

# IBM SPSS Missing Values 20



*Nota:* Prima di utilizzare queste informazioni e il relativo prodotto, leggere le informazioni generali disponibili in Note legali a pag. 92.

Questa versione si applica a IBM® SPSS® Statistics 20 e a tutte le successive versioni e modifiche fino a eventuali disposizioni contrarie indicate in nuove versioni.

Le schermate dei prodotti Adobe sono state ristampate su autorizzazione di Adobe Systems Incorporated.

Le schermate dei prodotti Microsoft sono state ristampate su autorizzazione di Microsoft Corporation.

Materiali concessi in licenza - Proprietà di IBM

© **Copyright IBM Corporation 1989, 2011.**

Tutti i diritti riservati.

---

# Prefazione

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

## **Informazioni su Business Analytics di IBM**

Il software IBM Business Analytics fornisce informazioni complete, coerenti e accurate a cui i responsabili delle decisioni possono affidarsi per ottimizzare le prestazioni dell'azienda. Un ampio portafoglio di applicazioni di [business intelligence](#), [analisi predittiva](#), [gestione delle prestazioni e delle strategie finanziarie](#) e [analisi](#) offre una panoramica chiara, istantanea e interattiva delle prestazioni attuali e la possibilità di prevedere i risultati futuri. Utilizzato in combinazione con potenti soluzioni di settore, prassi consolidate e servizi professionali, questo software consente alle aziende di tutte le dimensioni di ottimizzare la produttività, automatizzare le decisioni senza problemi e fornire risultati migliori.

Come parte di questo portafoglio, il software IBM SPSS Predictive Analytics consente alle aziende di prevedere gli eventi futuri e di agire tempestivamente in modo da migliorare i risultati delle attività aziendali. Le aziende, gli enti governativi e le università di tutto il mondo si affidano alla tecnologia IBM SPSS perché rappresenta un vantaggio concorrenziale in termini di attrazione, retention e aumento dei clienti, riducendo al tempo stesso le frodi e limitando i rischi. Incorporando il software IBM SPSS nelle attività quotidiane, le aziende diventano imprese in grado di effettuare previsioni e di gestire e automatizzare le decisioni, per raggiungere gli obiettivi aziendali e vantaggi tangibili sulla concorrenza. Per ulteriori informazioni o per contattare un rappresentante, visitare il sito <http://www.ibm.com/spss>.

## **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 IBM Corp. o per l'installazione di uno degli ambienti hardware supportati. Per contattare il supporto tecnico, visitare il sito Web di IBM Corp. all'indirizzo <http://www.ibm.com/support>. 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.

## **Supporto tecnico per studenti**

Gli studenti che utilizzano una versione accademica o grad pack di qualsiasi prodotto software IBM SPSS sono pregati di utilizzare le apposite pagine online per studenti [Solutions for Education](#) (<http://www.ibm.com/spss/rd/students/>). Gli studenti che utilizzano una copia del software IBM SPSS fornita dall'università, sono pregati di contattare il coordinatore del prodotto IBM SPSS presso l'università.

## ***Servizio clienti***

Per eventuali chiarimenti in merito alla spedizione o al proprio conto, rivolgersi alla sede locale. Tenere presente che sarà necessario fornire il numero di serie.

## ***Corsi di formazione***

IBM Corp. organizza corsi di formazione pubblici e onsite che includono esercitazioni pratiche. Tali corsi si terranno periodicamente nelle principali città. Per ulteriori informazioni su questi seminari, andare a <http://www.ibm.com/software/analytics/spss/training>.

## ***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>

---

# Contenuto

## **Parte I: Manuale dell'utente**

### **1 Introduzione ai Valori mancanti 1**

### **2 Analisi dati mancanti 2**

Visualizzazione dei modelli dei valori mancanti . . . . .	5
Visualizzazione delle statistiche descrittive dei valori mancanti. . . . .	6
Stima delle statistiche e assegnazione dei valori mancanti . . . . .	8
Opzioni di stima EM . . . . .	9
Opzioni di stima regressione . . . . .	10
Variabili prevista e predittore. . . . .	11
Opzioni aggiuntive del comando MVA. . . . .	12

### **3 Assegnazione multipla 13**

Analizza modelli . . . . .	14
Assegna valori di dati mancanti . . . . .	16
Metodo . . . . .	19
Vincoli . . . . .	21
Output . . . . .	23
Opzioni aggiuntive del comando MULTIPLE IMPUTATION . . . . .	24
Utilizzo dei dati ad assegnazione multipla . . . . .	24
Analisi di dati ad assegnazione multipla . . . . .	28
Opzioni Assegnazione multipla . . . . .	33

## **Parte II: Esempi**

### **4 Analisi dei valori mancanti 36**

Descrizione del modello dei dati mancanti . . . . .	36
Esecuzione dell'analisi per visualizzare le statistiche descrittive. . . . .	36
Valutazione delle statistiche descrittive . . . . .	37
Riesecuzione dell'analisi per visualizzare i modelli . . . . .	43

Valutazione della tabella dei modelli . . . . .	45
Riesecuzione dell'analisi per il test MCAR di Little . . . . .	46
<b>5 Assegnazione multipla</b>	<b>48</b>
Utilizzo dell'assegnazione multipla per completare e analizzare un insieme di dati . . . . .	48
Analisi dei modelli dei valori mancanti . . . . .	48
Assegnazione automatica dei valori mancanti . . . . .	52
Modello di assegnazione personalizzato . . . . .	59
Verifica della convergenza FCS . . . . .	67
Analisi dei dati completi. . . . .	70
Riepilogo . . . . .	81
 <b>Appendici</b>	
 <b>A File di esempio</b>	 <b>82</b>
 <b>B Note legali</b>	 <b>92</b>
 <b>Indice</b>	 <b>95</b>

***Parte I:***  
***Manuale dell'utente***



# Introduzione ai Valori mancanti

I casi con valori mancanti costituiscono una grande sfida in quanto le procedure tipiche di creazione dei modelli non tengono in considerazione questi casi nell'analisi. Quando i valori mancanti sono pochi (circa meno del 5% del totale dei casi) e tali valori possono essere considerati come mancanti in modo casuale (cioè quando la mancanza di un valore non dipende da altri valori), il metodo tipico di eliminazione listwise è relativamente "affidabile". L'opzione Valori mancanti consente di stabilire se l'eliminazione listwise è sufficiente e fornisce i metodi di gestione dei valori mancanti quando l'eliminazione listwise risulta inappropriata.

## **Confronto tra la procedura Analisi valori mancanti e la procedura Assegnazione multipla**

L'opzione Valori mancanti fornisce due serie di procedure per la gestione di tali valori:

- Le procedure di [Assegnazione multipla](#) consentono l'analisi dei modelli di dati mancanti e hanno come scopo l'eventuale assegnazione multipla dei valori mancanti. In altre parole, vengono prodotte più versioni dell'insieme di dati, ognuna contenente il proprio insieme di valori assegnati. Quando vengono eseguite le analisi statistiche, le stime dei parametri di tutti gli insiemi di dati assegnati vengono raggruppate, arrivando a creare stime che risultano solitamente più accurate di quanto non sarebbero state con una sola assegnazione.
- [Analisi dei valori mancanti](#) fornisce una serie di strumenti descrittivi leggermente diversa per l'analisi dei dati mancanti (in modo particolare il test MCAR di Little) e include diversi metodi di assegnazione singola. Si noti che l'assegnazione multipla è generalmente ritenuta più attendibile rispetto all'assegnazione singola.

## **Operazioni sui valori mancanti**

Per iniziare l'analisi dei valori mancanti, seguire questa procedura di base:

- ▶ **Esaminare la mancanza.** Utilizzare l'analisi dei valori mancanti e l'analisi dei modelli per esplorare i modelli dei valori mancanti nei dati e determinare se è necessario ricorrere all'assegnazione multipla.
- ▶ **Assegna i valori mancanti** Utilizzare Assegna valori di dati mancanti per eseguire l'assegnazione multipla dei dati mancanti.
- ▶ **Analizzare i dati "completi".** Utilizzare qualsiasi procedura che supporta i dati ad assegnazione multipla. Vedere [Analisi di dati ad assegnazione multipla](#) a pag. 28 per informazioni sull'analisi degli insiemi di dati ad assegnazione multipla e un elenco di procedure che supportano questi dati.

# ***Analisi dati mancanti***

Missing Value Analysis consente di eseguire tre operazioni principali:

- Descrivere il modello per i dati mancanti. La posizione dei valori mancanti. La loro estensione. Se le coppie di variabili tendono ad avere valori mancanti in più casi. Se i valori di dati sono estremi. Se i valori mancanti sono casuali.
- Stimare medie, deviazioni standard, covarianze e correlazioni per diversi metodi dei valori mancanti: listwise, pairwise, regressione o EM (Expectation-Maximization). Con il metodo pairwise vengono inoltre visualizzati i conteggi di casi completi pairwise.
- Compilare (assegnare) i valori mancanti mediante dei valori stimati utilizzando i metodi di regressione o EM; tuttavia, l'assegnazione multipla è solitamente ritenuta più attendibile dal punto di vista dell'accuratezza dei risultati.

L'analisi dei dati mancanti consente di risolvere diversi problemi dovuti alla presenza di dati incompleti. Se i casi con valori mancanti risultano sistematicamente diversi dai casi privi di valori mancanti, i risultati possono essere fuorvianti. I dati mancanti possono inoltre ridurre il grado di precisione di statistiche calcolate in quanto la quantità di informazioni risulta inferiore a quella prevista inizialmente. Un altro problema consiste nel fatto che le ipotesi su cui si fondano molte procedure statistiche sono basate su casi completi e di conseguenza i valori mancanti possono complicare la teoria richiesta.

**Esempio.** Nella valutazione di una cura contro la leucemia vengono misurate diverse variabili. Le varie misure tuttavia non sono tutte disponibili per ciascun paziente. I modelli dei dati mancanti, visualizzati in tabelle, risultano casuali. Per stimare medie, correlazioni e covarianze viene usata un'analisi EM. Questa analisi viene anche utilizzata per determinare la mancanza di dati in modo completamente casuale. I valori mancanti vengono sostituiti con altri valori e salvati in un nuovo file di dati per analisi successive.

**Statistiche.** Statistiche univariate, tra cui il numero di valori non mancanti, la media, la deviazione standard, il numero di valori mancanti e il numero di valori estremi. Medie stimate, matrice di covarianza e matrice di correlazione usando i metodi listwise, pairwise, EM o regressione. Test MCAR di Little con risultati EM. Riepilogo delle medie tramite vari metodi. Per i gruppi definiti in base a valori mancanti e valori non mancanti: test *t*. Per tutte le variabili: modelli dei valori mancanti visualizzati in base a casi per variabili.

## ***Considerazioni sui dati***

**Dati.** I dati possono essere categoriali o quantitativi (scala o continui). Tuttavia, è possibile stimare le statistiche e assegnare i dati mancanti solo alle variabili quantitative. Per ciascuna variabile, i valori mancanti non codificati come mancanti di sistema devono essere definiti come definibili dall'utente. Se, ad esempio, si desidera gestire come valore mancante la risposta *Non so* codificata come 5 e inclusa in un elemento di un questionario, è necessario associare a tale elemento il valore 5 come valore mancante definito dall'utente.

**Ponderazione.** La ponderazione della frequenza viene ignorata da questa procedura. I casi con valore di ponderazione della replica negativo o uguale a zero vengono ignorati. I pesi non interi vengono troncati.

**Assunzioni.** Le stime listwise, pairwise e regressione sono basate sull'ipotesi secondo cui il modello dei valori mancanti non dipende dai valori dei dati. Questa condizione è definita MCAR, acronimo dell'inglese **Missing Completely At Random**, ovvero completamente mancante in modo casuale. Quindi, tutti i metodi (incluso il metodo EM) per la stima forniscono stime coerenti e non parziali delle correlazioni e delle covarianze quando i dati sono di tipo MCAR. La violazione dell'assunzione MCAR può portare a stime imparziali generate dai metodi listwise, pairwise e regressione. Se i dati non sono di tipo MCAR, è necessario utilizzare la stima EM.

La stima EM è basata sull'ipotesi secondo cui il modello dei dati mancanti è correlato soltanto ai dati osservati. Questa condizione è definita **MAR**, acronimo dell'inglese Missing At Random, ovvero mancante in modo casuale. Questa ipotesi consente la correzione delle stime in base alle informazioni disponibili. Ad esempio, in uno studio relativo a istruzione e reddito, i soggetti con basso livello di istruzione possono avere più valori di reddito mancanti. In questo caso i dati sono di tipo MAR e non MCAR. In altre parole, per i dati MAR, la probabilità che il reddito venga registrato dipende dal livello di istruzione del soggetto. La probabilità può variare in base all'istruzione ma non in base al reddito *entro tale livello di educazione*. Se la probabilità che il reddito venga registrato varia anche in base al valore del reddito in ciascun livello di istruzione (ad esempio, le persone con reddito alto non vengono segnalate), i dati non sono né di tipo MCAR né di tipo MAR. Questo non è un caso comune e, nel caso, nessuno dei metodi è appropriato.

**Procedure correlate.** Molte procedure consentono di usare stime listwise o pairwise. Le procedure di analisi Regressione lineare e Fattoriale consentono di sostituire i valori mancanti con i valori medi. Nel modulo aggiuntivo Forecasting sono disponibili diversi metodi per la sostituzione dei dati mancanti in serie storiche.

#### ***Per ottenere l'analisi dei dati mancanti***

- ▶ Dai menu, scegliere:  
Analizza > Analisi dati mancanti...

Figura 2-1  
Finestra di dialogo *Analisi dati mancanti*



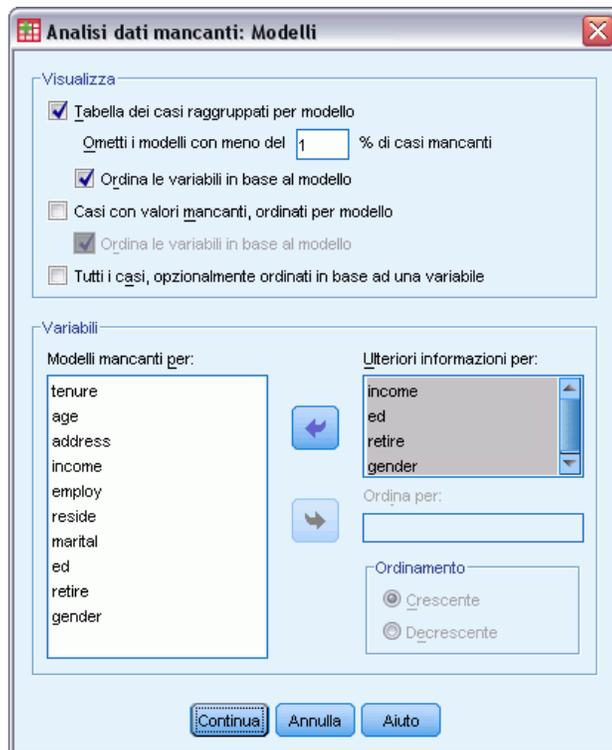
- Selezionare almeno una variabile quantitativa (scala) per la stima delle statistiche e facoltativamente per l'assegnazione dei valori mancanti.

Se lo si desidera, è possibile:

- Selezionare variabili categoriali (numeriche o stringa) e inserire un limite massimo per il numero di categorie nella casella Max. categorie.
- Fare clic su Modelli per rappresentare in una tabella i modelli dei dati mancanti. [Per ulteriori informazioni, vedere l'argomento Visualizzazione dei modelli dei valori mancanti a pag. 5.](#)
- Fare clic su Descrittive per vedere le statistiche descrittive dei valori mancanti. [Per ulteriori informazioni, vedere l'argomento Visualizzazione delle statistiche descrittive dei valori mancanti a pag. 6.](#)
- Selezionare un metodo di stima per le statistiche (medie, covarianze e correlazioni) e possibilmente per l'assegnazione dei valori mancanti. [Per ulteriori informazioni, vedere l'argomento Stima delle statistiche e assegnazione dei valori mancanti a pag. 8.](#)
- Se si seleziona EM o Regressione, fare clic su Variabili per specificare un sottoinsieme da usare per la stima. [Per ulteriori informazioni, vedere l'argomento Variabili prevista e predittore a pag. 11.](#)
- Selezionare una variabile per le etichette dei casi. Questa variabile viene utilizzata per assegnare etichette ai casi nelle tabelle di modelli che mostrano singoli casi.

## Visualizzazione dei modelli dei valori mancanti

Figura 2-2  
Finestra di dialogo Analisi dei dati mancanti: Modelli



È possibile scegliere di visualizzare varie tabelle che mostrano i modelli e la misura dei dati mancanti. Tali tabelle consentono di identificare:

- La posizione dei valori mancanti
- Se le coppie di variabili tendono ad avere valori mancanti in singoli casi
- Se i valori dei dati sono estremi

### Visualizza

Queste tabelle sono disponibili per la visualizzazione dei modelli per i dati mancanti.

**Tabella dei casi.** I modelli dei dati mancanti nelle variabili di analisi sono inseriti in tabelle, con le frequenze per ogni modello. Utilizzare Ordina le variabili in base al modello per specificare se i conteggi e le variabili sono ordinati per similarità di modelli. Utilizzare Ometti i modelli con meno del n% di casi per eliminare i modelli riscontrati non frequentemente.

**Casi con valori mancanti.** Ogni caso con un valore mancante o estremo viene inserito in tabella per ogni variabile di analisi. Utilizzare Ordina le variabili in base al modello per specificare se i conteggi e le variabili sono ordinati per similarità di modelli.

**Tutti i casi.** Ogni caso viene inserito in tabella e per ogni variabile vengono indicati i valori estremi e mancanti. I casi sono elencati nello stesso ordine in cui sono elencati nel file di dati, a meno che non si specifichi una variabile Ordina per.

Nelle tabelle che mostrano singoli casi, vengono utilizzati i seguenti simboli:

+	Valore estremamente alto
-	Valore estremamente basso
C	Valore mancante di sistema
T	Primo tipo di valore mancante definito dall'utente
P	Secondo tipo di valore mancante definito dall'utente
O	Terzo tipo di valore mancante definito dall'utente

### **Variabili**

È possibile visualizzare informazioni aggiuntive per le variabili incluse nell'analisi. Le variabili aggiunte a Ulteriori informazioni per vengono visualizzate singolarmente nella tabella dei modelli mancanti. Per le variabili quantitative (scala), viene visualizzata la media; per le variabili categoriali, viene visualizzato il numero di casi contenenti il modello in ogni categoria.

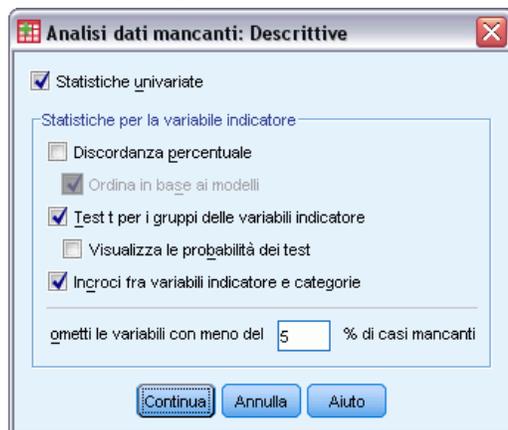
- **Ordina per.** I casi vengono elencati in base all'ordinamento crescente o decrescente dei valori della variabile specificata. Disponibile solo per l'opzione Tutti i casi.

### **Per visualizzare i modelli dei valori mancanti**

- ▶ Nella finestra di dialogo principale Analisi dei dati mancanti, selezionare le variabili di cui visualizzare i modelli dei dati mancanti.
- ▶ Fare clic su Modelli.
- ▶ Selezionare la tabella o le tabelle di modelli da visualizzare.

## **Visualizzazione delle statistiche descrittive dei valori mancanti**

Figura 2-3  
Finestra di dialogo Analisi dati mancanti: Descrittive



### **Statistiche univariate**

Le statistiche univariate consentono di identificare la misura generale dei dati mancanti. Per ogni variabile viene visualizzato quanto descritto di seguito:

- Numero di valori non mancanti
- Numero e percentuale di valori mancanti

Per le variabili quantitative (scala) viene inoltre visualizzato quanto descritto di seguito:

- Media
- Deviazione standard
- Numero di valori estremamente alti e bassi

### **Statistiche per la variabile indicatore**

Per ogni variabile, viene creata una variabile indicatore. Questa variabile categoriale indica se la variabile è presente o mancante relativamente a un singolo caso. Le variabili indicatore vengono utilizzate per creare tabelle di discordanza, di test  $t$  e di frequenza.

**Discordanza percentuale.** Per ciascuna coppia di variabili viene visualizzata la percentuale di casi in cui è presente una variabile con un valore mancante e una seconda variabile con un valore non mancante. Ciascun elemento sulla diagonale della tabella include la percentuale di valori mancanti di una singola variabile.

**Test  $t$  per i gruppi delle variabili indicatore.** Le medie di due gruppi vengono confrontate per ciascuna variabile quantitativa tramite la statistica  $t$  di Student. I gruppi specificano se una variabile è presente o mancante. Vengono visualizzati i gradi di libertà, la statistica  $t$ , i conteggi di valori mancanti e non mancanti e le medie dei due gruppi. È inoltre possibile visualizzare le eventuali probabilità a due code associate alla statistica  $t$ . Se l'analisi genera uno o più test, non utilizzare tali probabilità per il test di significatività. Le probabilità sono appropriate solo quando viene calcolato un singolo test.

**Incroci fra variabili indicatore e categorie.** Viene visualizzata una tabella per ogni variabile categoriale. Per ogni categoria, nella tabella sono riportate la frequenza e la percentuale di valori non mancanti per le altre variabili. Vengono inoltre visualizzate le percentuali di ciascun tipo di valore mancante.

**Ometti le variabili con meno del % di casi.** Per ridurre le dimensioni della tabella, è possibile omettere le statistiche che vengono calcolate esclusivamente per un numero limitato di casi.

### **Per visualizzare le statistiche descrittive**

- ▶ Nella finestra di dialogo principale Analisi dati mancanti, selezionare le variabili di cui visualizzare le statistiche descrittive dei valori mancanti.
- ▶ Fare clic su Descrittive.
- ▶ Scegliere le statistiche descrittive da visualizzare.

## **Stima delle statistiche e assegnazione dei valori mancanti**

È possibile stimare medie, deviazioni standard, covarianze e correlazioni tramite uno dei metodi listwise (solo casi completi), pairwise, EM (Expectation-Maximization) e/o il metodo di regressione. È inoltre possibile scegliere di assegnare i valori mancanti (stima dei valori di sostituzione). Si noti che l'**Assegnazione multipla** è generalmente ritenuta più efficace dell'assegnazione singola per risolvere il problema dei valori mancanti. Il test MCAR di Little è comunque utile per determinare se è necessario utilizzare l'assegnazione.

### **Metodo listwise**

Questo metodo utilizza solo casi completi. Se vi sono variabili con valori mancanti, il caso viene omesso dai calcoli.

### **Metodo pairwise**

Questo metodo esamina le coppie di variabili di analisi e utilizza un caso solo se dispone di valori non mancanti per entrambe le variabili. Le frequenze, le medie e le deviazioni standard sono calcolate separatamente per ogni coppia. Poiché altri valori mancanti nel caso vengono ignorati, le correlazioni e le covarianze per due variabili non dipendono dai valori mancanti in nessun'altra variabile.

### **Metodo EM**

Questo metodo presuppone una distribuzione per i dati parzialmente mancanti e per le inferenze di base della verosimiglianza di tale distribuzione. Ogni iterazione è composta da un passo E e da un passo M. Il passo E trova l'aspettativa condizionale dei dati "mancanti", in base ai valori osservati e alle stime correnti dei parametri. Queste aspettative vengono quindi sostituite per i dati "mancanti". Nel passo M, le stime di verosimiglianza massima dei parametri vengono calcolate come se i dati mancanti siano stati inseriti. "Mancante" è racchiuso tra virgolette poiché i valori mancanti non vengono inseriti direttamente. Le loro funzioni vengono invece utilizzate nella verosimiglianza.

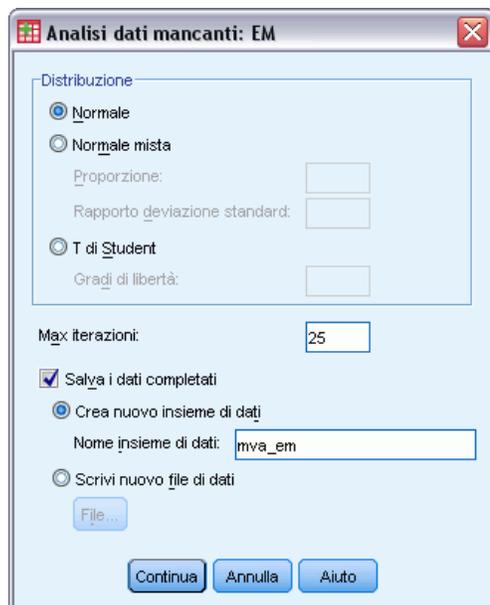
La statistica del chi-quadrato di Roderick J. A. Little per il test dei valori completamente mancanti in modo casuale (MCAR) viene stampata come un'annotazione nelle matrici EM. Per questo test, l'ipotesi nulla è che i dati sono completamente mancanti in modo casuale e che il valore  $p$  è significativo al livello 0,05. Se il valore è inferiore a 0,05, i dati non sono completamente mancanti in modo casuale. I dati potrebbero essere mancanti in modo casuale (MAR) o non mancanti in modo casuale (NMAR). Non è possibile presumere l'uno o l'altro ed è necessario analizzare i dati per determinare in che modo i dati sono mancanti.

