riportate di seguito. Per i grafici Peso dello spazio individuale, Distanze originali con distanze trasformate e Distanze trasformate con distanze, è possibile specificare le sorgenti per cui devono essere creati i grafici. I numeri delle sorgenti specificati devono essere valori della variabile sorgenti specificata nella finestra principale e devono essere compresi nell'intervallo da 1 al numero delle sorgenti.

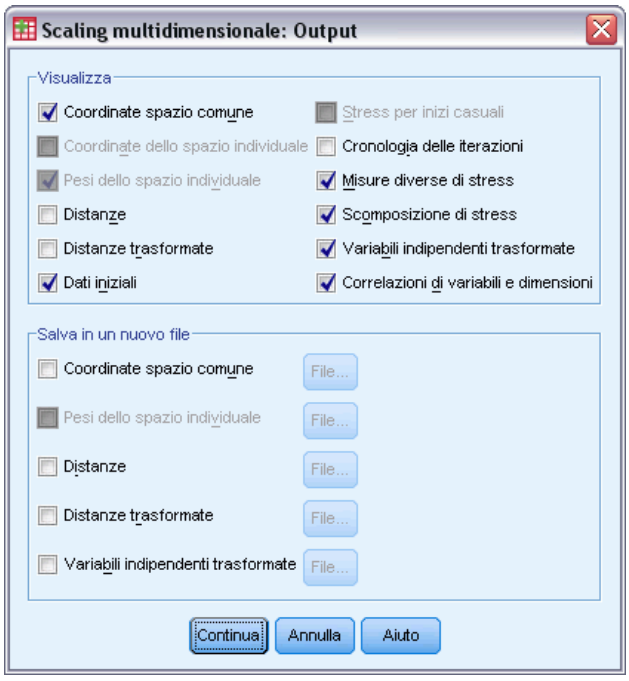
Figura 7-11
Finestra di dialogo Grafici, versione 2



Scaling multidimensionale: Output

Nella finestra di dialogo Output è possibile controllare la quantità di output visualizzata e salvarne una parte in file distinti.

Figura 7-12
Finestra di dialogo Output



Visualizzazione. Selezionare una o più delle seguenti opzioni per la visualizzazione:

- **Coordinate dello spazio comune.** Visualizza le coordinate dello spazio comune.
- **Coordinate dello spazio individuale.** Visualizza le coordinate degli spazi individuali, solo nel caso in cui il modello sia diverso da Identità.
- **Pesi dello spazio individuale.** Visualizza i pesi dello spazio individuale, solo nel caso in cui venga specificato uno dei modelli delle differenze individuali. A seconda del modello, i pesi dello spazio vengono scomposti in pesi di rotazione e pesi di dimensione, anch'essi visualizzati. nel grafico
- **Distanze.** Visualizza le distanze tra gli oggetti nella configurazione specificata.
- **Distanze trasformate.** Visualizza le distanze trasformate tra gli oggetti della configurazione.
- **Dati iniziali.** Include le distanze originali e, se presenti, i pesi dei dati, la configurazione iniziale e le coordinate fisse o le variabili indipendenti.
- **Stress per inizi casuali.** Visualizza il seme del numero casuale e il valore di Raw Stress normalizzato per ciascun inizio casuale.
- **Cronologia iterazioni.** Visualizza la cronologia delle iterazioni dell'algoritmo principale.
- **Misure diverse di stress.** Visualizza valori di stress diversi. La tabella include i valori di Raw Stress normalizzato, Stress-I, Stress-II, S-Stress, della dispersione spiegata (DAF) e del coefficiente di congruenza di Tucker.
- **Scomposizione di stress.** Visualizza una scomposizione degli oggetti e delle sorgenti del valore finale Raw Stress normalizzato, inclusa la media per oggetto e per sorgente.

- **Variabili indipendenti trasformate.** Se è stato selezionato un vincolo di combinazione lineare, vengono visualizzati le variabili indipendenti trasformate e i pesi della regressione corrispondenti.
- **Correlazioni di variabili e dimensioni.** Se è stato selezionato un vincolo di combinazione lineare, vengono visualizzate le correlazioni tra le variabili indipendenti e le dimensioni dello spazio comune.

Salva in un nuovo file. Le coordinate dello spazio comune, i pesi dello spazio individuale, le distanze, le distanze trasformate e le variabili indipendenti trasformate possono essere salvate in file di dati di IBM® SPSS® Statistics distinti.

Opzioni aggiuntive del comando PROXSCAL

Per personalizzare lo scaling multidimensionale dell'analisi delle distanze, è possibile incollare le impostazioni selezionate in una finestra di sintassi e modificare la sintassi del comando `PROXSCAL` risultante. Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare un elenco di variabili distinto per le trasformazioni e per i grafici dei residui (con il sottocomando `PLOT`).
- Specificare elenchi sorgente distinti per i pesi dello spazio individuale, le trasformazioni e i grafici dei residui (con il sottocomando `PLOT`).
- Specificare un sottoinsieme dei grafici di trasformazione delle variabili indipendenti che si desidera visualizzare (con il sottocomando `PLOT`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
8

Unfolding multidimensionale (PREFSCAL)

La procedura Unfolding multidimensionale tenta di individuare una scala quantitativa comune che consenta di analizzare visivamente le relazioni tra due insiemi di oggetti.

Esempi. È stato chiesto a 21 persone di disporre 15 alimenti da colazione in ordine di preferenza da 1 a 15. Grazie all'unfolding multidimensionale, è possibile stabilire che tali persone distinguono tra gli alimenti da colazione seguendo due criteri principali: pane fresco e pane tostato e alimenti ingrassanti e dietetici.

In alternativa, è stato chiesto a un gruppo di guidatori di valutare 10 caratteristiche di 26 modelli di auto adottando una scala da 6 punti, da 1= "assolutamente in disaccordo" a 6= "assolutamente d'accordo". Calcolando la media delle persone, tali valori vengono considerati come similarità. Utilizzando Unfolding multidimensionale, è possibile individuare raggruppamenti di modelli analoghi e gli attributi a cui vengono più strettamente associati.

Statistiche e grafici. La procedura Unfolding multidimensionale è in grado di generare cronologia delle iterazioni, misure di stress, scomposizione di stress, coordinate dello spazio comune, distanze degli oggetti entro la configurazione finale, pesi dello spazio individuale, spazi individuali, distanze trasformate, grafici di stress, grafici a dispersione dello spazio comune, grafici a dispersione dei pesi dello spazio individuale, grafici a dispersione degli spazi individuali, grafici di trasformazione e grafici dei residui di Shepard.

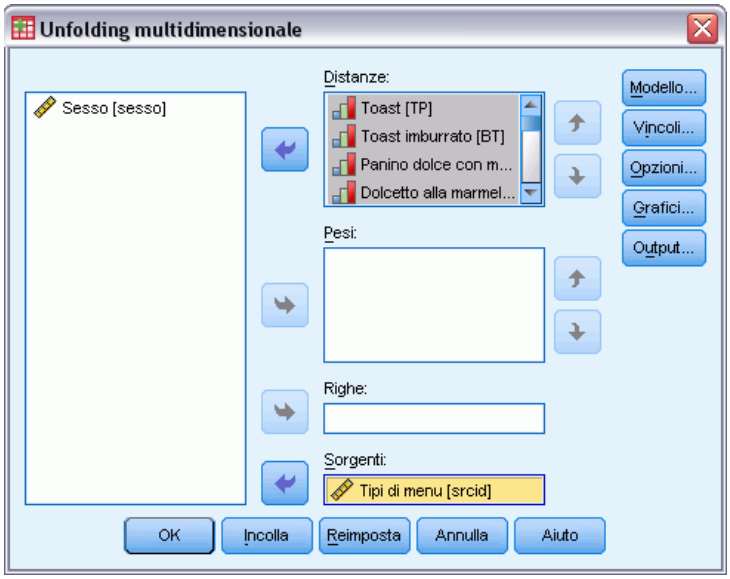
Dati. I dati vengono forniti sotto forma di matrici di distanza rettangolari. Ogni colonna viene considerata come un oggetto colonna separato. Ogni riga di una matrice di distanza viene considerata come un oggetto riga separato. Quando sono presenti più sorgenti delle distanze, le matrici vengono sovrapposte.

Assunzioni. È necessario specificare almeno due variabili. Il numero di dimensioni presenti nella soluzione non può essere maggiore del numero degli oggetti meno uno. Se viene specificata una sola origine, tutti i modelli sono equivalenti al modello di identità e pertanto l'analisi predefinita è il modello di identità.

Per ottenere un'analisi Unfolding multidimensionale

▶ Dai menu, scegliere:
Analizza > Scala > Unfolding multidimensionale (PREFSCAL)...

Figura 8-1
Finestra di dialogo principale Unfolding multidimensionale



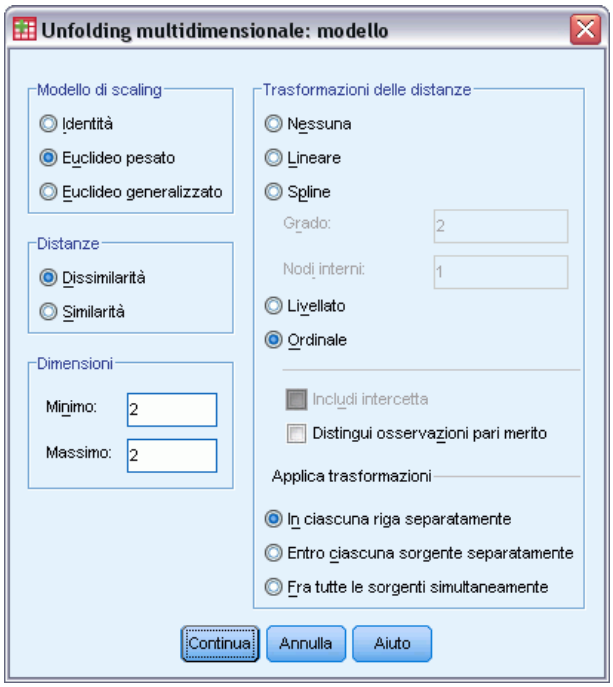
- Selezionare due o più variabili che identifichino le colonne nella matrice di distanza rettangolare. Ogni variabile rappresenta un oggetto colonna separato.
- Se necessario, selezionare un numero di variabili di ponderazione uguale al numero delle variabili dell'oggetto colonna. L'ordine delle variabili di ponderazione deve corrispondere all'ordine degli oggetti colonna da esse rappresentati.
- Se necessario, selezionare una variabile di riga. I valori (o le etichette dei valori) di questa variabile vengono utilizzati per etichettare gli oggetti riga nell'output.
- Se esistono più sorgenti, è possibile selezionare una variabile sorgente. Il numero dei casi nel file di dati deve essere uguale al numero degli oggetti riga per il numero delle sorgenti.

  È inoltre possibile definire un modello di unfolding multidimensionale, assegnare vincoli sullo spazio comune, impostare criteri di convergenza, specificare la configurazione iniziale che dovrà essere utilizzata e scegliere i grafici e l'output.

Definizione di un modello di unfolding multidimensionale

Nella finestra di dialogo Modello è possibile specificare un modello di scaling, il numero minimo e massimo delle dimensioni di tale modello, la struttura della matrice delle distanze, la trasformazione da utilizzare sulle distanze e se le distanze vengono trasformate in modo condizionale sulla riga e in modo condizionale o non condizionale sulla sorgente.

Figura 8-2
Finestra di dialogo Modello



Modello di scaling. Scegliere una delle seguenti opzioni:

- **Identità.** Tutte le sorgenti hanno la stessa configurazione.
- **Euclideo pesato.** È un modello per differenze individuali. Ciascuna sorgente ha uno spazio individuale in cui ogni dimensione dello spazio comune viene pesata in modo differenziale.
- **Euclideo generalizzato.** È un modello per differenze individuali. Ciascuna sorgente ha uno spazio individuale uguale alla rotazione dello spazio comune, seguito da una pesatura differenziale delle dimensioni.

Distanze. Specificare se la matrice di distanza contiene misure di similarità o di dissimilarità.

Dimensioni. Per impostazione predefinita, una soluzione viene calcolata in due dimensioni (Minimo=2, Massimo=2). È possibile scegliere un minimo e un massimo compresi tra 1 e il numero degli oggetto meno 1, a patto che il minimo sia minore o uguale al massimo. La procedura consente di calcolare una soluzione nelle dimensioni massime e di ridurre quindi la dimensionalità in passaggi, fino al raggiungimento di quello inferiore.

Trasformazioni delle distanze. Scegliere una delle seguenti opzioni:

- **Nessuna.** Le distanze non vengono trasformate. Se lo si desidera, è possibile selezionare Includi intercetta. In questo caso, le distanze possono essere spostate in base a un termine costante.
- **Lineare.** Le distanze trasformate sono proporzionali alle distanze originali, ovvero la funzione di trasformazione esegue la stima di un'inclinazione e l'intercetta viene fissata a 0. Questa operazione viene anche definita trasformazione del rapporto. Se lo si desidera, è possibile selezionare Includi intercetta. In questo caso, le distanze possono inoltre essere spostate in base a un termine costante. Questa operazione viene anche definita trasformazione di intervallo.

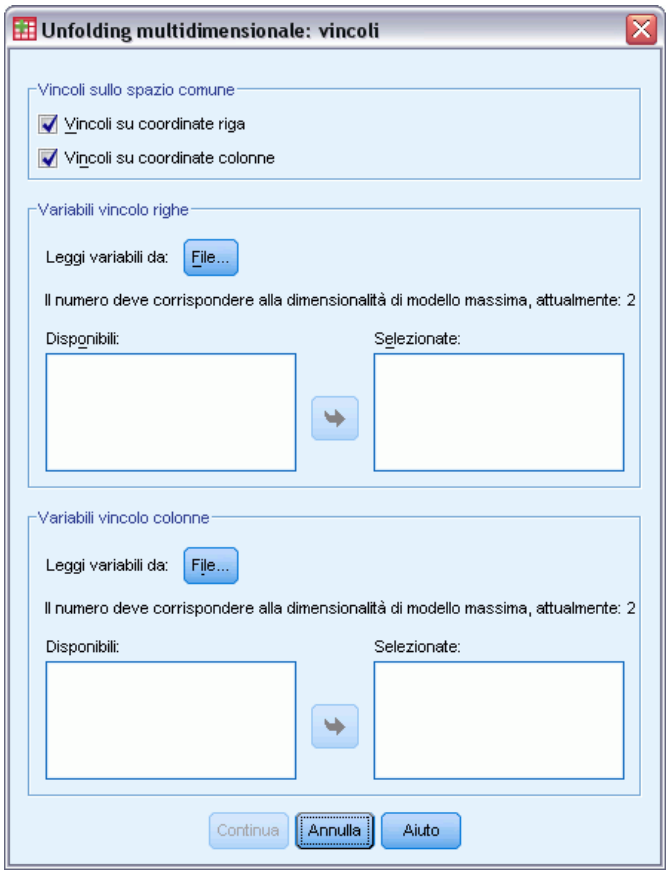
- **Spline.** Le distanze trasformate sono una trasformazione polinomiale non decrescente livellata delle distanze originali. È possibile specificare il grado del polinomio e il numero dei nodi interni. Se lo si desidera, è possibile selezionare Includi intercetta. In questo caso, le distanze possono inoltre essere spostate in base a un termine costante.
- **Livellamento.** Le distanze trasformate presentano lo stesso ordine delle distanze originali, incluso un vincolo che tiene in considerazione le differenze tra i valori successivi. Il risultato è una trasformazione di "livellamento ordinale". È possibile specificare se la distinzione delle distanze pari merito è consentita o meno.
- **Ordinale.** L'ordine delle distanze trasformate è uguale all'ordine delle distanze originali. È possibile specificare se la distinzione delle distanze pari merito è consentita o meno.

Applica trasformazioni. Specificare se confrontare l'una con l'altra solo le distanze presenti nelle singole righe o solo le distanze presenti in ogni sorgente o se i confronti sono non condizionali sulla riga o sulla sorgente, ovvero se le trasformazioni devono essere eseguite per riga, per sorgente o su tutte le distanze contemporaneamente.

Vincoli relativi all'unfolding multidimensionale

Nella finestra di dialogo Vincoli è possibile assegnare vincoli sullo spazio comune.

Figura 8-3
Finestra di dialogo Vincoli



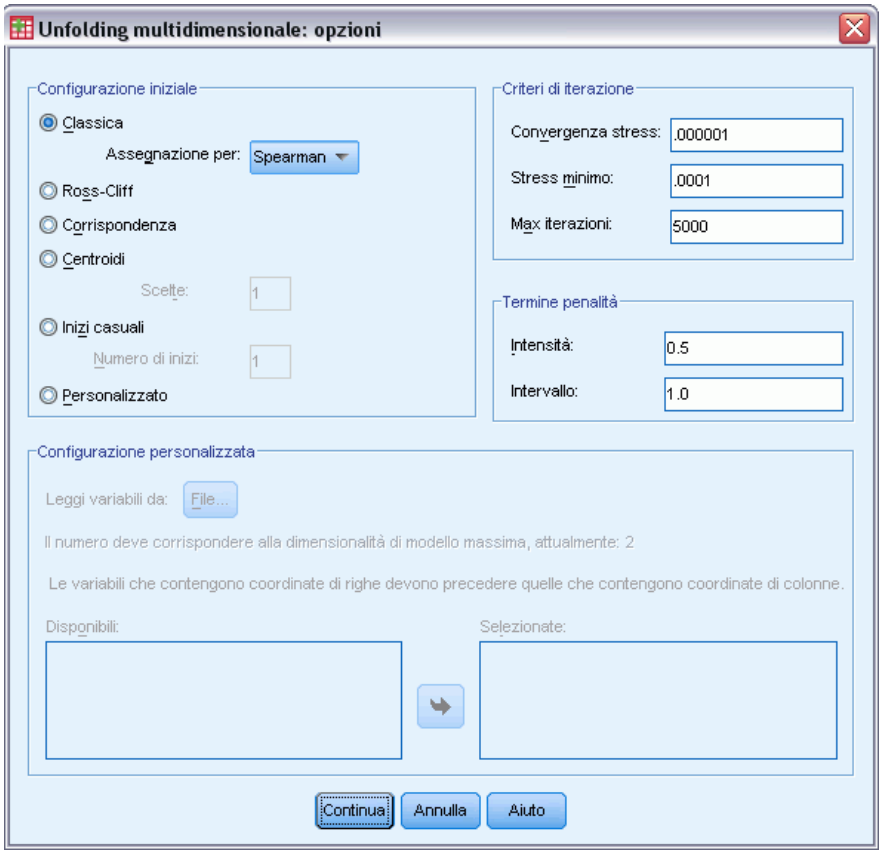
Vincoli sullo spazio comune. È possibile scegliere di definire le coordinate degli oggetti riga e/o colonna nello spazio comune.

Variabili vincolo di riga/colonna. Scegliere il file che contiene i vincoli e selezionare le variabili che definiscono i vincoli nello spazio comune. La prima variabile selezionata contiene le coordinate degli oggetti della prima dimensione, mentre la seconda corrisponde alle coordinate della seconda dimensione e così via. Un valore mancante indica che una coordinata su una dimensione è libera. Il numero di variabili selezionate deve essere uguale al numero massimo di dimensioni richiesto. Il numero dei casi di ciascuna variabile deve essere uguale al numero degli oggetti.

Opzioni di unfolding multidimensionale

Nella finestra di dialogo Opzioni è possibile selezionare lo stile di configurazione iniziale, specificare i criteri di iterazione e di convergenza e impostare il termine di penalità per stress.

Figura 8-4
Finestra di dialogo Opzioni



Configurazione iniziale. Selezionare una delle alternative seguenti:

- **Classica.** La matrice di distanza rettangolare viene utilizzata come supplemento per i valori tra i blocchi (ovvero i valori che si trovano tra le righe e tra le colonne) della matrice di scaling multidimensionale simmetrica completa. Dopo aver formato la matrice completa, come configurazione iniziale viene utilizzata una soluzione di scaling classica. I valori tra i blocchi

possono essere riempiti mediante assegnazione utilizzando l'ineguaglianza del triangolo o le distanze di Spearman.

- **Ross-Cliff.** Come valori iniziali per gli oggetti riga e colonna, l'inizio Ross-Cliff utilizza i risultati di una scomposizione di valori singoli sulla matrice di distanza quadrata e con doppia centratura.
- **Corrispondenza.** L'inizio corrispondenza utilizza i risultati di un'analisi di corrispondenza sui dati invertiti (similarità anziché dissimilarità) con normalizzazione simmetrica dei punteggi di riga e di colonna.
- **Centroidi.** La procedura inizia con il posizionamento degli oggetti riga nella configurazione utilizzando la scomposizione di un autovalore. A questo punto, gli oggetti colonna vengono posizionati in corrispondenza del centroide delle scelte specificate. Per il numero di scelte, specificare un numero intero positivo compreso tra 1 e il numero delle variabili di distanza.
- **Inizi casuali multipli.** Vengono calcolate soluzioni per numerose configurazioni iniziali scelte in modo casuale e quella che presenta lo stress penalizzato minore viene indicata come soluzione ottimale.
- **Personalizzata.** È possibile selezionare le variabili che contengono le coordinate della configurazione iniziale specificata. Il numero delle variabili selezionate deve essere uguale al numero massimo di dimensioni specificato. La prima variabile corrisponde alle coordinate sulla dimensione 1, la seconda variabile alle coordinate sulla dimensione 2 e così via. Il numero dei casi di ciascuna variabile deve essere uguale al numero combinato degli oggetti riga e colonna. Le coordinate di riga e colonna devono essere sovrapposte, con le coordinate della colonna che seguono le coordinate della riga.

Criteri di iterazione. Specificare i valori dei criteri di iterazione.

- **Convergenza stress.** L'algoritmo di iterazione si interrompe quando la differenza relativa tra valori di stress penalizzato consecutivi è inferiore al numero specificato, che deve essere non negativo.
- **Stress minimo.** L'algoritmo si interrompe quando lo stress penalizzato scende al di sotto del numero specificato, che deve essere non negativo.
- **Max iterazioni.** L'algoritmo esegue il numero di iterazioni specificato, a meno che non sia stato soddisfatto in precedenza uno dei criteri sopra riportati.

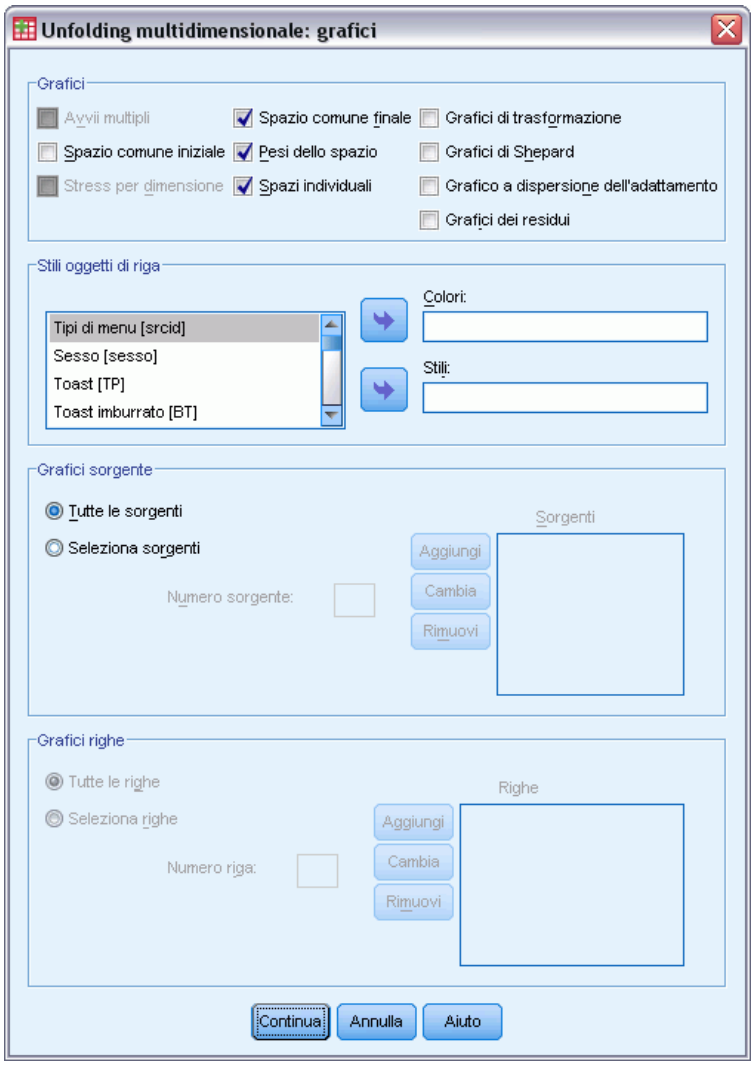
Termine di penalità. L'algoritmo tenta di ridurre al minimo lo stress penalizzato, una misura della bontà di adattamento equivalente al prodotto di Stress-I di Kruskal per un termine di penalità basato sul coefficiente di variazione delle distanze trasformate. Questi controlli consentono di impostare l'intensità e l'intervallo del termine di penalità.

- **Intensità.** Il valore del parametro dell'intensità è inversamente proporzionale alla penalità. Specificare un valore compreso tra 0.0 e 1.0.
- **Intervallo.** Questo parametro definisce il momento in cui la penalità diventa attiva. Se l'impostazione è 0.0, la penalità non è attiva. Se si aumenta il valore, l'algoritmo cerca una soluzione con una variazione maggiore tra le distanze trasformate. Specificare un valore non negativo.

Grafici di unfolding multidimensionale

Nella finestra di dialogo Grafici è possibile specificare i grafici che si desidera creare.

Figura 8-5
Finestra di dialogo Grafici



Grafici. Sono disponibili i seguenti grafici:

- **Inizi multipli.** Consente di visualizzare un istogramma sovrapposto di stress penalizzato, riportando sia lo stress che la penalità.
- **Spazio comune iniziale.** Consente di visualizzare una matrice di grafici a dispersione delle coordinate dello spazio comune iniziale.
- **Stress per dimensione.** Produce un grafico lineare di stress penalizzato rispetto alle dimensioni. Il grafico viene creato solo se il numero massimo delle dimensioni è maggiore del numero minimo delle dimensioni.

- **Spazio comune finale.** Viene visualizzata una matrice di grafici a dispersione delle coordinate dello spazio comune.
- **Pesi dello spazio.** Viene creato un grafico a dispersione dei pesi dello spazio individuale. Ciò avviene solo se nella finestra di dialogo Modello è specificato uno dei modelli delle differenze individuali. Per il modello Euclideo pesato, nei grafici vengono rappresentati i pesi relativi a tutte le sorgenti, con una dimensione su ciascun asse. Per il modello Euclideo generalizzato, viene creato un grafico per dimensione, in cui sono indicate sia la rotazione che il peso di tale dimensione per ogni sorgente.
- **Spazi individuali.** Viene visualizzata una matrice di grafici a dispersione delle coordinate dello spazio individuale di ogni sorgente. Ciò avviene solo se nella finestra di dialogo Modello è specificato uno dei modelli delle differenze individuali.
- **Grafici di trasformazione.** Vengono creati grafici a dispersione delle distanze originali rispetto alle distanze trasformate. A seconda della modalità di applicazione delle trasformazioni, a ogni riga o sorgente viene assegnato un diverso colore. Una trasformazione non condizionale produce un singolo colore.
- **Grafici Shepard.** Consente di confrontare le distanze originali con le distanze e le distanze trasformate. Le distanze sono indicate da punti, mentre le distanze trasformate sono indicate da una linea. A seconda della modalità di applicazione delle trasformazioni, per ogni riga o sorgente viene generata una linea distinta. Una trasformazione non condizionale produce una linea.
- **Grafico a dispersione dell'adattamento.** Viene visualizzato un grafico a dispersione che confronta le distanze e le distanze trasformate. Se si specificano più sorgenti, a ognuna di esse viene assegnato un diverso colore.
- **Grafici dei residui.** Viene visualizzato un grafico a dispersione che confronta le distanze trasformate e i residui (distanze trasformate meno le distanze). Se si specificano più sorgenti, a ognuna di esse viene assegnato un diverso colore.

Stili degli oggetti riga. Consente un maggiore controllo della visualizzazione degli oggetti riga nei grafici. I valori delle variabili di colore facoltative consentono di utilizzare tutti i colori. I valori delle variabili di simbolo facoltative consentono di utilizzare tutti i simboli possibili.

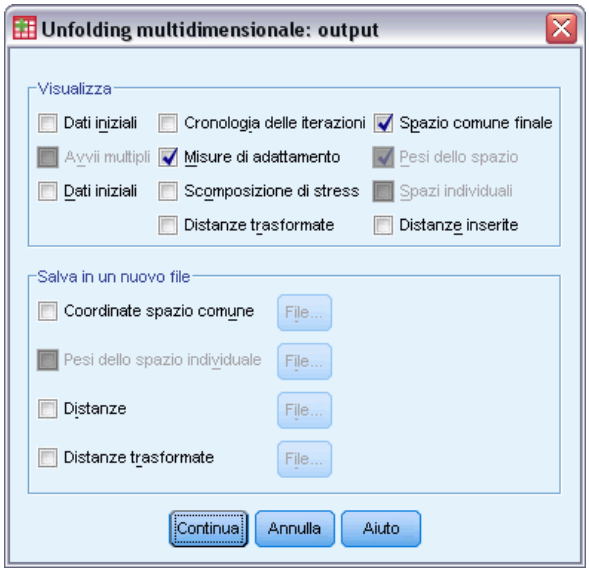
Grafici sorgente. Per i grafici Spazi individuali, Dispersione dell'adattamento e Residui e se le trasformazioni vengono applicate in base alla sorgente, per i grafici Trasformazione e Shepard è possibile specificare le sorgenti a cui tali grafici devono fare riferimento. I numeri delle sorgenti specificati devono essere valori della variabile sorgente specificata nella finestra di dialogo principale e devono essere compresi nell'intervallo da 1 al numero delle sorgenti.

Grafici righe. Se le trasformazioni vengono applicate per riga, per i grafici Trasformazione e Shepard è possibile specificare la riga a cui tali grafici devono fare riferimento. I numeri di riga immessi devono essere compresi tra 1 e il numero di righe.

Output dell'unfolding multidimensionale

Nella finestra di dialogo Output è possibile controllare la quantità di output visualizzata e salvarne una parte in file distinti.

Figura 8-6
Finestra di dialogo Output



Visualizzazione. Selezionare una o più delle seguenti opzioni per la visualizzazione:

- **Dati iniziali.** Include le distanze originali e, se presenti, i pesi dei dati, la configurazione iniziale e le coordinate fisse.
- **Inizi multipli.** Visualizza il seme del numero casuale e il valore di stress penalizzato per ciascun inizio casuale.
- **Dati iniziali.** Visualizza le coordinate dello spazio comune iniziale.
- **Cronologia iterazioni.** Visualizza la cronologia delle iterazioni dell'algoritmo principale.
- **Misure di adattamento.** Visualizza diverse misure. La tabella contiene varie misure di bontà dell'adattamento, inadeguatezza dell'adattamento, correlazione, variazione e non degenerazione.
- **Scomposizione di stress.** Visualizza una scomposizione di oggetti, righe e sorgenti di stress penalizzato, tra cui riga, colonna e medie e deviazioni standard della sorgente.
- **Distanze trasformate.** Visualizza le distanze trasformate.
- **Spazio comune finale.** Visualizza le coordinate dello spazio comune.
- **Pesi dello spazio.** Visualizza i pesi dello spazio individuale. Questa opzione è disponibile solo se si specifica uno dei modelli delle differenze individuali. A seconda del modello, i pesi dello spazio vengono scomposti in pesi di rotazione e pesi di dimensione, anch'essi visualizzati. nel grafico
- **Spazi individuali.** Vengono visualizzate le coordinate degli spazi individuali. Questa opzione è disponibile solo se si specifica uno dei modelli delle differenze individuali.
- **Distanze inserite.** Visualizza le distanze tra gli oggetti nella configurazione specificata.

Salva in un nuovo file. Le coordinate dello spazio comune, i pesi dello spazio individuale, le distanze e le distanze trasformate possono essere salvate in file di dati di IBM® SPSS® Statistics distinti.

Funzioni aggiuntive del comando PREFSCAL

Per personalizzare l'unfolding multidimensionale dell'analisi delle distanze, è possibile incollare le impostazioni selezionate in una finestra di sintassi e modificare la sintassi del comando PREFSCAL risultante. Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare diversi elenchi sorgente per i grafici Spazi individuali, Dispersione dell'adattamento e Residui—e in caso di trasformazioni condizionali della matrice, per i grafici Trasformazione e Shepard—quando sono disponibili più sorgenti (ricorrendo al sottocomando PLOT).
- Specificare diversi elenchi riga per i grafici Trasformazione e Shepard in caso di trasformazioni condizionali della riga (ricorrendo al sottocomando PLOT).
- Specificare un numero di righe anziché una variabile dell'ID di riga (ricorrendo al sottocomando INPUT).
- Specificare un numero di sorgenti anziché una variabile dell'ID di sorgente (ricorrendo al sottocomando INPUT).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Parte II: Esempi

Capitolo
9

Regressione categoriale

L'obiettivo della regressione categoriale con scaling ottimale è descrivere la relazione tra una variabile di risposta e un insieme di predittori. Quantificando tale relazione, è possibile stimare i valori della risposta per qualsiasi combinazione di predittori.

Nel presente capitolo, verranno utilizzati due esempi per illustrare le analisi relative alla regressione con scaling ottimale. Il primo esempio utilizza un limitato insieme di dati per illustrare i concetti di base. Il secondo esempio utilizza un insieme di variabili e di osservazioni più ampio in un esempio pratico.

Esempio: Dati relativi a un battitappeto

Come esempio tipico (Green e Wind, 1973), un'azienda interessata alla commercializzazione di un nuovo battitappeto desidera esaminare l'influenza di cinque fattori sulle preferenze del consumatore, ovvero design della confezione, marca, prezzo, la presenza di un *marchio di qualità* e una garanzia "Soddisfatti o rimborsati". Esistono tre livelli di fattore per il design della confezione, che differiscono per la posizione della spazzola dell'applicatore; tre marchi (*K2R*, *Glory* e *Bissell*); tre livelli di prezzo e due livelli (no o sì) per ciascuno degli ultimi due fattori. La tabella seguente mostra le variabili utilizzate nello studio relativo al battitappeto, con relative etichette e valori.

Tabella 9-1
Variabili esplicative nello studio relativo al battitappeto

Nome di variabile	Etichetta di valore	Etichetta del valore
confezione	Design confezione	A*, B*, C*
marca	Nome marca	K2R, Glory, Bissell
prezzo	Prezzo	$1.19, $1.39, $1.59
marchio di qualità	Presenza di un *marchio di qualità*	No, sì
garanzia	Garanzia "Soddisfatti o rimborsati"	No, sì

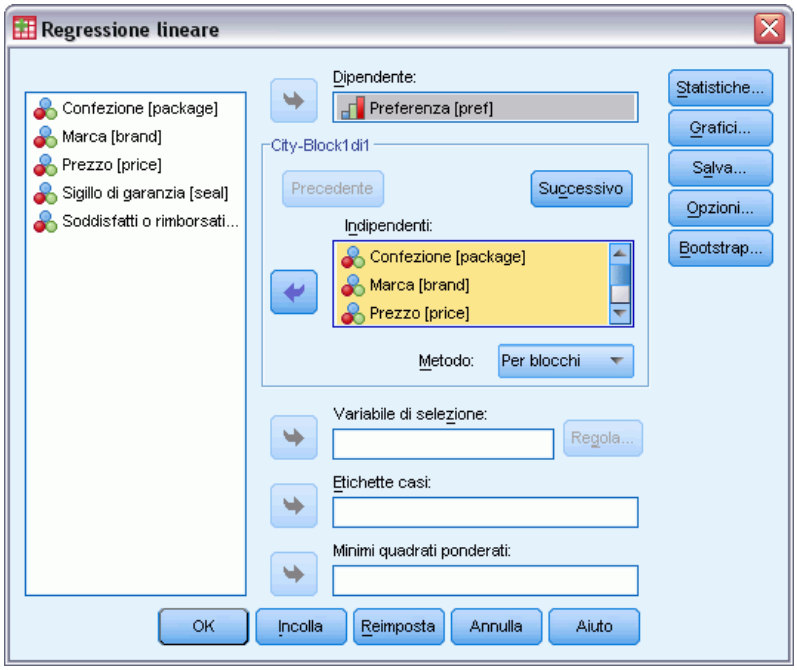
Dieci consumatori sono classificati in 22 profili definiti da questi fattori. La variabile *Preferenza* include il rango delle classificazioni medie per ogni profilo. Classificazioni basse corrispondono a una preferenza elevata. La variabile riflette una misura globale della preferenza per ogni profilo. Utilizzando la regressione categoriale si esaminerà la correlazione tra i cinque fattori e la preferenza. Questo insieme di dati è reperibile in *carpet.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*.

Analisi della regressione lineare standard

- Per generare l'output della regressione lineare standard, dai menu scegliere:
  Analizza > Regression > Lineare...

  Nota: Questa funzione richiede il modulo Statistics Base.

Figura 9-1
Finestra di dialogo Regressione lineare

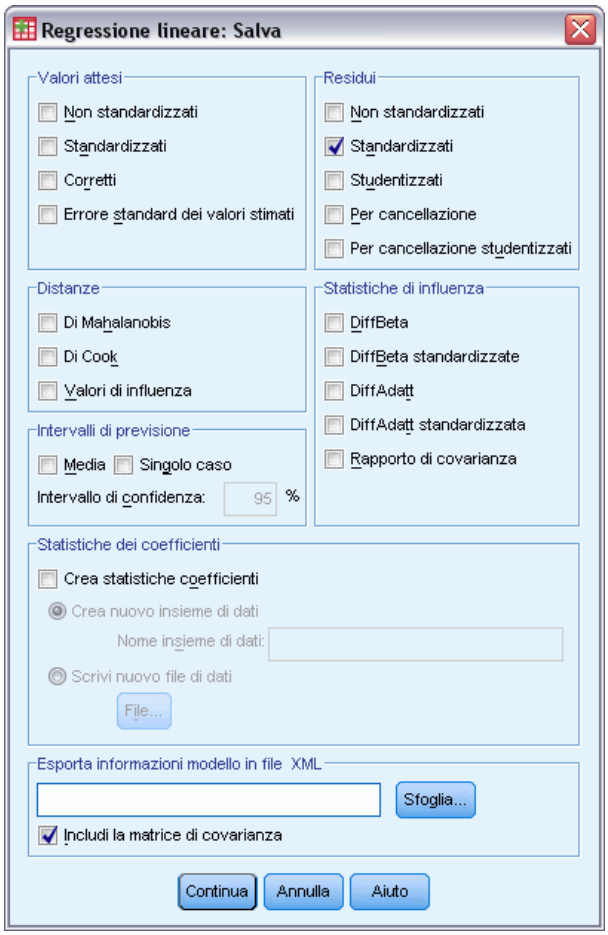


- Selezionare *Preferenza* come variabile dipendente.
- Selezionare da *Design confezione* a *Garanzia "Soddisfatti o rimborsati"* come variabili indipendenti.
- Fare clic su Grafici.

Figura 9-2
Finestra di dialogo Grafici

- Selezionare **ZRESID* come variabile dell'asse *y*.
- Selezionare **ZPRED* come variabile dell'asse *x*.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione lineare fare clic su Salva.

Figura 9-3
Salva



- Selezionare Standardizzati nel gruppo Residui.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione lineare scegliere OK.

Riepilogo del modello

Figura 9-4
Riepilogo del modello per regressione lineare standard

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	,841[a]	,707	,615	3,99810

a. Stimatori: (Costante), Residuo Standardizzato, Soddisfatti o rimborsati, Sigillo di garanzia, Marca, Prezzo, Confezione

L'approccio standard per descrivere le relazioni nel problema corrente è la regressione lineare. La misura più comune del grado di adattamento del modello di regressione ai dati è R^2. La statistica rappresenta la quantità di varianza nella risposta spiegata dalla combinazione ponderata dei

predittori. Maggiore sarà l'approssimazione di R^2 a 1, maggiore sarà l'adattamento del modello. Eseguendo la regressione di *Preferenza* sui cinque predittori si ottiene un R^2 pari a 0,707, a indicare che circa il 71% della varianza delle classificazioni di preferenza è spiegata dalle variabili indipendenti nella regressione lineare.

Coefficienti

I coefficienti standardizzati sono indicati nella tabella. Il segno del coefficiente indica se la risposta prevista aumenta o diminuisce quando il predittore aumenta, restando costanti tutti gli altri. Per i dati categoriali, la codificazione della categoria determina il significato dell'aumento di un predittore. Ad esempio, un aumento di *Garanzia "Soddisfatti o rimborsati"*, *Design confezione* o *Marchio di qualità* determinerà una riduzione della classificazione della preferenza prevista. La codifica di *Garanzia "Soddisfatti o rimborsati"* è 1 per *nessuna garanzia* e 2 per *garanzia presente*. Un aumento di *Garanzia "Soddisfatti o rimborsati"* corrisponde all'aggiunta di una garanzia. Di conseguenza, l'aggiunta di una garanzia "Soddisfatti o rimborsati" riduce la classificazione della preferenza prevista, che corrisponde a una preferenza prevista maggiore.

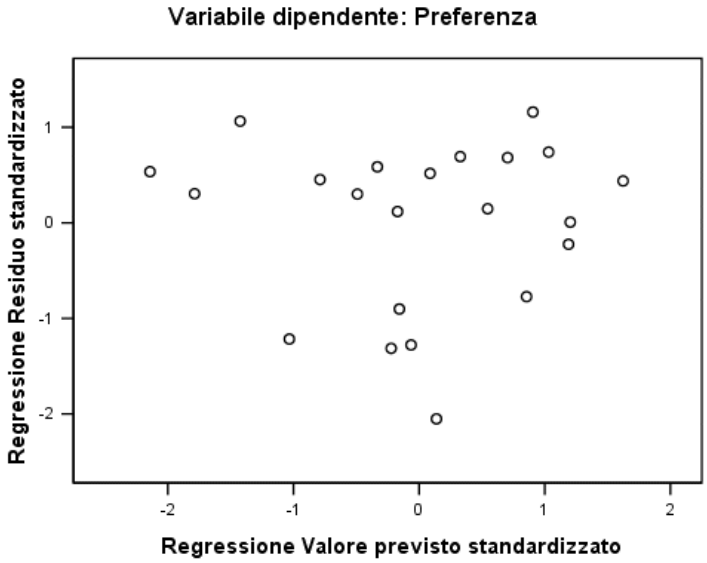
Figura 9-5
Coefficienti di regressione

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	22,529	5,177		4,352	,000
	Confezione	-4,159	1,036	-,560	-4,015	,001
	Marca	,429	1,054	,056	,407	,689
	Prezzo	2,703	1,009	,366	2,681	,016
	Sigillo di garanzia	-4,314	1,780	-,330	-2,423	,028
	Soddisfatti o rimborsati	-2,779	1,921	-,197	-1,447	,167

Il valore del coefficiente riflette la variazione nella classificazione della preferenza prevista. Utilizzando i coefficienti standardizzati, le interpretazioni si basano sulle deviazioni standard delle variabili. Ogni coefficiente indica il numero delle deviazioni standard di modifica della risposta prevista per una modifica della deviazione a uno standard in un predittore, restando costanti tutti gli altri. Ad esempio, una modifica della deviazione standard in *Nome marca* genera un aumento nella preferenza prevista pari a 0,056 deviazioni standard. La deviazione standard di *Preferenza* è 6,44, quindi *Preferenza* è aumentata di $0,056 \times 6,44 = 0,361$. Le modifiche di *Design confezione*generano le maggiori variazioni nella preferenza prevista.

Grafici a dispersione dei residui

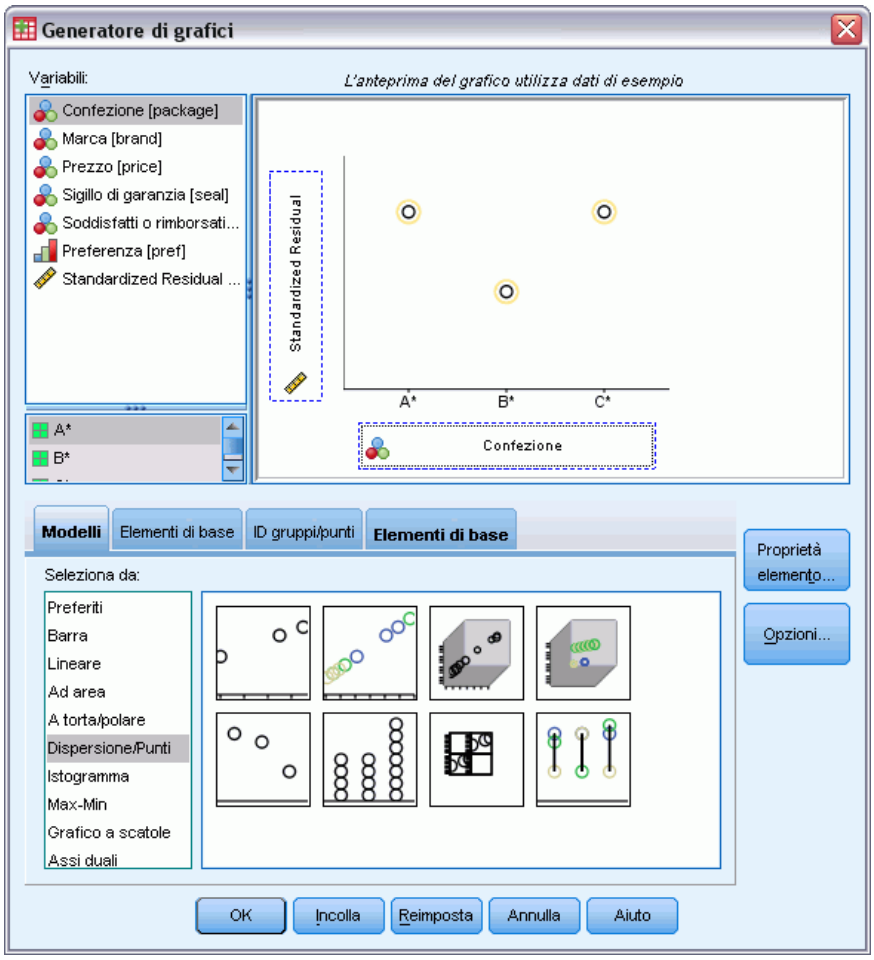
Figura 9-6
Residui e valori attesi



I residui standardizzati sono tracciati rispetto ai valori attesi standardizzati. Se l'adattamento del modello è buono, non dovrebbero essere presenti modelli. È visibile una forma a U in cui i valori attesi standardizzati inferiore e superiore hanno entrambi residui positivi. I valori attesi standardizzati vicini allo 0 tendono ad avere residui negativi.

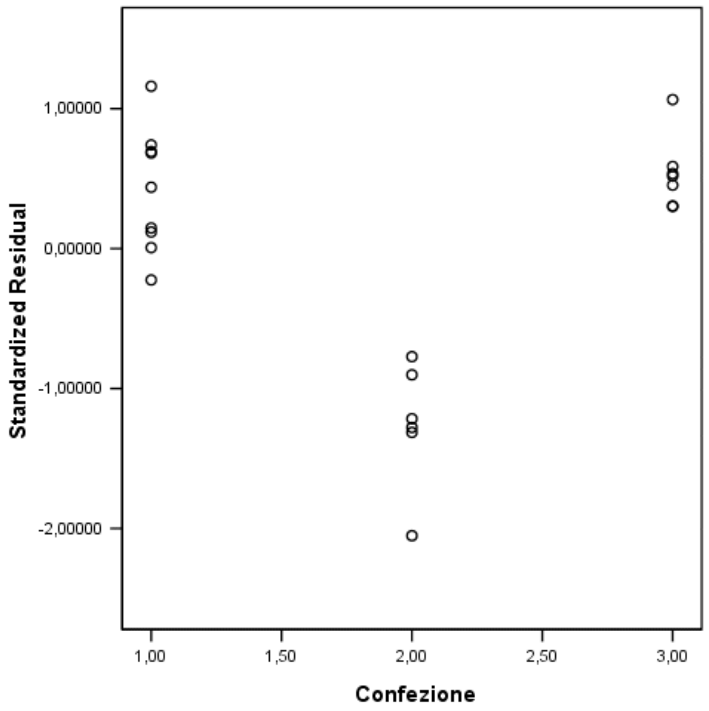
- Per generare un grafico a dispersione dei residui dal predittore *Design confezione*, dai menu scegliere:
  Grafici > Generatore di grafici...

Figura 9-7
Generatore di grafici



- Selezionare il modello Dispersione/Punti e scegliere Dispersione semplice.
- Selezionare *Residuo standardizzato* come variabile dell'asse *y* e *Design confezione* come variabile dell'asse *x*.
- Fare clic su OK.

Figura 9-8
Residui vs design confezione



La forma a U è più pronunciata nel grafico dei residui standardizzati rispetto alla confezione. Ogni residuo per Design B* è negativo, mentre tutti i residui eccetto uno sono positivi per gli altri due design. Poiché il modello di regressione lineare si adatta a un parametro per ogni variabile, il rapporto non può essere rilevato dall'approccio standard.

Analisi di regressione categoriale

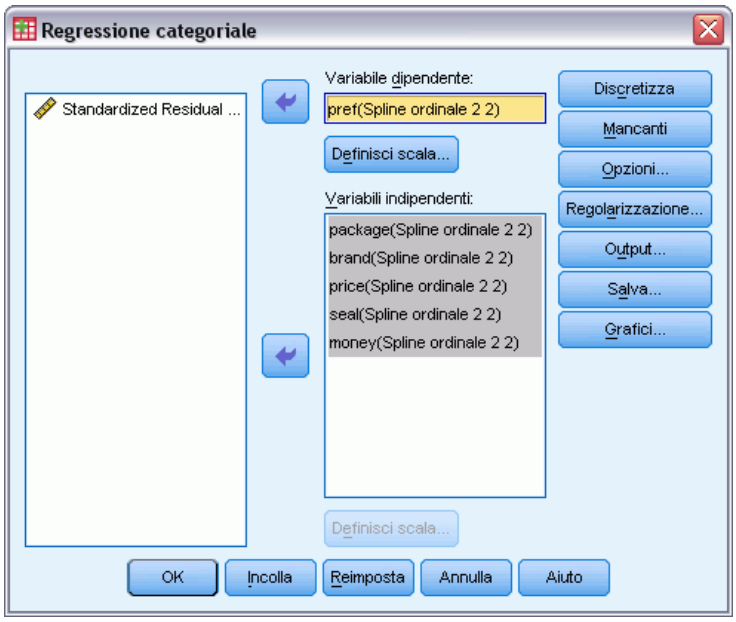
La natura categoriale delle variabili e il rapporto non lineare tra *Preferenza* e *Design confezione* suggerisce che la regressione su punteggi ottimali possa offrire prestazioni migliori di quella standard. La forma a U dei grafici dei residui indica che è consigliato l'utilizzo di un trattamento nominale di *Design confezione*. Tutti gli altri predittori saranno trattati a livello di scaling numerico.

La variabile di risposta merita una considerazione speciale. Poiché si desidera prevedere i valori di *Preferenze*. è consigliabile recuperare il maggior numero possibile di proprietà delle relative categorie nelle quantificazioni. Utilizzando un livello di scaling nominale o ordinale le differenze tra le categorie di risposta vengono ignorate. Tuttavia, la trasformazione lineare delle categorie di risposta conserva le differenze tra le categorie. Di conseguenza, lo scaling numerico delle risposte è generalmente preferibile e sarà utilizzato in questo caso.

Esecuzione dell'analisi

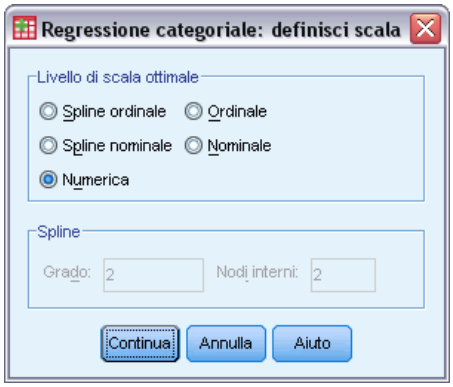
- Per eseguire un'analisi di regressione categoriale, dai menu scegliere:
  Analizza > Regression > Scaling ottimale (CATREG)...

Figura 9-9
Finestra di dialogo Regressione categoriale



- Selezionare *Preferenza* come variabile dipendente.
- Selezionare da *Design confezione* a *Garanzia "Soddisfatti o rimborsati"* come variabili indipendenti.
- Selezionare *Preferenza* e fare clic su Definisci scala.

Figura 9-10
Finestra di dialogo Definisci scala



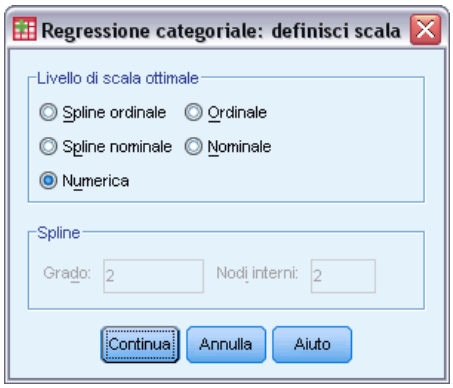
- Selezionare Numerico come livello di scaling ottimale.
- Fare clic su Continua.

- Nella finestra di dialogo Regressione categoriale selezionare *Design confezione* e fare clic su Definisci scala.

Figura 9-11
Finestra di dialogo Definisci scala

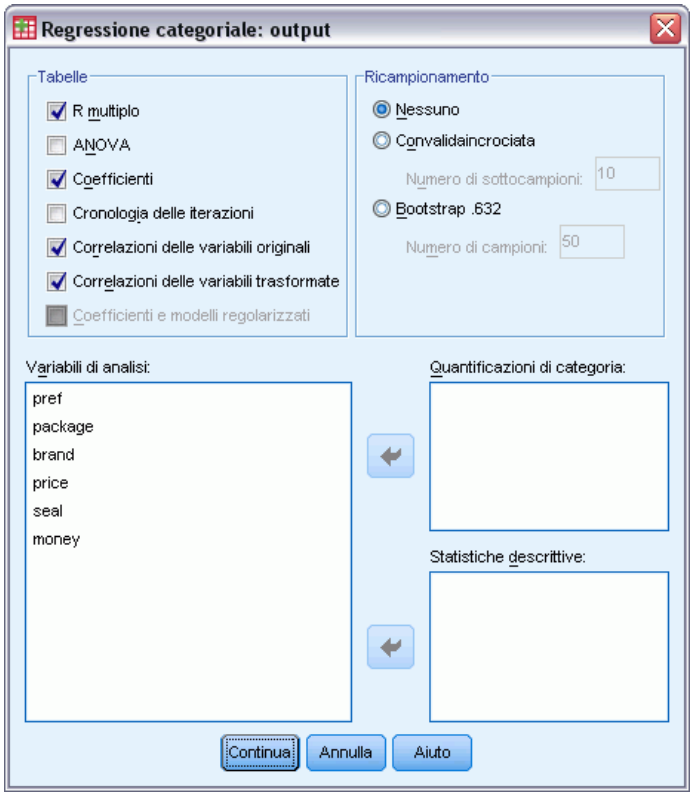
- Selezionare Nominale come livello di scaling ottimale.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale selezionare da *Nome marca* a *Garanzia "Soddisfatti o rimborsati"* e fare clic su Definisci scala.

Figura 9-12
Finestra di dialogo Definisci scala



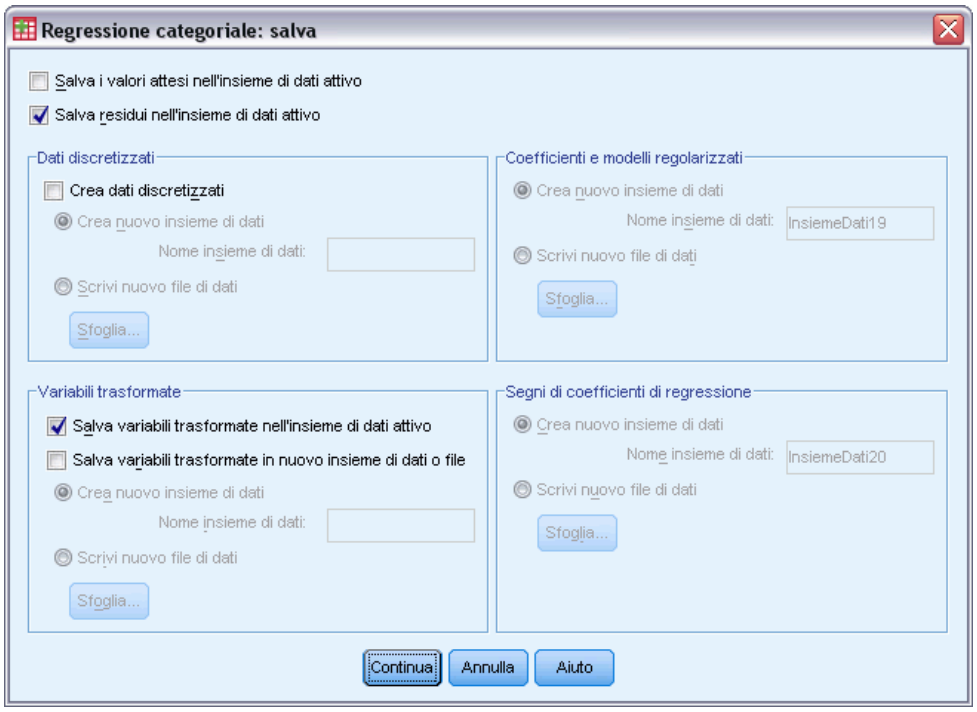
- Selezionare Numerico come livello di scaling ottimale.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale fare clic su Output.

Figura 9-13
Finestra di dialogo Output



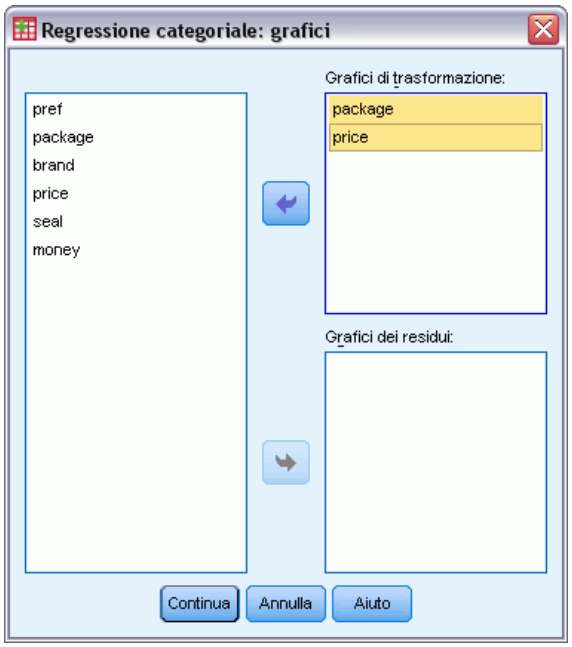
- Selezionare Correlazioni delle variabili originalie Correlazioni delle variabili trasformate .
- Deselezionare ANOVA.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale fare clic su Salva.

Figura 9-14
Salva



- Selezionare Salva residui nel file di dati attivo.
- Nel gruppo Variabili trasformate, selezionare Salva variabili trasformate nel file di dati attivo.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale fare clic su Grafici.

Figura 9-15
Finestra di dialogo Grafici



- Scegliere di creare grafici di trasformazione per *Design confezione* e *Prezzo*.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale scegliere OK.

Intercorrelazioni

Le intercorrelazioni tra i predittori sono utili per identificare la multicollinearità nella regressione. Le variabili strettamente correlate condurranno a stime di regressione instabili. Tuttavia, a causa dell'elevata correlazione, l'omissione di una di esse dal modello influenza la previsione in misura minima. La varianza nella risposta che può essere spiegata dalla variabile omessa rimane spiegata dalla variabile correlata rimanente. Tuttavia, le correlazioni di ordine zero sono sensibili ai valori anomali e inoltre non sono in grado di identificare la multicollinearità a causa dell'elevata correlazione tra un predittore e una combinazione degli altri predittori.

Figura 9-16
Correlazioni tra i predittori originali

	Confezione	Marca	Prezzo	Sigillo di garanzia	Soddisfatti o rimborsati
Confezione	1,000	-,189	-,126	,081	,066
Marca	-,189	1,000	,065	-,042	-,034
Prezzo	-,126	,065	1,000	,000	,000
Sigillo di garanzia	,081	-,042	,000	1,000	-,039
Soddisfatti o rimborsati	,066	-,034	,000	-,039	1,000
Dimensione	1	2	3	4	5
Autovalore	1,291	1,038	,980	,905	,785

Figura 9-17
Correlazioni tra i predittori trasformati

	Confezione	Marca	Prezzo	Sigillo di garanzia	Soddisfatti o rimborsati
Confezione	1,000	-,156	-,089	,032	,102
Marca	-,156	1,000	,065	-,042	-,034
Prezzo	-,089	,065	1,000	,000	,000
Sigillo di garanzia	,032	-,042	,000	1,000	-,039
Soddisfatti o rimborsati	,102	-,034	,000	-,039	1,000
Dimensione	1	2	3	4	5
Autovalore	1,248	1,043	,983	,905	,821

Vengono visualizzate le intercorrelazioni dei predittori per i predittori trasformati e non trasformati. Tutti i valori sono vicini allo 0, a indicare che la multicollinearità tra le singole variabili non rappresenta un problema.

Si noti che le sole correlazioni che si modificano riguardano *Design confezione*. Poiché tutti gli altri predittori sono trattati numericamente, le differenze tra le categorie e l'ordine di queste sono conservati per queste variabili. Di conseguenza, le correlazioni non possono modificarsi.

Adattamento del modello e coefficienti

La procedura di regressione categoriale genera un R^2 pari a 0,948, a indicare che circa il 95% della varianza delle classificazioni di preferenza trasformata è spiegata dalla regressione nei predittori trasformati in modo ottimale. La trasformazione dei predittori migliora l'adattamento rispetto all'approccio standard.

Figura 9-18
Riepilogo del modello per regressione categoriale

R multiplo	R quadrato	R quadrato corretto
,974	,948	,927

Variabile dipendente: Preferenza
Stimatori: Confezione Marca Prezzo
Sigillo di garanzia Soddisfatti o rimborsati

La seguente tabella mostra i coefficienti di regressione standardizzati. La regressione categoriale determina la standardizzazione delle variabili, di conseguenza solo i coefficienti standardizzati vengono riportati. Questi valori sono divisi per gli errori standard corrispondenti, generando un test *F* per ogni variabile. Tuttavia, il test per ogni variabile è contingente rispetto agli altri predittori nel modello. In altre parole, il test determina se l'omissione di una variabile di predittore dal modello in presenza di tutti gli altri predittori peggiora in modo significativo le capacità di previsione del modello stesso. Questi valori non dovrebbero essere utilizzati per l'omissione contemporanea di molte variabili per un modello successivo. Inoltre, il metodo dei minimi

quadrati alternati ottimizza le quantificazioni, il che implica che questi test devono essere interpretati in modo conservativo.

Figura 9-19
Coefficienti standardizzati per predittori trasformati

	Coefficienti standardizzati				
	Beta	Errore std	df	F	Sig.
Confezione	-,748	,060	2	155,289	,000
Marca	,045	,060	1	,578	,459
Prezzo	,371	,059	1	39,312	,000
Sigillo di garanzia	-,350	,059	1	35,299	,000
Soddisfatti o rimborsati	-,159	,059	1	7,175	,017

Variabile dipendente: Preferenza

Il coefficiente maggiore è relativo a *Design confezione*. Un aumento di deviazione standard di *Design confezione* genera una deviazione standard pari a 0,748 nella classificazione della preferenza prevista. Tuttavia, *Design confezione* viene trattato normalmente, quando un aumento delle quantificazioni non deve corrispondere a un aumento dei codici di categoria originali.

I coefficienti standardizzati sono spesso interpretati come indicativi dell'importanza di ogni predittore. Tuttavia, i coefficienti di regressione non possono descrivere completamente l'impatto di un predittore o le relazioni tra i predittori. È necessario utilizzare statistiche alternative in combinazione con i coefficienti standardizzati per esaminare in modo completo gli effetti dei predittori.

Correlazioni e importanza

Per interpretare i contributi dei predittori alla regressione, non è sufficiente limitarsi a esaminare i coefficienti di regressione. Inoltre è necessario esaminare le correlazioni, le correlazioni di parte e le correlazioni parziali. La seguente tabella include le misure delle correlazioni citate per ogni variabile.

La correlazione di ordine zero è quella tra il predittore trasformato e la risposta trasformata. Per questi dati, la correlazione maggiore si verifica per *Design confezione*. Tuttavia, se è possibile spiegare parte della variazione nel predittore o nella risposta, si otterrà una migliore rappresentazione delle prestazioni del predittore.

Figura 9-20
Correlazioni parziali, di parte e di ordine zero (variabili trasformate)

	Correlazioni				Tolleranza	
	Ordine zero	Parziali	Parziali indipendenti	Importanza	Dopo la trasformazione	Prima della trasformazione
Confezione	-,816	-,955	-,733	,644	,959	,942
Marca	,206	,193	,045	,010	,971	,961
Prezzo	,440	,851	,369	,172	,989	,982
Sigillo di garanzia	-,370	-,838	-,349	,137	,996	,991
Soddisfatti o rimborsati	-,223	-,569	-,158	,037	,987	,993

Variabile dipendente: Preferenza

Altre variabili nel modello possono creare confusione circa le prestazioni di un dato predittore per quanto concerne le previsioni della risposta. Il coefficiente di correlazione parziale rimuove gli effetti lineari di altri predittori dal predittore e dalla risposta. Questa misura è pari alla correlazione tra i residui derivanti dalla regressione del predittore sugli altri e i residui derivanti dalla regressione della risposta sugli altri predittori. La correlazione parziale quadrata corrisponde alla proporzione della varianza spiegata relativa alla varianza residua della risposta rimanente dopo la rimozione degli effetti delle altre variabili. Ad esempio, *Design confezione* ha una correlazione parziale di –0,955. Rimuovendo gli effetti delle altre variabili, *Design confezione* spiega $(–0,955)^2 = 0,91 = 91\%$ della variazione delle classificazioni della preferenza. Sia *Prezzo* che *Marchio di qualità* spiegano anch'essi una parte significativa della varianza se gli effetti delle altre variabili vengono rimossi.

In alternativa alla rimozione degli effetti delle variabili dalla risposta e da un predittore, è possibile rimuovere gli effetti solo dal predittore. La correlazione tra la risposta e i residui derivanti dalla regressione di un predittore sugli altri è la correlazione di parte. Elevando al quadrato tale valore si ottiene una misura della proporzione della varianza spiegata rispetto alla varianza totale della risposta. Se si rimuovono gli effetti di *Nome marca*, *Marchio di qualità*, *Garanzia "Soddisfatti o rimborsati"* e *Prezzo* da *Design confezione*, la parte restante di *Design confezione* spiega $(–0,733)^2 = 0,54 = 54\%$ della variazione nelle classificazioni della preferenza.

Importanza

Oltre ai coefficienti di regressione e alle correlazioni, la misura di importanza relativa di Pratt (Pratt, 1987) consente di interpretare i contributi dei predittori alla regressione. Singoli valori di importanza elevati rispetto ad altri corrispondono a predittori di importanza chiave per la regressione. Inoltre, la presenza di variabili di soppressore è indicata da un'importanza ridotta per una variabile con coefficiente di dimensioni analoghe ai predittori importanti.

In contrasto con i coefficienti di regressione, questa misura definisce l'importanza dei predittori additivamente, ovvero, l'importanza di un insieme di predittori è la somma delle importanze dei singoli predittori. La misura di importanza relativa di Pratt è pari al prodotto del coefficiente di regressione e alla correlazione di ordine zero per un predittore. Questi prodotti si aggiungono a R^2, quindi vengono divisi per R^2, generando una somma pari a 1. L'insieme di predittori*Design confezione* e *Nome marca*, ad esempio, hanno importanza pari a 0,654. L'importanza maggiore corrisponde a *Design confezione*, con *Design confezione*, *Prezzo* e *Marchio di qualità* che spiegano il 95% dell'importanza per questa combinazione di predittori.

Multicollinearità

Correlazioni ampie tra i predittori ridurranno notevolmente la stabilità di un modello di regressione. Predittori correlati determineranno stime dei parametri instabili. La tolleranza riflette il grado di reciproca relazione lineare tra le variabili indipendenti. Questa misura è la proporzione della varianza di una variabile non spiegata dalle altre variabili indipendenti dell'equazione. Se gli altri predittori possono spiegare una quantità elevata della varianza di un predittore, quest'ultimo non è necessario nel modello. Un valore di tolleranza vicino a 1 indica che la variabile non può essere prevista con grande affidabilità a partire dagli altri predittori. Per contro, una variabile con una tolleranza molto bassa apporta poche informazioni a un modello e può causare problemi di calcolo. Inoltre, elevati valori negativi della misura di importanza di Pratt sono indicativi di multicollinearità.

Tutte le misure di tolleranza sono molto elevate. Nessuno dei predittori è previsto con grande affidabilità dagli altri ed è presente multicollinearità.

