IBM SPSS Bootstrapping 20

Nota: Prima di utilizzare queste informazioni e il relativo prodotto, leggere le informazioni generali disponibili in Note legali a pag. 41.

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.

Prefazione

IBM® SPSS® Statistics è un sistema completo per l'analisi dei dati. Il modulo aggiuntivo opzionale Bootstrapping include le tecniche di analisi aggiuntive descritte nel presente manuale. Il modulo aggiuntivo Bootstrapping 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

Appendici

Parte I:
Manuale dell'utente

Capitolo

1

Introduzione a Bootstrapping

Quando si raccolgono dati, quello che interessa sono spesso le proprietà della popolazione da cui è stato tratto il campione. Si creano delle inferenze su questi parametri relativi alla popolazione mediante stime calcolate in base al campione. Per esempio, se l'insieme di dati *Employee data.sav* fornito con il prodotto è un campione casuale di una popolazione più estesa di dipendenti, la media del campione di 34.419,57 $ per *Stipendio corrente* è una stima dello stipendio corrente medio della popolazione di dipendenti. Inoltre, questa stima ha un errore standard di 784,311 $ per un campione di dimensione 474, pertanto un intervallo di confidenza del 95% per lo stipendio corrente medio nella popolazione di dipendenti è compreso tra 32.878,40 $ e 35.960,73 $. Ma quanto sono affidabili questi stimatori? Per determinate popolazioni "conosciute" e parametri affidabili, disponiamo di sufficienti informazioni sulle proprietà delle stime dei campioni e possiamo considerare affidabili i risultati. Bootstrapping cerca di scoprire ulteriori informazioni sulle proprietà degli stimatori per le popolazioni "sconosciute" e i parametri non affidabili.

Figura 1-1
Come creare inferenze parametriche sulla media della popolazione

			Statistica	Errore std.
Stipendio attuale (x000)	Media		34.419,57	784,311
	Intervallo di confidenza per la media al 95%	Limite inferiore	32.878,40	
		Limite superiore	35.960,73	
	Mediana		28.875,00	

Funzionamento di Bootstrapping

Nella sua versione più semplice, per un insieme di dati con una dimensione del campione pari a *N*, vengono estratti *B* campioni di "bootstrap" di dimensione *N* con sostituzione dall'insieme di dati originale e viene elaborato lo stimatore per ognuno di questi campioni di bootstrap *B*. Queste stime di bootstrap *B* sono un campione di dimensione *B* da cui è possibile creare delle inferenze sullo stimatore. Per esempio, se si prendono 1.000 campioni di bootstrap dall'insieme di dati *Employee data.sav*, l'errore standard stimato di bootstrap di 776,91 $ per la media del campione relativa a *Stipendio corrente* è un'alternativa alla stima di 784,311 $.

Inoltre, il bootstrap fornisce un errore standard e un intervallo di confidenza per la mediana, per la quale non sono disponibili stime parametriche.

Figura 1-2
Come creare inferenze di bootstrap sulla media del campione

					Bootstrap[a]			
							Intervallo di confidenza 95%	
			Statistica	Errore std.	Distorsione	Errore std.	Inferiore	Superiore
Stipendio attuale (x000)	Media		34.419,57	784,311	14,66	776,91	32.990,38	36.026,06
	Intervallo di confidenza per la media al 95%	Limite inferiore	32.878,40					
		Limite superiore	35.960,73					
	Mediana		28.875,00		-13,22	536,63	27.750,00	29.850,00

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Supporto per Bootstrapping nel prodotto

Bootstrapping è integrato sotto forma di finestra di dialogo secondaria nelle procedure che supportano il bootstrap. Vedere Procedure che supportano il bootstrap per informazioni sulle procedure che supportano il bootstrap.