### **Metodo di regressione (Analisi fattoriale)**

Questo metodo calcola le stime di regressione lineare multipla e dispone di opzioni per aumentare le stime con componenti casuali. Per ogni valore atteso, la procedura può aggiungere un residuo di un caso completo selezionato in modo casuale, una deviazione standard casuale o una deviazione casuale (scalata per la radice quadrata della media dei quadrati residua) dalla distribuzione  $t$ .

## Opzioni di stima EM

Figura 2-4  
Finestra di dialogo Analisi dati mancanti: EM



Utilizzando un processo iterativo, il metodo EM consente di stimare le medie, la matrice di covarianza e la correlazione di variabili quantitative (scala) con valori mancanti.

**Distribuzione.** EM crea inferenze in base alla verosimiglianza nella distribuzione specifica. Per impostazione predefinita, si suppone una distribuzione normale. Se le code di distribuzione sono più lunghe rispetto a quelle di una distribuzione normale, è possibile richiedere che la procedura generi la funzione di verosimiglianza da una distribuzione  $t$  di Student con  $n$  gradi di libertà. La distribuzione normale mista fornisce inoltre una distribuzione con code più lunghe. Specificare il rapporto delle deviazioni standard della distribuzione normale mista e la proporzione delle due distribuzioni. La distribuzione normale mista presuppone che solo le deviazioni standard delle distribuzioni siano differenti. Le medie devono essere uguali.

**Max iterazioni.** Imposta il numero massimo di iterazioni per stimare la reale covarianza. Questa procedura viene interrotta al raggiungimento del numero di iterazioni specificato, anche se per le stime non è stata ottenuta alcuna convergenza.

**Salva i dati completati.** È possibile salvare un insieme di dati con i valori assegnati al posto dei valori mancanti. Tenere presente che le statistiche basate sulla covarianza che utilizzano i valori assegnati sottostimeranno i relativi valori dei parametri. Il grado di sottostima è proporzionale al numero di casi inosservati.

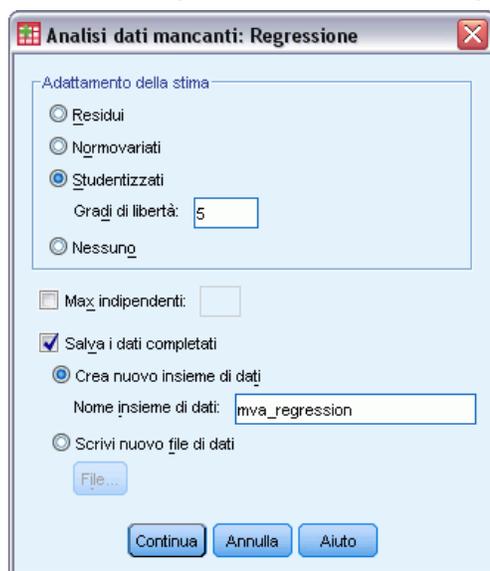
### Per specificare le opzioni EM

- ▶ Nella finestra di dialogo principale Analisi dati mancanti, selezionare le variabili di cui stimare i valori mancanti utilizzando il metodo EM.
- ▶ Selezionare EM nel gruppo Stima.

- ▶ Per specificare variabili dipendenti e indipendenti, fare clic su Variabili. [Per ulteriori informazioni, vedere l'argomento Variabili prevista e predittore a pag. 11.](#)
- ▶ Fare clic su EM.
- ▶ Selezionare le opzioni EM desiderate.

## Opzioni di stima regressione

Figura 2-5  
Finestra di dialogo Analisi dati mancanti: Regressione



Il metodo di regressione consente di stimare i valori mancanti utilizzando la regressione lineare multipla. Vengono visualizzate le medie, la matrice di covarianza e la matrice di correlazione delle variabili dipendenti.

**Adattamento della stima.** Con il metodo Regressione è possibile aggiungere un componente casuale nelle stime di regressione. È possibile selezionare residui, normovariati, variate  $t$  di Student oppure nessun adattamento.

- **Residui.** I termini di errore vengono scelti casualmente dai residui osservati dei casi completi da aggiungere alla stime di regressione.
- **Normovariati.** I termini di errore sono estratti casualmente da una distribuzione con valore atteso 0 e deviazione standard pari alla radice quadrata del quadrato dell'errore medio del modello di regressione.
- **Studentizzati.** I termini di errore vengono estratti casualmente dalla distribuzione  $t$  con i gradi di libertà specificati e scalati in base all'errore quadratico medio.

**Max indipendenti.** Consente di impostare il numero massimo di variabili stimatori, o indipendenti, usate nel processo di stima.

**Salva i dati completati.** Scrive un file di dati nel file di dati attivo o in un file di dati esterno in formato IBM® SPSS® Statistics, sostituendo i valori mancanti con valori stimati tramite il metodo di regressione.

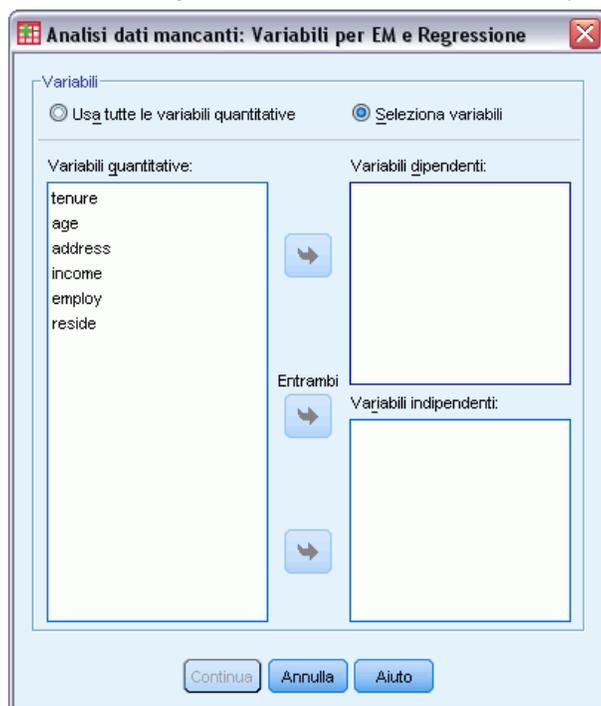
#### **Per specificare le opzioni di regressione**

- ▶ Nella finestra di dialogo principale Analisi dati mancanti, selezionare le variabili di cui stimare i valori mancanti utilizzando il metodo di regressione.
- ▶ Selezionare **Regressione** nel gruppo **Stima**.
- ▶ Per specificare variabili dipendenti e indipendenti, fare clic su **Variabili**. [Per ulteriori informazioni, vedere l'argomento Variabili prevista e predittore a pag. 11.](#)
- ▶ Fare clic su **Regressione**.
- ▶ Selezionare le opzioni di regressione desiderate.

### **Variabili prevista e predittore**

Figura 2-6

Finestra di dialogo Analisi dei dati mancanti: Variabili per EM e Regressione



Per impostazione predefinita, tutte le variabili quantitative vengono utilizzate per la stima EM e di regressione. Se necessario, nelle stime è possibile scegliere variabili specifiche come variabili prevista e predittore. Una determinata variabile può trovarsi in entrambi gli elenchi ma vi sono casi in cui è necessario limitare l'uso di una variabile. Ad esempio, alcuni analisti trovano scomodo stimare i valori delle variabili dei risultati. È inoltre possibile utilizzare variabili

differenti per stime differenti ed eseguire più volte la procedura. Ad esempio, se si dispone di un insieme di elementi che rappresentano le valutazioni delle infermiere e un altro insieme con le valutazioni dei medici, è possibile effettuare un'esecuzione per le infermiere per stimare i relativi elementi mancanti e un'altra esecuzione per la stima degli elementi dei medici.

Un'altra considerazione nasce quando si utilizza il metodo di regressione. Nella regressione multipla, l'utilizzo di un sottoinsieme di variabili indipendenti di grandi dimensioni può produrre valori attesi peggiori rispetto a un sottoinsieme più piccolo. Quindi, una variabile deve ottenere un limite  $4.0 F$  di inserimento per essere utilizzata. Questo limite può essere modificato con la sintassi.

### ***Per specificare le variabili prevista e predittore***

- ▶ Nella finestra di dialogo principale Analisi dati mancanti, selezionare le variabili di cui stimare i valori mancanti utilizzando il metodo di regressione.
- ▶ Dal gruppo Stima, selezionare EM o Regressione.
- ▶ Fare clic su Variabili.
- ▶ Per utilizzare variabili specifiche anziché tutte le variabili come prevista e predittore, selezionare Seleziona variabili e spostare le variabili negli elenchi appropriati.

## ***Opzioni aggiuntive del comando MVA***

Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare variabili descrittive distinte per i modelli di valori mancanti e di dati e i modelli a tabelle mediante la parola chiave `DESCRIBE` nei sottocomandi `MPATTERN`, `DPATTERN` o `TPATTERN`.
- Specificare più variabili di ordinamento per la tabella dei modelli dei dati tramite il sottocomando `DPATTERN`.
- Specificare più variabili di ordinamento per i modelli dei dati tramite il sottocomando `DPATTERN`.
- Specificare tolleranza e convergenza tramite il sottocomando `EM`.
- Specificare tolleranza e  $F$  di inserimento tramite il sottocomando `REGRESSION`.
- Specificare elenchi di variabili diversi per EM e Regressione tramite i sottocomandi `EM` e `REGRESSION`.
- Specificare percentuali diverse per la soppressione di casi con ciascuno dei sottocomandi `TTESTS`, `TABULATE` e `MISMATCH`.

Vedere *Command Syntax Reference* per informazioni dettagliate sulla sintassi.

# Assegnazione multipla

L'assegnazione multipla ha lo scopo di generare dei valori possibili per i valori mancanti, creando in questo modo diversi insiemi di dati "completi". Le procedure analitiche che funzionano con gli insiemi di dati delle assegnazioni multiple producono output per ogni insieme di dati "completo" nonché output raggruppato che stima i risultati che sarebbero stati prodotti se nell'insieme di dati originale non ci fossero stati valori mancanti. Questi risultati combinati sono solitamente più accurati di quelli forniti dai metodi di assegnazione singola.

**Variabili dell'analisi.** Le variabili dell'analisi possono essere:

- **Nominale.** Una variabile può essere considerata nominale quando i relativi valori rappresentano categorie prive di ordinamento intrinseco, per esempio l'ufficio di una società, Tra gli esempi di variabili nominali troviamo la regione, il codice postale e la religione.
- **Ordinale.** Una variabile può essere considerata ordinale quando i relativi valori rappresentano categorie con qualche ordinamento intrinseco, per esempio i gradi di soddisfazione per un servizio, da molto insoddisfatto a molto soddisfatto, i punteggi di atteggiamento corrispondenti a gradi di soddisfazione o fiducia e i punteggi di preferenza.
- **Scala.** Una variabile può essere considerata di scala (continua) quando i relativi valori rappresentano categorie ordinate con una metrica significativa, tale che i confronti fra le distanze dei relativi valori siano appropriati. Esempi di variabili di scala sono l'età espressa in anni o il reddito espresso in migliaia di Euro.

La procedura presume che il livello di misurazione appropriato sia stato assegnato a tutte le variabili; tuttavia, è possibile modificare temporaneamente il livello di misurazione di una variabile facendo clic con il pulsante destro del mouse sulla variabile nell'elenco delle variabili sorgente e scegliendo un livello di misurazione dal menu di scelta rapida.

L'icona accanto a ciascuna variabile nell'elenco delle variabili identifica il livello di misurazione e il tipo di dati.

	Numerico	Stringa	Data	Ora
Scala (continuo)		n/d		
Ordinale				
Nominale				

**Ponderazione.** La ponderazione della frequenza viene ignorata da questa procedura. I casi con valore di ponderazione della replica negativo o uguale a zero vengono ignorati. I pesi non interi vengono arrotondati al numero intero più vicino.

**Peso analisi.** I pesi di analisi (regressione o campionamento) sono incorporati nei riepiloghi dei valori mancanti e nei modelli di assegnazione adatti. I casi con un peso di analisi negativo o pari a zero vengono esclusi.

**Campioni complessi.** La procedura Assegnazione multipla non manipola esplicitamente gli strati, i cluster, o altre strutture di campionamento complesse, benché possa accettare pesi di campionamento finali sotto forma di variabile di peso dell'analisi. Si noti inoltre che le procedure di campionamento complesse attualmente non analizzano automaticamente gli insiemi di dati assegnati mediante assegnazione multipla. Per l'elenco completo delle procedure che supportano il raggruppamento, vedere [Analisi di dati ad assegnazione multipla](#) a pag. 28.

**Valori mancanti.** Sia i valori mancanti definiti dall'utente sia i valori mancanti di sistema vengono considerati come valori non validi; vale a dire, entrambi questi tipi di valori mancanti vengono sostituiti durante l'assegnazione dei valori ed entrambi sono considerati come valori non validi di variabili usate come predittori nei modelli di assegnazione. I valori mancanti di sistema e definiti dall'utente verranno considerati come mancanti anche nelle analisi dei valori mancanti.

**Replica dei risultati (Assegna valori di dati mancanti).** Se si desidera replicare esattamente i risultati dell'assegnazione, utilizzare lo stesso valore di inizializzazione per il generatore di numeri casuali, lo stesso ordine dei dati e delle variabili e le stesse impostazioni della procedura.

- **Generazione di numeri casuali.** La procedura utilizza la generazione di numeri casuali durante il calcolo dei valori assegnati. Per riprodurre gli stessi risultati randomizzati in futuro, utilizzare lo stesso valore di inizializzazione per il generatore di numeri casuali per ciascuna esecuzione della procedura Assegna valori di dati mancanti.
- **Ordine dei casi.** I valori vengono assegnati secondo l'ordine dei casi.
- **Ordine delle variabili.** Il metodo di assegnazione Specifica completamente condizionale assegna i valori nell'ordine specificato nell'elenco Variabili di analisi.

Sono disponibili due finestre di dialogo per l'assegnazione multipla.

- [Analizza modelli](#) fornisce misure descrittive dei modelli di valori mancanti e può essere utile come indagine prima dell'assegnazione.
- [Assegna valori di dati mancanti](#) viene utilizzata per generare assegnazioni multiple. Gli insiemi di dati completi possono essere analizzati mediante delle procedure che supportano gli insiemi di dati ad assegnazione multipla. Vedere [Analisi di dati ad assegnazione multipla](#) a pag. 28 per informazioni sull'analisi degli insiemi di dati ad assegnazione multipla e un elenco di procedure che supportano questi dati.

## ***Analizza modelli***

[Analizza modelli](#) fornisce le misure descrittive dei modelli dei valori mancanti nei dati e può essere utile come indagine prima dell'assegnazione.

**Esempio.** Un provider di telecomunicazioni desidera approfondire la conoscenza dei modelli di utilizzo dei servizi nel proprio database di clienti. Questo provider possiede dati completi circa i servizi utilizzati dai propri clienti, ma le informazioni di carattere demografico raccolte dall'azienda presentano alcuni valori mancanti. Dall'analisi dei modelli dei valori mancanti è possibile determinare i passaggi dell'assegnazione successivi. [Per ulteriori informazioni, vedere](#)

l'argomento Utilizzo dell'assegnazione multipla per completare e analizzare un insieme di dati in il capitolo 5 a pag. 48.

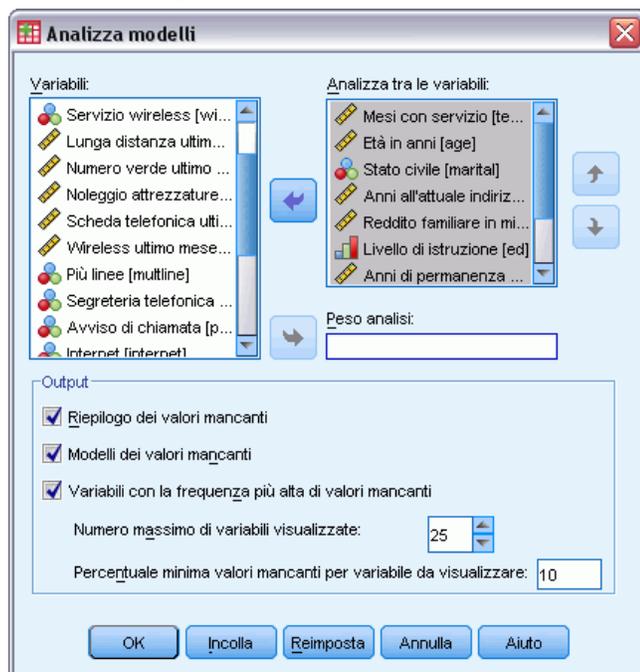
### **Per analizzare i modelli di dati mancanti**

Dai menu, scegliere:

Analizza > Assegnazione multipla > Analizza modelli...

Figura 3-1

Finestra di dialogo Analizza modelli



- Selezionare almeno due variabili di analisi. Questa procedura analizza i modelli di dati mancanti per queste variabili.

### **Impostazioni opzionali**

**Peso analisi.** Questa variabile contiene i pesi delle analisi (regressione o campionamento). La procedura incorpora i pesi di analisi nei riepiloghi dei valori mancanti. I casi con un peso di analisi negativo o pari a zero vengono esclusi.

**Output.** L'output opzionale disponibile è descritto di seguito:

- **Riepilogo dei valori mancanti.** Visualizza un grafico a torta a riquadri in cui sono illustrati il numero e la percentuale delle variabili di analisi, i casi e i singoli dati per cui mancano uno o più valori.
- **Modelli dei valori mancanti.** Visualizza i modelli di valori mancanti all'interno di tabelle. Ogni modello corrisponde a un gruppo di casi con lo stesso modello di dati incompleti e completi nelle variabili di analisi. Questo output consente di determinare se il metodo monotono di assegnazione può essere utilizzato con i propri dati e, in caso contrario, in che misura tali dati

si avvicinano a un modello monotono. La procedura ordina le variabili di analisi per rivelare o approssimare un modello monotono. Se dopo il riordinamento non rimane alcun modello non monotono, si può dedurre che i dati hanno un modello monotono quando le variabili di analisi sono ordinate in questo modo.

- **Variabili con la frequenza più alta di valori mancanti.** Visualizza una tabella di variabili di analisi ordinata in base alle percentuali di valori mancanti in ordine decrescente. La tabella include le statistiche descrittive (media e deviazione standard) per le variabili di scala.

È possibile controllare il numero massimo di variabili da visualizzare e la soglia percentuale minima di valori mancanti perché una variabile venga inclusa nella visualizzazione. Viene visualizzato il set di variabili che soddisfano entrambi i criteri. Ad esempio, l'impostazione del numero massimo di variabili su 50 e della percentuale di mancanza minima su 25 fa sì che la tabella visualizzi un massimo di 50 variabili con almeno il 25% di valori mancanti. Se sono presenti 60 variabili di analisi di cui solo 15 presentano una percentuale di valori mancanti pari al 25% o superiore, l'output visualizzerà solo 15 variabili.

## ***Assegna valori di dati mancanti***

Assegna valori di dati mancanti viene utilizzata per generare assegnazioni multiple. Gli insiemi di dati completi possono essere analizzati mediante delle procedure che supportano gli insiemi di dati ad assegnazione multipla. Vedere [Analisi di dati ad assegnazione multipla](#) a pag. 28 per informazioni sull'analisi degli insiemi di dati ad assegnazione multipla e un elenco di procedure che supportano questi dati.

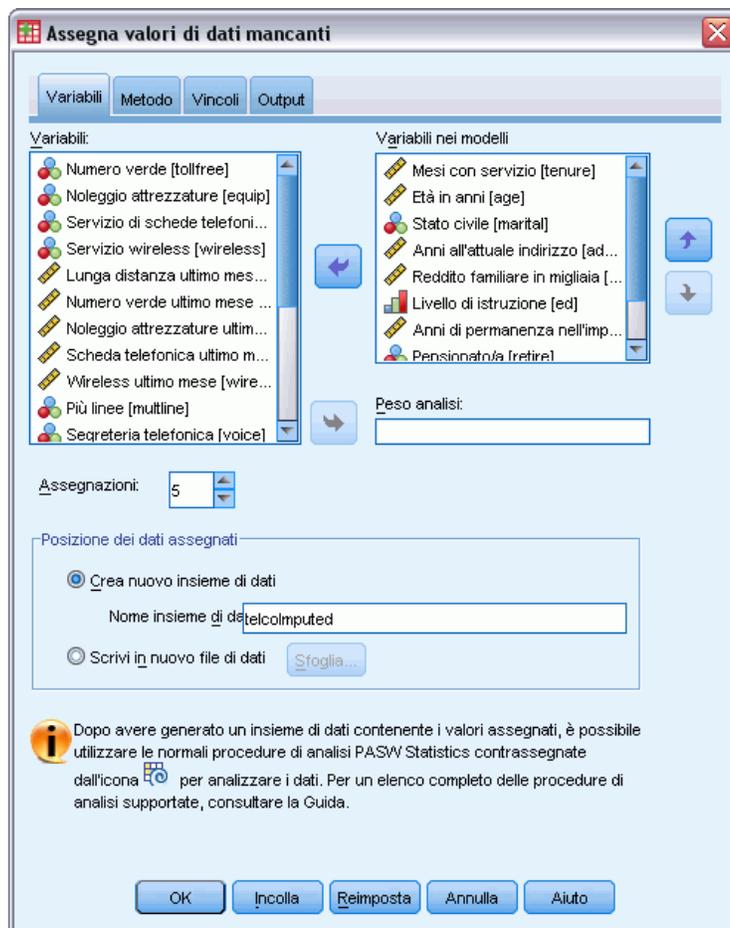
**Esempio.** Un provider di telecomunicazioni desidera approfondire la conoscenza dei modelli di utilizzo dei servizi nel proprio database di clienti. Questo provider possiede dati completi circa i servizi utilizzati dai propri clienti, ma le informazioni di carattere demografico raccolte dall'azienda presentano alcuni valori mancanti. Inoltre, tali valori non sono mancanti in modo interamente casuale, pertanto verrà utilizzata l'assegnazione multipla per completare l'insieme di dati. [Per ulteriori informazioni, vedere l'argomento Utilizzo dell'assegnazione multipla per completare e analizzare un insieme di dati in il capitolo 5 a pag. 48.](#)

### ***Per assegnare valori di dati mancanti***

Dai menu, scegliere:

Analizza > Assegnazione multipla > Assegna valori di dati mancanti...

Figura 3-2  
Finestra di dialogo Assegna valori di dati mancanti, scheda Variabili



- ▶ Selezionare almeno due variabili nel modello di assegnazione. La procedura assegna valori multipli per i dati mancanti di queste variabili.
- ▶ Specificare il numero delle assegnazioni da calcolare. Per impostazione predefinita, questo valore è 5.
- ▶ Specificare un insieme di dati o un file di dati in formato IBM® SPSS® Statistics nel quale scrivere i dati assegnati.

L'insieme di dati di output è composto dai dati del caso originale con dati mancanti a cui si aggiunge un insieme di casi con valori assegnati per ciascuna assegnazione. Ad esempio, se l'insieme di dati originale ha 100 casi e si hanno cinque assegnazioni, l'insieme di dati di output avrà 600 casi. Tutte le variabili nell'insieme di dati di input sono incluse nell'insieme di dati di output. Le proprietà dizionario (nomi, etichette, ecc.) delle variabili esistenti vengono copiate nel nuovo insieme di dati. Il file contiene anche una nuova variabile, *Imputation\_*, una variabile numerica che indica l'assegnazione (0 per i dati originali o 1..n per i casi che presentano valori assegnati).

Quando viene creato l'insieme di dati di output, la procedura definisce automaticamente la variabile *Imputation\_* come una variabile di distinzione. Se al momento dell'esecuzione della procedura sono attive le distinzioni, l'insieme di dati di output include un insieme di assegnazioni per ogni combinazione di valori delle variabili di distinzione.

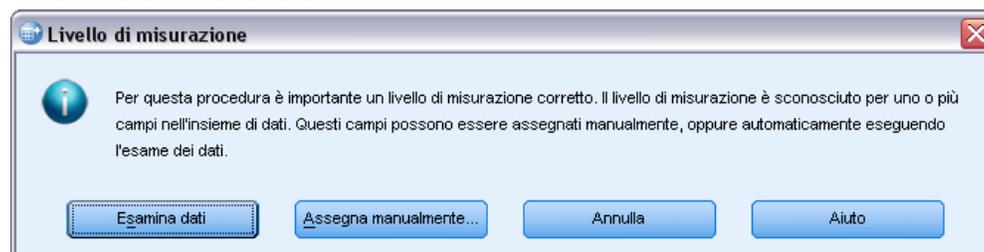
### **Impostazioni opzionali**

**Peso analisi.** Questa variabile contiene i pesi delle analisi (regressione o campionamento). La procedura incorpora i pesi di analisi nei modelli di regressione e classificazione utilizzati per assegnare i valori mancanti. I pesi di analisi vengono utilizzati anche nei riepiloghi dei valori assegnati, quali, ad esempio, media, deviazione standard ed errore standard. I casi con un peso di analisi negativo o pari a zero vengono esclusi.

### **Campi con livello di misurazione sconosciuto**

L'avviso Livello di misurazione viene visualizzato quando il livello di misurazione di una o più variabili (campi) dell'insieme di dati è sconosciuto. Poiché influisce sul calcolo dei risultati di questa procedura, il livello di misurazione deve essere definito per tutte le variabili.