Grafici di trasformazione

Tracciando i valori della categoria originale rispetto alle quantificazioni corrispondenti è possibile evidenziare trend che potrebbero non venire notati in un elenco delle quantificazioni. Tali grafici sono normalmente definiti grafici di trasformazione. Prestare attenzione alle categorie che ricevono quantificazioni simili. Queste categorie influenzano la risposta prevista nello stesso modo. Tuttavia, il tipo di trasformazione definisce l'aspetto di base del grafico.

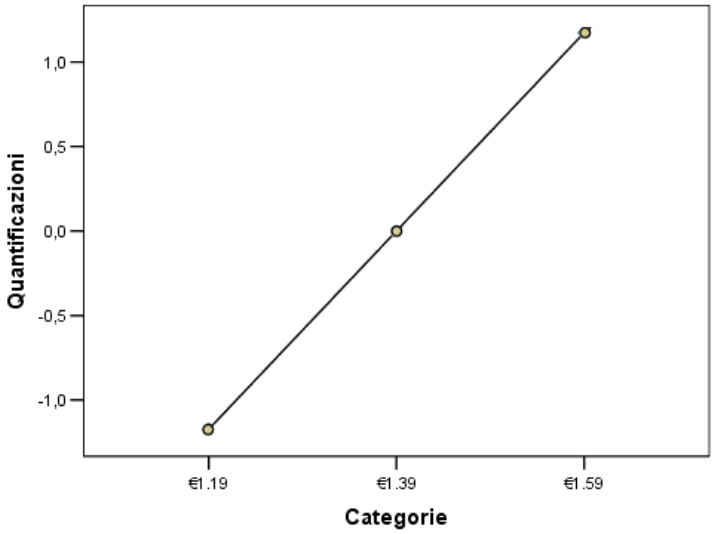
Le variabili trattate come numeriche determinano una relazione lineare tra le quantificazioni e le categorie originali, corrispondente a una linea retta nel grafico di trasformazione. L'ordine e la differenza tra le categorie originali vengono mantenuti nelle quantificazioni.

L'ordine delle quantificazioni per le variabili trattate come ordinali corrisponde all'ordine delle categorie originali. Tuttavia, le differenze tra le categorie non vengono mantenute. Di conseguenza, il grafico di trasformazione è non decrescente ma non deve essere necessariamente una linea retta. Se categorie consecutive corrispondono a quantificazioni simili, la distinzione tra categorie potrebbe essere superflua e le categorie combinate. Tali categorie danno come risultato un plateau nel grafico di trasformazione. Tuttavia, questo modello può anche derivare dall'imposizione di una struttura ordinale a una variabile che dovrebbe essere trattata come nominale. Se un successivo trattamento nominale della variabile presenta lo stesso modello, la combinazione delle categorie è opportuna. Inoltre, se le quantificazioni per una variabile trattata come ordinale corrispondono a una linea retta, una trasformazione numerica può essere più adatta.

Per le variabili trattate come nominali, l'ordine delle categorie lungo l'asse orizzontale corrisponde all'ordine dei codici utilizzati per rappresentare le categorie. Le interpretazioni dell'ordine delle categorie o della distanza tra le categorie sono infondate. Il grafico può assumere qualsiasi forma lineare o non lineare. Se è presente un trend crescente, tentare di eseguire un trattamento ordinale. Se il grafico di trasformazione nominale visualizza un trend lineare, una trasformazione numerica potrebbe essere più adatta.

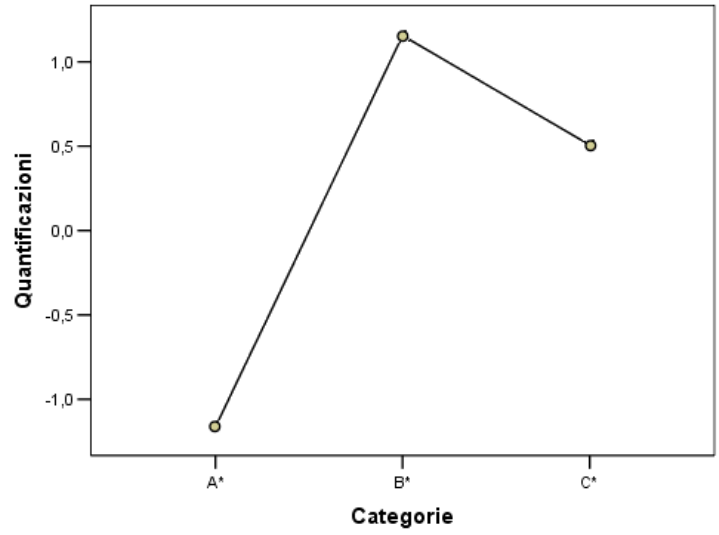
La figura seguente visualizza il grafico di trasformazione per *Prezzo*, trattato come numerico. Si noti che l'ordine delle categorie lungo la linea retta corrisponde all'ordine delle categorie originali. Inoltre, la differenza tra le quantificazioni per *$1,19* e *$1,39* (–1,173 e 0) è pari alla differenza tra le quantificazioni per *$1,39* e *$1,59* (0 e 1,173). Il fatto che la distanza delle categorie 1 e 3 dalla categoria 2 sia la stessa è mantenuto nelle quantificazioni.

Figura 9-21
Grafico di trasformazione del prezzo (numerico)



La trasformazione nominale di *Design confezione* genera il seguente grafico di trasformazione. Si noti la forma non lineare distinta in cui la seconda categoria ha la quantificazione maggiore. In termini di regressione, la seconda categoria riduce la classificazione della preferenza prevista, mentre la prima e la terza categoria hanno l'effetto opposto.

Figura 9-22
Grafico di trasformazione per Design confezione (nominale)

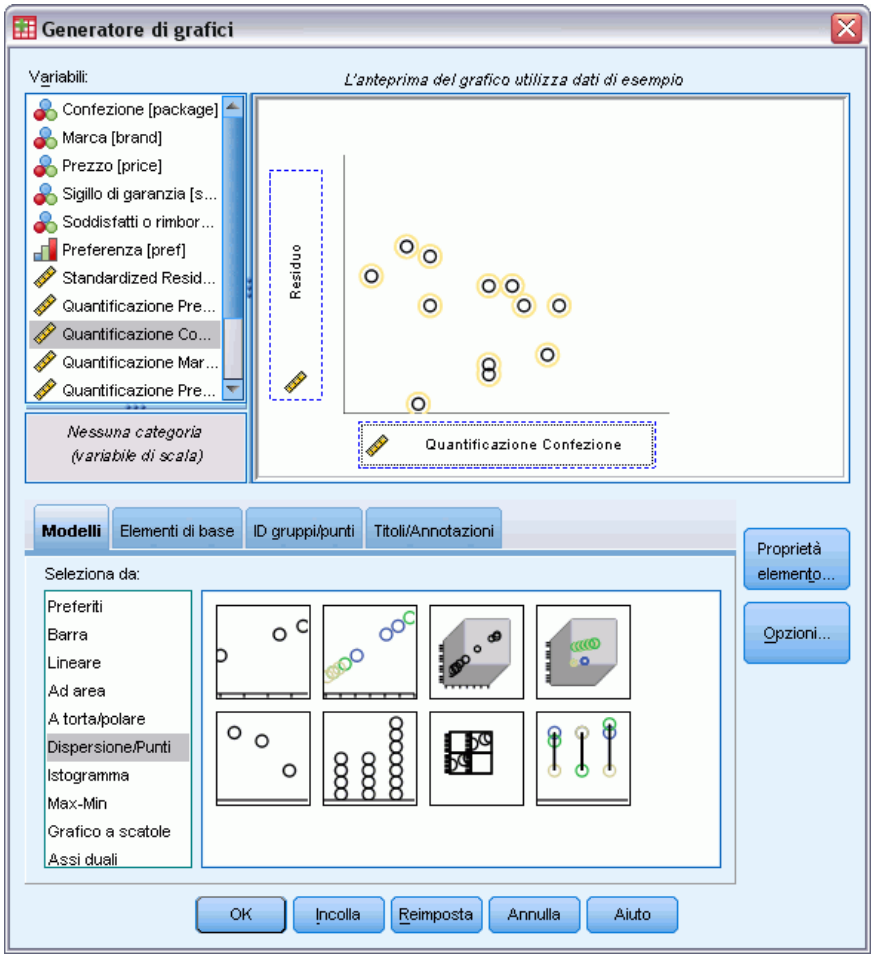


Analisi dei residui

Utilizzando i dati trasformati e i residui salvati nel file di dati attivo è possibile creare un grafico a dispersione dei valori attesi a partire dai valori trasformati di *Design confezione*.

Per ottenere tale grafico, richiamare Generatore di grafici e fare clic su Ripristina per annullare le selezioni precedenti e ripristinare le opzioni predefinite.

Figura 9-23
Generatore di grafici

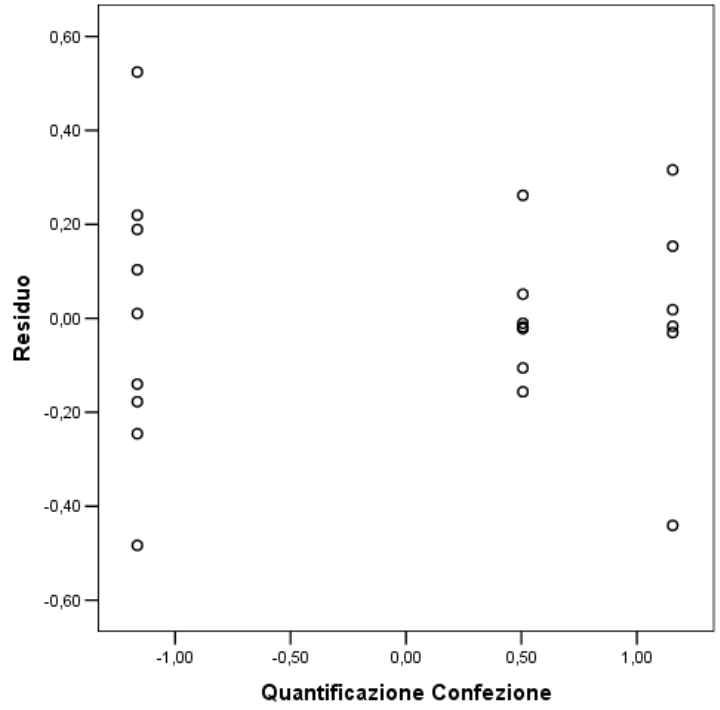


- Selezionare il modello Dispersione/Punti e scegliere Dispersione semplice.
- Selezionare *Residuo* come variabile dell'asse *y*.
- Selezionare *Quantificazione design confezione* come variabile dell'asse *x*.
- Fare clic su OK.

Il grafico a dispersione mostra i residui standardizzati tracciati rispetto ai punteggi ottimali per *Design confezione*. Tutti i residui sono compresi entro le due deviazioni standard di 0. Una dispersione casuale di punti sostituisce la forma a U nel grafico a dispersione derivato dalla

regressione lineare standard. Le capacità predittive vengono migliorate dalla quantificazione ottimale delle categorie.

Figura 9-24
Residui per regressione categoriale



Esempio: Dati sull'ozono

Nell'esempio, verrà utilizzato un insieme più ampio di dati per illustrare la selezione e gli effetti delle trasformazioni con scaling ottimale. I dati includono 330 osservazioni su sei variabili meteorologiche precedentemente analizzate, tra gli altri, da Breiman e Friedman (Breiman e Friedman, 1985) e da Hastie e Tibshirani (Hastie e Tibshirani, 1990). La seguente tabella descrive le variabili originali. La regressione categoriale tenta di prevedere la concentrazione di ozono dalle variabili restanti. I precedenti ricercatori hanno rilevato non linearità tra queste variabili, che impediscono un approccio di regressione standard.

Tabella 9-2
Variabili originali

Variabile	Descrizione
ozono	livello ozono giornaliero; categorizzato in una di 38 categorie
abi	altezza di base inversione
gp	gradiente pressione (mm Hg)
vis	visibilità (in miglia)

Variabile	Descrizione
temp	temperatura (gradi F)
gda	giorno dell'anno

Questo insiemi di dati è reperibile nel file *ozone.sav*.Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*.

Discretizzazione delle variabili

Se una variabile ha più categorie di quante siano effettivamente interpretabili, è necessario modificare le categorie utilizzando la finestra di dialogo Discretizzazione per ridurne la gamma a un numero più gestibile.

La variabile *Giorno dell'anno* ha un valore minimo di 3 e un valore massimo di 365. Il suo utilizzo in una regressione categoriale corrisponde all'utilizzo di una variabile con 365 categorie. Analogamente, *Visibilità (in miglia)* varia da 0 a 350. Per semplificare l'interpretazione delle analisi, discretizzare le variabili in intervalli uguali di lunghezza 10.

La variabile *Altezza di base inversione* varia da 111 a 5000. Una variabile con questo numero di categorie determinerà relazioni molto complesse. Tuttavia, la discretizzazione di questa variabile in intervalli uguali di lunghezza 100 genera circa 50 categorie. Utilizzando una variabile con 50 categorie anziché una variabile con 500 semplifica notevolmente le interpretazioni.

Gradiente pressione (mm Hg) varia da –69 a 107. La procedura esclude dall'analisi eventuali categorie codificate con numeri negativi, ma la discretizzazione della variabile in intervalli uguali di lunghezza 10 genera circa 19 categorie.

Temperatura (gradi F) varia da 25 a 93 sulla scala Fahrenheit. Per analizzare i dati come se fossero espressi sulla scala Celsius, discretizzare la variabile in intervalli uguali di lunghezza 1,8.

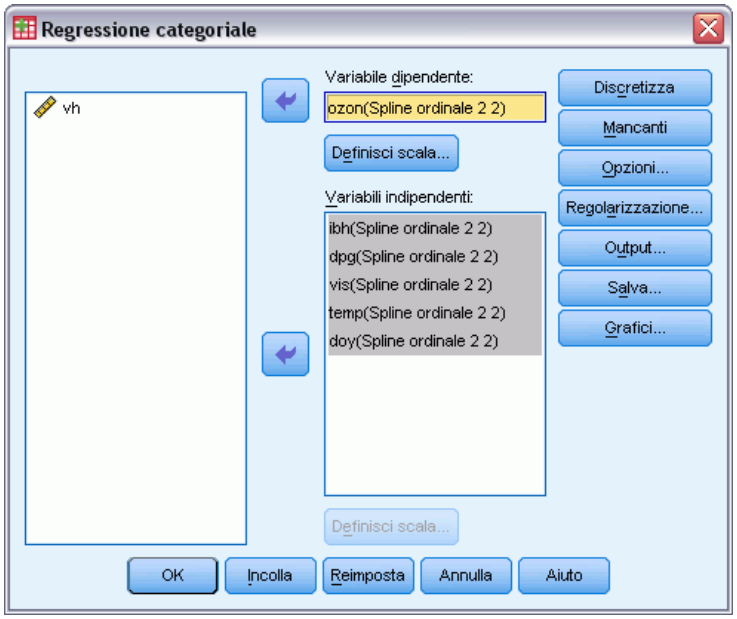
È possibile che siano necessarie discretizzazioni diverse per le variabili. Le scelte utilizzate nell'esempio sono puramente soggettive. Per ottenere un numero inferiore di categorie, scegliere intervalli più ampi. Ad esempio, *Giorno dell'anno* potrebbe essere diviso in mesi dell'anno o stagioni.

Selezione del tipo di trasformazione

Ciascuna variabile può essere analizzata a diversi livelli. Tuttavia, poiché l'obiettivo è la previsione della risposta, si consiglia di scalare la risposta "così com'è" utilizzando il livello di scaling ottimale numerico. Di conseguenza, l'ordine e le differenze tra le categorie saranno mantenuti nella variabile trasformata.

▶ Per eseguire un'analisi di regressione categoriale, dai menu scegliere:
Analizza > Regression > Scaling ottimale (CATREG)...

Figura 9-25
Finestra di dialogo Regressione categoriale



- Selezionare *Livello ozono giornaliero* come variabile dipendente.
- Selezionare da *Altezza di base inversione* a *Giorno dell'anno* come variabili indipendenti.
- Selezionare *Livello ozono giornaliero* e fare clic su Definisci scala.

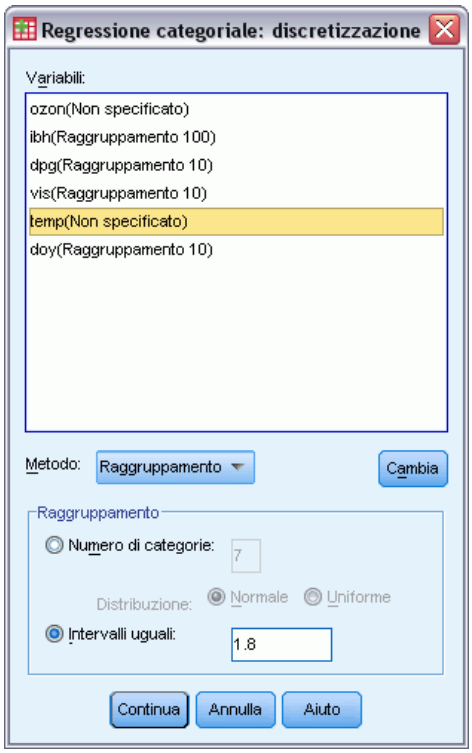
Figura 9-26
Finestra di dialogo Definisci scala

- Selezionare Numerico come livello di scaling ottimale.
- Fare clic su Continua.
- Selezionare da *Altezza di base inversione* a *Giorno dell'anno* e fare clic su Definisci scala nella finestra di dialogo Regressione categoriale.

Figura 9-27
Finestra di dialogo Definisci scala

- Selezionare Nominale come livello di scaling ottimale.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale scegliere Discretizza.

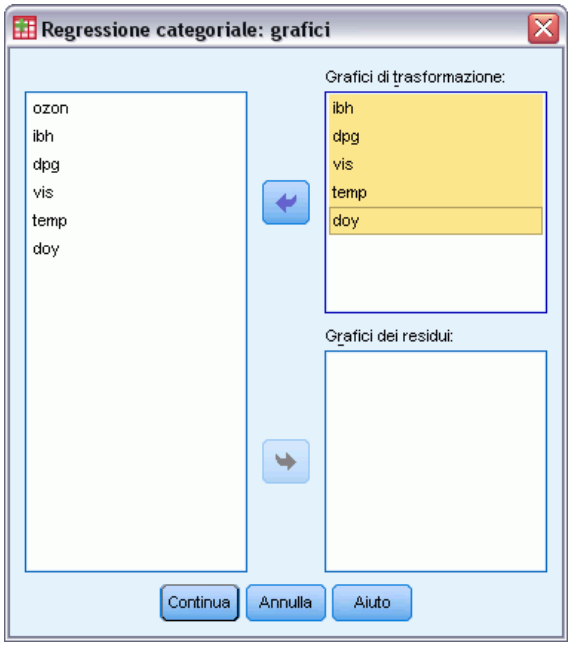
Figura 9-28
Finestra di dialogo Discretizza



- Selezionare *abi*.
- Selezionare Intervalli uguali e digitare 100 come lunghezza dell'intervallo.
- Fare clic su Cambia.

- Selezionare *gp*, *vis* e *gda*.
- Digitare 10 come lunghezza dell'intervallo.
- Fare clic su Cambia.
- Selezionare *temp*.
- Digitare 1.8 come lunghezza dell'intervallo.
- Fare clic su Cambia.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale fare clic su Grafici.

Figura 9-29
Finestra di dialogo Grafici



- Selezionare i grafici di trasformazione da *Altezza di base inversione* a *Giorno dell'anno*.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale scegliere OK.

Figura 9-30
Riepilogo modello

	R multiplo	R-quadrato	R-quadrato corretto	Errore di previsione apparente
Dati standardizzati	.938	.880	.785	.120

Variabile dipendente: Daily ozone level
Predittori: Inversion base height Pressure gradient (mm Hg) Visibility (miles) Temperature (degrees F) Day of the year

Il trattamento di tutti i predittori come nominali genera un R^2 pari a 0.880. Questa ampia porzione di varianza spiegata non è sorprendente perché il trattamento nominale non impone vincoli sulle quantificazioni. Tuttavia, l'interpretazione dei risultati può essere alquanto complessa.

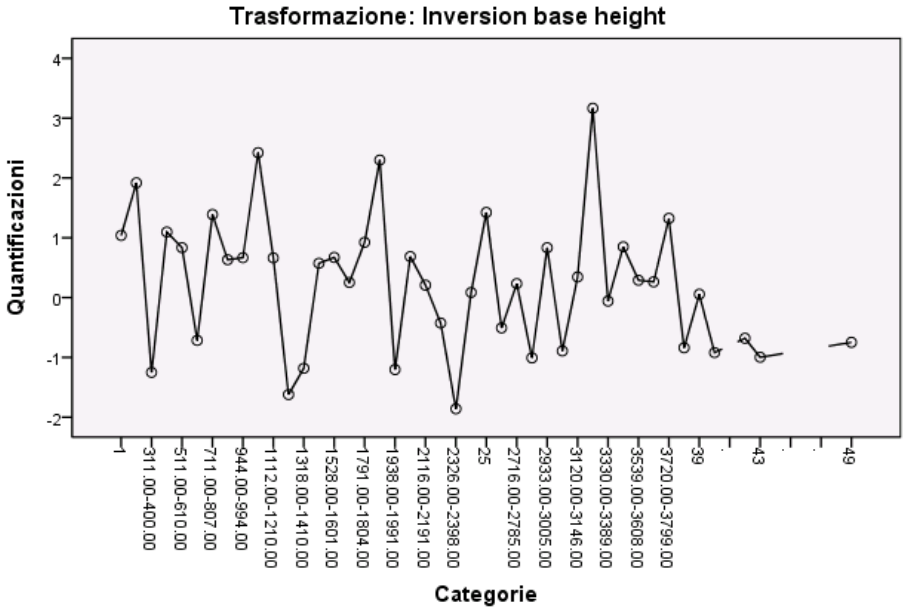
Figura 9-31
Coefficienti di regressione (tutti predittori nominali)

	Coefficienti standardizzati				
	Beta	Stima bootstrap (1000) dell'errore std.	df	F	Sig.
Inversion base height	.297	.053	42	31.047	.000
Pressure gradient (mm Hg)	.326	.055	16	34.793	.000
Visibility (miles)	.229	.050	17	21.465	.000
Temperature (degrees F)	.577	.091	35	40.562	.000
Day of the year	.420	.070	36	36.171	.000

Variabile dipendente: Daily ozone level

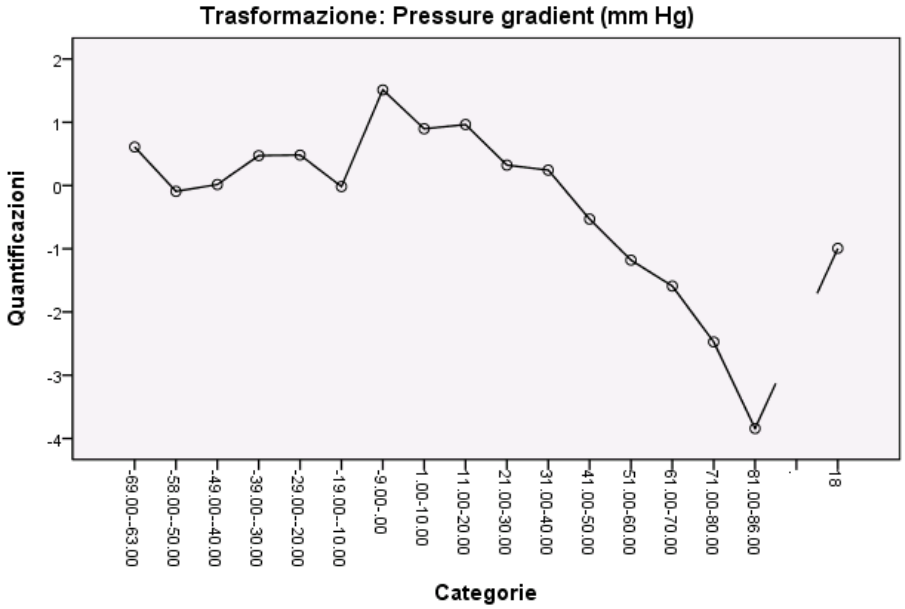
La seguente tabella mostra i coefficienti di regressione standardizzati dei predittori. Un errore comune nell'interpretazione di questi valori consiste nel concentrarsi sui coefficienti trascurando le quantificazioni. Non è possibile affermare semplicemente che un valore positivo di *Altezza di base inversione*, ad esempio, implica che quando il predittore aumenta, l'*ozono* previsto aumenta. Tutte le interpretazioni devono essere relative alle variabili trasformate, in modo che quando le quantificazioni per *Altezza di base inversione* aumentano, l'*ozono* previsto aumenta. Per esaminare gli effetti delle variabili originali, è necessario mettere in correlazione categorie e quantificazioni.

Figura 9-32
Grafico di trasformazione per Altezza di base inversione (nominale)



Il grafico di trasformazione per *Altezza di base inversione* non mostra modelli evidenti. Come evidenziato dalla natura irregolare del grafico, lo spostamento dalle categorie inferiori alle superiori genera fluttuazioni nelle quantificazioni in entrambe le direzioni. Di conseguenza, per descrivere gli effetti di questa variabile è necessario concentrarsi sulle singole categorie. L'imposizione di vincoli lineare o ordinali alle quantificazioni per questa variabile può ridurne significativamente l'adattamento.

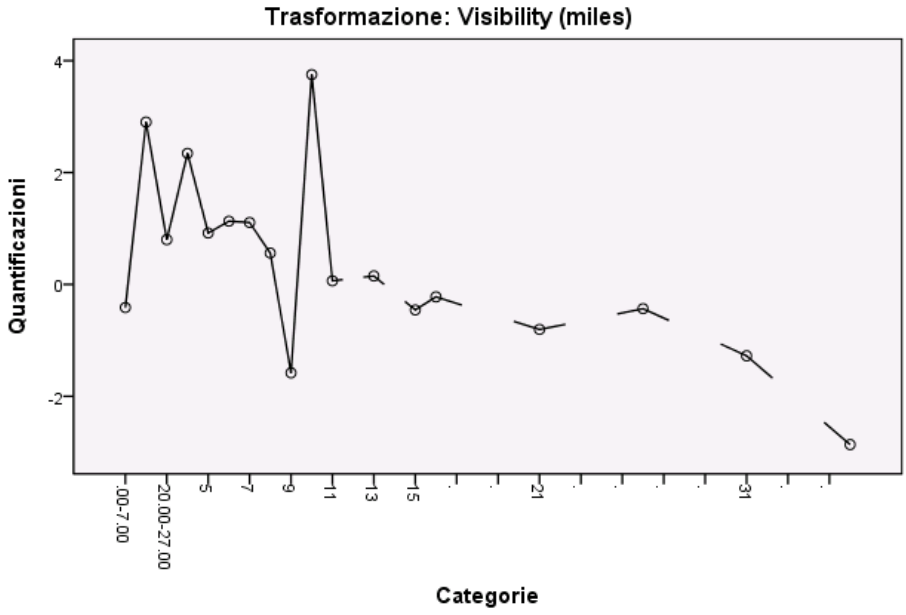
Figura 9-33
Grafico di trasformazione per Gradiente pressione (nominale)



La figura mostra il grafico di trasformazione per *Gradiente pressione*. Le categorie discretizzate iniziali (da *1* a *6*) ricevono quantificazioni limitate e quindi contribuiscono in modo ridotto alla risposta prevista. Le tre categorie successive ricevono valori più elevati e positivi, con conseguente aumento moderato del valore di ozono previsto.

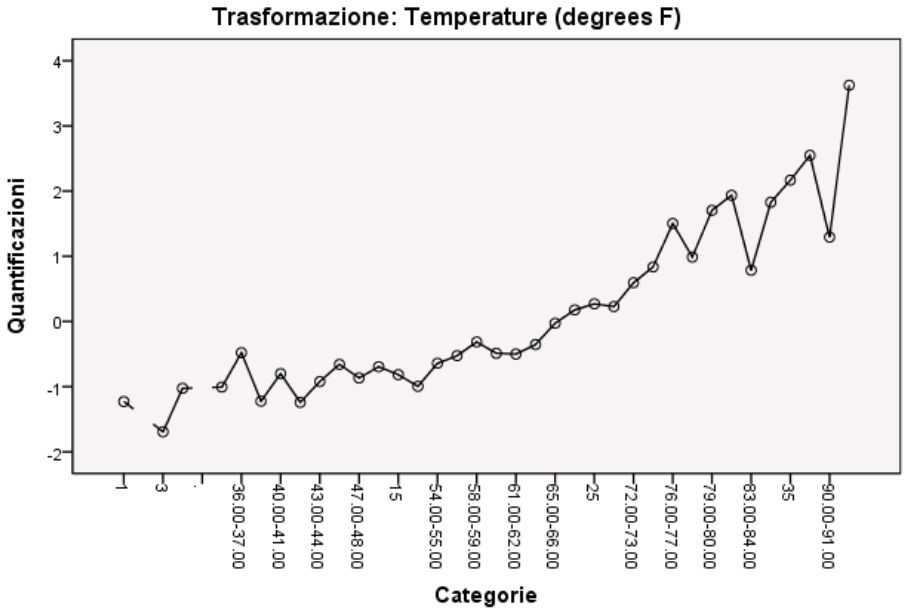
Le quantificazioni si riducono fino alla categoria *16*, per la quale *Gradiente pressione* ha l'effetto decrescente massimo sul valore di ozono previsto. Sebbene la linea aumenti dopo questa categoria, un livello di scaling ordinale per *Gradiente pressione* potrebbe non ridurre in modo significativo l'adattamento, semplificando al contempo le interpretazioni degli effetti. Tuttavia, la misura di importanza pari a 0,04 e il coefficiente di regressione per *Gradiente pressione* indica che questa variabile non è molto utile nella regressione.

Figura 9-34
Grafico di trasformazione per Visibilità (nominale)



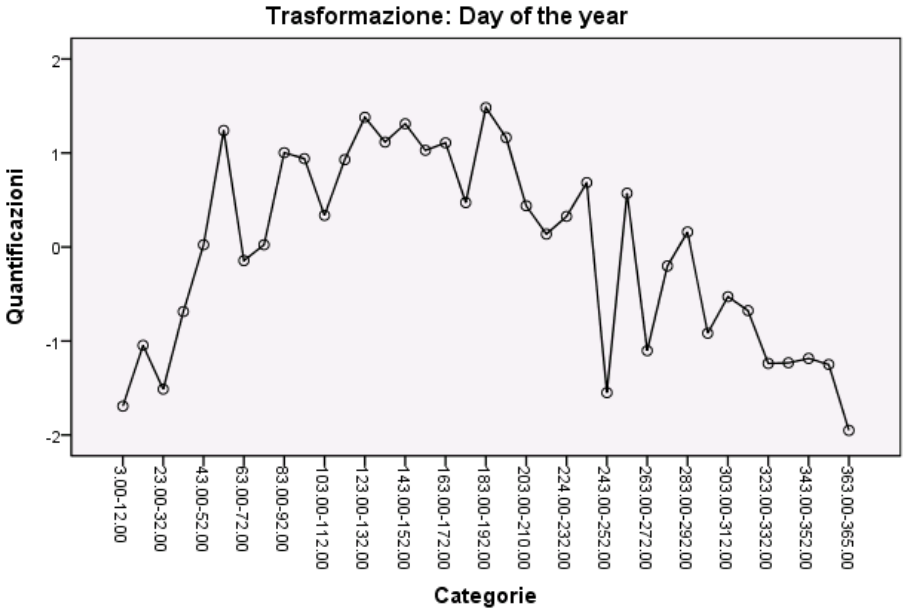
Il grafico di trasformazione per *Visibilità*, come quello per *Altezza di base inversione* non mostra modelli evidenti. L'imposizione di vincoli lineare o ordinali alle quantificazioni per questa variabile può ridurne significativamente l'adattamento.

Figura 9-35
Grafico di trasformazione per Temperatura (nominale)



Il grafico di trasformazione per *Temperatura* mostra un modello alternativo. Le quantificazioni tendono ad aumentare con l'aumento delle categorie. Come risultato, quando *Temperatura* aumenta, l'ozono previsto tende ad aumentare anch'esso. Questo modello suggerisce lo scaling di *Temperatura* a livello ordinale.

Figura 9-36
Grafico di trasformazione per Giorno dell'anno (nominale)



La figura mostra il grafico di trasformazione per *Giorno dell'anno*. Le quantificazioni tendono ad aumentare fino al punto centrale del grafico, in corrispondenza del quale tendono a diminuire, generando una forma a U invertita. Considerando il segno del coefficiente di regressione per *Giorno dell'anno*, le categorie iniziali ricevono quantificazioni con effetto decrescente sull'ozono previsto. Per le categorie centrali, l'effetto delle quantificazioni sull'ozono previsto aumenta, raggiungendo il valore massimo intorno al punto centrale del grafico.

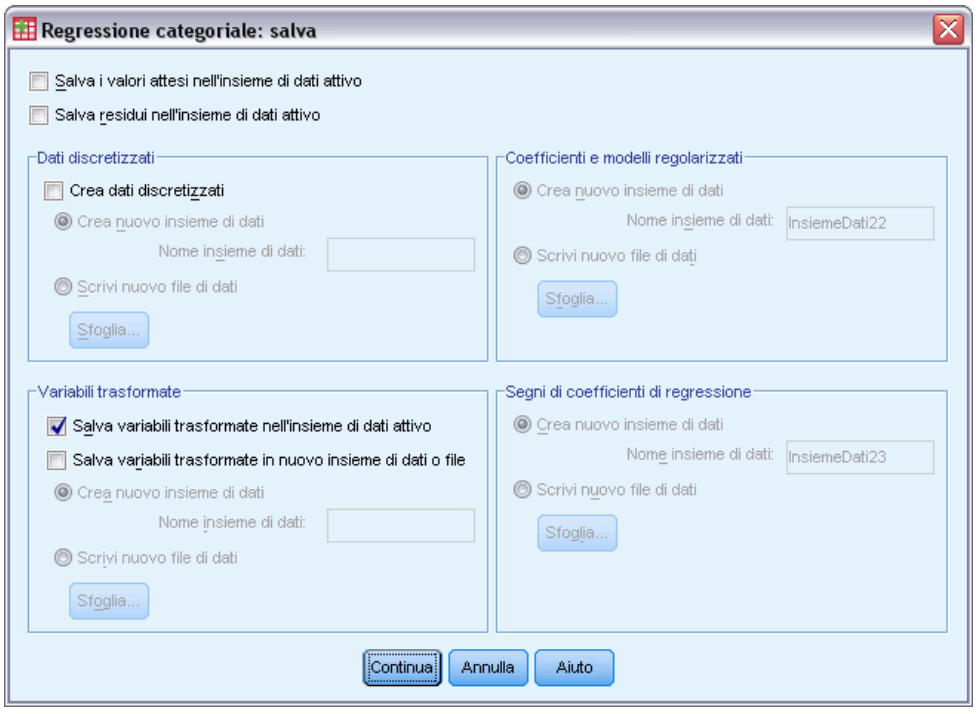
Oltre quel punto, le quantificazioni tendono a ridurre l'ozono previsto. Sebbene la linea sia piuttosto irregolare, la forma generale rimane identificabile. Di conseguenza, i grafici di trasformazione suggeriscono lo scaling di *Temperatura* a livello ordinale mantenendo al contempo lo scaling nominale di tutti gli altri predittori.

Per ricalcolare la regressione, eseguendo lo scaling di *Temperatura* a livello ordinale, richiamare la finestra di dialogo Regressione categoriale.

Figura 9-37
Finestra di dialogo Definisci scala

- Selezionare *Temperatura* e fare clic su Definisci scala.
- Selezionare Ordinale come livello di scaling ottimale.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale fare clic su Salva.

Figura 9-38
Salva



- Nel gruppo Variabili trasformate, selezionare Salva variabili trasformate nel file di dati attivo.

- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale scegliere OK.

Figura 9-39
Riepilogo del modello per regressione con Temperatura (ordinale)

Riepilogo del modello

	R multiplo	R-quadrato	R-quadrato corretto	Errore di previsione apparente
Dati standardizzati	.934	.872	.787	.128

Variabile dipendente: Daily ozone level
Predittori: Inversion base height Pressure gradient (mm Hg) Visibility (miles) Temperature (degrees F) Day of the year

Il modello ha come risultato un R^2 pari a 0,872, quindi la varianza spiegata si riduce in modo trascurabile quando le quantificazioni per *Temperatura* sono vincolate a essere ordinate.

Figura 9-40
Coefficienti di regressione con Temperatura (ordinale)

Coefficienti

	Coefficienti standardizzati				
	Beta	Stima bootstrap (1000) dell'errore std.	df	F	Sig.
Inversion base height	.298	.042	42	51.376	.000
Pressure gradient (mm Hg)	.301	.047	16	41.882	.000
Visibility (miles)	.224	.044	17	25.659	.000
Temperature (degrees F)	.609	.084	21	52.113	.000
Day of the year	.373	.053	36	50.330	.000

Variabile dipendente: Daily ozone level

Questa tabella visualizza i coefficienti per il modello in cui *Temperatura* è scalata come ordinale. Confrontando i coefficienti per il modello in cui *Temperatura* è scalata come nominale, non si evidenziano modifiche di rilievo.

Figura 9-41
Correlazioni, importanza e tolleranza

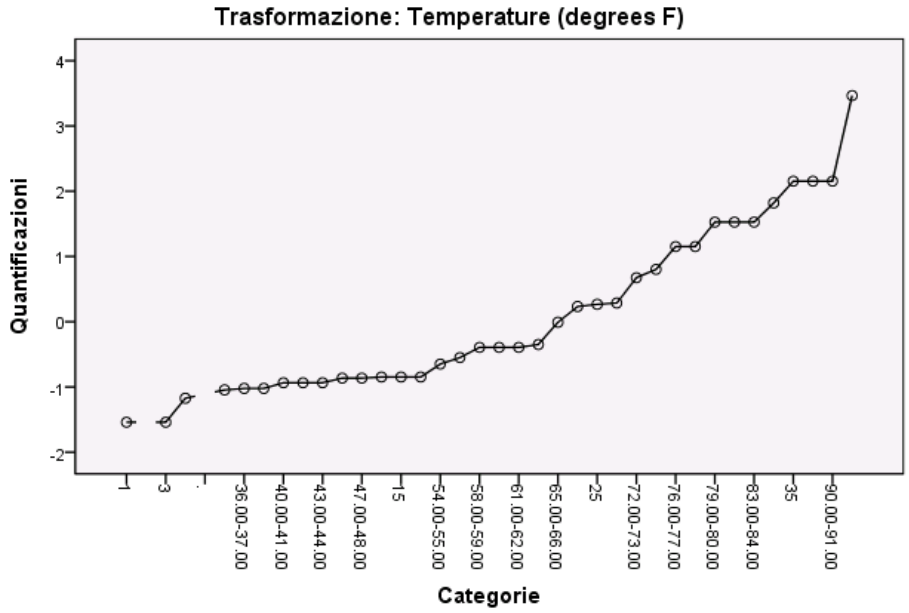
Correlazioni e tolleranza

	Correlazioni			Importanza	Tolleranza	
	Ordine zero	Parziali	Parziali indipendenti		Dopo la trasformazione	Prima della trasformazione
Inversion base height	.438	.627	.288	.150	.930	.596
Pressure gradient (mm Hg)	.128	.606	.272	.044	.815	.858
Visibility (miles)	.365	.518	.216	.094	.933	.752
Temperature (degrees F)	.804	.843	.559	.562	.842	.580
Day of the year	.352	.677	.329	.151	.777	.802

Variabile dipendente: Daily ozone level

Inoltre, le misure di importanza suggeriscono che *Temperatura* resti molto più importante della regressione che delle altre variabili. Ora, tuttavia, come risultato del livello di scaling ordinale di *Temperatura* e del coefficiente di regressione positivo, è possibile affermare che, se *Temperatura* aumenta, il valore atteso di ozono aumenta anch'esso.

Figura 9-42
Grafico di trasformazione per Temperatura (ordinale)



Il grafico di trasformazione illustra il vincolo ordinale sulle quantificazioni per *Temperatura*. La linea irregolare dalla trasformazione nominale viene qui sostituita da una linea crescente regolare. Inoltre, non sono presenti ampi plateau, a indicare che la compressione delle categorie non è necessaria.

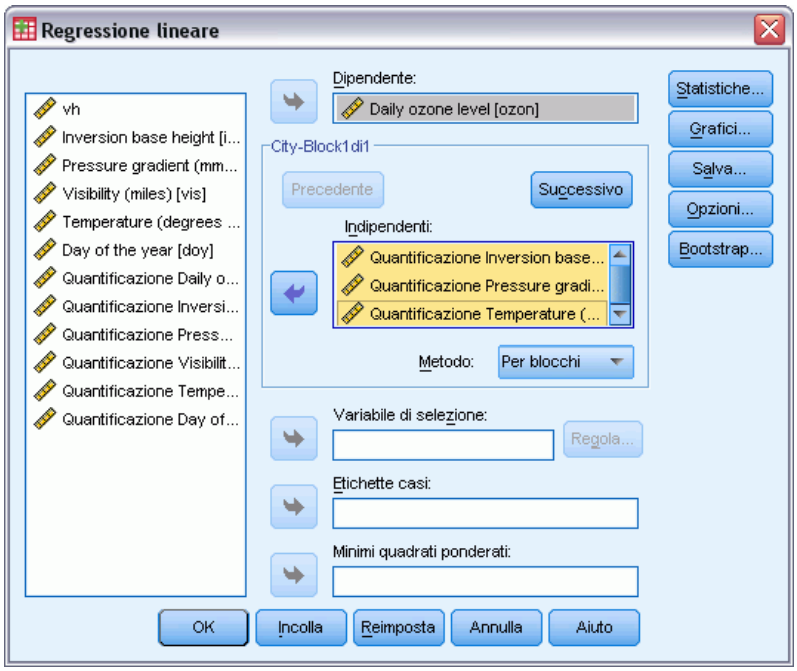
Ottimalità delle quantificazioni

Le variabili trasformate da una regressione categoriale possono essere utilizzate in una regressione lineare standard, generando risultati identici. Tuttavia, le quantificazioni sono ottimali solo per il modello che le ha prodotte. L'utilizzo di un sottoinsieme dei predittori nella regressione lineare non corrisponde a eseguire una regressione con scaling ottimale sullo stesso sottoinsieme.

Ad esempio, la regressione categoriale calcolata ha un R^2 pari a 0,875. Le variabili trasformate sono state salvate, perciò per adattare una regressione lineare utilizzando solo *Temperatura*, *Gradiente pressione* e *Altezza di base inversione* come predittori, dai menu scegliere:

Analizza > Regression > Lineare...

Figura 9-43
Finestra di dialogo Regressione lineare



- Selezionare *Quantificazione Valore ozono giornaliero* come variabile dipendente.
- Selezionare da *Quantificazione Altezza di base inversione*, *Quantificazione Gradiente pressione (mm Hg)* e *Temperatura (gradi F)* come variabili indipendenti.
- Fare clic su OK.

Figura 9-44
Riepilogo del modello per regressione con sottoinsieme di predittori con scaling ottimale

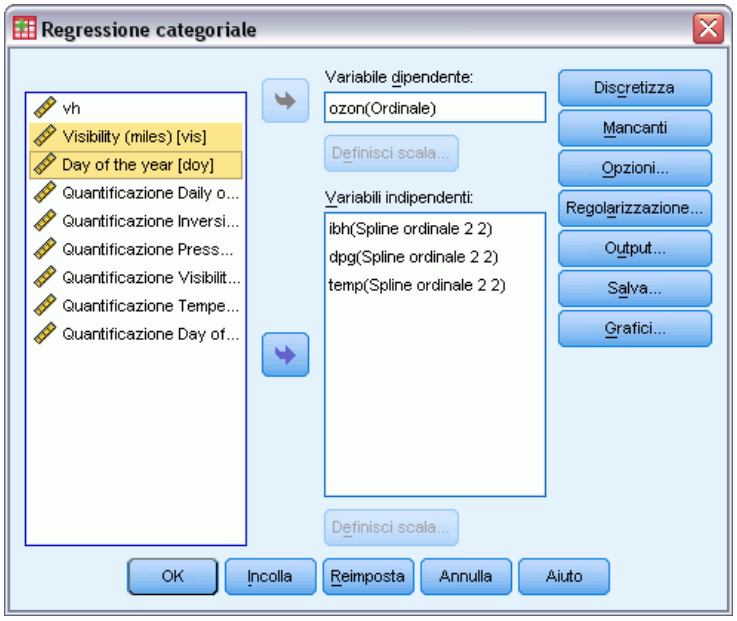
Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Deviazione standard Errore della stima
1	.856[a]	.732	.729	4.16711

a. Predittori: (Costante), Quantificazione Temperature (degrees F), Quantificazione Pressure gradient (mm Hg), Quantificazione Inversion base height

Utilizzando le quantificazioni per la risposta, *Temperatura*, *Gradiente pressione* e *Altezza di base inversione* in una regressione lineare standard determinano un adattamento pari a 0.732. Per confrontarlo con l'adattamento di una regressione categoriale utilizzando solo questi tre predittori, richiamare la finestra di dialogo Regressione categoriale.

Figura 9-45
Finestra di dialogo Regressione categoriale



- Deselezionare *Visibilità (miglia)* e *Giorno dell'anno* come variabili indipendenti.
- Fare clic su OK.

Figura 9-46
Riepilogo del modello per regressione categoriale sui tre predittori

	R multiplo	R-quadrato	R-quadrato corretto	Errore di previsione apparente
Dati standardizzati	.892	.796	.735	.204

Variabile dipendente: Daily ozone level
Predittori: Inversion base height Pressure gradient (mm Hg) Temperature (degrees F)

L'analisi di regressione categoriale ha un adattamento pari a 0.796, migliore di 0.732. Questo dimostra la proprietà degli scaling consistente nel fatto che le quantificazioni ottenute nella regressione originale sono ottimali solo quando tutte le cinque variabili sono incluse nel modello.

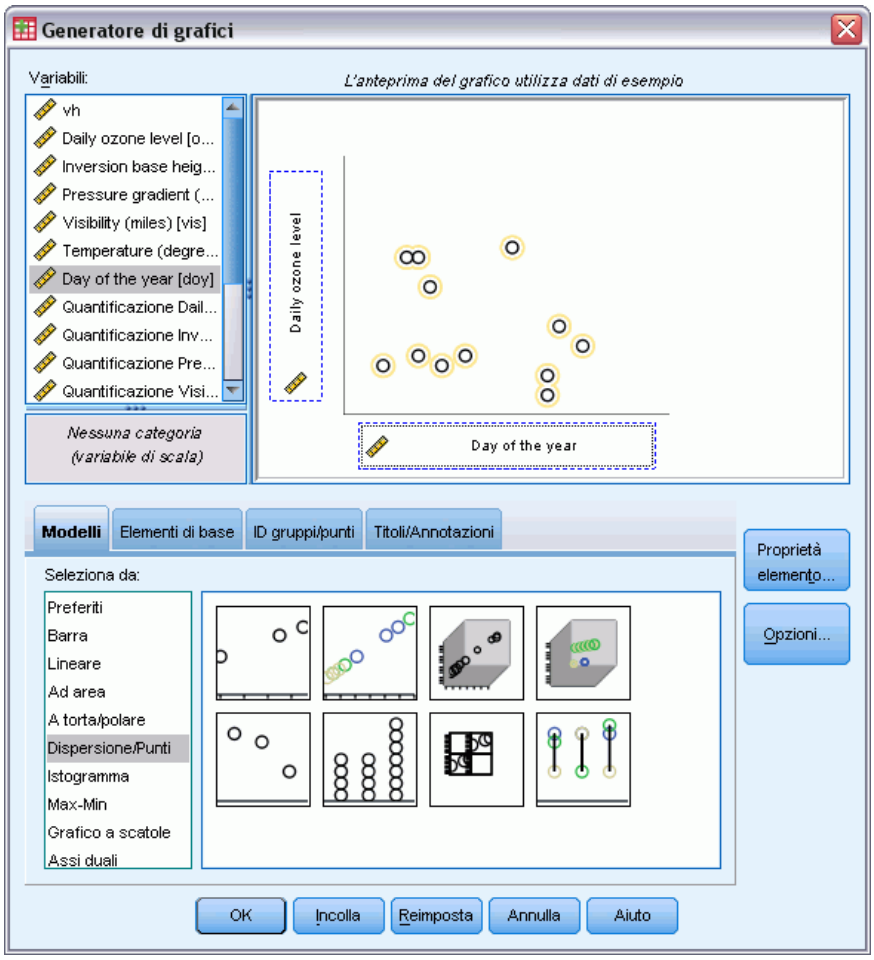
Effetti delle trasformazioni

La trasformazione delle variabili crea una relazione non lineare tra la risposta originale e l'insieme originale di predittori lineare per le variabili trasformate. Tuttavia, quando sono presenti più predittori, le altre variabili nel modello creano confusione circa le relazioni pairwise.

Per concentrare l'analisi sulla relazione tra *Livello giornaliero di ozono* e *Giorno dell'anno*, si esamini un grafico a dispersione. Dai menu, scegliere:

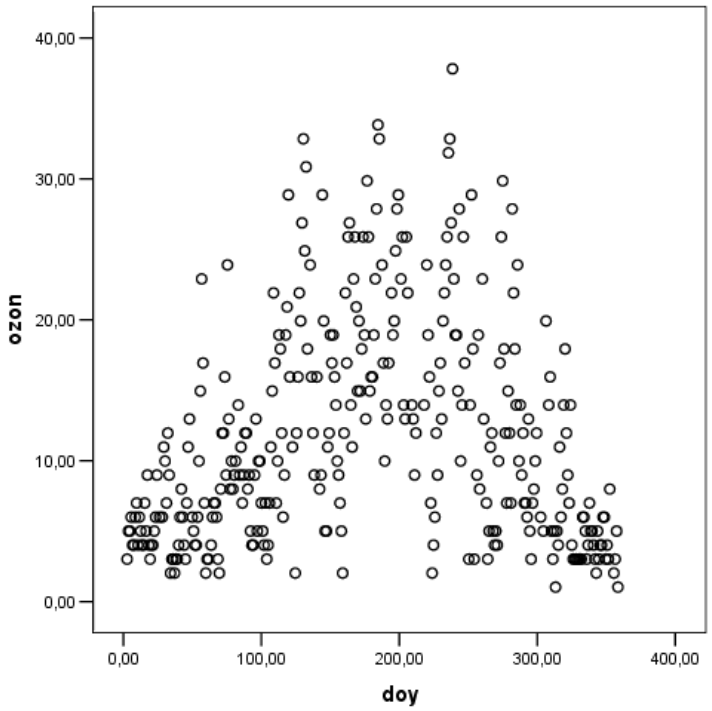
Grafici > Generatore di grafici...

Figura 9-47
Finestra di dialogo Generatore di grafici



- Selezionare il modello Dispersione/Punti e scegliere Dispersione semplice.
- Selezionare *Livello giornaliero di ozono* come variabile dell'asse *y* e *Giorno dell'anno* come variabile dell'asse *x*.
- Fare clic su OK.

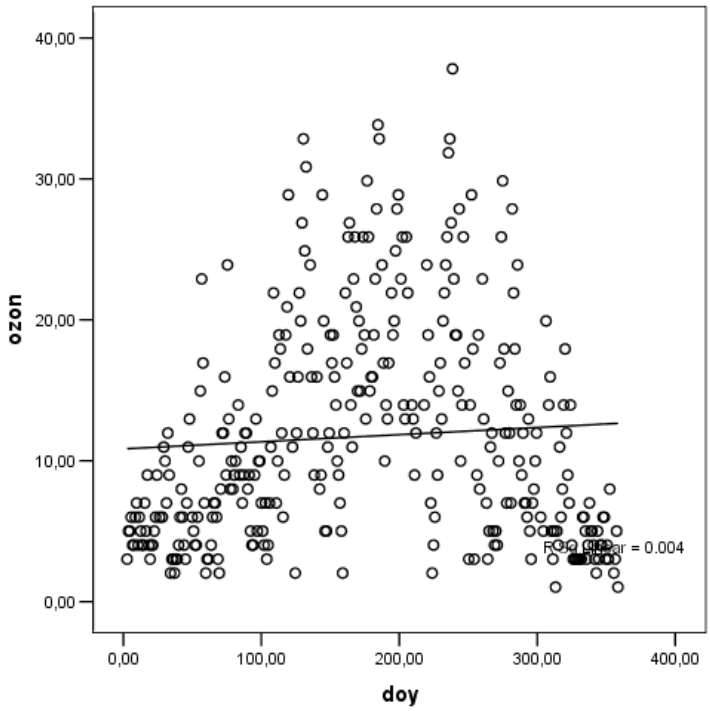
Figura 9-48
Grafico a dispersione per Livello giornaliero di ozono e Giorno dell'anno



La figura mostra la relazione tra *Livello giornaliero di ozono* e *Giorno dell'anno*. Se *Giorno dell'anno* aumenta all'incirca fino a 200, *Livello giornaliero di ozono* aumenta anch'esso. Tuttavia, per valori maggiori di 200 di *Giorno dell'anno*, *Livello giornaliero di ozono* si riduce. Questo modello a U invertito suggerisce una relazione quadratica tra le due variabili. Una regressione lineare non può rendere questa relazione.

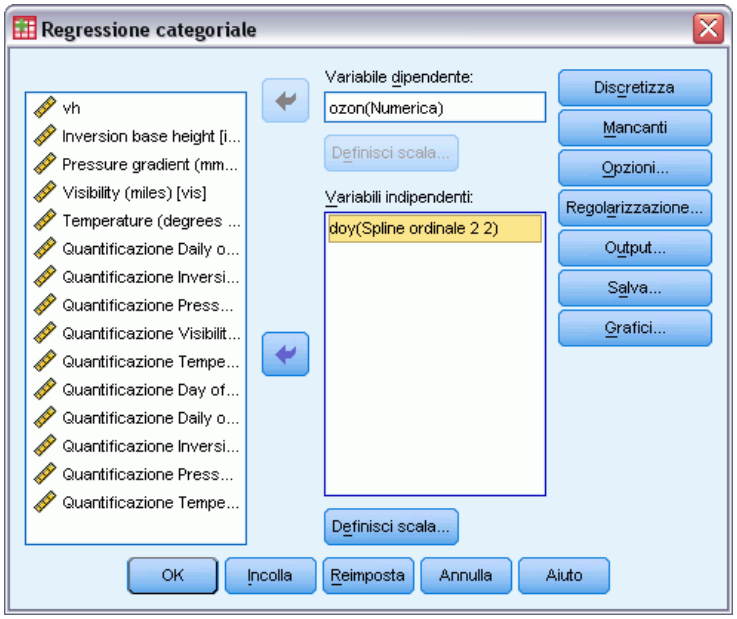
- Per visualizzare una curva di adattamento ottimale tracciata su punti del grafico a dispersione, attivare il grafico facendo doppio clic su di esso.
- Selezionare un punto nell'Editor dei dati.
- Fare clic sullo strumento Aggiungi curva di adattamento a totale e chiudere l'Editor dei grafici.

Figura 9-49
Grafico a dispersione con curva di adattamento ottimale



Una regressione lineare di *Livello giornaliero di ozono* su *Giorno dell'anno* genera un R^2 pari a 0,004. Questo adattamento suggerisce che *Giorno dell'anno* non abbia valore predittivo per *Livello giornaliero di ozono*. Questo non sorprende, dato il modello in figura. Utilizzando lo scaling ottimale, tuttavia, è possibile linearizzare la relazione quadratica e utilizzare *Giorno dell'anno* trasformato per prevedere la risposta.

Figura 9-50
Finestra di dialogo Regressione categoriale



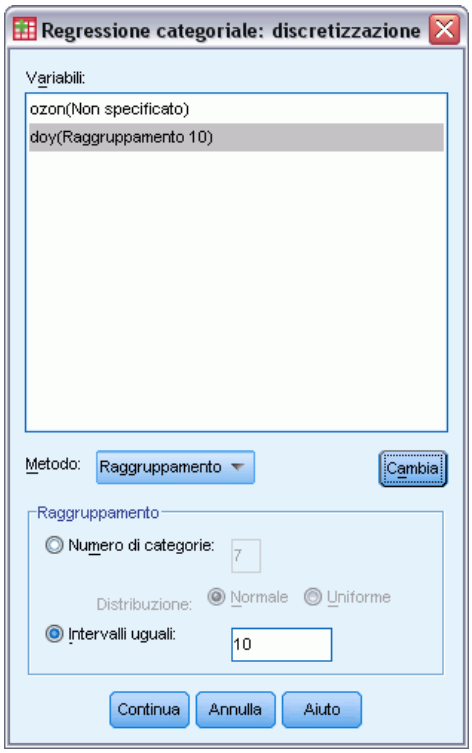
Per ottener una regressione categoriale di *Livello giornaliero di ozono* su *Giorno dell'anno*, richiamare la finestra di dialogo Regressione categoriale.

- Deselezionare da *Altezza di base inversione* a *Temperatura (gradi F)* come variabili indipendenti.
- Selezionare *Giorno dell'anno* come variabile indipendente.
- Fare clic su Definisci scala.

Figura 9-51
Finestra di dialogo Definisci scala

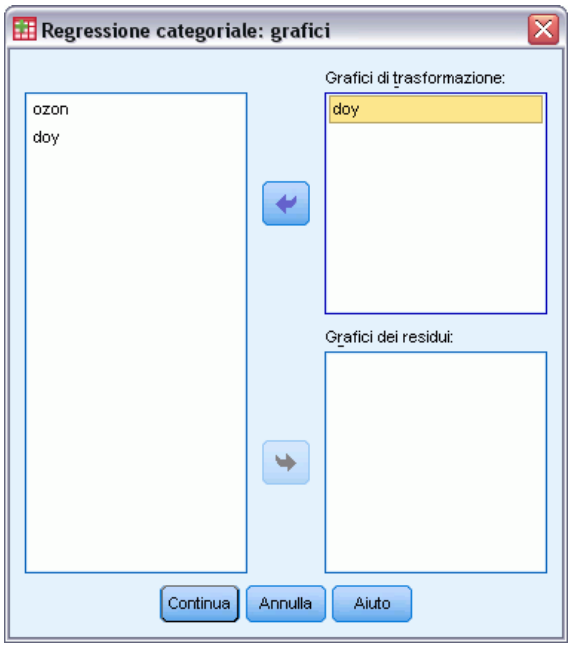
- Selezionare Nominale come livello di scaling ottimale.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale scegliere Discretizza.

Figura 9-52
Finestra di dialogo Discretizza



- Selezionare *gda*.
- Selezionare Intervalli uguali.
- Digitare 10 come lunghezza dell'intervallo.
- Fare clic su Cambia.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale fare clic su Grafici.

Figura 9-53
Finestra di dialogo Grafici



- Selezionare *gda* per i grafici di trasformazione.
- Fare clic su Continua.
- Nella finestra di dialogo Regressione categoriale scegliere OK.

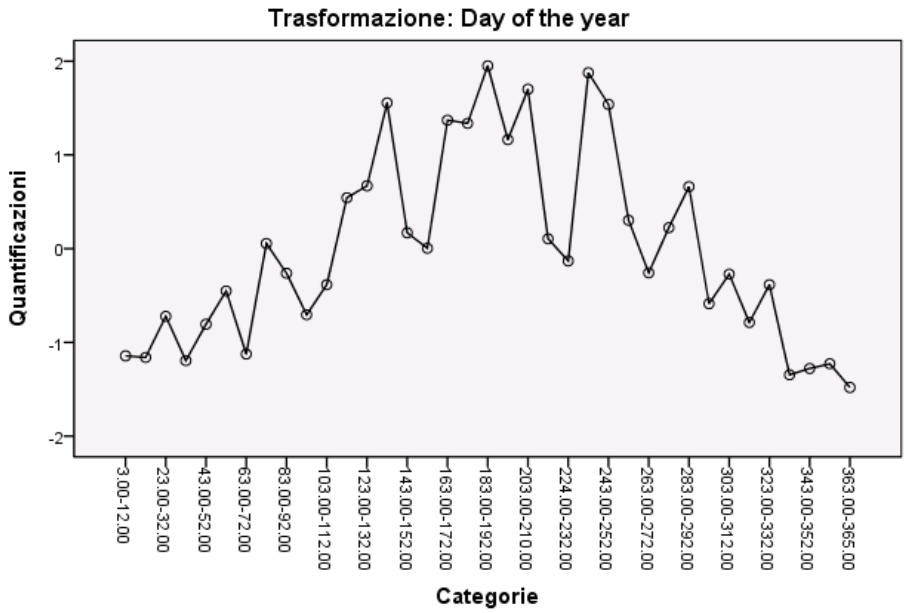
Figura 9-54
Riepilogo del modello per regressione categoriale di Livello giornaliero di ozono su Giorno dell'anno.

	R multiplo	R-quadrato	R-quadrato corretto	Errore di previsione apparente
Dati standardizzati	.741	.549	.494	.451

Variabile dipendente: Daily ozone level
Predittore: Day of the year

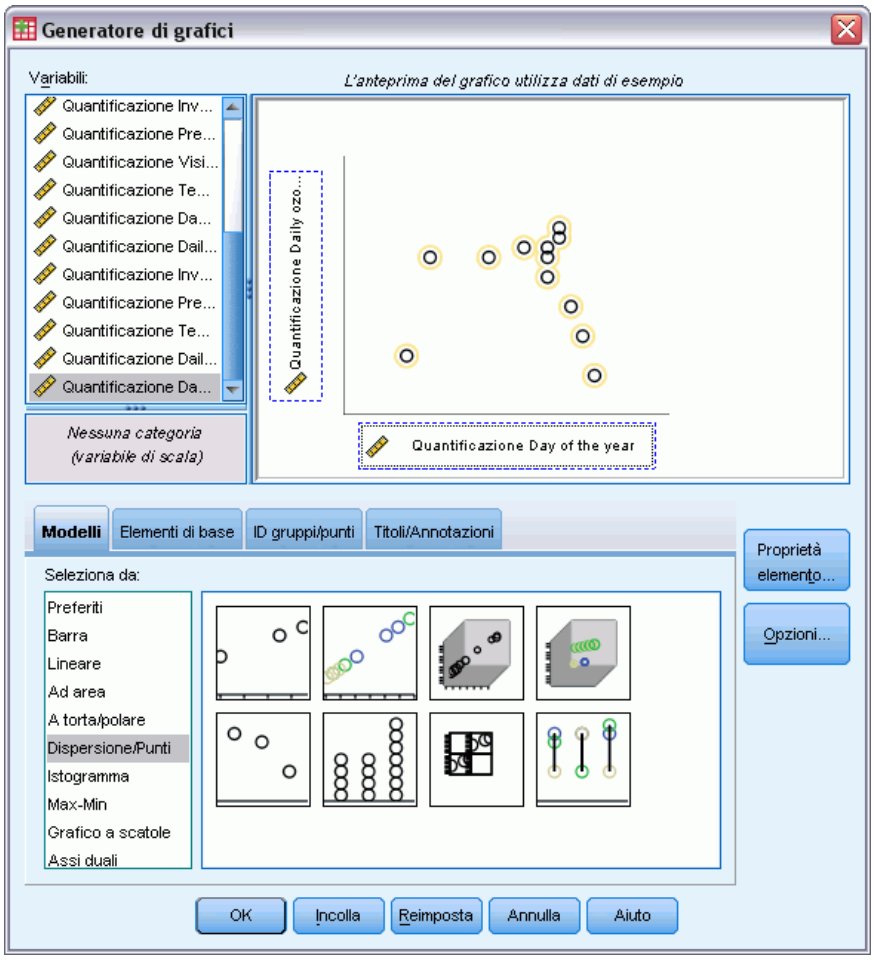
La regressione con scaling ottimale tratta *Livello giornaliero di ozono* come numerico e *Giorno dell'anno* come nominale. Questo determina un R^2 pari a 0,549. Sebbene solo il 55% della variazione di *Livello giornaliero di ozono* sia spiegata dalla regressione categoriale, si tratta di un miglioramento significativo rispetto alla regressione originale. La trasformazione di *Giorno dell'anno* consente la previsione di *Livello giornaliero di ozono*.

Figura 9-55
Grafico di trasformazione per Giorno dell'anno (nominale)



La figura mostra il grafico di trasformazione per *Giorno dell'anno*. Gli estremi di *Giorno dell'anno* ricevono entrambi quantificazioni negative, mentre i valori centrali hanno quantificazioni positive. Applicando questa trasformazione, i valori inferiore e superiore di *Giorno dell'anno* hanno effetti simili sul *Livello giornaliero di ozono* previsto.

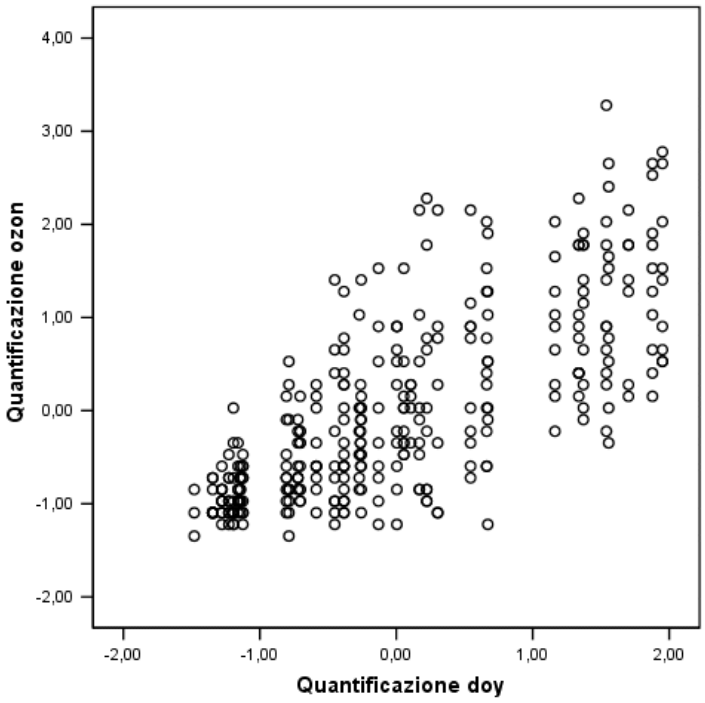
Figura 9-56
Generatore di grafici



Per vedere un grafico a dispersione delle variabili trasformate, richiamare Generatore di grafici e fare clic su Ripristina per annullare le selezioni precedenti.

- Selezionare il modello Dispersione/Punti e scegliere Dispersione semplice.
- Selezionare *Quantificazione Livello giornaliero di ozono [TRA1_3* come variabile dell'asse *y* e *Quantificazione Giorno dell'anno [TRA2_3]* come variabile dell'asse *x*.
- Fare clic su OK.

Figura 9-57
Grafico a dispersione delle variabili trasformate.



La figura mostra la relazione tra le variabili trasformate. Un trend crescente sostituisce la U invertita. La linea di regressione ha una pendenza positiva, a indicare che l'aumento di *Giorno dell'anno* trasformato corrisponde all'aumento di *Livello giornaliero di ozono* atteso. Utilizzando lo scaling ottimale si linearizza la relazione rendendo possibili interpretazioni che diversamente passerebbero inosservate.

Letture consigliate

Consultare i testi seguenti per ulteriori informazioni sulla regressione categoriale:

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, 18, .

Hastie, T., R. Tibshirani, e A. Buja. 1994. Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, .

Hayashi, C. 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statitical Mathematics*, 2, .

Kruskal, J. B. 1965. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B*, 27, .

Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, 4, .

Ramsay, J. O. 1989. Monotone regression splines in action. *Statistical Science*, 4, .

Van der Kooij, A. J., e J. J. Meulman. 1997. MURALS: Multiple regression and optimal scaling using alternating least squares. In: *Softstat '97,* F. Faulbaum, e W. Bandilla, ed. Stuttgart: Gustav Fisher.

Winsberg, S., e J. O. Ramsay. 1980. Monotonic transformations to additivity using splines. *Biometrika*, 67, .

Winsberg, S., e J. O. Ramsay. 1983. Monotone spline transformations for dimension reduction. *Psychometrika*, 48, .

Young, F. W., J. De Leeuw, e Y. Takane. 1976. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .

Capitolo
10

Analisi Componenti principali categoriale

analisi Componenti principali categoriale

L'analisi componenti principali categoriale può essere vista come un metodo di riduzione del numero delle dimensioni. Un gruppo di variabili viene analizzato per rivelare le dimensioni principali della variazione. L'insieme di dati originale può essere quindi sostituito da un insieme nuovo e di dimensioni inferiori, con una minima perdita di informazioni. Il metodo rivela le relazioni tra le variabili, tra i casi e tra le variabili e i casi.

Il criterio utilizzato dall'analisi componenti principali categoriale per la quantificazione dei dati osservati consiste nel fatto che i punteggi degli oggetti (punteggi dei componenti) abbiano correlazioni elevate con ciascuna delle variabili quantificate. Una soluzione è valida nella misura in cui tale criterio viene soddisfatto.

Verranno illustrati due esempi di analisi componenti principali categoriale. Il primo utilizza un insieme di dati piuttosto piccolo, utile per illustrare i concetti e le interpretazioni di base associati alla procedura. Il secondo esempio esamina un'applicazione pratica.

Esempio: Esame delle interrelazioni tra sistemi sociali

L'esempio esamina l'adattamento di Guttman (Guttman, 1968) di una tabella di Bell (Bell, 1961). I dati vengono anche discussi da Lingoes (Lingoes, 1968).