Quando nelle finestre di dialogo è richiesto il bootstrap, un nuovo comando `BOOTSTRAP` separato viene incollato in aggiunta alla normale sintassi generata dalla finestra di dialogo. Il comando `BOOTSTRAP` crea i campioni di bootstrap in base alle specifiche indicate. A livello interno, il prodotto tratta questi campioni di bootstrap come distinzioni, anche se non sono esplicitamente mostrati nell'Editor dei dati. Ciò significa che, a livello interno, ci sono effettivamente *B***N* casi, quindi il contatore dei casi nella barra di stato effettua il conteggio da 1 a *B***N* quando vengono elaborati i dati durante il bootstrap. Il Sistema di gestione dell'output (OMS, Output Management System) viene utilizzato per raccogliere i risultati dell'esecuzione dell'analisi su ciascuno "split di bootstrap". Questi risultati vengono raggruppati, e i risultati di bootstrap raggruppati vengono visualizzati nel Viewer con il resto del normale output generato dalla procedura. In determinati casi, è possibile incontrare un riferimento a "bootstrap split 0"; si tratta dell'insieme di dati originale.

Capitolo
2

Bootstrapping

Il bootstrap è un metodo utilizzato per derivare delle stime affidabili su errori standard e intervalli di confidenza per stime quali media, mediana, proporzione, rapporto odd, coefficiente di correlazione o coefficiente di regressione. Può essere utilizzato anche per creare dei test di ipotesi. Il bootstrap è particolarmente utile come alternativa alle stime parametriche in caso di dubbio sulle supposizioni di questi metodi (come nel caso dei modelli di regressione con residui eteroschedastici adattati a campioni di piccole dimensioni) o quando l'inferenza parametrica è impossibile o richiede formule molto complicate per il calcolo degli errori standard (come nel caso del calcolo degli intervalli di confidenza per la mediana, i quartili e altri percentili).

Esempi. Un'azienda di telecomunicazioni perde ogni mese circa il 27% dei suoi clienti a causa della disdetta di un servizio da parte del cliente. Per analizzare correttamente gli sforzi da compiere al fine di ridurre la disdetta dei servizi, la direzione desidera conoscere se questa percentuale varia in gruppi di clienti predefiniti. Utilizzando il bootstrap è possibile determinare se un singolo tasso di disdetta descrive adeguatamente i quattro tipi di clienti principali. Per ulteriori informazioni, vedere l'argomento Utilizzo del bootstrap per ottenere intervalli di confidenza per proporzioni in il capitolo 3 in *IBM SPSS Bootstrapping 20*.

Nel corso di una revisione dei record dei dipendenti, la direzione desidera maggiori dettagli sulle loro precedenti esperienze lavorative. L'esperienza lavorativa è asimmetrica verso destra, caratteristica che rende la mediana preferibile alla media come stima della precedente esperienza lavorativa "tipica" tra i dipendenti. Non sono però disponibili intervalli di confidenza parametrici per la mediana nel prodotto. Per ulteriori informazioni, vedere l'argomento Utilizzo del bootstrap per ottenere intervalli di confidenza per mediane in il capitolo 3 in *IBM SPSS Bootstrapping 20*.

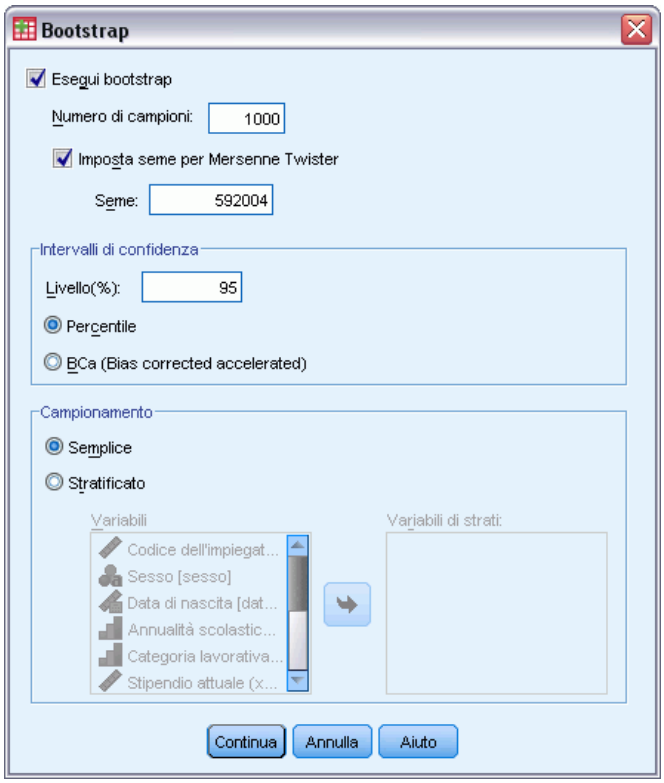
La direzione desidera inoltre determinare quali fattori sono associati agli aumenti di stipendio dei dipendenti adattando un modello lineare alla differenza tra stipendio corrente e iniziale. Nel bootstrap di un modello lineare è possibile utilizzare speciali metodi di ricampionamento (bootstrap residuale e casuale) per ottenere risultati più accurati. Per ulteriori informazioni, vedere l'argomento Utilizzo del bootstrap per scegliere predittori migliori in il capitolo 3 in *IBM SPSS Bootstrapping 20*.