Figura 3-3  
Avviso Livello di misurazione

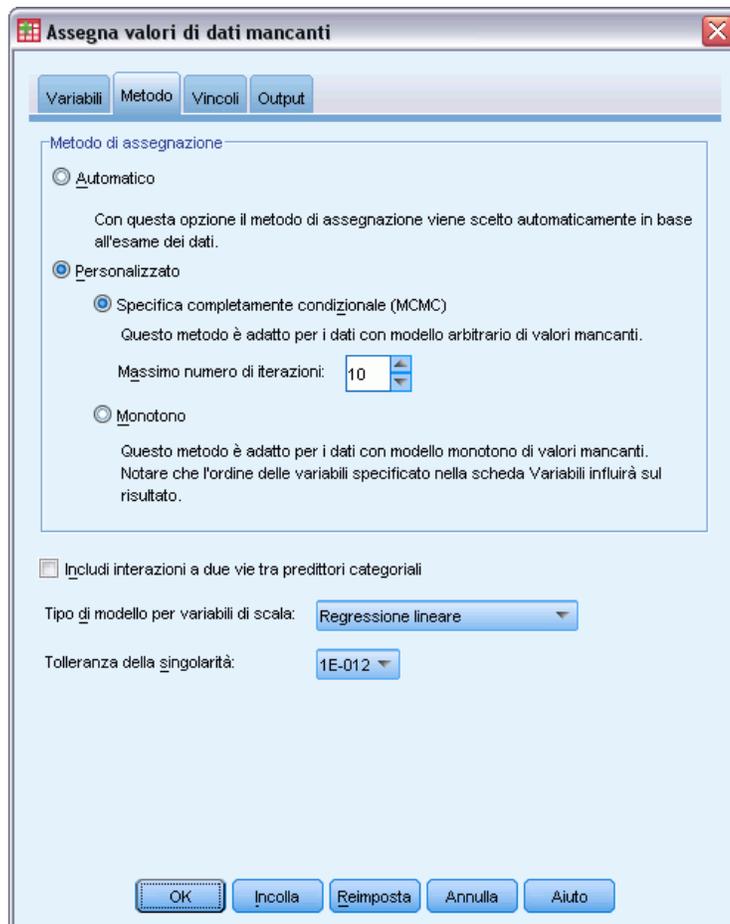


- **Esamina dati.** Legge i dati dell'insieme di dati attivo e assegna un livello di misurazione predefinito a tutti i campi con livello di misurazione sconosciuto. Con insiemi di dati di grandi dimensioni, questa operazione può richiedere del tempo.
- **Assegna manualmente.** Apre una finestra di dialogo che elenca tutti i campi con livello di misurazione sconosciuto, mediante la quale è possibile assegnare un livello di misurazione a questi campi. Il livello di misurazione si può assegnare anche nella Visualizzazione variabili dell'Editor dei dati.

Dal momento che il livello di misurazione è importante per questa procedura, è possibile accedere alla finestra di dialogo per la sua esecuzione solo quando per tutti i campi è stato definito un livello di misurazione.

## Metodo

Figura 3-4  
Finestra di dialogo Assegna valori di dati mancanti, scheda Metodo



La scheda Metodo specifica in che modo i valori mancanti verranno assegnati e include i tipi di modelli utilizzati. I predittori categoriali sono codificati in modo fittizio.

**Metodo di assegnazione.** Il metodo Automatico analizza i dati e utilizza il metodo monotono se i dati evidenziano un modello monotono di valori mancanti; in caso contrario, viene utilizzata la specifica completamente condizionale. Se si è certi del metodo che si vuole utilizzare, è possibile specificarlo come metodo Personalizzato.

- **Specifica completamente condizionale.** Si tratta di un metodo MCMC (Markov chain Monte Carlo) iterativo che può essere utilizzato quando il modello di dati mancanti è arbitrario (monotono o non monotono).

Per ciascuna iterazione e per ciascuna variabile nell'ordine specificato nell'elenco delle variabili, il metodo Specifica completamente condizionale (FCS) adatta un modello univariato (variabile dipendente singola) utilizzando tutte le altre variabili disponibili nel modello come predittori, quindi assegna i valori mancanti per la variabile che si sta adattando. Il metodo

continua finché non viene raggiunto il numero massimo di iterazioni e non vengono salvati nell'insieme di dati assegnato i valori assegnati all'iterazione massima.

**Max iterazioni.** Indica il numero di iterazioni, o fasi, compiute dalla catena di Markov utilizzata dal metodo Specifica completamente condizionale. Se il metodo della specifica completamente condizionale viene scelto automaticamente, utilizza il numero predefinito di 10 iterazioni. Al contrario, quando questo metodo viene scelto esplicitamente, è possibile specificare un numero personalizzato di iterazioni. Se la catena di Markov non converge, potrebbe essere necessario aumentare il numero di iterazioni. Nella scheda Output, è possibile salvare i dati cronologici delle iterazioni del metodo di specifica completamente condizionale ed eseguire un grafico per valutare la convergenza.

- **Monotono.** Si tratta di un metodo non iterativo che può essere utilizzato solo quando i dati presentano un modello di valori mancanti monotono. Per modello monotono si intende un caso in cui è possibile ordinare le variabili in modo che se una variabile ha un valore non mancante, tutte le variabili precedenti avranno dei valori non mancanti. Quando si specifica che si tratta di un metodo Personalizzato, assicurarsi di specificare le variabili nell'elenco in base a un ordine che mostri un motivo monotono.

Per ciascuna variabile nell'ordine monotono, il metodo monotono adatta un modello univariato (variabile dipendente singola) utilizzando tutte le variabili precedenti nel modello come predittori, quindi assegna i valori mancanti per la variabile che si sta adattando. Tali valori assegnati vengono salvati nell'insieme di dati assegnato.

**Includi interazioni a due vie.** Quando il metodo di assegnazione viene scelto automaticamente, il modello di assegnazione per ogni variabile include un termine costante ed effetti principali per le variabili predittore. Quando si sceglie un metodo specifico, è possibile, se lo si desidera, includere tutte le interazioni a due vie tra variabili predittore categoriali.

**Tipo di modello per variabili di scala.** Quando il metodo di assegnazione viene scelto automaticamente, per le variabili di scala viene utilizzata la regressione lineare come modello univariato. Quando si sceglie un metodo specifico, è possibile scegliere PMM (Predictive Mean Matching) come modello per le variabili di scala. PMM è una variante di regressione lineare che crea una corrispondenza tra i valori assegnati calcolati dal modello di regressione e il valore osservato più vicino.

Per le variabili categoriali viene sempre utilizzata la regressione logistica come modello univariato. Indipendentemente dal tipo di modello, i predittori categoriali vengono gestiti mediante una codifica fittizia.

**Tolleranza della singolarità.** Le matrici singole (o non reversibili) hanno colonne linearmente dipendenti che possono creare gravi problemi all'algoritmo di stima. Poiché anche le matrici quasi singolari possono generare risultati inadeguati, la procedura considera le matrici i cui determinanti sono inferiori alla tolleranza come singolari. Specificare un valore positivo.

## Vincoli

Figura 3-5  
Finestra di dialogo Assegna valori di dati mancanti, scheda Vincoli

Esame dati per riepilogo variabili

Limita numero di casi esaminati

Riepilogo variabili:

Variabili nei modelli	Percentuale dati ma...	Numero minimo dati...	Numero massimo d...
tenure	3,20	1	72
age	2,50	18	77
marital	11,50	0	1
address	15,00	0	55

Casi esaminati: 1000

Definisci vincoli:

Variabili nei modelli	Ruolo	Min	Max	Arrotond...
retire	Assegna e usa come pr...			
gender	Assegna e usa come pr...			
reside	Assegna e usa come pr...	1		1
lninc	Assegna e usa come pr...	0		

Escludi variabili con grandi quantità di dati mancanti

Percentuale massima dati mancanti:

Numero massimo estrazioni casi:

Numero massimo estrazioni parametri:

L'aumento del numero massimo di estrazioni di parametri può aumentare notevolmente la durata dell'analisi.

La scheda Vincoli consente di limitare il ruolo di una variabile durante l'assegnazione e di ridurre l'intervallo dei valori assegnati di una variabile di scala in modo che risultino plausibili. Inoltre, è possibile limitare l'analisi alle variabili con una percentuale inferiore a una percentuale massima di valori mancanti.

**Esame dati per riepilogo variabili.** Se si fa clic su Esamina dati l'elenco visualizza le variabili di analisi e la percentuale osservata per i valori mancanti, minimi e massimi di ciascuna. I riepiloghi possono essere basati su tutti i casi oppure limitati a un'analisi dei primi  $n$  casi, come specificato nella casella di testo Casi. Per aggiornare i riepiloghi delle distribuzioni, fare clic su Riesamina dati.

### Definisci vincoli

- Ruolo.** Consente di personalizzare l'insieme di variabili da assegnare e/o considerare come predittori. Tipicamente, ogni variabile di analisi viene considerata sia come una variabile dipendente sia come un predittore nel modello di assegnazione. Il Ruolo può essere utilizzato per disattivare l'assegnazione per le variabili che si desidera utilizzare con Usa come predittore solo o per impedire che le variabili vengano utilizzate come predittori (Assegna solo) al fine

di rendere più compatto il modello di previsione. Questo è l'unico vincolo che può essere specificato per le variabili categoriali, o per le variabili che vengono utilizzate solo come predittori.

- **Min e Max.** Queste colonne consentono di specificare i valori minimo e massimo di assegnazione consentiti per le variabili di scala. Se un valore assegnato non rientra in questo intervallo, la procedura continua a estrarre valori finché ne viene rilevato uno che rientra nell'intervallo o finché non viene raggiunto il numero massimo di estrazioni consentite (vedere la sezione relativa al numero massimo di estrazioni riportata di seguito). Queste colonne sono disponibili solo se la Regressione lineare è stata selezionata come modello di variabile di scala nella scheda Metodo.
- **Arrotondamento.** Alcune variabili possono essere utilizzate come scala, ma presentano valori che sono ulteriormente limitati per natura; ad esempio, il numero di persone che compongono un nucleo familiare deve essere un intero e la quantità di tempo impiegata per andare dal fruttivendolo non può avere frazioni in centesimi. Questa colonna consente di specificare la denominazione più piccola da accettare. Ad esempio, per ottenere dei valori interi, si dovrà specificare 1 come denominazione di arrotondamento, mentre per ottenere valori arrotondati al centesimo più vicino si dovrà specificare 0,01. In generale, i valori vengono arrotondati al multiplo intero più vicino della denominazione di arrotondamento. La tabella che segue mostra in che modo i diversi valori di arrotondamento agiscono su un valore assegnato pari a 6,64823 (prima dell'arrotondamento).

Denominazione di arrotondamento	Valore a cui viene arrotondato 6,64832
10	10
1	7
0.25	6.75
0.1	6.6
0.01	6.65

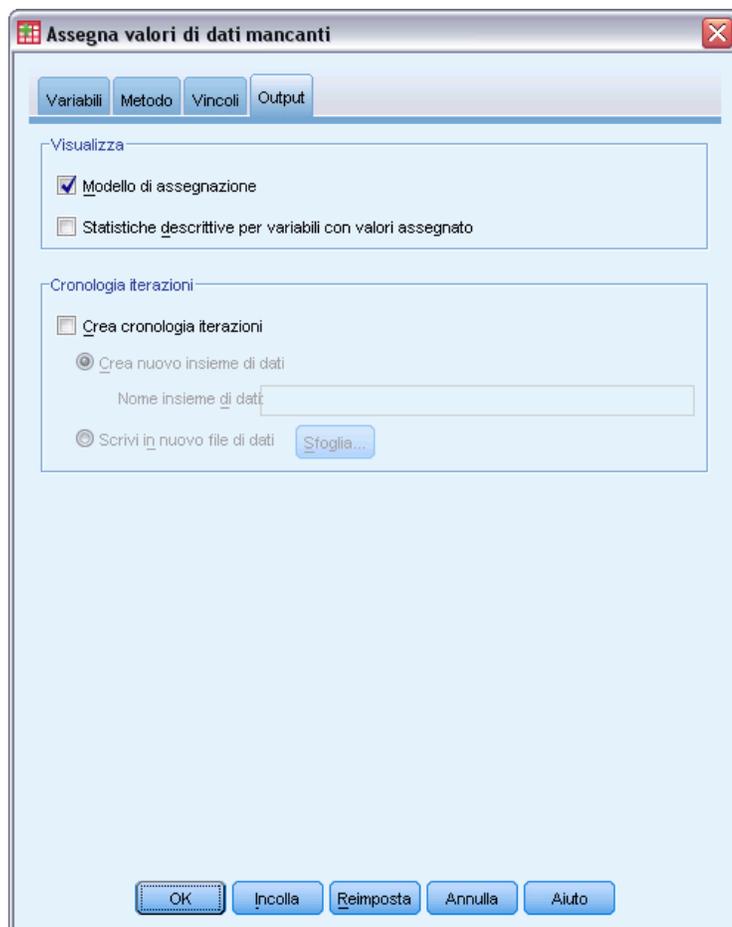
**Escludi variabili con grandi quantità di dati mancanti.** Solitamente, le variabili di analisi vengono assegnate e utilizzate come predittori indipendentemente dal numero di valori mancanti, a condizione che contengano un numero di dati sufficiente per eseguire la stima di un modello di assegnazione. È possibile scegliere di escludere le variabili che presentano un'alta percentuale di valori mancanti. Ad esempio, se si specifica 50 come Percentuale massima dati mancanti, le variabili di analisi che presentano più del 50% di valori mancanti non vengono assegnate né utilizzate come predittori nei modelli di assegnazione.

**Numero massimo di estrazioni.** Se sono stati specificati dei valori minimo e massimo per i valori assegnati delle variabili di scala (vedere la sezione precedente Min e Max), la procedura continua a estrarre valori per un caso finché non trova un insieme di valori che rientrano nell'intervallo specificato. Se con il numero di estrazioni per caso specificato non si ottiene un insieme di valori, la procedura estrae un altro insieme di parametri del modello e ripete il processo di estrazione dei casi. Se non si ottiene un insieme di valori compresi negli intervalli per il numero di estrazioni di casi e parametri specificato, si verifica un errore.

Si noti che l'incremento di questi valori può prolungare i tempi di elaborazione. Se la procedura impiega molto tempo o non è in grado di trovare estrazioni adatte, controllare che i valori minimo e massimo specificati siano appropriati.

## Output

Figura 3-6  
Finestra di dialogo *Assegna valori di dati mancanti*, scheda *Output*



**Visualizzazione.** Controlla la visualizzazione dell'output. Viene sempre visualizzato un riepilogo di assegnazione complessivo, contenente delle tabelle relative alle specifiche di assegnazione, le iterazioni (per il metodo di specifica completamente condizionale), le variabili dipendenti assegnate, le variabili dipendenti escluse dall'assegnazione e la sequenza di assegnazione. Inoltre, se sono stati specificati, vengono visualizzati anche i vincoli relativi alle variabili di analisi.

- **Modello di assegnazione.** Visualizza il modello di assegnazione per le variabili dipendenti e i predittori; include il tipo di modello univariato, gli effetti del modello e il numero di valori assegnati.
- **Statistiche descrittive.** Visualizza le statistiche descrittive relative alle variabili dipendenti per le quali sono stati assegnati i valori. Nel caso delle variabili di scala, le statistiche descrittive includono media, conteggio, deviazione standard, valore minimo e valore massimo per i dati di input originali (precedenti all'assegnazione), valori assegnati (dalla procedura di assegnazione) e dati completi (dati originali e assegnati insieme, —in base all'assegnazione). Nel caso delle variabili categoriali, le statistiche descrittive includono conteggio e percentuale per categoria per i dati di input originali (precedenti all'assegnazione), valori assegnati (dalla

procedura di assegnazione) e dati completi (dati originali e assegnati insieme, —in base all'assegnazione).

**Cronologia iterazioni.** Quando viene utilizzato il metodo di assegnazione della specifica completamente condizionale, è possibile richiedere un insieme di dati contenente i dati sulla cronologia di iterazione per questo tipo di assegnazione. L'insieme di dati contiene le medie e le deviazioni standard in base all'iterazione e all'assegnazione per ogni variabile dipendente di scala per la quale sono stati assegnati i valori. Per valutare la convergenza del modello, è possibile eseguire un grafico dei dati. [Per ulteriori informazioni, vedere l'argomento Verifica della convergenza FCS in il capitolo 5 a pag. 67.](#)

## ***Opzioni aggiuntive del comando MULTIPLE IMPUTATION***

Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare un sottoinsieme di variabili per le quali vengono visualizzate le statistiche descrittive (sottocomando `IMPUTATIONSUMMARIES`).
- Specificare sia un'analisi dei modelli mancanti sia l'assegnazione in un'unica esecuzione della procedura.
- Specificare il numero massimo di parametri del modello consentiti quando si assegna qualsiasi variabile (parola chiave `MAXMODELPARAM` ).

Vedere *Command Syntax Reference* per informazioni dettagliate sulla sintassi.

## ***Utilizzo dei dati ad assegnazione multipla***

Quando viene creato un insieme di dati ad assegnazione multipla, viene aggiunta una variabile chiamata *Imputation\_* con un'etichetta variabile *Numero di assegnazione* e l'insieme di dati viene ordinato in modo crescente. I casi dell'insieme di dati originale hanno un valore di 0. I casi dei valori assegnati sono numerati da 1 a *M*, dove *M* indica il numero delle assegnazioni.

Quando si apre un insieme di dati, la presenza di *Imputation\_* lo identifica come un insieme di dati potenzialmente ad assegnazione multipla.

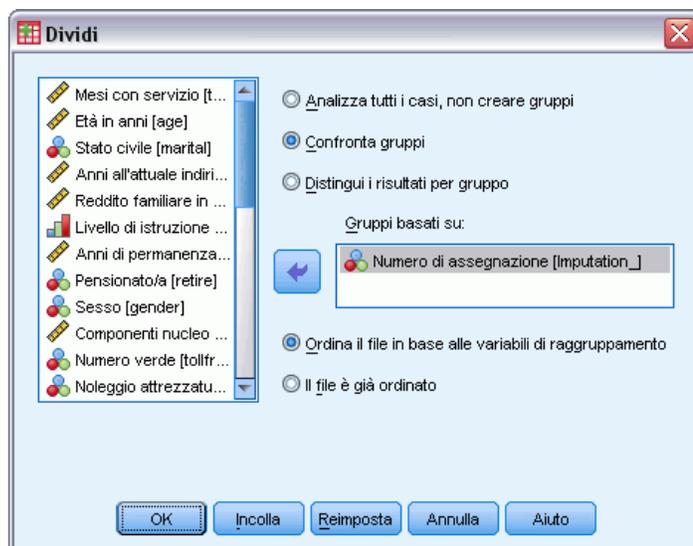
### ***Attivazione di un insieme di dati ad assegnazione multipla per l'analisi***

L'insieme di dati deve essere suddiviso mediante l'opzione *Confronta gruppi*, con *Imputation\_* come variabile di raggruppamento, per potere essere trattato come un insieme di dati ad assegnazione multipla nelle analisi. È inoltre possibile definire le suddivisioni in altre variabili.

Dai menu, scegliere:

Dati > Dividi...

Figura 3-7  
Finestra di dialogo Dividi



- ▶ Selezionare Confronta gruppi.
- ▶ Selezionare *Numero di assegnazione [Imputation\_]* come variabile in base alla quale raggruppare i casi.

In alternativa, quando vengono attivati i contrassegni (vedere sotto) il file viene diviso in base a *Numero di assegnazione [Imputation\_]*.

### ***Distinzione tra valori assegnati e valori osservati***

È possibile distinguere i valori assegnati dai valori osservati in base al colore dello sfondo delle celle, al carattere e all'uso del grassetto (per i valori assegnati). Per informazioni dettagliate sui tipi di contrassegni attivi, vedere [Opzioni Assegnazione multipla](#) a pag. 33. Quando si crea un nuovo insieme di dati nella sessione corrente con Assegna i valori mancanti, i contrassegni vengono attivati per impostazione predefinita. Quando si apre un file di dati salvato che include delle assegnazioni, i contrassegni vengono disattivati.

Figura 3-8  
 Editor dei dati con contrassegni di assegnazione disattivati

	Imputation_	tenure	age	marital	address	ii
1034	1	11	27	1	16	
1035	1	60	46	0	13	
1036	1	20	35	1	2	
1037	1	66	60	0	38	
1038	1	44	57	1	1	
1039	1	11	41	1	0	
1040	1	72	57	0	17	

Per attivare i contrassegni, dai menu dell'Editor dei dati, scegliere:  
 Visualizza > Contrassegna dati assegnati...

Figura 3-9  
 Editor dei dati con contrassegni di assegnazione attivati

	Imputation_	tenure	age	marital	address	income
1034	1	11	27	1	16	55,00
1035	1	60	46	0	13	163,00
1036	1	20	35	1	2	52,00
1037	1	66	60	0	38	211,00
1038	1	44	57	1	1	186,00
1039	1	11	41	1	0	39,00

In alternativa, è possibile attivare i contrassegni facendo clic sul pulsante dei contrassegni di assegnazione sul lato destro della barra di modifica nella Visualizzazione dati dell'Editor dei dati.

### ***Spostamento tra assegnazioni***

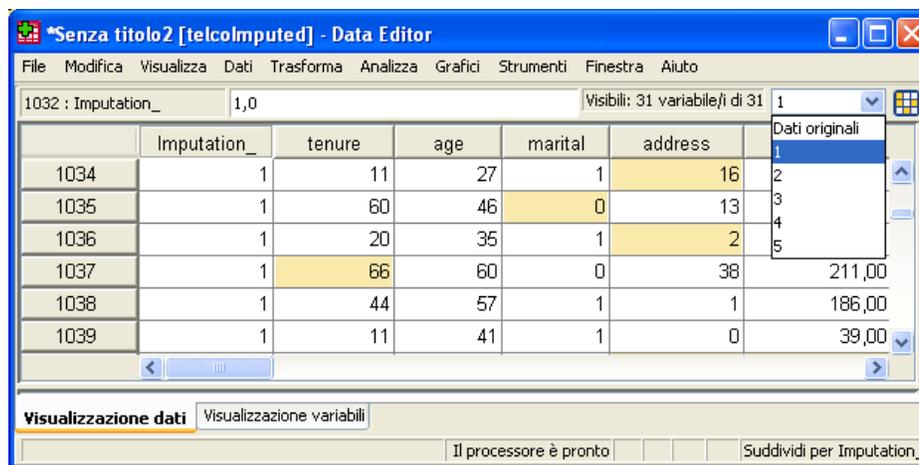
- ▶ Dai menu, scegliere:  
 Modifica > Vai all'assegnazione...
- ▶ Selezionare l'assegnazione (o Dati originali) dall'elenco a discesa.

Figura 3-10  
Finestra di dialogo Vai a



In alternativa, è possibile selezionare l'assegnazione dall'elenco a discesa nella barra di modifica nella Visualizzazione dati dell'Editor dei dati.

Figura 3-11  
Editor dei dati con contrassegni di assegnazione attivati



La posizione relativa dei casi viene mantenuta quando si selezionano le assegnazioni. Ad esempio, se ci sono 1000 casi nell'insieme di dati originale, il caso 1034, il trentaquattresimo nella prima assegnazione, viene visualizzato nella parte superiore della griglia. Se si seleziona l'assegnazione 2 nell'elenco a discesa, il caso 2034, il trentaquattresimo dell'assegnazione 2, viene visualizzato nella parte superiore della griglia. Se si seleziona Dati originali nell'elenco a discesa, il caso 34 viene visualizzato nella parte superiore della griglia. Quando si passa da un'assegnazione all'altra, viene mantenuta anche la posizione delle colonne, in modo da agevolare il confronto dei valori tra le assegnazioni.

### **Trasformazione e modifica dei valori assegnati**

Talvolta sarà necessario eseguire le trasformazioni dei dati assegnati. Ad esempio, potrebbe essere necessario utilizzare il registro di tutti i valori di una variabile stipendio e salvare il risultato in una nuova variabile. Un valore calcolato utilizzando i dati assegnati verrà trattato come assegnato se è diverso dal valore calcolato utilizzando i dati originali.

Se si modifica un valore assegnato in una cella dell'Editor dei dati, questa cella viene comunque trattata come assegnata. Si consiglia di non modificare in questo modo i valori assegnati.

## **Analisi di dati ad assegnazione multipla**

Molte procedure supportano il raggruppamento dei risultati delle analisi di insiemi di dati ad assegnazione multipla. Quando sono attivati i contrassegni di assegnazione, un'icona speciale viene visualizzata accanto alle procedure che supportano il raggruppamento. Nel sottomenu Statistiche descrittive del menu Analizza, ad esempio, Frequenze, Descrittive, Esplora e Tavole di contingenza supportano il raggruppamento, mentre Rapporto, Grafici P-P e Grafici Q-Q non lo supportano.

**Figura 3-12**  
Menu Analizza con contrassegni di assegnazione attivati



L'output tabulare e i file PMML di modelli possono essere raggruppati. Non è stata prevista una nuova procedura per la richiesta di output raggruppato; tuttavia, una nuova scheda nella finestra di dialogo Opzioni fornisce il controllo totale dell'output ad assegnazione multipla.