Bell ha presentato una tabella per illustrare i possibili gruppi sociali. Guttman ha utilizzato un parte di tale tabella, in cui cinque variabili che descrivono elementi come l'interazione sociale, i sentimenti di appartenenza a un gruppo, la vicinanza fisica dei membri e il grado di formalità della relazione, sono state incrociate con cinque gruppi sociali teorici, compresi folla (ad esempio, le persone presenti a una partita di calcio), uditorio (ad esempio, di uno spettacolo teatrale o di una lezione universitaria), pubblico (ad esempio televisivo), calca (come una folla, ma con un'interazione molto maggiore), gruppi primari (intimi), gruppi secondari (volontari) e la comunità moderna (unione non stretta derivante da una vicinanza fisica elevata e dall'esigenza di servizi specializzati).

La seguente tabella mostra le variabili nell'insieme di dati derivante dalla classificazione in sette gruppi sociali utilizzata nei dati di Guttman-Bell, con le etichette delle variabili e dei valori (categorie) associate ai livelli di ciascuna variabile. Questo insiemi di dati è reperibile nel file *guttman.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*. Oltre a selezionare le variabili da includere nel calcolo dell'analisi componenti principali categoriale, è possibile selezionare variabili utilizzate per assegnare

etichette agli oggetti nei grafici. Nell'esempio, le prime cinque variabili nei dati sono incluse nell'analisi, mentre il cluster è utilizzato esclusivamente come variabile di etichetta. Quando si specifica un'analisi componenti principali categoriale, è necessario specificare il livello di scaling ottimale per ogni variabile dell'analisi. Nell'esempio, è specificato un livello ordinale per tutte le variabili dell'analisi.

Tabella 10-1
Variabili nell'insieme di dati di Guttman-Bell

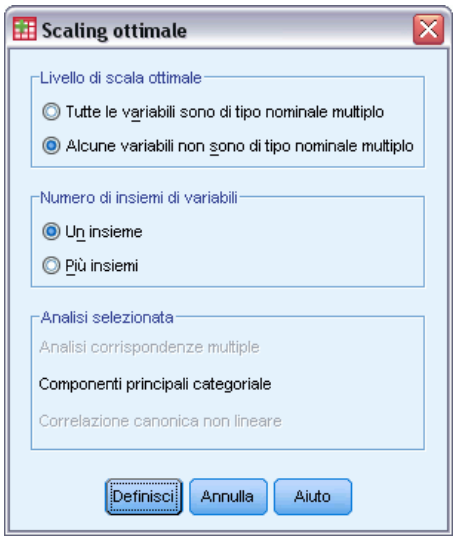
Nome di variabile	Etichetta di valore	Etichetta del valore
intensità	intensità dell'interazione	Leggera, bassa, moderata, alta
frequenza	Frequenza dell'interazione	Leggera, non ricorrente, non frequente, frequente
appartenenza	Sentimento di appartenenza	Nessuna, leggera, variabile, alta
vicinanza	Vicinanza fisica.	Limitata, elevata
formalità	Formalità della relazione	Nessuna relazione, formale, informale
cluster		Folla, spettatori, uditorio, pubblico, calca, gruppi primari, gruppi secondari, comunità moderna

Esecuzione dell'analisi

- Per generare il risultato dei componenti principali categoriale per questo insieme di dati, dai menu scegliere:

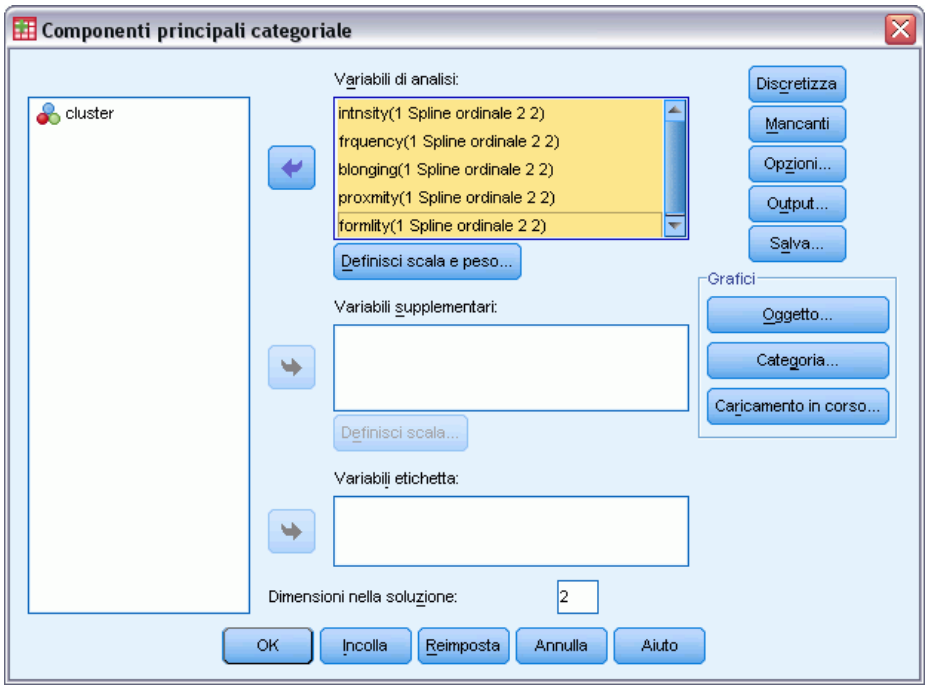
  Analizza > Riduzioni dimensione > Scaling ottimale...

Figura 10-1
Finestra di dialogo Scaling ottimale



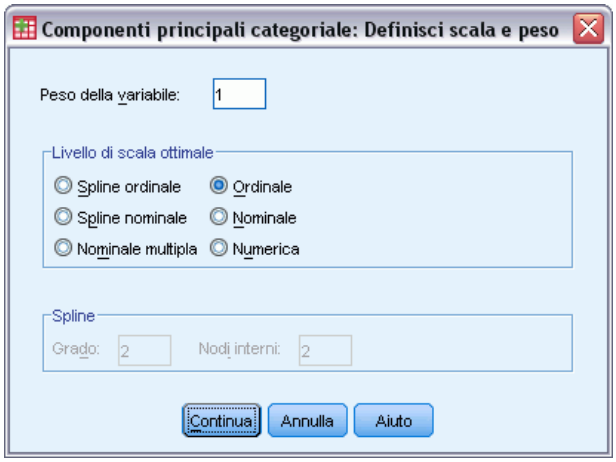
- Selezionare Una o più variabili non nominali multiple nel gruppo Livello di scaling ottimale.
- Fare clic su Definisci.

Figura 10-2
Finestra di dialogo Componenti principali categoriale



- Selezionare da *Intensità dell'interazione* a *Formalità della relazione* come variabili di analisi.
- Fare clic su Definisci scala e peso.

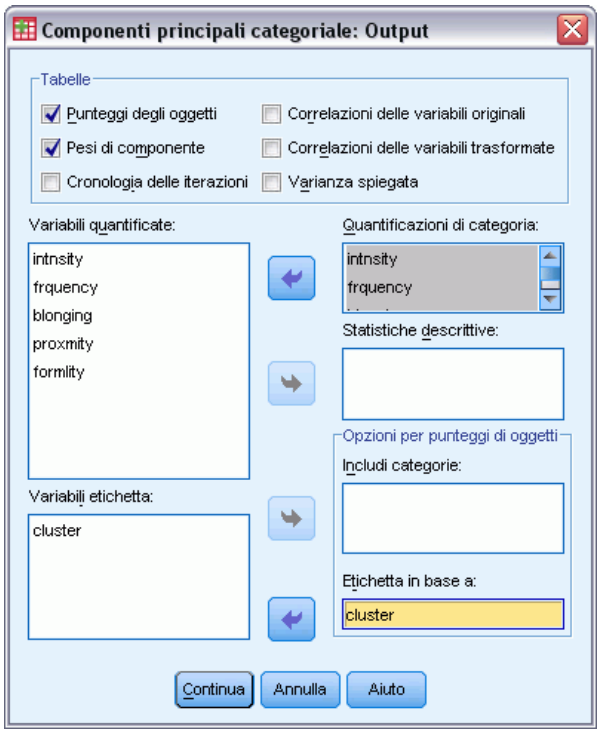
Figura 10-3
Finestra di dialogo Definisci scala e peso



- Selezionare Ordinale nel gruppo Livello di scaling ottimale.
- Fare clic su Continua.
- Selezionare *cluster* come variabile di etichetta nella finestra di dialogo Componenti principali categoriale.

▶ Fare clic su Output.

Figura 10-4
Finestra di dialogo Output



▶ Selezionare Punteggi degli oggetti e deselezionare Correlazioni delle variabili trasformate nel gruppo Tabelle.

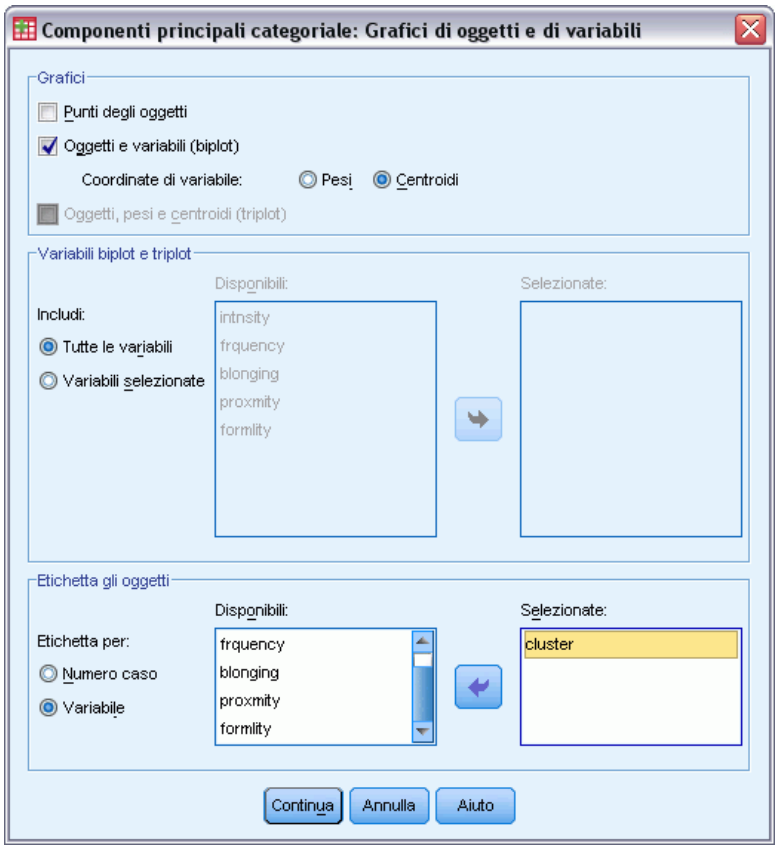
▶ Scegliere di generare quantificazioni di categoria per le variabili da *intensità (Intensità dell'interazione)* a *formalità (Formalità della relazione)*.

▶ Scegliere di etichettare i punteggi degli oggetti in base a *cluster*.

▶ Fare clic su Continua.

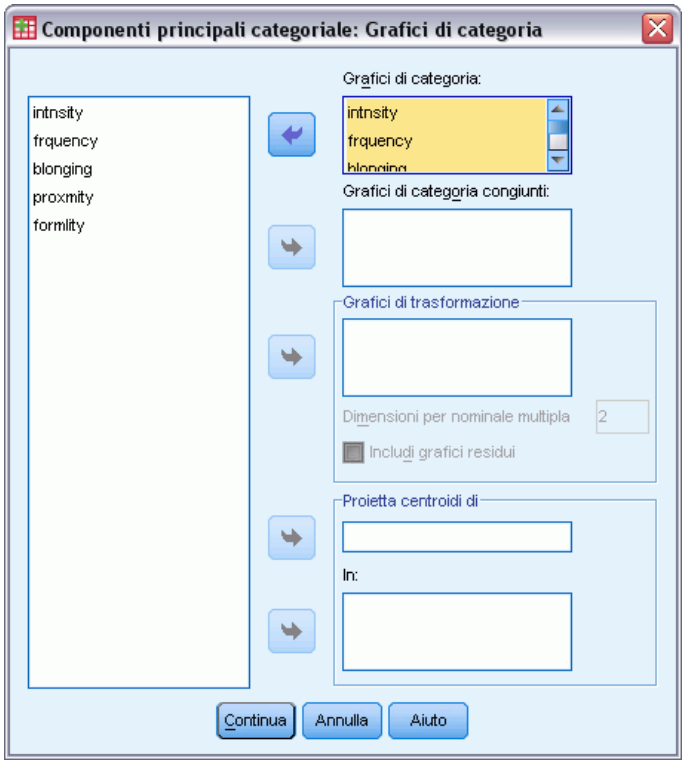
▶ Fare clic su Oggetto nel gruppo Grafici della finestra di dialogo Analisi componenti principali categoriale.

Figura 10-5
Finestra di dialogo Grafici di oggetti e di variabili



- Selezionare Oggetti e variabili (biplot) nel gruppo Grafici.
- Scegliere di etichettare gli oggetti in base a Variabile nel gruppo Etichetta gli oggetti, quindi selezionare *cluster* come variabile in base alla quale etichettare gli oggetti.
- Fare clic su Continua.
- Fare clic su Categoria nel gruppo Grafici della finestra di dialogo Analisi componenti principali categoriale.

Figura 10-6
Finestra di dialogo Grafici di categoria



- Scegliere di generare grafici di categoria congiunti per le variabili da *intensità (Intensità dell'interazione)* a *formalità (Formalità della relazione)*.
- Fare clic su Continua.
- Fare clic su OK nella finestra di dialogo Analisi componenti principali categoriale.

Numero di dimensioni

Questi dati mostrano parte dell'output iniziale dell'analisi componenti principali categoriale. Dopo la cronologia iterazioni dell'algoritmo, viene visualizzato il riepilogo del modello, compresi gli autovalori di ciascuna dimensione. Tali autovalori sono equivalenti a quelli dell'analisi componenti principali classica. Rappresentano una misura della quantità di varianza spiegata per ogni dimensione.

Figura 10-7
Cronologia delle iterazioni

	Varianza spiegata		Perdita		
Numero di iterazione	Totale	Aumento	Totale	Coordinate del centroide	Restrizione coordinate centroide rispetto a coordinate vettore
0	4,515315	,000000	5,484685	4,075583	1,409101
31[a]	4,726009	,000008	5,273991	4,273795	1,000196

a. Processo di iterazione interrotto perché è stato raggiunto il valore di controllo per la convergenza.

Figura 10-8
Riepilogo modello

		Varianza spiegata	
Dimensione	Alfa di Cronbach	Totale (autovalore)	% di varianza
1	,881	3,389	67,774
2	,315	1,337	26,746
Totale	,986[a]	4,726	94,520

a. Il totale di Alfa di Cronbach è basato sul totale dell'autovalore.

Gli autovalori possono essere utilizzati come indicazione del numero di dimensioni necessarie. Nell'esempio, è stato utilizzato il numero predefinito di dimensioni (2). Si tratta del numero corretto? Come regola generale, quando tutte le variabili sono nominali singole, ordinali o numeriche, l'autovalore per una dimensione deve essere maggiore di 1. Poiché la soluzione a due dimensioni spiega il 94,52% della varianza, una terza dimensione probabilmente non aggiungerebbe molte informazioni.

Per variabili nominali multiple, non esiste una regola semplice per determinare il numero adeguato di dimensioni. Se il numero delle variabili è sostituito dal numero totale di categorie meno il numero di variabili, la regola sopra illustrata resta applicabile. Tale regola da sola, tuttavia, consentirebbe probabilmente più dimensioni del necessario. Quando si sceglie il numero di dimensioni, la regola pratica più utile è mantenerlo sufficientemente basso da consentire interpretazioni significative. La tabella di riepilogo del modello mostra inoltre l'Alfa di Cronbach (una misura di affidabilità) massimizzata dalla procedura.

Quantificazioni

Per ciascuna variabile vengono presentate le quantificazioni, le coordinate del vettore e del centroide per ogni dimensione. Le quantificazioni sono i valori assegnati a ciascuna categoria. Le coordinate del centroide sono la media dei punteggi degli oggetti per gli oggetti della stessa categoria. Le coordinate del vettore sono le coordinate delle categorie quando è necessario che siano presenti su una linea, a rappresentazione della variabile nello spazio. Questo è necessario per le variabili con livello di scaling numerico e ordinale.

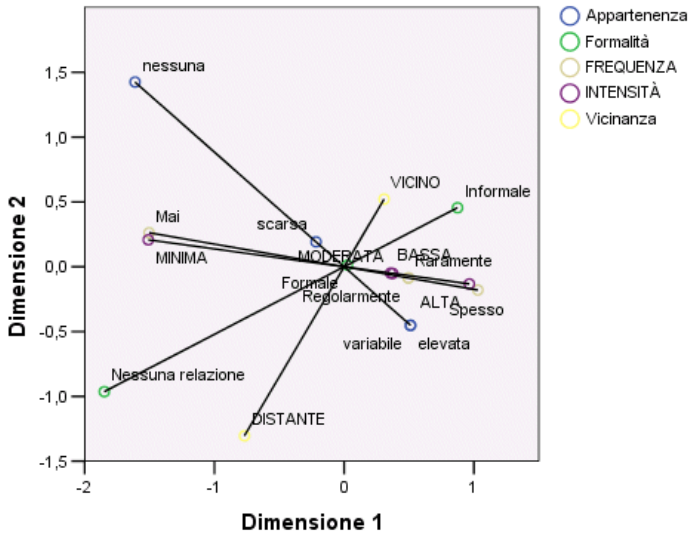
Figura 10-9
Quantificazioni dell'intensità dell'interazione

Categoria	Frequenza	Quantificazione	Coordinate del centroide		Coordinate del vettore	
			Dimensione		Dimensione	
			1	2	1	2
MINIMA	2	-1,530	-1,496	,308	-1,510	,208
BASSA	2	,362	,392	,202	,358	-,049
MODERATA	1	,379	,188	-1,408	,374	-,051
ALTA	2	,978	1,010	,194	,965	-,133

Normalizzazione principale per variabile.

Considerando le quantificazioni nel grafico congiunto dei punti delle categorie, è possibile vedere che alcune delle categorie di alcune variabili non sono state distinte dall'analisi componenti principali categoriale tanto nettamente quanto previsto se il livello fosse stato completamente ordinale. Le variabili *Intensità dell'interazione* e *Frequenza dell'interazione*, ad esempio, hanno quantificazioni pari o quasi pari per le due relative categorie centrali. Questo tipo di risultati può suggerire di tentare analisi componenti principali categoriali alternative, eventualmente con alcune categorie compresse o con un diverso livello di analisi, ad esempio nominale (multiplo).

Figura 10-10
Punti delle categorie dei grafici congiunti



Il grafico congiunto dei punti delle categorie è analogo al grafico per i pesi di componente, ma mostra in aggiunta la posizione dei punti finali corrispondenti alle quantificazioni più basse (ad esempio, *leggera* per *Intensità dell'interazione* e *nessuna* per *Sentimento di appartenenza*). Le due variabili che misurano l'interazione, *Intensità dell'interazione* e *Frequenza dell'interazione*, vengono visualizzate una accanto all'altra e spiegano gran parte della varianza nella dimensione 1. *Formalità della relazione* appare anch'essa accanto a *Vicinanza fisica*.

Concentrando l'attenzione sui punti delle categorie, è possibile vedere i rapporti in modo ancora più chiaro. Non solo *Intensità dell'interazione* e *Frequenza dell'interazione* sono vicine, ma le direzioni delle relative scale sono simili, ovvero intensità leggera e frequenza scarsa sono analoghe; ovvero, intensità leggera e frequenza scarsa sono vicine e interazione frequente e intensità di interazione elevata sono vicine. È possibile inoltre vedere che la vicinanza fisica elevata sembra andare di pari passo con un tipo di relazione informale e che la distanza fisica è correlata all'assenza di relazione.

Punteggi oggetto

Si può inoltre richiedere un elenco e il grafico dei punteggi degli oggetti. Il grafico dei punteggi degli oggetti può essere utile per rilevare valori anomali, gruppi tipici di oggetti o per evidenziare alcuni modelli speciali.

La tabella dei punteggi degli oggetti mostra l'elenco dei punteggi etichettati per gruppo sociale per i dati di Guttman-Bell. Esaminando i valori per i punti degli oggetti, è possibile identificare oggetti specifici all'interno del grafico.

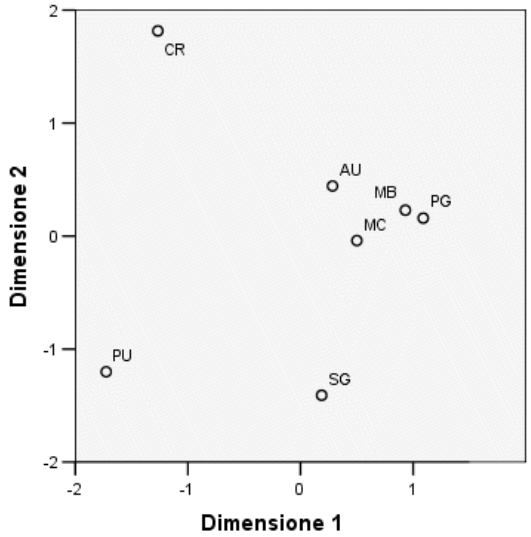
Figura 10-11
Punteggi degli oggetti

	Dimensione	
Gruppo	1	2
CR	-1,266	1,816
AU	,284	,444
PU	-1,726	-1,201
MB	,931	,229
PG	1,089	,159
SG	,188	-1,408
MC	,500	-,039

Normalizzazione principale per variabile.

La prima dimensione sembra dividere *FOLLA* e *UDITORIO*, che hanno punteggi negativi relativamente elevati, da *PUBBLICO* e *GRUPPI PRIMARI*, che hanno punteggi positivi relativamente elevati. La seconda dimensione include tre gruppi: *UDITORIO* e *GRUPPI SECONDARI* con valori negativi elevati, *FOLLA* con valori positivi elevati e gli altri gruppi sociali compresi tra di essi. Esaminando il grafico dei punteggi degli oggetti questo risulta più evidente.

Figura 10-12
Grafico dei punteggi degli oggetti

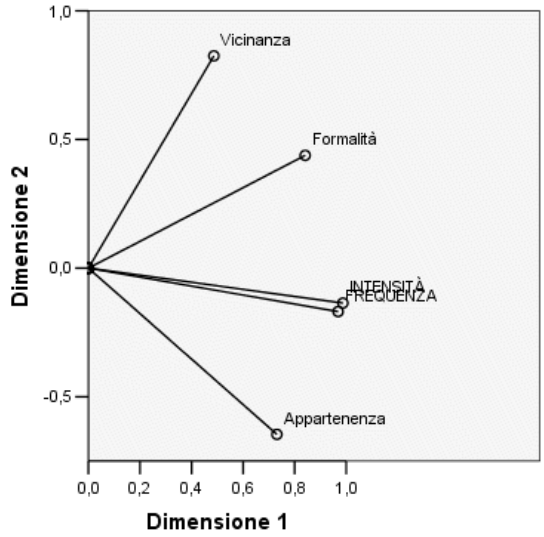


Nel grafico, è possibile vedere *UDITORIO* e *GRUPPI SECONDARI* nella parte inferiore, *FOLLA* nella parte superiore e gli altri gruppi sociali nel mezzo. L'esame dei modelli tra i singoli oggetti dipende dalle informazioni aggiuntive disponibili per le unità di analisi. In questo caso, è nota la classificazione degli oggetti. In altri casi, è possibile utilizzare variabili supplementari per etichettare gli oggetti. È inoltre possibile vedere che l'analisi componenti principali categoriale non divide *PUBBLICO* da *GRUPPI PRIMARI*. Sebbene la maggioranza delle persone normalmente non ' pensi alla propria famiglia come a una calca di persone, nelle variabili utilizzate i due gruppi hanno ricevuti lo stesso punteggio per quattro variabili su cinque! Normalmente si desidera esaminare possibili difetti delle variabili e delle categorie utilizzate. Ad esempio, un'elevata intensità dell'interazione e relazioni informali probabilmente indicano cose diverse per questi due gruppi. In alternativa, è possibile prendere in considerazione una soluzione con un maggiore numero di dimensioni.

Pesi di componente

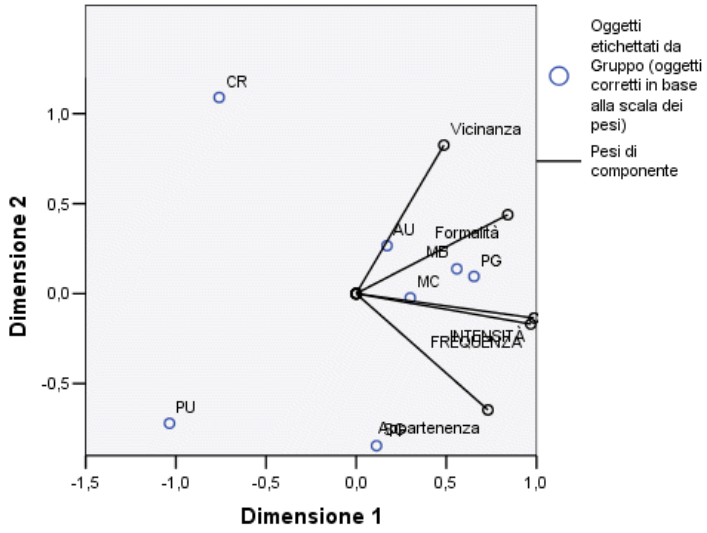
La figura mostra il grafico dei pesi di componente. I vettori (linee) sono relativamente lunghi, a indicare di nuovo che le prime due dimensioni spiegano la maggior parte della varianza di tutte le variabili quantificate. Nella prima dimensione, tutte le variabili hanno pesi di componenti elevati (positivi). La seconda dimensione è correlata principalmente alle variabili quantificate *Sentimento di appartenenza* e *Vicinanza fisica*, in direzioni opposte. Questo significa che gli oggetti con punteggio negativo elevato nella dimensione 2 avranno un punteggio elevato per sentimento di appartenenza e ridotto per vicinanza fisica. La seconda dimensione, quindi, rivela un contrasto tra queste due variabili, con al contempo una limitata relazione con le variabili quantificate *Intensità dell'interazione* e *Frequenza dell'interazione*.

Figura 10-13
Pesi di componente



Per esaminare la relazione tra gli oggetti e le variabili, si consideri il biplot di oggetti e di pesi di componente. Il vettore di una variabile è orientato nella direzione della categoria massima della variabile. Ad esempio, per *Vicinanza fisica* e *Sentimento di appartenenza* le categorie massime sono rispettivamente *elevata* e *alta*. Di conseguenza, il gruppo *FOLLA* è caratterizzato da un'elevata vicinanza fisica e dall'assenza di sentimento di appartenenza; *GRUPPI SECONDARI* da una ridotta vicinanza fisica e da un elevato sentimento di appartenenza.

Figura 10-14
Biplot



Dimensioni aggiuntive

L'aumento del numero delle dimensioni aumenta la quantità di variazioni considerate e può rivelare differenze non evidenti nelle soluzioni con un numero di dimensioni minore. Come notato in precedenza, in presenza di due dimensioni *PUBBLICO* e *GRUPPI PRIMARI* non possono essere separati. Tuttavia, l'aumento della dimensionalità può consentire una differenziazione tra i due gruppi.

Esecuzione dell'analisi

- Per ottenere una soluzione a tre dimensioni, richiamare la finestra di dialogo Componenti principali categoriale.
- Digitare 3 come numero di dimensioni per la soluzione.
- Fare clic su OK nella finestra di dialogo Analisi componenti principali categoriale.

Riepilogo del modello

Figura 10-15
Riepilogo modello

Dimensione	Alfa di Cronbach	Varianza spiegata	
		Totale (autovalore)	% di varianza
1	,885	3,424	68,480
2	-,232	,844	16,871
3	-,459	,732	14,649
Totale	1,000[a]	5,000	99,999

a. Il totale di Alfa di Cronbach è basato sul totale dell'autovalore.

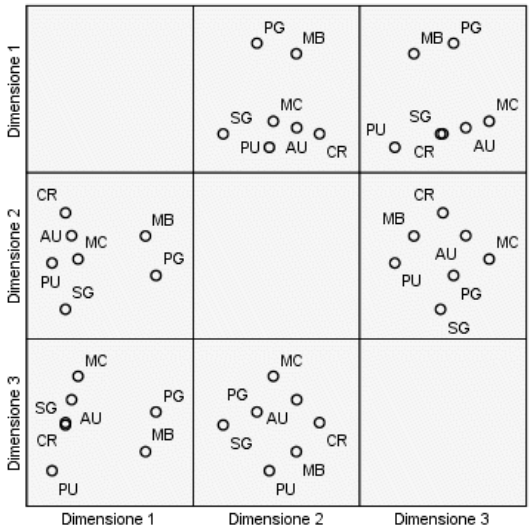
Una soluzione a tre dimensioni ha autovalori pari a 3,424, 0,844 e 0,732, che spiegano la quasi totalità della varianza.

Punteggi oggetto

I punteggi degli oggetti per la soluzione a tre dimensioni sono tracciati in una matrice di grafici a dispersione, nella quale ogni dimensione viene tracciata rispetto alle altre in una serie di grafici a dispersione a due dimensioni. Si noti che i primi due autovalori in presenza di tre dimensioni non sono uguali agli autovalori nella soluzione a tre; in altre parole, le soluzioni non sono nidificate. Poiché gli autovalori nelle dimensioni 2 e 3 sono ora inferiori rispetto alla dimensione 1 (con Alfa

di Cronbach negativa), è consigliabile optare per la soluzione a due dimensioni. La soluzione a tre dimensioni viene inclusa a scopo illustrativo:

Figura 10-16
Matrice di grafici a dispersione dei punteggi degli oggetti a tre dimensioni



La riga superiore dei grafici rivela che la prima dimensione separa *GRUPPI PRIMARI* e *PUBBLICO* dagli altri gruppi. Si noti che l'ordine degli oggetti lungo l'asse verticale non cambia in alcuno dei grafici della riga superiore; ciascuno di tali grafici utilizza la dimensione 1 come asse *y*.

La riga centrale dei grafici consente di interpretare la dimensione 2. La seconda dimensione si è leggermente modificata rispetto alla soluzione a due dimensioni. In precedenza, la seconda dimensione includeva tre gruppi distinti; ora gli oggetti sono maggiormente distribuiti lungo l'asse.

La terza dimensione consente di separate *PUBBLICO* da *GRUPPI PRIMARI*, il che non avviene nella soluzione a due dimensioni.

Si osservino più in dettaglio i grafici della dimensione 2 rispetto alla 3 e della dimensione 1 rispetto alla 2. Nel piano definito dalle dimensioni 2 e 3, gli oggetti formano un rettangolo, con *FOLLA*, *COMUNITÀ MODERNA*, *GRUPPI SECONDARI* e *UDITORIO* ai vertici. Su questo piano, *PUBBLICO* e *GRUPPI PRIMARI* risultano rispettivamente combinazioni convesse di *UDITORIO-FOLLA* e *GRUPPI SECONDARI-COMUNITÀ MODERNA*. Tuttavia, come già indicato, sono separati dagli altri gruppi nella dimensione 1. *SPETTATORI* non è separato dagli altri gruppi nella dimensione 1 e sembra rappresentare una combinazione di *FOLLA* e di *COMUNITÀ MODERNA*.

Pesi di componente

Figura 10-17
Pesi di componenti a tre dimensioni

	Dimensione		
	1	2	3
INTENSITÀ	,980	-,005	-,201
FREQUENZA	,521	-,643	,561
Appartenenza	,980	-,002	-,197
Vicinanza	,519	,656	,549
Formalità	,981	,004	-,193

Normalizzazione principale per variabile.

La modalità di separazione degli oggetti non indica però quali variabili corrispondono a quali dimensioni. Questo risultato si raggiunge utilizzando i pesi componente. La prima dimensione corrispondente essenzialmente a *Sentimento di appartenenza*, *Intensità dell'interazione* e *Formalità della relazione*; la seconda dimensione separa *Frequenza dell'interazione* e *Vicinanza fisica*; la terza separa queste ultime dalle altre.

Esempio: Sintomatologia dei disturbi dell'alimentazione

I disturbi dell'alimentazione sono malattie debilitanti associate ad anomalie nelle abitudini alimentari, a una percezione gravemente distorta del proprio corpo e a un'ossessione per il peso che influenza contemporaneamente mente e corpo. Milioni di persone ne vengono colpite ogni anno; particolarmente a rischio sono gli adolescenti. Esistono delle cure, la maggioranza delle quali particolarmente utili quando il problema viene identificato nelle prime fasi.

Un sanitario può tentare di diagnosticare un disturbo dell'alimentazione tramite valutazione medica e psicologica. Tuttavia, può essere difficile assegnare un paziente a una delle diverse classi di disturbi dell'alimentazione, in quanto non esiste una sintomatologia standardizzata del comportamento anoressico/bulimico. Esistono sintomi che differenziano chiaramente i pazienti dei quattro gruppi? Quali sintomi hanno in comune?

Per tentare di rispondere a queste domande, i ricercatori (Van der Ham, Meulman, Van Strien, e Van Engeland, 1997) hanno condotto uno studio su 55 adolescenti con disturbi alimentari noti, come illustrato nella tabella seguente.

Tabella 10-2
Diagnosi dei pazienti

Diagnosi	**Numero di pazienti**
Anoressia nervosa	25
Anoressia con bulimia nervosa	9
Bulimia nervosa post anoressia	14
Disturbo dell'alimentazione atipico	7
Totale	55

Ogni paziente è stato visitato quattro volte in quattro anni, per un totale di 220 visite. Durante ciascuna visita, ai pazienti è stato assegnato un punteggio per ognuno dei 16 sintomi indicati nella tabella seguente. I punteggi relativi ai sintomi sono assenti per il paziente 71 alla visita 2, il

paziente 76 alla visita 2 e il paziente 47 alla visita 3, con 217 osservazioni valide. Questi dati sono reperibili in *anorectic.sav*.Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*.

Tabella 10-3
Sottoscale di Morgan-Russell modificate per la misura del benessere

Nome di variabile	Etichetta di valore	Estremità inferiore (punteggio1)	Estremità superiore (punteggio 3 o 4)
peso	Peso corporeo	Esterno all'intervallo normale	Normale
ciclo	Ciclo mestruale	Amenorrea	Ciclo regolare
digiuno	Limitata ingestione di cibo (digiuno)	Minore di 1200 calorie	Pasti normali/regolari
eccessi	Eccessi alimentari	Più spesso di una volta a settimana	No
vomito	Episodi di vomito	Più spesso di una volta a settimana	No
lassativi	Uso di lassativi	Più spesso di una volta a settimana	No
iper	Iperattività	Impossibilità di stare fermo/a	No
fami	Rapporti famigliari	Insufficiente	Buona
eman	Emancipazione dalla famiglia	Elevato grado di dipendenza	Adeguato
amici	Relazioni amicali	Assenza di amici intimi	Due o più amici intimi
scuola	Risultati lavorativi/scolastici	Lavoro/scuola interrotto	Risultati da discreti a buoni
attses	Atteggiamento sessuale	Non adeguato	Adeguato
comses	Comportamento sessuale	Non adeguato	In grado di apprezzare il sesso
umore	Stato mentale (umore)	Molto depresso	Normale
preo	Preoccupazione legata a cibo e peso	Completo	Nessuna
corpo	Percezione del proprio corpo	Distorta	Normale

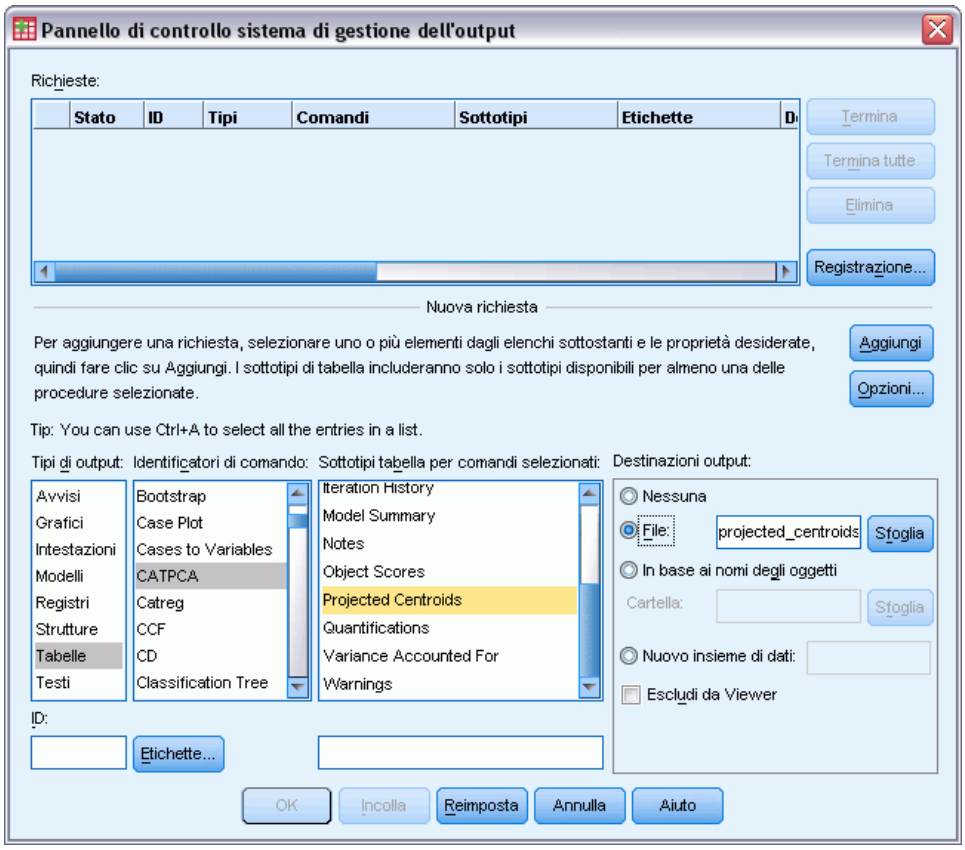
L'analisi componenti principali è ideale per questa situazione, perché lo scopo dello studio è accertare le relazioni tra i sintomi e le diverse classi di disturbi dell'alimentazione. Inoltre, l'analisi componenti principali categoriale è probabilmente più utile di quella classica, in quanto ai sintomi viene assegnato un punteggio su una scala ordinale.

Esecuzione dell'analisi

Per esaminare correttamente la struttura dello sviluppo della malattia per ogni diagnosi, sarà opportuno fare in modo che i risultati della tabella dei centroidi proiettati siano disponibili come dati per grafici a dispersione. È possibile farlo utilizzando il Sistema di gestione dell'output.

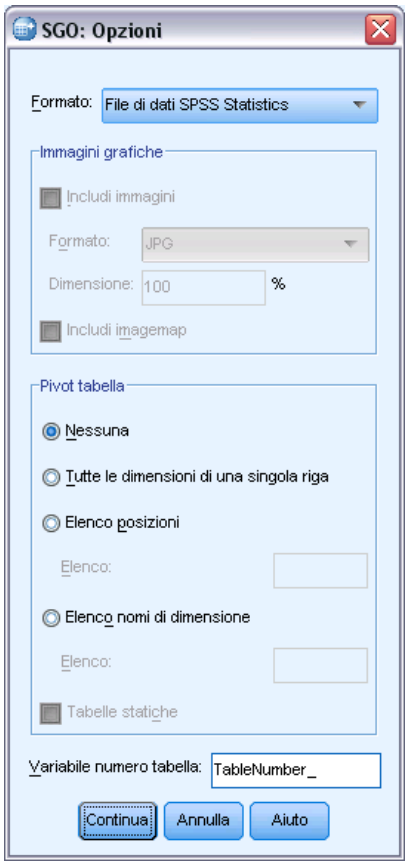
▶ Per avviare una richiesta SGO, dai menu scegliere:
Strumenti > Pannello di controllo SGO...

Figura 10-18
Pannello di controllo sistema di gestione dell'output



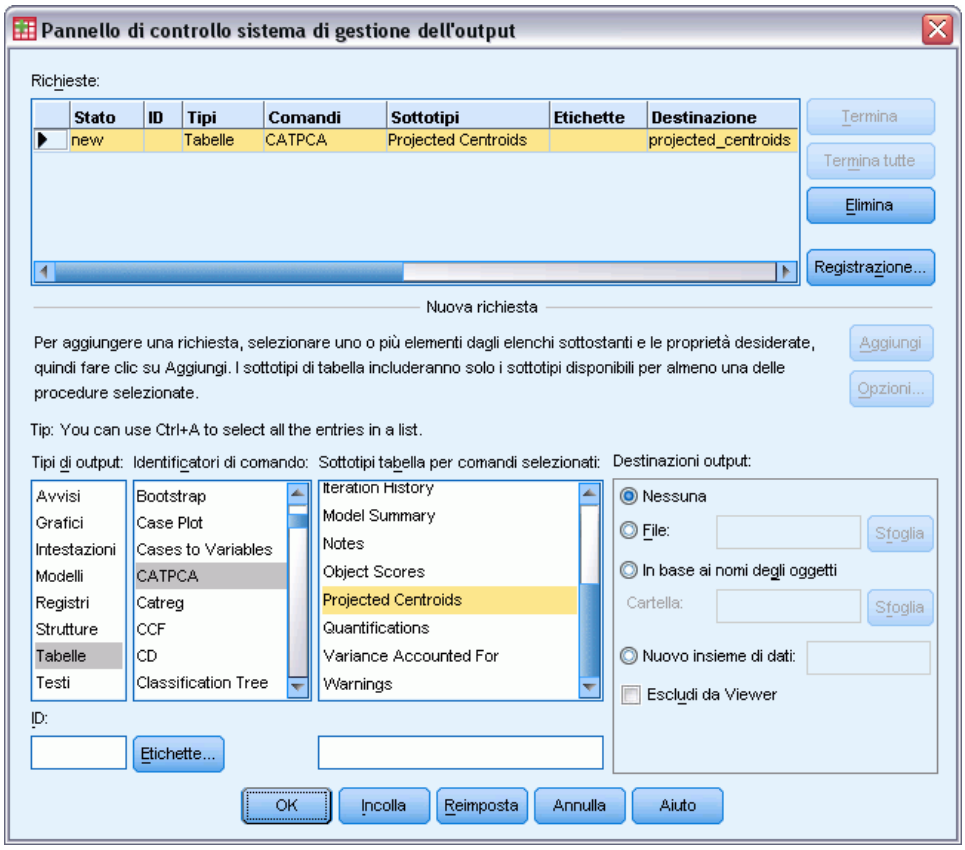
- Selezionare Tabelle come tipo di output.
- Selezionare CATPCA come comando.
- Selezionare Centroidi proiettati come tipo di tabella.
- Selezionare File nel gruppo Destinazioni output e digitare projected_centroids.sav come nome del file.
- Fare clic su Opzioni.

Figura 10-19
Finestra di dialogo Opzioni



- Selezionare File di dati IBM® SPSS® Statistics come formato dell'output.
- Digitare NumeroTabella_1 come variabile di numero di tabella.
- Fare clic su Continua.

Figura 10-20
Pannello di controllo sistema di gestione dell'output

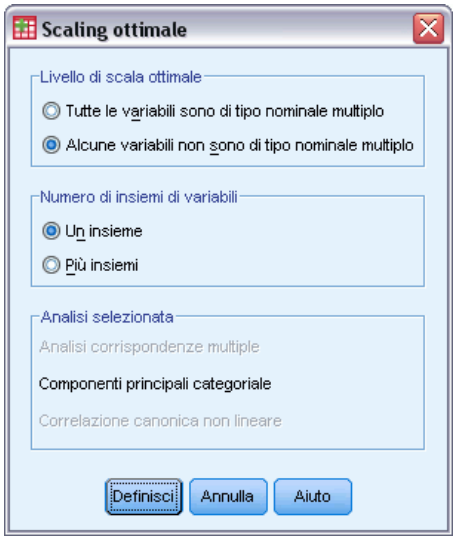


- Fare clic su Aggiungi.
- Fare clic su OK e quindi di nuovo su OK per confermare la sessione SGO.

  Il Sistema di gestione dell"output è ora impostato per scrivere i risultati della tabella dei centroidi proiettati nel file *projected_centroids.sav*.
- Per generare il risultato dei componenti principali categoriale per questo insieme di dati, dai menu scegliere:

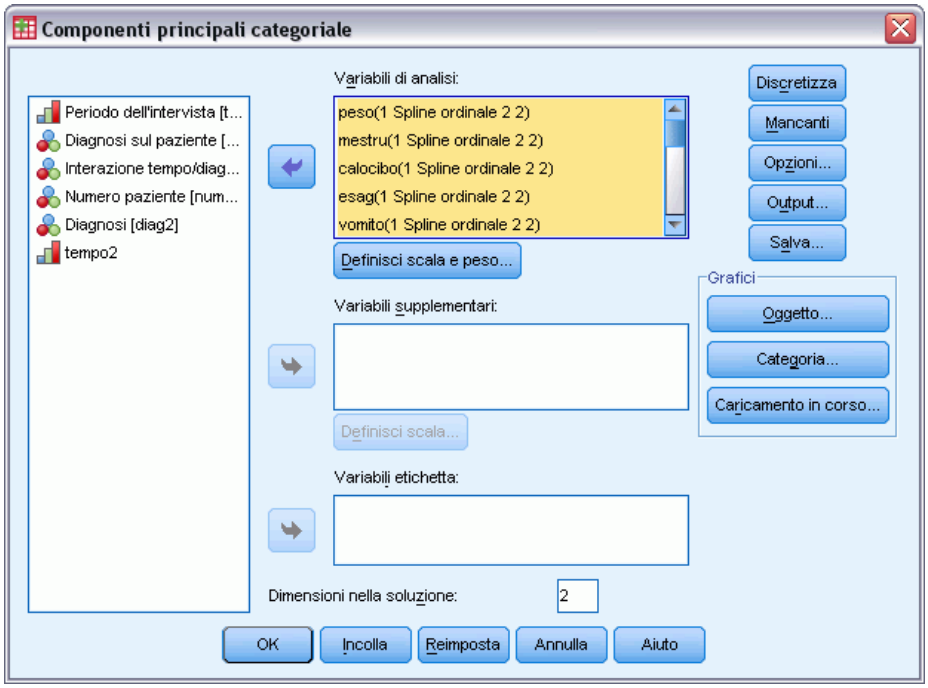
  Analizza > Riduzioni dimensione > Scaling ottimale...

Figura 10-21
Finestra di dialogo Scaling ottimale



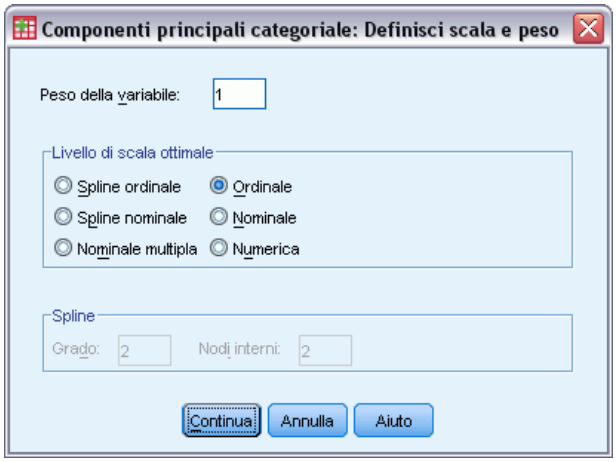
- Selezionare Una o più variabili non nominali multiple nel gruppo Livello di scaling ottimale.
- Fare clic su Definisci.

Figura 10-22
Finestra di dialogo Componenti principali categoriale



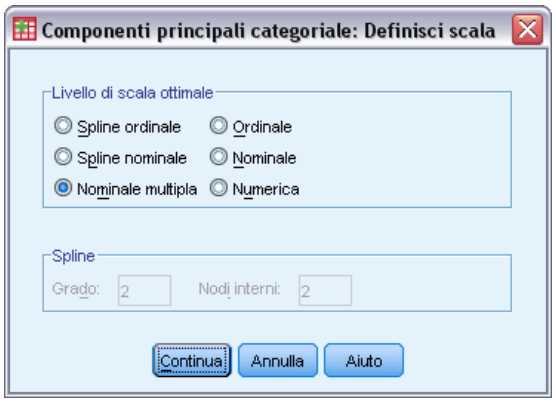
- Selezionare da *Peso corporeo* a *Percezione del proprio corpo* come variabili di analisi.
- Fare clic su Definisci scala e peso.

Figura 10-23
Finestra di dialogo Definisci scala e peso



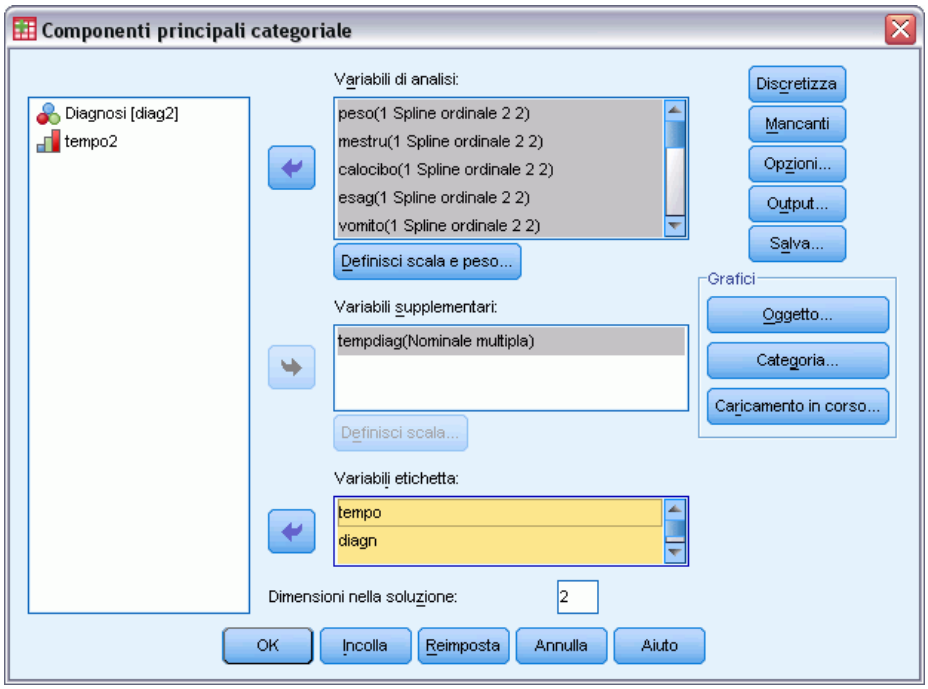
- Selezionare Ordinale come livello di scaling ottimale.
- Fare clic su Continua.
- Selezionare *Interazione diagnosi/tempo* come variabile supplementare e fare clic su Definisci scala nella finestra di dialogo Analisi componenti principali categoriale.

Figura 10-24
Finestra di dialogo Definisci scala



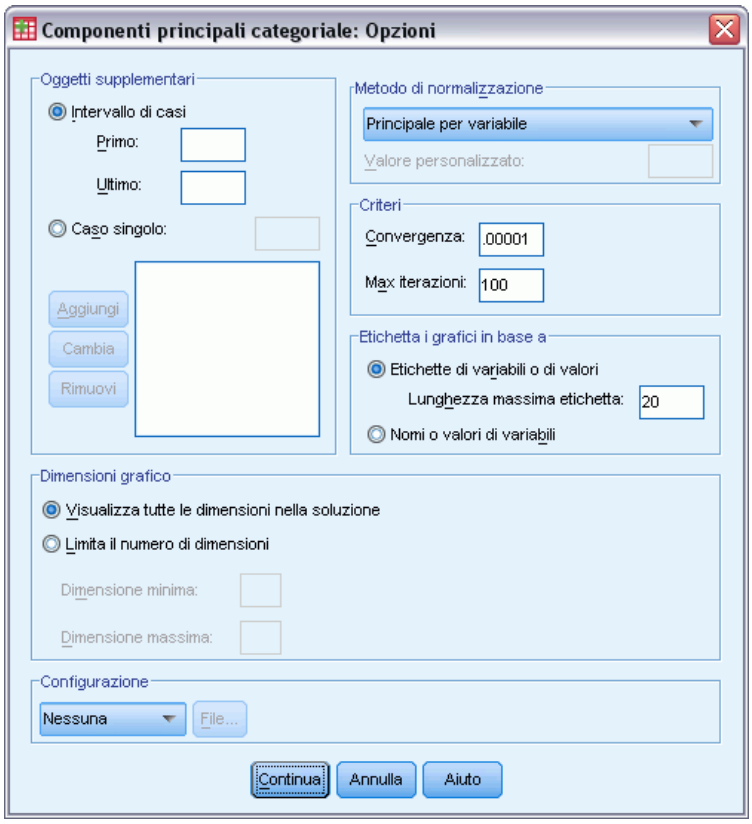
- Selezionare Nominale multiplo come livello di scaling ottimale.
- Fare clic su Continua.

Figura 10-25
Finestra di dialogo Componenti principali categoriale



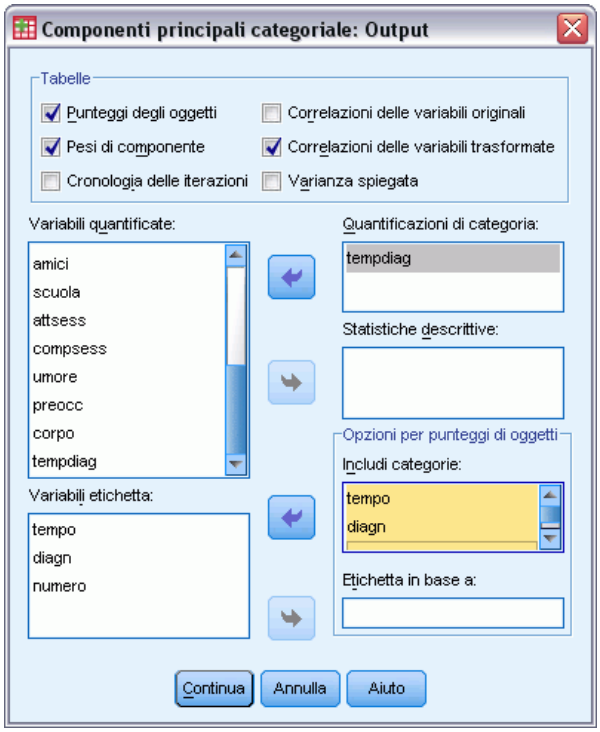
- Selezionare da *Numero colloquio* a *Numero paziente* come variabili di etichetta.
- Fare clic su Opzioni.

Figura 10-26
Finestra di dialogo Opzioni



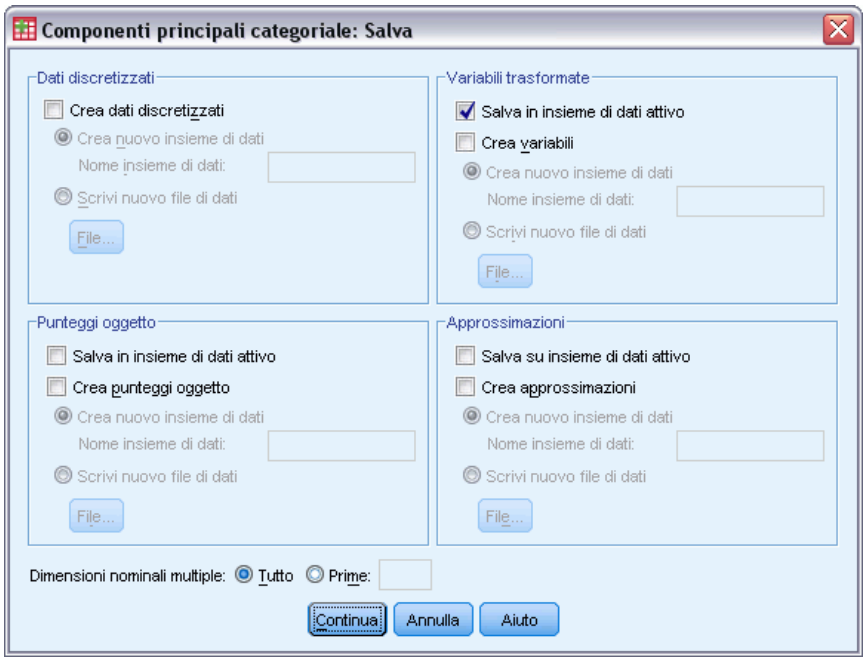
- Scegliere di etichettare i grafici in base a Nomi o valori di variabili.
- Fare clic su Continua.
- Fare clic su Output nella finestra di dialogo Analisi componenti principali categoriale.

Figura 10-27
Finestra di dialogo Output



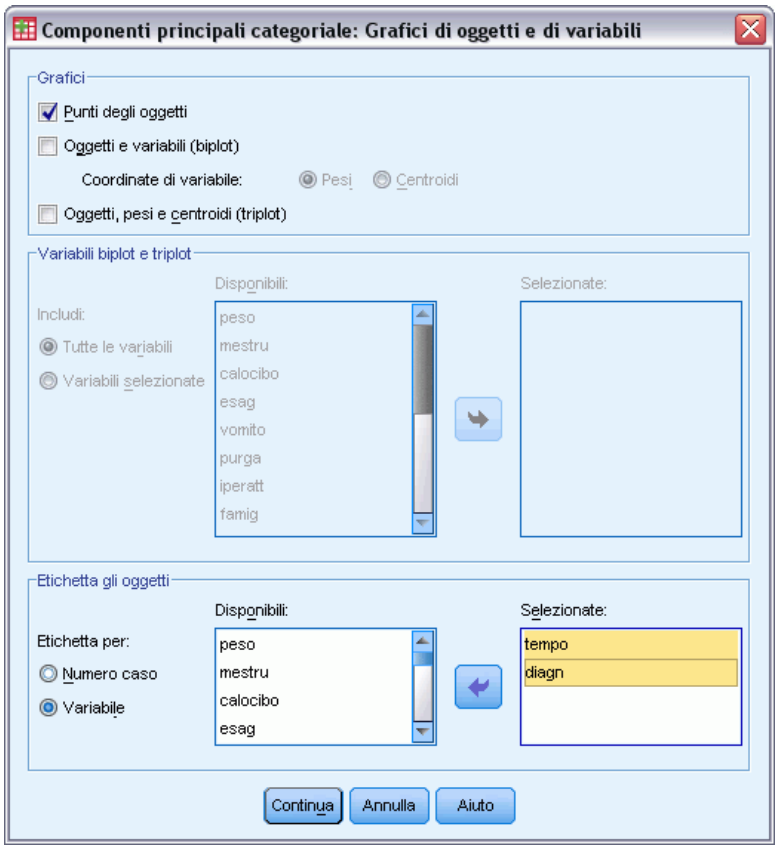
- Selezionare Punteggi degli oggetti nel gruppo Tabelle.
- Richiedere le quantificazioni di categoria per *tidi*
- Scegliere di includere le categorie *ora*, *diag* e *numero*.
- Fare clic su Continua.
- Fare clic su Salva nella finestra di dialogo Analisi componenti principali categoriale.

Figura 10-28
Salva



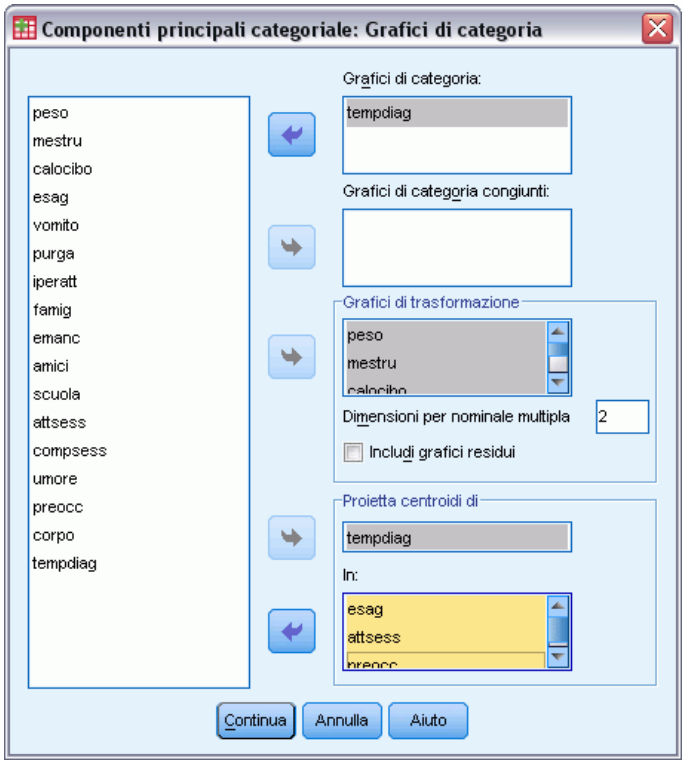
- Nel gruppo Variabili trasformate, selezionare Salva nel file di dati attivo.
- Fare clic su Continua.
- Fare clic su Oggetto nella finestra di dialogo Analisi componenti principali categoriale.

Figura 10-29
Finestra di dialogo Grafici di oggetti e di variabili



- Scegliere di etichettare gli oggetti in base a Variabile.
- Selezionare *ora* e *diag* come variabili in base alle quali etichettare gli oggetti.
- Fare clic su Continua.
- Fare clic su Categoria nella finestra di dialogo Analisi componenti principali categoriale.

Figura 10-30
Finestra di dialogo Grafici di categoria

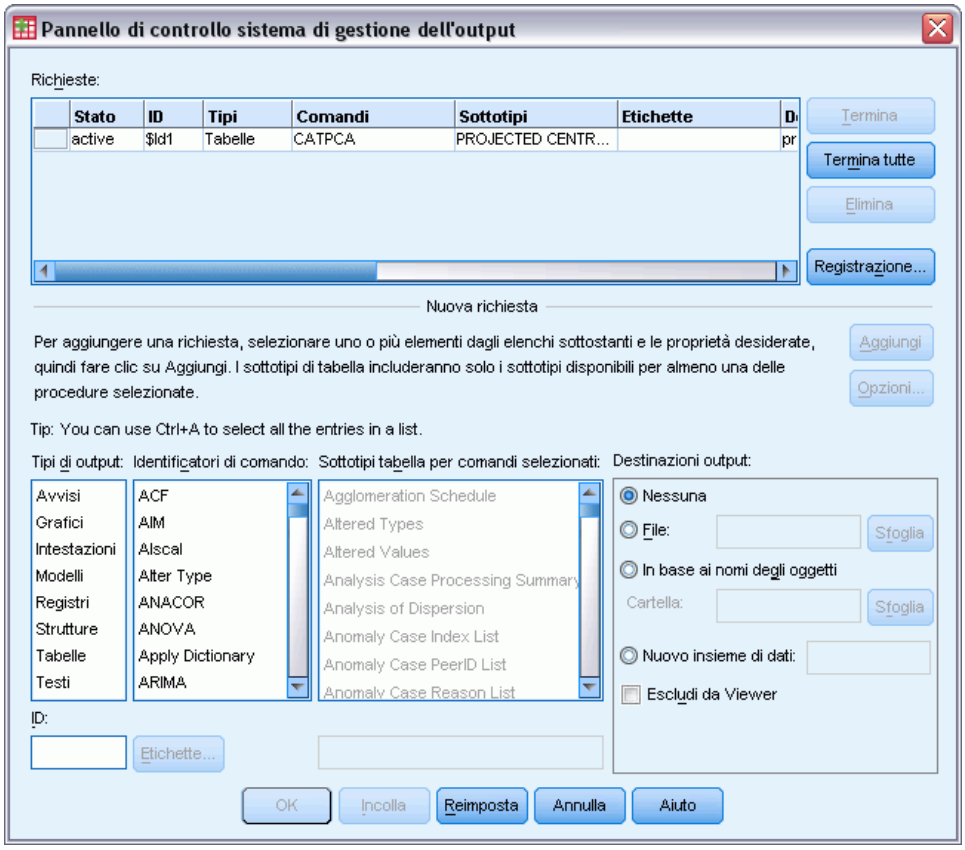


- Richiedere i grafici di categoria per *tidi*
- Richiedere i grafici di trasformazione per le variabili da *peso* a *corpo*.
- Scegliere di proiettare i centroidi di *tidi* su *eccesso*, *attses* e *preo*.
- Fare clic su Continua.
- Fare clic su OK nella finestra di dialogo Analisi componenti principali categoriale.

  La procedura dà come risultato punteggi per i soggetti (con media 0 e varianza di unità) e quantificazioni delle categorie che massimizzano la correlazione quadratica media tra i punteggi dei soggetti e le variabili trasformate. Nell'analisi corrente, le quantificazioni di categoria sono state limitate per riflettere le informazioni ordinali.

Infine, per scrivere le informazioni della tabella dei centroidi proiettati nel file *projected_centroids.sav*, è necessario terminare la richiesta SGO. Richiamare il Pannello di controllo SGO.

Figura 10-31
Pannello di controllo sistema di gestione dell'output

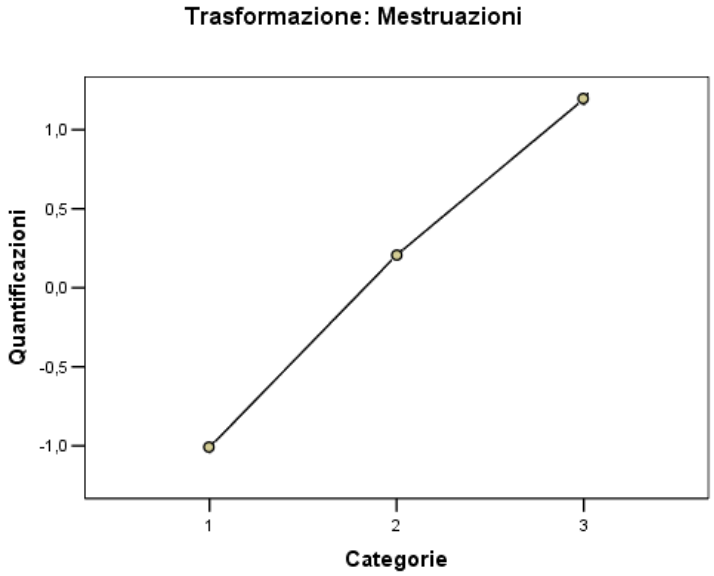


- Fare clic su Termina.
- Fare clic su OK e quindi di nuovo su OK per confermare.

Grafici di trasformazione

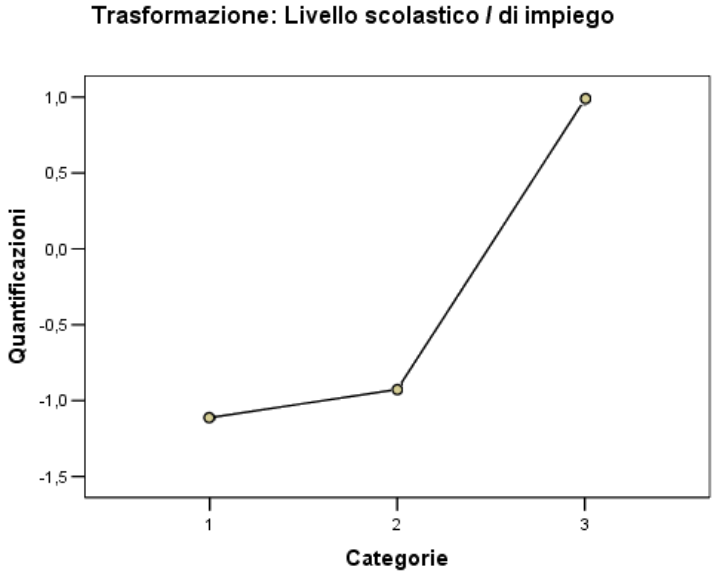
I grafici di trasformazione visualizzano il numero della categoria originale sugli assi orizzontali; gli assi verticali indicano le quantificazioni ottimali.

Figura 10-32
Grafico di trasformazione per ciclo mestruale



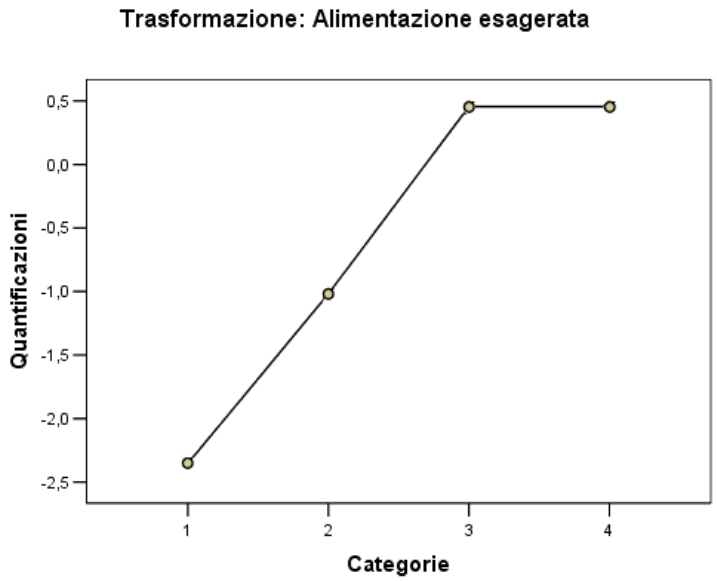
Alcune variabili, come *Ciclo mestruale*, hanno ottenuto trasformazioni quasi lineari, perciò in questa analisi è possibile interpretarle come numeriche.