Molte procedure supportano il campionamento di bootstrap e il raggruppamento dei risultati derivati dall'analisi dei campioni di bootstrap. I controlli per specificare le analisi di bootstrap sono integrati direttamente sotto forma di una normale finestra di dialogo secondaria nelle procedure che supportano il bootstrap. Poiché le impostazioni della finestra di dialogo del bootstrap sono costanti in tutte le procedure, se si esegue un'analisi Frequenze con bootstrap tra le varie finestre di dialogo, per impostazione predefinita il bootstrap viene attivato per le altre procedure che lo supportano.

Per ottenere un'analisi di bootstrap

- Dai menu, scegliere una procedura che supporti il bootstrap e fare clic su Bootstrap.

Figura 2-1
Finestra di dialogo Bootstrap



▶ Selezionare Esegui bootstrap.

Se necessario, è possibile controllare le opzioni seguenti:

Numero di campioni. Per il percentile e gli intervalli BCa prodotti, si consiglia di utilizzare almeno 1000 campioni bootstrap. Specificare un intero positivo.

Imposta seme per Mersenne Twister. Impostando un seme è possibile replicare le analisi. L'utilizzo di questo controllo è analogo all'impostazione di Mersenne Twister come generatore attivo specificando un punto di partenza fisso nella finestra di dialogo Generatori di numeri casuali, con un'importante differenza: impostando il seme in questa finestra di dialogo si conserva lo stato corrente del generatore di numeri casuali e lo si ripristina una volta terminata l'analisi.

Intervalli di confidenza. Specificare un livello di confidenza superiore a 50 e inferiore a 100. Gli intervalli di percentile utilizzano semplicemente i valori di bootstrap ordinati corrispondenti ai percentili dell'intervallo di confidenza desiderato. Per esempio, un intervallo di confidenza dei percentili del 95% utilizza il 2,5esimo e il 97,5esimo percentile dei valori di bootstrap come limite inferiore e superiore dell'intervallo (interpolando i valori di bootstrap, se necessario). Gli intervalli BCa (Bias corrected and accelerated) sono intervalli corretti che risultano più accurati, ma che richiedono più tempo per il calcolo.

Campionamento. Il metodo Semplice è un ricampionamento del caso con sostituzione dall'insieme di dati originale. Il metodo Stratificato è un ricampionamento del caso con sostituzione dall'insieme di dati originale, *all'interno* degli strati definiti dalla classificazione incrociata delle variabili di stratificazione. Il campionamento di bootstrap stratificato può rivelarsi utile quando le unità all'interno degli strati sono relativamente omogenee mentre quelle tra i vari strati sono molto diverse.

Procedure che supportano il bootstrap

Le procedure seguenti supportano il bootstrap.

Nota:

- il bootstrap non funziona con gli insiemi di dati assegnati mediante assegnazione multipla. Se nell'insieme di dati è presente una variabile *Imputation_*, la finestra di dialogo Bootstrap viene disabilitata.
- Il bootstrap utilizza l'eliminazione listwise per determinare la base di un caso; in altre parole, i casi con valori mancanti in una delle variabili di analisi vengono eliminati dall'analisi, pertanto quando il bootstrap è attivo, l'eliminazione listwise è attiva anche se la procedura di analisi specifica un altro modo per gestire i valori mancanti.