- **Raggruppamento di output tabulare.** Per impostazione predefinita, quando si esegue una procedura supportata su un insieme di dati ad assegnazione multipla, vengono automaticamente prodotti i risultati per ogni assegnazione, i dati originali (non assegnati) e i risultati raggruppati (finali) che tengono conto della variazione tra assegnazioni. Le statistiche raggruppate variano in base alla procedura.
- **Raggruppamento di PMML.** Le procedure supportate di esportazione PMML consentono di ottenere PMML raggruppati. La richiesta di PMML raggruppati avviene allo stesso modo della richiesta di PMML non raggruppati e viene salvata, a differenza di quest'ultima.

Le procedure non supportate non consentono di produrre output raggruppato né file PMML raggruppati.

### ***Livelli di raggruppamento***

L'output viene raggruppato mediante uno dei due livelli seguenti.

- **Combinazione naïve.** È disponibile solo il parametro che viene raggruppato.
- **Combinazione univariata.** Quando disponibili, vengono visualizzati anche il parametro raggruppato, l'errore standard, le statistiche del test e i gradi di libertà effettivi, il valore  $p$ , l'intervallo di confidenza e la diagnostica di raggruppamento (frazione delle informazioni mancanti, efficienza relativa, incremento relativo della varianza).

Solitamente vengono raggruppati i coefficienti (regressione e correlazione), le medie (e differenza media) e i conteggi. Quando è disponibile l'errore standard della statistica, viene utilizzato il raggruppamento univariato; in caso contrario si utilizza il raggruppamento naïve.

### ***Procedure che supportano il raggruppamento***

Le seguenti procedure supportano gli insiemi di dati ad assegnazione multipla ai livelli di raggruppamento specificati per le singole porzioni di output.

#### **Frequenze**

- La tabella Statistiche supporta le medie al livello di raggruppamento univariato (se è richiesto anche l'errore standard della media) e casi validi e casi mancanti al livello di raggruppamento naïve.
- La tabella Frequenze supporta Frequenza al livello di raggruppamento naïve.

#### **Descrittive**

- La tabella Statistiche descrittive supporta le medie al livello di raggruppamento univariato (se è richiesto anche l'errore standard della media) e N al livello di raggruppamento naïve.

#### **Tavole di contingenza**

- La tabella Tavola di contingenza supporta i conteggi al livello di raggruppamento naïve.

#### **Medie**

- La tabella Rapporti supporta la media al livello di raggruppamento univariato (se è richiesto anche l'errore standard della media) e N al livello di raggruppamento naïve.

**Test T per un campione**

- La tabella Statistiche supporta le medie al livello di raggruppamento univariato e N al livello di raggruppamento naïve.
- La tabella Test supporta la differenza media al livello di raggruppamento univariato.

**T per campioni indipendenti**

- La tabella Statistiche di gruppo supporta le medie al livello di raggruppamento univariato e N al livello di raggruppamento naïve.
- La tabella Test supporta la differenza media al livello di raggruppamento univariato.

**T per campioni appaiati**

- La tabella Statistiche supporta le medie a livello di raggruppamento univariato e N al livello di raggruppamento naïve.
- La tabella Correlazioni supporta le correlazioni e N al livello di raggruppamento naïve.
- La tabella Test supporta la media al livello di raggruppamento univariato.

**ANOVA univariata**

- La tabella Statistiche descrittive supporta la media al livello di raggruppamento univariato e N al livello di raggruppamento naïve.
- La tabella Test di contrasto supporta il valore di contrasto al livello di raggruppamento univariato.

**Modelli misti lineari**

- La tabella Statistiche descrittive supporta la media e N al livello di raggruppamento naïve.
- La tabella Stime degli effetti fissi supporta la stima al livello di raggruppamento univariato.
- La tabella Stime dei parametri di covarianza supporta la stima al livello di raggruppamento univariato.
- La tabella Medie marginali stimate: La tabella Stime supporta la media al livello di raggruppamento univariato.
- La tabella Medie marginali stimate: La tabella Confronti pairwise supporta la differenza media al livello di raggruppamento univariato.

**Modelli lineari generalizzati ed Equazioni di stima generalizzate.** Queste procedure supportano i file PMML raggruppati.

- La tabella Informazioni variabile categoriale supporta N e le percentuali al livello di raggruppamento naïve.
- La tabella Informazioni variabile continua supporta N e la media al livello di raggruppamento naïve.
- La tabella Stime dei parametri supporta il coefficiente, B, al livello di raggruppamento univariato.
- La tabella Medie marginali stimate: La tabella Coefficienti di stima supporta la media al livello di raggruppamento naïve.

- La tabella Medie marginali stimate: La tabella Stime supporta la media al livello di raggruppamento univariato.
- La tabella Medie marginali stimate: La tabella Confronti pairwise supporta la differenza media al livello di raggruppamento univariato.

#### **Correlazioni bivariate**

- La tabella Statistiche descrittive supporta la media e N al livello di raggruppamento naïve.
- La tabella Correlazioni supporta le correlazioni e N al livello di raggruppamento univariato. Si noti che le correlazioni vengono trasformate mediante la trasformazione  $z$  di Fisher prima del raggruppamento e trasformate all'indietro dopo il raggruppamento.

#### **Correlazioni parziali**

- La tabella Statistiche descrittive supporta la media e N al livello di raggruppamento naïve.
- La tabella Correlazioni supporta le correlazioni al livello di raggruppamento naïve.

**Regressione lineare.** Questa procedura supporta i file PMML raggruppati.

- La tabella Statistiche descrittive supporta la media e N al livello di raggruppamento naïve.
- La tabella Correlazioni supporta le correlazioni e N al livello di raggruppamento naïve.
- La tabella Coefficienti supporta B al livello di raggruppamento univariato e le correlazioni al livello di raggruppamento naïve.
- La tabella Coefficienti di correlazione supporta le correlazioni al livello di raggruppamento naïve.
- La tabella Statistiche dei residui supporta la media e N al livello di raggruppamento naïve.

**Regressione logistica binaria.** Questa procedura supporta i file PMML raggruppati.

- La tabella Variabili nell'equazione supporta B al livello di raggruppamento univariato.

**Regressione logistica multinomiale.** Questa procedura supporta i file PMML raggruppati.

- La tabella Stime dei parametri supporta il coefficiente, B, al livello di raggruppamento univariato.

#### **Regressione ordinale**

- La tabella Stime dei parametri supporta il coefficiente, B, al livello di raggruppamento univariato.

**Analisi discriminante.** Questa procedura supporta i file XML del modello raggruppati.

- La tabella Statistiche di gruppo supporta la media e i casi validi al livello di raggruppamento naïve.
- La tabella Matrici entro i gruppi raggruppate supporta le correlazioni al livello di raggruppamento naïve.
- La tabella Coefficienti della funzione discriminante canonica supporta i coefficienti non standardizzati al livello di raggruppamento naïve.

- La tabella Funzioni dei centroidi di gruppo supporta i coefficienti non standardizzati al livello di raggruppamento naïve.
- La tabella Coefficienti della funzione discriminante supporta i coefficienti al livello di raggruppamento naïve.

**Test Chi-quadrato**

- La tabella Descrittive supporta la media e N al livello di raggruppamento naïve.
- La tabella Frequenze supporta osservate N al livello di raggruppamento naïve.

**Test binomiale**

- La tabella Descrittive supporta le medie e N al livello di raggruppamento naïve.
- La tabella Test supporta N, la proporzione osservata e la proporzione attesa al livello di raggruppamento naïve.

**Test delle successioni**

- La tabella Descrittive supporta le medie e N al livello di raggruppamento naïve.

**Test di Kolmogorov-Smirnov per un campione**

- La tabella Descrittive supporta le medie e N al livello di raggruppamento naïve.

**Test per due campioni indipendenti**

- La tabella Ranghi supporta il rango medio e N al livello di raggruppamento naïve.
- La tabella Frequenze supporta N al livello di raggruppamento naïve.

**Test per diversi campioni indipendenti**

- La tabella Ranghi supporta il rango medio e N al livello di raggruppamento naïve.
- La tabella Frequenze supporta i conteggi al livello di raggruppamento naïve.

**Test per due campioni dipendenti**

- La tabella Ranghi supporta il rango medio e N al livello di raggruppamento naïve.
- La tabella Frequenze supporta N al livello di raggruppamento naïve.

**Test per diversi campioni dipendenti**

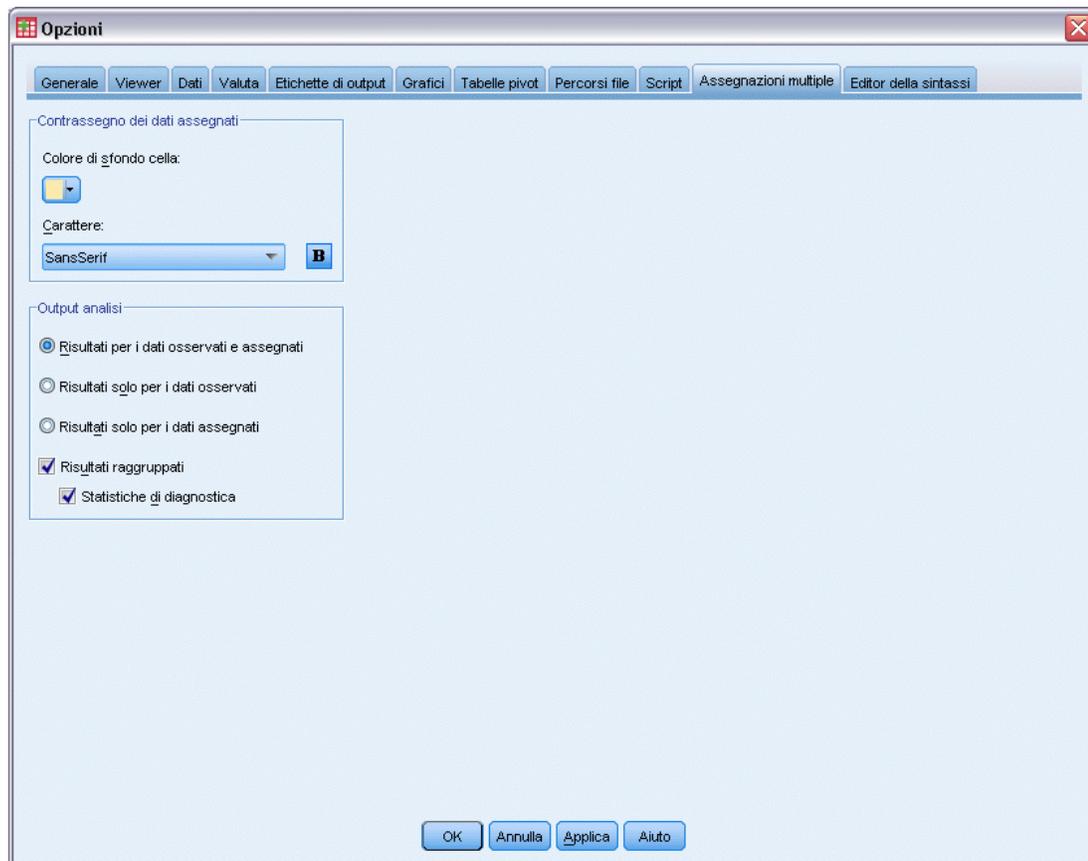
- La tabella Ranghi supporta il rango medio al livello di raggruppamento naïve.

**Regressione di Cox.** Questa procedura supporta i file PMML raggruppati.

- La tabella Variabili nell'equazione supporta B al livello di raggruppamento univariato.
- La tabella Medie covariate supporta la media al livello di raggruppamento naïve.

## Opzioni Assegnazione multipla

Figura 3-13  
Finestra di dialogo Opzioni: Scheda Assegnazioni multiple



La scheda Assegnazioni multiple consente di controllare due tipi di preferenze relative alle assegnazioni multiple:

**Aspetto dei dati assegnati.** Per impostazione predefinita, le celle che contengono i dati assegnati hanno uno sfondo di colore diverso rispetto alle celle contenenti dati non assegnati. L'aspetto facilmente distinguibile dei dati assegnati dovrebbe facilitare la consultazione di un insieme di dati e l'individuazione di tali celle. È possibile cambiare il colore predefinito dello sfondo della cella e il carattere, nonché visualizzare i dati assegnati in grassetto.

**Output analisi.** Questo gruppo controlla il tipo di output del Viewer prodotto ogni volta che viene analizzato un insieme di dati ad assegnazione multipla. Per impostazione predefinita, verrà prodotto output per l'insieme di dati originale (precedente all'assegnazione) e per ogni insieme di dati assegnato. Inoltre, nel caso delle procedure che supportano il raggruppamento in pool dei dati assegnati, verranno generati risultati raggruppati finali. Quando viene eseguito il raggruppamento univariato, viene visualizzata anche la diagnostica del raggruppamento. Tuttavia, è possibile sopprimere qualsiasi output che non si desidera produrre.

***Per specificare le impostazioni di assegnazione multipla***

Dai menu, scegliere:

Modifica > Opzioni

Fare clic sulla scheda Assegnazione multipla.

# ***Parte II: Esempi***

# Analisi dei valori mancanti

## Descrizione del modello dei dati mancanti

Un provider di telecomunicazioni desidera approfondire la conoscenza dei modelli di utilizzo dei servizi nel proprio database di clienti. La società vuole garantire che i dati sono mancanti in modo del tutto casuale prima di eseguire ulteriori analisi.

Un campione casuale dal database dei clienti è contenuto nel file *telco\_missing.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in *IBM SPSS Missing Values 20*.

## Esecuzione dell'analisi per visualizzare le statistiche descrittive

- Per eseguire l'analisi dei valori mancanti, dai menu scegliere:  
Analizza > Analisi dei valori mancanti...

Figura 4-1  
Finestra di dialogo Analisi dei valori mancanti



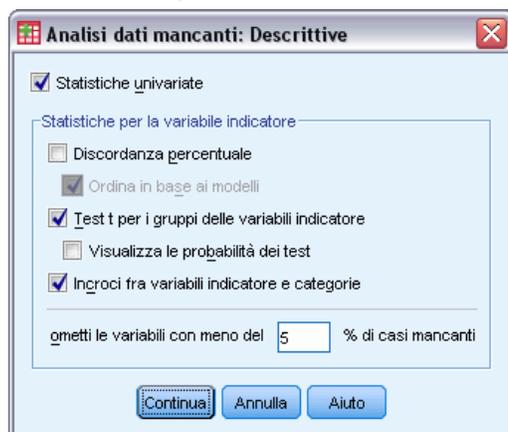
- Selezionare *Stato civile [stato civ]*, *Titolo di studio [istruz]*, *Pensionato [pensione]* e *Sesso [sesso]* come variabili categoriali.
- Selezionare da *Mesi con servizio [durata]* a *Componenti nucleo familiare [resid]* come variabili quantitative (di scala).

A questo punto, è possibile eseguire la procedura e ottenere statistiche univariate, ma ora si selezioneranno statistiche descrittive aggiuntive.

- Fare clic su Descrittive.

Figura 4-2

Finestra di dialogo Analisi dei valori mancanti: Descrittive



Nella finestra di dialogo Descrittive è possibile specificare varie statistiche descrittive da visualizzare nell'output. Le statistiche univariate predefinite possono aiutare a determinare la misura generale dei dati mancanti, ma le statistiche per la variabile indicatore offrono maggiori informazioni su come il modello dei dati mancanti in una variabile può influenzare i valori di un'altra variabile.

- Selezionare Test t per i gruppi delle variabili indicatore.
- Selezionare Incroci fra variabili indicatore e categorie.
- Fare clic su Continua.
- Nella finestra di dialogo principale Analisi valori mancanti, fare clic su OK.

## ***Valutazione delle statistiche descrittive***

Per questo esempio, l'output include:

- Statistiche univariate
- Tabella dei test  $t$  di varianza separati, incluse le medie dei sottogruppi quando un'altra variabile è presente o mancante
- Tabelle per ogni variabile categoriale che mostra le frequenze dei dati mancanti per ogni categoria in base alla variabile quantitativa (scala)

Figura 4-3  
Tabella Statistiche univariate

	N	Media	Deviazione std.	Mancante		N. di estremi <sup>a</sup>	
				Conteggio	Percentuale	Min	Max
MesiDiServizio	968	35,56	21,268	32	3,2	0	0
Età	975	41,75	12,573	25	2,5	0	0
AnniAlnirizzo	850	11,47	9,965	150	15,0	0	9
Reddito	821	71,1462	83,14424	179	17,9	0	71
AnniConDatore	904	11,00	10,113	96	9,6	0	15
NumNucleoFamigliare	966	2,32	1,431	34	3,4	0	33
StatoCivile	885			115	11,5		
Istruzione	965			35	3,5		
Pensionato	916			84	8,4		
Sesso	958			42	4,2		

a. Numero di casi non compresi nell'intervallo (Q1 - 1.5\*IQR, Q3 + 1.5\*IQR).

Le statistiche univariate danno un'indicazione, variabile per variabile, sulla misura dei dati mancanti. Il numero di valori non mancanti per ogni variabile è riportato nella colonna *N* e il numero di valori mancanti nella colonna del conteggio mancante. La colonna relativa alla *percentuale di valori mancanti* mostra la percentuale dei casi con valori mancanti e fornisce una buona indicazione per confrontare la misura dei dati mancanti tra le variabili. *reddito (Reddito familiare in migliaia)* ha il maggior numero di casi con valori mancanti (17,9%), mentre *età (Età in anni)* ha il minor numero di casi mancanti (2,5%). *reddito* ha inoltre il maggior numero di valori estremi.

Figura 4-4  
Tabella dei test *t* di varianza separati

	MesIDiservizio	Età	AnniAnirizzo	Reddito	AnniConDatore	NumNucleoFamiliare
<b>AnniAnirizzo</b>						
t	.4	.3	.	3,5	1,4	1,0
df	202,2	192,5	.	313,6	191,1	199,5
# Presente	819	832	850	693	766	824
# Mancante	149	143	0	128	138	142
Media(Presente)	35,68	41,79	11,47	74,0779	11,20	2,34
Media(Mancante)	34,91	41,49	.	55,2734	9,86	2,21
<b>Reddito</b>						
t	-5,0	-8,3	-3,9	.	-5,9	3,6
df	249,5	222,8	191,1	.	203,3	315,2
# Presente	793	801	693	821	741	792
# Mancante	175	174	157	0	163	174
Media(Presente)	33,93	40,01	10,67	71,1462	9,91	2,39
Media(Mancante)	42,97	49,73	14,97	.	15,93	2,02
<b>AnniConDatore</b>						
t	-1,0	-.4	-.7	.5	.	-.3
df	110,5	110,2	97,6	114,9	.	110,9
# Presente	877	881	766	741	904	874
# Mancante	91	94	84	80	0	92
Media(Presente)	35,34	41,69	11,37	71,4953	11,00	2,31
Media(Mancante)	37,70	42,27	12,32	67,9125	.	2,37
<b>StatoCivile</b>						
t	.0	1,8	1,2	-.8	.9	-2,2
df	148,1	149,5	138,8	121,2	128,3	134,2
# Presente	856	862	748	728	805	857
# Mancante	112	113	102	93	99	109
Media(Presente)	35,56	42,00	11,61	70,3887	11,10	2,28
Media(Mancante)	35,57	39,85	10,43	77,0753	10,17	2,61
<b>Pensionato</b>						
t	-.6	-.4	-.4	.3	..	.2
df	95,4	94,4	84,0	93,2	..	99,0
# Presente	888	893	777	751	904	885
# Mancante	80	82	73	70	0	81
Media(Presente)	35,44	41,70	11,42	71,3356	11,00	2,32
Media(Mancante)	36,89	42,29	11,96	69,1143	.	2,30

La tabella dei test *t* di varianza separati consente di identificare le variabili il cui modello di valori mancanti può influenzare le variabili quantitative (scala). Il test *t* viene calcolato utilizzando una variabile indicatore che specifica se una variabile è presente o mancante relativamente a un singolo caso. Nella tabella vengono inoltre riportate le medie del sottogruppo per la variabile indicatore. Tenere presente che una variabile indicatore viene creata solo se una variabile presenta valori mancanti in almeno il 5% dei casi.

Sembra che rispondenti più anziani siano meno propensi a indicare i livelli di reddito. Quando manca il valore di *reddito*, la media *età* è 49,73, rispetto a 40,01 quando il valore di *reddito* non è mancante. Infatti, la mancanza del valore *reddito* sembra influenzare le medie di diverse variabili quantitative (scala). Questa è un'indicazione che i dati potrebbero non essere mancanti completamente in modo casuale.

Figura 4-5  
Tavola di contingenza per Stato civile [statociv]

			Totale	Non sposato	Sposato	Mancanti
						SysMis
AnniAnirizzo	Presente	Conteggio	850	390	358	102
		Percentuale	85,0	85,5	83,4	88,7
	Mancanti	% SysMis	15,0	14,5	16,6	11,3
Reddito	Presente	Conteggio	821	380	348	93
		Percentuale	82,1	83,3	81,1	80,9
	Mancanti	% SysMis	17,9	16,7	18,9	19,1
AnniConDatore	Presente	Conteggio	904	418	387	99
		Percentuale	90,4	91,7	90,2	86,1
	Mancanti	% SysMis	9,6	8,3	9,8	13,9
Pensionato	Presente	Conteggio	916	423	392	101
		Percentuale	91,6	92,8	91,4	87,8
	Mancanti	% SysMis	8,4	7,2	8,6	12,2

Le tavole di contingenza delle variabili categoriali rispetto alle variabili indicatore mostrano informazioni simili a quelle della tabella dei test  $t$  della varianza separati. Le variabili indicatore vengono nuovamente create, ma in questo caso vengono utilizzate per calcolare le frequenze in ogni categoria di ciascuna variabile categoriale. I valori consentono di determinare se vi sono differenze nei valori mancanti tra le categorie.

Esaminando la tabella per *statociv* (*Stato civile*), il numero di valori mancanti nelle variabili indicatore sembra non variare molto tra le categorie *statociv*. Il fatto che qualcuno sia sposato o meno sembra non influenzare la mancanza di dati delle variabili quantitative (scala). Ad esempio, le persone non sposate hanno indicato *indirizzo* (*Anni all'attuale indirizzo*) nell'85,5% delle volte e le persone sposate hanno indicato la stessa variabile nell'83,4% delle volte. La differenza è minima e probabilmente dovuta al caso.

Figura 4-6  
Tavola di contingenza per Titolo di studio [istruz.]

			Totale	Non ha completato le scuole superiori	Diploma di scuola superiore	Alcuni anni all'università	Diploma universitario	Laurea specialistica	Mancanti SysMis
AnniAlnirizzo	Presente	Conteggio	850	163	240	175	186	56	30
		Percentuale	85,0	83,2	85,7	88,4	81,9	87,5	85,7
	Mancanti	% SysMis	15,0	16,8	14,3	11,6	18,1	12,5	14,3
Reddito	Presente	Conteggio	821	155	229	165	193	50	29
		Percentuale	82,1	79,1	81,8	83,3	85,0	78,1	82,9
	Mancanti	% SysMis	17,9	20,9	18,2	16,7	15,0	21,9	17,1
AnniConDatore	Presente	Conteggio	904	178	254	178	204	60	30
		Percentuale	90,4	90,8	90,7	89,9	89,9	93,8	85,7
	Mancanti	% SysMis	9,6	9,2	9,3	10,1	10,1	6,2	14,3
StatoCivile	Presente	Conteggio	885	193	278	148	184	52	30
		Percentuale	88,5	98,5	99,3	74,7	81,1	81,2	85,7
	Mancanti	% SysMis	11,5	1,5	,7	25,3	18,9	18,8	14,3
Pensionato	Presente	Conteggio	916	180	259	180	207	60	30
		Percentuale	91,6	91,8	92,5	90,9	91,2	93,8	85,7
	Mancanti	% SysMis	8,4	8,2	7,5	9,1	8,8	6,2	14,3

Considerare ora la tavola di contingenza per *istruz.* (*Titolo di studio*). Se un rispondente ha avuto perlomeno un'istruzione universitaria, è più probabile che non risponda relativamente allo stato civile. Almeno il 98,5% dei rispondenti senza un'istruzione universitaria ha indicato lo stato civile. D'altra parte, solo l'81,1% delle persone con una laurea ha indicato lo stato civile. Il numero è anche inferiore per le persone con un'istruzione universitaria ma senza laurea.

Figura 4-7  
Tavola di contingenza per Pensionato [pensione]

			Totale	No	00 "S"	Mancanti
						SysMis
AnniAnirizzo	Presente	Conteggio	850	744	33	73
		Percentuale	85,0	85,0	80,5	86,9
	Mancanti	% SysMis	15,0	15,0	19,5	13,1
Reddito	Presente	Conteggio	821	732	19	70
		Percentuale	82,1	83,7	46,3	83,3
	Mancanti	% SysMis	17,9	16,3	53,7	16,7
AnniConDatore	Presente	Conteggio	904	864	40	0
		Percentuale	90,4	98,7	97,6	,0
	Mancanti	% SysMis	9,6	1,3	2,4	100,0
StatoCivile	Presente	Conteggio	885	777	38	70
		Percentuale	88,5	88,8	92,7	83,3
	Mancanti	% SysMis	11,5	11,2	7,3	16,7

Una differenza ancora più evidente può essere vista in *pensione (Pensionato)*. Le persone in pensione sono molto meno propensi a indicare il proprio reddito rispetto alle persone che ancora non sono in pensione. Solo il 46,3% dei pensionati ha indicato il livello di reddito, mentre la percentuale delle persone non in pensione che ha indicato il reddito è stata dell'83,7.