Figura 10-33
Grafico di trasformazione per Risultati lavorativi/scolastici



Le quantificazioni per le altre variabili, come *Risultati lavorativi/scolastici* non hanno ottenuto trasformazioni lineari e dovrebbero essere interpretate a livello di scaling ordinale. La differenza tra la seconda e la terza categoria è molto più significativa di quella tra la prima e la seconda.

Figura 10-34
Grafico di trasformazione per Eccessi alimentari



Un caso interessante si verifica per le quantificazioni di *Eccessi alimentari*. La trasformazione ottenuta è lineare per le categorie da 1 a 3, ma i valori quantificati per le categorie 3 e 4 sono uguali. Questo risultato mostra che i punteggi 3 e 4 non fanno differenza tra i pazienti e suggerisce che sia possibile utilizzare il livello di scaling numerico in una soluzione a due componenti ricodificando 4' come 3'.

Riepilogo del modello

Figura 10-35
Riepilogo modello

Dimensione	Alfa di Cronbach	Varianza spiegata	
		Totale (autovalore)	% di varianza
1	,874	5,550	34,690
2	,522	1,957	12,234
Totale	,925[a]	7,508	46,924

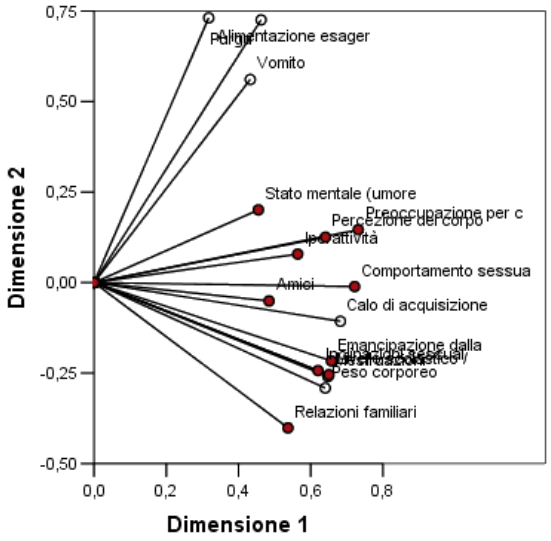
a. Il totale di Alfa di Cronbach è basato sul totale dell'autovalore.

Per verificare l'attendibilità del modello rispetto ai dati, si veda il riepilogo del modello. Circa il 47% della varianza totale è spiegata dal modello a due componenti, il 35% dalla prima dimensione e il 12% dalla seconda. Di conseguenza, quasi la metà della variabilità a livello di singoli oggetti è spiegata dal modello a due componenti.

Pesi di componente

Per iniziare l'interpretazione delle due dimensioni della soluzione, si esaminino i pesi di componente. Tutte le variabili hanno un peso di componente positivo nella prima dimensione: questo significa che esiste un fattore comune di correlazione positiva con tutte le variabili.

Figura 10-36
Grafico dei pesi di componente



La seconda dimensione separa le variabili: Le variabili *Eccessi alimentari*, *Episodi di vomito* e *Uso di lassativi* formano un insieme con elevati pesi positivi nella seconda dimensione. Questi sintomi sono tipicamente considerati rappresentativi di un comportamento bulimico.

Le variabili *Emancipazione dalla famiglia*, *Risultati lavorativi/scolastici*, *Atteggiamento sessuale*, *Peso corporeo* e *Ciclo mestruale* formano un altro insieme, ed è possibile includere *Limitata ingestione di cibo (digiuno)* e *Rapporti famigliari* nel medesimo insieme, in quanto i relativi vettori sono vicini al cluster principale, e queste variabili sono considerate sintomi di anoressia (digiuno, peso, ciclo mestruale) o sono di natura psicosociale (emancipazione, risultati lavorativi/scolastici, atteggiamento sessuale, rapporti famigliari). I vettori di questo insieme sono ortogonali (perpendicolari) ai vettori di eccessi, vomito e lassativi, il che significa che questo gruppo di variabili è privo di correlazione con l'insieme delle variabili indicative di bulimia.

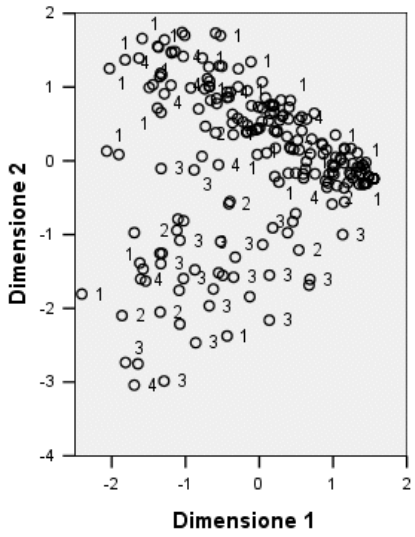
Le variabili *Amici*, *Stato mentale (umore)* e *Iperattività* non sembrano adattarsi particolarmente bene alla soluzione. È possibile vederlo nel grafico osservando la lunghezza di ciascun vettore. La lunghezza del vettore di una data variabile corrisponde al suo adattamento e i vettori di queste variabili sono i più corti. In una soluzione a due componenti, queste variabili verrebbero probabilmente eliminate da una proposta di sintomatologia relativa ai disturbi alimentari. Esse potrebbero tuttavia adattarsi meglio a una soluzione con un maggiore numero di dimensioni.

Le variabili *Comportamento sessuale*, *Preoccupazione legata a cibo e peso* e *Percezione del proprio corpo* formano un altro gruppo teorico di simboli, relativo alla percezione del proprio corpo da parte del paziente. Sebbene correlate con i due insiemi ortogonali di variabili, queste variabili hanno vettori piuttosto lunghi e sono strettamente associate alla prima dimensione; di conseguenza, possono fornire informazioni utili circa il fattore "comune".

Punteggi oggetto

La figura seguente mostra un grafico dei punteggi degli oggetti, nel quale i soggetti sono etichettati in base alla categoria diagnostica.

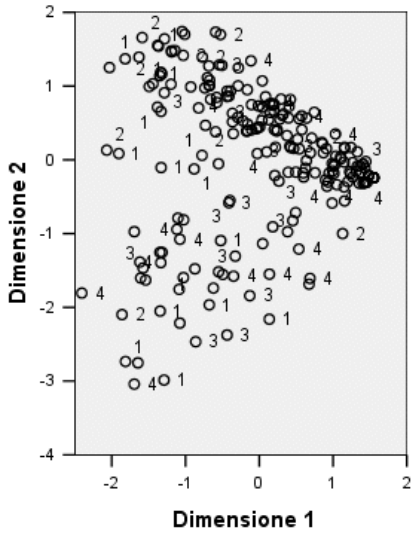
Figura 10-37
Grafico dei punteggi degli oggetti etichettati per diagnosi



Questo grafico non consente di interpretare la prima dimensione perché in esso i pazienti non sono separati per diagnosi. Tuttavia, sono presenti alcune informazioni sulla seconda dimensione. I soggetti anoressici (1) e i pazienti con disturbi dell'alimentazione atipici (4) formano un gruppo, collocato sopra i soggetti con una forma di bulimia (2 e 3). Di conseguenza, la seconda dimensione divide i pazienti bulimici dagli altri, come visto anche nella sezione precedente (le variabili dell'insieme relativo alla bulimia hanno elevati pesi di componente positivi nella seconda dimensione). Questo ha senso in quanto i pesi di componente dei sintomi generalmente associati alla bulimia hanno valori elevati nella seconda dimensione.

La figura mostra un grafico dei punteggi degli oggetti, nel quale i soggetti sono etichettati in base al momento della diagnosi.

Figura 10-38
Punteggi degli oggetti etichettati in base al numero del colloquio



Le etichette dei punteggi degli oggetti in base alla progressione temporale indicano che la prima dimensione ha una correlazione con quest'ultima: sembra infatti che vi sia una progressione dei tempi diagnostici dall'1 in maggioranza a sinistra e gli altri a destra. Si noti che è possibile collegare i punti temporali nel grafico salvando i punteggi degli oggetti e creando un grafico a dispersione utilizzando i punteggi della dimensione 1 sull'asse *x*, i punteggi della dimensione 2 sull'asse *y* e impostando i simboli utilizzando i numeri dei pazienti.

Confrontando il grafico dei punteggi degli oggetti in base al tempo con quello etichettato per diagnosi è possibile ottenere alcune indicazioni su oggetti insoliti. Ad esempio, nel grafico etichettato in base al tempo, è presente un paziente la cui diagnosi in corrispondenza del quarto incontro si trova a sinistra di tutti gli altri punti del grafico. Questo è insolito in quanto il trend generale dei punti relativi ai colloqui successivi nel tempo è di trovarsi più a destra. È interessante notare come questo punto apparentemente fuori posto dal punto di vista temporale corrisponda anche a una diagnosi insolita, nel senso che il paziente è un soggetto anoressico i cui punteggi lo inseriscono nel cluster relativo alla bulimia. Esaminando la tabella dei punteggi degli oggetti, si vedrà che si tratta del paziente numero 43, cui è stata diagnosticata un'anoressia nervosa e i cui punteggi sono indicati nella tabella seguente.

Tabella 10-4
Punteggi degli oggetti per il paziente n. 43

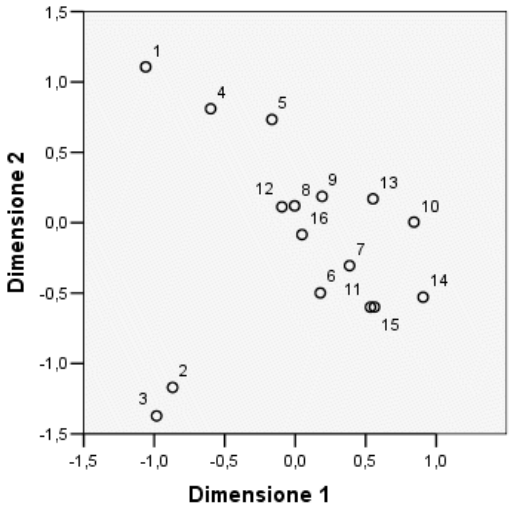
Ora	Dimensione 1	Dimensione 2
1	–2.031	1.250
2	–2.067	0.131
3	–1.575	–1.467
4	–2.405	–1.807

I punteggi del paziente in corrispondenza del colloquio numero 1 sono tipici dei soggetti anoressici, con un elevato punteggio negativo nella dimensione 1, corrispondente a una percezione negativa del proprio corpo, e un punteggio positivo per la dimensione 2, indicativo di sintomi di anoressia o di comportamento psicosociale inadeguato. Tuttavia, diversamente dalla maggioranza dei pazienti, nella dimensione 1 i progressi sono scarsi o assenti. Nella dimensione 2, apparentemente sono presenti dei progressi verso "normale" (attorno allo 0, tra comportamento anoressico e bulimico), ma successivamente il paziente inizia a mostrare sintomi bulimici.

Esame della struttura dell'andamento della malattia

Per reperire maggiori informazioni sulla connessione tra le due dimensioni e le quattro categorie diagnostiche e i quattro punti temporali, è stata creata la variabile supplementare *Interazione diagnosi/tempo* tramite una classificazione incrociata delle quattro categorie di *Diagnosi paziente* e le quattro categorie di *Numero colloquio*. Di conseguenza, *Interazione diagnosi/tempo* ha 16 categorie, la prima delle quali indica i pazienti diagnosticati con anoressia nervosa alla prima visita. La quinta categoria indica i pazienti diagnosticati con anoressia nervosa al punto temporale 2 e così via; la sedicesima categoria indica i pazienti con disturbi alimentari atipici al punto temporale 4. L'utilizzo della variabile supplementare *Interazione diagnosi/tempo* consente di studiare l'andamento della malattia per i vari gruppi nel tempo. Alla variabile è assegnato un livello di scaling nominale multiplo e i punti di categoria sono visualizzati nella figura seguente.

Figura 10-39
Punti di categoria per interazione diagnosi/tempo

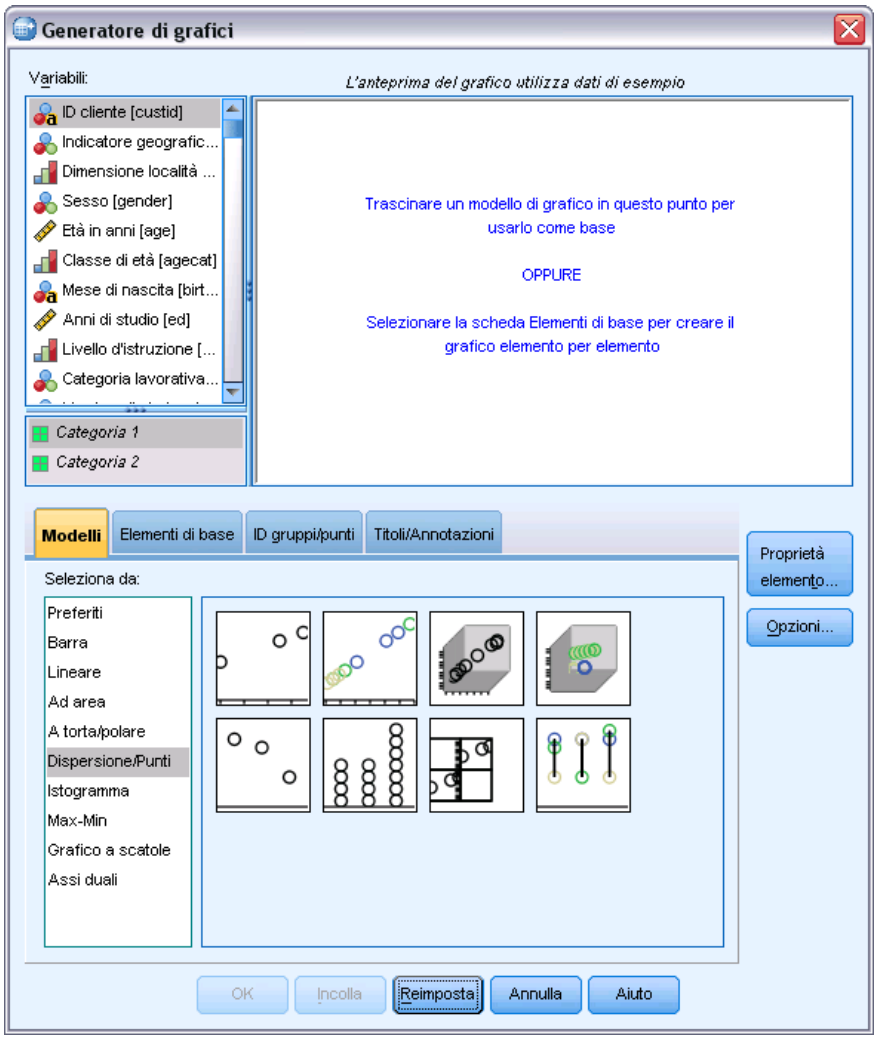


Parte della struttura è evidente dal grafico corrente: le categorie diagnostiche in corrispondenza del punto temporale 1 separano chiaramente l'anoressia nervosa e il disturbo alimentare atipico dall'anoressia nervosa con bulimia nervosa e dalla bulimia nervosa post anoressia nervosa nella seconda dimensione. A parte questo, evidenziare i modelli risulta leggermente più difficile.

Tuttavia, è possibile rendere tali modelli più visibili creando un grafico a dispersione basato sulle quantificazioni. Per farlo, dai menu scegliere:

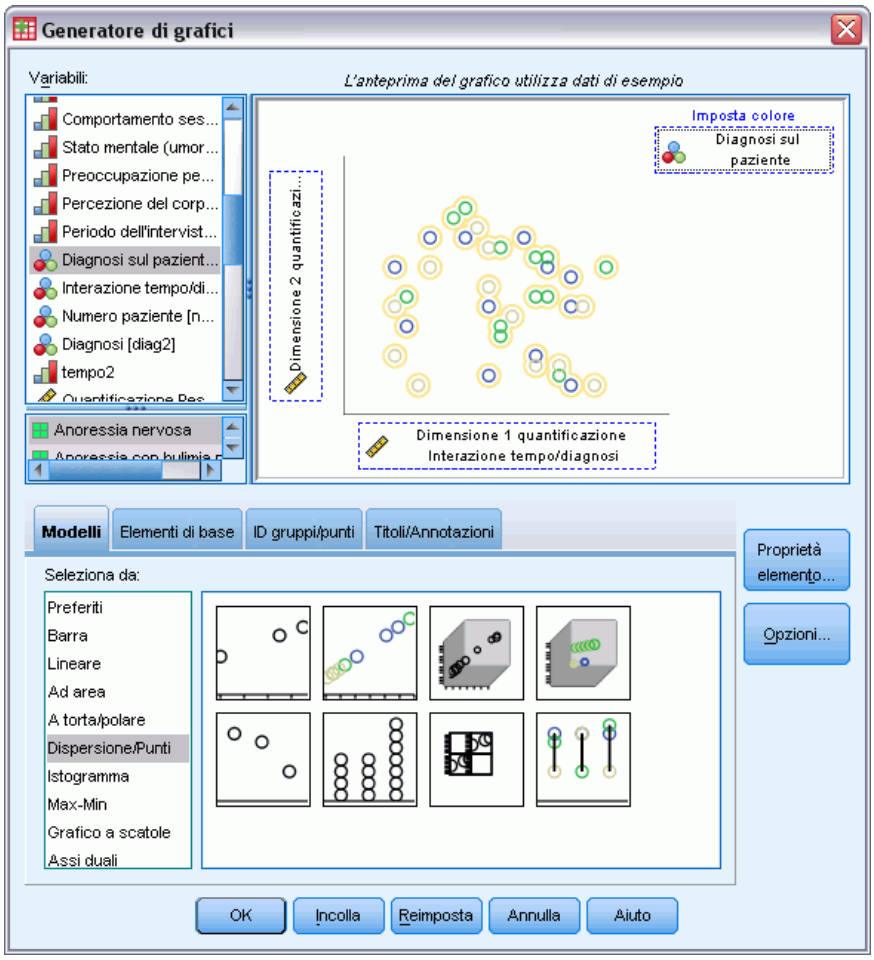
Grafici > Generatore di grafici...

Figura 10-40
Modello Dispersione/Punti



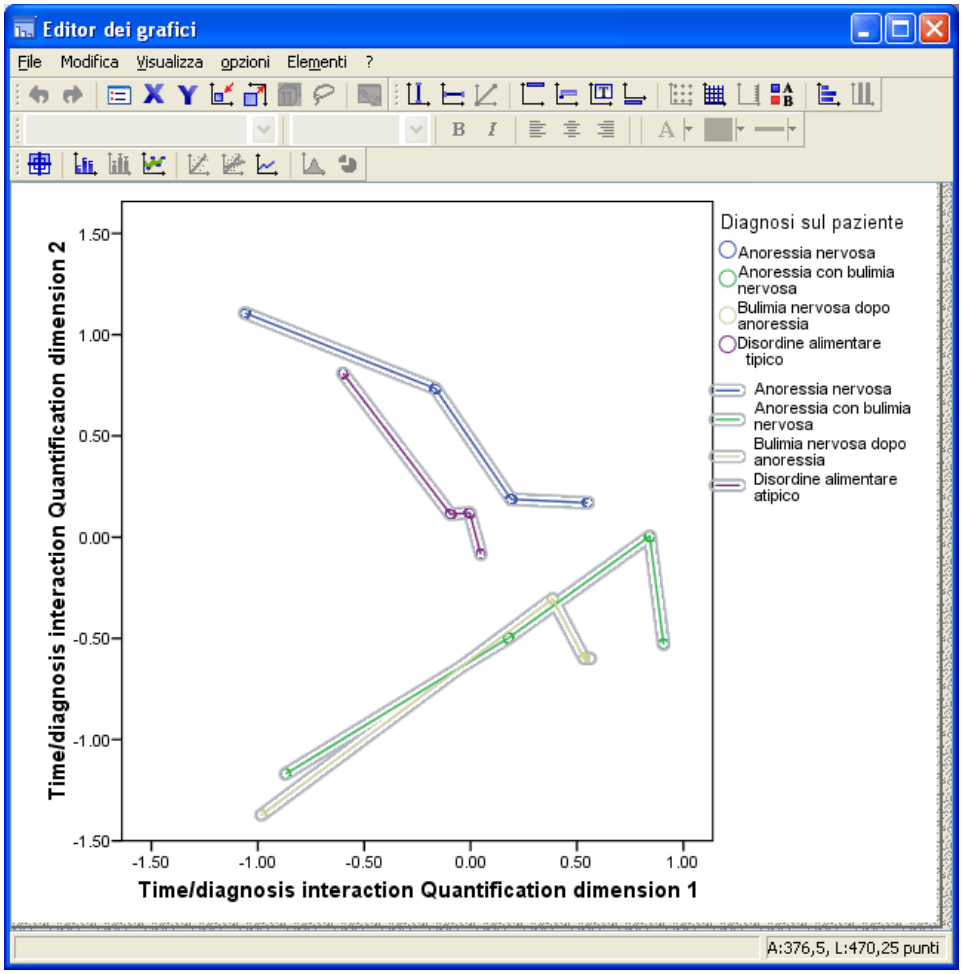
▶ Selezionare il modello Dispersione/Punti e scegliere A dispersione raggruppato.

Figura 10-41
Generatore di grafici



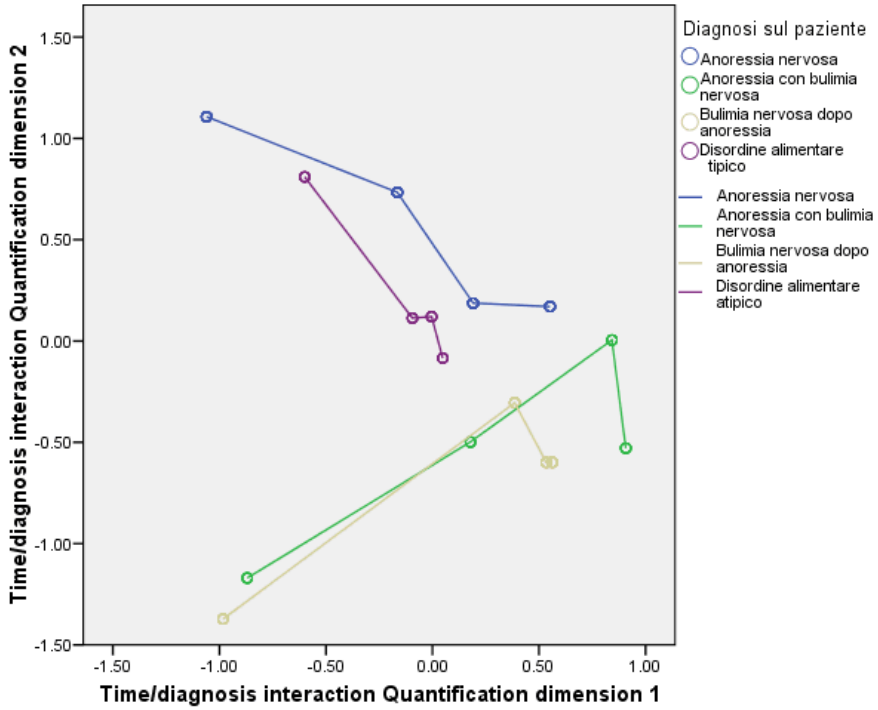
- Selezionare*Quantificazione interazione diagnosi/tempo dimensione 2* come variabile dell'asse *y* e *Quantificazione diagnosi/tempo dimensione 1* come variabile dell'asse *x*.
- Scegliere di impostare il colore in base a *Diagnosi paziente*.
- Fare clic su OK.

Figura 10-42
Strutture dell'andamento della malattia



- Quindi, per collegare i punti, fare doppio clic sul grafico e quindi fare clic sullo strumento Aggiungi linea di interpolazione nell'Editor dei grafici.
- Chiudere l'Editor dei grafici.

Figura 10-43
Strutture dell'andamento della malattia



Collegando i punti di categoria per ogni categoria diagnostica nel tempo, i modelli suggeriscono immediatamente che la prima dimensione è correlata al tempo e la seconda alla diagnosi, come determinato in precedenza per i grafici dei punteggi degli oggetti.

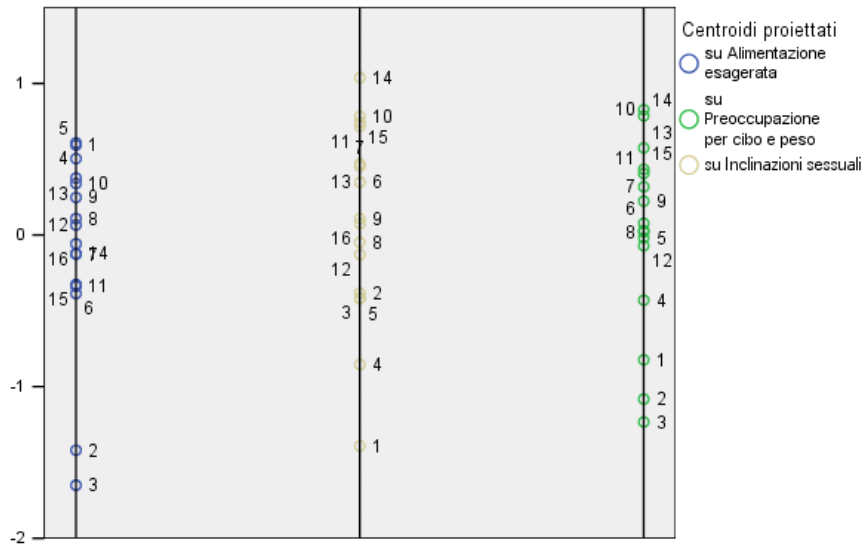
Tuttavia, questo grafico mostra che, nel tempo, i disturbi tendono a diventare più simili tra loro. Inoltre, per tutti i gruppi, i progressi sono maggiori tra i punti temporali 1 e 2; i pazienti anoressici mostrano più progressi da 2 a 3, ma per gli altri gruppi i progressi sono scarsi.

Sviluppo differenziale per variabili selezionate

Una variabile da ogni insieme di sintomi identificata dai pesi di componente è stata selezionata come "rappresentativa" dell'insieme. La variabile Eccessi alimentari è stata selezionata per l'insieme bulimico, Atteggiamento sessuale per l'insieme anoressico/psicosociale e preoccupazione legata al corpo per il terzo insieme.

Per esaminare i possibili andamenti differenziali della malattia, le proiezioni di *Interazione diagnosi/tempo* per *Eccessi alimentari*, *Atteggiamento sessuale* e *Preoccupazione legata a cibo e peso* sono state calcolate e tracciate in un grafico nella seguente figura.

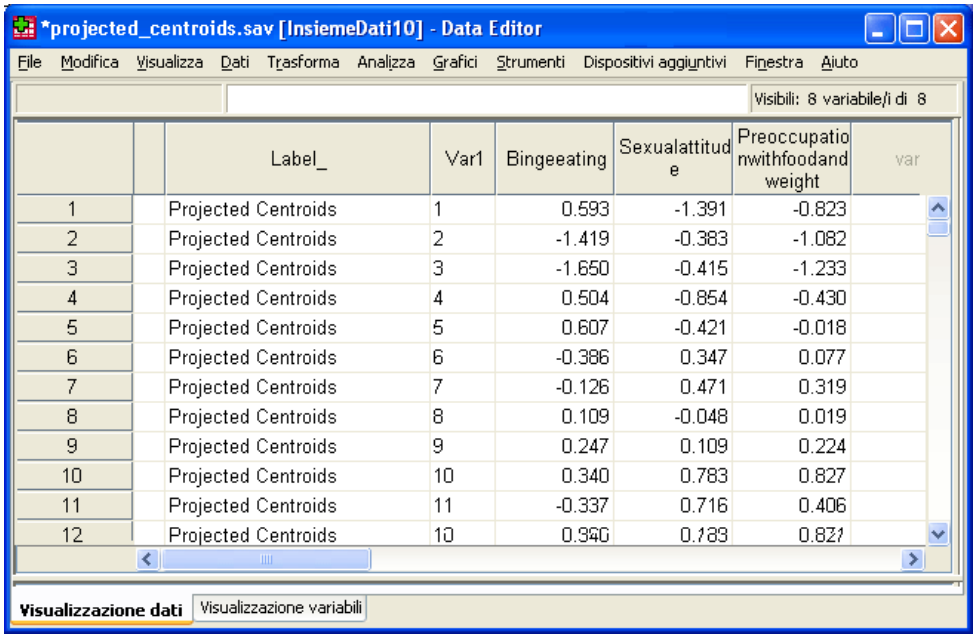
Figura 10-44
Centroidi proiettati di Interazione diagnosi/tempo su Eccessi alimentari, Atteggiamento sessuale e Preoccupazione legata a cibo e peso



Il grafico mostra che, in corrispondenza del primo punto temporale, il sintomo Eccessi alimentari separa i pazienti bulimici (2 e 3) dagli altri (1 e 4); Atteggiamento sessuale separa gli anoressici e i pazienti atipici (1 e 4) dagli altri (2 e 3); Preoccupazione legata a cibo e peso non separa i pazienti in modo significativo. In molte applicazioni, questo grafico sarebbe sufficiente per descrive la relazione tra i sintomi e la diagnosi, ma a causa della complicazione rappresentata dai punti di tempo multipli, il quadro diventa più confuso.

Per visualizzare queste proiezioni nel tempo, è necessario poter tracciare i contenuti della tabella dei centroidi proiettati in un grafico. Questo è reso possibile dalla richiesta SGO che ha salvato tali informazioni nel file *projected_centroids.sav*.

Figura 10-45
Projected_centroids.sav

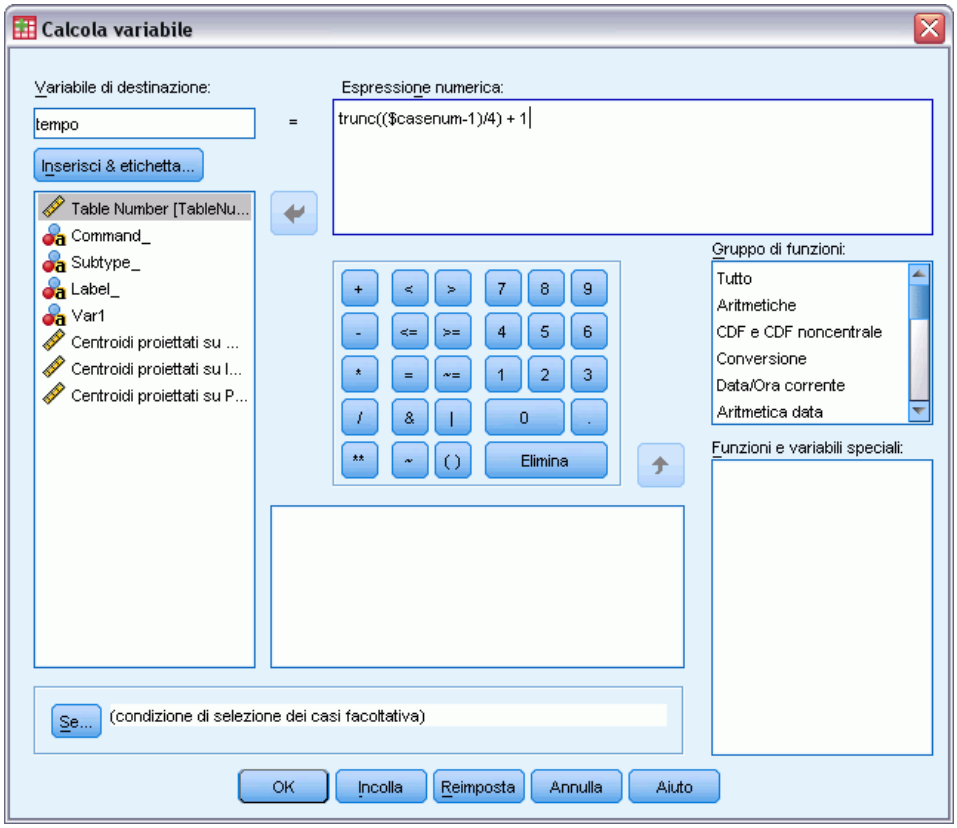


Le variabili *Eccessialimentari*, *Atteggiamentosessuale* e *Preoccupazionecibopeso* contengono i valori dei centroidi proiettati su ciascuno dei sintomi di interesse. Il numero dei casi (da 1 a 16) corrisponde all'interazione diagnosi/tempo. Sarà necessario calcolare le nuove variabili che separano i valori Tempo e Diagnosi.

▶ Dai menu, scegliere:

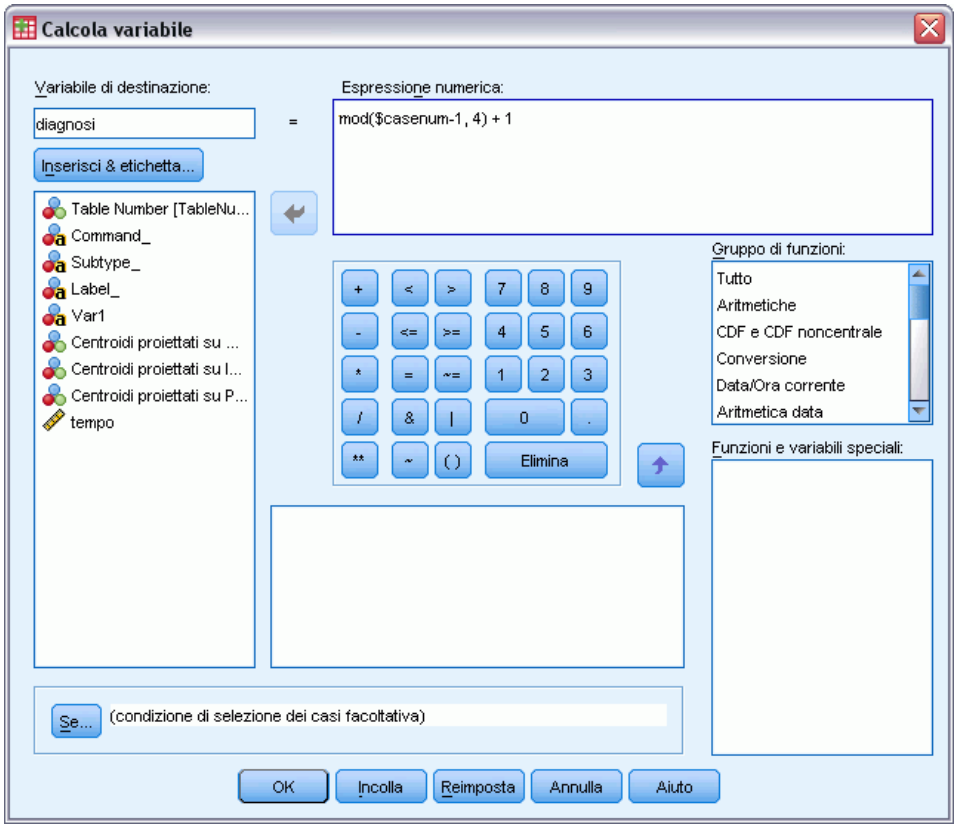
Trasforma > Calcola variabile...

Figura 10-46
Finestra di dialogo Calcola variabile



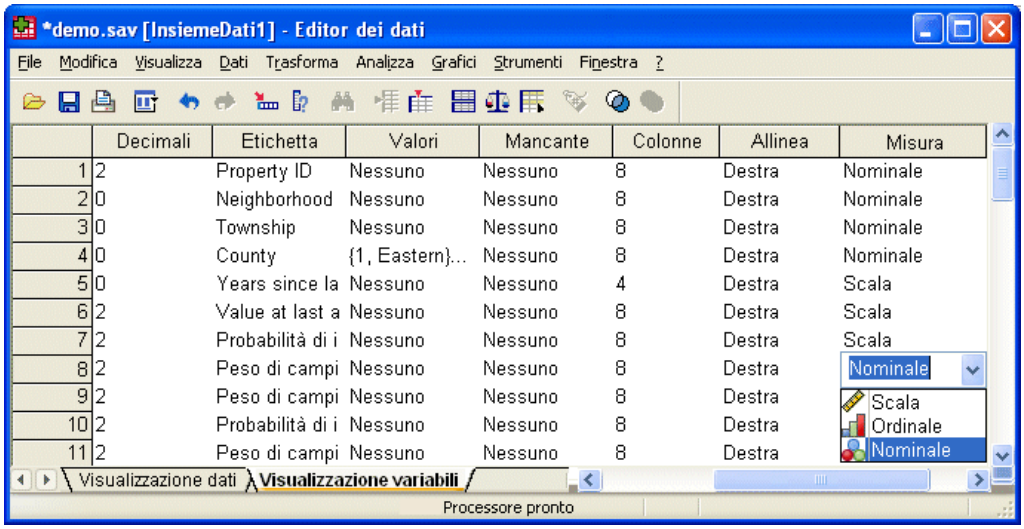
- Digitare *tempo* come variabile di destinazione.
- Digitare trunc(($casenum-1)/4) + 1 come espressione numerica.
- Fare clic su OK.

Figura 10-47
Finestra di dialogo Calcola variabile



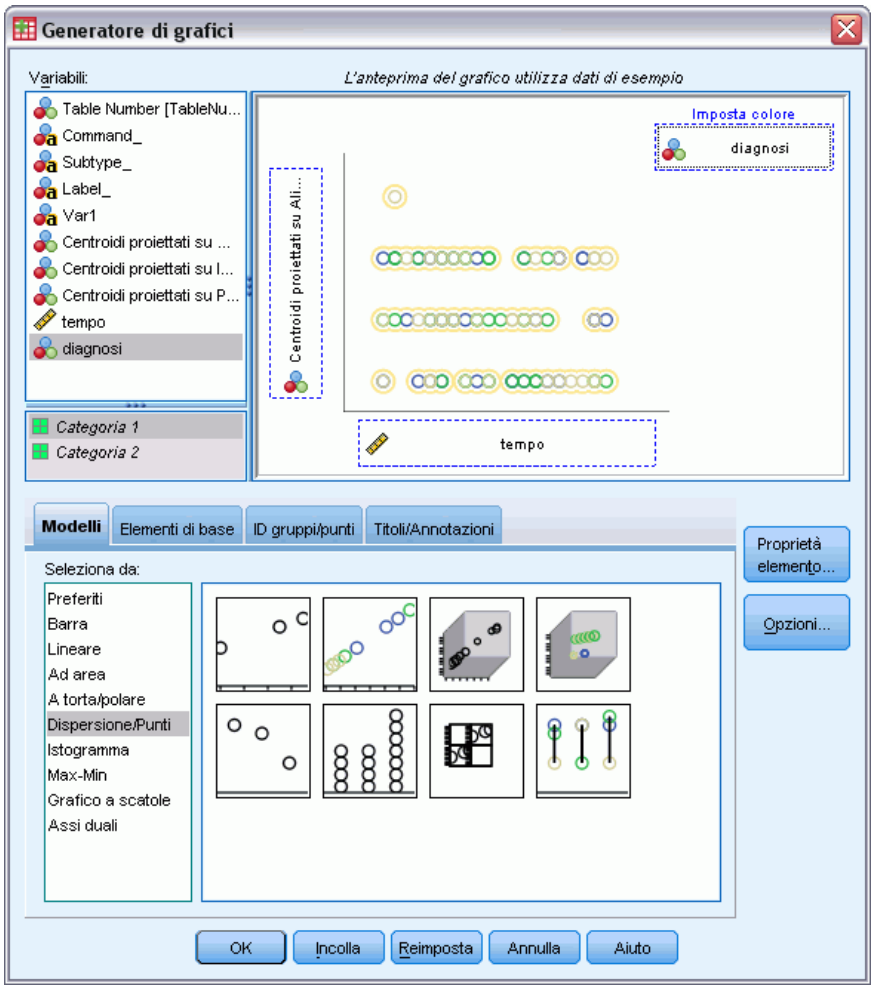
- Richiamare la finestra di dialogo Calcola variabile.
- Digitare *diagnosi* come variabile di destinazione.
- Digitare mod($casenum-1,4) + 1 come espressione numerica.
- Fare clic su OK.

Figura 10-48
Projected_centroids.sav



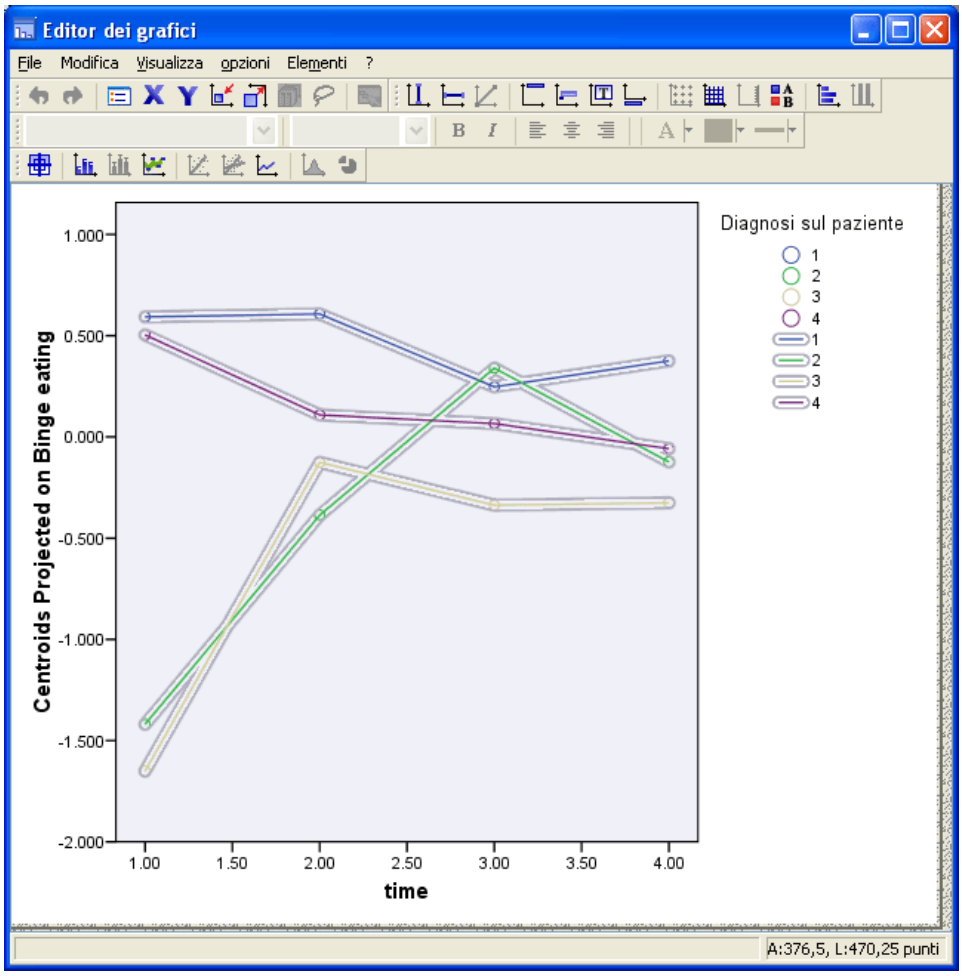
Nella Visualizzazione variabili modificare la misura di *diagnosi* da Scala to Nominale.

Figura 10-49
Generatore di grafici



- Infine, per visualizzare i centroidi proiettati del momento temporale della diagnosi su eccessi alimentari nel tempo, richiamare Generatore di grafici e fare clic su Ripristina per annullare le selezioni precedenti.
- Selezionare il modello Dispersione/Punti e scegliere A dispersione raggruppato.
- Selezionare*Centroidi proiettati su Eccessi alimentari* come variabile dell'asse *y* e *tempo* come variabile dell'asse *x*.
- Scegliere di impostare i colori in base a *diagnosi*.
- Fare clic su OK.

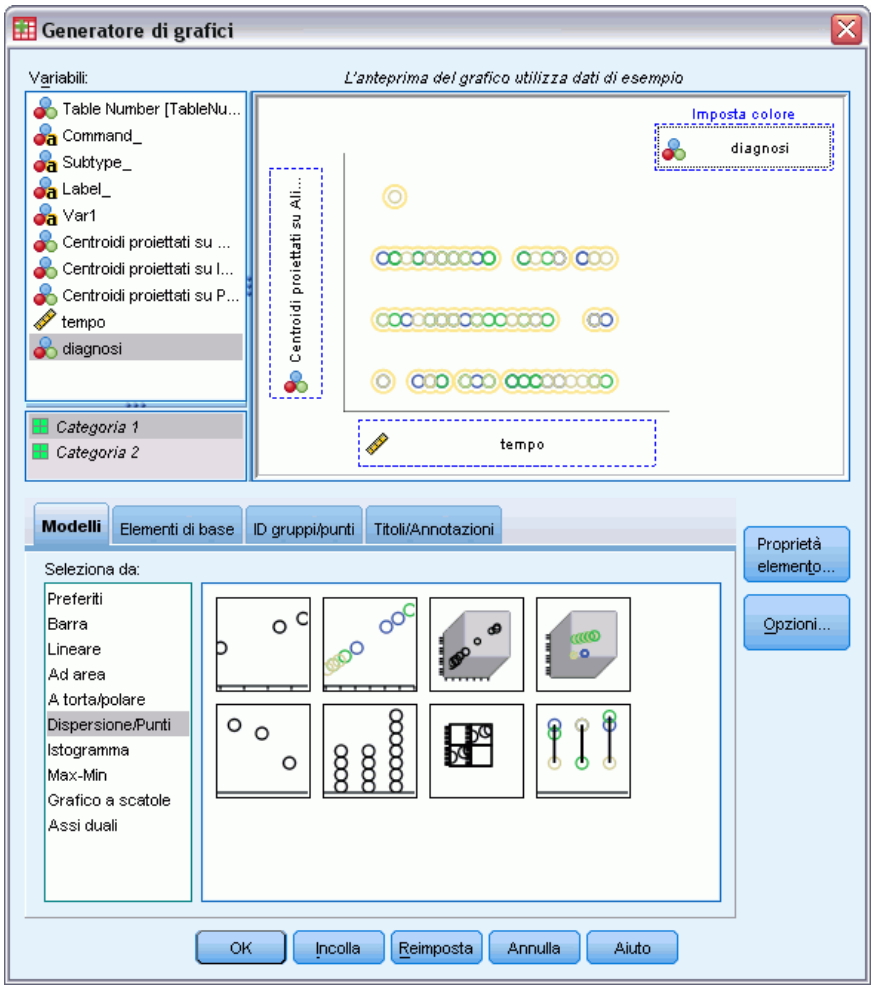
Figura 10-50
Centroidi proiettati del momento della diagnosi su Eccessi alimentari nel tempo



- Quindi, per collegare i punti, fare doppio clic sul grafico e quindi fare clic sullo strumento Aggiungi linea di interpolazione nell'Editor dei grafici.
- Chiudere l'Editor dei grafici.

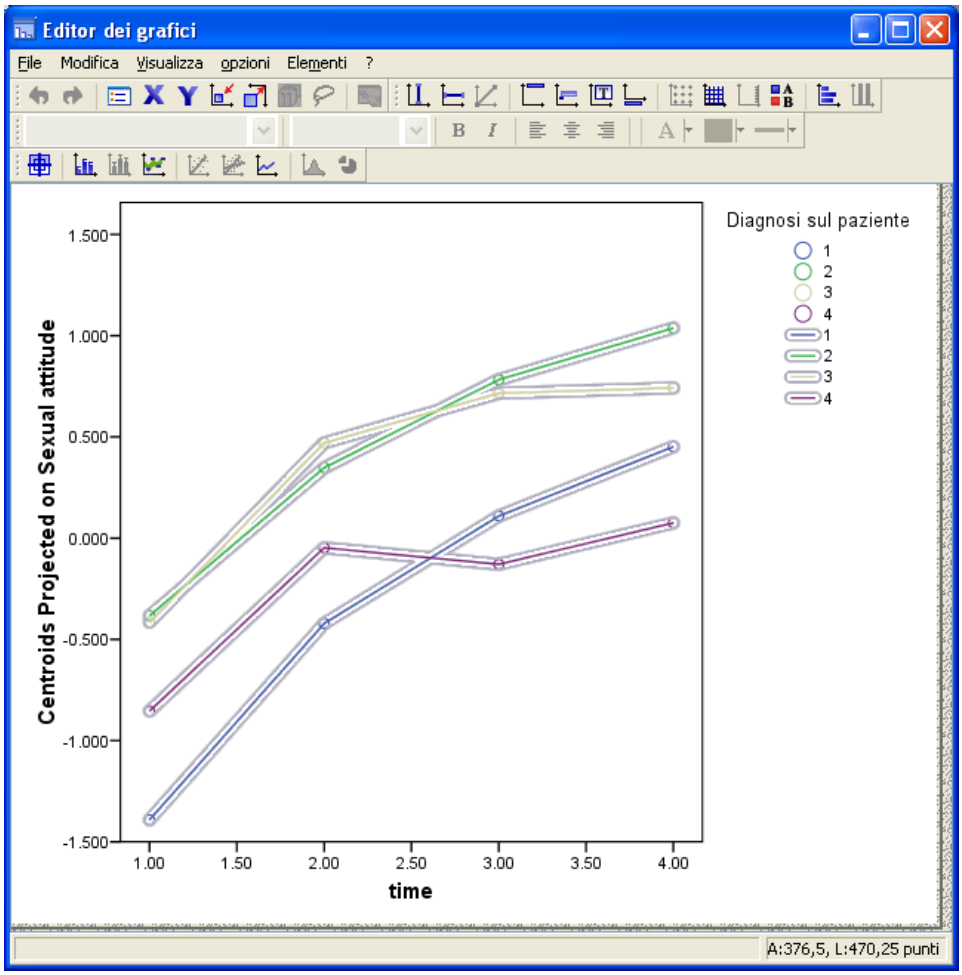
Rispetto agli eccessi alimentari è chiaro che i gruppi anoressici hanno valori iniziali diversi dai gruppi bulimici. La differenza si riduce nel tempo, in quanto i gruppi anoressici si modificano solo leggermente, mentre i gruppi bulimici mostrano progressi.

Figura 10-51
Generatore di grafici



- Richiamare Generatore grafici.
- Deselezionare*Centroidi proiettati su Eccessi alimentari* come variabile dell'asse *y* e selezionare *Centroidi proiettati su Atteggiamento sessuale* come variabile dell'asse *x*.
- Fare clic su OK.

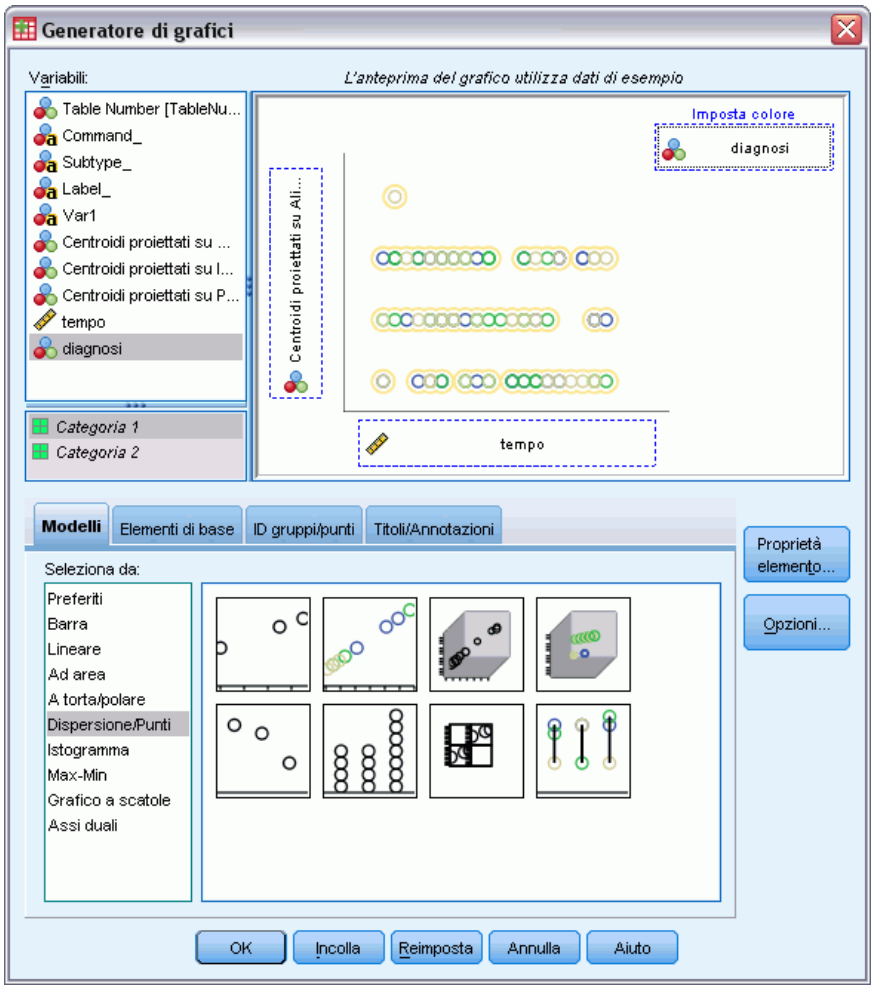
Figura 10-52
Centroidi proiettati del momento della diagnosi su Atteggiamento sessuale nel tempo



- Quindi, per collegare i punti, fare doppio clic sul grafico e quindi fare clic sullo strumento Aggiungi linea di interpolazione nell'Editor dei grafici.
- Chiudere l'Editor dei grafici.

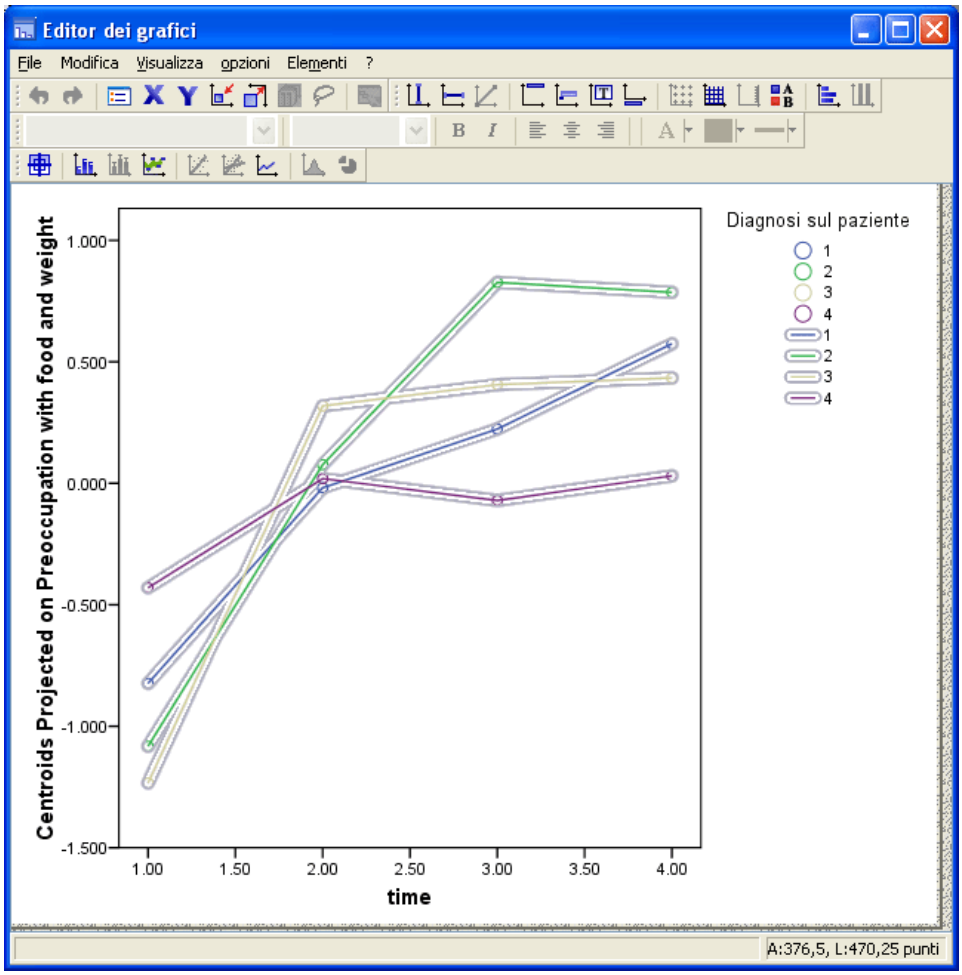
Rispetto all'atteggiamento sessuale, le quattro traiettorie sono più o meno parallele nel tempo e tutti i gruppi mostrano dei progressi. I gruppi bulimici, tuttavia, hanno punteggi più elevati (migliori) del gruppo anoressico.

Figura 10-53
Generatore di grafici



- Richiamare Generatore grafici.
- Deselezionare*Centroidi proiettati su Atteggiamento sessuale* come variabile dell'asse *y* e selezionare *Centroidi proiettati su Preoccupazione legata a cibo e peso* come variabile dell'asse *x*.
- Fare clic su OK.

Figura 10-54
Centroidi proiettati del momento della diagnosi su Preoccupazione legata al corpo nel tempo



- Quindi, per collegare i punti, fare doppio clic sul grafico e quindi fare clic sullo strumento Aggiungi linea di interpolazione nell'Editor dei grafici.
- Chiudere l'Editor dei grafici.

La preoccupazione relativa al corpo è una variabile che rappresenta i sintomi chiave, condivisi dai quattro diversi gruppi. Oltre ai pazienti con disturbi alimentari atipici, il gruppo anoressico e i due gruppi bulimici hanno livelli molto simili sia all'inizio sia alla fine.

Letture consigliate

Consultare i testi seguenti per maggiori informazioni sull'analisi componenti principali categoriale:

De Haas, M., J. A. Algera, H. F. J. M. Van Tuijl, e J. J. Meulman. 2000. Macro and micro goal setting: In search of coherence. *Applied Psychology*, 49, .

De Leeuw, J. 1982. Nonlinear principal components analysis. In: *COMPSTAT Proceedings in Computational Statistics,* Vienna: Physica Verlag.

Eckart, C., e G. Young. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, .

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58, .

Gifi, A. 1985. *PRINCALS. Research Report UG-85-02*. Leiden: Department of Data Theory, University of Leiden.

Gower, J. C., e J. J. Meulman. 1993. The treatment of categorical information in physical anthropology. *International Journal of Anthropology*, 8, .

Heiser, W. J., e J. J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. In: *Correspondence Analysis in the Social Sciences: Recent Developments and Applications,* M. Greenacre, e J. Blasius, ed. New York: Academic Press.

Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. In: *International Encyclopedia of Statistics,* W. H. Kruskal, e J. M. Tanur, ed. New York: The Free Press.

Kruskal, J. B., e R. N. Shepard. 1974. A nonmetric variety of linear factor analysis. *Psychometrika*, 39, .

Meulman, J. J. 1993. Principal coordinates analysis with optimal transformations of the variables: Minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46, .

Meulman, J. J., e P. Verboon. 1993. Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika*, 58, .

Meulman, J. J., A. J. Van der Kooij, e A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. In: *Classification, Automation and New Media,* W. Gaul, e G. Ritter, ed. Berlin: Springer-Verlag.

Meulman, J. J., A. J. Van der Kooij, e W. J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, ed. Thousand Oaks, Calif.: Sage Publications, Inc..

Theunissen, N. C. M., J. J. Meulman, A. L. Den Ouden, H. M. Koopman, G. H. Verrips, S. P. Verloove-Vanhorick, e J. M. Wit. 2003. Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology*, 4, .

Tucker, L. R. 1960. Intra-individual and inter-individual multidimensionality. In: *Psychological Scaling: Theory & Applications,* H. Gulliksen, e S. Messick, ed. New York: John Wiley and Sons.

Vlek, C., e P. J. Stallen. 1981. Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance*, 28, .

Wagenaar, W. A. 1988. *Paradoxes of gambling behaviour*. London: Lawrence Erlbaum Associates, Inc.

Young, F. W., Y. Takane, e J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, .

Zeijl, E., Y. te Poel, M. du Bois-Reymond, J. Ravesloot, e J. J. Meulman. 2000. The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research*, 32, .

Capitolo
11

Analisi della correlazione canonica non lineare (OVERALS)

Lo scopo dell'analisi della correlazione canonica non lineare è determinare la similarità reciproca di due o più insiemi di variabili. Come nell'analisi della correlazione canonica lineare, l'obiettivo è quello di spiegare la maggior parte dei valori di varianza osservati nelle relazioni tra gli insiemi in uno spazio dimensionale ridotto. Diversamente dall'analisi della correlazione canonica lineare, tuttavia, l'analisi della correlazione canonica non lineare non presume un livello di intervallo di misurazione o che le relazioni siano lineari. Un'altra importante differenza è costituita dal fatto che l'analisi della correlazione canonica non lineare determina la similarità tra gli insiemi confrontando contemporaneamente le combinazioni lineari delle variabili di ogni insieme con un insieme sconosciuto di punteggi degli oggetti.

Esempio un'analisi dei risultati dell'indagine

Gli esempi di questo capitolo derivano da un'indagine (Verdegaal, 1985). Sono state registrate le risposte di quindici soggetti a otto variabili. Le variabili, le etichette delle variabili e le etichette di valore (categorie) dell'insieme di dati sono visualizzate nella seguente tabella.

Tabella 11-1
Dati dell'indagine

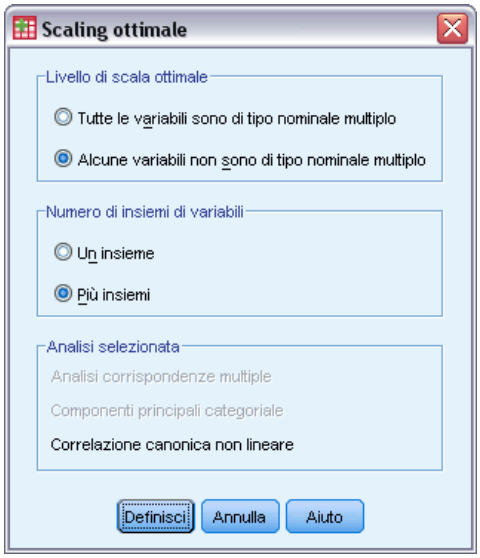
Nome di variabile	Etichetta della variabile	Etichetta del valore
età	*Età in anni*	20–25, 26–30, 31–35, 36–40, 41–45, 46–50, 51–55, 56–60, 61–65, 66–70
statociv	*Stato civile*	Single, coniugato/a, altro
andom	*Animali domestici*	Nessuno, Gatto(i), Cane(i), Altro diverso da gatto o cane, Animali domestici vari
giornale	*Giornale letto più spesso*	Nessuno, il Corriere della Sera, la Repubblica, La Stampa, Altro
musica	*Musica preferita*	Classica, New wave, Popolare, Varietà, Non ama la musica
vicinato	*Preferenze vicinato*	Città, Paese, Campagna
mate	*Punteggio test matematico*	0–5, 6–10, 11–15
lingua	*Punteggio test linguistico*	0–5, 6–10, 11–15, 16–20

Questi insiemi di dati sono reperibili nel file *verd1985.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*. Le variabili di interesse sono le prime sei variabili e sono divise in tre insiemi. L'insieme 1 include *età* e *statociv*, l'insieme 2 include *andom* e *giornale* e l'insieme 3 include *musica* e *vicinato*. *Andom* viene scalata come nominale multipla ed *età* come ordinale; tutte le altre variabili vengono scalate come nominali singole. L'analisi richiede una configurazione iniziale casuale. Per impostazione predefinita, la configurazione iniziale è numerica. Tuttavia, quando alcune variabili vengono elaborate come nominale singola senza possibilità di ordinamento,è consigliabile scegliere una configurazione iniziale casuale. È il caso della maggioranza delle variabili di questo studio.

Esame dei dati

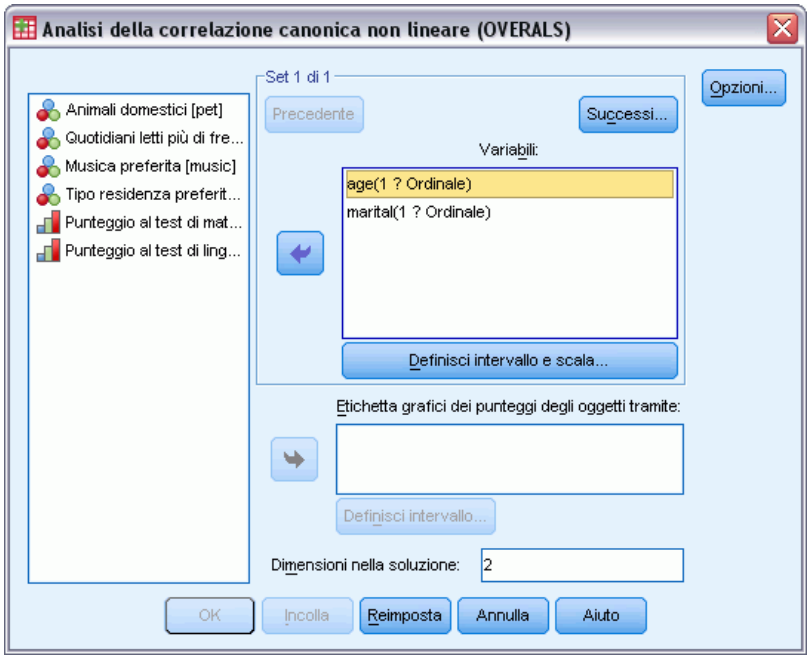
- Per ottenere un'analisi della correlazione canonica non lineare per questo insieme di dati, dai menu scegliere:
  Analizza > Riduzione dimensionale > Scaling ottimale...

Figura 11-1
Finestra di dialogo Scaling ottimale



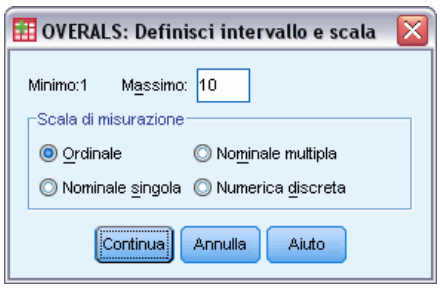
- Selezionare Una o più variabili non nominali multiple nel gruppo Livello di scaling ottimale.
- Selezionare Più insiemi nel gruppo Numero di insiemi di variabili.
- Fare clic su Definisci.

Figura 11-2
Finestra di dialogo Analisi della correlazione canonica non lineare



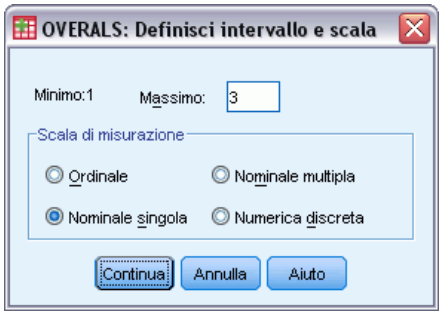
- Selezionare *Età in anni* e *Stato civile* come variabili per il primo insieme.
- Selezionare *età* e fare clic su Definisci intervallo e scala.

Figura 11-3
Finestra di dialogo Definisci intervallo e scala



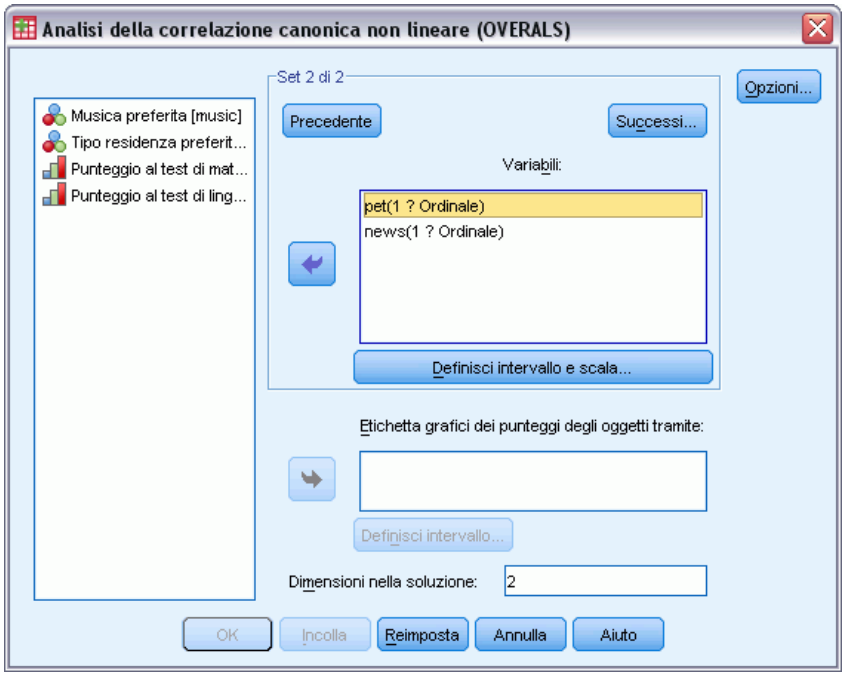
- Digitare 10 come valore massimo per questa variabile.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della correlazione canonica non lineare, selezionare *statociv* e fare clic su Definisci intervallo e scala.

Figura 11-4
Finestra di dialogo Definisci intervallo e scala



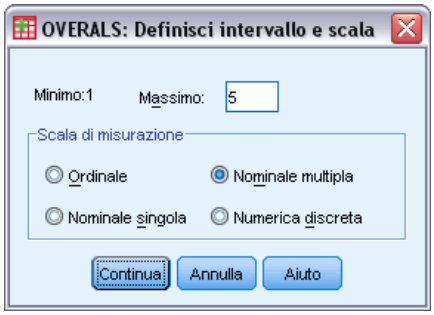
- Digitare 3 come valore massimo per questa variabile.
- Selezionare Nominale singola come scala di misurazione.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della correlazione canonica non lineare, fare clic su Avanti per definire l'insieme di variabili successivo.