Modulo Statistics Base

Frequenze

- La tabella Statistiche supporta le stime bootstrap per media, deviazione standard, varianza, mediana, asimmetria, curtosi e percentili.
- La tabella Frequenze supporta le stime bootstrap per percentuale.

Descrittive

- La tabella Statistiche descrittive supporta le stime bootstrap per media, deviazione standard, varianza, asimmetria e curtosi.

Esplora

- La tabella Descrittive supporta le stime bootstrap per media, media 5% trim, deviazione standard, varianza, mediana, asimmetria, curtosi e distanza interquartilica.
- La tabella Stimatori M supporta le stime bootstrap per Stimatore M di Huber, Stimatore di Tukey a doppio peso, Stimatore M di Hampel e Onda di Andrew.
- La tabella Percentili supporta le stime bootstrap per percentili.

Tavole di contingenza

- La tabella Misure di direzione supporta le stime bootstrap per Lambda, Tau di Goodman e Kruskal, Coefficiente di incertezza e D di Somers.
- La tabella Misure simmetriche supporta le stime bootstrap per Phi, V di Cramer, Coefficiente di contingenza, Tau-b di Kendall, Tau-c di Kendall, Gamma, Correlazione di Spearman e R di Pearson.

- La tabella Stima del rischio supporta le stime bootstrap per il rapporto odd.
- La tabella Rapporto odd comune di Mantel-Haenszel supporta le stime bootstrap e i test di significatività per ln(Stima).

Medie

- La tabella Rapporto supporta le stime bootstrap per media, mediana, mediana dei gruppi, deviazione standard, varianza, curtosi, asimmetria, media armonica e media geometrica.

Test T per un campione

- La tabella Statistiche supporta le stime bootstrap per media e deviazione standard.
- La tabella Test supporta le stime bootstrap e i test di significatività per la differenza media.

T per campioni indipendenti

- La tabella Statistiche di gruppo supporta le stime bootstrap per media e deviazione standard.
- La tabella Test supporta le stime bootstrap e i test di significatività per la differenza media.

T per campioni appaiati

- La tabella Statistiche supporta le stime bootstrap per media e deviazione standard.
- La tabella Correlazioni supporta le stime bootstrap per le correlazioni.
- La tabella Test supporta le stime bootstrap per la media.

ANOVA univariata

- La tabella Statistiche descrittive supporta le stime bootstrap per media e deviazione standard.
- La tabella Confronti multipli supporta le stime bootstrap per la differenza media.
- La tabella Test di contrasto supporta le stime bootstrap e i test di significatività per il valore di contrasto.

GLM univariato

- La tabella Statistiche descrittive supporta le stime bootstrap per media e deviazione standard.
- La tabella Stime di parametri supporta le stime bootstrap e i test di significatività per coefficiente, B.
- La tabella Risultati del contrasto supporta le stime bootstrap e i test di significatività per la differenza.
- La tabella Medie marginali stimate: Stime supporta le stime bootstrap per la media.
- La tabella Medie marginali stimate: Confronti pairwise supporta le stime bootstrap per la differenza media.
- La tabella Test post hoc: Confronti multipli supporta le stime bootstrap per la differenza media.

Correlazioni bivariate

- La tabella Statistiche descrittive supporta le stime bootstrap per media e deviazione standard.
- La tabella Correlazioni supporta le stime bootstrap e i test di significatività per le correlazioni.

Note:

se oltre alle correlazioni di Pearson sono richieste le correlazioni non parametriche (tau-b di Kendall o Spearman), la finestra di dialogo incolla i comandi `CORRELATIONS` e `NONPAR CORR` con un comando `BOOTSTRAP` separato per ciascuno. Gli stessi campioni bootstrap verranno utilizzati per calcolare tutte le correlazioni.

Prima del raggruppamento, alle correlazioni viene applicata la trasformata *Z* di Fisher. Dopo il raggruppamento, viene applicata la trasformata *Z* inversa.

Correlazioni parziali

- La tabella Statistiche descrittive supporta le stime bootstrap per media e deviazione standard.
- La tabella Correlazioni supporta le stime bootstrap per le correlazioni.