Figura 4-8  
Tavola di contingenza per Sesso [sesso]

			Totale	Maschio	Femmina	Mancanti
						SysMis
AnniAnirizzo	Presente	Conteggio	850	363	456	31
		Percentuale	85,0	78,6	91,9	73,8
	Mancanti	% SysMis	15,0	21,4	8,1	26,2
Reddito	Presente	Conteggio	821	381	406	34
		Percentuale	82,1	82,5	81,9	81,0
	Mancanti	% SysMis	17,9	17,5	18,1	19,0
AnniConDatore	Presente	Conteggio	904	412	457	35
		Percentuale	90,4	89,2	92,1	83,3
	Mancanti	% SysMis	9,6	10,8	7,9	16,7
StatoCivile	Presente	Conteggio	885	400	445	40
		Percentuale	88,5	86,6	89,7	95,2
	Mancanti	% SysMis	11,5	13,4	10,3	4,8
Pensionato	Presente	Conteggio	916	420	461	35
		Percentuale	91,6	90,9	92,9	83,3
	Mancanti	% SysMis	8,4	9,1	7,1	16,7

Un'altra discrepanza è evidente per *sesso (Sesso)*. Le informazioni sull'indirizzo sono mancanti più spesso nelle persone di sesso maschile rispetto a quelle di sesso femminile. Sebbene queste discrepanze possano essere dovute al caso, sembra tuttavia improbabile. I dati non sembrano essere mancanti completamente in modo casuale.

Verranno esaminati i modelli dei dati mancanti per analizzare ulteriormente questo caso.

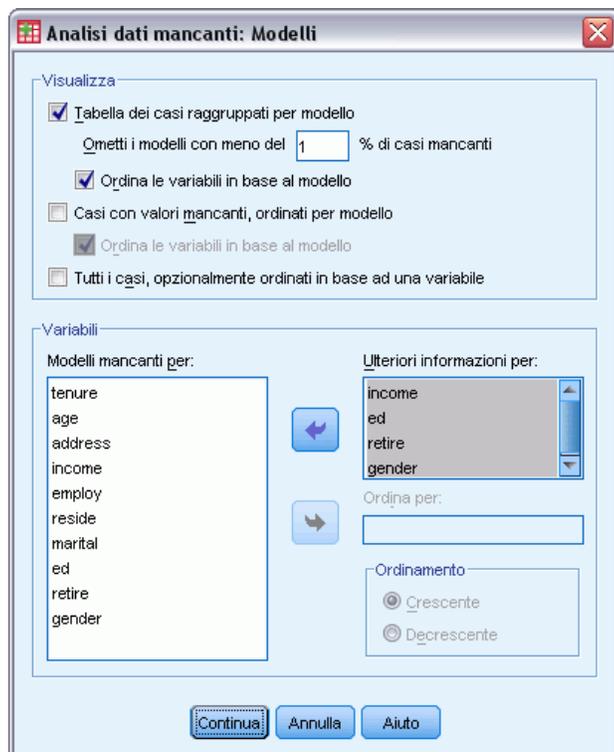
## Riesecuzione dell'analisi per visualizzare i modelli

Figura 4-9  
Finestra di dialogo Analisi dei valori mancanti



- Richiamare la finestra di dialogo Analisi dei valori mancanti. La finestra di dialogo indica le variabili utilizzate nella precedente analisi. Non modificarle.
- Fare clic su Modelli.

Figura 4-10  
Finestra di dialogo Analisi dei valori mancanti: Modelli



Nella finestra di dialogo Modelli è possibile selezionare varie tabelle di modelli. Verranno ora visualizzati i modelli inseriti in tabella raggruppati per modelli mancanti. Poiché i modelli mancanti in *istruz.* (*Titolo di studio*), *pensione* (*Pensionato*) e  *Sesso* (*Sesso*) sembravano influenzare i dati, verranno visualizzate informazioni aggiuntive per tali variabili. Verranno inoltre incluse informazioni aggiuntive per *reddito* (*Reddito familiare in migliaia*) a causa del relativo numero elevato di valori mancanti.

- ▶ Selezionare Tabella dei casi raggruppati per modello.
- ▶ Selezionare *reddito*, *istruz.*, *pensione* e  *Sesso*, quindi aggiungere tali valori all'elenco Ulteriori informazioni per.
- ▶ Fare clic su Continua.
- ▶ Nella finestra di dialogo principale Analisi valori mancanti, fare clic su OK.

## Valutazione della tabella dei modelli

Figura 4-11  
Tabella Modelli in tabella

Numero di assegnazione	Numero di casi	Modelli di valori mancanti <sup>a</sup>										Completo se ... <sup>b</sup>	Reddito	Istruzione <sup>d</sup>				Pensionato <sup>d</sup>		Sesso <sup>d</sup>		
		Imputation_	Età	NumNucleoFamiliare	MesiDiServizio	Istruzione	Sesso	Pensionato	AnniConDatore	StatoCivile	AnniInIrrizzo			Reddito	Non ha completato le scuole superiori	Diploma di scuola superiore	Alcuni anni all'università	Diploma universitario	Laurea specialistica	No	00 "Sp"	Maschio
Dati originali	475											475	76,5853	99	157	87	101	31	463	12	201	274
	109									X	X	584	.	27	35	19	17	11	95	14	47	62
	16											687	.	5	9	0	1	1	12	4	12	4
	87			X								562	54,4368	21	27	9	24	6	85	2	86	21
	13											488	56,0000	4	3	2	3	1	13	0	4	9
	60				X				X			535	77,2167	1	2	27	24	6	59	1	35	25
	16											491	47,8125	0	0	0	0	0	16	0	6	10
	17				X							492	76,2353	2	7	3	4	1	17	0	7	10
	18					X						493	54,1111	3	7	4	4	0	17	1	0	0
	16							X				660	.	0	0	7	8	1	14	2	6	10
	37						X	X		X		520	59,4595	9	14	5	8	1	0	0	15	22

I modelli con meno del 1% di casi (10 o meno) non sono visualizzati.

a. Le variabili sono ordinate per modelli dei valori mancanti.

b. Numero di casi completi se le variabili mancanti in quel modello (contrassegnato con X) non sono utilizzate.

d. Distribuzione della frequenza in ogni modello univoco

La tabella dei modelli in tabella mostra, in singoli casi, se i dati tendono a essere mancanti per più variabili. Ovvero, può essere di aiuto a determinare se i dati sono mancanti.

Vi sono tre modelli di dati mancanti in oltre l'1% dei casi. Le variabili *impiego* (*Anni di permanenza nell'impiego attuale*) e *pensione* (*Pensionato*) insieme mancano più spesso rispetto alle altre coppie di variabili. Questo fatto non sorprende poiché *pensione* e *impiego* indicano informazioni simili. Se non si è a conoscenza se un rispondente sia in pensione, probabilmente non si conoscerà neanche il numero di anni di permanenza nell'impiego corrente del rispondente.

La media *reddito* (*Reddito familiare in migliaia*) sembra variare notevolmente in base al modello di dati mancanti. In particolare, la media *Reddito* è molto superiore al 6% (60 su 1000) dei casi, quando manca il valore di *statociv* (*Stato civile*). (È superiore anche quando manca il valore di *durata* (*Mesi di servizio*), ma questo modello è valido solo nell'1,7% dei casi). Tenere presente che le persone con un livello di istruzione più alto, sono meno propense a rispondere alla domanda relativa allo stato civile. È possibile vedere questo trend nelle frequenze mostrato per *istruz.* (*Titolo di studio*). È possibile tenere conto dell'aumento in *reddito* presumendo che le persone con un livello di istruzione più alto guadagnino più soldi e siano meno propense a indicare lo stato civile.

Considerando le statistiche descrittive e i modelli dei dati mancanti, è possibile concludere che i dati non sono completamente mancanti in modo casuale. È possibile confermare questa conclusione mediante un test MCAR di Little, che riporta le stime EM.

## Riesecuzione dell'analisi per il test MCAR di Little

Figura 4-12  
Finestra di dialogo Analisi dei valori mancanti



- ▶ Richiamare la finestra di dialogo Analisi dei valori mancanti.
- ▶ Fare clic su EM.
- ▶ Fare clic su OK.

Figura 4-13  
Tabelle delle medie marginali

MesiDiServizio	Età	AnniAllInizio	Reddito	AnniConDatore	NumNucleoFamigliare
36,12	41,91	11,58	77,3941	11,22	2,29

a. Test MCAR di Little: Chi-quadrato = 179,836, DF = 107, Sig. = ,000

I risultati del test MCAR di Little sono riportati nelle annotazioni di ogni tabella delle stime EM. L'ipotesi nulla del test MCAR di Little indica che i dati sono completamente mancanti in modo casuale (MCAR). I dati sono definiti MCAR quando il modello dei valori mancanti non dipende dai valori dei dati. Poiché nell'esempio il valore di significatività è inferiore a 0,05, è possibile concludere che i dati *non* sono completamente mancanti in modo casuale. Questo conferma la conclusione tratta per le statistiche descrittive e per i modelli in tabella.

A questo punto, poiché i dati non sono completamente mancanti in modo casuale, non è sicuro effettuare l'eliminazione listwise dei casi con valori mancanti o assegnare singoli valori mancanti. Tuttavia, è possibile utilizzare l'[assegnazione multipla](#) per approfondire l'analisi di questo insieme di dati.

# ***Assegnazione multipla***

## ***Utilizzo dell'assegnazione multipla per completare e analizzare un insieme di dati***

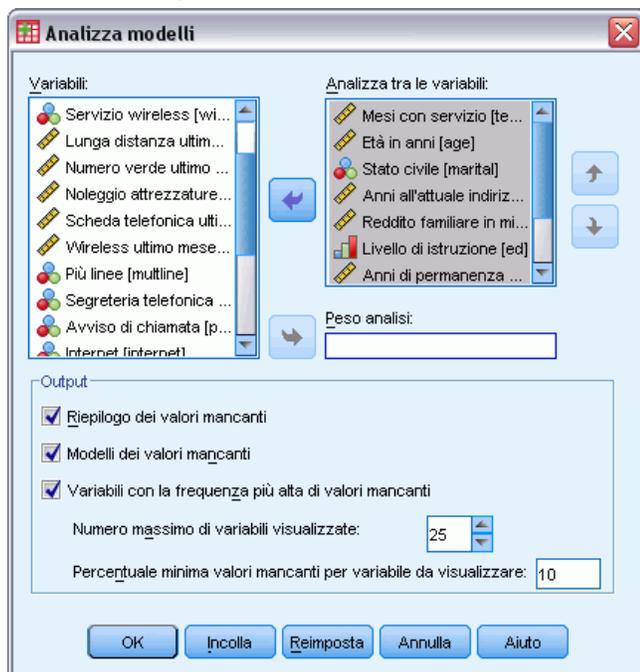
Un provider di telecomunicazioni desidera approfondire la conoscenza dei modelli di utilizzo dei servizi nel proprio database di clienti. Questo provider possiede dati completi circa i servizi utilizzati dai propri clienti, ma le informazioni di carattere demografico raccolte dall'azienda presentano alcuni valori mancanti. Inoltre, tali valori non sono mancanti in modo interamente casuale, pertanto verrà utilizzata l'assegnazione multipla per completare l'insieme di dati.

Un campione casuale dal database dei clienti è contenuto nel file *telco\_missing.sav*. [Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A in IBM SPSS Missing Values 20.](#)

### ***Analisi dei modelli dei valori mancanti***

- ▶ Come prima fase, osservare i modelli dei dati mancanti. Dai menu, scegliere:  
Analizza > Assegnazione multipla > Analizza modelli...

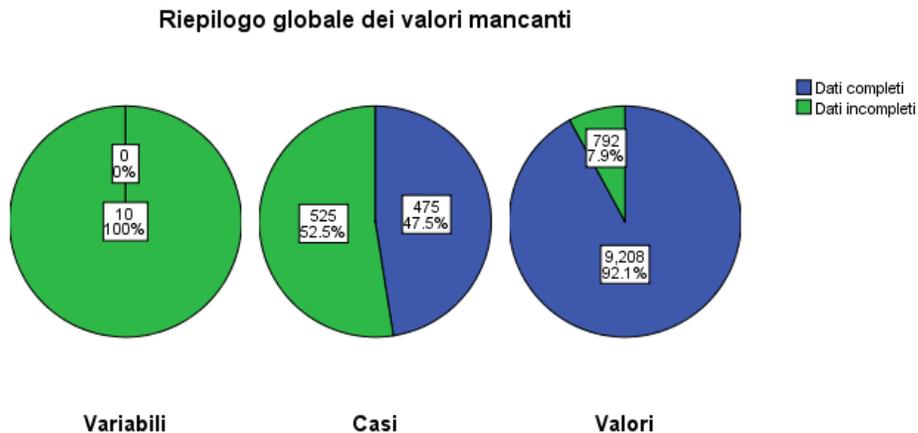
Figura 5-1  
Finestra di dialogo *Analizza modelli*



- Selezionare da *Mesi di servizio [durata]* fino a *Componenti nucleo familiare [nfam]* come variabili di analisi.

### Riepilogo generale

Figura 5-2  
*Riepilogo generale dei valori mancanti*



Il riepilogo generale dei valori mancanti visualizza tre grafici a torta che mostrano i diversi aspetti dei valori mancanti nei dati.

- Il grafico *Variabili* indica che ciascuna delle 10 variabili di analisi ha almeno un valore mancante su un caso.
- Il grafico *Casi* indica che 525 dei 1000 casi ha almeno un valore mancante su una variabile.
- Il grafico *Valori* indica che 792 dei 10.000 valori (casi × variabili) sono mancanti.

Ciascun caso con valori mancanti ha, in media, dei valori mancanti su circa 1,5 delle 10 variabili. Questo suggerisce che **eliminazione listwise** perderà molte delle informazioni presenti nell'insieme di dati.

### **Riepilogo variabili**

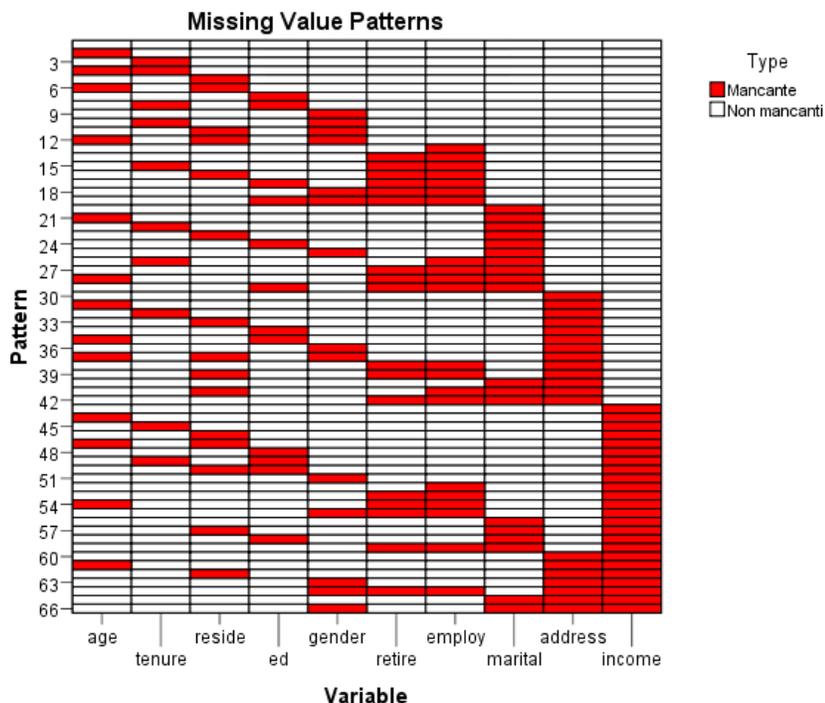
Figura 5-3  
*Riepilogo variabili*

	Mancante		N valido	Media	Deviazione standard Variabile
	N	Percentuale			
Reddito familiare in ...	179	17,9%	821	71,1462	83,14424
Anni all'attuale indirizzo	150	15,0%	850	11,47	9,965
Stato civile	115	11,5%	885		

Il riepilogo variabili viene visualizzato per le variabili con almeno il 10% di valori mancanti e mostra il numero e la percentuali di valori mancanti per ciascuna variabile nella tabella. Visualizza anche la media e la deviazione standard per i valori validi delle variabili di scala e il numero dei valori validi per tutte le variabili. *Reddito familiare in migliaia (reddito)*, *Anni all'attuale indirizzo* e *Stato civile* presentano il maggior numero di valori mancanti, in questo ordine.

**Motivi**

Figura 5-4  
Motivi valori mancanti

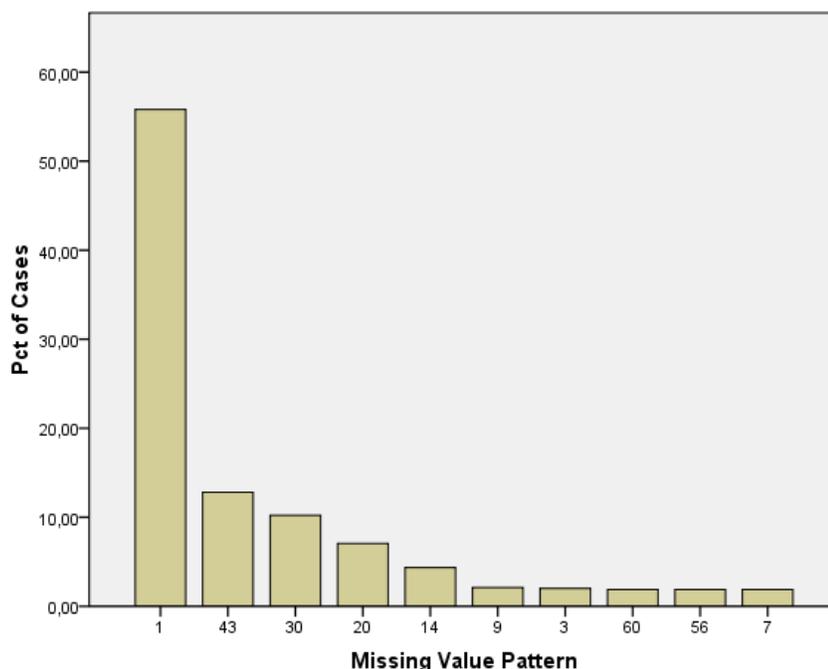


Il grafico dei modelli visualizza i motivi dei valori mancanti per le variabili di analisi. Ogni modello corrisponde a un gruppo di casi con lo stesso modello di dati incompleti e completi. Ad esempio, Modello 1 rappresenta dei casi che non hanno valori mancanti, mentre Modello 33 rappresenta dei casi che hanno valori mancanti in *nfma* (*Componenti nucleo familiare*) e *indirizzo* (*Anni all'attuale indirizzo*) e Modello 66 rappresenta dei casi che hanno dei valori mancanti su  *sesso* (*Sesso*), *stato* (*Stato civile*), *indirizzo* e *reddito* (*Reddito familiare in migliaia*). Un insieme di dati potenzialmente può avere 2 modelli numero di variabili. Per 10 variabili di analisi avremo  $2^{10}=1024$ ; tuttavia, solo 66 modelli sono rappresentati nei 1000 casi dell'insieme di dati.

Il grafico ordina le variabili di analisi e i modelli per far emergere la monotonicità, dove presente. Nello specifico, le variabili vengono ordinate da sinistra a destra in ordine crescente di valori mancanti. I modelli vengono quindi ordinati prima in base all'ultima variabile (per primi i valori non-mancanti, quindi i valori mancanti), successivamente in base alla penultima variabile e così via, andando da destra a sinistra. In questo modo si determina se il metodo monotono di assegnazione può essere utilizzato con i propri dati o, in caso contrario, in che misura tali dati si avvicinano a un modello monotono. Se i dati sono monotoni, tutte le celle mancanti e le celle non mancanti nel grafico saranno contigue, ovvero non ci saranno "isole" di celle non mancanti nella parte inferiore destra del grafico, così come non ci saranno "isole" di celle mancanti nella parte superiore sinistra del grafico.

Questo insieme di dati è non monotono e sono presenti molti dati che dovrebbero essere assegnati per ottenere la monotonicità.

Figura 5-5  
Frequenze dei motivi



Quando sono richiesti i motivi, un grafico a barre visualizza la percentuale di casi per ciascun motivo. In questo caso più di metà dei casi nell'insieme di dati hanno il Modello 1 e il grafico motivi valori mancanti mostra che questo è il motivo per i casi con valori non mancanti. Il Modello 43 rappresenta i casi con un valore mancante in *reddito*, il Modello 30 rappresenta i casi con un valore mancante in *indirizzo* e il Modello 20 rappresenta i casi con un valore mancante in *stato civ.* La maggior parte dei casi, circa 4 su 5, è rappresentata da questi quattro motivi. I Modelli 14, 60 e 56 sono gli unici modelli tra i dieci modelli che ricorrono più di frequente per rappresentare i casi con valori mancanti in più di una variabile.

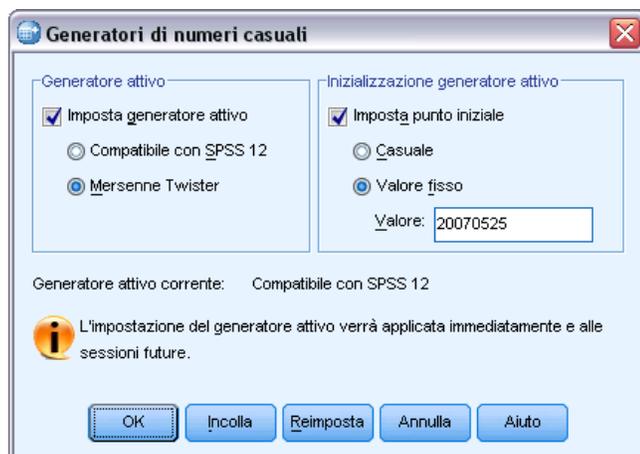
L'analisi dei modelli mancanti non ha rivelato ostacoli particolari all'assegnazione multipla, ad eccezione del fatto che l'uso del metodo monotono non sarà realmente fattibile.

### **Assegnazione automatica dei valori mancanti**

Ora si è pronti a iniziare l'assegnazione di valori. Inizieremo con un'esecuzione con impostazioni automatiche, ma prima di richiedere le assegnazioni, imposteremo un seme casuale. L'impostazione del seme casuale consente di replicare esattamente l'analisi.

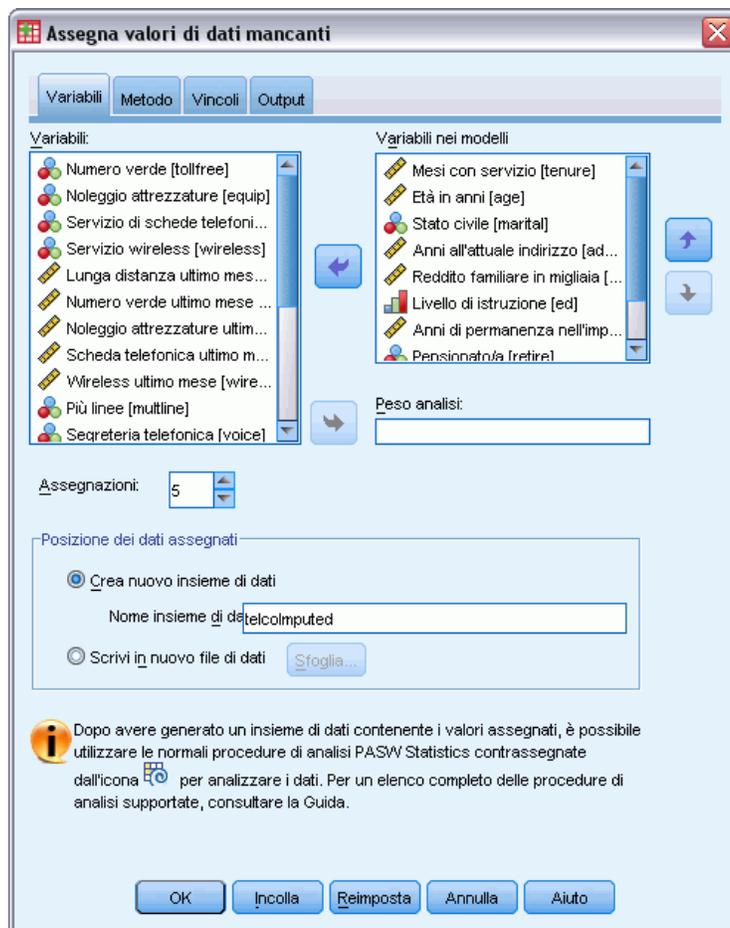
- Per impostare il seme casuale, dai menu scegliere:  
Trasforma > Generatori numeri casuali...