Figura 11-5
Finestra di dialogo Analisi della correlazione canonica non lineare



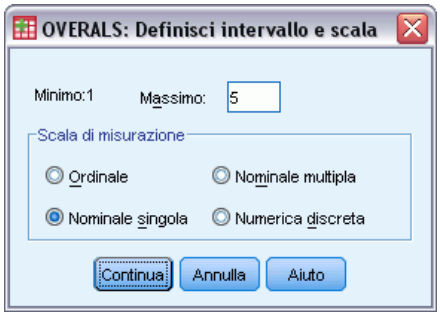
- Selezionare *Animali domestici* e *Giornale letto più spesso* come variabili per il secondo insieme.
- Selezionare *andom* e fare clic su Definisci intervallo e scala.

Figura 11-6
Finestra di dialogo Definisci intervallo e scala



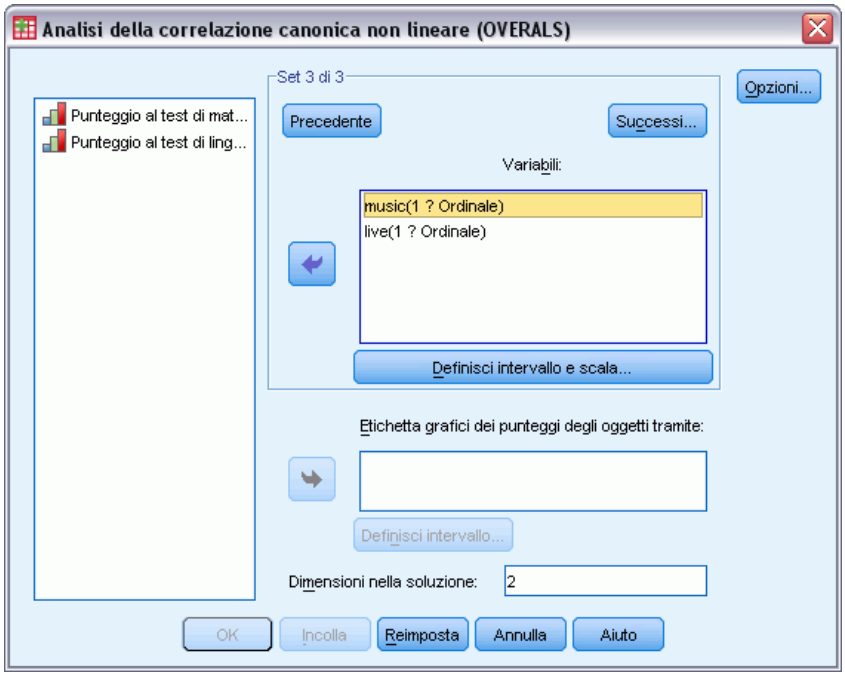
- Digitare 5 come valore massimo per questa variabile.
- Selezionare Nominale multipla come scala di misurazione.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della correlazione canonica non lineare, selezionare *giornale* e fare clic su Definisci intervallo e scala.

Figura 11-7
Finestra di dialogo Definisci intervallo e scala



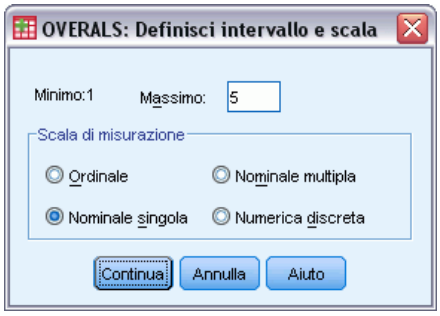
- Digitare 5 come valore massimo per questa variabile.
- Selezionare Nominale singola come scala di misurazione.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della correlazione canonica non lineare, fare clic su Avanti per definire l'ultimo insieme di variabili.

Figura 11-8
Finestra di dialogo Analisi della correlazione canonica non lineare



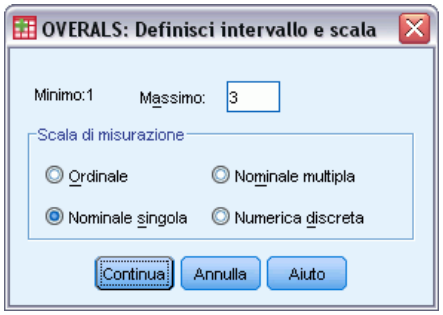
- Selezionare *Musica preferita* e *Preferenze vicinato* come variabili per il terzo insieme.
- Selezionare *musica* e fare clic su Definisci intervallo e scala.

Figura 11-9
Finestra di dialogo Definisci intervallo e scala



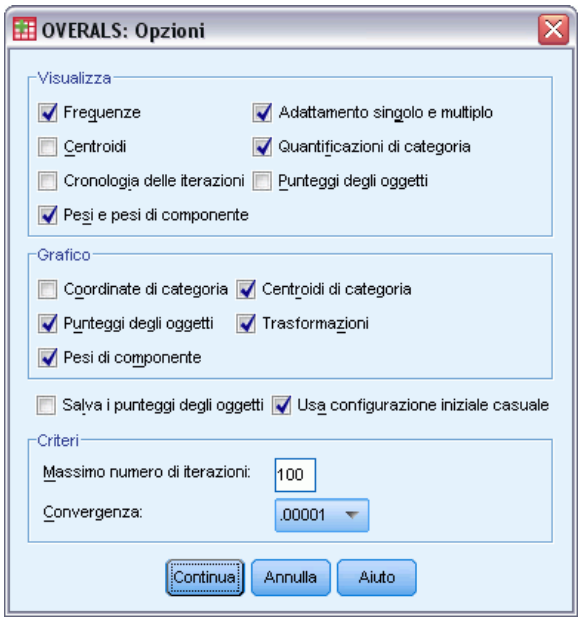
- Digitare 5 come valore massimo per questa variabile.
- Selezionare Nominale singola come scala di misurazione.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della correlazione canonica non lineare, selezionare *vicinato* e fare clic su Definisci intervallo e scala.

Figura 11-10
Finestra di dialogo Definisci intervallo e scala



- Digitare 3 come valore massimo per questa variabile.
- Selezionare Nominale singola come scala di misurazione.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della correlazione canonica non lineare, fare clic su Opzioni.

Figura 11-11
Finestra di dialogo Opzioni



- Deselezionare Centroidi e selezionare Pesi e pesi di componente nel gruppo Visualizza.
- Selezionare Centroidi di categoria e Trasformazioni nel gruppo Grafici.
- Selezionare Usa configurazione iniziale casuale.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della correlazione canonica non lineare, fare clic su OK.

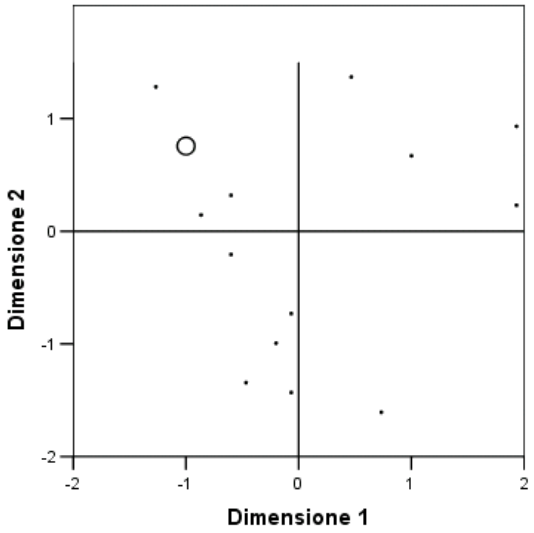
Dopo un elenco delle variabili con i relativi livelli di scaling ottimale, l'analisi della correlazione canonica categoriale con scaling ottimale genera una tabella che mostra le frequenze degli oggetti nelle categorie. Questa tabella è particolarmente importante in presenza di dati mancanti, in quanto ci sono maggiori probabilità che le categorie quasi vuote dominino la soluzione. In questo esempio non ci sono dati mancanti.

Una seconda verifica preliminare consiste nell'esaminare il grafico dei punteggi degli oggetti alla ricerca di valori anomali. I valori anomali hanno quantificazioni diverse dagli altri oggetti, tali che si trovano ai limiti del grafico, dominando di conseguenza una o più dimensioni.

Se vengono reperiti valori anomali, è possibile gestirli in due modi. È possibile eliminarli semplicemente dai dati ed eseguire di nuovo l'analisi della correlazione canonica non lineare. In alternativa, è possibile provare a ricodificare le risposte estreme degli oggetti anomali comprimendo (unendo) alcune categorie.

Come indicato nel grafico dei punteggi degli oggetti, non ci sono valori anomali per i dati dell'indagine.

Figura 11-12
Punteggi degli oggetti



Casi pesati per il numero di oggetti.

Spiegazione della similarità tra gli insiemi

Esistono molti modi di misurare l'associazione tra insiemi in un'analisi della correlazione canonica non lineare (ciascuno dei quali viene illustrato in dettaglio in una tabella separata o in un insieme di tabelle).

Riepilogo dell'analisi

I valori di perdita e di adattamento indicano la bontà dell'adattamento della soluzione con analisi della correlazione canonica non lineare rispetto ai dati con quantificazione ottimale in relazione all'associazione tra gli insiemi. Il riepilogo della tabella di analisi mostra i valori di adattamento, i valori di perdita e gli autovalori per l'indagine di esempio.

Figura 11-13
Riepilogo dell'analisi

		Dimensione		
		1	2	Somma
Perdita	Insieme 1	,240	,183	,423
	Insieme 2	,184	,408	,593
	Insieme 3	,171	,205	,376
	Media	,199	,265	,464
Autovalore		,801	,735	
Adattamento				1,536

La perdita è suddivisa tra le dimensioni e gli insiemi. Per ogni dimensione e insieme, la perdita rappresenta la proporzione di variabilità nei punteggi degli oggetti che non può essere spiegata dalla combinazione ponderata delle variabili nell'insieme. La perdita media viene etichettata come Media. Nell'esempio, la perdita media negli insiemi è pari a 0,464. Si noti che per la seconda dimensione è presente una perdita maggiore rispetto alla prima.

L'autovalore per ogni dimensione è pari a 1 meno la perdita media per la dimensione e indica quanta parte della relazione viene indicata da ogni dimensione. Gli autovalori si aggiungono all'adattamento totale. Per i dati di Verdegaal, 0,801/1,536 = 52% dell'adattamento effettivo viene spiegato dalla prima dimensione.

Il valore di adattamento massimo è pari al numero delle dimensioni e, se ottenuto, indica che la relazione è perfetta. Il valore di perdita media negli insiemi e nelle dimensioni indica la differenza tra l'adattamento massimo e quello reale. L'adattamento più la perdita media è pari al numero delle dimensioni. Una similarità perfetta si verifica raramente e in genere riguarda aspetti trascurabili dei dati.

Un'altra statistica diffusa relativa a due insiemi di variabili è la correlazione canonica. Poiché la correlazione canonica è correlata all'autovalore e di conseguenza non fornisce informazioni aggiuntive, non viene inclusa nell'output dell'analisi della correlazione canonica non lineare. Per due insiemi di variabili, la correlazione canonica per dimensione si ottiene dalla seguente formula:

$$\rho_d = 2 \times E_d - 1$$

dove *d* è il numero delle dimensioni e *E* è l'autovalore.

È possibile generalizzare la correlazione canonica per più di due insiemi attraverso la seguente formula:

$$\rho_d = ((K \times E_d) - 1)/(K - 1)$$

dove *d* è il numero delle dimensioni, *K* è il numero degli insiemi ed *E* è l'autovalore. Nell'esempio,

$$\rho_1 = ((3 \times 0.801) - 1)/2 = 0.702$$

e

$$\rho_2 = ((3 \times 0.735) - 1)/2 = 0.603$$

Pesi e pesi di componente (Categories)

Un'altra misura dell'associazione è la correlazione multipla tra combinazioni lineari da ogni insieme e da punteggi degli oggetti. Qualora nessuna variabile di un insieme sia nominale multipla, è possibile calcolare la misura moltiplicando il peso e il peso di componente di ciascuna variabile all'interno dell'insieme, sommando questi prodotti e calcolando la radice quadrata della somma.

Figura 11-14
Pesi

Insieme		Dimensione 1	Dimensione 2
1	Età in anni	,680	,789
	Stato civile	,296	-1,016
2	Quotidiani letti più spesso	-,845	-,361
3	Musica preferita	,631	-,749
	Preferenze abitative	-,484	-,780

Figura 11-15
Pesi di componente

Insieme			Dimensione 1	Dimensione 2
1	Età in anni[a,b]		,834	,259
	Stato civile[c,b]		,651	-,604
2	Animali domestici[d,e]	Dimensione 1	,397	-,431
		2	-,277	,680
	Quotidiani letti più spesso[c,b]		-,667	-,391
3	Musica preferita[c,b]		,786	-,500
	Preferenze abitative[c,b]		-,687	-,540

a. Livello di scaling ottimale: Ordinale
b. Proiezioni delle variabili quantificate singole nell'area dell'oggetto
c. Livello di scaling ottimale: Nominale singola
d. Livello di scaling ottimale: Nominale multipla
e. Proiezioni delle variabili quantificate multiple nell'area dell'oggetto

Questi dati forniscono i pesi e i pesi di componente delle variabili dell'esempio. La correlazione multipla (*R*) per la prima somma ponderata di variabili con scaling ottimale (*Età in anni* e *Stato civile*) con la prima dimensione dei punteggi di oggetti è la seguente:

$$\begin{aligned} R &= \sqrt{(0.701 \times 0.841 + (-0.273 \times -0.631))} \\ &= \sqrt{(0.5895 + 0.1723)} \\ &= 0.873 \end{aligned}$$

Per ogni dimensione, 1 – perdita = R^2. Ad esempio, dalla tabella Riepilogo dell'analisi 1 – 0,238 = 0,762, pari a 0,873 elevato al quadrato (tenendo conto di un certo grado di errore di arrotondamento). Di conseguenza, valori di perdita limitati indicano elevate correlazioni multiple tra le somme ponderate delle variabili e delle dimensioni con scaling ottimale. I pesi non sono univoci per le variabili nominali multiple. Per le variabili nominali multiple, utilizzare 1 – perdita per insieme.

Ripartizione dell'adattamento e perdita

La perdita di ogni insieme viene ripartita dall'analisi della correlazione canonica non lineare in vari modi. La tabella di adattamento presenta le tabelle di adattamento multiplo, di adattamento singolo e di perdita singola generate dall'analisi della correlazione canonica non lineare per l'indagine di esempio. Si noti che l'adattamento multiplo meno l'adattamento singolo è pari alla perdita singola.

Figura 11-16
Ripartizione dell'adattamento e perdita

		Adattamento multiplo			Adattamento singolo			Perdita singola		
		Dimensione			Dimensione			Dimensione		
Insieme		1	2	Somma	1	2	Somma	1	2	Somma
1	Età in anni[a]	,494	,676	1,170	,462	,622	1,085	,032	,054	,085
	Stato civile[b]	,089	1,033	1,122	,088	1,033	1,120	,001	,000	,001
2	Animali domestici[c]	,402	,439	,841						
	Quotidiani letti più spesso[b]	,724	,187	,911	,714	,130	,844	,010	,057	,067
3	Musica preferita[b]	,421	,577	,998	,398	,561	,960	,022	,016	,039
	Preferenze abitative[b]	,234	,609	,843	,234	,608	,843	,000	,000	,000

a. Livello di scaling ottimale: Ordinale

b. Livello di scaling ottimale: Nominale singola

c. Livello di scaling ottimale: Nominale multipla

La perdita singola indica la perdita risultante dalla limitazione delle variabili a un insieme di quantificazioni (ovvero, nominale singola, ordinale o nominale). Se la perdita singola è elevata, è preferibile trattare le variabili come nominali multiple. Nell'esempio, tuttavia, l'adattamento singolo e multiplo sono pressoché uguali, il che significa che le coordinate multiple si trovano quasi su una linea retta nella direzione indicata dai pesi.

L'adattamento multiplo è pari alla varianza delle coordinate della categoria multipla per ciascuna variabile. Queste misure sono analoghe alle misure di discriminazione rilevate nell'analisi dell'omogeneità. È possibile esaminare la tabella dell'adattamento multiplo per verificare quali variabili comportano la migliore discriminazione. Ad esempio, si veda la tabella dell'adattamento multiplo per *Stato civile* e *Giornale letto più spesso*. I valori di adattamento, sommati nelle due dimensioni, sono pari a 1,122 per *Stato civile* e a 0,911 per *Giornale letto più spesso*. Questa informazione indica che lo stato civile di una persona fornisce un potere di discriminazione maggiore rispetto alle sue preferenze di lettura.

L'adattamento singolo corrisponde al peso quadrato per ogni variabile ed è pari alla varianza delle coordinate della categoria singola. Di conseguenza, i pesi sono pari alle deviazioni standard delle coordinate della categoria singola. Esaminando la ripartizione dell'adattamento singolo tra le dimensioni, è possibile vedere che la variabile *Giornale letto più spesso* comporta una discriminazione principalmente nella prima dimensione e che la variabile *Stato civile* comporta una discriminazione pressoché totale nella seconda. In altre parole, le categorie di *Giornale letto*

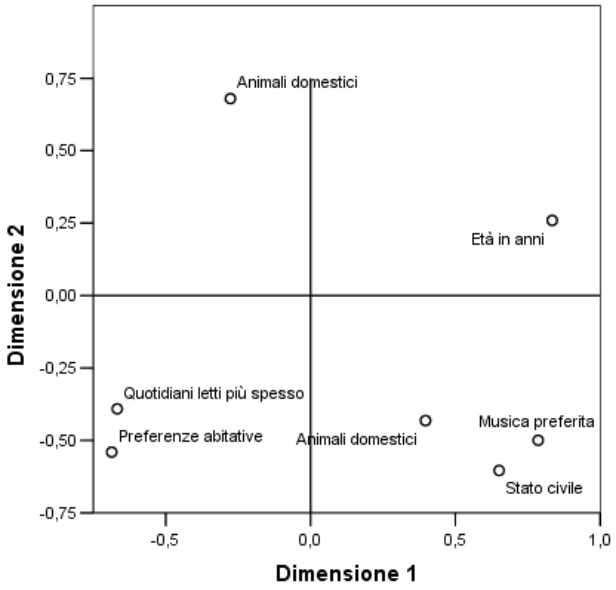
più spesso sono maggiormente separate nella prima dimensione rispetto alla seconda, mentre il modello è invertito per *Stato civile*. Per contro, *Età in anni* comporta una discriminazione nella prima e nella seconda dimensione; di conseguenza la distribuzione delle categorie è uguale in entrambe le dimensioni.

Pesi di componente

La figura seguente mostra il grafico dei pesi di componente per i dati dell'indagine. In assenza di dati mancanti, i pesi di componente sono equivalenti alle correlazioni di Pearson tra le variabili quantificate e i punteggi degli oggetti.

La a distanza dall'origine a ogni punto di variabile è approssimativamente pari all'importanza di tale variabile. Le variabili canoniche non sono inserite nel grafico ma possono essere rappresentate tramite linee verticali e orizzontali tracciate a partire dall'origine.

Figura 11-17
Pesi di componente



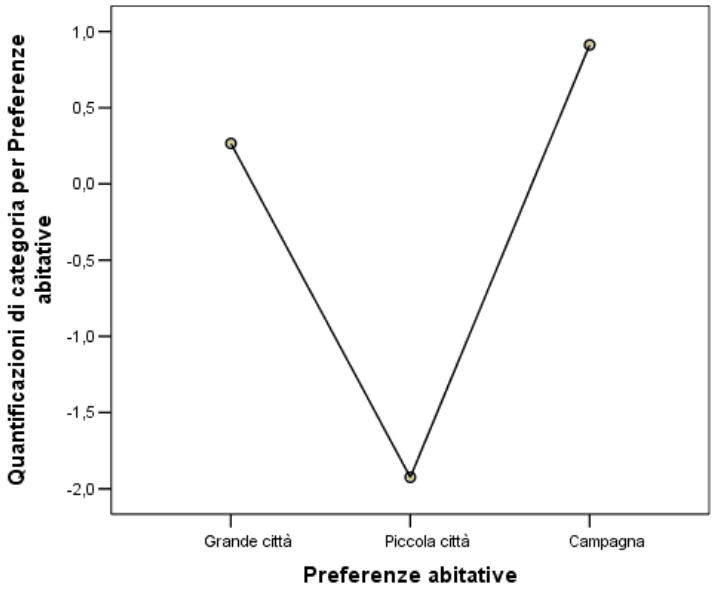
Le relazioni tra le variabili sono evidenti. Esistono due direzioni che non coincidono con gli assi verticale e orizzontale. Una direzione è determinata da *Età in anni*, *Giornale letto più spesso* e *Preferenze vicinato*. L'altra direzione è definita dalle variabili *Stato civile*, *Musica preferita* e *Animali domestici*. La variabile *Animali domestici* è una variabile nominale multipla, quindi per essa sono inseriti nel grafico due punti. Ogni quantificazione viene interpretata come una variabile singola.

Grafici di trasformazione

I diversi livelli di scaling di ciascuna variabile determinano l'applicazione di vincoli alle quantificazioni. I grafici di trasformazione illustrano la relazione tra le quantificazioni e le categorie originali risultanti dal livello di scaling ottimale selezionato.

Il grafico di trasformazione per *Preferenze vicinato*, trattata come nominale, visualizza un modello con forma a U, nel quale la categoria centrale riceve la quantificazione minore e le categorie alle estremità valori simili tra loro. Questo modello indica una relazione quadratica tra la variabile originale e la variabile trasformata. L'utilizzo di un livello di scaling ottimale alternativo non è consigliabile per *Preferenze vicinato*.

Figura 11-18
Grafico di trasformazione per Preferenze vicinato (nominale)



Le quantificazioni per *Giornale letto più spesso*, per contro, corrispondono a un trend crescente nelle tre categorie con casi osservati. La prima categoria riceve la quantificazione minore, la seconda un valore maggiore e la terza il valore massimo. Sebbene la variabile venga scalata come nominale, l'ordine delle categorie viene recuperato nelle quantificazioni.

Figura 11-19
Grafico di trasformazione per Giornale letto più spesso (nominale)

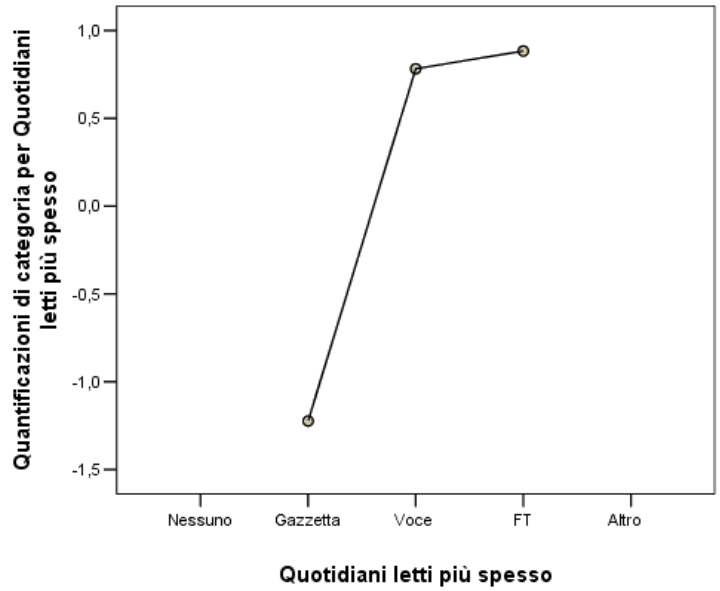
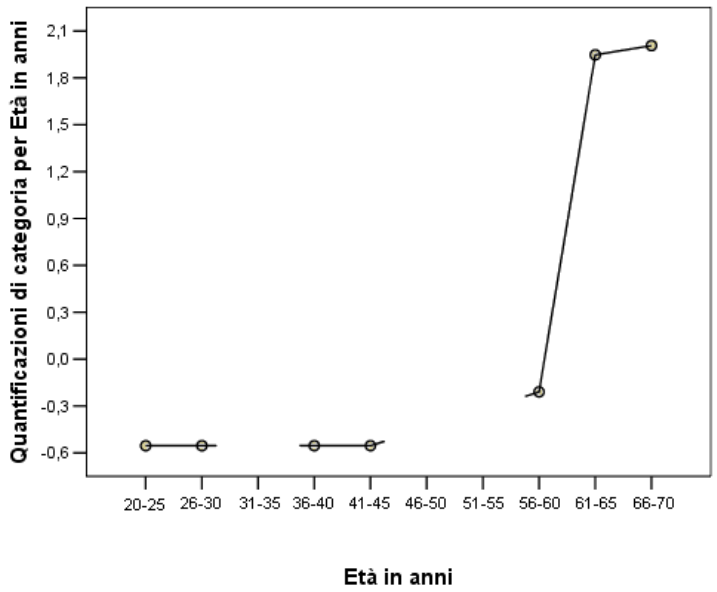


Figura 11-20
Grafico di trasformazione per Età in anni (ordinale)



Il grafico di trasformazione per *Età in anni* mostra una curva a forma di S. Le quattro categorie relative alle fasce di età più giovani osservate ricevono tutte la stessa quantificazione negativa, mentre le due categorie relative alle fasce di età più anziane ricevono valori positivi analoghi. Di conseguenza, è possibile tentare la compressione di tutte le età più giovani in una categoria comune (ovvero, di età inferiore a 50 anni) e la compressione delle due categorie più anziane in una sola. Tuttavia, l'esatta uguaglianza delle quantificazioni per i gruppi più giovani indica che la limitazione dell'ordine delle quantificazioni all'ordine delle categorie originali potrebbe non essere consigliabile. Poiché le quantificazioni per i gruppi di 26–30, 36–40 e 41–45 non possono essere minori della quantificazione per il gruppo 20–25, questi valori vengono impostati come uguali al valore limite. Se si consente che questi valori siano minori della quantificazione per il gruppo più giovane (ovvero, se si tratta la variabile età come nominale), è possibile ottenere un miglioramento dell'adattamento. Di conseguenza, sebbene l'età sia considerata una variabile ordinale, trattarla come tale non sembra appropriato in questo caso. Inoltre, trattando l'età come variabile numerica e quindi mantenendo le distanze tra le categorie, si determinerebbe una significativa riduzione dell'adattamento.

Coordinate della categoria multipla vs categoria singola

Per ogni variabile trattata come nominale singola, ordinale o numerica, sono determinate le quantificazioni, le coordinate della categoria singola e le coordinate della categoria multipla. Queste statistiche sono illustrate per *Età in anni*.

Figura 11-21
Coordinate per Età in anni

	Frequenza marginale	Quantificazione	Coordinate della categoria singola		Coordinate della categoria multipla	
			Dimensione		Dimensione	
			1	2	1	2
20-25	3	-,554	-,377	-,437	-,192	-,139
26-30	5	-,554	-,377	-,437	-,404	-,623
31-35	0	,000				
36-40	1	-,554	-,377	-,437	-,318	-,733
41-45	1	-,554	-,377	-,437	-,356	-,534
46-50	0	,000				
51-55	0	,000				
56-60	2	-,209	-,142	-,165	-,435	,087
61-65	1	1,947	1,324	1,536	1,710	1,204
66-70	2	2,006	1,364	1,583	1,215	1,711
Mancante	0					

Ogni categoria per la quale non sono stati registrati casi riceve una quantificazione pari a 0. Per *Età in anni*, le categorie prive di casi includono 31–35, 46–50 e 51–55. Queste categorie non sono limitate all'ordinamento con altre categorie e non influenzano alcun calcolo.

Per le variabili nominali multiple, ogni categoria riceve una quantificazione diversa per ciascuna dimensione. Per tutti gli altri tipi di trasformazioni, una categoria ha una sola quantificazione, indipendentemente dalla dimensionalità della soluzione. Ciascun insieme di coordinate della categoria singola rappresenta la posizione delle categorie su una linea nello spazio dell'oggetto. Le coordinata di una data categoria equivalgono alla quantificazione moltiplicata per i pesi di dimensione della variabile. Ad esempio, nella tabella *Età in anni*, le coordinate della categoria

singola per la categoria 56-60 (-0,142, -0,165) sono pari alla quantificazione (-0,209) moltiplicata per i pesi di dimensione (0,680, 0,789).

Le coordinate della categoria multipla per le variabili trattate come nominali singole, ordinali o numeriche, rappresentano le coordinate delle categorie nello spazio dell'oggetto prima dell'applicazione di vincoli ordinali o lineari. Questi valori sono riduttori non vincolati della perdita. Per le variabili nominali multiple, queste coordinate rappresentano le quantificazioni delle categorie.

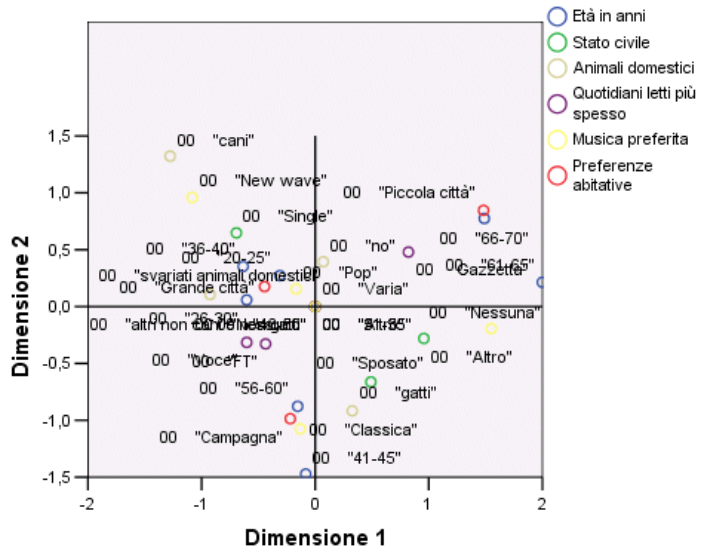
Gli effetti dell'imposizione di vincoli alla relazione tra le categorie e le relative quantificazioni si evidenziano confrontando le coordinate della categoria singola con quelle della categoria multipla. Nella prima dimensione, le coordinate della categoria multipla per *Età in anni* si riducono fino alla categoria 2 e rimangono relativamente allo stesso livello fino alla categoria 9, in corrispondenza della quale si verifica un significativo aumento. Un modello simile viene evidenziato per la seconda dimensione. Queste relazioni vengono rimosse nelle coordinate della categoria singola, cui è applicato il vincolo ordinale. In entrambe le dimensioni, le coordinate sono ora non decrescenti. La diversa struttura dei due insiemi di coordinate suggerisce che un trattamento nominale potrebbe essere più appropriato.

Centroidi e centroidi proiettati

Il grafico dei centroidi etichettati in base alle variabili dovrebbe essere interpretato in modo analogo al grafico delle quantificazioni di categoria nell'analisi dell'omogeneità o alle coordinate della categoria multipla nell'analisi componenti principali non lineare. Di per se stesso, tale grafico mostra il grado di separazione tra i gruppi di oggetti a opera delle variabili (i centroidi si trovano nel centro di gravità degli oggetti).

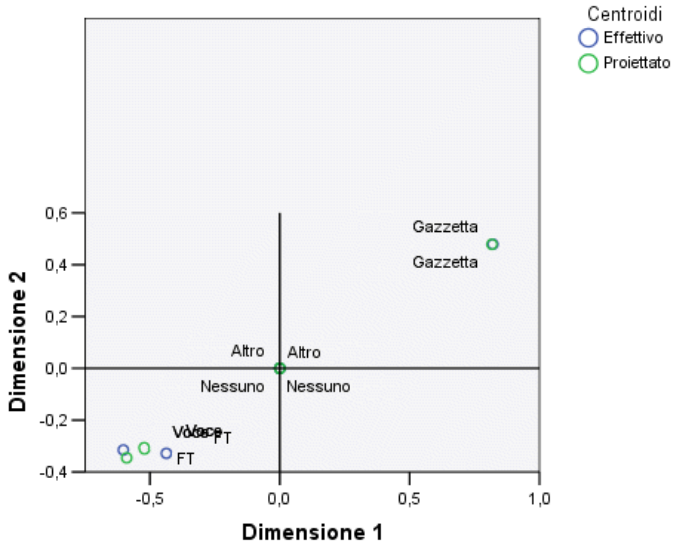
Si noti che le categorie per *Età in anni* non sono separate molto nettamente. Le categorie relative a fasce di età più giovane sono raggruppate a sinistra del grafico. Come suggerito in precedenza, quello ordinale potrebbe essere un livello di scaling ordinale troppo rigido da applicare a *Età in anni*.

Figura 11-22
Centroidi etichettati in base a variabili



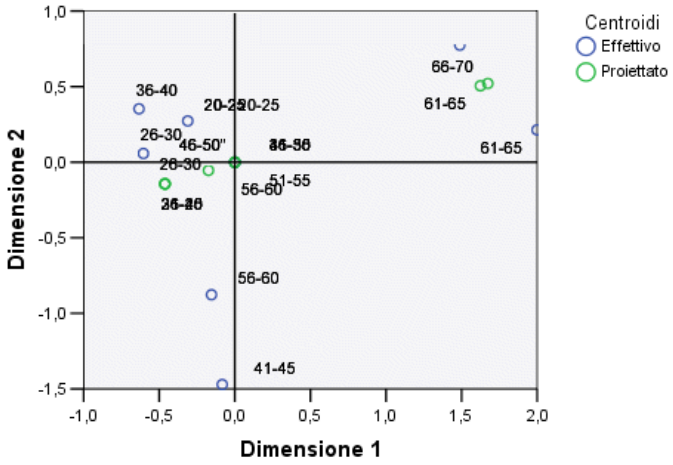
Quando si richiede il grafico dei centroidi, vengono generati anche i grafici dei singoli centroidi e dei centroidi proiettati per ogni variabile etichettata in base alle etichette dei valori. I centroidi proiettati si trovano su una linea nello spazio dell'oggetto.

Figura 11-23
Centroidi e centroidi proiettati per Giornale letto più spesso



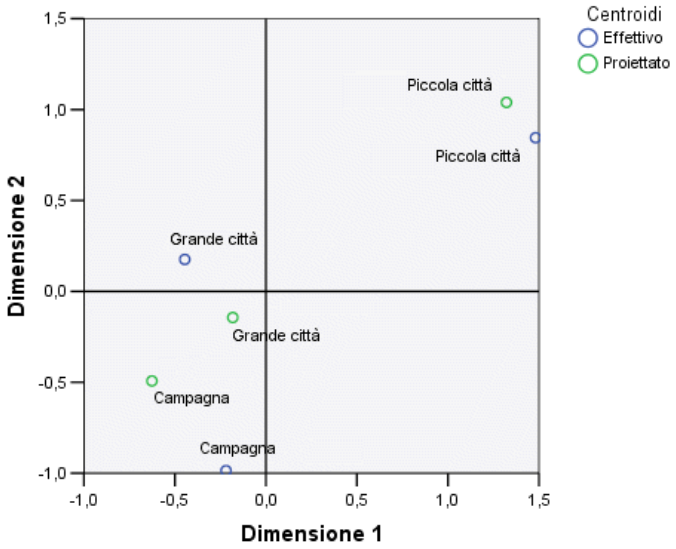
I centroidi reali sono proiettati sui vettori definiti dai pesi di componente. Questi vettori sono stati aggiunti ai grafici dei centroidi per semplificare la distinzione tra centroidi proiettati e reali. I centroidi proiettati sono compresi in uno dei quattro quadranti formati prolungando le due linee di riferimento perpendicolari fino all'origine. L'interpretazione della direzione delle variabili nominali singole, ordinali o numeriche viene ottenuta dalla posizione dei centroidi proiettati. Ad esempio, la variabile *Giornale letto più spesso* è specificata come nominale singola. I centroidi proiettati mostrano che *La Repubblica* e *La Stampa* sono in contrasto con *Corriere della Sera*.

Figura 11-24
Centroidi e centroidi proiettati per Età in anni



Il problema con *Età in anni* è evidente dai centroidi proiettati. Il trattamento di *Età in anni* come ordinale, implica che l'ordine dei gruppi di età debba essere preservato. Per soddisfare questo vincolo, tutti i gruppi di età inferiori a 45 anni sono proiettati sullo stesso punto. Lungo la direzione definita da *Età in anni*, *Giornale letto più spesso* e *Preferenze vicinato*, non esiste separazione dei gruppi di età più giovane. Questo risultato suggerisce che la variabile debba essere trattata come nominale.

Figura 11-25
Centroidi e centroidi proiettati per Preferenze vicinato



Per comprendere le relazioni tra le variabili, individuare le categorie specifiche (valori) per i cluster di categorie nei grafici dei centroidi. Le relazioni tra *Età in anni*, *Giornale letto più spesso* e *Preferenze vicinato*, possono essere descritte esaminando la parte superiore destra e inferiore sinistra dei grafici. Nella parte superiore destra si trovano i rispondenti più anziani, che leggono il Corriere della Sera e preferiscono vivere in un paese. Nell'angolo inferiore sinistro di ciascun grafico, è possibile vedere che i rispondenti di mezza età e più giovani leggono La Repubblica o La Stampa e preferiscono vivere in campagna o in una città. Tuttavia, la separazione tra i gruppi più giovani è alquanto complessa.

Gli stessi tipi di interpretazione possono essere effettuati in relazione all'altra direzione (*Stato civile*, *Musica preferita* e *Animali domestici*), concentrandosi sulla parte superiore sinistra e inferiore destra dei grafici dei centroidi. Nell'angolo superiore sinistro, è possibile notare che i single tendono ad avere un cane e a preferire la musica new wave. Le persone sposate e con stato civile diverso hanno un gatto; il primo gruppo preferisce la musica classica e il secondo non ama la musica.

Un'analisi alternativa

I risultati dell'analisi suggeriscono che il trattamento di *Età in anni* come ordinale non sia adeguato. Sebbene *Età in anni* sia misurata a livello ordinale, le sue relazioni con altre variabili non sono monotone. Per esaminare gli effetti della modifica del livello di scaling ottimale in nominale singolo, ripetere l'analisi.

Per eseguire l'analisi

- ▶ Richiamare la finestra di dialogo Analisi della correlazione canonica non lineare e accedere al primo insieme.
- ▶ Selezionare *età* e fare clic su Definisci intervallo e scala.
- ▶ Nella finestra di dialogo Definisci intervallo e scala, selezionare Nominale singola come intervallo di scala.
- ▶ Fare clic su Continua.
- ▶ Nella finestra di dialogo Analisi della correlazione canonica non lineare, fare clic su OK.

Gli autovalori per una soluzione a due dimensioni sono pari rispettivamente a 0,806 e 0,757, con adattamento totale pari a 1,564.

Figura 11-26
Autovalori per una soluzione a due dimensioni

		Dimensione		
		1	2	Somma
Perdita	Insieme 1	,249	,115	,363
	Insieme 2	,176	,408	,584
	Insieme 3	,157	,205	,363
	Media	,194	,243	,436
Autovalore		,806	,757	
Adattamento				1,564

Le tabelle di adattamento singolo e multiplo mostrano che *Età in anni* continua a essere una variabile a elevata discriminazione, come evidenziato dalla somma dei valori di adattamento multiplo. In contrasto con i risultati precedenti, tuttavia, l'esame dei valori di adattamento singolo rivela che la discriminazione appartiene quasi totalmente alla seconda dimensione.

Figura 11-27
Ripartizione dell'adattamento e perdita

		Adattamento multiplo			Adattamento singolo			Perdita singola		
		Dimensione			Dimensione			Dimensione		
Insieme		1	2	Somma	1	2	Somma	1	2	Somma
1	Età in anni[a]	,246	1,197	1,443	,195	1,188	1,384	,051	,008	,059
	Stato civile[a]	,273	1,136	1,409	,272	1,135	1,407	,001	,000	,002
2	Animali domestici[b]	,530	,392	,921						
	Quotidiani letti più spesso[a]	,639	,185	,824	,631	,149	,780	,008	,036	,044
3	Musica preferita[a]	,604	,438	1,041	,603	,437	1,040	,000	,001	,001
	Preferenze abitative[a]	,075	,822	,897	,075	,822	,897	,000	,000	,000

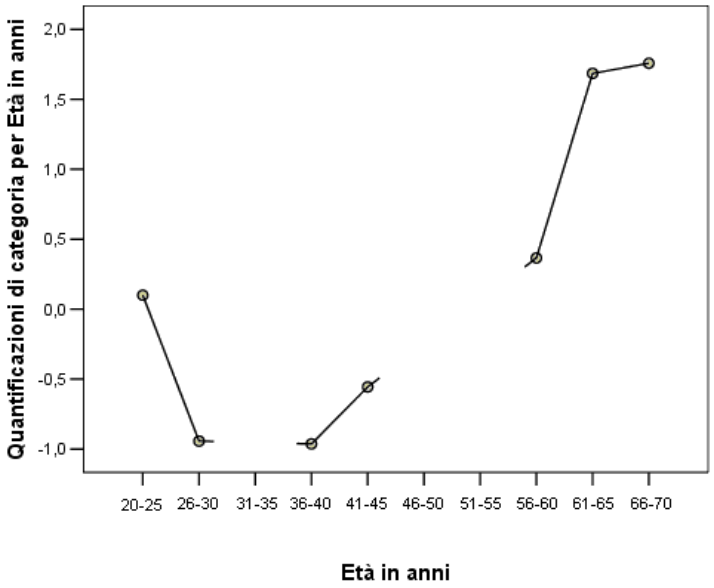
a. Livello di scaling ottimale: Nominale singola

b. Livello di scaling ottimale: Nominale multipla

Tornare al grafico di trasformazione per *Età in anni*. Le quantificazioni per una variabile nominale sono non vincolate, quindi il trend non decrescente visualizzato quando *Età in anni* è stata trattata originariamente non è più presente. Esiste quindi un trend non decrescente fino all'età di 40 anni e un trend crescente a partire da quell'età, corrispondente a una relazione a forma di U

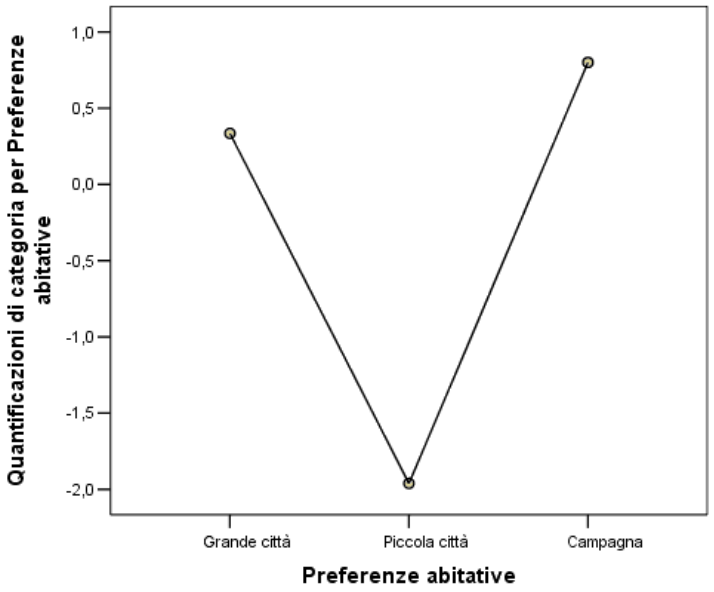
(quadratica). Le due categorie più anziane ricevono ancora punteggi simili ed eventuali analisi possono includere la combinazione di queste categorie.

Figura 11-28
Grafico di trasformazione per Età in anni (nominale)



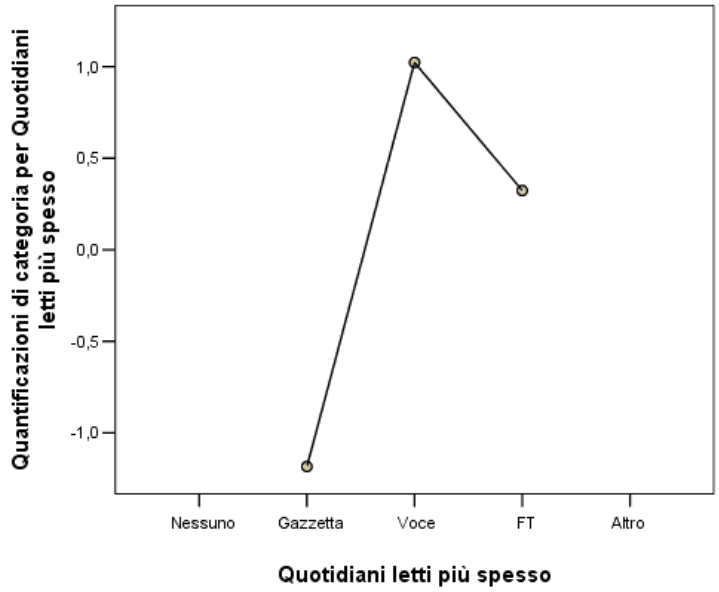
Il grafico di trasformazione per *Preferenze vicinato* viene illustrato di seguito. Il trattamento di *Età in anni* come nominale non influenza le quantificazioni per *Preferenze vicinato* da nessun punto di vista significativo. La categoria centrale riceve la quantificazione minima, quelle alle estremità elevati valori positivi.

Figura 11-29
Grafico di trasformazione per Preferenze vicinato (età nominale)



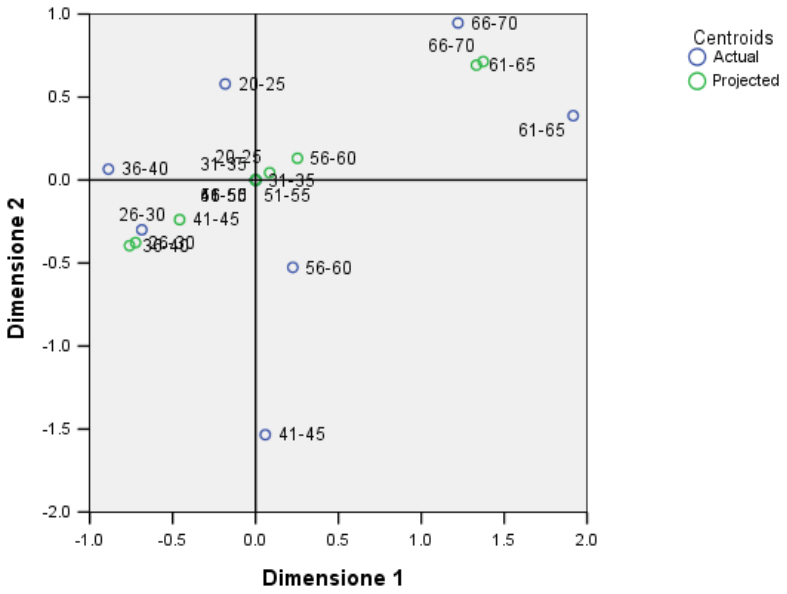
Nel grafico di trasformazione per *Giornale letto più spesso* viene rilevata una variazione. In precedenza, era presente un trend crescente nelle quantificazioni, che poteva suggerire un trattamento ordinale per questa variabile. Tuttavia, il trattamento di *Età in anni* come nominale rimuove tale trend dalle quantificazioni relative a giornale.

Figura 11-30
Grafico di trasformazione per Giornale letto più spesso (nominale)



Quello visualizzato è il grafico dei centroidi per *Età in anni*. Si noti che le categorie non hanno ordine cronologico lungo la linea che congiunge i centroidi proiettati. Il gruppo 20–25 si trova nella parte centrale invece che alla fine. La distribuzione delle categorie è molto migliorata rispetto alla controparte ordinale illustrata in precedenza.

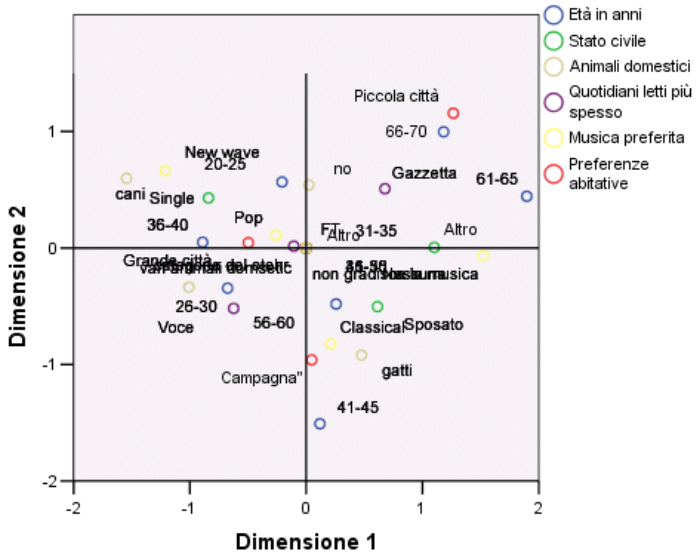
Figura 11-31
Centroidi e centroidi proiettati per Età in anni (nominale)



L'interpretazione dei gruppi di età più giovani è ora possibile dal grafico dei centroidi. Le categorie *La Repubblica* e *La Stampa* sono inoltre più distanti rispetto all'analisi precedente, il che consente l'interpretazione separata di ciascuna categoria. I gruppi di età tra 26 e 45 anni leggono la Repubblica e preferiscono vivere in campagna. I gruppi di età 20–25 e 56–60 leggono La Stampa; il primo gruppo preferisce vivere in città, il secondo in campagna. I gruppi più anziani leggono il Corriere della Sera e preferiscono vivere in un paese.

L'interpretazione dell'altra direzione (*Stato civile*, *Musica preferita* e *Animali domestici*) resta fondamentalmente invariata rispetto all'analisi precedente. L'unica differenza evidente consiste nel fatto che le persone con stato civile *Diverso* hanno gatti oppure non hanno animali domestici.

Figura 11-32
Centroidi etichettati in base a variabili (età nominale)



Suggerimenti generali

Una volta esaminati i risultati iniziali, si desidererà probabilmente perfezionare l'analisi modificando alcune delle specifiche dell'analisi della correlazione canonica non lineare. Di seguito vengono illustrati alcuni suggerimenti relativi a come definire la struttura dell'analisi:

- Creare quanti più insiemi è possibile. Inserire una variabile importante della quale si desidera prevedere il valore da sola in un insieme separato.
- Raggruppare le variabili considerate predittori in un unico insieme. Se sono presenti molti predittori, tentare di ripartirli in più insiemi.
- Inserire ciascuna variabile nominale multipla da sola in un insieme separato.
- Se tra le variabili è presente un elevato grado di correlazione e non si desidera evitare che questa relazione domini la soluzione, raggruppare tali variabili nello stesso insieme.

Letture consigliate

Consultare i testi seguenti per ulteriori informazioni sull'analisi della correlazione canonica non lineare:

Carroll, J. D. 1968. Generalization of canonical correlation analysis to three or more sets of variables. In: *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* Washington, D.C.: American Psychological Association.

De Leeuw, J. 1984. *Canonical analysis of categorical data*, 2nd ed. Leiden: DSWO Press.

Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, .

Horst, P. 1961. Relations among m sets of measures. *Psychometrika*, 26, .

Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58, .

Van der Burg, E. 1988. *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.

Van der Burg, E., e J. De Leeuw. 1983. Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, .

Van der Burg, E., J. De Leeuw, e R. Verdegaal. 1988. Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 53, .

Verboon, P., e I. A. Van der Lans. 1994. Robust canonical discriminant analysis. *Psychometrika*, 59, .

Capitolo
12

Analisi corrispondenze

Una **tabella di corrispondenza** è una tabella a due vie le cui celle contengono alcune misure della corrispondenza tra righe e colonne. La misura della corrispondenza può essere un qualsiasi indicatore di similarità, affinità, confusione, associazione o interazione tra le variabili di riga e di colonna. Un tipo molto comune di tabella di corrispondenza è la tavola di contingenza, in cui le celle contengono conteggi di frequenza.

Queste tavole possono essere ottenute facilmente grazie alla procedura Tavole di contingenza. Tuttavia, una tavola di contingenza non genera sempre un quadro chiaro della natura della relazione tra le due variabili. Questo è particolarmente vero se le variabili di interesse sono nominali (senza ordine inerente o rango) e contengono numerose categorie. La tavola di contingenza può indicare che le frequenze di celle osservate differiscono notevolmente dai valori attesi in un tavola di contingenza $10x9$ di *occupazione* e *cereali da colazione*, ma può essere difficile discernere quali gruppi occupazionali hanno gusti analoghi o quali siano questi gusti.

L'analisi delle corrispondenze consente di esaminare graficamente la relazione esistente fra due variabili nominali in uno spazio multidimensionale. Essa calcola i punteggi di righe e colonne generando grafici in base a tali punteggi. Le categorie simili tra loro sono visualizzate nel grafico vicine le une alle altre. In questo modo, è facile vedere quali categorie di una variabile sono simili tra loro o quali categorie delle due variabili sono correlate. L'analisi delle corrispondenze consente inoltre di adattare punti supplementari allo spazio definito dai punti attivi.

Se l'ordinamento delle categorie in base ai relativi punteggi non corrisponde alle aspettative o è di difficile comprensione, è possibile imporre vincoli all'ordine imponendo che i punteggi siano uguali per alcune categorie. Ad esempio, si supponga che si preveda che la variabile *Tabagismo* con le categorie *nessuno*, *lieve*, *medio* e *forte* abbia punteggi corrispondenti a questo ordinamento. Tuttavia, se l'analisi ordina le categorie *nessuno*, *lieve*, *intenso* e *medio*, imponendo che i punteggi per *intenso* e *medio* siano uguali si mantiene l'ordinamento delle categorie nei rispettivi punteggi.

L'interpretazione dell'analisi delle corrispondenze in termini di distanze dipende dal metodo di normalizzazione utilizzato. L'analisi delle corrispondenze può essere utilizzata per analizzare le differenze tra le categorie di una variabile o tra le variabili. Con la normalizzazione predefinita, essa analizza le differenze tra le variabili di riga e di colonna.

L'algoritmo dell'analisi delle corrispondenze è in grado di eseguire vari tipi di analisi. La centratura delle righe e delle colonne e l'utilizzo delle distanze chi-quadrato corrisponde all'analisi delle corrispondenze standard. Tuttavia, l'utilizzo delle opzioni di centratura alternative combinate con le distanze euclidee consente una rappresentazione alternativa di una matrice in uno spazio a ridotto numero di dimensioni.

Verranno illustrati tre esempi: il primo impiega una tabella di corrispondenza relativamente limitata e illustra i concetti legati all'analisi delle corrispondenze. Il secondo esempio illustra un'applicazione di marketing pratica. L'esempio finale utilizza una tabella delle distanze in un approccio di scaling multidimensionale.

Normalizzazione

La normalizzazione è utilizzata per distribuire l'inerzia nei punteggi sia di riga che di colonna. Alcuni aspetti della soluzione con analisi delle corrispondenze, come i singoli valori, l'inerzia per dimensione e i contributi, non cambiano nelle varie normalizzazioni. I punteggi di riga e di colonna e le loro varianze ne vengono influenzate. L'analisi delle corrispondenze include vari modi per distribuire l'inerzia. I tre più comuni includono la distribuzione dell'inerzia solo sui punteggi di riga, la distribuzione dell'inerzia solo su punteggi di colonna o la distribuzione dell'inerzia in modo simmetrico sui punteggi di riga e di colonna.

Principale per riga. Nella normalizzazione principale per riga, le distanze euclidee fra i punti di riga sono approssimazioni delle distanze chi-quadrato tra le righe della tabella di corrispondenza. I punteggi di riga sono la media ponderata dei punteggi di colonna. I punteggi di colonna sono standardizzati in modo da avere una somma ponderata delle distanze quadrate al centroide 1. Poiché questo metodo massimizza le distanze tra le categorie di riga, si consiglia di utilizzare la normalizzazione principale per riga se si è interessati principalmente a evidenziare le differenze tra le categorie della variabile di riga.

Principale per colonna. D'altro lato, se si desidera che approssimare le distanze chi-quadrato tra le colonne della tabella di corrispondenza, i punteggi di colonna dovranno essere la media ponderata dei punteggi di riga. I punteggi di riga sono standardizzati in modo da avere una somma ponderata delle distanze quadrate al centroide 1. Poiché questo metodo massimizza le distanze tra le categorie di colonna, si consiglia di utilizzarlo se si è interessati principalmente a evidenziare le differenze tra le categorie della variabile di colonna.

Simmetrico. È anche possibile trattare righe e colonne in modo simmetrico. La normalizzazione distribuisce l'inerzia in modo uniforme sui punteggi di riga e di colonna. Si noti che né le distanze tra i punti di riga né le distanze tra i punti di colonna sono approssimazioni delle distanze chi-quadrato, in questo caso. Utilizzare questo metodo se si è interessati principalmente alle differenze o alle somiglianze tra le due variabili. Questo è in genere il metodo di elezione per generare biplot.

Principale. Una quarta opzione è denominata normalizzazione principale e prevede la distribuzione dell'inerzia due volte nella soluzione, una sui punteggi di riga e una sui punteggi di colonna. Utilizzare questo metodo se si è interessati principalmente alle distanze tra i punti di riga e di colonna separatamente, ma non alle correlazioni tra punti di riga e di colonna. I biplot non sono adatti per questa opzione di normalizzazione e di conseguenza non sono disponibili se è stato specificato il metodo di normalizzazione principale.

Esempio: Percezione delle marche di caffè

Il precedente esempio riguardava una tabella di piccole dimensioni di dati ipotetici. Le applicazioni reali spesso riguardano tabelle molto più ampie. Nell'esempio, verranno utilizzati dati relativi alle immagini percepite di sei marche di caffè freddo (Kennedy, Riquier, e Sharp, 1996). Questi insiemi di dati sono reperibili nel file *coffee.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*.

Per ciascuno dei 23 attributi dell'immagine del caffè freddo, sono state selezionate tutte le marche descritte da tale attributo. Le sei marche sono indicate dalle sigle *AA*, *BB*, *CC*, *DD*, *EE* e *FF* per tutelare la confidenzialità dei dati.

Tabella 12-1
Attributi del caffè freddo

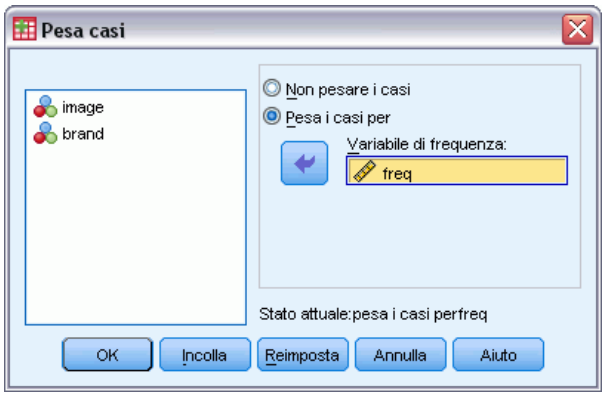
Attributo immagine	**Etichetta**	**Attributo immagine**	**Etichetta**
prodotto valido post-ubriacatura	*cura*	marca che fa ingrassare	*ingrassante*
marca a quantità ridotta di grassi/calorie	*pochi grassi*	attrae gli uomini	*uomini*
marca per bambini	*bambini*	Marca dell'Australia del sud	*Australia del Sud*
marca da classe lavoratrice	*classe lavoratrice*	marca tradizionale/vecchio stampo	*tradizionale*
gusto ricco/dolce	*dolce*	marca di alta qualità	*alta qualità*
marca non diffusa	*non diffusa*	marca salutare	*salutare*
marca per persone grasse/di aspetto sgradevole	*aspetto sgradevole*	marca ad alto contenuto di caffeina	*caffeina*
molto fresca	*fresca*	marca nuova	*nuovo*
marca per yuppie	*yuppie*	marca per persone attraenti	*attraente*
marca a elevato valore nutritivo	*nutriente*	gusto forte	*forte*
marca per donne	*donne*	marca diffusa	*diffusa*
marca minore	*minore*		

Inizialmente, l'attenzione sarà dedicata alla relazione tra gli attributi e le marche. L'utilizzo della normalizzazione principale distribuisce l'inerzia totale una volta tra le righe e una volta tra le colonne. Sebbene questo impedisca l'interpretazione biplot, è possibile esaminare le distanze tra le categorie per ogni variabile.

Esecuzione dell'analisi

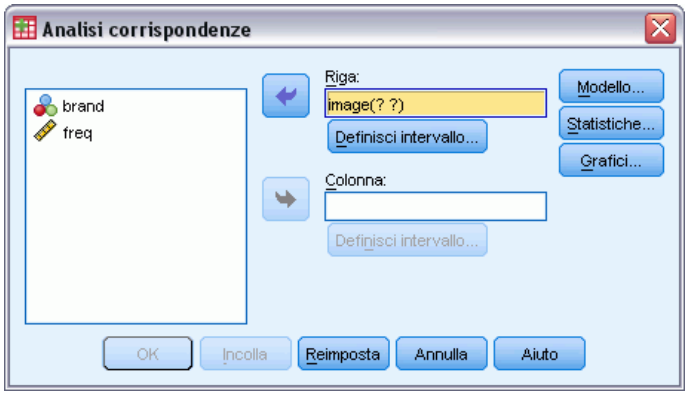
- L'impostazione dei dati richiede che i casi siano pesati tramite la variabile *freq*. Per farlo, dai menu scegliere:
  Dati > Pesa casi...

Figura 12-1
Finestra di dialogo Pesa casi



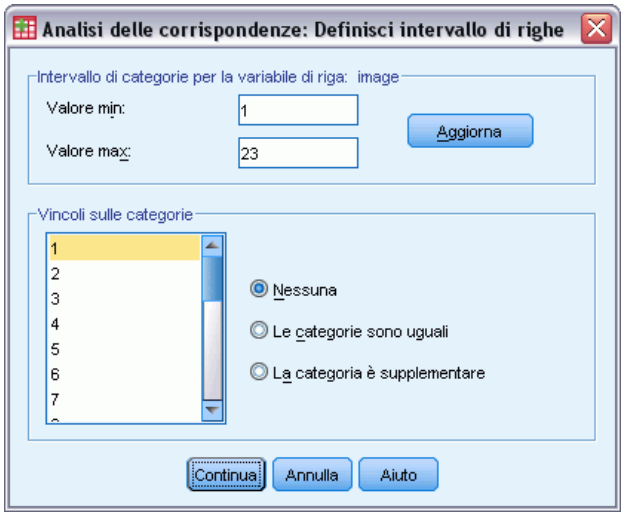
- Pesa i casi per *freq*.
- Fare clic su OK.
- Per ottenere una soluzione iniziale in cinque dimensioni con normalizzazione principale, dai menu scegliere:
  Analizza > Riduzioni dimensione > Analisi corrispondenze...

Figura 12-2
Finestra di dialogo Analisi della corrispondenze



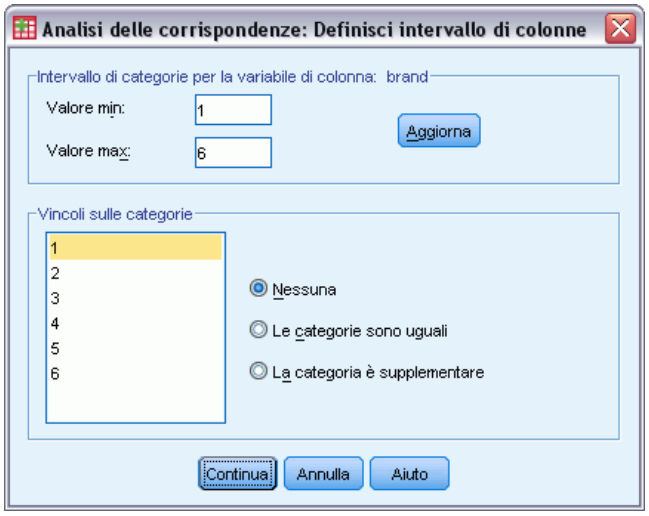
- Selezionare *immagine* come variabile di riga.
- Fare clic su Definisci intervallo.

Figura 12-3
Finestra di dialogo Definisci intervallo di righe



- Digitare 1 come valore minimo.
- Digitare 23 come valore massimo.
- Fare clic su Aggiorna.
- Fare clic su Continua.
- Selezionare *marca* come variabile di colonna.
- Fare clic su Definisci intervallo nella finestra di dialogo Analisi delle corrispondenze.

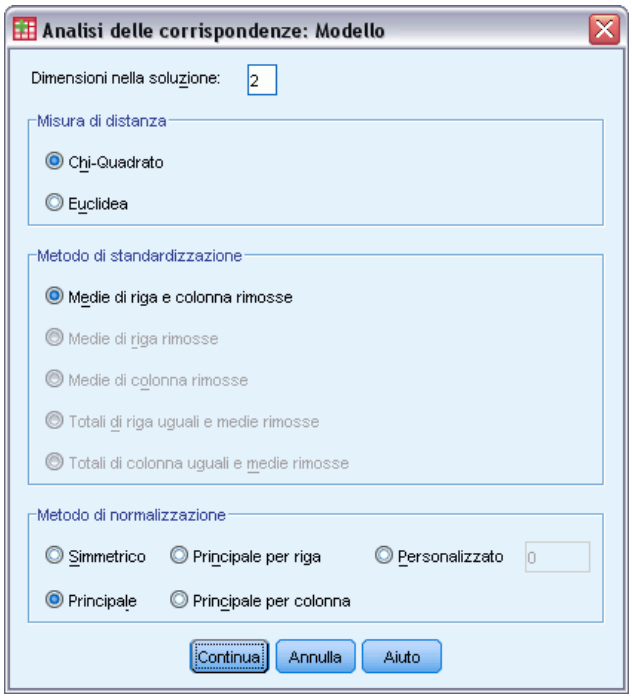
Figura 12-4
Finestra di dialogo Definisci intervallo di colonne



- Digitare 1 come valore minimo.
- Digitare 6 come valore massimo.

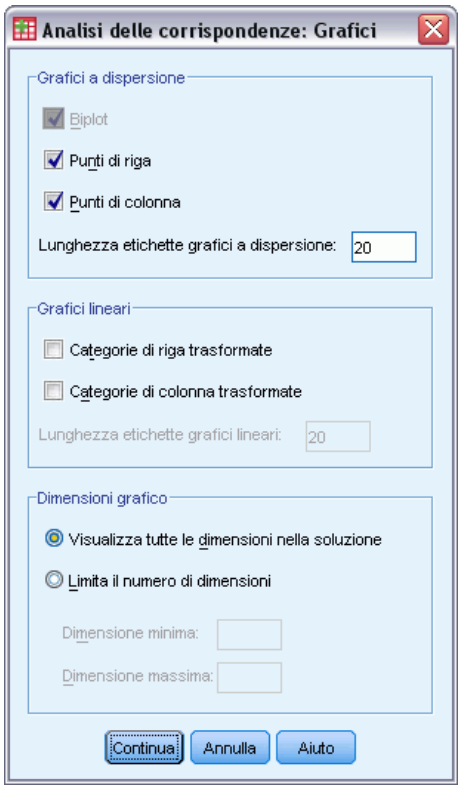
- Fare clic su Aggiorna.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della corrispondenze fare clic su Modello.

Figura 12-5
Finestra di dialogo Modello



- Selezionare Principale come metodo di normalizzazione.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della corrispondenze fare clic su Grafici.

Figura 12-6
Finestra di dialogo Grafici



- Selezionare Profili di riga e Profili di colonna nel gruppo Grafici a dispersione.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della corrispondenze fare clic su OK.

Dimensionalità

L'inerzia per dimensione mostra la scomposizione dell'inerzia totale per ogni dimensione. Due dimensioni spiegano l'83% dell'inerzia totale. Aggiungendo una terza dimensione si aggiunge solo l'8,6% di inerzia spiegata. Di conseguenza, si opta per utilizzare una rappresentazione a due dimensioni.

Figura 12-7
Inerzia per dimensione

Dimensione	Valore singolare	Inerzia	Chi-Quadrato	Sig.	Proporzione di inerzia		Confidenza del valore singolare	
					Spiegata	Cumulata	Deviazione standard	Correlazione 2
1	,711	,506			,629	,629	,009	,132
2	,399	,159			,198	,827	,014	
3	,263	,069			,086	,913		
4	,234	,055			,068	,982		
5	,121	,015			,018	1,000		
Totale		,804	3746,968	,000[a]	1,000	1,000		

a. 110 gradi di libertà

Contributi (Analisi delle corrispondenze)

La panoramica dei punti di riga mostra i contributi dei punti di riga all'inerzia delle dimensioni e i contributi delle dimensioni all'inerzia dei punti di riga. Se tutti i punti contribuissero in pari misura all'inerzia, i contributi sarebbero pari a 0,043. *Salutare* e *pochi grassi* contribuiscono entrambe in modo significativo all'inerzia della prima dimensione. *Uomini* e *forte* contribuiscono per le porzioni maggiori all'inerzia della seconda dimensione. Sia *aspetto sgradevole* che *fresca* contribuiscono in modo molto limitato a ciascuna dimensione.