Regressione lineare

- La tabella Statistiche descrittive supporta le stime bootstrap per media e deviazione standard.
- La tabella Correlazioni supporta le stime bootstrap per le correlazioni.
- La tabella Riepilogo del modello supporta le stime bootstrap per Durbin-Watson.
- La tabella Coefficienti supporta le stime bootstrap e i test di significatività per il coefficiente B.
- La tabella Coefficienti di correlazione supporta le stime bootstrap per le correlazioni.
- La tabella Statistiche dei residui supporta le stime bootstrap per media e deviazione standard.

Regressione ordinale

- La tabella Stime di parametri supporta le stime bootstrap e i test di significatività per coefficiente, B.

Analisi discriminante

- La tabella Coefficienti della funzione discriminante canonica standardizzata supporta le stime bootstrap per i coefficienti standardizzati.
- La tabella Coefficienti della funzione discriminante canonica supporta le stime bootstrap per i coefficienti non standardizzati.
- La tabella Coefficienti della funzione discriminante supporta le stime bootstrap per i coefficienti.

Modulo Advanced Statistics

GLM - Multivariata

- La tabella Stime di parametri supporta le stime bootstrap e i test di significatività per coefficiente, B.

Modelli misti lineari

- La tabella Stime degli effetti fissi supporta le stime bootstrap e i test di significatività per la stima.
- La tabella Stime dei parametri di covarianza supporta le stime bootstrap e i test di significatività per la stima.

Modelli lineari generalizzati

- La tabella Stime di parametri supporta le stime bootstrap e i test di significatività per coefficiente, B.

Regressione di Cox

- La tabella Variabili nella tabella Equazione supporta le stime bootstrap e i test di significatività per coefficiente, B.

Modulo Regression

Regressione logistica binaria

- La tabella Variabili nella tabella Equazione supporta le stime bootstrap e i test di significatività per coefficiente, B.

Regressione logistica multinomiale

- La tabella Stime di parametri supporta le stime bootstrap e i test di significatività per coefficiente, B.

Opzioni aggiuntive del comando BOOTSTRAP

Il linguaggio della sintassi dei comandi consente inoltre di:

- Eseguire il campionamento bootstrap residuale e casuale (sottocomando `SAMPLING`)

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

Parte II: Esempi

Capitolo
3

Bootstrapping

Il bootstrap è un metodo utilizzato per derivare delle stime affidabili su errori standard e intervalli di confidenza per stime quali media, mediana, proporzione, rapporto odd, coefficiente di correlazione o coefficiente di regressione. Può essere utilizzato anche per creare dei test di ipotesi. Il bootstrap è particolarmente utile come alternativa alle stime parametriche in caso di dubbio sulle supposizioni di questi metodi (come nel caso dei modelli di regressione con residui eteroschedastici adattati a campioni di piccole dimensioni) o quando l'inferenza parametrica è impossibile o richiede formule molto complicate per il calcolo degli errori standard (come nel caso del calcolo degli intervalli di confidenza per la mediana, i quartili e altri percentili).

Utilizzo del bootstrap per ottenere intervalli di confidenza per proporzioni

Un'azienda di telecomunicazioni perde ogni mese circa il 27% dei suoi clienti a causa della disdetta di un servizio da parte del cliente. Per analizzare correttamente gli sforzi da compiere al fine di ridurre la disdetta dei servizi, la direzione desidera conoscere se questa percentuale varia in gruppi di clienti predefiniti.

Tali informazioni vengono raccolte nel file *telco.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 31. Utilizzare il bootstrap per determinare se un singolo tasso di disdetta descrive adeguatamente i quattro tipi di clienti principali.