Figura 5-6  
Finestra di dialogo Generatori di numeri casuali



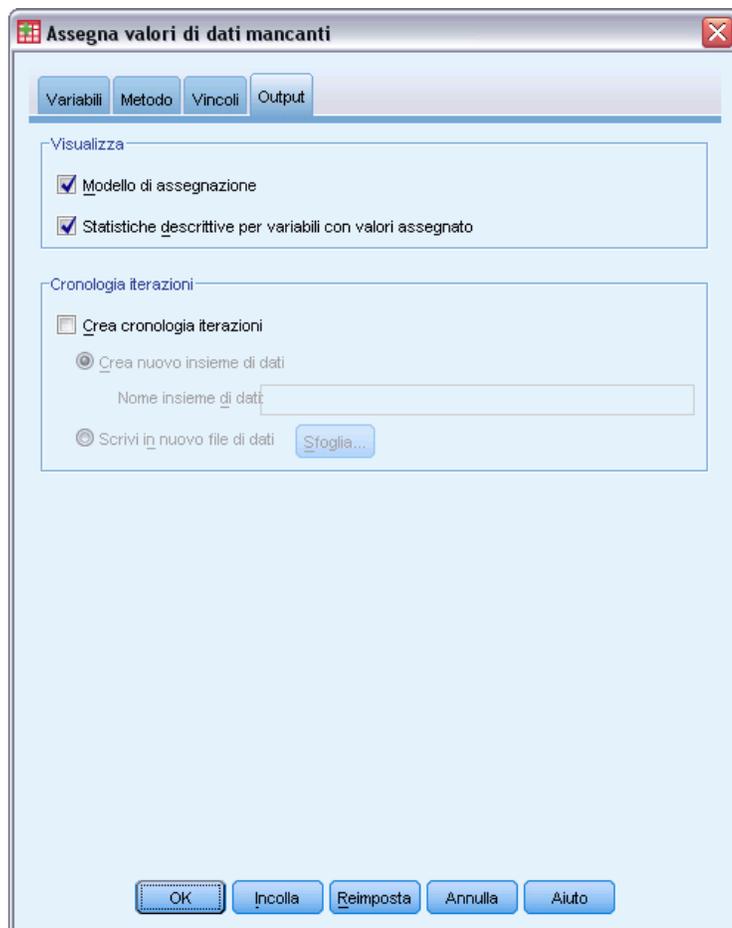
- ▶ Selezionare Imposta generatore attivo.
- ▶ Selezionare Mersenne Twister.
- ▶ Selezionare Imposta punto iniziale.
- ▶ Selezionare Valore fisso e digitare il valore 20070525.
- ▶ Fare clic su OK.
- ▶ Per l'assegnazione multipla dei valori di dati mancanti, dai menu scegliere:  
Analizza > Assegnazione multipla > Assegna valori di dati mancanti...

Figura 5-7  
Finestra di dialogo *Assegna valori di dati mancanti*



- ▶ Selezionare da *Mesi di servizio [durata]* fino a *Componenti nucleo familiare [nfam]* come variabili nel modello di assegnazione.
- ▶ Digitare *telcolcomputed* come l'insieme di dati in cui salvare i dati assegnati.
- ▶ Fare clic sulla scheda *Output*.

Figura 5-8  
scheda Output



- ▶ Selezionare Statistiche descrittive per variabili con valori assegnati.
- ▶ Fare clic su OK.

### **Specifiche di assegnazione**

Figura 5-9  
Specifiche di assegnazione

Metodo di assegnazione	Automatico	
Numero di assegnazioni		5
Modello per le variabili ...	Regressione lineare	
Interazioni incluse nei ...	(nessuno)	
Percentuale massima ...		100,0%

La tabella Specifiche di assegnazione rappresenta un'utile revisione degli elementi richiesti in modo che sia possibile confermare la correttezza delle specifiche.

### Risultati di assegnazione

Figura 5-10  
Risultati di assegnazione

Metodo di assegnazione	Specifica interamente condizionale	
Iterazioni del metodo di specifica interamente condizionale		10
Variabili dipendenti	Assegnato	tenure,age,marital,address,income,ed,employ,retire,gender,reside
	Non assegnato(troppi valori mancanti)	
	Non assegnato(nessun valore mancante)	
Sequenza di assegnazione		age,tenure,reside,ed,gender,retire,employ,marital,address,income

I Risultati di assegnazione offrono una panoramica di ciò che è effettivamente accaduto durante il processo di assegnazione. Si noti in particolare che:

- Il metodo di assegnazione nella tabella delle specifiche era impostato su Automatico e il metodo effettivamente scelto dalla selezione automatica del metodo è stato Specifica completamente condizionale.
- Sono state assegnate tutte le variabili richieste.
- La sequenza di assegnazione è l'ordine in cui le variabili appaiono sull'asse delle x nel grafico Motivi valori mancanti.

### Modelli di assegnazione

Figura 5-11  
Modelli di assegnazione

	Modello		Valori mancanti	Valori assegnati
	Tipo	Effetti		
Età in anni	Regressione lineare	ed,gender,retire,marital,tenure,reside,employ,address,income	25	125
Mesi con servizio	Regressione lineare	ed,gender,retire,marital,age,reside,employ,address,income	32	160
Componenti nucleo familiare	Regressione lineare	ed,gender,retire,marital,age,tenure,employ,address,income	34	170
Livello di istruzione	Regressione logistica	gender,retire,marital,age,tenure,reside,employ,address,income	35	175
Sesso	Regressione logistica	ed,retire,marital,age,tenure,reside,employ,address,income	42	210
Pensionato/a	Regressione logistica	ed,gender,marital,age,tenure,reside,employ,address,income	84	420
Anni di permanenza nell'impiego attuale	Regressione lineare	ed,gender,retire,marital,age,tenure,reside,address,income	96	480
Stato civile	Regressione logistica	ed,gender,retire,age,tenure,reside,employ,address,income	115	575
Anni all'attuale indirizzo	Regressione lineare	ed,gender,retire,marital,age,tenure,reside,employ,income	150	750
Reddito familiare in migliaia	Regressione lineare	ed,gender,retire,marital,age,tenure,reside,employ,address	179	895

La tabella Modelli di assegnazione fornisce ulteriori dettagli sul modo in cui è stata assegnata ciascuna variabile. Si noti in particolare che:

- Le variabili vengono elencate nell'ordine della sequenza di assegnazione.
- Le variabili di scala sono modellate con una regressione lineare, mentre le variabili categoriali con una regressione logistica.
- Ciascun modello utilizza tutte le altre variabili come effetti principali.
- Viene riportato il numero di valori mancanti per ciascuna variabile, insieme al numero totale di valori assegnati per quella variabile (numero mancanti × numero di assegnazioni).

### Statistiche descrittive

Figura 5-12  
Statistiche descrittive per la durata (Mesi con servizio)

Dati	Assegnazione	N	Media	Deviazione standard Variabile	Minimo	Massimo
Dati originali		968	35,56	21,268	1,00	72,00
Valori assegnati	1	32	39,17	21,610	1,52	93,69
	2	32	36,63	16,958	4,29	85,29
	3	32	42,39	23,733	4,82	90,89
	4	32	39,97	21,813	10,43	95,04
	5	32	41,86	21,703	5,18	88,84
Dati completi dopo l'assegnazione	1	1000	35,68	21,278	1,00	93,69
	2	1000	35,60	21,138	1,00	85,29
	3	1000	35,78	21,372	1,00	90,89
	4	1000	35,70	21,289	1,00	95,04
	5	1000	35,76	21,300	1,00	88,84

Le tabelle delle statistiche descrittive mostrano dei riepiloghi di variabili con valori assegnati. Viene prodotta una tabella separata per ogni variabile. I tipi di statistiche mostrate dipendono dal tipo di variabile, di scala o categoriale.

Le statistiche per le variabili di scala includono il conteggio, la media, la deviazione standard, il valore minimo e massimo, visualizzati per i dati originali, ciascun set di valori assegnati e ciascun insieme di dati completo (che combina i dati originali con i valori assegnati).

La tabella delle statistiche descrittive per *durata (Mesi con servizio)* mostra le medie e le deviazioni standard in ciascun insieme di valori assegnati, equivalenti all'incirca a quelle nei dati originali. Tuttavia, sorge immediatamente un problema quando si osserva il valore minimo e ci si accorge che sono stati assegnati i valori negativi per *durata*.

Figura 5-13  
Statistiche descrittive per *statociv* (*Stato civile*)

Dati	As...	Ca...	N	Percentuale
Dati originali		0	456	51,5
		1	429	48,5
Valori assegnati	1	0	51	44,3
		1	64	55,7
	2	0	49	42,6
		1	66	57,4
	3	0	51	44,3
		1	64	55,7
	4	0	50	43,5
		1	65	56,5
	5	0	54	47,0
		1	61	53,0
Dati completi dopo l'assegnazione	1	0	507	50,7
		1	493	49,3
	2	0	505	50,5
		1	495	49,5
	3	0	507	50,7
		1	493	49,3
	4	0	506	50,6
		1	494	49,4
	5	0	510	51,0
		1	490	49,0

Per le variabili categoriali, le statistiche includono il conteggio e la percentuale in base alla categoria per i dati originali, i valori assegnati e i dati completi. La tabella per *statociv* (*Stato civile*) mostra un risultato interessante in quanto, per i valori assegnati, viene stimata una percentuale maggiore dei casi con stato civile coniugato rispetto ai dati originali. Ciò può essere dovuto a una variazione casuale, in alternativa la possibilità di essere mancante può essere ricondotta al valore di questa variabile.

Figura 5-14  
Statistiche descrittive per *reddito* (*Reddito familiare in migliaia*)

Dati	Assegnazione	N	Media	Deviazione standard Variabile	Minimo	Massimo
Dati originali		821	71,1462	83,14424	9,0000	944,0000
Valori assegnati	1	179	116,4104	74,76155	1,1206	283,4055
	2	179	115,2794	81,58019	1,6164	372,4810
	3	179	119,4888	77,46842	1,0957	392,0048
	4	179	111,8589	74,05791	1,0153	335,2312
	5	179	105,6987	77,29500	,2568	434,6000
Dati completi dopo l'assegnazione	1	1000	79,2485	83,49607	1,1206	944,0000
	2	1000	79,0460	84,53795	1,6164	944,0000
	3	1000	79,7995	84,18673	1,0957	944,0000
	4	1000	78,4337	83,03836	1,0153	944,0000
	5	1000	77,3311	83,15323	,2568	944,0000

Come accade per *durata*, e per tutte le altre variabili di scala, *reddito* (*Reddito familiare in migliaia*) mostra chiaramente dei valori assegnati negativi, pertanto dovremo eseguire un modello personalizzato con alcuni vincoli su certe variabili. Tuttavia, *reddito* mostra altri potenziali

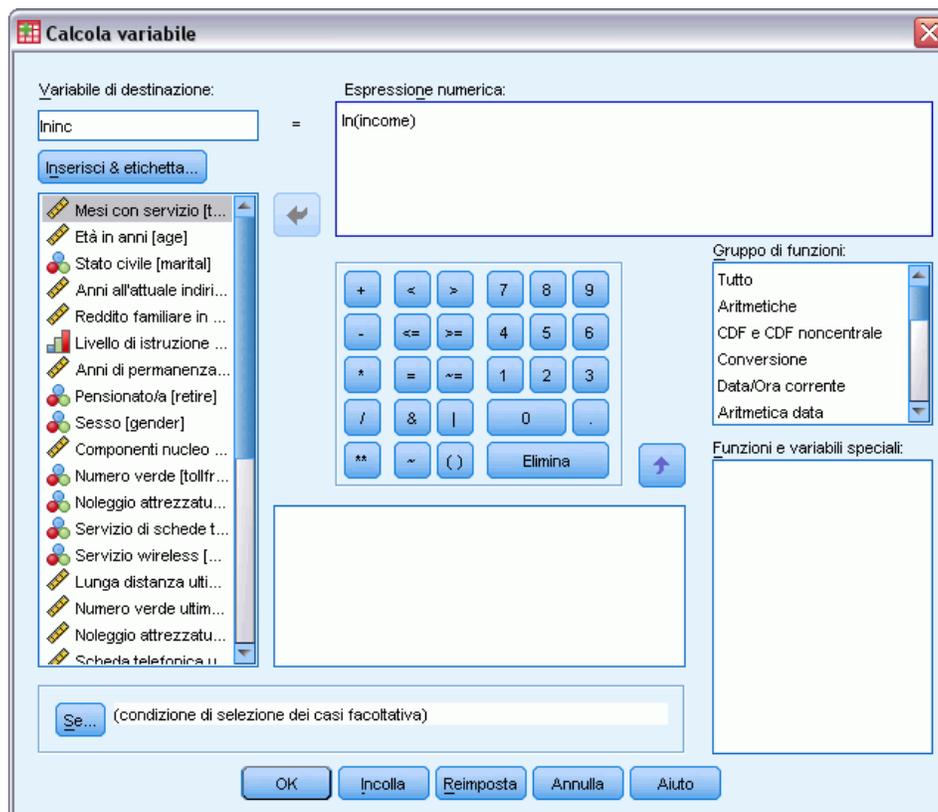
problemi. I valori delle medie per ciascuna assegnazione sono considerevolmente più alti rispetto ai dati originali e i valori massimi per ciascuna assegnazione sono considerevolmente più bassi rispetto ai dati originali. La distribuzione del reddito tende a essere altamente asimmetrica verso destra e questo potrebbe essere l'origine del problema.

## Modello di assegnazione personalizzato

Per impedire che i valori assegnati ricadano all'esterno dell'intervallo ragionevole di valori per ciascuna variabile, specificheremo un modello di assegnazione personalizzato con vincoli sulle variabili. Inoltre, *Reddito familiare in migliaia* è altamente asimmetrico verso destra e l'ulteriore analisi utilizzerà probabilmente il logaritmo di *reddito*, pertanto è ragionevole assegnare direttamente il valore logaritmico del reddito.

- ▶ Assicurarsi che l'insieme di dati originale sia attivo.
- ▶ Per creare una variabile del reddito logaritmico, dai menu scegliere:  
Trasforma > Calcola variabile...

Figura 5-15  
Finestra di dialogo Calcola variabile



- ▶ Digitare *Ininc* come variabile di destinazione.

- ▶ Digitare  $\ln(\text{reddito})$  come espressione numerica.
- ▶ Fare clic su Inserisci & etichetta..

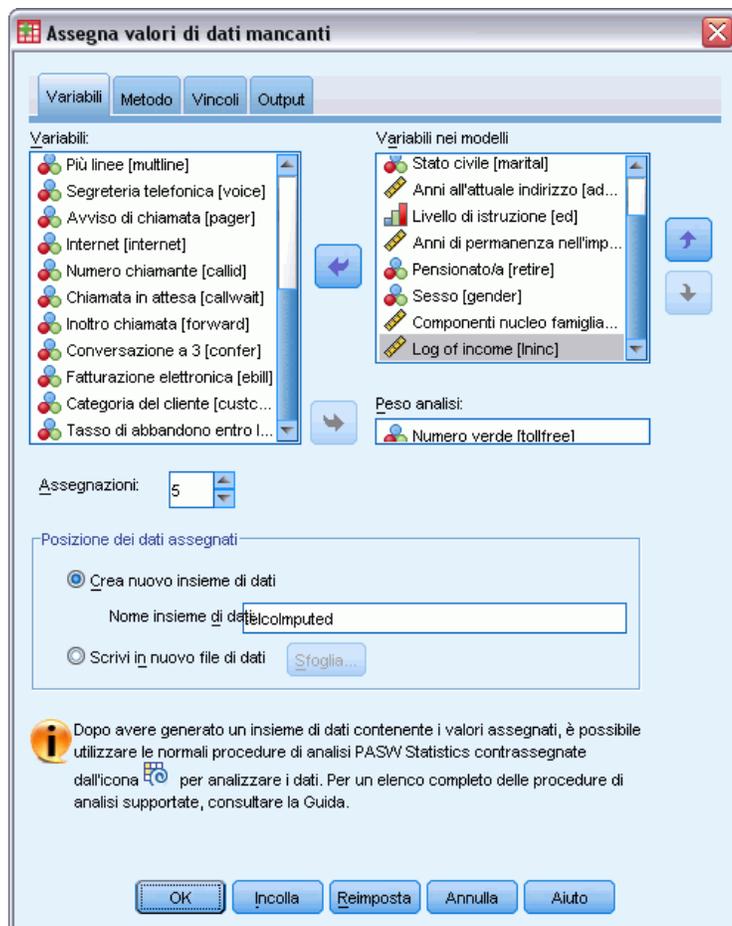
Figura 5-16

Finestra di dialogo Tipo ed etichetta



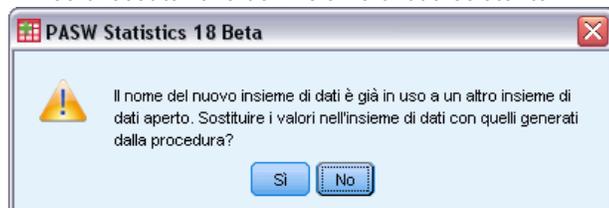
- ▶ Digitare *Logaritmo del reddito* come etichetta.
- ▶ Fare clic su Continua.
- ▶ Fare clic su OK nella finestra di dialogo Calcola variabile.

Figura 5-17  
 Scheda Variabili con Logaritmo del reddito che sostituisce Reddito familiare in migliaia nel modello di assegnazione



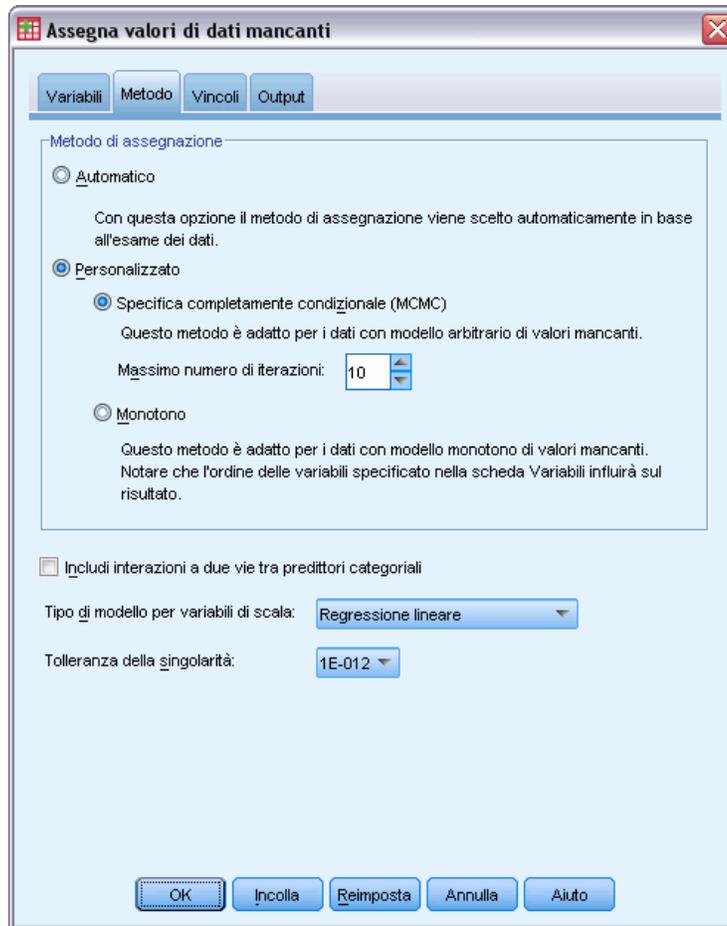
- ▶ Richiamare la finestra di dialogo Valori di dati mancanti e fare clic sulla scheda Variabili.
- ▶ Deselezionare *Reddito familiare in migliaia [reddito]* e selezionare *Logaritmo del reddito [lninc]* come variabile nel modello.
- ▶ Fare clic sulla scheda Metodo.

Figura 5-18  
 Avviso di sostituzione dell'insieme di dati esistente



- ▶ Fare clic su Sì nell'avviso visualizzato.

Figura 5-19  
Scheda Metodo



- ▶ Selezionare Personalizzata e lasciare selezionato Specifica completamente condizionale come metodo di assegnazione.
- ▶ Fare clic sulla scheda Vincoli.

Figura 5-20  
Scheda Vincoli

Assegna valori di dati mancanti

Variabili Metodo Vincoli Output

Esame dati per riepilogo variabili

Limita numero di casi esaminati Casi: 5000

Riepilogo variabili:

Variabili nei modelli	Percentuale dati ma...	Numero minimo dati...	Numero massimo d...
tenure	3,20	1	72
age	2,50	18	77
marital	11,50	0	1
address	15,00	0	55

Casi esaminati: 1000

Definisci vincoli:

Variabili nei modelli	Ruolo	Min	Max	Arrotond...
retire	Assegna e usa come pr...			
gender	Assegna e usa come pr...			
reside	Assegna e usa come pr...	1		1
lninc	Assegna e usa come pr...	0		

Escludi variabili con grandi quantità di dati mancanti

Percentuale massima dati mancanti:

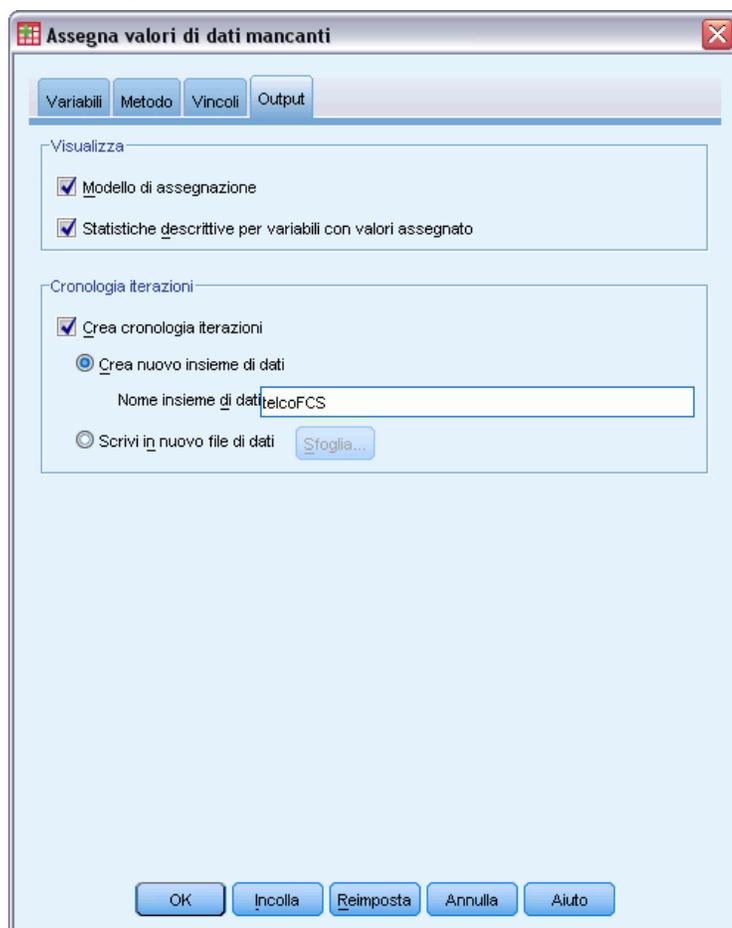
Numero massimo estrazioni casi:

Numero massimo estrazioni parametri:

L'aumento del numero massimo di estrazioni di parametri può aumentare notevolmente la durata dell'analisi.

- ▶ Fare clic su Esamina dati.
- ▶ Nella griglia Definisci vincoli, digitare 1 come valore minimo per *Mesi con servizio [durata]*.
- ▶ Digitare 18 come valore minimo per *età (Età in anni)*.
- ▶ Digitare 0 come valore minimo per *indirizzo (Anni all'attuale indirizzo)*.
- ▶ Digitare 0 come valore minimo per *impiego (Anni di permanenza nell'impiego attuale)*.
- ▶ Digitare 1 come valore minimo e 1 come il livello di arrotondamento per *nfam (Componenti nucleo familiare)*. Si noti che mentre molte delle altre variabili di scala vengono riportate con valori interi, è ragionevole presupporre che qualcuno abbia vissuto per 13,8 anni all'indirizzo attuale, ma non che 2,2 persone vivano a quell'indirizzo.
- ▶ Digitare 0 come valore minimo per *lninc (Logaritmo del reddito)*.
- ▶ Fare clic sulla scheda Output.

Figura 5-21  
scheda Output



- ▶ Selezionare Crea cronologia iterazioni e digitare telcoFCS come nome del nuovo insieme di dati.
- ▶ Fare clic su OK.

### Vincoli di assegnazione

Figura 5-22  
Vincoli di assegnazione

	Ruolo in assegnazione		Valori assegnati		
	Dipendente	Predittore	Minimo	Massimo	Arrotondamento
Mesi con servizio	Sì	Sì	1	(nessuno)	
Età in anni	Sì	Sì	18	(nessuno)	
Stato civile	Sì	Sì			
Anni all'attuale indirizzo	Sì	Sì	0	(nessuno)	
Livello di istruzione	Sì	Sì			
Anni di permanenza ...	Sì	Sì	0	(nessuno)	
Pensionato/a	Sì	Sì			
Sesso	Sì	Sì			
Componenti nucleo ...	Sì	Sì	1	(nessuno)	Intero
Log of Income	Sì	Sì	0	(nessuno)	

Il modello di assegnazione personalizzato genera una tabella nuova che riesamina i vincoli imposti al modello di assegnazione. Tutto sembra essere conforme alle specifiche indicate.

### Statistiche descrittive

Figura 5-23  
Statistiche descrittive per la durata (Mesi con servizio)

Dati	Assegnazione	N	Media	Deviazione standard Variabile	Minimo	Massimo
Dati originali		968	35,56	21,268	1,00	72,00
Valori assegnati	1	32	36,53	23,000	1,12	89,97
	2	32	41,53	27,370	1,24	126,01
	3	32	39,10	19,183	14,73	93,25
	4	32	40,17	22,113	7,78	89,33
	5	32	41,34	23,936	3,05	87,90
Dati completi dopo l'assegnazione	1	1000	35,59	21,314	1,00	89,97
	2	1000	35,75	21,499	1,00	126,01
	3	1000	35,67	21,205	1,00	93,25
	4	1000	35,71	21,300	1,00	89,33
	5	1000	35,75	21,370	1,00	87,90

La tabella delle statistiche descrittive per *durata (Mesi con servizio)* sotto il modello di assegnazione personalizzato con vincoli mostra che il problema dei valori assegnati negativi per *durata* è stato risolto.