Figura 12-8
Contributi degli attributi

impress	Massa	Punteggio nella dimensione		Inerzia	Contributo				
					del punto all'inerzia della dimensione		della dimensione all'inerzia del punto		
		1	2		1	2	1	2	Totale
grasso	,080	-,514	-,265	,033	,042	,035	,652	,173	,825
uomini	,051	-,852	,825	,072	,073	,219	,512	,480	,992
sud australiano	,057	-,303	-,350	,046	,010	,044	,114	,152	,266
tradizionale	,040	-,703	-,532	,043	,039	,071	,454	,260	,715
premium	,042	-,444	-,582	,028	,016	,090	,296	,509	,805
salutare	,053	1,200	,174	,081	,152	,010	,953	,020	,973
caffeina	,047	-,452	,124	,014	,019	,005	,702	,053	,755
nuovo	,047	,960	,147	,048	,086	,006	,893	,021	,914
interessante	,041	,657	-,056	,019	,035	,001	,911	,007	,918
pesante	,039	-,850	1,002	,070	,056	,246	,404	,560	,964
comune	,060	-,697	-,042	,038	,058	,001	,771	,003	,774
affumicato	,026	-,389	,266	,009	,008	,011	,446	,209	,655
magro	,052	1,305	,196	,094	,175	,013	,941	,021	,962
bambini	,024	-,352	-,513	,017	,006	,041	,179	,380	,559
funziona	,045	-,785	,477	,040	,055	,064	,693	,255	,948
dolce	,038	-,519	-,683	,048	,020	,112	,212	,368	,580
non comune	,024	,489	,186	,010	,011	,005	,585	,085	,670
brutto	,030	,006	-,109	,003	,000	,002	,000	,131	,131
fresco	,036	-,096	-,100	,002	,001	,002	,196	,214	,410
rampante	,034	,380	-,301	,012	,010	,019	,392	,246	,637
nutriente	,040	,722	,055	,022	,041	,001	,946	,006	,951
donne	,054	,758	-,063	,032	,062	,001	,965	,007	,972
minimo	,040	,579	,063	,023	,027	,001	,593	,007	,600
Totale attivi	1,000			,804	1,000	1,000			

Due dimensioni contribuiscono a una proporzione molto ampia dell'inerzia per la maggioranza dei punti di riga. I maggiori contributi della prima dimensione a *salutare*, *nuova*, *attraente*, *pochi grassi*, *nutriente* e *donne* indica che questi punti sono ben rappresentati in una dimensione. Di conseguenza, le dimensioni più elevate contribuiscono poco all'inerzia di questi punti, che si troveranno molto vicino all'asse orizzontale. La seconda dimensione contribuisce per la maggior parte a *uomini* , *alta qualità* e *forte*. Entrambe le dimensioni contribuiscono in modo molto limitato all'inerzia per *Australia del Sud* e *aspetto sgradevole*, perciò tali punti sono rappresentati in modo scarso.

La panoramica dei punti di colonna mostra i contributi relativi ai punti di colonna. Le marche *CC* e *DD* contribuiscono per la maggioranza alla prima dimensione, mentre *EE* e *FF* spiegano un'ampia porzione dell'inerzia per la seconda dimensione. *AA* e *BB* contribuiscono in modo molto limitato a ciascuna dimensione.

Figura 12-9
Contributi delle marche

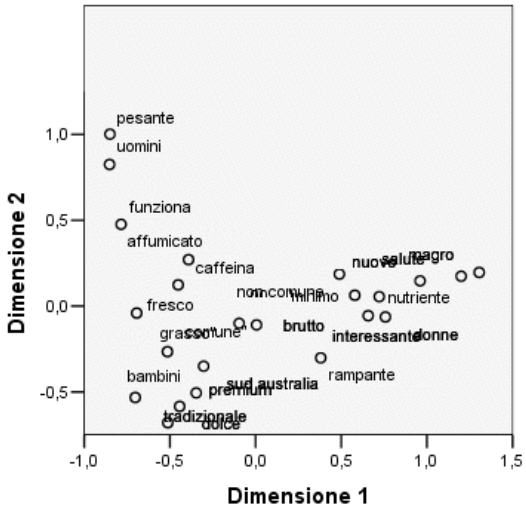
		Punteggio nella dimensione			Contributo				
					del punto all'inerzia della dimensione		della dimensione all'inerzia del punto		
marca	Massa	1	2	Inerzia	1	2	1	2	Totale
AA	,217	-,659	,046	,127	,187	,003	,744	,004	,748
BB	,131	-,284	-,404	,078	,021	,134	,135	,272	,407
CC	,185	,996	,076	,193	,362	,007	,951	,006	,957
DD	,162	,915	,101	,146	,267	,010	,928	,011	,939
EE	,152	-,651	,706	,153	,127	,477	,420	,494	,914
FF	,153	-,343	-,618	,107	,036	,369	,169	,550	,718
Totale attivi	1,000			,804	1,000	1,000			

In due dimensioni, tutte le marche eccetto *BB* sono ben rappresentate. *CC* e *DD* sono ben rappresentate in una dimensione. La seconda dimensione contribuisce per le porzioni maggiori per *EE* e *FF*. Si noti che *AA* è rappresentata nella prima dimensione, ma non contribuisce in modo significativo a essa.

Grafici

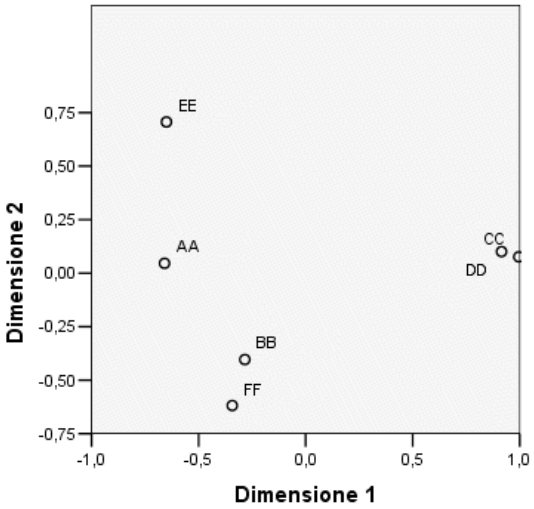
Il grafico dei punti di riga mostra che *fresca* e *aspetto sgradevole* sono entrambe molto vicine all'origine, a indicare che differiscono poco dal profilo di riga medio. A emergere sono tre classificazioni generali. Situati nella parte superiore sinistra del grafico, *forte*, *uomini* e *classe lavoratrice* sono tutti simili tra loro. La parte inferiore sinistra include *dolce*, *ingrassante*, *bambini* e *alta qualità*. Di contro, *salutare* , *pochi grassi*, *nutriente* e *nuova* sono raggruppati nella parte destra del grafico.

Figura 12-10
Grafico degli attributi dell'immagine (normalizzazione principale)



Si noti nei punti di colonna che tutte le marche sono lontane dall'origine, perciò nessuna marca è simile al centroide globale. Le marche *CC* e *DD* sono raggruppate a destra, mentre *BB* e *FF* sono raggruppate nella metà inferiore del grafico. Le marche *AA* e *EE* non sono simili ad alcuna altra marca.

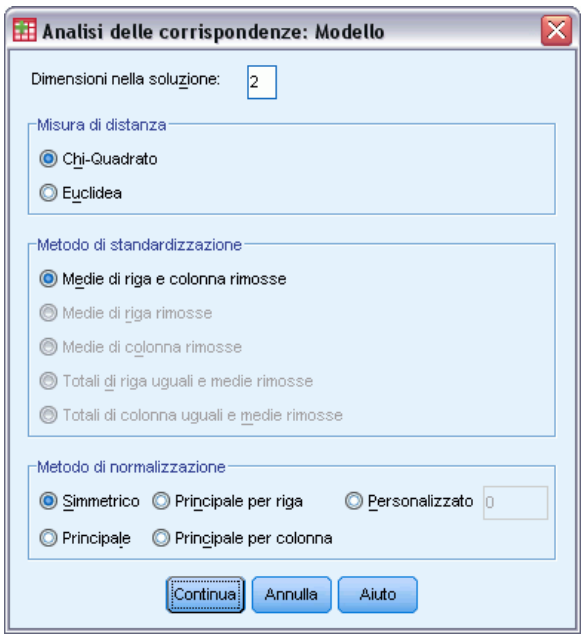
Figura 12-11
Grafico delle marche (normalizzazione principale)



Normalizzazione simmetrica

Qual è la correlazione tra marche e attributi dell'immagine? La normalizzazione principale non è in grado di evidenziare queste relazioni. Per concentrarsi sulle correlazioni tra le variabili, è necessario utilizzare la normalizzazione simmetrica. Anziché distribuire l'inerzia due volte (come nella normalizzazione principale), la normalizzazione simmetrica divide l'inerzia in parti uguali tra righe e colonne. Le distanze tra le categorie per una singola variabile non possono essere interpretate, ma la distanze tra le categorie per le diverse variabili sono significative.

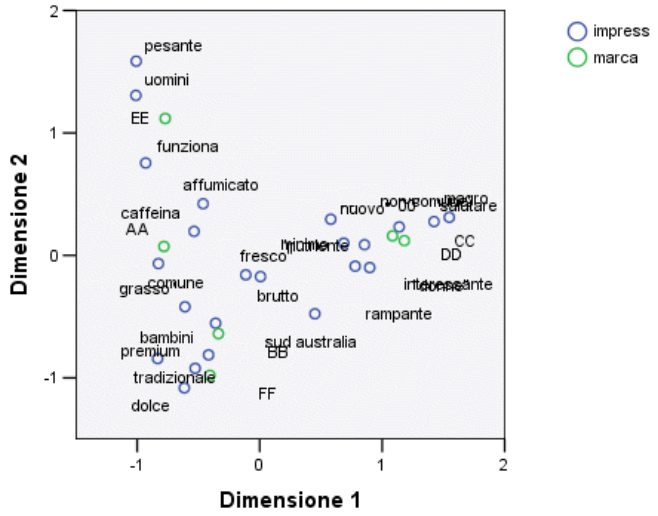
Figura 12-12
Finestra di dialogo Modello



- Per generare la seguente soluzione con la normalizzazione simmetrica, richiamare la finestra di dialogo Analisi delle corrispondenze e fare clic su Modello.
- Selezionare Simmetrico come metodo di normalizzazione.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della corrispondenze fare clic su OK.

Nella parte superiore sinistra del biplot risultante, la marca *EE* è l'unica forte, per la classe lavoratrice e che attrae gli uomini. La marca *AA* è la più diffusa e quella percepita come contenente la maggiore percentuale di caffeina. Le marche dolci e che fanno ingrassare includono *BB* e *FF*. Le marche *CC* e *DD* , sebbene percepite come nuove e salutari, sono anche le meno diffuse.

Figura 12-13
Biplot delle marche e degli attributi (normalizzazione simmetrica)



Per un'interpretazione ulteriore, è possibile estrarre una linea attraverso l'origine e i due attributi dell'immagine *uomini* e *yuppie* proiettare le marche su tale linea. I due attributi sono uno il contrario dell'altro, a indicare che il modello di associazione delle marche per *uomini* è invertito rispetto al modello per *yuppie*. Ovvero, la categoria uomini viene associata con la maggiore frequenza alla marca *EE* e con la frequenza minore alla marca *CC*, laddove la categoria yuppie è associata alla marca *CC* con la frequenza maggiore e alla marca *EE* con quella minore.

Letture consigliate

Consultare i testi seguenti per ulteriori informazioni sull'analisi delle corrispondenze:

Fisher, R. A. 1938. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics*, 10, .

Gilula, Z., e S. J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association*, 83, .

Capitolo
13

Analisi corrispondenze multiple

Lo scopo dell'analisi delle corrispondenze multiple, nota anche come analisi di omogeneità, è individuare le quantificazioni ottimali, nel senso che le categorie vengono separate le une dalle altre nella misura più ampia possibile. Questo implica che gli oggetti all'interno della stessa categoria vengono inseriti nel grafico gli uni accanto agli altri, mentre gli oggetti di categorie diverse sono inseriti in posizioni distanti. Il termine **omogeneità** fa inoltre riferimento al fatto che la correttezza dell'analisi è maggiore quando le variabili sono omogenee, ovvero quando suddividono gli oggetti in cluster con categorie uguali o simili.

Esempio: Caratteristiche degli articoli da ferramenta

Per esaminare il funzionamento dell'analisi della corrispondenza multipla, si utilizzano i dati ricavati da Hartigan (Hartigan, 1975), riportati in *screws.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*. Questo insieme di dati contiene informazioni sulle caratteristiche di viti, bulloni, dadi e puntine. La seguente tabella mostra le variabili, insieme alle relative etichette, e le etichette di valore assegnate alle categorie di ciascuna variabile nel file di dati relativi agli articoli da ferramenta di Hartigan.

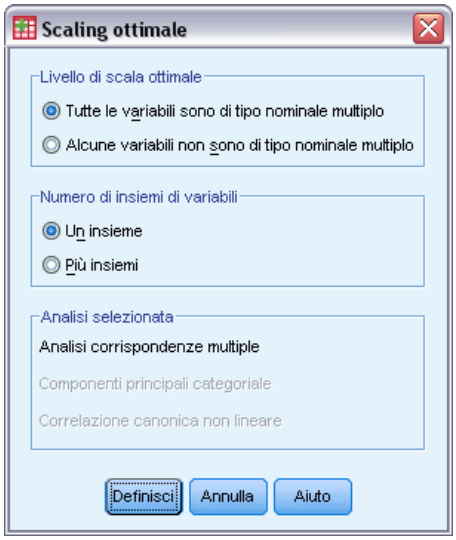
Tabella 13-1
File di dati relativi agli articoli da ferramenta di Hartigan

Nome di variabile	Etichetta di valore	Etichetta del valore
filettatura	*Filettatura*	*Sì_Filettatura*, *No_Filettatura*
testa	*Forma della testa*	*Piatta*, *A coppa*, *Conica*, *Arrotondata*, *Cilindrica*
rientes	*Rientro della testa*	*Nessuno*, *A stella*, *A feritoia*
inferiore	*Forma parte inferiore*	*A punta*, *Piatta*
lunghezza	*Lunghezza in mezzi pollici*	*1/2_in*, *1_in*, *1_1/2_ in*, *2_in*, *2_1/2_in*
ottone	*Ottone*	*Sì_Ot*, *No_Ot*
oggetto	*Oggetto*	*puntina, chiodo1, chiodo2, chiodo3, chiodo4, chiodo5, chiodo6, chiodo7, chiodo8, vite1, vite2, vite3, vite4, vite5, bullone1, bullone2, bullone3, bullone4, bullone5, bullone6, puntina1, puntina2, chiodob, viteb*

Esecuzione dell'analisi

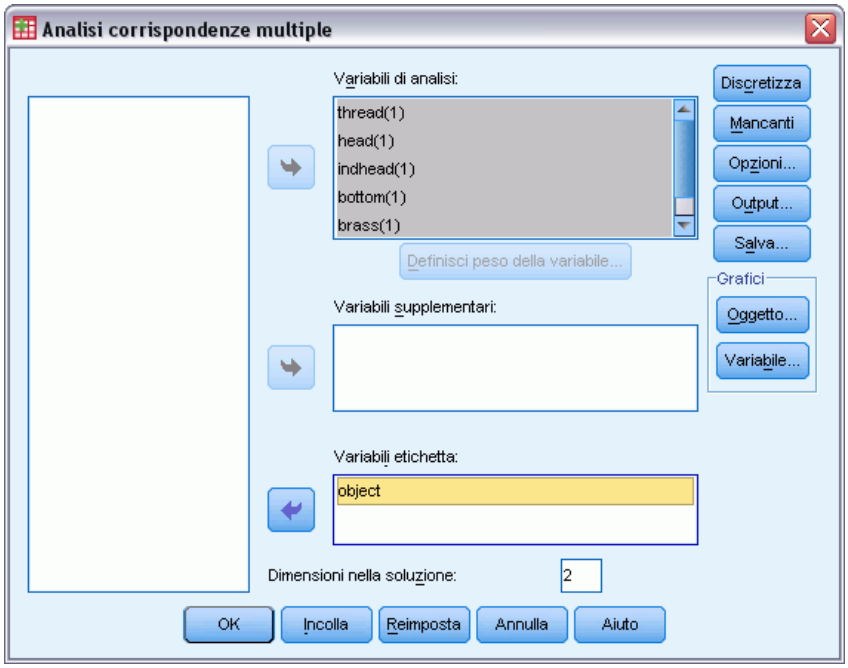
- Per eseguire un'analisi delle corrispondenze multiple, dai menu scegliere:
  Analizza > Riduzioni dimensione > Scaling ottimale...

Figura 13-1
Finestra di dialogo Scaling ottimale



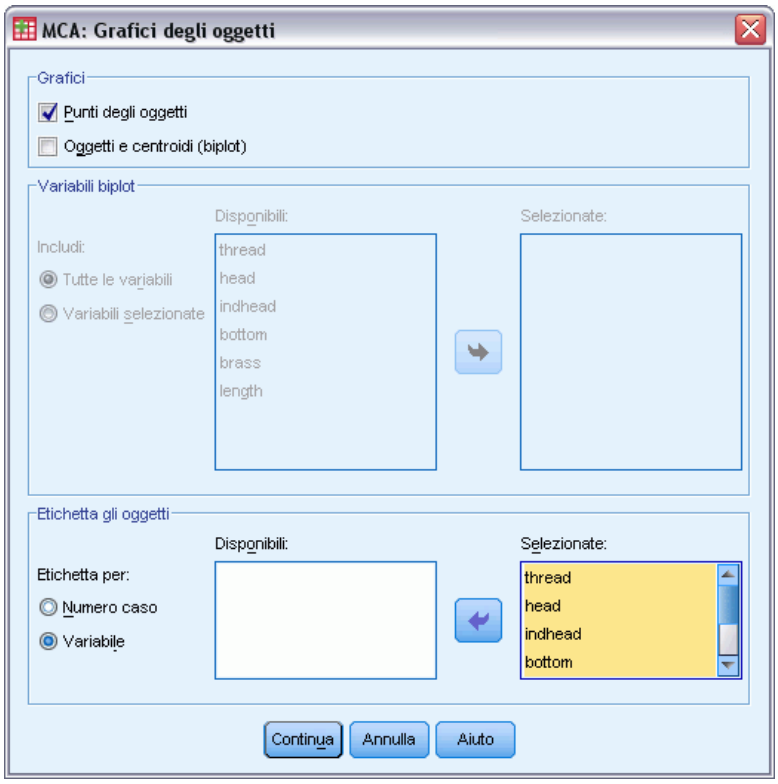
- Assicurarsi che le opzioni Tutte le variabili nominali multiple e Un insieme siano selezionate e fare clic su Definisci.

Figura 13-2
Finestra di dialogo Analisi delle corrispondenze multiple



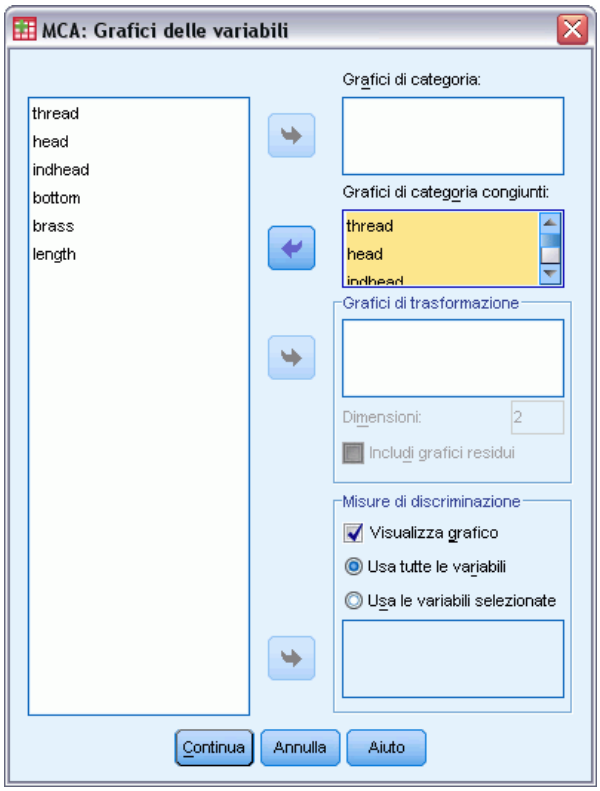
- Selezionare da *Filettatura* a *Lunghezza in mezzi pollici* come variabili di analisi.
- Scegliere *oggetto* come variabile di etichettatura.
- Nel gruppo Grafici fare clic su Oggetto.

Figura 13-3
Finestra di dialogo Grafici: Oggetto



- Scegliere di etichettare gli oggetti in base a Variabile.
- Scegliere da *filettatura* a *oggetto* come variabili di etichettatura.
- Nel gruppo Grafici della finestra di dialogo Analisi della corrispondenze multiple fare clic su Continua e quindi su Variabile.

Figura 13-4
Finestra di dialogo Grafici delle variabili



- Scegliere di generare un grafico di categoria congiunto per le variabili da *filettatura* a *lunghezza*.
- Fare clic su Continua.
- Nella finestra di dialogo Analisi della corrispondenze multiple fare clic su OK.

Riepilogo del modello

L'analisi di omogeneità può calcolare una soluzione per numerose dimensioni. Il numero massimo di dimensioni è pari al numero di categorie meno il numero delle variabili senza dati mancanti oppure al numero di osservazioni meno 1, a seconda di quale dei due sia il valore minore. Tuttavia, è consigliabile utilizzare il numero massimo di dimensioni solo raramente. Un numero inferiore di dimensioni è più facile da interpretare e, dopo un certo numero di dimensioni, la quantità di associazione aggiuntiva spiegata diventa trascurabile. In un'analisi di omogeneità una soluzione a una, due o tre dimensioni è molto comune.

Figura 13-5
Riepilogo modello

Dimensione	Alfa di Cronbach	Varianza spiegata		
		Totale (autovalore)	Inerzia	% di varianza
1	,878	3,727	,621	62,123
2	,657	2,209	,368	36,809
Media	,796[a]	2,968	,495	49,466

a. La media di Alfa di Cronbach è basata sulla media dell'autovalore.

La quasi totalità della varianza nei dati è spiegata dalla soluzione: il 62,1% dalla prima dimensione e il 36,8% dalla seconda.

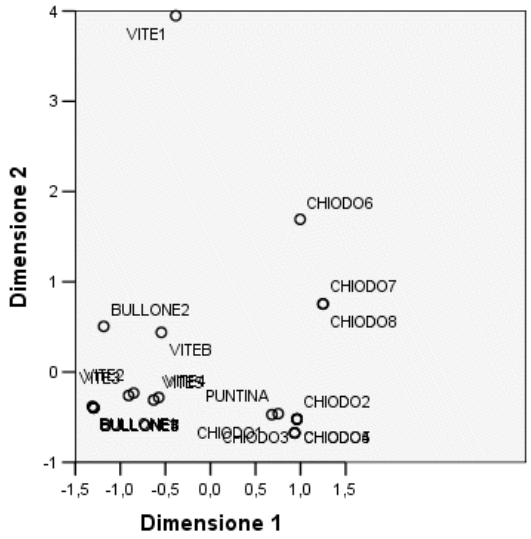
Le due dimensioni insieme forniscono un'interpretazione in termini di distanze. Se la discriminazione di una variabile è elevata, gli oggetti saranno vicini alle categorie cui appartengono. Idealmente, gli oggetti nella stessa categoria saranno vicini gli uni agli altri (ovvero, avranno punteggi simili) e le categorie di variabili diverse saranno vicine se appartengono agli stessi oggetti (ovvero, due oggetti con punteggi simili per una variabile avranno anche punteggi analoghi per le altre nella soluzione).

Punteggi oggetto

Dopo avere esaminato il riepilogo del modello, verificare i punteggi degli oggetti. È possibile specificare una o più variabili per etichettare il grafico dei punteggi degli oggetti. Per ciascuna variabile di etichettatura viene generato un grafico distinto, etichettato in base ai valori della specifica variabile. Verrà inoltre esaminato il grafico dei punteggi degli oggetti etichettati in base all'oggetto della variabile. Si tratta di una variabile di identificazione dei casi non utilizzata in nessun calcolo.

La distanza di un oggetto dall'origine riflette la variazione rispetto al modello di risposta "media". Questo modello di risposta media corrisponde alla categoria più frequente per ogni variabile. Gli oggetti con un numero elevato di caratteristiche corrispondenti alle categorie più frequenti si trovano in prossimità dell'origine. Per contro, gli oggetti con caratteristiche uniche sono posizionati lontani dall'origine.

Figura 13-6
Grafico dei punteggi degli oggetti etichettati per oggetto



Esaminando il grafico, si può vedere che la prima dimensione (l'asse orizzontale) separa viti e bulloni (dotati di filettatura) da chiodi e puntine (privi di filettatura). Questo è evidente nel grafico in quanto viti e bulloni si trovano su un'estremità dell'asse orizzontale, puntine e chiodi sull'altra. In misura minore, la prima dimensione separa inoltre i bulloni (con parte inferiore piatta) da tutti gli altri oggetti (con parte inferiore a punta).

La seconda dimensione (l'asse verticale) separa *VITE1* e *CHIODO6* da tutti altri oggetti. L'elemento in comune tra *VITE1* e *CHIODO6* sono i valori sulla lunghezza della variabile: sono gli oggetti più lunghi presenti nei dati. Inoltre, *VITE1* si trova molto più lontano dall'origine rispetto agli altri oggetti, il che suggerisce che, considerate nel complesso, molte caratteristiche di questo oggetto non sono condivise dagli altri.

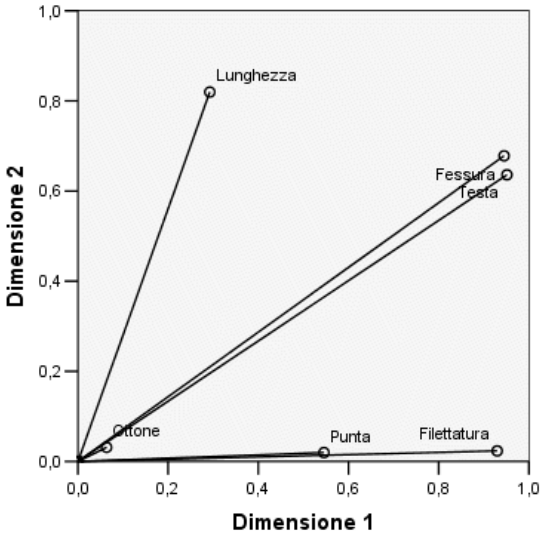
Il grafico dei punteggi degli oggetti è particolarmente utile per individuare visivamente i valori anomali. *VITE1* potrebbe essere considerato un valore anomalo. Più oltre, considereremo cosa avviene escludendo questo oggetto.

Misure di discriminazione

Prima di esaminare il resto dei grafici dei punteggi delle oggetti, verificheremo se le misure di discriminazione concordano con quanto detto finora. Per ogni variabile, una misura di discriminazione, che può essere considerata come un peso di componente quadrato, viene calcolata per ogni dimensione. Questa misura rappresenta anche la varianza della variabile quantificata in quella dimensione. Il suo valore massimo è 1, ottenuto se i punteggi degli oggetti sono compresi in gruppi mutuamente esclusivi e se tutti i punteggi di oggetti all'interno di una categoria sono identici (*Nota*: Questa misura può avere un valore maggiore di 1 in presenza di dati mancanti). Misure di discriminazione elevate corrispondono a una distribuzione ampia tra le categorie della variabile e, di conseguenza, indicano un grado elevato di discriminazione tra le categorie di una variabile in quella dimensione.

La media delle misure di discriminazione per qualsiasi dimensione è pari alla percentuale della varianza spiegata per quella dimensione. Di conseguenza, le dimensioni vengono ordinate in base alla discriminazione media. La prima dimensione ha la discriminazione media più ampia, la seconda dimensione il valore di discriminazione successivo e così via per tutte le dimensioni della soluzione.

Figura 13-7
Grafico delle misure di discriminazione



Come evidenziato nel grafico dei punteggi degli oggetti, il grafico delle misure di discriminazione mostra che la prima dimensione è relativa alle variabili *Filettatura* e *Forma parte inferiore*. Queste variabili hanno elevate misure di discriminazione nella prima dimensione e misure di discriminazione ridotte nella seconda. Di conseguenza, per entrambe queste variabili, le categorie vengono distribuite in posizioni distanti solo lungo la prima dimensione. *Lunghezza in mezzi pollici* ha un valore elevato nella seconda dimensione, ma un valore ridotto nella prima. Di conseguenza, *lunghezza* è più vicina alla seconda dimensione, il che corrisponde all'osservazione fatta sul grafico dei punteggi degli oggetti relativa al fatto che la seconda dimensione sembra dividere gli oggetti più lunghi agli altri. *Rientro della testa* e *Forma della testa* hanno valori relativamente ampi in entrambe le dimensioni, a indicare una discriminazione della prima della seconda dimensione. La variabile *Ottone*, situata in una posizione molto vicino all'origine, non determina nessuna discriminazione nelle prime due dimensioni. Questo ha senso in quanto tutti gli oggetti possono essere fatti di ottone o meno.

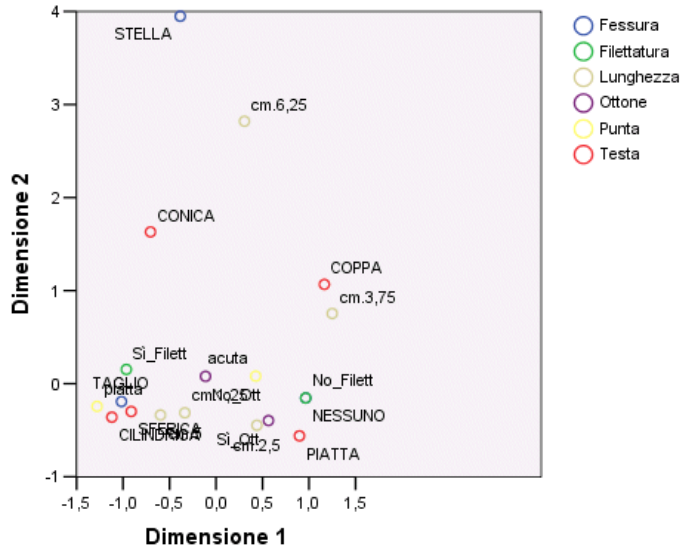
Quantificazioni di categoria (Categories: opzioni Visualizza)

Si ricordi che una discriminazione rappresenta la varianza della variabile quantificata in una particolare dimensione. Il grafico delle misure di discriminazione include queste varianze, a indicare quali variabili comportano una discriminazione e in quale dimensione. Tuttavia, la stessa varianza può corrispondere a tutte le categorie distribuite in una posizione relativamente lontana o alla maggioranza delle categorie vicine tra loro, con un numero limitato di categorie

che differiscono da questo gruppo. Il grafico di discriminazione non può differenziare queste due condizioni.

I grafici delle quantificazioni di categoria offrono un metodo alternativo di visualizzazione della discriminazione tra le variabili in grado di identificare le relazioni tra categorie. In questo grafico, sono visualizzate le coordinate di ciascuna categoria in ciascuna dimensione. Di conseguenza, è possibile determinare quali categorie sono simili per ciascuna variabile.

Figura 13-8
Quantificazioni di categoria



Lunghezza in mezzi pollici ha cinque categorie, tre delle quali sono raggruppate in prossimità della parte superiore del grafico. Le due categorie rimanenti in si trovano nella metà inferiore del grafico, con la categoria *2_1/2_in* posizionata molto lontana dal gruppo. L'elevata discriminazione relativa alla lunghezza nella dimensione 2 è il risultato della forte differenziazione di questa categoria rispetto alle altre categorie di lunghezza. Analogamente, per *Forma della testa*, la categoria *A STELLA* è molto lontana delle altre categorie e genera una misura discriminazione elevata nella seconda dimensione. Questi modelli non possono essere illustrati in un grafico delle misure discriminazione.

La distribuzione delle quantificazioni di categoria per una variabile riflette la varianza e quindi indica la correttezza della discriminazione di tale variabile in ciascuna dimensione. Concentrando l'attenzione sulla dimensione 1, le categorie per *Filettatura* sono molto lontane. Tuttavia, nella dimensione 2, le categorie per questa variabile sono molto vicine. Di conseguenza, *Filettattura* comporta una discriminazione migliore nella dimensione 1 rispetto alla dimensione 2. Per contro, le categorie per *Forma della testa* sono distribuite in posizioni lontane in entrambe le dimensioni, a suggerire che questa variabile comporti una discriminazione corretta in entrambe le dimensioni..

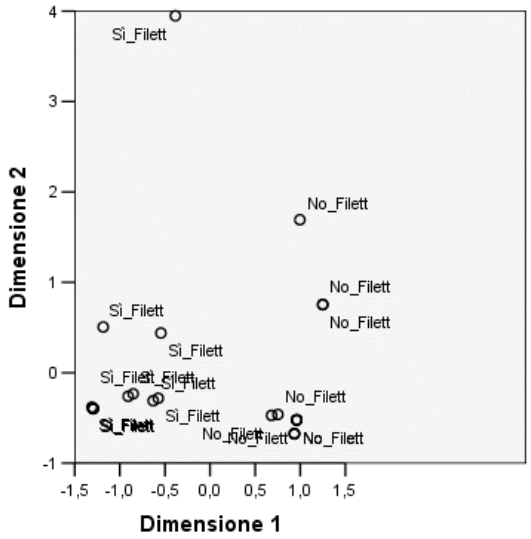
Oltre a determinare le dimensioni lungo le quali una variabile comporta una discriminazione e le modalità di quest'ultima,il grafico delle quantificazioni di categoria confronta anche la discriminazione della variabile. Una variabile con categorie lontane comporta una discriminazione migliore rispetto a una variabile con categorie vicine tra loro. Ad esempio, nella dimensione 1, le due categorie di *Ottone* sono più vicine tra loro delle due categorie di *Filettatura*, a

indicare che *Filettatura* comporta una discriminazione migliore rispetto a *Ottone* in questa dimensione. Tuttavia, nella dimensione 2, le distanze sono molto simili, a suggerire che il grado di discriminazione di queste variabili è lo stesso nella dimensione corrente. Il grafico delle misure di discriminazione illustrato sopra identifica queste stesse relazioni utilizzando le varianze per riflettere la distribuzione delle categorie.

Un esame più dettagliato dei punteggi degli oggetti

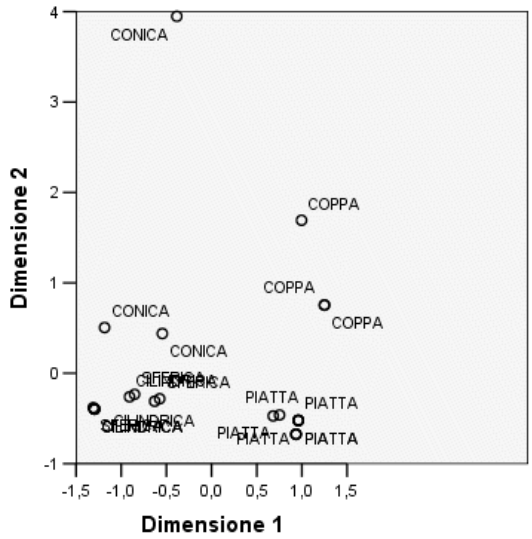
È possibile ottenere una migliore comprensione dei dati esaminando i grafici dei punteggi degli oggetti etichettati in base a ciascuna variabile. Idealmente, oggetti simili dovrebbero formare gruppi esclusivi e questi gruppi dovrebbero essere lontani tra loro.

Figura 13-9
Punteggi degli oggetti etichettati in base a Filettatura



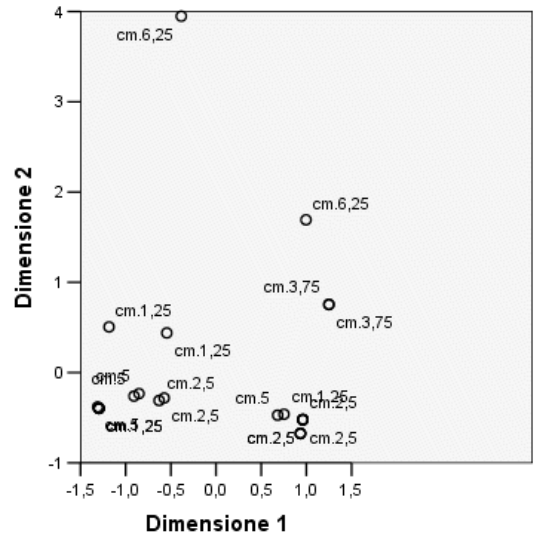
Il grafico etichettato con *Filettatura* mostra che la prima dimensione separa perfettamente *Sì_Filettatura* e *No_Filettatura* . Tutti gli oggetti con filettature hanno punteggi degli oggetti negativi, mentre tutti gli oggetti privi di filettatura hanno punteggi degli oggetti positivi. Sebbene le due categorie non formino gruppi compatti, la perfetta differenziazione tra le categorie è generalmente considerata un buon risultato.

Figura 13-10
Punteggi degli oggetti etichettati in base a Forma della testa



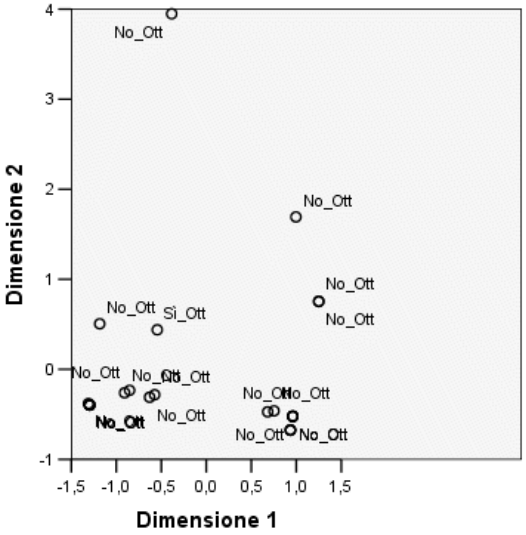
Il grafico etichettato con *Forma della testa* mostra che questa variabile comporta una discriminazione in entrambe le dimensioni. Gli oggetti *PIATTI* sono raggruppati nell'angolo inferiore destro del grafico, mentre gli oggetti *A COPPA* sono raggruppati nella parte superiore destra. Gli oggetti *CONICA* sono tutti posizionati nella parte superiore sinistra. Tuttavia, tali oggetti sono più distribuiti rispetto agli altri gruppi e, di conseguenza, non altrettanto omogenei. Infine, gli oggetti *CILINDRICA* non possono essere separati dagli oggetti *ROTONDI* ; entrambi sono posizionati nell'angolo inferiore destro del grafico.

Figura 13-11
Punteggi degli oggetti etichettati in base a Lunghezza in mezzi pollici



Il grafico etichettato con *Lunghezza in mezzi pollici* mostra che questa variabile non comporta una discriminazione nella prima dimensione. Le sue categorie non visualizzano alcun raggruppamento se proiettate su una linea orizzontale. Tuttavia, *Lunghezza in mezzi pollici* non comporta una discriminazione nella seconda dimensione. Gli oggetti più corti corrispondono a punteggi positivi, gli oggetti più lunghi a elevati punteggi negativi.

Figura 13-12
Punteggi degli oggetti etichettati in base a Ottone



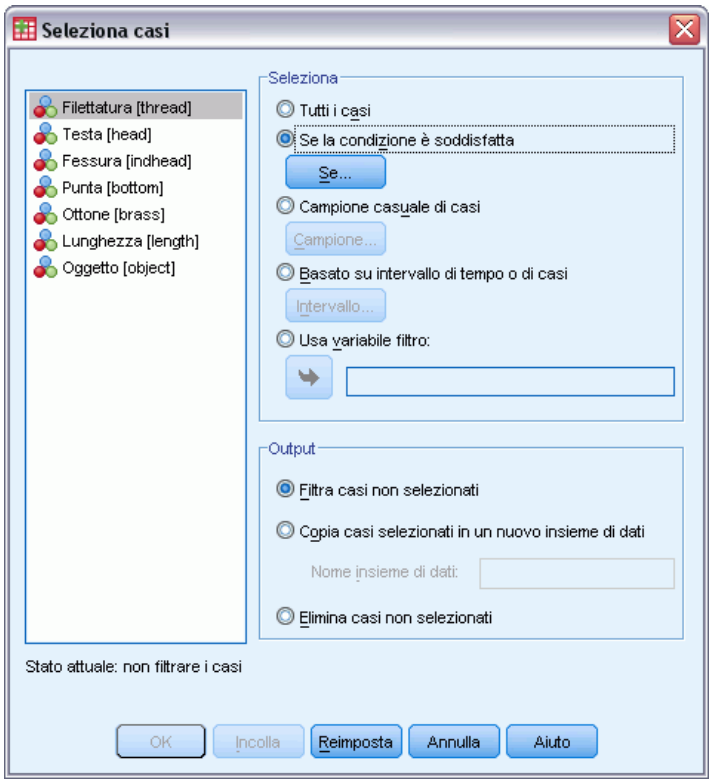
Il grafico etichettato con *Ottone* mostra che questa variabile include categorie che non è possibile separare nettamente nella prima o nella seconda dimensione. I punteggi degli oggetti sono ampiamente distribuiti nello spazio. Gli oggetti in ottone non possono essere differenziati rispetto agli oggetti non in ottone.

Omissione di valori anomali

Nell'analisi di omogeneità, i valori anomali sono oggetti con un numero eccessivo di caratteristiche uniche. Come notato in precedenza, *VITE1* può essere considerato un valore anomalo.

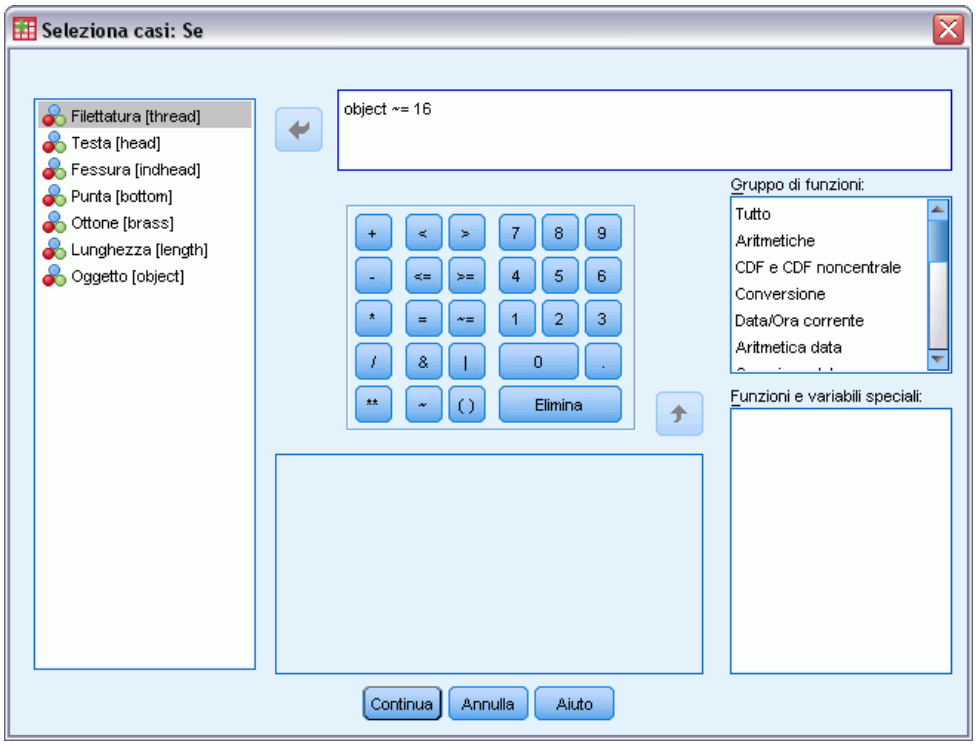
Per eliminare questo oggetto ed eseguire di nuovo l'analisi, dai menu scegliere:
Dati > Seleziona casi...

Figura 13-13
Finestra di dialogo Seleziona casi



- Selezionare Se la condizione è soddisfatta.
- Fare clic su Se.

Figura 13-14
Se la finestra di dialogo



- Digitare oggetto ~= 16 come condizione.
- Fare clic su Continua.
- Fare clic su OK nella finestra di dialogo Seleziona casi.
- Infine, richiamare la finestra di dialogo Analisi delle corrispondenze multiple e fare clic su OK.

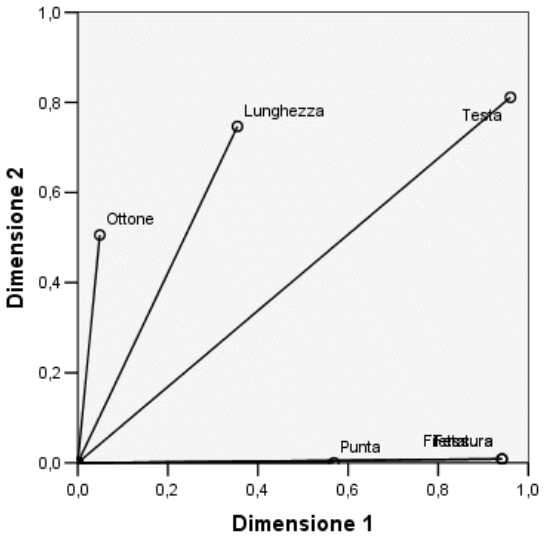
Figura 13-15
Riepilogo del modello (con valore anomalo rimosso)

Dimensione	Alfa di Cronbach	Varianza spiegata		
		Totale (autovalore)	Inerzia	% di varianza
1	,885	3,815	,636	63,591
2	,623	2,081	,347	34,676
Media	,793[a]	2,948	,491	49,133

a. La media di Alfa di Cronbach è basata sulla media dell'autovalore.

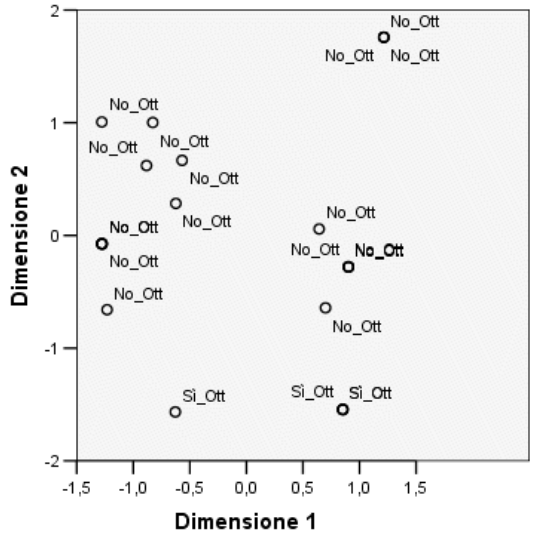
Gli autovalori si modificano leggermente. La prima dimensione spiega ora una porzione leggermente superiore della varianza.

Figura 13-16
Misure di discriminazione



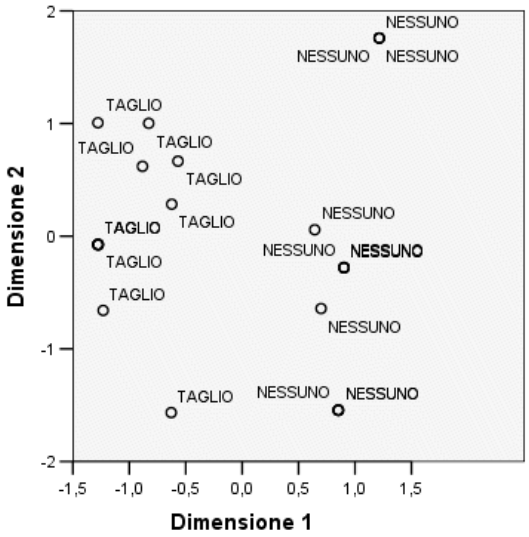
Come indicato nel grafico di discriminazione, *Rientro della testa* non comporta più una discriminazione nella seconda dimensione, mentre *Ottone* passa da nessuna discriminazione in tutte le dimensioni a una discriminazione nella seconda. La discriminazione per le altre variabili resta fondamentalmente invariata.

Figura 13-17
Punteggi degli oggetti etichettati in base a Ottone (con valore anomalo rimosso)



Il grafico dei punteggi degli oggetti etichettato in base alla variabile *Ottone* mostra che i quattro oggetti in ottone sono tutti visualizzati in prossimità della parte inferiore del grafico (tre oggetti occupano posizioni identiche), a indicare un'elevata discriminazione nella seconda dimensione. Come nel caso di *Filettatura* nell'analisi precedente, gli oggetti non formano gruppi compatti, ma la differenziazione degli oggetti per categorie è perfetta.

Figura 13-18
Punteggi degli oggetti etichettati in base a Rientro della testa (con valore anomalo rimosso)



Il grafico dei punteggi degli oggetti etichettato in base alla variabile *Rientro della testa* mostra che la prima dimensione comporta la discriminazione perfetta tra gli oggetti senza rientro e gli oggetti con rientro, come nell'analisi precedente. In contrasto con l'analisi precedente, tuttavia, nella seconda dimensione non è ora possibile distinguere tra le due categorie.

Di conseguenza, l'omissione di *VITE1*, l'unico oggetto con una testa a stella, influenza in modo significativo l'interpretazione della seconda dimensione. Questa dimensione differenzia ora gli oggetti in base alle variabili *Ottone*, *Forma della testa* e *Lunghezza in mezzi pollici*i

Letture consigliate

Consultare i testi seguenti per ulteriori informazioni sull'analisi delle corrispondenze multiple:

Benzécri, J. P. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.

Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. In: *The Prediction of Personal Adjustment,* P. Horst, ed. New York: Social Science Research Council.

Meulman, J. J. 1982. *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.

Meulman, J. J. 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification*, 13, .

Meulman, J. J., e W. J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. In: *Visual Display of Categorical Data,* M. Greenacre, e J. Blasius, ed. New York: Academic Press.

Nishisato, S. 1984. Forced classification: A simple application of a quantification method. *Psychometrika*, 49, .

Tenenhaus, M., e F. W. Young. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, .

Van Rijckevorsel, J. 1987. *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. Leiden: DSWO Press.

Capitolo
14

Scaling multidimensionale

Dato un insieme di oggetti, l'obiettivo dello scaling multidimensionale è individuare una rappresentazione degli oggetti in uno spazio dimensionale ridotto. Questa soluzione viene ottenuta utilizzando le **distanze** tra gli oggetti. La procedura riduce al minimo le deviazioni quadrate tra le distanze degli oggetti originali, o trasformati, e le relative distanze euclidee nello spazio dimensionale ridotto.

Lo scopo dello spazio dimensionale ridotto è evidenziare le relazioni tra gli oggetti,. Limitando la soluzione in modo che sia una combinazione lineare di variabili indipendenti, è possibile interpretare le dimensioni della soluzione alla luce di tali variabili. Nel seguente esempio, si vedrà come è possibile rappresentare 15 diversi termini indicanti parentela in tre dimensioni e come tale spazio può essere interpretato in relazione a sesso, generazione e grado di separazione di ciascuno di tali termini.

Esempio un esame dei termini indicanti parentela

Rosenberg e Kim (Rosenberg e Kim, 1975) si prefiggono di analizzare 15 termini indicanti parentela (zia, fratello, cugino, padre, nipote femmina di nonni, nonno, nonna, nipote maschio di nonni, madre, nipote maschio di zii), nipote femmina di zii, sorella, figlio, zio). Hanno richiesto a quattro gruppi di studenti universitari (due composti da femmine e due da maschi) di ordinare questi termini in base alla similiarità. A due gruppi (uno femminile e uno maschile) è stato richiesto di effettuare l'ordinamento due volte, con il secondo ordinamento basato su un criterio diverso rispetto al primo. Di conseguenza, sono state ottenute sei "sorgenti" totali, come indicato nella tabella seguente.

Tabella 14-1
Struttura sorgente dei dati di parentela

Sorgente	Sesso	Condizione	Dimensione campione
1	Femmina	Ordinamento singolo	85
2	Maschio	Ordinamento singolo	85
3	Femmina	Primo ordinamento	80
4	Femmina	Secondo ordinamento	80
5	Maschio	Primo ordinamento	80
6	Maschio	Secondo ordinamento	80

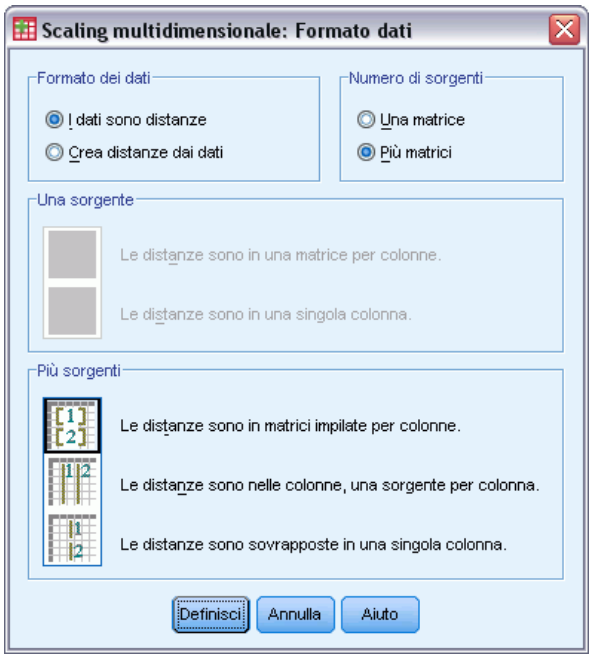
Ogni sorgente corrisponde a una matrice di prossimità 15×15, le cui celle sono uguali al numero delle persone in una sorgente meno il numero di volte in cui gli oggetti sono stati ripartiti insieme nella sorgente. Questo insiemi di dati è reperibile nel file *kinship_dat.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*.

Scelta del numero di dimensioni

Sta all'utente decidere quante dimensioni deve avere la soluzione. Il grafico decrescente aiuta a prendere tale decisione.

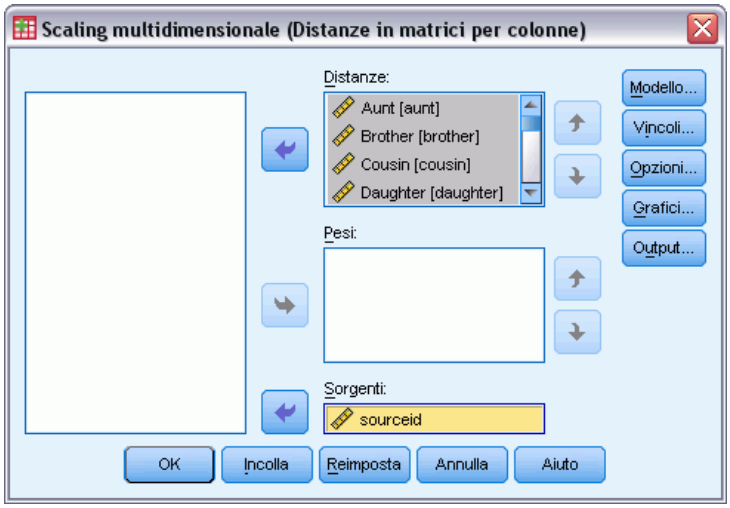
- Per creare un grafico decrescente degli autovalori, dai menu scegliere:
  Analizza > Scala > Scaling multidimensionale (PROXSCAL)...

Figura 14-1
Finestra di dialogo Formato dati



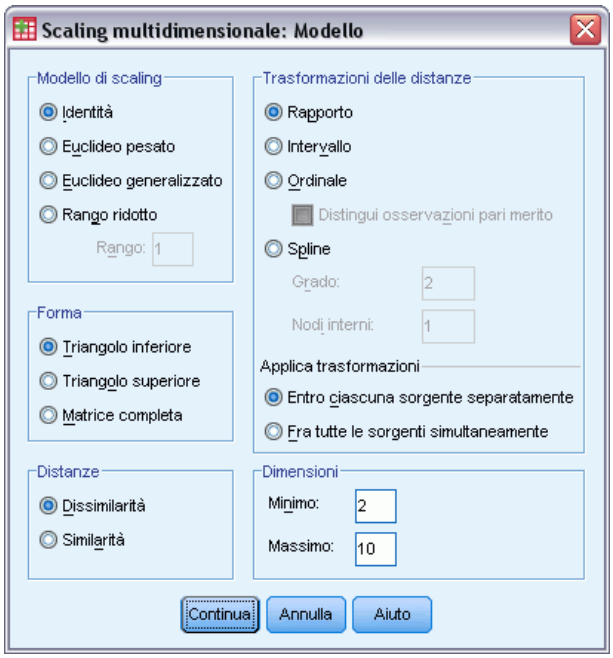
- Selezionare Più matrici nel gruppo Numero di sorgenti.
- Fare clic su Definisci.

Figura 14-2
Finestra di dialogo Scaling multidimensionale



- Selezionare da *Zia* a *Zio* come variabili di distanza.
- Selezionare *idsorgente* come variabile di identificazione della sorgente.
- Fare clic su Modello.

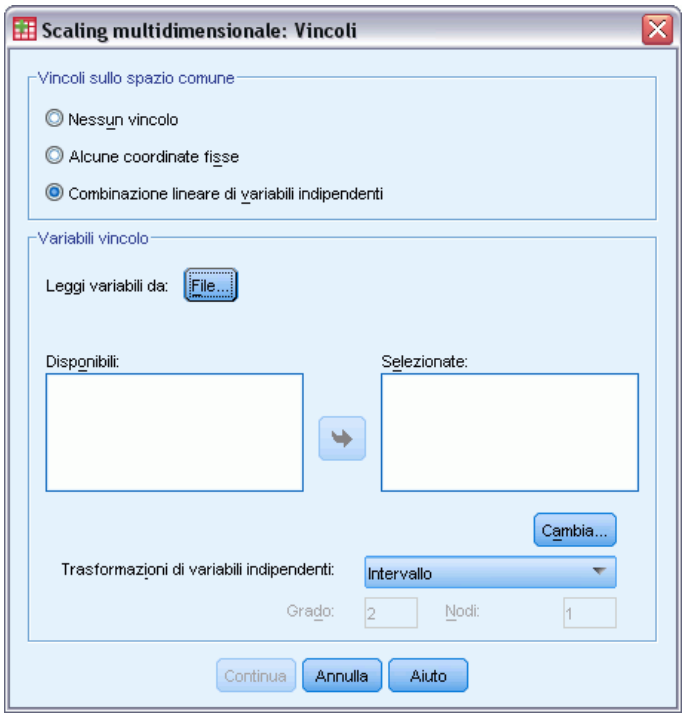
Figura 14-3
Finestra di dialogo Modello



- Digitare 10 come numero massimo delle dimensioni.
- Fare clic su Continua.

▶ Nella finestra di dialogo Scaling multidimensionale fare clic su Vincoli.

Figura 14-4
Finestra di dialogo Vincoli

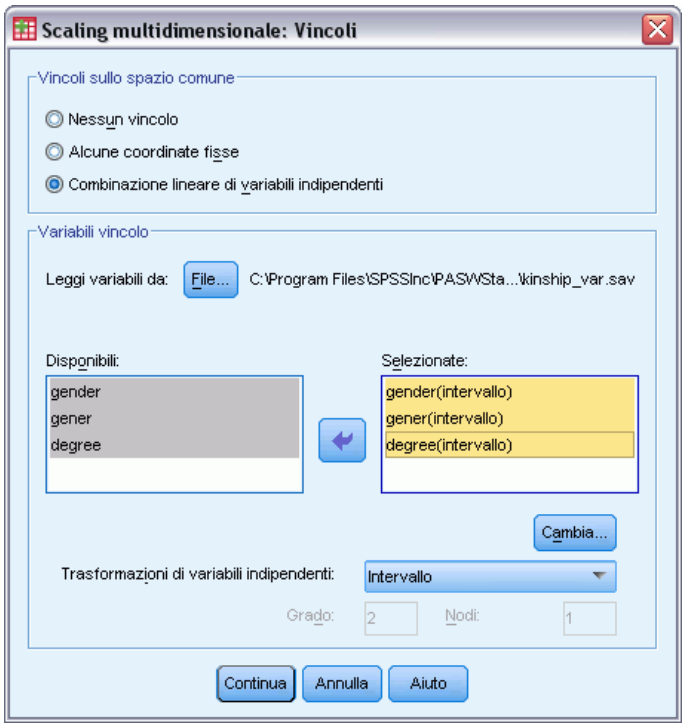


▶ Selezionare Combinazione lineare di variabili indipendenti.

▶ Fare clic su File per selezionare la sorgente delle variabili indipendenti.

▶ Selezionare *kinship_var.sav*.

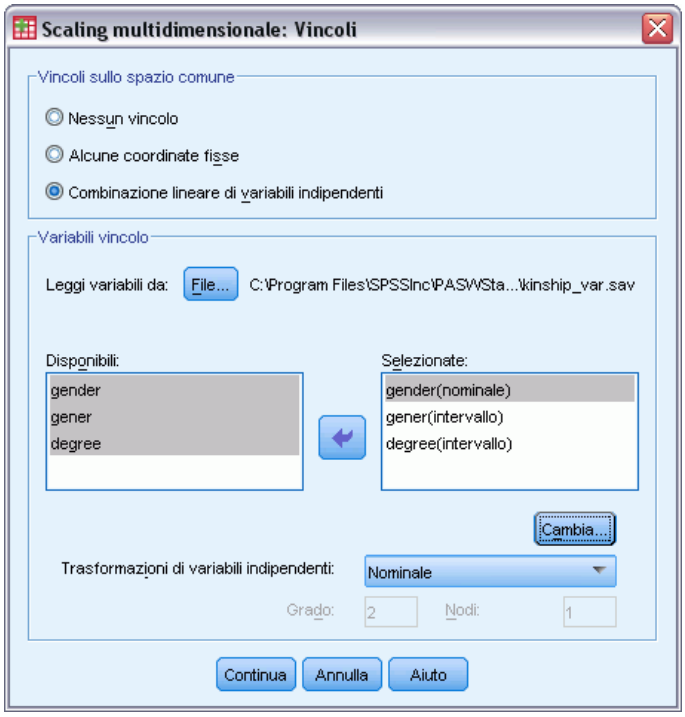
Figura 14-5
Finestra di dialogo Vincoli



► Selezionare *sesso*, *gener* e *grado* come variabili vincolo.

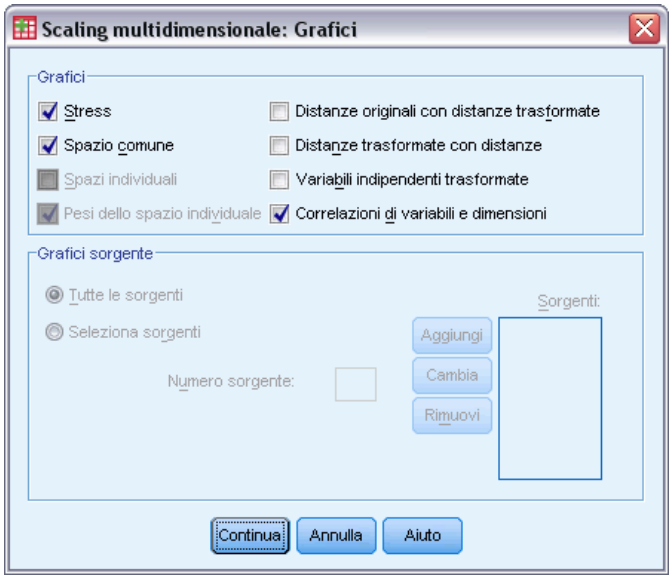
Si noti che la variabile *sesso* ha un valore mancante definito dall'utente—9=mancante (per cugino). La procedura considera tale valore come una categoria valida. Di conseguenza, è improbabile che la trasformazione lineare predefinita risulti appropriata. Utilizzare invece una trasformazione nominale.

Figura 14-6
Finestra di dialogo Vincoli



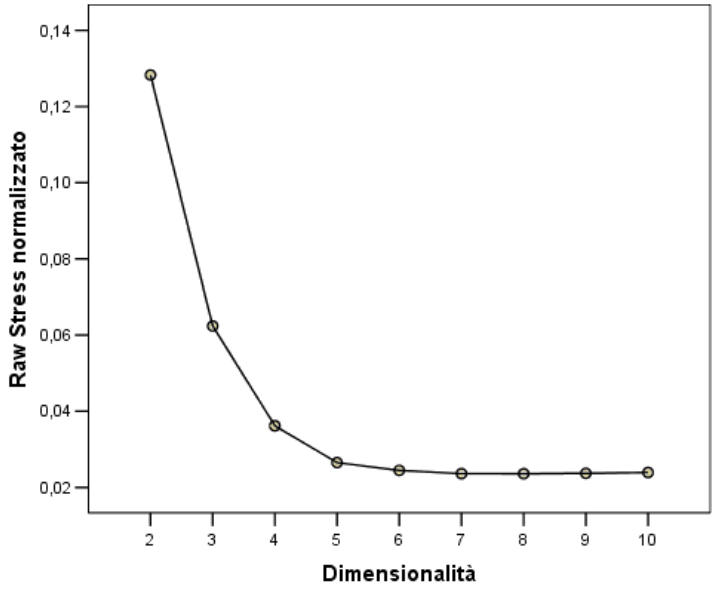
- Selezionare *sesso*.
- Selezionare Nominale dall'elenco a discesa Trasformazioni di variabili indipendenti.
- Fare clic su Cambia.
- Fare clic su Continua.
- Nella finestra di dialogo Scaling multidimensionale fare clic su Grafici.

Figura 14-7
Finestra di dialogo Grafici



- Nel gruppo Grafici fare clic su Stress.
- Fare clic su Continua.
- Nella finestra di dialogo Scaling multidimensionale fare clic su OK.

Figura 14-8
Grafico decrescente autovalori



La procedura inizia con una soluzione a dieci dimensioni, che si riducono a 2. Il grafico decrescente degli autovalori mostra il raw stress normalizzato della soluzione per ogni dimensione. È possibile vedere dal grafico che, aumentando la dimensionalità da 2 a 3 e da 3 a 4, lo stress viene notevolmente migliorato. Per valori successivi a 4, i miglioramenti sono limitati. Si sceglierà di analizzare i dati utilizzando una soluzione a tre dimensioni, in quanto l'interpretazione dei dati risulta facilitata.

Una soluzione a tre dimensioni

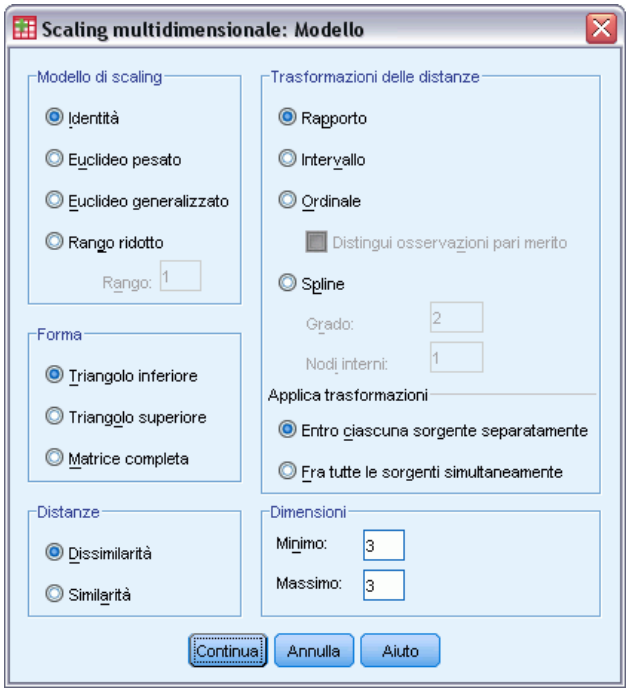
Le variabili indipendenti *sesso*, *gener*(generazione) e *grado* (di separazione) sono state strutturare con l'intenzione di utilizzarle per interpretare le dimensioni della soluzione. Le variabili indipendenti sono state strutturate come segue:

sesso	1 = maschile, 2 = femminile, 9 = mancante (per cugino)
gener	Il numero di generazioni tra il rispondente (se il termine si riferisce alla sua parentela), dove numeri più bassi corrispondono a generazioni più vecchie. Di conseguenza, i valori per nonni, nipoti e fratelli sono rispettivamente 2, 2 e 0.
grado	Il numero di gradi di separazione lungo l'albero genealogico. Di conseguenza, i genitori del rispondente si trovano sul nodo superiore, i suoi figli sul nodo inferiore. I fratelli del rispondente si trovano su un nodo più in alto rispetto ai suoi genitori e quindi di un altro nodo più in basso, per un totale di 2 gradi di separazione. Il cugino del rispondente è a 4 gradi di separazione—2 fino ai nonni, quindi altri due verso il basso attraverso la zia o lo zio.

Le variabili esterne sono reperibili in *kinship_var.sav*. Inoltre, una configurazione iniziale da un'analisi precedente è fornita in *kinship_ini.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*.