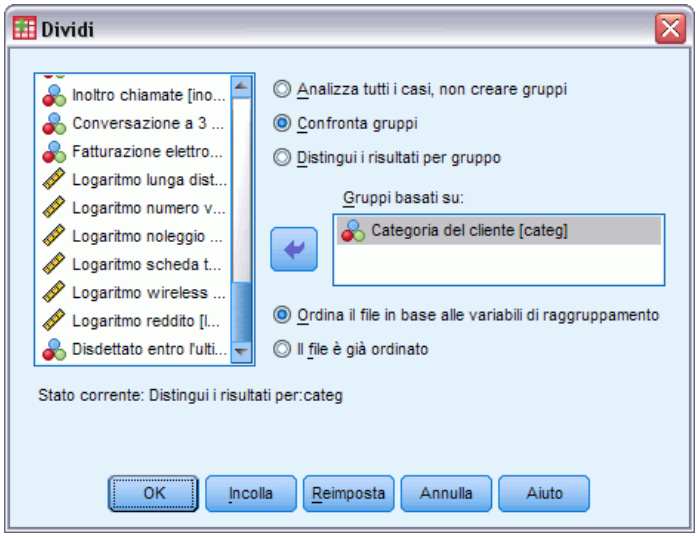
Nota: questo esempio utilizza la procedura Frequenze e richiede il modulo Statistics Base.

Preparazione dei dati

Per prima cosa è necessario suddividere il file per *Categoria del cliente*.

- Per suddividere il file, dai menu dell'Editor dei dati scegliere:
  Dati > Dividi...

Figura 3-1
Finestra di dialogo Dividi

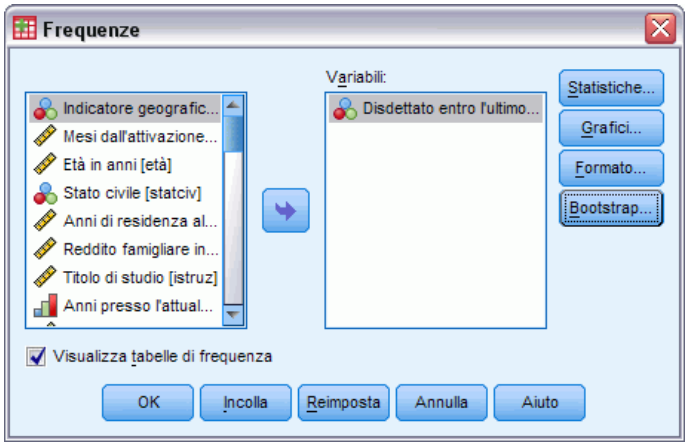


- Selezionare Confronta gruppi.
- Selezionare *Categoria del cliente* come variabile su cui basare i gruppi.
- Fare clic su OK.

Esecuzione dell'analisi

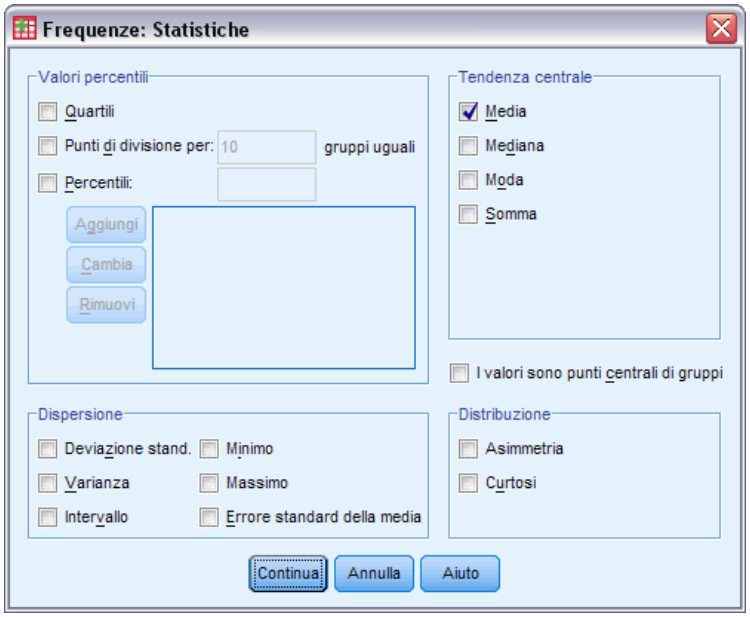
- Per ottenere gli intervalli di confidenza di bootstrap per le proporzioni, dai menu scegliere: Analizza > Statistiche descrittive > Frequenze...