Figura 5-24  
Statistiche descrittive per *statociv* (Stato civile)

Dati	As...	Ca...	N	Percentuale
Dati originali		0	456	51,5
		1	429	48,5
Valori assegnati	1	0	56	48,7
		1	59	51,3
	2	0	41	35,7
		1	74	64,3
	3	0	52	45,2
		1	63	54,8
	4	0	51	44,3
		1	64	55,7
	5	0	49	42,6
		1	66	57,4
Dati completi dopo l'assegnazione	1	0	512	51,2
		1	488	48,8
	2	0	497	49,7
		1	503	50,3
	3	0	508	50,8
		1	492	49,2
	4	0	507	50,7
		1	493	49,3
	5	0	505	50,5
		1	495	49,5

La tabella per *statociv* (Stato civile) ora presenta un'assegnazione (3) la cui distribuzione è più in linea con i dati originali, ma la maggior parte mostra ancora una percentuale maggiore di casi stimati come stato civile coniugato rispetto ai dati originali. Ciò può essere dovuto a una variazione casuale, ma potrebbe essere necessario studiare i dati in modo più approfondito per determinare se questi valori non sono mancanti in modo casuale. Noi ci fermeremo qui con l'analisi.

Figura 5-25  
Statistiche descrittive per *lninc* (Logaritmo del reddito)

Dati	Assegnazione	N	Media	Deviazione standard Variabile	Minimo	Massimo
Dati originali		821	3,9291	,75305	2,1972	6,8501
Valori assegnati	1	179	4,1101	,90290	,9481	6,1834
	2	179	4,1957	,93415	2,1554	7,0682
	3	179	4,1496	,94692	1,3730	6,7174
	4	179	4,1591	,87912	2,0418	6,4203
	5	179	4,2245	,90974	1,9923	6,6590
Dati completi dopo l'assegnazione	1	1000	3,9615	,78457	,9481	6,8501
	2	1000	3,9768	,79462	2,1554	7,0682
	3	1000	3,9685	,79523	1,3730	6,8501
	4	1000	3,9702	,78164	2,0418	6,8501
	5	1000	3,9819	,79106	1,9923	6,8501

Analogamente alla variabile *durata* e a tutte le altre variabili di scala, *lninc* (Logaritmo del reddito) non mostra valori negativi assegnati. Inoltre, i valori delle medie per le assegnazioni sono più vicini alla media per i dati originali rispetto a quanto accade nell'esecuzione dell'assegnazione automatica — nella scala *reddito*, la media dei dati originali per *lninc* è all'incirca  $e^{3,9291}=50,86$ ,

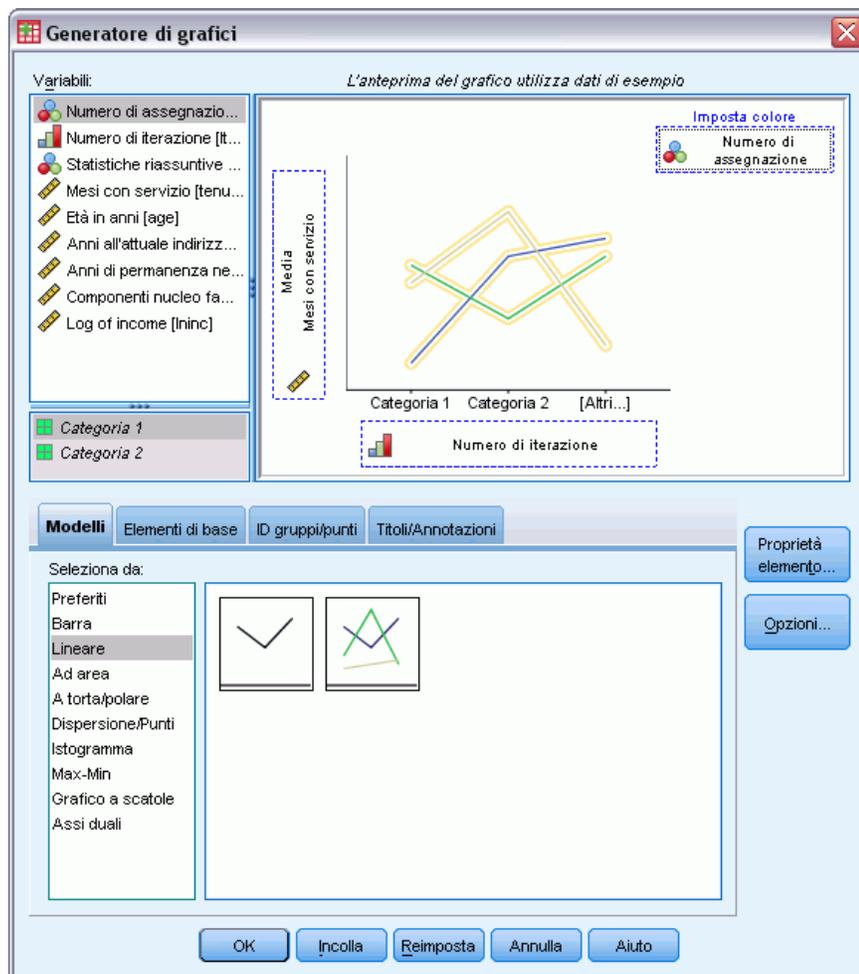
mentre il tipico valore della media tra le assegnazioni è grosso modo  $e^{4,2}=66,69$ . In aggiunta, i valori massimi per ciascuna assegnazione sono prossimi al valore massimo per i dati originali.

## Verifica della convergenza FCS

Quando si utilizza il metodo di specifica completamente condizionale, è buona norma controllare i grafici delle medie e delle deviazioni standard in base alle iterazioni per ciascuna variabile dipendente di scala per la quale i valori vengono assegnati in modo da consentire la valutazione della convergenza del modello.

- Per creare questo tipo di grafico, attivare l'insieme di dati *telcoFCS*, quindi dai menu scegliere: Grafici > Generatore di grafici...

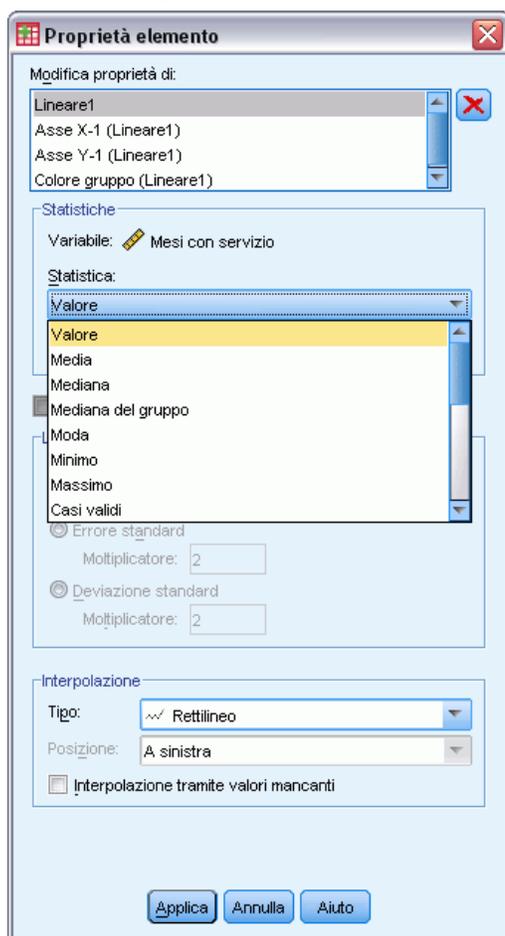
Figura 5-26  
Generatore di grafici, grafico Più linee



- Selezionare il modello Lineare e scegliere Più linee.

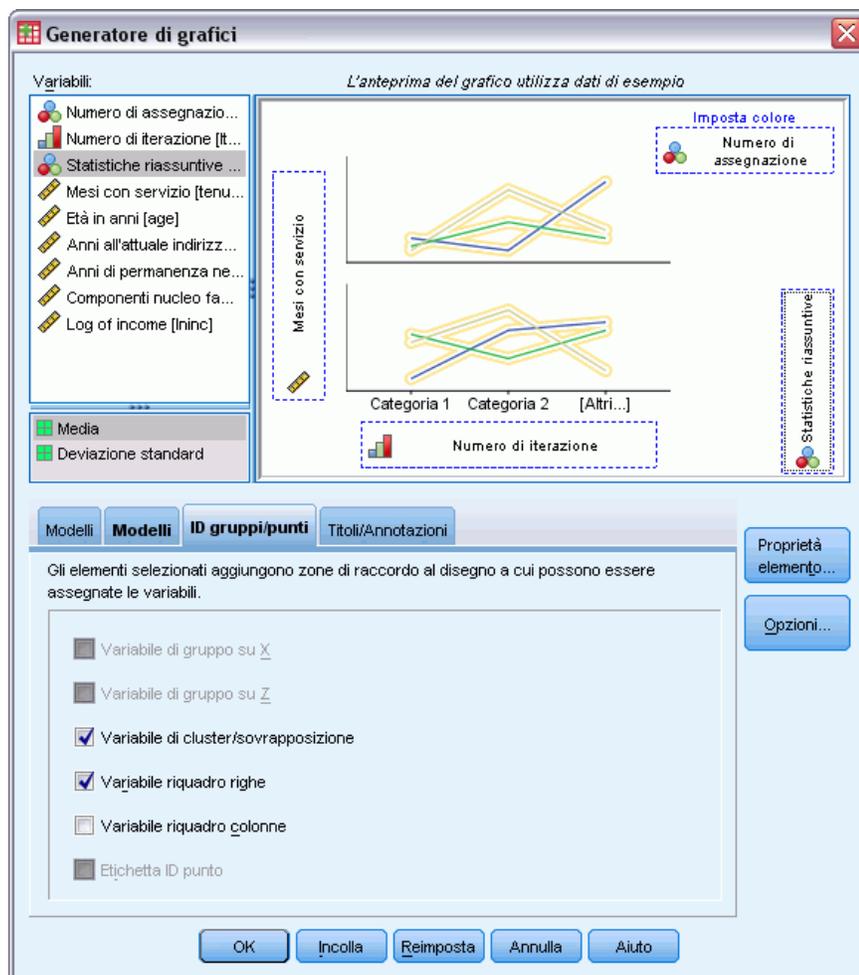
- ▶ Selezionare *Mesi con servizio [durata]* come la variabile da rappresentare graficamente sull'asse delle *Y*.
- ▶ Selezionare *Numero di iterazione [Iteration\_]* come variabile da rappresentare graficamente nell'asse delle *X*.
- ▶ Selezionare *Numero di assegnazione [Imputations\_]* come variabile in base alla quale impostare i colori.

Figura 5-27  
Generatore di grafici, Proprietà elemento



- ▶ In Proprietà elemento, selezionare Valore come la statistica da visualizzare.
- ▶ Fare clic su Applica.
- ▶ Nel Generatore di grafici, fare clic sulla scheda ID gruppi/punti.

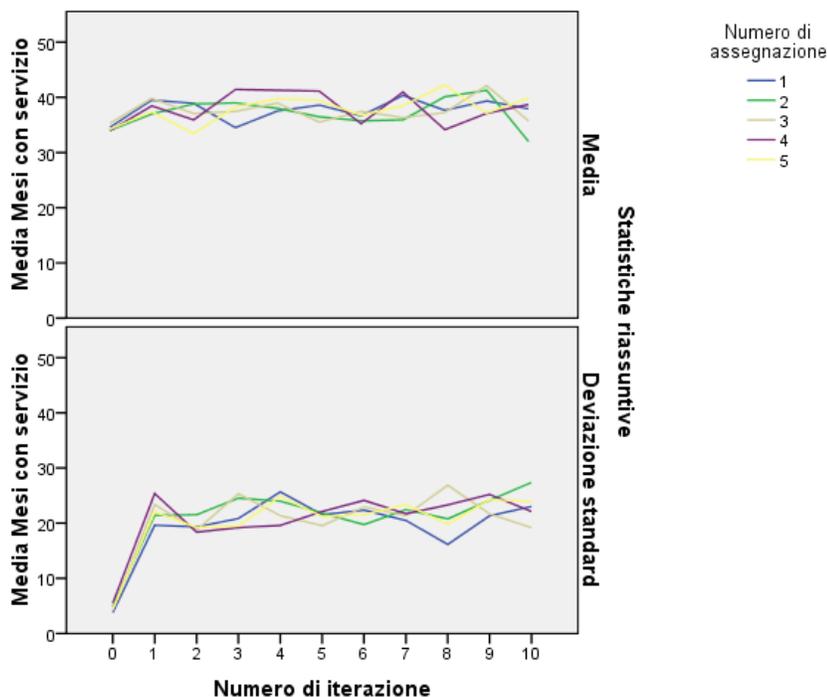
Figura 5-28  
Generatore di grafici, scheda ID gruppi/punti



- ▶ Selezionare Variabile riquadro righe.
- ▶ Selezionare *Statistiche riassuntive [SummaryStatistic]* come variabile riquadro.
- ▶ Fare clic su OK.

### Grafici di convergenza FCS

Figura 5-29  
Grafico di convergenza FCS



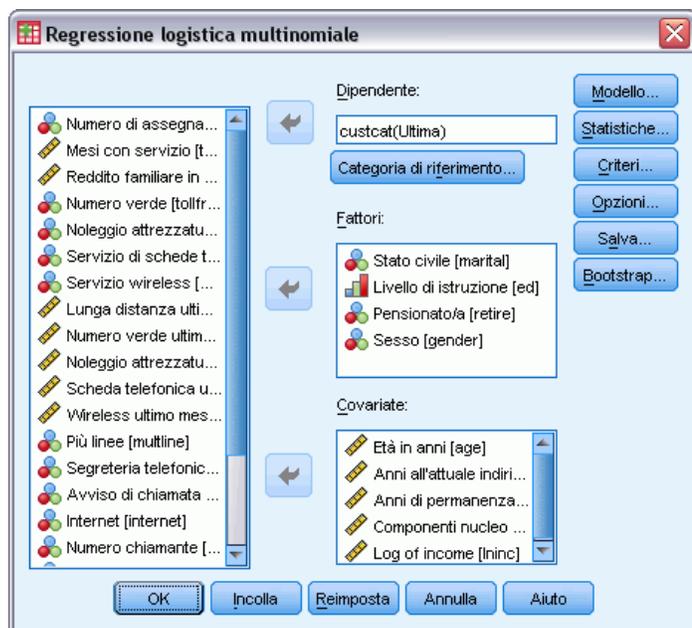
È stata creata una coppia di grafici a più linee che mostrano la media e la deviazione standard dei valori assegnati di *Mesi con servizio [durata]* a ciascuna iterazione del metodo di assegnazione FCS (Specificamente completamente condizionale) per ciascuna delle 5 assegnazioni richieste. Lo scopo di questo grafico è la ricerca dei motivi nelle linee. Non dovrebbero essercene e sembrano “casuali” nel modo appropriato. È possibile creare grafici simili per altre variabili di scala e facciamo notare che quei grafici, inoltre, non mostrano motivi rilevabili.

### Analisi dei dati completi

Ora che i valori assegnati sembrano soddisfacenti, si è pronti a eseguire un’analisi sui dati “completi”. L’insieme di dati contiene una variabile *Categoria del cliente [catcli]* che segmenta i clienti in base ai modelli di utilizzo dei servizi, suddividendo i clienti in quattro gruppi. Se si può adattare un modello utilizzando i dati demografici per prevedere l’appartenenza al gruppo, è possibile personalizzare le offerte per potenziali clienti individuali.

- Attivare l’insieme di dati *telcoImputed*. Per creare un modello di regressione logistica multinomiale per i dati completi, dai menu scegliere:  
Analizza > Regression > Logistica multinomiale...

Figura 5-30  
Finestra di dialogo *Regressione logistica multinomiale*



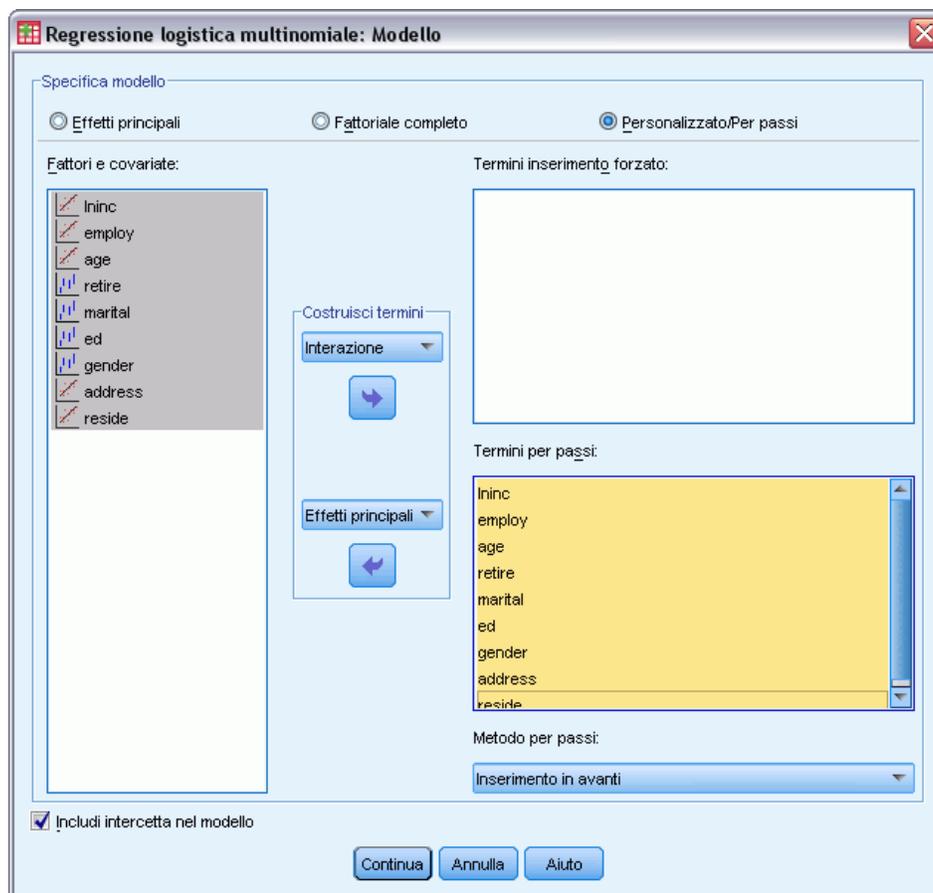
- ▶ Selezionare *Categoria del cliente* come variabile dipendente.
- ▶ Selezionare *Stato civile*, *Istruzione*, *Pensionato* e *Sesso* come fattori.
- ▶ Selezionare *Età in anni*, *Anni all'attuale indirizzo*, *Anni di permanenza nell'impiego attuale*, *Componenti nucleo familiare* e *Logaritmo del reddito* come covariate.
- ▶ Per confrontare altri clienti con quelli abbonati al Servizio base, selezionare *Categoria del cliente* e fare clic su *Categoria di riferimento*.

Figura 5-31  
Finestra di dialogo *Categoria di riferimento*



- ▶ Selezionare Prima categoria.
- ▶ Fare clic su Continua.
- ▶ Nella finestra di dialogo Regressione logistica multinomiale, fare clic su Modello.

Figura 5-32  
Finestra di dialogo Modello



- ▶ Selezionare Personalizzato/Per passi.
- ▶ Selezionare Effetti principali dall'elenco a discesa Termini per passi, Costruisci termine/i.
- ▶ Selezionare da *lninc* fino a *nfam* come Termini per passi.
- ▶ Fare clic su Continua.
- ▶ Nella finestra di dialogo Regressione logistica multinomiale, fare clic su OK.

**Riepilogo passi**

Figura 5-33  
Riepilogo passi

Numero di assegnazione	Modello	Operazione	Effetti	Criteria di adattament o del modello	Test di selezione degli effetti		
				-2 Log verosimiglia nza	Chi- quadrato <sup>a</sup>	df	Sig.
Dati originali	0	Aggiunti	Intercetta	1307,711			
	1	Aggiunti	tenure	1220,010	87,701	3	,000
	2	Aggiunti	ed	1127,923	92,087	12	,000
	3	Aggiunti	marital	1229,808	7,856	3	,049
1	0	Aggiunti	Intercetta	2762,531			
	1	Aggiunti	ed	2606,291	156,240	12	,000
	2	Aggiunti	tenure	2472,334	133,956	3	,000
	3	Aggiunti	employ	2458,684	13,650	3	,003
	4	Aggiunti	reside	2446,139	12,545	3	,006
2	0	Aggiunti	Intercetta	2762,531			
	1	Aggiunti	ed	2601,995	160,536	12	,000
	2	Aggiunti	tenure	2472,375	129,620	3	,000
	3	Aggiunti	Lninc	2454,220	18,155	3	,000
	4	Aggiunti	reside	2443,437	10,783	3	,013
	5	Aggiunti	retire	2434,487	8,951	3	,030
3	0	Aggiunti	Intercetta	2762,531			
	1	Aggiunti	ed	2603,012	159,519	12	,000
	2	Aggiunti	tenure	2477,040	125,972	3	,000
	3	Aggiunti	Lninc	2458,693	18,347	3	,000
	4	Aggiunti	reside	2445,799	12,894	3	,005
4	0	Aggiunti	Intercetta	2762,531			
	1	Aggiunti	ed	2600,174	162,358	12	,000
	2	Aggiunti	tenure	2470,463	129,710	3	,000
	3	Aggiunti	Lninc	2448,059	22,404	3	,000
	4	Aggiunti	reside	2434,805	13,254	3	,004
	5	Aggiunti	retire	2426,597	8,208	3	,042
5	0	Aggiunti	Intercetta	2762,531			
	1	Aggiunti	ed	2600,384	162,147	12	,000
	2	Aggiunti	tenure	2472,527	127,857	3	,000
	3	Aggiunti	Lninc	2452,739	19,788	3	,000
	4	Aggiunti	reside	2440,803	11,936	3	,008

Metodo stepwise: stepwise in avanti

a. Il chi-quadrato per l'immissione si basa sul test del rapporto di verosimiglianza.

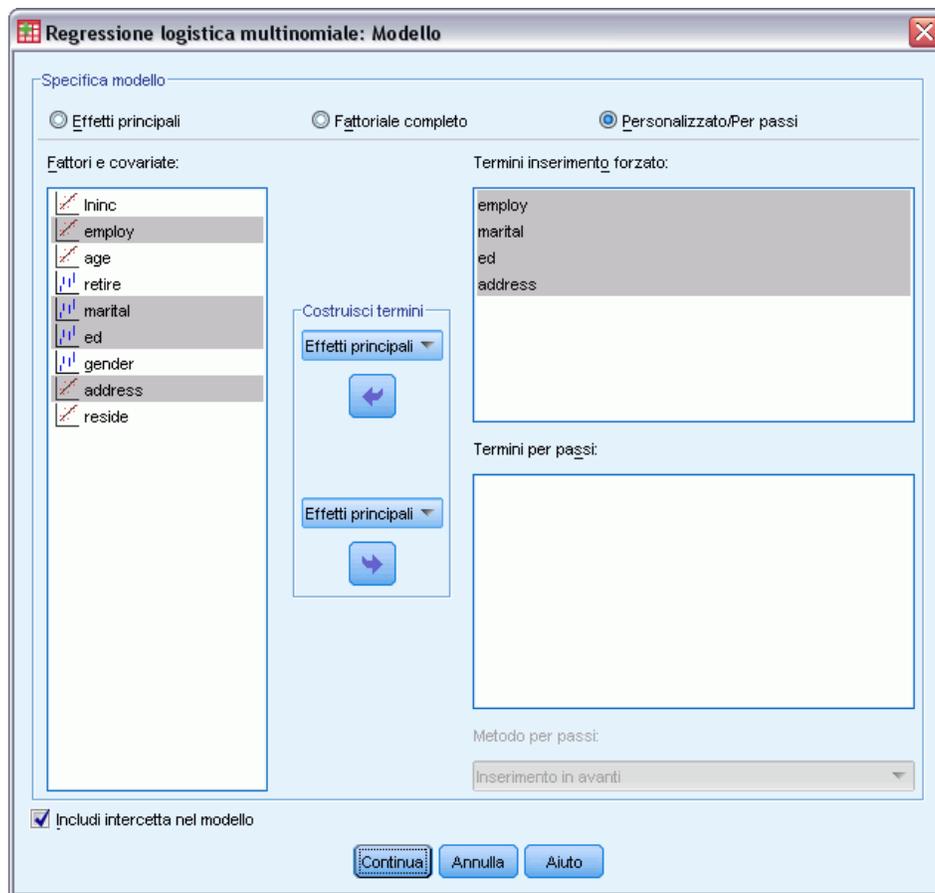
La Regressione logistica multinomiale supporta il raggruppamento dei coefficienti di regressione; tuttavia, si può notare come note *tutte* le tabelle nell'output mostrino i risultati per ciascuna assegnazione e i dati originali. Questo si spiega con il fatto che il file è distinto in *Imputation\_*, cosicché tutte le tabelle che rispettano la variabile di distinzione presenteranno i gruppi del file distinto insieme in un'unica tabella.