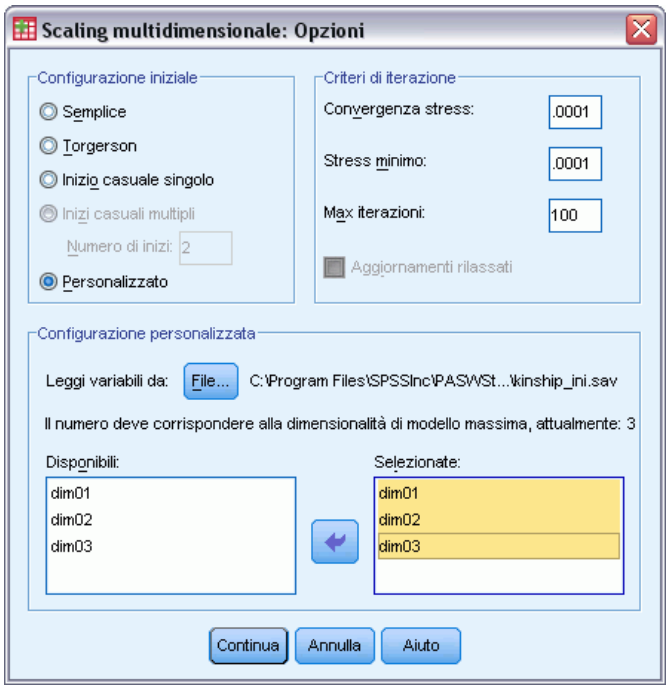
Esecuzione dell'analisi

Figura 14-9
Finestra di dialogo Modello



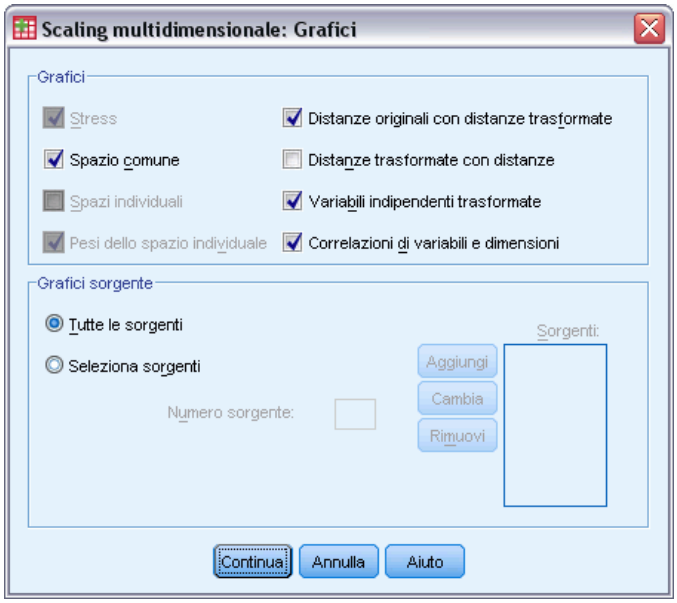
- Per ottenere una soluzione a tre dimensioni, richiamare la finestra di dialogo Scaling multidimensionale e fare clic su Modello.
- Digitare 3 come numero massimo e minimo di dimensioni.
- Fare clic su Continua.
- Nella finestra di dialogo Scaling multidimensionale fare clic su Opzioni.

Figura 14-10
Finestra di dialogo Opzioni



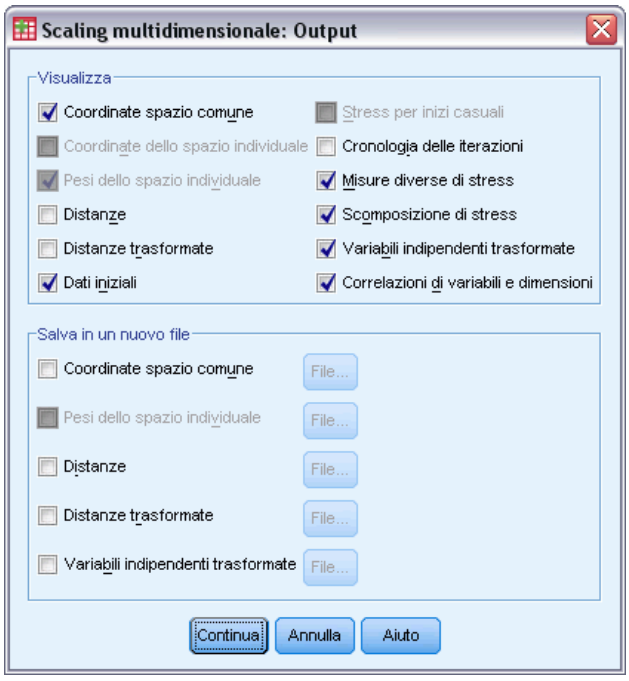
- Selezionare Personalizzata come configurazione iniziale.
- Selezionare *kinship_ini.sav* come file da cui leggere le variabili.
- Selezionare *dim01*, *dim02* e *dim03* come variabili.
- Fare clic su Continua.
- Nella finestra di dialogo Scaling multidimensionale fare clic su Grafici.

Figura 14-11
Finestra di dialogo Grafici



- Selezionare Distanze originali con distanze trasformatee Variabili indipendenti trasformate.
- Fare clic su Continua.
- Nella finestra di dialogo Scaling multidimensionale fare clic su Output.

Figura 14-12
Finestra di dialogo Output



- Selezionare Dati iniziali, Scomposizione di stress e Correlazioni di variabili e dimensioni.
- Fare clic su Continua.
- Nella finestra di dialogo Scaling multidimensionale fare clic su OK.

Misure di stress

Le misure di stress e di adattamento forniscono un'indicazione dell'approssimazione delle distanze nella soluzione rispetto alle distanze originali.

Figura 14-13
Misure di stress e di adattamento

Stress non trasformato normalizzato	,06234
Stress-I	,24968[a]
Stress-II	,87849[a]
S-stress	,14716[b]
Dispersione spiegata	,93766
Coefficiente di convergenza di Tucker	,96833

PROXSCAL minimizza lo stress non trasformato normalizzato.

a. Livello di scala ottimale = 1,066.

b. Livello di scala ottimale = ,984.

Ciascuna delle quattro statistiche di Stress misura il non adattamento dei dati, mentre la dispersione spiegata e il coefficiente di congruenza di Tucker misurano l'adattamento. Misure di stress inferiori (fino a un minimo di 0) e misure di adattamento superiori (fino a un massimo di 1) indicano soluzioni migliori.

Figura 14-14
Scomposizione del raw stress normalizzato

		Sorgente						
		SRC_1	SRC_2	SRC_3	SRC_4	SRC_5	SRC_6	Media
Oggetto	Zia	,0991	,0754	,0629	,0468	,0391	,0489	,0620
	Fratello	,1351	,0974	,0496	,0813	,0613	,0597	,0807
	Cugino	,0325	,0336	,0480	,0290	,0327	,0463	,0370
	Figlia	,0700	,0370	,0516	,0229	,0326	,0207	,0391
	Padre	,0751	,0482	,0521	,0225	,0272	,0298	,0425
	Pronipote femmina	,1410	,0736	,0801	,0707	,0790	,0366	,0802
	Nonno	,1549	,1057	,0858	,0821	,0851	,0576	,0952
	Nonna	,1550	,0979	,0858	,0844	,0816	,0627	,0946
	Pronipote maschio	,1374	,0772	,0793	,0719	,0791	,0382	,0805
	Madre	,0813	,0482	,0526	,0229	,0260	,0227	,0423
	Nipote maschio	,0843	,0619	,0580	,0375	,0317	,0273	,0501
	Nipote femmina	,0850	,0577	,0503	,0353	,0337	,0260	,0480
	Sorella	,1361	,0946	,0496	,0816	,0629	,0588	,0806
	Figlio	,0689	,0373	,0456	,0242	,0337	,0253	,0392
	Zio	,0977	,0761	,0678	,0489	,0383	,0498	,0631
Media		,1035	,0681	,0613	,0508	,0496	,0407	,0623

La decomposizione dello stress consente di identificare le sorgenti e gli oggetti che forniscono il maggiore contributo allo stress complessivo della soluzione. In questo caso, la maggioranza dello stress tra le sorgenti è attribuibile alle sorgenti 1 e 2, mentre tra gli oggetti, la maggioranza dello stress è attribuibile a *Fratello*, *Nipote femmina di nonni*, *Nonno*, *Nonna*, *Nipote maschio di nonni* e *Sorella*.

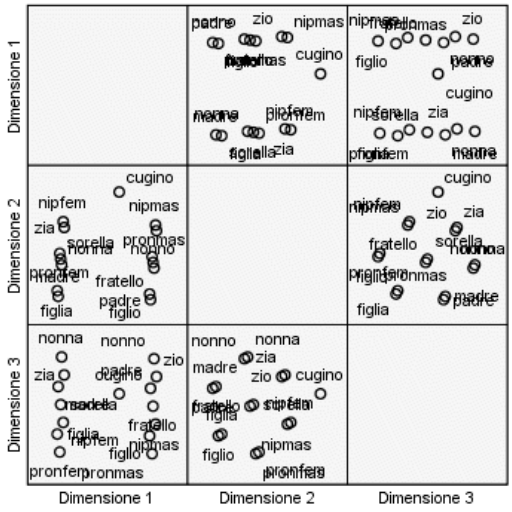
Le due sorgenti all'origine della maggioranza dello stress sono i due gruppi che hanno ordinato i termini solo una volta. Tale informazione suggerisce che tutti gli studenti hanno considerato fattori multipli nell'ordinamento dei termini e che coloro che hanno potuto eseguire l'ordinamento due volte si sono concentrati su una parte di tali fattori per il primo ordinamento, considerando poi i fattori restanti durante il secondo.

Gli oggetti che spiegano la maggioranza dello stress sono quelli con *grado* pari a 2. Tali persone rappresentano le relazioni che non sono parte della famiglia "nucleare" (*Madre*, *Padre*, *Figlia*, *Figlio*), ma che sono comunque più strette di altre. Questa posizione centrale potrebbe facilmente determinare qualche tipo di ordinamento differenziale dei termini.

Coordinate finali dello spazio comune

Il grafico dello spazio comune fornisce una rappresentazione visuale delle relazioni tra gli oggetti.

Figura 14-15
Coordinate spazio comune



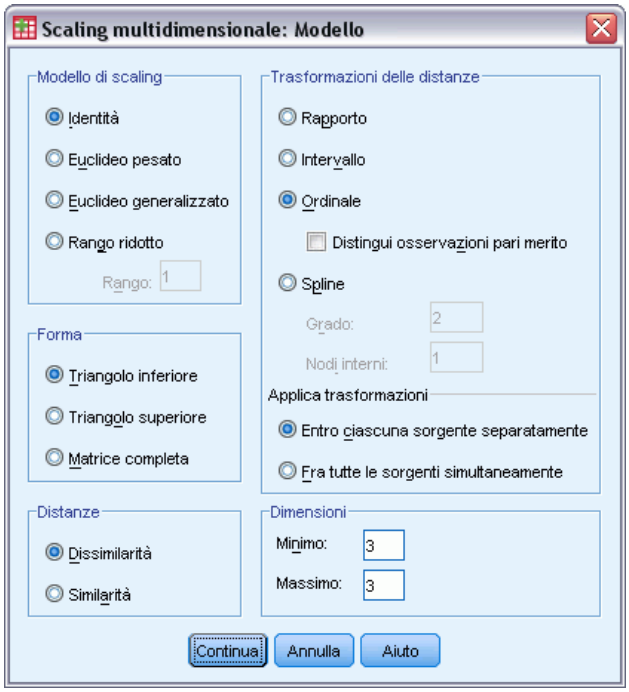
Si esaminino le coordinate finali per gli oggetti nelle dimensioni 1 e 3; si tratta del grafico nell'angolo inferiore sinistro della matrice di grafici a dispersione. Questo grafico mostra che la dimensione 1 (sull'asse *x*) è correlata con la variabile *sesso* e che la dimensione 3 (sull'asse *y*) è correlata con *gener*. Da sinistra a destra, è possibile vedere che la dimensione 1 separa i termini femminili e maschili, con il termine neutro *Cugino* nel mezzo. Dal basso del grafico verso l'alto, valori crescenti lungo l'asse corrispondono ai termini più vecchi.

Si esaminino ora le coordinate finali per gli oggetti nelle dimensioni 2 e 3; si tratta del grafico nella parte centrale destra della matrice di grafici a dispersione. Questo grafico mostra che la seconda dimensione (sull'asse *y*) corrisponde alla variabile *grado*, con valori più elevati lungo l'asse corrispondente ai termini più lontani dalla famiglia "nucleare".

Una soluzione a tre dimensioni con trasformazioni non predefinite

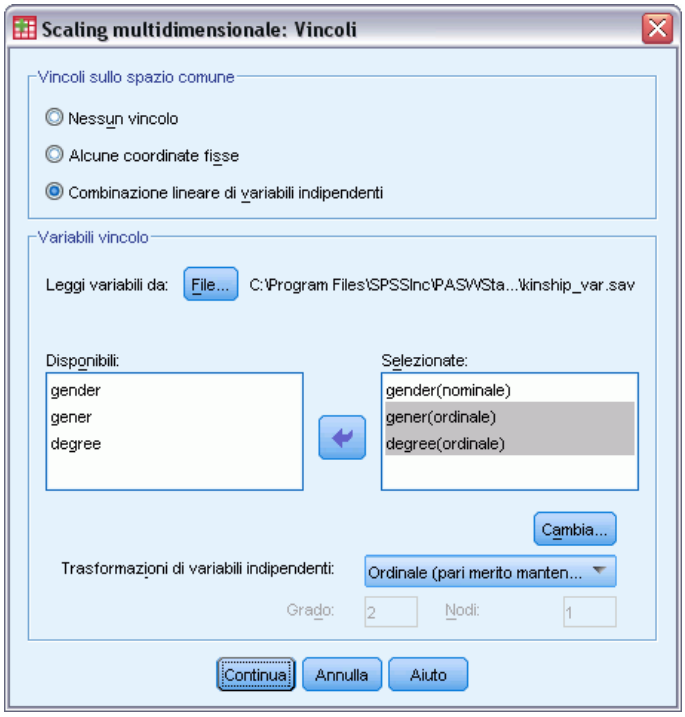
La soluzione precedente è stata calcolata utilizzando la trasformazione del rapporto predefinita per le distanze e le trasformazioni di intervallo per le variabili indipendenti *gener* e *grado*. I risultati sono buoni, ma potrebbe essere possibile migliorarli ulteriormente utilizzando altre trasformazioni. Ad esempio, le distanze, *gener* e *grado* hanno tutte ordinamenti naturali, ma potrebbe essere possibile creare un modello migliore tramite una trasformazione ordinale, anziché lineare.

Figura 14-16
Finestra di dialogo Modello



- Per ripetere l'analisi, scalando le distanze, *gener* e *grado* a livello ordinale (mantenendo i pari merito), richiamare la finestra di dialogo Scaling multidimensionale e fare clic su Modello:
- Selezionare Ordinale come trasformazione della distanza.
- Fare clic su Continua.
- Nella finestra di dialogo Scaling multidimensionale fare clic su Vincoli.

Figura 14-17
Finestra di dialogo Vincoli



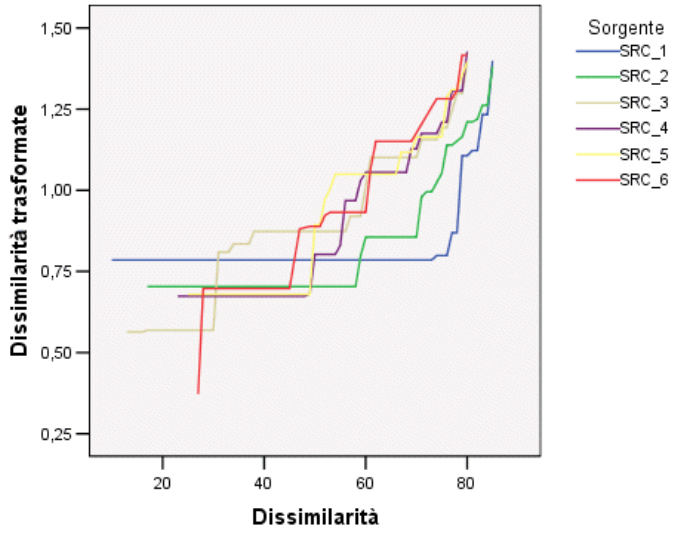
- Selezionare *gener* e *grado*.
- Selezionare Ordinale (mantieni pari merito) dall'elenco a discesa Trasformazioni di variabili indipendenti.
- Fare clic su Cambia.
- Fare clic su Continua.
- Nella finestra di dialogo Scaling multidimensionale fare clic su OK.

Grafici di trasformazione

I grafici di trasformazione sono utili per una prima verifica circa l'adeguatezza delle trasformazioni originali. Se i grafici sono approssimativamente lineari, l'ipotesi lineare è adeguata. In caso contrario, è necessario verificare le misure di stress per vedere se è presente un miglioramento nell'adattamento e controllare il grafico dello spazio comune per verificare se l'interpretazione risulta più utile.

Le variabili indipendenti ottengono ciascuna trasformazioni approssimativamente lineari, perciò potrebbe essere indicato interpretarle come numeriche. Tuttavia, le distanze non ottengono una trasformazione lineare, quindi è possibile che per esse la trasformazione ordinale sia più adatta.

Figura 14-18
Distanze trasformate



Misure di stress

Lo stress per la soluzione corrente supporta l'argomento relativo allo scaling delle distanze a livello ordinale.

Figura 14-19
Misure di stress e di adattamento

Stress non trasformato normalizzato	,03137
Stress-I	,17712[a]
Stress-II	,61987[a]
S-stress	,07953[b]
Dispersione spiegata	,96863
Coefficiente di convergenza di Tucker	,98419

PROXSCAL minimizza lo stress non trasformato normalizzato.

a. Livello di scala ottimale = 1,032.

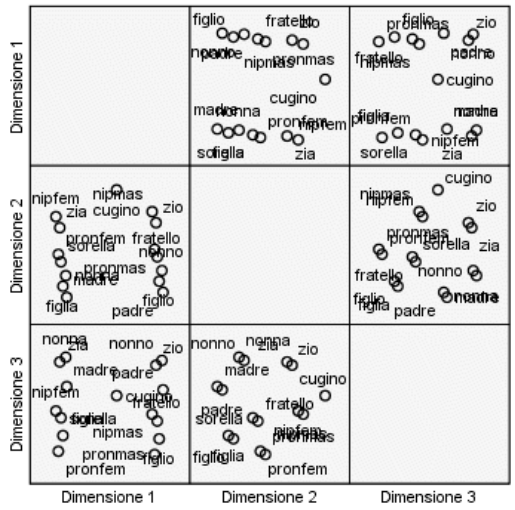
b. Livello di scala ottimale = ,980.

Il raw stress normalizzato per la soluzione precedente è 0,06234. Lo scaling delle variabili utilizzando trasformazioni non predefinite determina uno stress pari a 0,03137.

Coordinate finali dello spazio comune

I grafici dello spazio comune offrono essenzialmente la stessa interpretazione delle dimensioni della soluzione precedente.

Figura 14-20
Coordinate spazio comune



Discussione

Il metodo ottimale è trattare le distanze come variabili ordinali, perché si ottiene così un notevole miglioramento nelle misure di stress. Come passaggio successivo, è possibile che si desideri "distinguere" le variabili ordinali—ovvero, consentire che valori equivalenti delle variabili originali ottengano diversi valori trasformati. Ad esempio, nella prima sorgente, le distanze tra *Zia* e *Figlio* e tra *Zia* e *Nipote maschio di nonni*sono pari a 85. L'approccio "a pari merito" alle variabili ordinali forza l'equivalenza tra i valori trasformati di queste distanze, ma non c'è alcuna ragione particolare per presumere che questo sia corretto. In questo caso, consentendo la distinzione delle distanze si elimina un vincolo superfluo.

Letture consigliate

Consultare i testi seguenti per ulteriori informazioni sullo scaling multidimensionale:

Commandeur, J. J. F., e W. J. Heiser. 1993. *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices.* Leiden: Department of Data Theory, University of Leiden.

De Leeuw, J., e W. J. Heiser. 1980. Multidimensional scaling with restrictions on the configuration. In: *Multivariate Analysis, Vol. V,* P. R. Krishnaiah, ed. Amsterdam: North-Holland.

Heiser, W. J. 1981. *Unfolding analysis of proximity data.* Leiden: Department of Data Theory, University of Leiden.

Heiser, W. J., e F. M. T. A. Busing. 2004. Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, ed. Thousand Oaks, Calif.: Sage Publications, Inc..

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, .

Kruskal, J. B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, .

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika*, 27, .

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function II. *Psychometrika*, 27, .

Capitolo
15

Unfolding multidimensionale

La procedura Unfolding multidimensionale tenta di individuare una scala quantitativa comune che consenta di analizzare visivamente le relazioni tra due insiemi di oggetti.

Esempio preferenze relative ai cibi da colazione

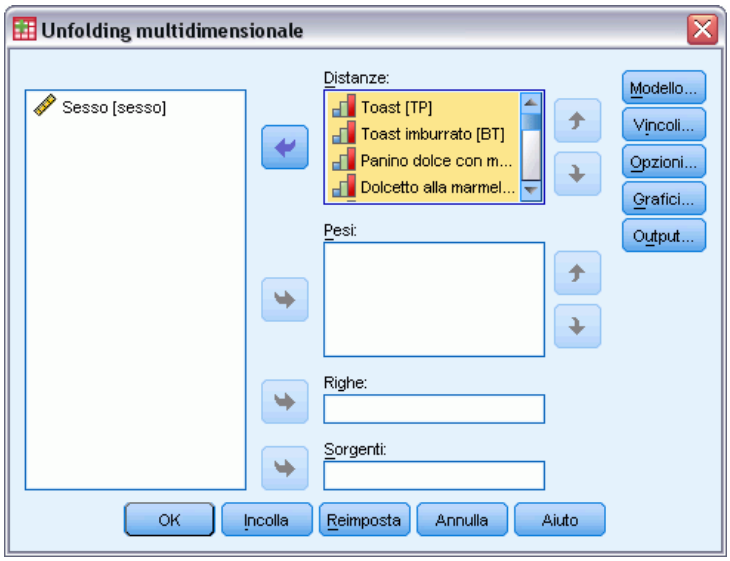
In uno studio classico (Green e Rao, 1972), è stato chiesto a 21 studenti MBA della Wharton School e ai loro consorti di classificare 15 cibi da colazione in ordine di preferenza, dove 1 era l'alimento preferito in assoluto e 15 quello meno preferito. Tali informazioni vengono raccolte nel file *breakfast_overall.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*.

I risultati dello studio forniscono un tipico esempio del problema della degenerazione tipico di molti algoritmi impiegati per l'unfolding multidimensionale, che viene generalmente risolto penalizzando il coefficiente di variazione delle distanze trasformate (Busing, Groenen, e Heiser, 2005). L'esempio mostra come individuare la soluzione degenerata e come risolvere il problema utilizzando l'unfolding multidimensionale, che permette di stabilire in che modo i singoli classificano i cibi da colazione. La sintassi per l'esecuzione di queste analisi è contenuta nel file *prefscal_breakfast-overall.sps*.

Creazione di una soluzione degenerata

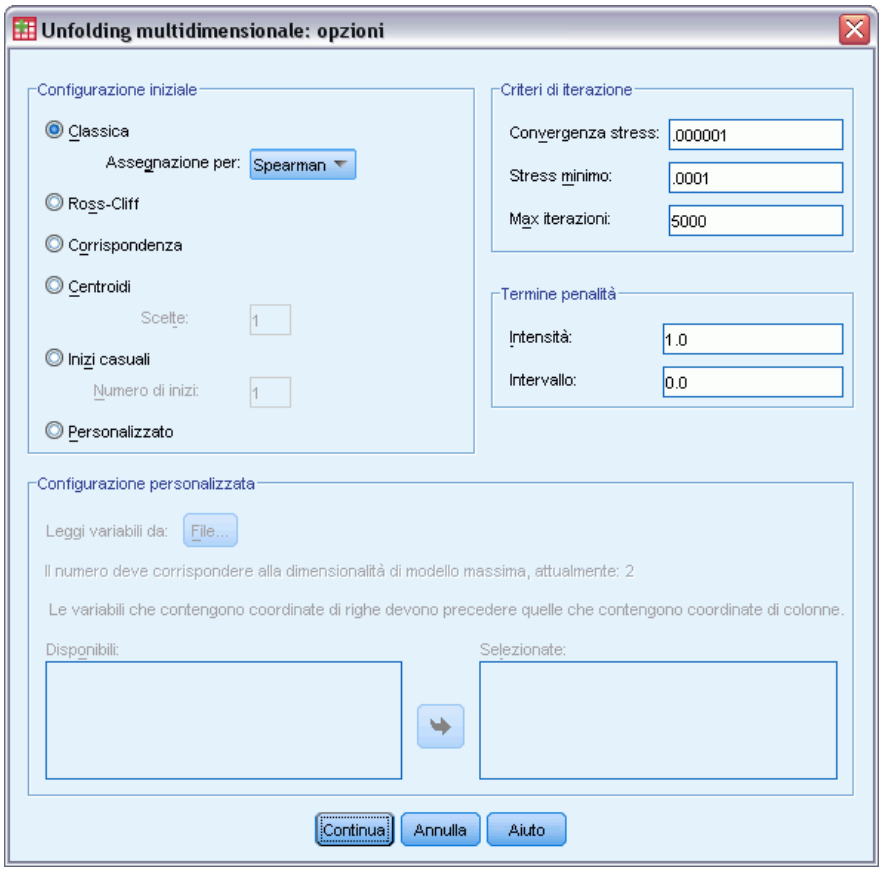
- Per eseguire un'analisi di unfolding multidimensionale, dai menu scegliere:
  Analizza > Scala > Unfolding multidimensionale (PREFSCAL)...

Figura 15-1
Finestra di dialogo principale Unfolding multidimensionale



- Selezionare le variabili di distanza da *Pane da tostare* a *Muffin e burro*.
- Fare clic su Opzioni.

Figura 15-2
Finestra di dialogo Opzioni



- Selezionare Spearman come metodo di assegnazione per il punto iniziale tradizionale.
- Nel gruppo Termine penalità digitare 1.0 come valore del parametro Intensità e 0.0 come valore del parametro Intervallo. In tal modo il termine di penalità viene disattivato.
- Fare clic su Continua.
- Nella finestra di dialogo Unfolding multidimensionale fare clic su OK.

Segue la sintassi di comandi generata da queste selezioni:

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INITIAL=CLASSICAL (SPEARMAN)
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(1.0) OMEGA(0.0)
  /PRINT=MEASURES COMMON
  /PLOT=COMMON .
```

- Questa sintassi specifica un'analisi sulle variabili da *tp (pane da tostare)* a *cmb (muffin con burro)*.

- Il sottocomando `INITIAL` specifica che i valori iniziali possono essere immessi usando le distanze di Spearman.
- I valori specificati nel sottocomando `PENALTY` disabilitano la penalità, quindi la procedura minimizza lo stress I di Kruskal, provocando la creazione di una soluzione degenerata.
- Il sottocomando `PLOT` richiede i grafici per lo spazio comune.
- Tutti gli altri parametri vengono impostati sui valori predefiniti.

Misure

Figura 15-3
Misure per la soluzione degenerata

Iterazioni		154
Valore finale della funzione		,0000990
Parti dei valori della funzione	Parte dello stress	,0000990
	Parte penalità	1,0000000
Cattivo adattamento	Stress normalizzato	,0000000
	Stress I di Kruskal	,0000990
	Stress II di Kruskal	,6129749
	S-stress I di Young	,0001980
	S-stress II di Young	,7703817
Bontà dell'adattamento	Dispersione spiegata in base a	1,0000000
	Varianza spiegata in base a	,6230788
	Ordini delle preferenze recuperate	,7074830
	Rho di Spearman	,7450748
	Tau-b di Kendall	,6218729
Coefficienti di variazione	Vicinanze delle variazioni	,5590170
	Vicinanze trasformate delle variazioni	,0000924
	Distanze delle variazioni	,1808765
Indici di degenerazione	Somma dei quadrati degli indici di amalgamazione di DeSarbo	117,3115413
	Indice di non degenerazione approssimativo di Shepard	,0000000

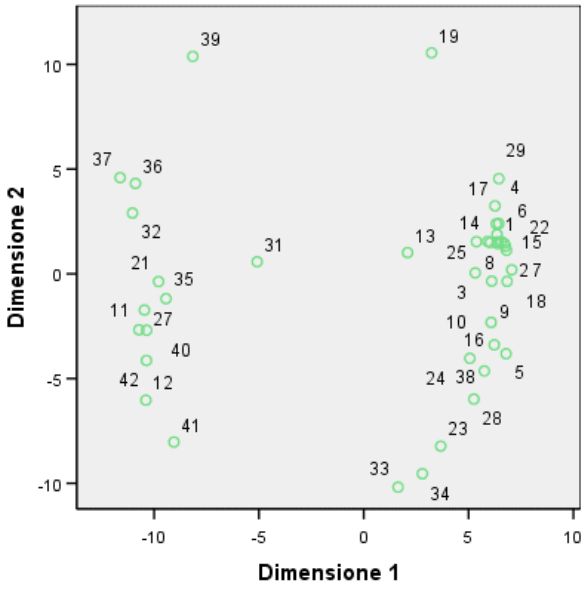
L'algoritmo arriva a una soluzione dopo 154 iterazioni con una misura dello stress penalizzata (contrassegnata come valore funzionale finale) pari allo 0,0000990. Poiché la penalità è stata disabilitata, lo stress penalizzato è uguale allo stress I di Kruskal (la parte dello stress del valore della funzione è equivalente all'inadeguatezza dell'adattamento di Kruskal). Valori bassi di stress

indicano che la soluzione si adatta bene ai dati, ma che ci sono numerosi segnali che indicano la presenza di una soluzione degenerata:

- il coefficiente di variazione delle distanze trasformate è molto piccolo rispetto al coefficiente di variazione delle distanze originali. Ciò suggerisce che le distanze trasformate di ciascuna riga sono quasi costanti e che, conseguentemente, la soluzione non consente di discriminare gli oggetti.
- La somma dei quadrati degli indici di intervariabilità di DeSarbo indica il livello di variabilità dei punti dei diversi insiemi. Se non ci sono oggetti distribuiti in modo variabile, è possibile che la soluzione sia degenerata. Più vicino il risultato è a 0 e maggiore è la probabilità che la soluzione contenga oggetti con una distribuzione variabile. Poiché in questo caso il valore risultante è molto alto, la soluzione non contiene oggetti con una distribuzione variabile.
- L'indice di non-degenerazione approssimativo di Shepard, espresso come percentuale delle diverse distanze, è uguale a 0. Ciò indica che le distanze non sono abbastanza diverse e che la soluzione è probabilmente degenerata.

Spazio comune

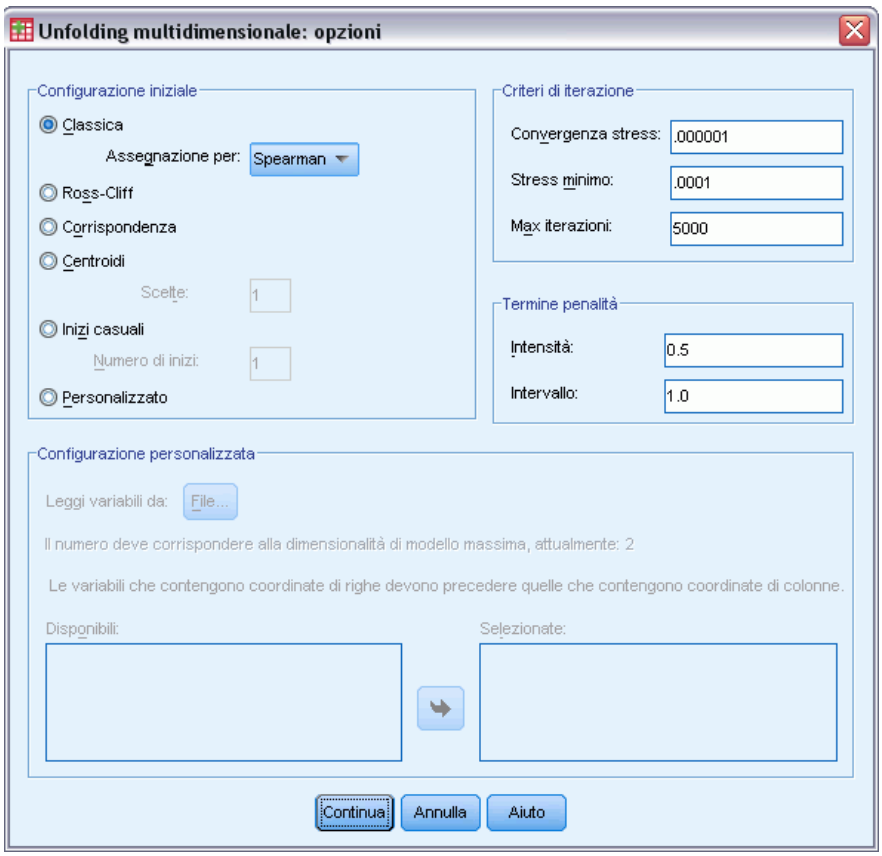
Figura 15-4
Grafico congiunto dello spazio comune per la soluzione degenerata



La conferma visiva della degenerazione della soluzione viene fornita dal grafico congiunto dello spazio comune degli oggetti riga e colonna. Gli oggetti riga (singoli) sono situati lungo la circonferenza di un cerchio centrato sugli oggetti colonna (cibi per colazione), le cui coordinate sono state compresse in un unico punto.

Esecuzione di un'analisi Non degenerata

Figura 15-5
Finestra di dialogo Opzioni



- Per produrre una soluzione non degenerata, fare clic sullo strumento Richiama finestra e selezionare Unfolding multidimensionale.
- Nella finestra di dialogo Unfolding multidimensionale fare clic su Opzioni.
- Nel gruppo Termine penalità digitare 0.5 come valore del parametro Intensità e 1.0 come valore del parametro Intervallo. In tal modo il termine di penalità viene disattivato.
- Fare clic su Continua.
- Nella finestra di dialogo Unfolding multidimensionale fare clic su OK.

Segue la sintassi di comandi generata da queste selezioni:

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INITIAL=CLASSICAL (SPEARMAN)
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(0.5) OMEGA(1.0)
  /PRINT=MEASURES COMMON
```

```
/PLOT=COMMON .
```

- L'unica variazione è contenuta nel sottocomando PENALTY. LAMBDA è stato impostato su 0.5 e OMEGA su 1.0 (valori predefiniti).

Misure

Figura 15-6
Misure per la soluzione non degenerata

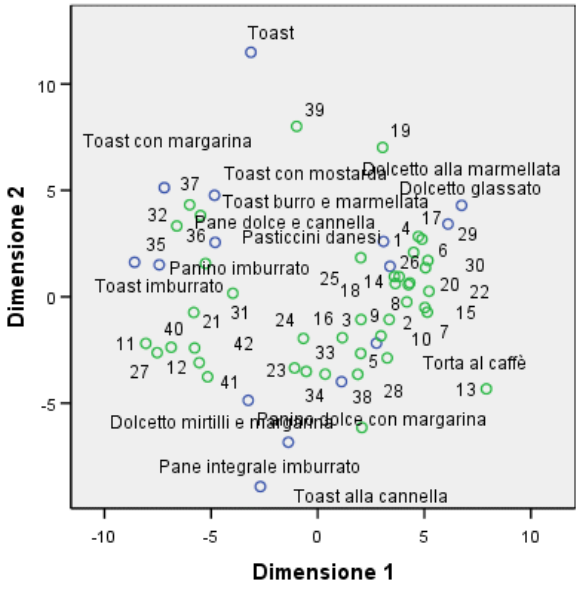
Iterazioni		157
Valore finale della funzione		,6848930
Parti dei valori della funzione	Parte dello stress	,2428268
	Parte penalità	1,9317409
Cattivo adattamento	Stress normalizzato	,0583589
	Stress I di Kruskal	,2415758
	Stress II di Kruskal	,5875599
	S-stress I di Young	,3446361
	S-stress II di Young	,5030127
Bontà dell'adattamento	Dispersione spiegata in base a	,9416411
	Varianza spiegata in base a	,7651552
	Ordini delle preferenze recuperate	,7818594
	Rho di Spearman	,8179181
	Tau-b di Kendall	,6916725
Coefficienti di variazione	Vicinanze delle variazioni	,5590170
	Vicinanze trasformate delle variazioni	,6006156
	Distanze delle variazioni	,4833617
Indici di degenerazione	Somma dei quadrati degli indici di amalgamazione di DeSarbo	,1590979
	Indice di non degenerazione approssimativo di Shepard	,7895692

Nell'esempio i problemi notati nelle misure relative alla soluzione degenerata sono stati corretti.

- Lo stress normalizzato non è più 0.
- Il coefficiente di variazione delle distanze trasformate ha adesso un valore molto simile al coefficiente di variazione delle distanze originali.
- Gli indici di intervariabilità di DeSarbo sono molto più vicini a 0, ad indicare che la soluzione ha un grado di intervariabilità migliore.
- L'indice di non-degenerazione generale di Shepard, espresso come percentuale delle diverse distanze, è ora pari a circa l'80%. Poiché le distanze sono sufficientemente diverse, la soluzione è probabilmente non degenerata.

Spazio comune

Figura 15-7
Grafico congiunto dello spazio comune per la soluzione degenerata



Il grafico congiunto dello spazio comune consente di interpretare le dimensioni. La dimensione orizzontale sembra mettere in evidenza una differenza tra pane morbido e duro o tostato, poiché i cibi morbidi tendono ad aumentare man mano che ci si sposta a destra dell'asse. La dimensione verticale non offre un'interpretazione chiara, sebbene fornisce probabilmente informazioni sulle differenze in termini di convenienza, considerato che i cibi più "tradizionali" tendono ad aumentare man mano che ci sposta lungo l'asse.

Ciò contribuisce a creare vari cluster di cibi per colazione. I bomboloni, i biscotti alla cannella e i pasticcini formano un cluster di cibi morbidi e non tradizionali. I muffin e il pane tostato alla cannella formano un cluster di cibi più duri ma più tradizionali. Gli altri tipi di pane e panini tostati formano un cluster di cibi duri e meno tradizionali. Il pane da tostare rientra nel cluster dei cibi duri e decisamente meno tradizionali.

Gli individui rappresentati dagli oggetti riga sono chiaramente suddivisi in cluster in base alle loro preferenze in termini di cibi duri o morbidi, con variazioni significative all'interno del cluster lungo la dimensione verticale.

Esempio unfolding a tre vie delle preferenze relative ai cibi da colazione

In uno studio classico (Green et al., 1972), è stato chiesto a 21 studenti MBA della Wharton School e ai loro consorti di classificare 15 cibi da colazione in ordine di preferenza, dove 1 era l'alimento preferito in assoluto e 15 quello meno preferito. Le loro preferenze sono state registrate per sei diversi scenari, che comprendevano tutti gli scenari compresi tra "Preferenza generale" e

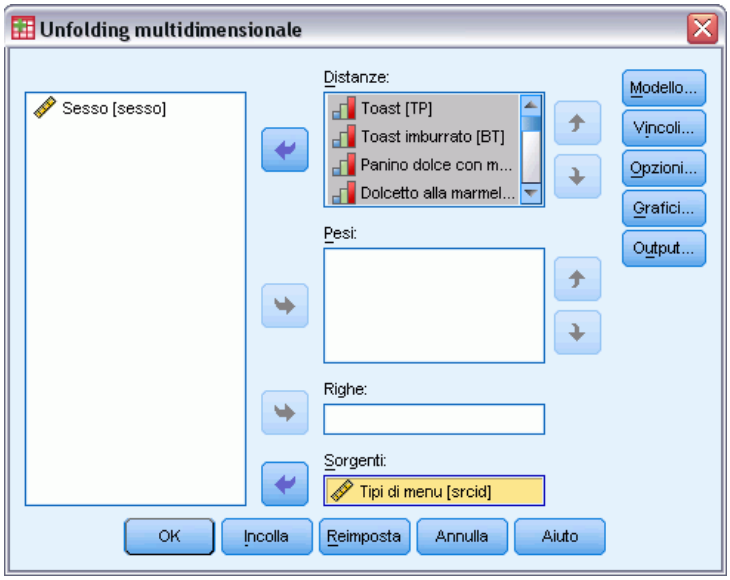
"Solo snack con bibita". Tali informazioni vengono raccolte nel file *breakfast.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*.

I sei scenari possono essere considerati sorgenti diverse. Usare PREFSCAL per eseguire l'unfolding a tre vie delle righe, colonne e sorgenti. La sintassi utilizzabile per riprodurre queste analisi è contenuta in *prefscal_breakfast.sps*.

Esecuzione dell'analisi

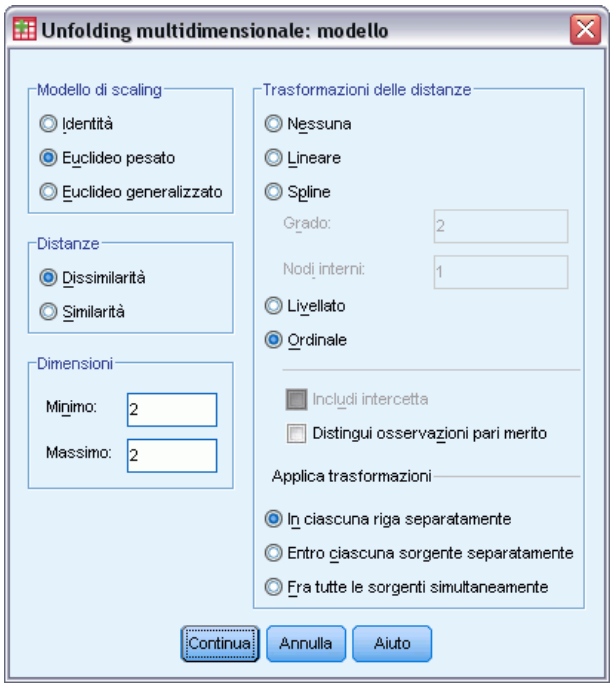
- Per eseguire un'analisi di unfolding multidimensionale, dai menu scegliere:
  Analizza > Scala > Unfolding multidimensionale (PREFSCAL)...

Figura 15-8
Finestra di dialogo principale Unfolding multidimensionale



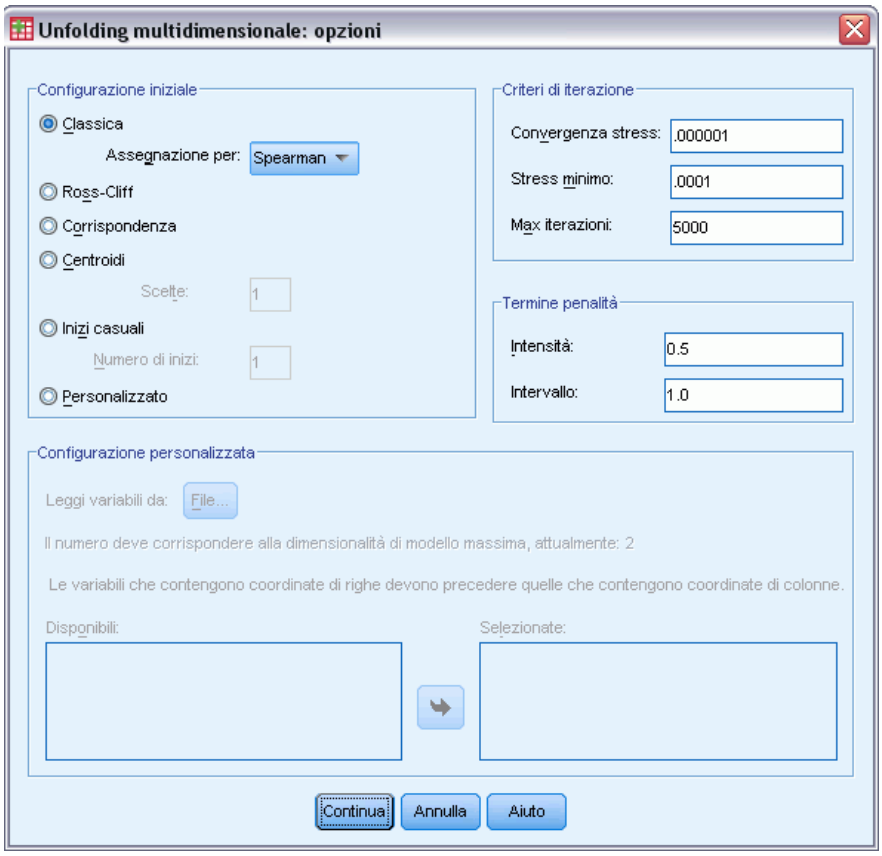
- Selezionare le variabili di distanza da *Pane da tostare* a *Muffin e burro*.
- Selezionare *$$$Menu scenarios* come variabile sorgente.
- Fare clic su Modello.

Figura 15-9
Finestra di dialogo Modello



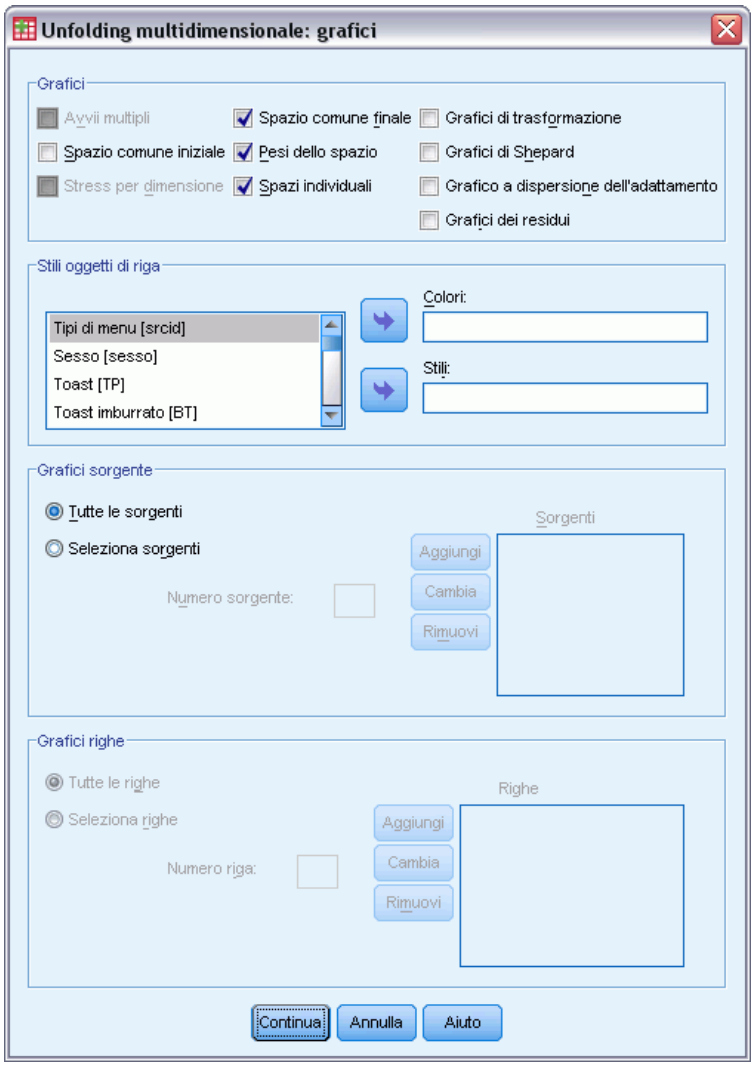
- Selezionare Euclideo pesato come modello di scaling.
- Fare clic su Continua.
- Nella finestra di dialogo Unfolding multidimensionale fare clic su Opzioni.

Figura 15-10
Finestra di dialogo Opzioni



- Selezionare Spearman come metodo di assegnazione per il punto iniziale tradizionale.
- Fare clic su Continua.
- Nella finestra di dialogo Unfolding multidimensionale fare clic su Grafici.

Figura 15-11
Finestra di dialogo Grafici



- Nel gruppo Grafici selezionare Spazi individuali.
- Fare clic su Continua.
- Nella finestra di dialogo Unfolding multidimensionale fare clic su OK.

Segue la sintassi di comandi generata da queste selezioni:

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INPUT=SOURCES(srcid )
  /INITIAL=CLASSICAL (SPEARMAN)
  /CONDITION=ROW
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /MODEL=WEIGHTED
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
```

```
/PENALTY=LAMBDA(0.5) OMEGA(1.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON WEIGHTS INDIVIDUAL ( ALL ) .
```

- Questa sintassi specifica un'analisi sulle variabili da *tp (pane da tostare)* a *cmb (muffin con burro)*. La variabile *srcid* viene usata per identificare le sorgenti.
- Il sottocomando `INITIAL` specifica che i valori iniziali possono essere immessi usando le distanze di Spearman.
- Il sottocomando `MODEL` specifica un modello euclideo ponderato che permette a ciascuno spazio individuale di ponderare in modo diverso le dimensioni dello spazio comune.
- `PLOT` richiede i grafici per lo spazio comune, gli spazi individuali e i pesi dello spazio individuale.
- Tutti gli altri parametri vengono impostati sui valori predefiniti.

Misure

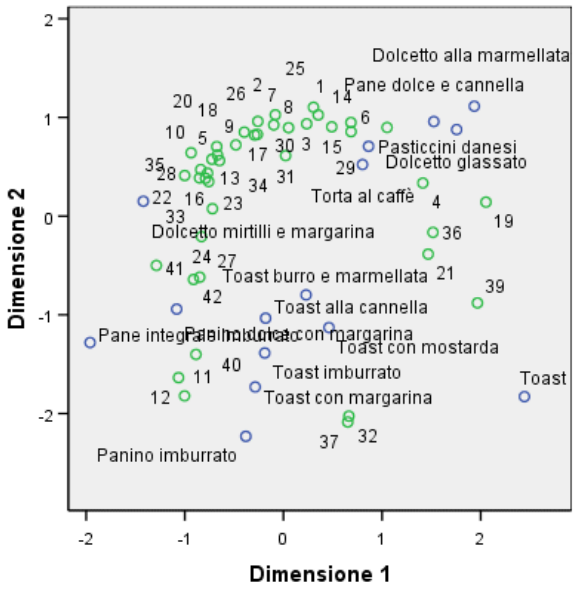
Figura 15-12
Misure

Iterazioni		481
Valore finale della funzione		,8199642
Parti dei valori della funzione	Parte dello stress	,3680994
	Parte penalità	1,8265211
Cattivo adattamento	Stress normalizzato	,1335343
	Stress I di Kruskal	,3654234
	Stress II di Kruskal	,9780824
	S-stress I di Young	,4938016
	S-stress II di Young	,6912352
Bontà dell'adattamento	Dispersione spiegata in base a	,8664657
	Varianza spiegata in base a	,5024853
	Ordini delle preferenze recuperate	,7025321
	Rho di Spearman	,6271702
	Tau-b di Kendall	,4991188
Coefficienti di variazione	Vicinanze delle variazioni	,5590170
	Vicinanze trasformate delle variazioni	,6378878
	Distanze delle variazioni	,4484515
Indici di degenerazione	Somma dei quadrati degli indici di amalgamazione di DeSarbo	,2199287
	Indice di non degenerazione approssimativo di Shepard	,7643613

L'algoritmo converge dopo 481 interazioni, con uno stress penalizzato finale pari a 0,8199642. I coefficienti di variazione e l'indice di Shepard sono sufficientemente grandi e gli indici di DeSarbo sono sufficientemente piccoli, ad indicare che non ci sono problemi di degenerazione.

Spazio comune

Figura 15-13
Grafico congiunto dello spazio comune



Il grafico congiunto dello spazio comune mostra una configurazione finale che è molto simile all'analisi a due vie relativa alle preferenze generali, in cui la soluzione risulta capovolta rispetto alla riga dei 45 gradi. Quindi, la dimensione orizzontale mette in evidenza una differenza tra pane morbido e duro o tostato, poiché i cibi morbidi tendono ad aumentare man mano che ci si sposta verso la parte superiore dell'asse. La dimensione orizzontale non offre un'interpretazione chiara, anche se fornisce informazioni sulle differenze in termini di convenienza, poiché i cibi più "tradizionali" tendono ad aumentare man mano che ci sposta verso il lato sinistro dell'asse.

Gli individui rappresentati dagli oggetti riga sono chiaramente suddivisi in cluster in base alle loro preferenze in termini di cibi "duri" o "morbidi", con variazioni significative all'interno del cluster lungo la dimensione orizzontale.

Spazi individuali

Figura 15-14
Pesi di dimensione

		Dimensione		
		1	2	Specificità[a]
Sorgente	1	3,235	4,297	,186
	2	4,883	2,193	,457
	3	4,131	3,438	,109
	4	4,291	3,267	,164
	5	3,124	4,413	,223
	6	2,750	4,541	,313
Importanza[b]		.504	.496	

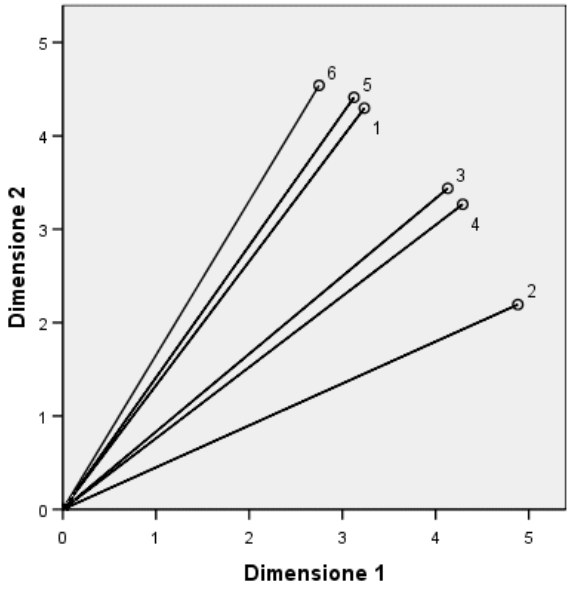
a. La specificità indica la tipicità di una sorgente. Il range della specificità è compreso tra zero e uno, dove zero indica una sorgente media con pesi di dimensioni identici e uno indica una sorgente molto specifica con un peso di dimensione eccezionale e altri pesi vicini a zero.

b. Importanza relativa di ciascuna dimensione, indicata come rapporto tra la somma dei quadrati di una dimensione e il totale della somma dei quadrati.

Gli spazi individuali vengono calcolati per ciascuna sorgente. I pesi di dimensione mostrano il peso dei singoli spazi sulle dimensioni dello spazio comune. Un peso alto indica una maggiore distanza dallo spazio individuale e quindi una maggiore differenza in termini di spazio individuale tra gli oggetti della dimensione.

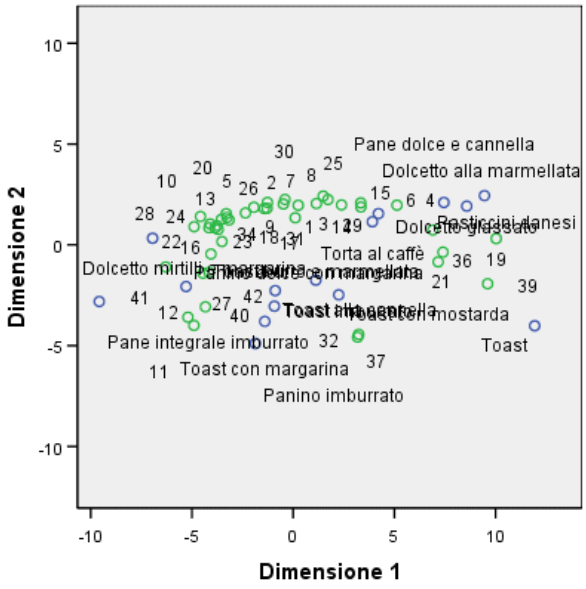
- La **specificità** indica il grado di diversità dello spazio individuale rispetto a quello comune. Uno spazio individuale identico a quello comune ha generalmente pesi di dimensione identici e una specificità pari a 0, mentre uno spazio individuale riferito a una dimensione specifica ha generalmente pesi di dimensione maggiori e una specificità pari a 1. In questo caso le origini più divergenti sono *Colazione con succo, pancetta, uova e bibita* e *Snack con bibita*.
- L'**importanza** è la misura del contributo relativo di ciascuna dimensione alla soluzione. In questo caso le dimensioni hanno la stessa importanza.

Figura 15-15
Pesi di dimensione



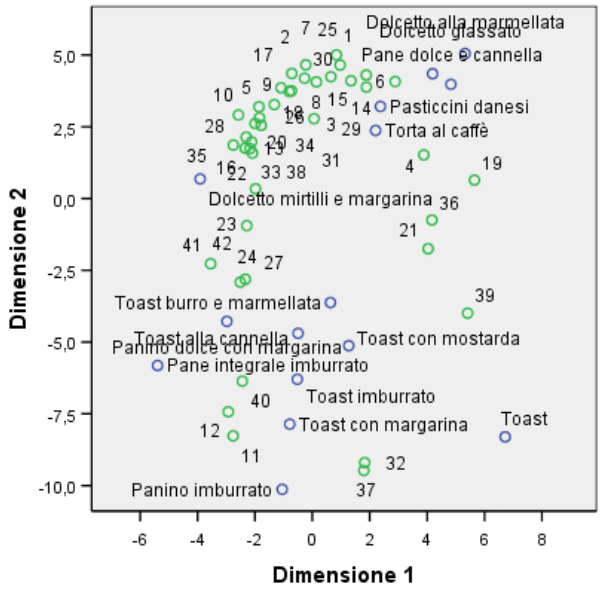
Il grafico Pesi di dimensione fornisce un'indicazione visiva della tabella dei pesi. Le origini *Colazione con succo, pancetta, uova e bibita* e *Snack con bibita* sono quelle più vicine agli assi delle dimensioni, ma nessuna delle due è specifica per una dimensione.

Figura 15-16
Grafico congiunto dello spazio individuale "Colazione con succo, pancetta, uova e bibita"



Il grafico congiunto dello spazio individuale *Colazione con succo, pancetta, uova e bibita* mostra l'effetto dello scenario sulle preferenze. Questa sorgente pesa molto di più sulla prima dimensione, quindi la differenziazione tra gli oggetti dipende soprattutto da questa dimensione.

Figura 15-17
Grafico congiunto dello spazio individuale "Snack con bibita"



Il grafico congiunto dello spazio individuale*Snack con bibita* mostra l'effetto di questo scenario sulle preferenze. Questa sorgente pesa molto di più sulla seconda dimensione, quindi la differenziazione tra gli oggetti dipende soprattutto da questa dimensione. Tuttavia, è presente una differenziazione piuttosto marcata rispetto alla prima dimensione, soprattutto a causa della bassa specificità dell'origine.

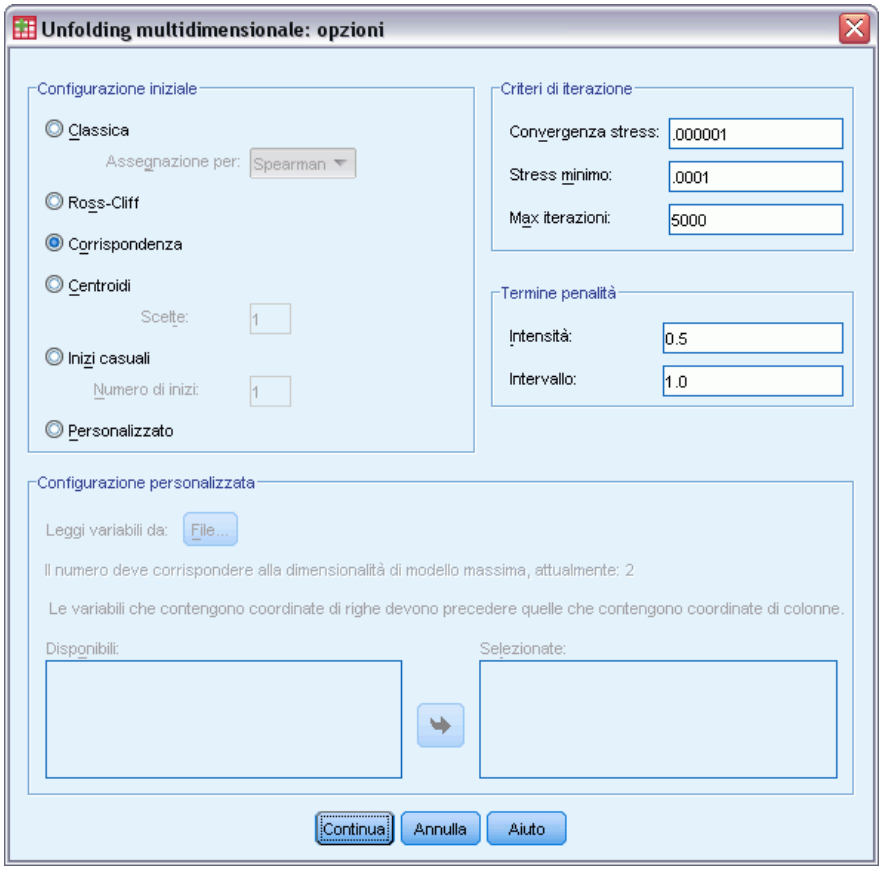
Uso di una configurazione iniziale diversa

La configurazione finale dipende dai punti iniziali assegnati all'algoritmo. La struttura generale di una soluzione dovrebbe idealmente rimanere invariata per consentire l'identificazione della soluzione corretta. Tuttavia, è possibile mettere in evidenza dettagli specifici provando a usare configurazioni iniziali diverse; ad esempio usando un inizio di corrispondenza per l'analisi a tre vie dei dati relativi alla colazione.

- Per produrre una soluzione con un inizio di corrispondenza, fare clic sullo strumento Richiama finestra e selezionare Unfolding multidimensionale.

▶ Nella finestra di dialogo Unfolding multidimensionale fare clic su Opzioni.

Figura 15-18
Finestra di dialogo Opzioni



▶ Selezionare Corrispondenza nel gruppo Configurazione iniziale.

▶ Fare clic su Continua.

▶ Nella finestra di dialogo Unfolding multidimensionale fare clic su OK.

Segue la sintassi di comandi generata da queste selezioni:

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INPUT=SOURCES(srcid )
  /INITIAL=CORRESPONDENCE
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(0.5) OMEGA(1.0)
  /PRINT=MEASURES COMMON
```

```
/PLOT=COMMON WEIGHTS INDIVIDUAL ( ALL ) .
```

- L'unica variazione è contenuta nel sottocomando `INITIAL`. La configurazione iniziale è stata impostata su `CORRESPONDENCE`, che usa i risultati dell'analisi della corrispondenza sui dati invertiti (similarità anziché dissimilarità) insieme a una normalizzazione simmetrica dei punteggi delle righe e delle colonne.

Misure

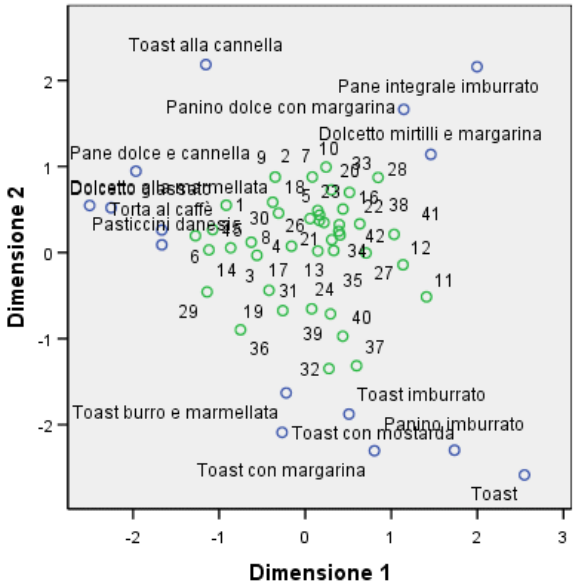
Figura 15-19
Misure per la configurazione iniziale della corrispondenza

Iterazioni		385
Valore finale della funzione		,8140741
Parti dei valori della funzione	Parte dello stress	,3493640
	Parte penalità	1,8969229
Cattivo adattamento	Stress normalizzato	,1212145
	Stress I di Kruskal	,3481587
	Stress II di Kruskal	1,0770522
	S-stress I di Young	,4812632
	S-stress II di Young	,6871733
Bontà dell'adattamento	Dispersione spiegata in base a	,8787855
	Varianza spiegata in base a	,5183498
	Ordini delle preferenze recuperate	,7174981
	Rho di Spearman	,6446272
	Tau-b di Kendall	,5165230
Coefficienti di variazione	Vicinanze delle variazioni	,5590170
	Vicinanze trasformate delle variazioni	,6122308
	Distanze delle variazioni	,4043695
Indici di degenerazione	Somma dei quadrati degli indici di amalgamazione di DeSarbo	1,7571887
	Indice di non degenerazione approssimativo di Shepard	,7532124

L'algoritmo converge dopo 385 interazioni, con uno stress penalizzato finale pari a 0,8140741. Questa statistica, l'inadeguatezza dell'adattamento, la bontà dell'adattamento, i coefficienti di variazione e l'indice di Shepard sono tutti molti simili a quelli riferiti alla soluzione ottenuta con il punto di inizio tradizionale di Spearman. Gli indici di DeSarbo sono diversi, poiché hanno un valore pari a 1,7571887 anziché 0,2199287, ad indicare che la soluzione che utilizza l'inizio di corrispondenza non è ben distribuita. Per vedere in che misura ciò influisce sulla soluzione, è sufficiente osservare il grafico congiunto dello spazio comune.

Spazio comune

Figura 15-20
Grafico congiunto dello spazio comune per la configurazione iniziale della corrispondenza



Il grafico congiunto dello spazio comune mostra una configurazione finale che è simile a quella risultante dall'analisi eseguita con la configurazione iniziale tradizionale di Spearman. Tuttavia, gli oggetti colonna (alimenti consumati a colazione) appaiono posizionati intorno agli oggetti riga (individui) anziché essere distribuiti.

Spazi individuali

Figura 15-21
Pesi di dimensioni per la configurazione iniziale della corrispondenza

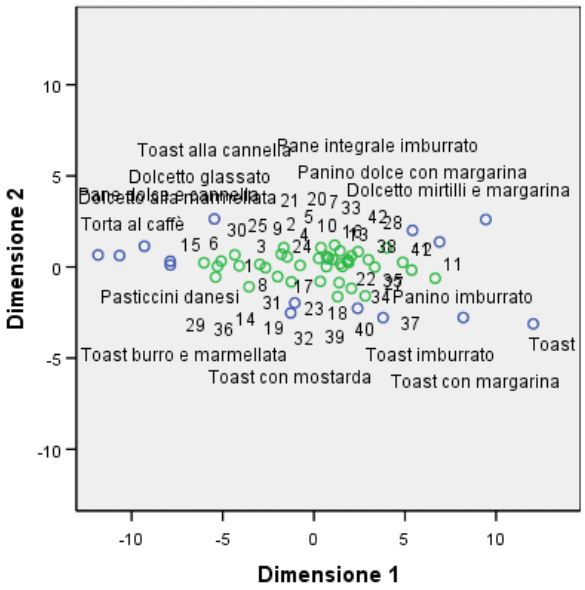
		Dimensione		
		1	2	Specificità[a]
Sorgente	1	2,836	3,877	,279
	2	4,727	1,207	,636
	3	4,183	2,377	,263
	4	4,412	1,993	,389
	5	2,605	4,050	,351
	6	1,864	4,415	,552
Importanza[b]		,556	,444	

a. La specificità indica la tipicità di una sorgente. Il range della specificità è compreso tra zero e uno, dove zero indica una sorgente media con pesi di dimensioni identici e uno indica una sorgente molto specifica con un peso di dimensione eccezionale e altri pesi vicini a zero.

b. Importanza relativa di ciascuna dimensione, indicata come rapporto tra la somma dei quadrati di una dimensione e il totale della somma dei quadrati.

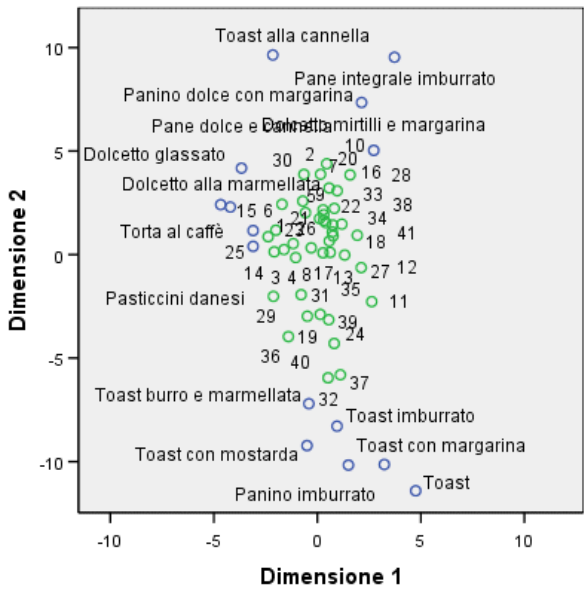
Nella configurazione di corrispondenza iniziale, ciascuno degli spazi individuali ha una specificità più alta. Ciò vuol dire che ciascun caso in cui i partecipanti hanno fornito delle preferenze relativamente ai cibi consumati a colazione sono più marcatamente associati a una dimensione specifica. Le sorgenti più divergenti rimangono *Colazione con succo, pancetta, uova e bibita* e *Snack con bibita*.

Figura 15-22
Grafico congiunto dello spazio individuale "Colazione con succo, pancetta, uova e bibita" per la configurazione iniziale della corrispondenza



La maggiore specificità è evidente nel grafico congiunto dello spazio individuale *Colazione con succo, pancetta, uova e bibita*. La sorgente pesa ancora più significativamente sulla prima dimensione rispetto al punto iniziale tradizionale di Spearman. Di conseguenza, gli oggetti riga e colonna mostrano una variazione minore lungo l'asse verticale e una variazione più marcata lungo l'asse orizzontale.

Figura 15-23
Grafico congiunto dello spazio individuale "Snack con bibita" per la configurazione iniziale della corrispondenza



Il grafico congiunto dello spazio individuale*Snack con bibita* mostra che gli oggetti riga e colonna sono più vicini alla riga verticale rispetto a quelli del grafico ottenuto con il punto iniziale tradizionale di Spearman.