Si noterà anche che la tabella Stime dei parametri non mostra le stime raggruppate. Per conoscerne il motivo, osservare il Riepilogo passi. Abbiamo richiesto la selezione per passi degli effetti del modello e non è stato scelto lo stesso insieme di effetti per tutte le assegnazioni, pertanto non è possibile eseguire il raggruppamento. Tuttavia, si ottengono comunque informazioni utili, in quanto vediamo che *istrucz.* (*Livello di istruzione*), *impiego* (*Anni di permanenza nell'impiego attuale*), *statociv* (*Stato civile*) e *indirizzo* (*Anni all'attuale indirizzo*) vengono scelti

frequentemente dalla selezione per passi tra le assegnazioni. Adatteremo un altro modello utilizzando solo questi predittori.

### Esecuzione del modello con un sottoinsieme di predittori

Figura 5-34  
Finestra di dialogo Modello



- ▶ Richiamare la finestra di dialogo Regressione logistica multinomiale e fare clic su Modello.
- ▶ Deselezionare le variabili dall'elenco Termini per passi.
- ▶ Selezionare Effetti principali dall'elenco a discesa Termini inserimento forzato, Costruisci termine/i.
- ▶ Selezionare *impiego*, *stato civ.*, *istruz.* e *indirizzo* come Termini inserimento forzato.
- ▶ Fare clic su Continua.
- ▶ Nella finestra di dialogo Regressione logistica multinomiale, fare clic su OK.

### Stime dei parametri raggruppate

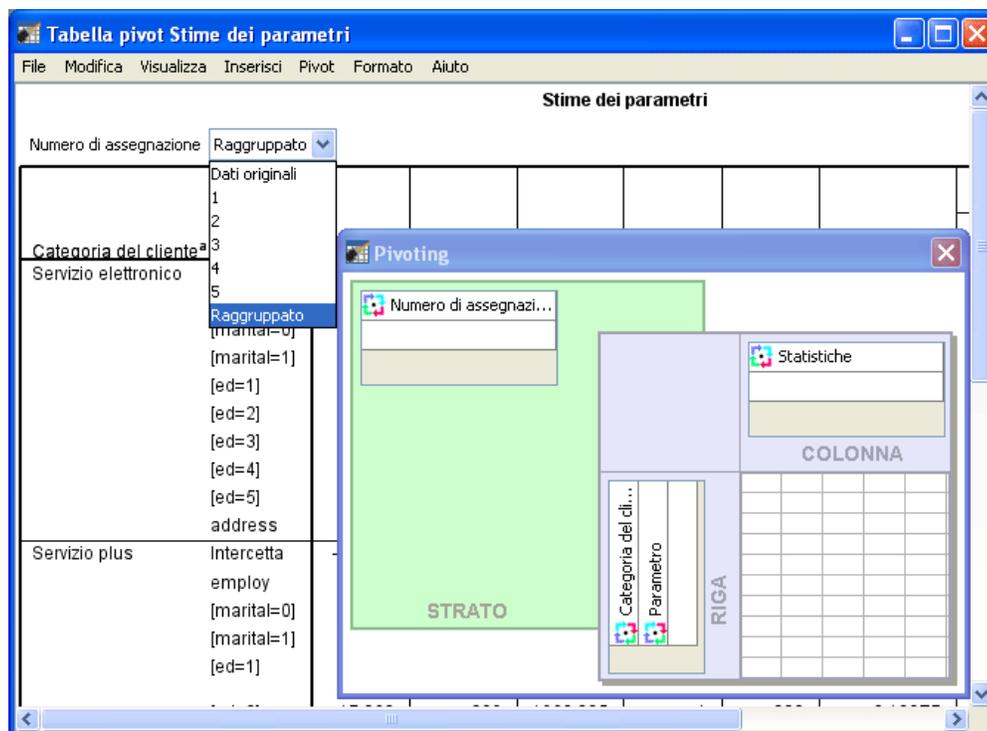
Questa tabella è piuttosto ampia, ma il pivoting ci fornirà un paio di diverse panoramiche utili dell'output.

Figura 5-35  
Stime dei parametri raggruppate

Numero di assegnazione	Categoria di...	Coefficiente	Wald	df	Sig.	Exp(1)
Dati originali	Servizio elet...		1,630	1	,202	
			7,819	1	,005	
			6,914	1	,009	
				0		
			11,263	1	,001	
			7,161	1	,007	
			1,857	1	,173	
			1,350	1	,245	
				0		
	Servizio plus		2171,237	1	,000	
			11,037	1	,001	
			4,368	1	,037	
				0		
		[ed=1]	1809,447	1	,000	3,8
		[ed=2]	1968,235	1	,000	3,1
		[ed=3]	1471,655	1	,000	4,7
		[ed=4]		1		5,0

- Attivare (con il doppio clic) la tabella, quindi selezionare Pivoting dal menu di scelta rapida.

Figura 5-36  
Stime dei parametri raggruppate



- ▶ Spostare *Numero di assegnazione* dalla riga allo strato.
- ▶ Selezionare Raggruppato dall'elenco a discesa Numero di assegnazione.

Figura 5-37  
Stime dei parametri raggruppate

Categoria del cliente <sup>a,b,c,d,e,f</sup>	B	Errore std	Sig.	Exp(B)	Intervallo di confidenza al 95% per Exp(B)		Info mancanti frazione	Varianza aumento relativo	Efficienza relativa	
					Limite inferiore	Limite superiore				
Servizio elettronico	Intercetta	,583	,427	,172			,035	,035	,993	
	employ	,028	,012	,017	1,029	1,005	1,053	,054	,056	,989
	[marital=0]	-,556	,197	,005	,573	,389	,844	,070	,072	,986
	[marital=1]	0 <sup>g</sup>								
	[ed=1]	-2,102	,471	,000	,122	,049	,308	,046	,047	,991
	[ed=2]	-1,367	,448	,002	,255	,106	,614	,059	,060	,988
	[ed=3]	-,786	,452	,082	,455	,188	1,104	,037	,037	,993
	[ed=4]	-,625	,451	,166	,536	,221	1,296	,047	,048	,991
	[ed=5]	0 <sup>g</sup>								
address	,032	,012	,006	1,033	1,009	1,057	,084	,088	,984	
Servizio plus	Intercetta	-1,190	,627	,058			,066	,069	,987	
	employ	,049	,011	,000	1,050	1,029	1,073	,067	,069	,987
	[marital=0]	-,330	,182	,070	,719	,503	1,027	,048	,049	,990
	[marital=1]	0 <sup>g</sup>								
	[ed=1]	,529	,652	,418	1,697	,470	6,122	,112	,120	,978
	[ed=2]	,784	,633	,215	2,190	,633	7,580	,069	,072	,986
	[ed=3]	,760	,653	,245	2,137	,593	7,706	,089	,093	,983
	[ed=4]	,468	,655	,475	1,597	,441	5,780	,075	,078	,985
	[ed=5]	0 <sup>g</sup>								
address	,020	,011	,060	1,020	,999	1,041	,042	,043	,992	
Servizio totale	Intercetta	1,078	,408	,008			,034	,035	,993	
	employ	,040	,012	,001	1,041	1,017	1,066	,029	,030	,994
	[marital=0]	-,575	,196	,003	,562	,383	,826	,045	,046	,991
	[marital=1]	0 <sup>g</sup>								
	[ed=1]	-3,531	,512	,000	,029	,011	,080	,016	,016	,997
	[ed=2]	-1,825	,426	,000	,161	,070	,372	,042	,043	,992
	[ed=3]	-1,321	,439	,003	,267	,113	,632	,053	,054	,990
	[ed=4]	-,585	,426	,170	,557	,241	1,286	,056	,058	,989
	[ed=5]	0 <sup>g</sup>								
address	,016	,012	,189	1,016	,992	1,040	,027	,027	,995	

Questa visualizzazione mostra tutte le statistiche per i risultati raggruppati. È possibile utilizzare e interpretare questi coefficienti nello stesso modo in cui si utilizzerebbe questa tabella per un insieme di dati che non presenta valori mancanti.

La tabella delle stime dei parametri riepiloga l'effetto di ogni predittore. Il rapporto fra il coefficiente e il rispettivo errore standard, quadrato, equivale alla statistica di Wald. Se il livello di significatività della statistica di Wald è basso (inferiore a 0,05), il parametro risulta diverso da 0.

- I parametri con coefficienti significativamente negativi riducono la verosimiglianza di tale categoria di risposta rispetto alla categoria di riferimento.
- I parametri con coefficienti negativi aumentano la verosimiglianza di tale categoria di risposta.
- I parametri associati all'ultima categoria di ogni fattore è ridondante, dato il termine di intercetta.

Sono presenti tre colonne aggiuntive nella tabella che offrono più informazioni relative all'output raggruppato. La stima **info mancanti frazione** è una stima del rapporto di informazioni mancanti rispetto alle informazioni "complete", in base alla **varianza aumento relativo** dovuto a una

non-risposta, che a sua volta è un rapporto (modificato) della varianza tra-assegnazioni e della varianza media entro-l'assegnazione del coefficiente di regressione. L'**efficienza relativa** è un confronto di questa stima con una stima (teoretica) calcolata mediante un numero infinito di assegnazioni. L'efficienza relativa è determinata dalla frazione di informazioni mancanti e dal numero di assegnazioni utilizzato per ottenere il risultato raggruppato; quando la frazione di informazioni mancanti è ampia, è necessario un numero maggiore di assegnazioni per portare l'efficienza relativa più vicina a 1 e la stima raggruppata più vicina alla stima idealizzata.

Figura 5-38  
Stime dei parametri raggruppate

The screenshot shows the SPSS 'Tabella pivot Stime dei parametri' window. The main window displays a pivot table with the following data:

Parametro	[marital=1]	[ed=1]	[ed=2]	[ed=3]	[ed=4]	[ed=5]	address
Intercetta	-17						
[ed=1]	17,262	,793	,553	,842	,767	,965	,7
[ed=2]	17,680	,775	,490	,847	,728	,958	,7
[ed=3]	17,734	,490	,206	,555	,458	,631	,4
[ed=4]	0 <sup>a</sup>						
[ed=5]							
address	,022	,022	,020	,021	,016	,020	,0

The 'Pivoting' dialog box is open, showing the following configuration:

- STRATO:** Statistiche
- COLONNA:** Numero di assegnazioni...
- RIGA:** Categoria del cliente, Parametro

- ▶ Attivare ora (con il doppio clic) la tabella, quindi selezionare Pivoting dal menu di scelta rapida.
- ▶ Spostare *Numero di assegnazione* dallo strato alla colonna.
- ▶ Spostare *Statistiche* dalla colonna allo strato.
- ▶ Selezionare B dall'elenco a discesa Statistiche.

Figura 5-39  
Stime dei parametri raggruppate, Numero di assegnazione in Colonne e Statistiche nello strato

Statistiche= B		Numero di assegnazione						
Categoria del cliente <sup>a,b,c,d,e,f</sup>		Dati originali	1	2	3	4	5	Raggruppato
Servizio elettronico	Intercetta	,817	,643	,589	,571	,644	,468	,583
	employ	,051	,028	,028	,025	,032	,028	,028
	[marital=0]	-,760	-,577	-,491	-,529	-,571	-,612	-,556
	[marital=1]	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>
	[ed=1]	-2,376	-2,162	-2,167	-2,114	-2,123	-1,943	-2,102
	[ed=2]	-1,754	-1,412	-1,434	-1,362	-1,427	-1,200	-1,367
	[ed=3]	-,943	-,777	-,878	-,771	-,835	-,671	-,786
	[ed=4]	-,798	-,699	-,659	-,659	-,633	-,473	-,625
	[ed=5]	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>
	address	,030	,032	,035	,035	,027	,032	,032
Servizio plus	Intercetta	-17,768	-1,190	-,953	-1,278	-1,196	-1,332	-1,190
	employ	,054	,048	,052	,047	,052	,047	,049
	[marital=0]	-,542	-,330	-,364	-,295	-,292	-,367	-,330
	[marital=1]	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>
	[ed=1]	17,479	,488	,232	,655	,524	,745	,529
	[ed=2]	17,262	,793	,553	,842	,767	,965	,784
	[ed=3]	17,680	,775	,490	,847	,728	,958	,760
	[ed=4]	17,734	,490	,206	,555	,458	,631	,468
	[ed=5]	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>
	address	,022	,022	,020	,021	,016	,020	,020
Servizio totale	Intercetta	1,528	1,100	1,116	1,062	1,145	,968	1,078
	employ	,038	,041	,040	,037	,041	,042	,040
	[marital=0]	-,459	-,524	-,603	-,616	-,551	-,583	-,575
	[marital=1]	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>
	[ed=1]	-3,664	-3,537	-3,579	-3,508	-3,589	-3,442	-3,531
	[ed=2]	-2,305	-1,855	-1,882	-1,758	-1,903	-1,725	-1,825
	[ed=3]	-1,295	-1,327	-1,380	-1,250	-1,434	-1,212	-1,321
	[ed=4]	-,710	-,640	-,610	-,571	-,669	-,435	-,585
	[ed=5]	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>	0 <sup>g</sup>
	address	,005	,015	,018	,017	,015	,014	,016

Questa visualizzazione della tabella è utile per confrontare i valori tra le assegnazioni, per effettuare una rapida verifica visiva della variazione nelle stime del coefficiente di regressione di assegnazione in assegnazione e addirittura rispetto ai dati originali. In particolare, cambiando la statistica nello strato in Errore std. consente di vedere in che modo l'assegnazione multipla ha ridotto la variabilità nelle stime dei coefficienti rispetto all'eliminazione listwise (dati originali).

Figura 5-40  
Avvisi

Le variabili seguenti: retire, gender, tenure, age, reside, Lninc sono usate solo per definire le sottopopolazioni ma non per costruire il modello.

Per il file distinto Numero di assegnazione = Dati originali, si rilevano delle singolarità impreviste nella matrice hessiana. Questo significa che o alcune variabili stimatori devono essere escluse o alcune categoria devono essere unite.

La procedura NOMREG prosegue nonostante gli avvisi. I risultati visualizzati sono basati sull'ultima iterazione. La validità dell'adattamento del modello è incerta.

Tuttavia, in questo esempio, l'insieme di dati originale causa effettivamente un errore, il che spiega le stime molto grandi dei parametri per l'intercetta *Servizio plus* e i livelli non ridondanti di *istruz.* (*Livello di istruzione*) nella colonna dei dati originali della tabella.

## ***Riepilogo***

Utilizzando le procedure di assegnazione multipla, abbiamo analizzato i motivi dei valori mancanti e scoperto che molte delle informazioni hanno una probabilità di andare perse se venisse utilizzata l'eliminazione listwise semplice. Dopo un'esecuzione automatica iniziale dell'assegnazione multipla, abbiamo scoperto che i vincoli erano necessari per mantenere i valori assegnati entro certi limiti ragionevoli. L'esecuzione con vincoli ha prodotto buoni valori e non abbiamo avuto prove immediate della mancata convergenza del metodo FCS. Utilizzando l'insieme di dati "completo" con i valori di assegnazione multipli, abbiamo adattato una Regressione logistica multinomiale ai dati e ottenuto delle stime di regressione raggruppate, oltre ad avere scoperto che l'adattamento del modello finale, in realtà, non sarebbe stato possibile con l'eliminazione listwise sui dati originali.

## ***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).
- **anorectic.sav.** Per trovare una sintomatologia standardizzata del comportamento anoressico/bulimico, i ricercatori 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.
- **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 , 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 , 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 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 , 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. 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*.
- **debate.sav.** File di dati ipotetici che prende in esame le risposte appaite 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". 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 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 ha presentato una tabella per illustrare i possibili gruppi sociali. Guttman 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 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 similarità. 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  $15 \times 15$ , 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 *sexo*, *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/](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, e , 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.

- **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 .
- **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 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.
- **stocks.sav** Questo file di dati ipotetici contiene i prezzi e i volumi delle scorte riferiti a un anno.
- **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 miocardico, 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 .

- **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 .
- **verd1985.sav.** Questo file di dati prende in esame un'indagine . 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 . 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 selezionati in modo casuale e altri no. Ogni caso rappresenta un diverso partecipante al programma.
- **worldsales.sav** Questo file di dati ipotetici contiene i ricavi suddivisi per continenti e prodotti.

## ***Note legali***

Queste informazioni sono state preparate per prodotti e servizi offerti in tutto il mondo.

IBM potrebbe non offrire i prodotti, i servizi o le funzionalità di cui si tratta nel presente documento in altri paesi. Contattare il rappresentante IBM locale per informazioni sui prodotti e i servizi attualmente disponibili nella propria zona. Qualsiasi riferimento a un prodotto, programma o servizio IBM non intende dichiarare o implicare che sia possibile utilizzare esclusivamente tale prodotto, programma o servizio IBM. Potrà invece essere utilizzato qualsiasi prodotto, programma o servizio con funzionalità equivalente e che non violi i diritti di proprietà intellettuale di IBM. Tuttavia, è responsabilità dell'utente valutare e verificare il funzionamento di qualsiasi prodotto, programma o servizio non IBM.

IBM può essere titolare di brevetti o domande di brevetto relativi alla materia oggetto del presente documento. La consegna del presente documento non conferisce alcuna licenza rispetto a questi brevetti. Rivolgere per iscritto i quesiti sulle licenze a:

*IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.*

Per richieste di informazioni sulle licenze riguardanti il set di caratteri a byte doppio (DBCS), contattare l'Intellectual Property Department di IBM del proprio paese, oppure inviare le richieste in forma scritta all'indirizzo:

*Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Giappone.*

**Il seguente paragrafo non si applica per il Regno Unito o altri paesi in cui le presenti disposizioni non sono conformi alle leggi locali:** INTERNATIONAL BUSINESS MACHINES FORNISCE QUESTA PUBBLICAZIONE “COSÌ COM'È” SENZA GARANZIA DI ALCUN TIPO, SIA ESSA ESPRESSA O IMPLICITA, INCLUSE, MA NON LIMITATE A, LE GARANZIE IMPLICITE DI NON VIOLAZIONE, COMMERCIALIZZABILITÀ O IDONEITÀ A UNO SCOPO SPECIFICO. Alcuni stati non consentono limitazioni di garanzie espresse o implicite in determinate transazioni, pertanto quanto sopra potrebbe non essere applicabile.

Le presenti informazioni possono includere imprecisioni tecniche o errori tipografici. Le modifiche periodiche apportate alle informazioni contenute in questa pubblicazione verranno inserite nelle nuove edizioni della pubblicazione. IBM può apportare miglioramenti e/o modifiche al/ai prodotto/i e/o al/ai programma/i descritti nella presente pubblicazione in qualsiasi momento senza preavviso.

Qualsiasi riferimento nelle presenti informazioni a siti Web non IBM viene fornito esclusivamente per facilitare la consultazione e non rappresenta in alcun modo un'approvazione o sostegno da parte nostra di tali siti Web. I materiali contenuti in tali siti Web non fanno parte dei materiali di questo prodotto IBM e il loro utilizzo è esclusivamente a rischio dell'utente.

IBM può utilizzare o distribuire eventuali informazioni fornite dall'utente nei modi che ritiene appropriati senza incorrere in alcun obbligo nei confronti dell'utente.

I licenziatari del programma che desiderassero informazioni su di esso allo scopo di abilitare: (i) lo scambio di informazioni tra programmi creati indipendentemente e altri programmi (questo compreso) e (ii) l'utilizzo in comune delle informazioni scambiate, dovranno rivolgersi a:

*IBM Software Group, All'attenzione di: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.*

Tali informazioni saranno fornite in conformità ai termini e alle condizioni in vigore e, in alcuni casi, dietro pagamento.

Il programma concesso in licenza descritto nel presente documento e tutto il materiale correlato disponibile sono forniti da IBM in base ai termini del contratto di licenza cliente IBM, del contratto di licenza internazionale IBM o del contratto equivalente esistente tra le parti.

le informazioni relative a prodotti non IBM sono state ottenute dai fornitori di tali prodotti, da loro annunci pubblicati e da altre fonti disponibili al pubblico. IBM non ha verificato tali prodotti e non può confermare l'accuratezza delle prestazioni, la compatibilità o qualsiasi altra dichiarazione relativa a prodotti non IBM. Eventuali domande in merito alle funzionalità dei prodotti non IBM vanno indirizzate ai fornitori di tali prodotti.

Le presenti informazioni includono esempi di dati e report utilizzati in operazioni aziendali quotidiane. Per fornire una descrizione il più possibile esaustiva, gli esempi includono nomi di persone, società, marchi e prodotti. Tutti questi nomi sono fittizi e ogni somiglianza a nomi e indirizzi utilizzati da aziende reali è puramente casuale.

Per chi visualizza queste informazioni a video: le fotografie e le illustrazioni a colori potrebbero non essere disponibili.

### ***Marchi commerciali***

IBM, il logo IBM, ibm.com e SPSS sono marchi di IBM Corporation, registrati in numerose giurisdizioni nel mondo. Un elenco aggiornato dei marchi IBM è disponibile sul Web all'indirizzo <http://www.ibm.com/legal/copytrade.shtml>.

Adobe, il logo Adobe, PostScript e il logo PostScript sono marchi o marchi registrati di Adobe Systems Incorporated negli Stati Uniti e/o in altri paesi.

Intel, il logo Intel, Intel Inside, il logo Intel Inside, Intel Centrino, il logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium sono marchi o marchi registrati di Intel Corporation o delle sue consociate negli Stati Uniti e in altri paesi.

Java e tutti i marchi e i logo basati su Java sono marchi di Sun Microsystems, Inc. negli Stati Uniti e/o negli altri paesi.

Linux è un marchio registrato di Linus Torvalds negli Stati Uniti e/o negli altri paesi.

Microsoft, Windows, Windows NT e il logo Windows sono marchi di Microsoft Corporation negli Stati Uniti e/o negli altri paesi.

UNIX è un marchio registrato di The Open Group negli Stati Uniti e in altri paesi.

Questo prodotto utilizza WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, <http://www.winwrap.com>.

Altri nomi di prodotti e servizi possono essere marchi commerciali di IBM o di altre aziende.

Le schermate dei prodotti Adobe sono state ristampate su autorizzazione di Adobe Systems Incorporated.

Le schermate dei prodotti Microsoft sono state ristampate su autorizzazione di Microsoft Corporation.



- Analisi dati mancanti, 2
  - assegnazione dei valori mancanti, 8
  - EM, 9
  - EM (Expectation-Maximization), 11
  - funzioni aggiuntive del comando, 12
  - metodi, 8
  - modelli, 5
  - regressione, 10
  - statistiche descrittive, 6
  - stima delle statistiche, 8
  - test MCAR, 8
- Analisi dei valori mancanti, 36
  - modelli, 43
  - statistiche descrittive, 36
- Analizza modelli, 14
- Assegna valori di dati mancanti, 16
  - metodo di assegnazione, 19
  - output, 23
  - vincoli, 21
- assegnazione monotonica
  - in assegnazione multipla, 19
- assegnazione multipla, 13, 48
  - analizza modelli, 14
  - assegna valori di dati mancanti, 16
  - Grafico di convergenza FCS, 70
  - modelli, 56
  - motivi valori mancanti, 51
  - riepilogo generale dei valori mancanti, 49
  - riepilogo variabili, 50
  - risultati di assegnazione, 56
  - risultati raggruppati, 70
  - specifiche di assegnazione, 55
  - statistiche descrittive, 57, 65
  - stime raggruppate, 76
  - vincoli, 65
- Assegnazione multipla, 24, 28
  - opzioni, 33
- conteggi di valori estremi
  - in analisi dei dati mancanti, 6
- correlazioni
  - in analisi dei dati mancanti, 9–10
- covarianza
  - in analisi dei dati mancanti, 9–10
- cronologia iterazioni
  - in assegnazione multipla, 23
- dati incompleti
  - vedere Analisi dei dati mancanti, 2
- deviazione standard
  - in analisi dei dati mancanti, 6
- discordanza
  - in analisi dei dati mancanti, 6
- eliminazione listwise
  - in analisi dei dati mancanti, 2
- eliminazione pairwise
  - in analisi dei dati mancanti, 2
- EM
  - in analisi dei dati mancanti, 9
- file di esempio
  - posizione, 82
- Grafico di convergenza FCS
  - in assegnazione multipla, 70
- marchi commerciali, 93
- media
  - in analisi dei dati mancanti, 6, 9–10
- motivi valori mancanti, 45
- normovariati
  - in analisi dei dati mancanti, 10
- note legali, 92
- opzioni
  - assegnazione multipla, 33
- ordinamento dei casi
  - in analisi dei dati mancanti, 5
- regressione
  - in analisi dei dati mancanti, 10
- residui
  - in analisi dei dati mancanti, 10
- risultati raggruppati
  - in assegnazione multipla, 70
- specifica completamente condizionale
  - in assegnazione multipla, 19
- statistiche univariate
  - in analisi dei dati mancanti, 38
- stime raggruppate
  - in assegnazione multipla, 76
- tabella di frequenza
  - in analisi dei dati mancanti, 6
- tabelle dei casi
  - in analisi dei dati mancanti, 5
- tabelle di categorie
  - in analisi dei dati mancanti, 6
  - in analisi dei valori mancanti, 40
- test MCAR
  - in analisi dei dati mancanti, 2
  - in analisi dei valori mancanti, 46
- test MCAR di Little, 8
  - in analisi dei dati mancanti, 2

- in analisi dei valori mancanti, 46
- test t
  - in analisi dei dati mancanti, 6
  - in analisi dei valori mancanti, 39
- test t di Student
  - in analisi dei dati mancanti, 10
- Test *T* di Student
  - in analisi dei valori mancanti, 39
  
- valori mancanti
  - statistiche univariate, 6, 38
- variabili indicatore
  - in analisi dei dati mancanti, 6
- variabili indicatore mancanti
  - in analisi dei dati mancanti, 6