Esempio analisi della correttezza dei comportamenti

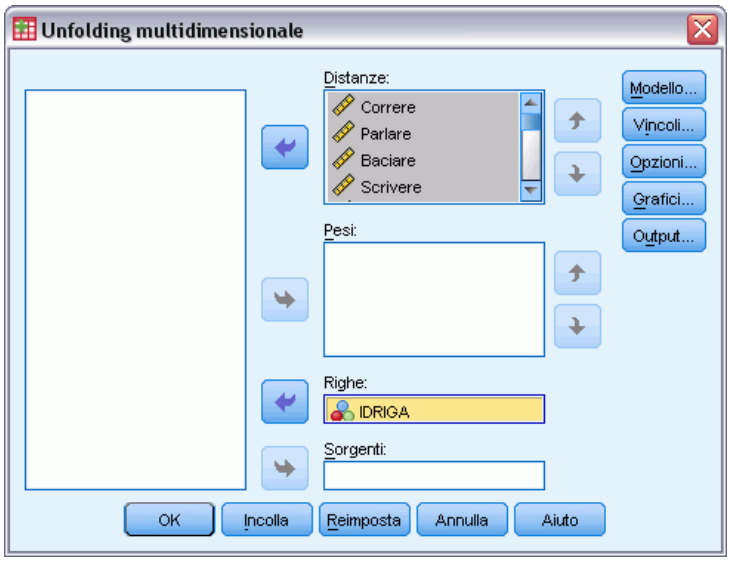
In un classico esempio (Prezzo e Bouffard, 1974), è stato chiesto a 52 studenti di classificare una combinazione di 15 situazioni e 15 comportamenti utilizzando una scala da 0="molto appropriato" a 9="molto inadeguato". I valori medi riferiti ai partecipanti sono stati considerati dissimilarità.

Tali informazioni vengono raccolte nel file *behavior.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*. Usare l'Unfolding multidimensionale per trovare i raggruppamenti di situazioni simili e i comportamenti più direttamente associati agli stessi. La sintassi utilizzabile per riprodurre queste analisi è contenuta in *prefscal_behavior.sps*.

Esecuzione dell'analisi

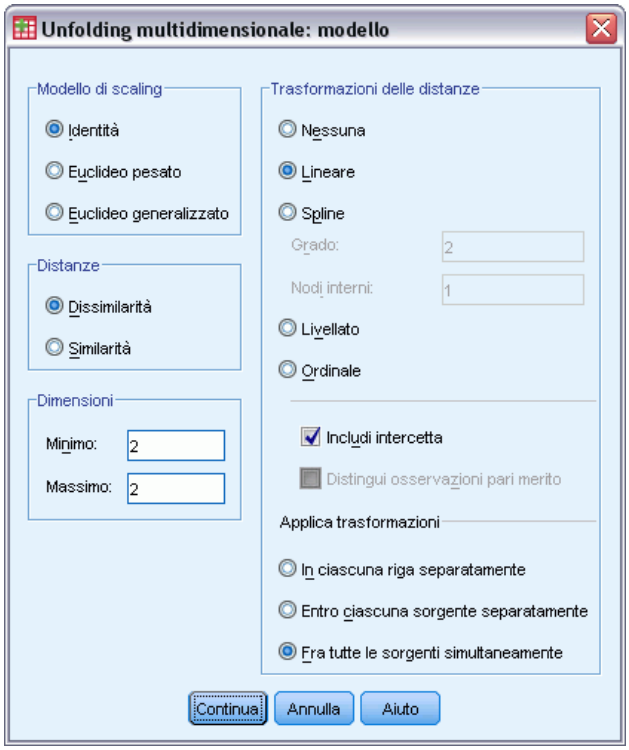
- Per eseguire un'analisi di unfolding multidimensionale, dai menu scegliere:
  Analizza > Scala > Unfolding multidimensionale (PREFSCAL)...

Figura 15-24
Finestra di dialogo principale Unfolding multidimensionale



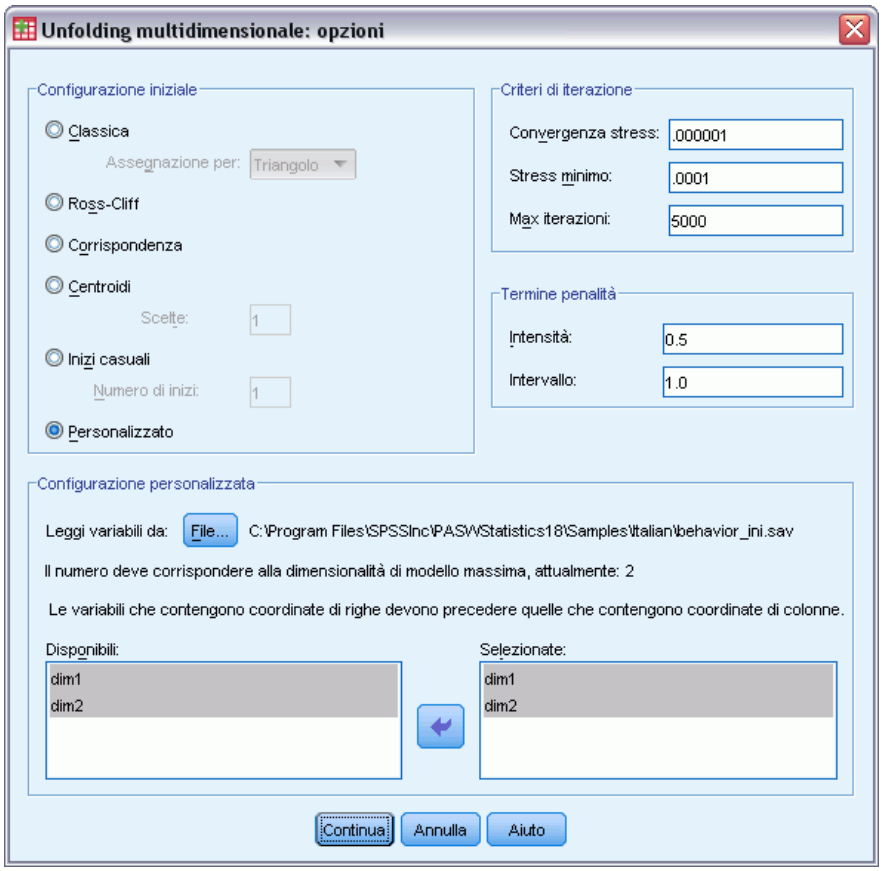
- Selezionare le variabili di distanza da *$$$Run* a *$$$Shout*.
- Selezionare *ROWID* come variabile di riga.
- Fare clic su Modello.

Figura 15-25
Finestra di dialogo Modello



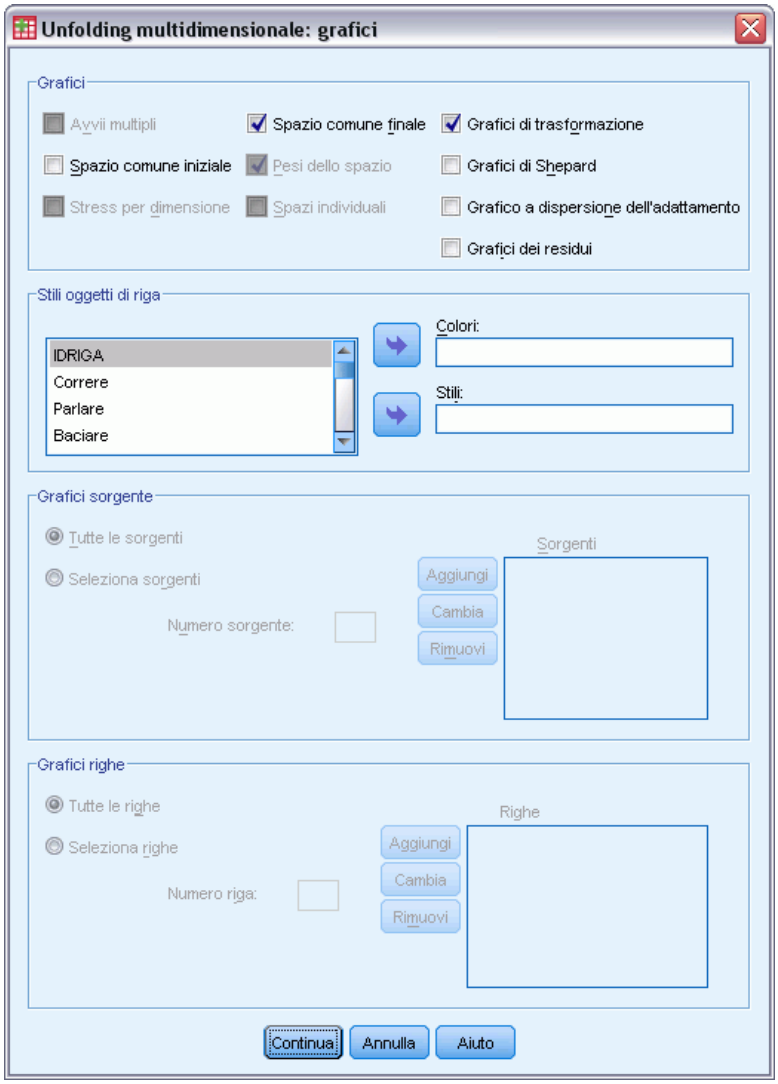
- Selezionare Lineare come trasformazione della distanza e scegliere Includi intercetta.
- Scegliere di applicare le trasformazioni Fra tutte le sorgenti simultaneamente.
- Fare clic su Continua.
- Nella finestra di dialogo Unfolding multidimensionale fare clic su Opzioni.

Figura 15-26
Finestra di dialogo Opzioni



- Selezionare Personalizzata nel gruppo Configurazione iniziale.
- Scegliere *behavior_ini.sav* come file contenente la configurazione personalizzata iniziale. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Categories 19*.
- Selezionare *dim1* e *dim2* come variabili che specificano la configurazione iniziale.
- Fare clic su Continua.
- Nella finestra di dialogo Unfolding multidimensionale fare clic su Grafici.

Figura 15-27
Finestra di dialogo Grafici



- Selezionare Grafici di trasformazione nel gruppo Grafici.
- Fare clic su Continua.
- Nella finestra di dialogo Unfolding multidimensionale fare clic su OK.

Segue la sintassi di comandi generata da queste selezioni:

```
PREFSCAL
  VARIABLES=Run Talk Kiss Write Eat Sleep Mumble Read Fight Belch Argue Jump
  Cry Laugh Shout
  /INPUT=ROWS(ROWID )
  /INITIAL=( 'samplesDirectory/behavior_ini.sav' )
  dim1 dim2
  /CONDITION=UNCONDITIONAL
  /TRANSFORMATION=LINEAR (INTERCEPT)
  /PROXIMITIES=DISSIMILARITIES
```

```
/MODEL=IDENTITY
/CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
MAXITER(5000)
/PENALTY=LAMBDA(0.5) OMEGA(1.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON TRANSFORMATIONS .
```

- La sintassi specifica un'analisi sulle variabili da *correre* a *saltare*. La variabile *idriga* viene usata per identificare le righe.
- Il sottocomando `INITIAL` specifica che i valori iniziali devono essere acquisiti dal file *behavior_ini.sav*. Le coordinate delle righe e delle colonne sono impilate, in modo che le coordinate delle colonne seguano quelle delle righe.
- Il sottocomando `CONDITION` specifica che tutte le distanze possono essere confrontate le une con le altre. Tutto ciò si applica alla nostra analisi, poiché è possibile confrontare le distanze per comportamenti quali correre nel parco e correre in chiesa, e stabilire quale comportamento è più appropriato.
- Il sottocomando `TRANSFORMATION` specifica una trasformazione lineare delle distanze con intercetta. Ciò è appropriato se la differenza pari a 1 punto nelle distanze è equivalente su tutta la scala dei 10 punti. In altre parole, la trasformazione lineare è appropriata se gli studenti hanno assegnato i punteggi in modo che la differenza tra 0 e 1 sia uguale alla differenza tra 5 e 6.
- Il sottocomando `PLOT` richiede i grafici per lo spazio comune e i grafici di trasformazione.
- Tutti gli altri parametri vengono impostati sui valori predefiniti.

Misure

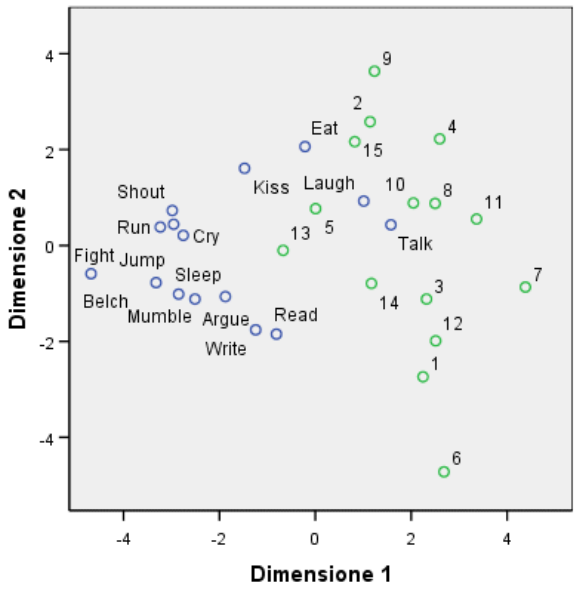
Figura 15-28
Misure

Iterazioni		169
Valore finale della funzione		,6427725
Parti dei valori della funzione	Parte dello stress	,1900001
	Parte penalità	2,1745069
Cattivo adattamento	Stress normalizzato	,0361000
	Stress I di Kruskal	,1900001
	Stress II di Kruskal	,5224668
	S-stress I di Young	,2760971
	S-stress II di Young	,4525933
Bontà dell'adattamento	Dispersione spiegata in base a	,9639000
	Varianza spiegata in base a	,8082862
	Ordini delle preferenze recuperate	,8608333
	Rho di Spearman	,8981120
	Tau-b di Kendall	,7202452
Coefficienti di variazione	Vicinanze delle variazioni	,5138436
	Vicinanze trasformate delle variazioni	,4751934
	Distanze delle variazioni	,3912592
Indici di degenerazione	Somma dei quadrati degli indici di amalgamazione di DeSarbo	,4957969
	Indice di non degenerazione approssimativo di Shepard	,7173810

L'algoritmo converge dopo 169 interazioni, con uno stress penalizzato finale pari a 0,6427725. I coefficienti di variazione e l'indice di Shepard sono sufficientemente grandi e gli indici di DeSarbo sono sufficientemente piccoli, ad indicare che non ci sono problemi di degenerazione.

Spazio comune

Figura 15-29
Grafico congiunto dello spazio comune

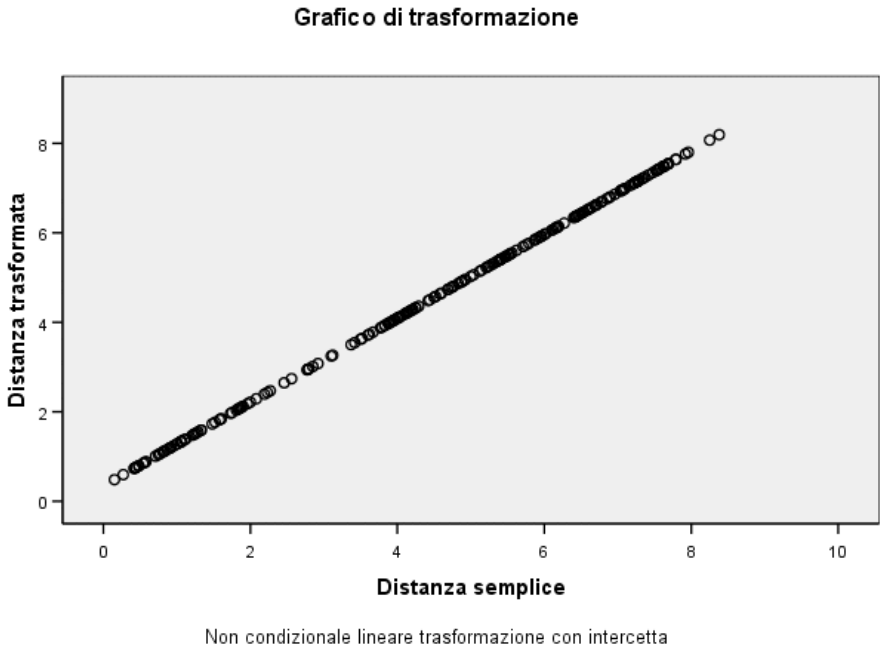


La dimensione orizzontale appare significativamente associata agli oggetti colonna (comportamenti) e consente di distinguere tra comportamenti inadeguati (litigare, fare rutti) e quelli più appropriati. La dimensione verticale appare significativamente correlata agli oggetti riga (situazioni) e definisce più limitazioni per i comportamenti relativi a situazioni specifiche.

- La parte finale della dimensione verticale riporta le situazioni (chiesa, classe) che limitano i comportamenti, ovvero che impongono comportamenti più posati (leggere, scrivere). Questi comportamenti compaiono verso l'estremità inferiore dell'asse verticale.
- La parte superiore della dimensione verticale mostra le situazioni (film, giochi, appuntamenti) che limitano i comportamenti, ovvero che impongono comportamenti più socievoli/estroversi (mangiare, baciare, ridere). Questi comportamenti compaiono verso l'estremità superiore dell'asse verticale.
- Al centro della dimensione verticale, le situazioni sono separate nella dimensione orizzontale in base alle caratteristiche limite della situazione. Quindi le situazioni più lontane dai comportamenti (intervista) sono quelli associate a situazioni più limitative, mentre quelle più vicine ai comportamenti (stanza, parco) sono meno limitative.

Trasformazioni delle distanze

Figura 15-30
Grafico di trasformazione



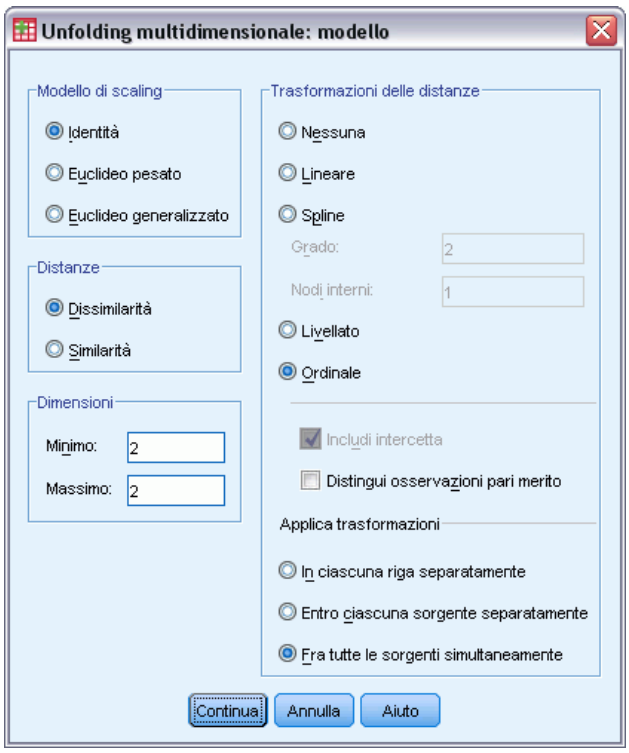
Le distanze sono state considerate lineari in quest'analisi, quindi il grafico che mostra il confronto tra i valori trasformati e le distanze originali sotto forma di riga lineare. L'adattamento di questa soluzione è buono, ma non esclude che si possa ottenere un adattamento migliore con una trasformazione diversa delle distanze.

Modifica delle trasformazioni delle distanze (ordinali)

- Per produrre una soluzione con una trasformazione ordinale delle distanze, fare clic sullo strumento Richiama finestra e selezionare Unfolding multidimensionale.

▶ Nella finestra di dialogo Unfolding multidimensionale fare clic su Modello.

Figura 15-31
Finestra di dialogo Modello



▶ Selezionare Ordinale come trasformazione della distanza.

▶ Fare clic su Continua.

▶ Nella finestra di dialogo Unfolding multidimensionale fare clic su OK.

Segue la sintassi di comandi generata da queste selezioni:

```
PREFSCAL
  VARIABLES=Run Talk Kiss Write Eat Sleep Mumble Read Fight Belch Argue Jump
  Cry Laugh Shout
  /INPUT=ROWS(ROWID )
  /INITIAL=( 'samplesDirectory/behavior_ini.sav' )
  dim1 dim2
  /CONDITION=UNCONDITIONAL
  /TRANSFORMATION=ORDINAL (KEEPTIES)
  /PROXIMITIES=DISSIMILARITIES
  /MODEL=IDENTITY
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(0.5) OMEGA(1.0)
  /PRINT=MEASURES COMMON
  /PLOT=COMMON TRANSFORMATIONS .
```

- L'unica variazione è contenuta nel sottocomando `TRANSFORMATION`. La trasformazione è stata impostata su `ORDINAL` che mantiene l'ordine delle distanze, ma non richiede che i valori trasformati siano proporzionali ai valori originali.

Misure

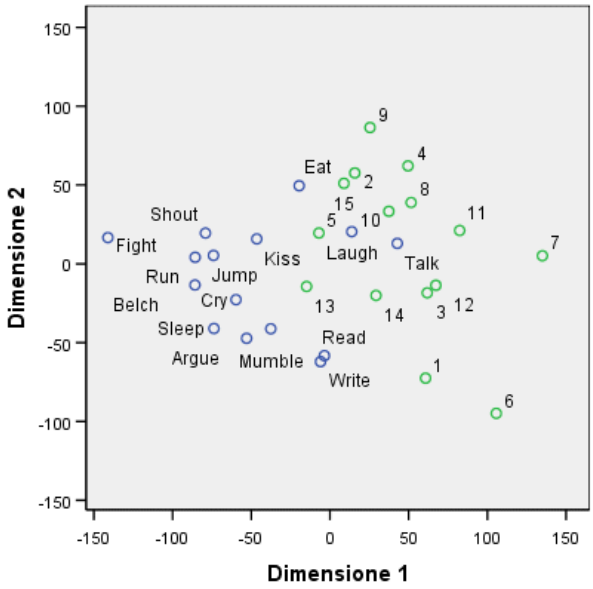
Figura 15-32
Misure per la soluzione con trasformazione ordinale

Iterazioni		268
Valore finale della funzione		,6044671
Parti dei valori della funzione	Parte dello stress	,1747239
	Parte penalità	2,0911875
Cattivo adattamento	Stress normalizzato	,0305285
	Stress I di Kruskal	,1747239
	Stress II di Kruskal	,4444641
	S-stress I di Young	,2707147
	S-stress II di Young	,3978003
Bontà dell'adattamento	Dispersione spiegata in base a	,9694715
	Varianza spiegata in base a	,8454488
	Ordini delle preferenze recuperate	,8574206
	Rho di Spearman	,9032676
	Tau-b di Kendall	,7532788
Coefficienti di variazione	Vicinanze delle variazioni	,5138436
	Vicinanze trasformate delle variazioni	,4930018
	Distanze delle variazioni	,4284849
Indici di degenerazione	Somma dei quadrati degli indici di amalgamazione di DeSarbo	,3610680
	Indice di non degenerazione approssimativo di Shepard	,7469048

L'algoritmo converge dopo 268 interazioni, con uno stress penalizzato finale pari a 0,6044671. Questa statistica e le altre misure sono leggermente migliori per questa soluzione rispetto a quella che prevede la trasformazione lineare delle distanze.

Spazio comune

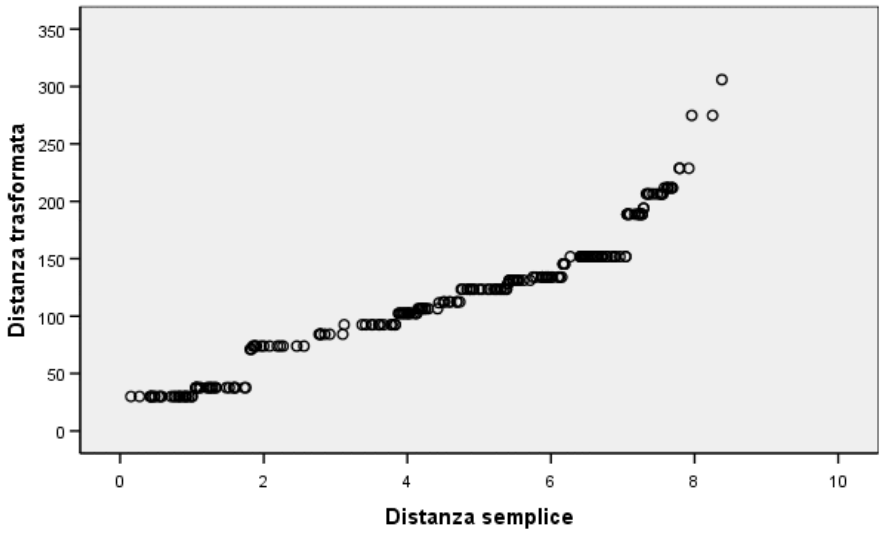
Figura 15-33
Grafico congiunto dello spazio comune per la soluzione con trasformazione ordinale



L'interpretazione dello spazio comune è lo stesso in entrambe le soluzioni. Questa soluzione (con la trasformazione ordinale) mostra probabilmente un grado di variazione minore sulla dimensione verticale rispetto a quella orizzontale, che è invece evidente nella soluzione con trasformazione lineare.

Trasformazioni delle distanze

Figura 15-34
Grafico di trasformazione per la soluzione con trasformazione ordinale



Fatta eccezione per i valori con le distanze maggiori, che tendono a curvarsi verso l'altro rispetto al resto dei valori, la trasformazione ordinale delle distanze è abbastanza lineare. Queste distanze spiegano probabilmente il motivo della maggior parte delle differenze tra le soluzioni con trasformazione ordinale e lineare. Tuttavia, in questo caso le informazioni non sono sufficienti per stabilire se il trend non lineare presente nei valori maggiori è un trend reale o un'anomalia.

Letture consigliate

Per ulteriori informazioni, consultare i seguenti testi:

Busing, F. M. T. A., P. J. F. Groenen, e W. J. Heiser. 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, 70, .

Green, P. E., e V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Prezzo, R. H., e D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Appendice

A

File di esempio

Il file di esempio installato con il prodotto si trova nella sottodirectory *Samples* della directory di installazione. La sottodirectory Samples contiene cartelle separate per ciascuna delle seguenti lingue: Inglese, Francese, Tedesco, Italiano, Giapponese, Coreano, Polacco, Russo, Cinese semplificato, Spagnolo e Cinese tradizionale.

Non tutti i file di esempio sono disponibili in tutte le lingue. Se un file di esempio non è disponibile in una lingua, la cartella di tale lingua contiene una versione inglese del file.

Descrizioni

Questa sezione contiene brevi descrizioni dei file di esempio utilizzati negli esempi riportati in tutta la documentazione.

- **accidents.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nello studio dei fattori di rischio correlati all'età e al sesso per gli incidenti automobilistici che si verificano in una determinata regione. Ciascun caso corrisponde a una classificazione incrociata della categoria relativa età e del sesso.
- **adl.sav.** File di dati ipotetici che prende in esame l'impegno richiesto per determinare i vantaggi di un tipo di terapia proposto per i pazienti con problemi di cuore. I medici hanno assegnato in modo casuale i pazienti con problemi di cuore di sesso femminile a uno di due gruppi. Al primo gruppo è stata assegnata la terapia fisica standard; al secondo gruppo, un'ulteriore terapia di supporto psicologico. Dopo tre mesi di trattamenti, a ciascuna capacità dei pazienti che consente di riprendere le normali attività giornaliere è stato assegnato un punteggio come variabile ordinale.
- **advert.sav.** File di dati ipotetici che prende in esame l'impegno di un rivenditore al dettaglio che desidera esaminare la relazione tra il denaro speso per la pubblicità e le vendite risultanti. Finora sono stati raccolti i dati delle vendite precedenti e i relativi costi pubblicitari.
- **aflatoxin.sav.** File di dati ipotetici che prende in esame il test di raccolti di mais con presenza di Aflatossina, un veleno la cui concentrazione varia notevolmente nei raccolti. Una macchina per la lavorazione dei cereali ha ricevuto 16 campioni da ciascuno degli otto raccolti di mais e ha misurato i livelli di Aflatossina in parti per miliardo (PPB).
- **aflatoxin20.sav.** Questo file di dati contiene le misurazioni di Aflatossina di ciascuno dei 16 campioni di quattro raccolti e otto campioni dal file di dati *aflatoxin.sav*.
- **anorectic.sav.** Per trovare una sintomatologia standardizzata del comportamento anoressico/bulimico, i ricercatori (Van der Ham, Meulman, Van Strien, e Van Engeland, 1997) hanno condotto uno studio basato su 55 adolescenti affetti da disordini alimentari conosciuti. Ogni paziente è stato visitato quattro volte in quattro anni, per un totale di 220 visite. Durante ogni visita, ai pazienti sono stati assegnati punteggi per ciascuno dei 16 sintomi. I punteggi

relativi ai sintomi sono assenti per il paziente 71 alla visita 2, il paziente 76 alla visita 2 e il paziente 47 alla visita 3, con 217 osservazioni valide.

- **autoaccidents.sav.** File di dati ipotetici che prende in esame l'impegno di un analista che opera nel campo delle assicurazioni per creare un modello del numero di incidenti automobilistici per conducente. Il modello prende in esame anche l'età e il sesso del conducente. Ciascun caso rappresenta un diverso conducente e riporta il sesso e l'età (in anni) del conducente e il numero di incidenti automobilistici negli ultimi cinque anni.
- **band.sav.** Questo file di dati ipotetici contiene le cifre sulle vendite settimanali di CD conseguite da un gruppo musicale. Il file include anche i dati di tre possibili variabili predittore.
- **bankloan.sav.** File di dati ipotetici che prende in esame l'impegno di una banca nel tentativo di ridurre il tasso di inadempienza nel rimborso di un prestito. Il file contiene informazioni finanziarie e demografiche su 850 vecchi e potenziali clienti. I primi 700 casi riguardano i clienti a cui sono stati concessi dei prestiti precedentemente. Gli ultimi 150 casi riguardano i potenziali clienti che la banca deve classificare come rischi di credito positivi o negativi.
- **bankloan_binning.sav.** File di dati ipotetici che contiene informazioni finanziarie e demografiche su 5000 vecchi clienti.
- **behavior.sav.** In un classico esempio (Prezzo e Bouffard, 1974), è stato chiesto a 52 studenti di classificare una combinazione di 15 situazioni e 15 comportamenti utilizzando una scala da 0="molto appropriato" a 9="molto inadeguato". I valori medi riferiti ai partecipanti sono stati considerati dissimilarità.
- **behavior_ini.sav.** Questo file di dati contiene la configurazione iniziale di una soluzione a due dimensioni per *behavior.sav*.
- **brakes.sav.** File di dati ipotetici che prende in esame il controllo di qualità di un'industria che produce freni a disco per automobili con elevate prestazioni. Il file di dati contiene le misurazioni del diametro di 16 dischi da ciascuna delle otto macchine di produzione. L'obiettivo finale è ottenere un diametro dei dischi pari a 322 millimetri.
- **breakfast.sav.** In uno studio classico (Green e Rao, 1972), è stato chiesto a 21 studenti MBA della Wharton School e ai loro consorti di classificare 15 cibi da colazione in ordine di preferenza, dove il valore 1 corrispondeva all'alimento preferito in assoluto e il valore 15 a quello meno preferito. Le loro preferenze sono state registrate per sei diversi scenari, che comprendevano tutti gli scenari compresi tra "Preferenza generale" e "Solo snack con bibita".
- **breakfast-overall.sav.** Questo file contiene le preferenze degli alimenti della colazione solo per il primo scenario, "Preferenza generale".
- **broadband_1.sav.** File di dati ipotetici che contiene il numero di sottoscrittori, per area, di un provider di servizi a banda larga nazionale. Il file di dati contiene il numero dei sottoscrittori mensili di 85 aree in un periodo di quattro anni.
- **broadband_2.sav.** Questo file è identico al file *broadband_1.sav*, ma contiene i dati per ulteriori tre mesi.
- **car_insurance_claims.sav.** Un insieme di dati presentato e analizzato altrove (McCullagh e Nelder, 1989) riguarda le richieste di risarcimento auto. La quantità media di richieste di risarcimento può essere adattata come avente una distribuzione gamma, utilizzando una funzione di collegamento inverso per correlare la media della variabile dipendente a una

combinazione lineare di età del contraente della polizza e tipo e anni del veicolo. Il numero delle richieste di risarcimento specificato può essere utilizzato come peso scalato.

- **car_sales.sav.** Questo file di dati ipotetici contiene le stime sulle vendite, i prezzi di listino e le specifiche fisiche di numerose marche e modelli di veicoli. I prezzi di listino e le specifiche fisiche sono state ottenute dal sito *edmunds.com* e dai siti dei produttori.
- **car_sales_uprepared.sav.** Questa è una versione modificata di *car_sales.sav* che non comprende versioni trasformate dei campi.
- **carpet.sav.** Come esempio tipico (Green e Wind, 1973), un'azienda interessata alla commercializzazione di un nuovo battitappeto desidera esaminare l'influenza di cinque fattori sulle preferenze del consumatore, ovvero design della confezione, marca, prezzo, la presenza di un *marchio di qualità* e una garanzia "Soddisfatti o rimborsati". Esistono tre livelli di fattore per il design della confezione, che differiscono per la posizione della spazzola dell'applicatore; tre marchi (*K2R*, *Glory* e *Bissell*); tre livelli di prezzo e due livelli (no o sì) per ciascuno degli ultimi due fattori. Dieci consumatori sono classificati in 22 profili definiti da questi fattori. La variabile *Preferenza* include il rango delle classificazioni medie per ogni profilo. Classificazioni basse corrispondono a una preferenza elevata. La variabile riflette una misura globale della preferenza per ogni profilo.
- **carpet_prefs.sav.** Questo file di dati si basa sullo stesso esempio del file *carpet.sav*, ma contiene le classificazioni effettive raccolte da ciascuno dei 10 clienti. Ai clienti è stato chiesto di classificare 22 profili di prodotti in ordine di preferenza. Le variabili da *PREF1* a *PREF22* contengono gli ID dei profili associati, come definito nel file *carpet_plan.sav*.
- **catalog.sav.** File di dati ipotetico che contiene le cifre sulle vendite mensili di tre prodotti venduti da una società di vendita per corrispondenza. Il file include anche i dati di cinque possibili variabili predittore.
- **catalog_seasfac.sav.** Questo file di dati è uguale al file *catalog.sav* con l'eccezione che contiene un insieme di fattori stagionali calcolati dalla procedura Decomposizionale stagionale insieme a variabili di dati.
- **cellular.sav.** File di dati ipotetici che prende in esame l'impegno di un'azienda di telefonia cellulare nel tentativo di ridurre il churn, ovvero l'abbandono dei clienti. Agli account vengono applicati i punteggi relativi alla propensione al churn, con valori compresi tra 0 e 100. Gli account con punteggio pari a 50 o superiore è probabile che stiano cercando nuovi provider.
- **ceramics.sav.** File di dati ipotetici che prende in esame l'impegno di un produttore che desidera stabilire se una nuova lega premium ha una maggiore resistenza al calore rispetto alla lega standard. Ciascun caso rappresenta il test separato di una delle leghe. È indicata la temperatura massima alla quale può essere sottoposto il cuscinetto.
- **cereal.sav.** File di dati ipotetici che prende in esame le preferenze relative agli alimenti della colazione di un campione di 880 persone. Il file riporta anche l'età, il sesso e lo stato civile del campione e se le persone conducono uno stile di vita attivo (in base a un'attività sportiva con frequenza di due volte alla settimana). Ogni caso rappresenta un rispondente separato.
- **clothing_defects.sav.** File di dati ipotetici che prende in esame il processo di controllo di qualità di un'industria di abbigliamento. Per ciascun lotto prodotto nella fabbrica, gli ispettori prelevano un campione di abiti per contare il numero dei capi che non sono accettabili per la vendita.

- **coffee.sav.** Questo file di dati contiene informazioni sulle immagini percepite di sei marche di caffè freddo (Kennedy, Riquier, e Sharp, 1996). Per ciascuno dei 23 attributi dell'immagine del caffè freddo, sono state selezionate tutte le marche descritte da tale attributo. Le sei marche sono indicate dalle sigle AA, BB, CC, DD, EE e FF per tutelare la confidenzialità dei dati.
- **contacts.sav.** File di dati ipotetici che prende in esame l'elenco dei contatti di un gruppo di rappresentanti di vendita di computer aziendali. Ciascun contatto è classificato in base al reparto della società in cui lavora e dalle relative categorie aziendali. Il file riporta anche l'importo dell'ultima vendita effettuata, il tempo trascorso dall'ultima vendita e le dimensioni della società del contatto.
- **creditpromo.sav.** File di dati ipotetici che prende in esame l'impegno di un grande magazzino nel tentativo di valutare l'efficacia di una recente promozione con carta di credito. A tale scopo, sono stati selezionati 500 titolari di carta in modo casuale. Alla metà di questi è stato inviato un annuncio promozionale che comunica la riduzione del tasso d'interesse nel caso di acquisti effettuati entro i tre mesi successivi. All'altra metà è stato inviato un annuncio stagionale standard.
- **customer_dbase.sav.** File di dati ipotetico che prende in esame l'impegno di una società nel tentativo di utilizzare le informazioni contenute nel proprio database dei dati per creare offerte speciali per i clienti che più probabilmente risponderanno all'offerta. È stato selezionato in modo casuale un sottoinsieme della base dei clienti a cui è stata inviata l'offerta speciale e sono state registrate le risposte ricevute.
- **customer_information.sav.** File di dati ipotetici contenente le informazioni postali del cliente, ad esempio il nome e l'indirizzo.
- **customer_subset.sav.** Un sottoinsieme di 80 casi da *customer_dbase.sav*.
- **customers_model.sav.** File di dati ipotetici che contiene il nominativo delle persone a cui è stata inviata una campagna di marketing. I dati includono informazioni demografiche, un riepilogo della cronologia degli acquisti e se ciascuna persona ha risposto alla campagna. Ogni caso rappresenta una persona separata.
- **customers_new.sav.** File di dati ipotetici che contiene i nominativi delle persone che sono state evidenziate come potenziali candidati per una campagna di marketing. I dati includono informazioni demografiche e un riepilogo sulla cronologia degli acquisti di ciascuna persona. Ogni caso rappresenta una persona separata.
- **debate.sav.** File di dati ipotetici che prende in esame le risposte appaiate a un'indagine da parte dei partecipanti a un dibattito politico prima e dopo il dibattito. Ogni caso rappresenta un rispondente separato.
- **debate_aggregate.sav.** File di dati ipotetici che aggrega le risposte contenute nel file *debate.sav*. Ciascun caso corrisponde a una classificazione incrociata della preferenza prima e dopo il dibattito.
- **demo.sav.** File di dati ipotetici che prende in esame un database di clienti che hanno fatto acquisti al fine di inviare offerte mensili tramite il metodo del direct mailing. Viene registrata la risposta dei clienti, sia che abbiano aderito all'offerta o meno, insieme a diverse informazioni demografiche.
- **demo_cs_1.sav.** File di dati ipotetici che prende in esame il primo passo che una società intraprende per compilare un database con informazioni ricavate dai sondaggi. Ogni caso rappresenta una diversa città. Sono registrate anche le informazioni sulla regione, provincia, distretto e città.

- **demo_cs_2.sav.** File di dati ipotetici che prende in esame il secondo passo che una società intraprende per compilare un database con informazioni ricavate dai sondaggi. Ogni caso rappresenta una diversa unità di abitazione, ricavata dalle città selezionate nel primo passo. Sono registrate anche le informazioni sulla regione, provincia, distretto, città, suddivisione e unità. Il file include inoltre informazioni sul campionamento ottenute dai primi due stadi del disegno.
- **demo_cs.sav.** File di dati ipotetici che contiene informazioni sulle indagini raccolte utilizzando un disegno di campionamento complesso. Ogni caso rappresenta una diversa unità di abitazione. Sono registrate diverse informazioni demografiche e sul campionamento.
- **dmdata.sav.** File di dati ipotetici che contiene informazioni demografiche e di acquisto di una società di direct marketing. *dmdata2.sav* contiene informazioni su un sottoinsieme di contatti che hanno ricevuto un mailing di prova e *dmdata3.sav* contiene informazioni sui contatti rimanenti che non hanno ricevuto il mailing di prova.
- **dietstudy.sav.** File di dati ipotetici che contiene il risultato di uno studio ipotetico sulla dieta chiamato "Stillman diet" (Rickman, Mitchell, Dingman, e Dalen, 1974). Ogni caso rappresenta un diverso soggetto e ne riporta il peso prima e dopo la dieta in libbre e i livelli dei trigliceridi in mg/100 ml.
- **dvdplayer.sav.** File di dati ipotetici che prende in esame lo sviluppo di un nuovo lettore DVD. Utilizzando un prototipo, il personale addetto al marketing ha raccolto dati sui gruppi di interesse. Ogni caso rappresenta un diverso utente che è stato sottoposto all'indagine e include informazioni demografiche personali dell'utente e sulle risposte che ha fornito riguardo al prototipo.
- **german_credit.sav.** Questo file di dati contiene informazioni ricavate dall'insieme di dati "German Credit" del Repository of Machine Learning Databases (Blake e Merz, 1998) presso la University of California, Irvine.
- **grocery_1month.sav.** Questo file di dati ipotetici corrisponde al file di dati *grocery_coupons.sav* con gli acquisti settimanali organizzati in modo che ogni caso corrisponda a un cliente separato. Alcune delle variabili che cambiano settimanalmente non vengono riportate nei risultati; l'importo speso registrato corrisponde ora alla somma degli importi spesi durante le quattro settimane dello studio.
- **grocery_coupons.sav.** File di dati ipotetici che contiene i dati sui sondaggi raccolti da una catena di drogherie interessata alle abitudini di acquisto dei suoi clienti. Ciascun cliente viene seguito per quattro settimane e ciascun caso corrisponde a una settimana per cliente con informazioni sul luogo degli acquisti e i tipi di acquisti, incluso l'importo speso nelle drogherie durante la settimana.
- **guttman.sav.** Bell (Bell, 1961) ha presentato una tabella per illustrare i possibili gruppi sociali. Guttman (Guttman, 1968) ha utilizzato una parte di tale tabella, in cui cinque variabili che descrivono elementi come l'interazione sociale, i sentimenti di appartenenza a un gruppo, la vicinanza fisica dei membri e il grado di formalità della relazione, sono state incrociate con cinque gruppi sociali teorici, compresi folla (ad esempio, le persone presenti a una partita di calcio), uditorio (ad esempio, di uno spettacolo teatrale o di una lezione universitaria), pubblico (ad esempio televisivo), calca (come una folla, ma con un'interazione molto maggiore), gruppi primari (intimi), gruppi secondari (volontari) e la comunità moderna (unione non stretta derivante da una vicinanza fisica elevata e dall'esigenza di servizi specializzati).

- **health_funding.sav.** File di dati ipotetici che contiene i dati sui fondi di assistenza sanitaria (importo per 100 persone), sui tassi di malattie (tasso per 10.000 persone) e sulle visite ai fornitori di assistenza sanitaria (tasso per 10.000 persone). Ogni caso rappresenta una diversa città.
- **hivassay.sav.** File di dati ipotetici che prende in esame l'impegno di un'industria farmaceutica nel tentativo di sviluppare un'analisi che riesca a rilevare in tempi brevi l'infezione da virus HIV. I risultati dell'analisi sono otto sfumature di colore rosso sempre più intenso; le sfumature più intense indicano la maggiore probabilità di infezione. Un esperimento di laboratorio è stato condotto su 2000 campioni di sangue. La metà di questi è risultata infetta al virus HIV, l'altra metà non è risultata infetta.
- **hourlywagedata.sav.** File di dati ipotetici che prende in esame la paga oraria degli infermieri occupati presso uffici e ospedali e in base ai diversi livelli di esperienza.
- **insurance_claims.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nella creazione di un modello per contrassegnare le richieste di risarcimento sospette e potenzialmente fraudolente. Ogni caso rappresenta una richiesta di risarcimento separata.
- **insure.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nello studio dei fattori di rischio, che indicano l'eventualità che un cliente presenti una domanda di indennizzo in un contratto assicurativo sulla vita della durata di dieci anni. Ogni caso nel file di dati rappresenta una coppia di contratti. In un contratto sono contenute informazioni su una richiesta di risarcimento, l'altro sull'età e sul sesso.
- **judges.sav.** File di dati ipotetici che prende in esame il punteggio assegnato, da giurie qualificate (più un appassionato) a 300 prestazioni sportive. Ciascuna riga rappresenta una diversa prestazione; i giudici hanno esaminato le stesse prestazioni.
- **kinship_dat.sav.** Rosenberg e Kim (Rosenberg e Kim, 1975) si prefiggono di analizzare 15 termini indicanti parentela (zia, fratello, cugino, padre, nipote femmina, di nonni, nonno, nonna, nipote maschio di nonni, madre, nipote maschio di zii), nipote femmina di zii, sorella, figlio, zio). Hanno richiesto a quattro gruppi di studenti universitari (due composti da femmine e due da maschi) di ordinare questi termini in base alla similiarità. A due gruppi (uno femminile e uno maschile) è stato richiesto di effettuare l'ordinamento due volte, con il secondo ordinamento basato su un criterio diverso rispetto al primo. Di conseguenza, sono state ottenute sei "sorgenti" in totale. Ogni sorgente corrisponde a una matrice di prossimità 15×15, le cui celle sono uguali al numero delle persone in una sorgente meno il numero di volte in cui gli oggetti sono stati ripartiti insieme nella sorgente.
- **kinship_ini.sav.** Questo file di dati contiene la configurazione iniziale di una soluzione a tre dimensioni per *kinship_dat.sav*.
- **kinship_var.sav.** Questo file di dati contiene variabili indipendenti relative a *sesso*, *generazione* e *grado* di separazione che possono essere utilizzate per interpretare le dimensioni di una soluzione per *kinship_dat.sav*. In modo specifico, tali variabili possono essere utilizzate per limitare lo spazio della soluzione a una combinazione lineare di tali variabili.
- **marketvalues.sav.** File di dati che prende in esame le vendite di abitazioni in un nuovo centro abitato in Algonquin, Ill., durate gli anni 1999–2000. Tali vendite sono una questione di dominio pubblico.

- **nhis2000_subset.sav.** Il National Health Interview Survey (NHIS) è un sondaggio di grandi dimensioni condotto sulla popolazione civile americana. Le interviste vengono realizzate di persona e si basano su un campione rappresentativo di famiglie a livello nazionale. Per ogni membro di una famiglia vengono raccolte osservazioni e informazioni di carattere demografico relative allo stato di salute. Questo file di dati contiene un sottoinsieme delle informazioni ottenute dall'indagine del 2000. National Center for Health Statistics. National Health Interview Survey, 2000. File di dati e documentazione di dominio pubblico. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accesso 2003.
- **ozone.sav** I dati includono 330 osservazioni basate su sei variabili meteorologiche per quantificare la concentrazione dell'ozono dalle variabili rimanenti. I precedenti ricercatori, (Breiman e Friedman, 1985) e (Hastie e Tibshirani, 1990), hanno rilevato non linearità tra queste variabili, che impediscono un approccio di regressione standard.
- **pain_medication.sav.** File di dati ipotetici che contiene i risultati di un test clinico per stabilire la cura antinfiammatoria per il trattamento del dolore generato dall'artrite cronica. Di particolare interesse, il test ha evidenziato il tempo che impiega il farmaco ad avere effetto e il confronto con altri farmaci esistenti.
- **patient_los.sav.** File di dati ipotetici che contiene informazioni sul trattamento dei pazienti ricoverati per sospetto di infarto del miocardio. Ogni caso corrisponde a un diverso paziente e contiene diverse variabili correlate alla degenza nell'ospedale.
- **patlos_sample.sav.** File di dati ipotetici che contiene informazioni sul trattamento di un campione di pazienti curato con trombolitici durante la degenza per infarto del miocardio. Ogni caso corrisponde a un diverso paziente e contiene diverse variabili correlate alla degenza nell'ospedale.
- **polishing.sav.** File di dati "Nambeware Polishing Times" di Data and Story Library. Prende in esame l'impegno di un'industria di stoviglie in metallo (Nambe Mills, Santa Fe, N. M.) nel tentativo di pianificare il proprio piano di produzione. Ogni caso rappresenta un diverso articolo nella linea dei prodotti. Per ciascun articolo sono indicati il diametro, il tempo di lucidatura, il prezzo e il tipo di prodotto.
- **poll_cs.sav.** File di dati ipotetici che prende in esame i sondaggi per stabilire il livello di sostegno pubblico nei confronti di un disegno di legge prima che diventi una legge vera e propria. I casi corrispondono ai votanti registrati. Ciascun caso riporta informazioni sulla contea, sul comune e sul quartiere in cui vive il votante.
- **poll_cs_sample.sav.** File di dati ipotetici che contiene un campione dei votanti elencati nel file *poll_cs.sav*. Il campione è stato selezionato in base al disegno specificato nel file di piano *poll.csplan* e questo file di dati contiene le probabilità di inclusione e i pesi del campione. Tuttavia, notare che poiché fa uso del metodo PPS (probability-proportional-to-size, probabilità proporzionale alla dimensione), esiste anche un file contenente le probabilità di selezione congiunte (*poll_jointprob.sav*). Le ulteriori variabili corrispondenti ai dati demografici dei votanti e alla loro opinione sul disegno di legge, sono state raccolte e aggiunte al file di dati dopo aver acquisito il campione.
- **property_assess.sav.** File di dati ipotetici che prende in esame l'impegno di un perito di una contea nel tentativo di mantenere gli accertamenti sui valori delle proprietà aggiornati in base alle risorse limitate. I casi rappresentano le proprietà vendute nella contea nello scorso anno. Ogni caso nel file di dati contiene informazioni sul comune in cui si trova la proprietà, il perito che per ultimo ha visitato la proprietà, il tempo trascorso dall'accertamento, la valutazione fatta in tale momento e il valore di vendita della proprietà.

- **property_assess_cs.sav.** File di dati ipotetici che prende in esame l'impegno di un perito di uno stato nel tentativo di mantenere aggiornati gli accertamenti sui valori delle proprietà in base alle risorse limitate. I casi corrispondono alle proprietà nello stato. Ogni caso nel file di dati include informazioni sulla contea, il comune e il quartiere in cui risiede la proprietà, la data dell'ultimo accertamento e la valutazione fatta in tale data.
- **property_assess_cs_sample.sav.** File di dati ipotetici che contiene un campione delle proprietà elencate nel file *property_assess_cs.sav*. Il campione è stato selezionato in base al disegno specificato nel file di piano *property_assess.csplan* e questo file di dati contiene le probabilità di inclusione e i pesi del campione. L'ulteriore variabile *Valore corrente* è stata raccolta e aggiunta al file di dati dopo aver acquisito il campione.
- **recidivism.sav.** File di dati ipotetici che prende in esame l'impegno delle Forze dell'Ordine nel tentativo di valutare il tasso di recidività nella propria area di giurisdizione. Ogni caso corrisponde a un precedente trasgressore e include le informazioni demografiche, alcuni dettagli sul primo crimine, il tempo trascorso fino al secondo arresto e se tale arresto è avvenuto entro due anni dal primo.
- **recidivism_cs_sample.sav.** File di dati ipotetici che prende in esame l'impegno delle Forze dell'Ordine nel tentativo di valutare il tasso di recidività nella propria area di giurisdizione. Ogni caso corrisponde a un trasgressore precedente, rilasciato dopo il primo arresto durante il mese di giugno del 2003 e registra le relative informazioni demografiche, alcuni dettagli sul primo crimine commesso e i dati del secondo arresto, se si è verificato prima della fine di giugno del 2006. I trasgressori sono stati selezionati dai dipartimenti sottoposti a campione in base al piano di campionamento specificato nel file *recidivism_cs.csplan*. Poiché viene utilizzato un metodo PPS (Probability-Proportional-to-Size, probabilità proporzionale alla dimensione), esiste anche un file contenente le probabilità di selezione congiunte (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** File di dati ipotetici contenente i dati delle transazioni di acquisto, inclusa la data di acquisto, gli articoli acquistati e il valore monetario di ciascuna transazione.
- **salesperformance.sav.** File di dati ipotetici che prende in esame la valutazione di due nuovi corsi di formazione alle vendite. Sessanta dipendenti, divisi in tre gruppi, ricevono tutti la formazione standard. In più, al gruppo 2 viene assegnato un corso di formazione tecnica e al gruppo 3 un'esercitazione pratica. Alla fine del corso di formazione, ciascun dipendente viene sottoposto a un esame e il punteggio conseguito viene registrato. Ciascun caso nel file di dati rappresenta un diverso partecipante. Il file di dati include il gruppo a cui è assegnato il partecipante e il punteggio conseguito all'esame finale.
- **satisf.sav.** File di dati ipotetico che prende in esame un'indagine sulla soddisfazione dei clienti condotta da una società di vendita al dettaglio presso 4 negozi. Sono stati intervistati 582 clienti e ciascun caso rappresenta le risposte ottenute da un singolo cliente.
- **screws.sav.** Questo file di dati contiene informazioni sulle caratteristiche di viti, bulloni, dadi e puntine (Hartigan, 1975).
- **shampoo_ph.sav.** File di dati ipotetici che prende in esame il processo di controllo di qualità di un'industria di prodotti per capelli. A intervalli di tempo regolari, vengono misurati sei diversi lotti prodotti e ne viene registrato il relativo pH. I valori accettati sono compresi tra 4,5 e 5,5.
- **ships.sav.** Ad esempio, un insieme di dati presentato e analizzato altrove (McCullagh et al., 1989) riguarda i danni subiti dalle navi da carico a causa delle onde. I conteggi degli incidenti possono essere presentati con un tasso di Poisson in base al tipo di nave, al periodo di

costruzione e al periodo di servizio. I mesi di servizio aggregati di ciascuna cella della tabella generata dalla classificazione incrociata dei fattori fornisce i valori di esposizione al rischio.

- **site.sav.** File di dati ipotetici che prende in esame l'impegno di una società nella scelta di nuovi siti in cui espandere la propria presenza. La società ha incaricato due consulenti separati che, oltre a valutare i siti e presentare un report completo, devono classificarli come potenzialmente "molto adatti", "adatti" o "poco adatti".
- **smokers.sav.** Questo file di dati è un estratto del 1998 National Household Survey of Drug Abuse e rappresenta un campione probabile di famiglie americane. (*http://dx.doi.org/10.3886/ICPSR02934*) Il primo passo nell'analisi di questo file di dati consiste quindi nel pesare i dati per rispecchiare le tendenze della popolazione.
- **stroke_clean.sav.** File di dati ipotetici che riporta lo stato di un database medico dopo averne eseguito la pulizia utilizzando le procedure del modulo Data Preparation.
- **stroke_invalid.sav.** File di dati ipotetici che riporta lo stato iniziale di un database medico e contiene numerosi errori di immissione dati.
- **stroke_survival.** Questo file di dati ipotetici riguarda i tempi di sopravvivenza per i pazienti che, dopo avere completato un programma riabilitativo in seguito a un ictus postischemico, affrontano alcune sfide. Dopo l'attacco, viene annotata l'occorrenza dell'infarto miocardiaco, dell'ictus ischemico o emorragico e viene registrata l'ora dell'evento. Questo campione viene troncato a sinistra perché include solo i pazienti che sono sopravvissuti fino alla fine del programma riabilitativo post-ictus.
- **stroke_valid.sav.** File di dati ipotetici che riporta lo stato di un database medico dopo il controllo dei valori eseguito con la procedura Convalida i dati. Il database contiene comunque casi potenzialmente anomali.
- **survey_sample.sav.** File di dati che contiene i dati dell'indagine, compresi i dati demografici e varie misure dell'atteggiamento. Si basa su un sottoinsieme di variabili tratte dal 1998 NORC General Social Survey, benché i valori di alcuni dati siano stati modificati e siano state aggiunte variabili fittizie a scopo dimostrativo.
- **telco.sav.** File di dati ipotetici che prende in esame l'impegno di un'azienda di telecomunicazioni nel tentativo di ridurre il churn, ovvero l'abbandono dei propri clienti. Ciascun caso rappresenta un cliente separato e riporta diverse informazioni demografiche e sull'uso del servizio.
- **telco_extra.sav.** Questo file di dati è simile al file *telco.sav*, ma le variabili "tenure" e spesa del cliente trasformata tramite logaritmo sono state sostituite dalle variabili di spesa del cliente trasformata tramite logaritmo standardizzate.
- **telco_missing.sav.** Questo file di dati è un sottoinsieme del file di dati *telco.sav*, ma alcuni dei valori di dati demografici sono stati sostituiti con valori mancanti.
- **testmarket.sav.** File di dati ipotetici che prende in esame i piani di una catena di fast food per aggiungere un nuovo prodotto al proprio menu. Sono previste tre campagne promozionali del nuovo prodotto. Il prodotto viene introdotto in diversi mercati selezionati in modo casuale. Per ogni sede viene utilizzata una promozione differente registrando le vendite settimanali della nuova voce per le prime quattro settimane. Ogni caso rappresenta un luogo e una settimana diversi.

- **testmarket_1month.sav.** Questo file di dati ipotetici corrisponde al file *testmarket.sav* con le vendite settimanali organizzate in modo che ogni caso corrisponda a un luogo separato. Alcune delle variabili che cambiano settimanalmente non vengono riportate nei risultati; le vendite registrate corrispondono ora alla somma delle vendite conseguite durante le quattro settimane dello studio.
- **tree_car.sav.** File di dati ipotetici che contiene dati demografici e sul prezzo di acquisto dei veicoli.
- **tree_credit.sav.** File di dati ipotetici che contiene dati demografici e sulla cronologia dei mutui di una banca.
- **tree_missing_data.sav.** File di dati ipotetici che contiene dati demografici e sulla cronologia dei mutui di una banca con un numero elevato di valori mancanti.
- **tree_score_car.sav.** File di dati ipotetici che contiene dati demografici e sul prezzo di acquisto dei veicoli.
- **tree_textdata.sav.** File di dati semplice con due variabili destinato principalmente per mostrare lo stato predefinito delle variabili prima dell'assegnazione dei livelli di misurazione e delle etichette dei valori.
- **tv-survey.sav.** File di dati ipotetici che prende in esame un sondaggio condotto da una emittente televisiva che deve stabilire se estendere la durata di un programma di successo. A un campione di 906 intervistati è stato chiesto se preferisce guardare il programma con diverse condizioni. Ciascuna riga rappresenta un diverso intervistato e ciascuna colonna una diversa condizione.
- **ulcer_recurrence.sav.** Questo file contiene informazioni parziali su uno studio svolto per mettere a confronto l'efficacia di due terapie preventive per la recidiva delle ulcere. Fornisce un ottimo esempio di dati acquisiti a intervalli ed è stato presentato e analizzato in altri luoghi (Collett, 2003).
- **ulcer_recurrence_recoded.sav.** In questo file sono contenute le informazioni del file *ulcer_recurrence.sav* riorganizzate per consentire di presentare la probabilità degli eventi per ciascun intervallo dello studio, anziché solo alla fine. È stato presentato e analizzato in altri luoghi (Collett et al., 2003).
- **verd1985.sav.** Questo file di dati prende in esame un'indagine (Verdegaal, 1985). Sono state registrate le risposte di quindici soggetti a otto variabili. Le variabili di interesse sono suddivise in tre insiemi. L'insieme 1 include *età* e *statociv*, l'insieme 2 include *andom* e *giornale* e l'insieme 3 include *musica* e *vicinato*. *Andom* viene scalata come nominale multipla ed *età* come ordinale; tutte le altre variabili vengono scalate come nominali singole.
- **virus.sav.** File di dati ipotetici che prende in esame l'impegno di un ISP (Internet Service Provider) nel tentativo di determinare gli effetti che un virus può generare nelle sue reti. Si è tenuta traccia della percentuale (approssimativa) di traffico e-mail infettato da virus sulla rete in un lasso di tempo, dal momento dell'individuazione fino alla soppressione della minaccia.
- **wheeze_steubenville.sav.** Questo file è un sottoinsieme di uno studio longitudinale degli effetti che l'inquinamento provoca sulla salute dei bambini (Ware, Dockery, Spiro III, Speizer, e Ferris Jr., 1984). I dati contengono misure binarie ripetute del livello di asma dei bambini

della città di Steubenville, Ohio, di 7, 8, 9 e 10 anni. I dati indicano anche se la mamma dei bambini era fumatrice durante il primo anno dello studio.

- **workprog.sav.** File di dati ipotetici che prende in esame un programma di lavoro governativo il cui obiettivo è fornire attività più adatte alle persone diversamente abili. È stato seguito un campione di potenziali partecipanti al programma, alcuni dei quali sono stati selezionai in modo casuale e altri no. Ogni caso rappresenta un diverso partecipante al programma.

Appendice

B

Notices

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing,

using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Bibliografia

Barlow, R. E., D. J. Bartholomew, D. J. Bremner, e H. D. Brunk. 1972. *Statistical inference under order restrictions*. New York: John Wiley and Sons.

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. In: *Methodologies of Pattern Recognition,* S. Watanabe, ed. New York: Academic Press.

Benzécri, J. P. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.

Bishop, Y. M., S. E. Feinberg, e P. W. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass.: MIT Press.

Blake, C. L., e C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., e J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, .

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, 18, .

Busing, F. M. T. A., P. J. F. Groenen, e W. J. Heiser. 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, 70, .

Carroll, J. D. 1968. Generalization of canonical correlation analysis to three or more sets of variables. In: *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* Washington, D.C.: American Psychological Association.

Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

Commandeur, J. J. F., e W. J. Heiser. 1993. *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices*. Leiden: Department of Data Theory, University of Leiden.

De Haas, M., J. A. Algera, H. F. J. M. Van Tuijl, e J. J. Meulman. 2000. Macro and micro goal setting: In search of coherence. *Applied Psychology*, 49, .

De Leeuw, J. 1982. Nonlinear principal components analysis. In: *COMPSTAT Proceedings in Computational Statistics,* Vienna: Physica Verlag.

De Leeuw, J. 1984. *Canonical analysis of categorical data*, 2nd ed. Leiden: DSWO Press.

De Leeuw, J. 1984. The Gifi system of nonlinear multivariate analysis. In: *Data Analysis and Informatics III,* E. Diday, et al., ed..

De Leeuw, J., e W. J. Heiser. 1980. Multidimensional scaling with restrictions on the configuration. In: *Multivariate Analysis, Vol. V,* P. R. Krishnaiah, ed. Amsterdam: North-Holland.

De Leeuw, J., e J. Van Rijckevorsel. 1980. HOMALS and PRINCALS—Some generalizations of principal components analysis. In: *Data Analysis and Informatics,* E. Diday, et al., ed. Amsterdam: North-Holland.

De Leeuw, J., F. W. Young, e Y. Takane. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .

De Leeuw, J. 1990. Multivariate analysis with optimal scaling. In: *Progress in Multivariate Analysis,* S. Das Gupta, e J. Sethuraman, ed. Calcutta: Indian Statistical Institute.

Eckart, C., e G. Young. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, .

Fisher, R. A. 1938. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics*, 10, .

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58, .

Gifi, A. 1985. *PRINCALS. Research Report UG-85-02*. Leiden: Department of Data Theory, University of Leiden.

Gifi, A. 1990. *Nonlinear multivariate analysis*. Chichester: John Wiley and Sons.

Gilula, Z., e S. J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association*, 83, .

Gower, J. C., e J. J. Meulman. 1993. The treatment of categorical information in physical anthropology. *International Journal of Anthropology*, 8, .

Green, P. E., e V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., e Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. In: *The Prediction of Personal Adjustment,* P. Horst, ed. New York: Social Science Research Council.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, .

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, T., e R. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

Hastie, T., R. Tibshirani, e A. Buja. 1994. Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, .

Hayashi, C. 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statitical Mathematics*, 2, .

Heiser, W. J. 1981. *Unfolding analysis of proximity data*. Leiden: Department of Data Theory, University of Leiden.

Heiser, W. J., e F. M. T. A. Busing. 2004. Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, ed. Thousand Oaks, Calif.: Sage Publications, Inc..

Heiser, W. J., e J. J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. In: *Correspondence Analysis in the Social Sciences: Recent Developments and Applications,* M. Greenacre, e J. Blasius, ed. New York: Academic Press.

Heiser, W. J., e J. J. Meulman. 1995. Nonlinear methods for the analysis of homogeneity and heterogeneity. In: *Recent Advances in Descriptive Multivariate Analysis,* W. J. Krzanowski, ed. Oxford: Oxford University Press.

Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, .

Horst, P. 1961. Relations among m sets of measures. *Psychometrika*, 26, .

Israëls, A. 1987. *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Kennedy, R., C. Riquier, e B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement and Analysis for Marketing*, 5, .

Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58, .

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, .

Kruskal, J. B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, .

Kruskal, J. B. 1965. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B*, 27, .

Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. In: *International Encyclopedia of Statistics,* W. H. Kruskal, e J. M. Tanur, ed. New York: The Free Press.

Kruskal, J. B., e R. N. Shepard. 1974. A nonmetric variety of linear factor analysis. *Psychometrika*, 39, .

Krzanowski, W. J., e F. H. C. Marriott. 1994. *Multivariate analysis: Part I, distributions, ordination and inference*. London: Edward Arnold.

Lebart, L., A. Morineau, e K. M. Warwick. 1984. *Multivariate descriptive statistical analysis*. New York: John Wiley and Sons.

Lingoes, J. C. 1968. The multivariate analysis of qualitative data. *Multivariate Behavioral Research*, 3, .

Max, J. 1960. Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, 6, .

McCullagh, P., e J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Meulman, J. J. 1982. *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.

Meulman, J. J. 1986. *A distance approach to nonlinear multivariate analysis*. Leiden: DSWO Press.

Meulman, J. J. 1992. The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57, .

Meulman, J. J. 1993. Principal coordinates analysis with optimal transformations of the variables: Minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46, .

Meulman, J. J. 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification*, 13, .

Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, 4, .

Meulman, J. J., e W. J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. In: *Visual Display of Categorical Data,* M. Greenacre, e J. Blasius, ed. New York: Academic Press.

Meulman, J. J., e P. Verboon. 1993. Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika*, 58, .

Meulman, J. J., A. J. Van der Kooij, e A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. In: *Classification, Automation and New Media,* W. Gaul, e G. Ritter, ed. Berlin: Springer-Verlag.

Meulman, J. J., A. J. Van der Kooij, e W. J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, ed. Thousand Oaks, Calif.: Sage Publications, Inc..

Nishisato, S. 1980. *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S. 1984. Forced classification: A simple application of a quantification method. *Psychometrika*, 49, .

Nishisato, S. 1994. *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Pratt, J. W. 1987. Dividing the indivisible: Using simple symmetry to partition variance explained. In: *Proceedings of the Second International Conference in Statistics,* T. Pukkila, e S. Puntanen, ed. Tampere, Finland: University of Tampere.

Prezzo, R. H., e D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Ramsay, J. O. 1989. Monotone regression splines in action. *Statistical Science*, 4, .

Rao, C. R. 1973. *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley and Sons.

Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In: *Multivariate Analysis, Vol. 5,* P. R. Krishnaiah, ed. Amsterdam: North-Holland.

Rickman, R., N. Mitchell, J. Dingman, e J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

Rosenberg, S., e M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, .

Roskam, E. E. 1968. *Metric analysis of ordinal data in psychology*. Voorschoten: VAM.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika*, 27, .

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function II. *Psychometrika*, 27, .

Shepard, R. N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, .

Tenenhaus, M., e F. W. Young. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, .

Theunissen, N. C. M., J. J. Meulman, A. L. Den Ouden, H. M. Koopman, G. H. Verrips, S. P. Verloove-Vanhorick, e J. M. Wit. 2003. Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology*, 4, .

Tucker, L. R. 1960. Intra-individual and inter-individual multidimensionality. In: *Psychological Scaling: Theory & Applications*, H. Gulliksen, e S. Messick, ed. New York: John Wiley and Sons.

Van der Burg, E. 1988. *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.

Van der Burg, E., e J. De Leeuw. 1983. Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, .

Van der Burg, E., J. De Leeuw, e R. Verdegaal. 1988. Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 53, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, e H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, .

Van der Kooij, A. J., e J. J. Meulman. 1997. MURALS: Multiple regression and optimal scaling using alternating least squares. In: *Softstat '97,* F. Faulbaum, e W. Bandilla, ed. Stuttgart: Gustav Fisher.

Van Rijckevorsel, J. 1987. *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. Leiden: DSWO Press.

Verboon, P., e I. A. Van der Lans. 1994. Robust canonical discriminant analysis. *Psychometrika*, 59, .

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (in Dutch)*. Leiden: Department of Data Theory, University of Leiden.

Vlek, C., e P. J. Stallen. 1981. Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance*, 28, .

Wagenaar, W. A. 1988. *Paradoxes of gambling behaviour*. London: Lawrence Erlbaum Associates, Inc.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, e B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, .

Winsberg, S., e J. O. Ramsay. 1980. Monotonic transformations to additivity using splines. *Biometrika*, 67, .

Winsberg, S., e J. O. Ramsay. 1983. Monotone spline transformations for dimension reduction. *Psychometrika*, 48, .

Wolter, K. M. 1985. *Introduction to variance estimation*. Berlin: Springer-Verlag.

Young, F. W. 1981. Quantitative analysis of qualitative data. *Psychometrika*, 46, .

Young, F. W., J. De Leeuw, e Y. Takane. 1976. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .

Young, F. W., Y. Takane, e J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, .

Zeijl, E., Y. te Poel, M. du Bois-Reymond, J. Ravesloot, e J. J. Meulman. 2000. The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research*, 32, .

Indice