Figura 3-2
Finestra di dialogo principale Frequenze



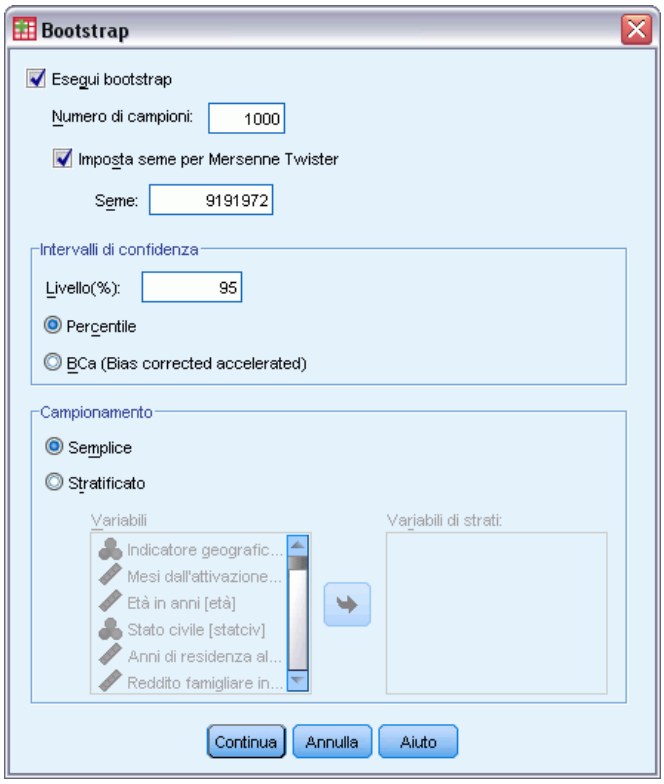
- Selezionare *Disdettato entro l'ultimo mese [churn]* come variabile nell'analisi.
- Fare clic su Statistiche.

Figura 3-3
Finestra di dialogo Statistiche



- Selezionare Media nel gruppo Tendenza centrale.
- Fare clic su Continua.
- Fare clic su Bootstrap nella finestra di dialogo Frequenze.

Figura 3-4
Finestra di dialogo Bootstrap



- Selezionare Esegui bootstrap.
- Per replicare esattamente i risultati di questo esempio, selezionare Imposta seme per Mersenne Twister e digitare 9191972 come seme.
- Fare clic su Continua.
- Fare clic su OK nella finestra di dialogo Frequenze.

Tali selezioni generano la sintassi seguente:

```
SORT CASES  BY custcat.
SPLIT FILE LAYERED BY custcat.
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES INPUT=churn
  /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
FREQUENCIES VARIABLES=churn
  /STATISTICS=MEAN
  /ORDER=ANALYSIS.
RESTORE.
```

- I comandi `SORT CASES` e `SPLIT FILE` suddividono il file in base alla variabile *custcat*.

- I comandi `PRESERVE` e `RESTORE` memorizzano lo stato corrente del generatore di numeri casuali e ripristinano tale stato del sistema al termine del bootstrap.
- Il comando `SET` imposta il generatore di numeri casuali sul Mersenne Twister e l'indice su 9191972, consentendo la replica esatta dei risultati del bootstrap. Il comando `SHOW` visualizza l'indice nell'output per riferimento.
- Il comando `BOOTSTRAP` richiede 1,000 campioni di bootstrap con un semplice ricampionamento.
- La variabile *churn* viene utilizzata per determinare la base di casi per il ricampionamento. I record con valori mancanti in base a questa variabile vengono eliminati dall'analisi.
- La procedura `FREQUENCIES` successiva a `BOOTSTRAP` viene eseguita su ognuno dei campioni di bootstrap.
- Il sottocomando `STATISTICS` genera la media per la variabile *churn* in base ai dati originali. Inoltre, vengono prodotte statistiche raggruppate per la media e le percentuali nella tabella di frequenza.

Specifiche di bootstrap

Figura 3-5
Specifiche di bootstrap

Metodo di campionamento	Semplice
Numero di campioni	1000
Livello dell'intervallo di confidenza	95.0%
Tipo di intervallo di confidenza	Percentile

La tabella Specifiche di bootstrap contiene le impostazioni utilizzate durante il ricampionamento e costituisce un utile riferimento per verificare se l'analisi desiderata è stata eseguita.

Statistiche

Figura 3-6
Tabella Statistiche con intervallo di confidenza di bootstrap per la proporzione

Disdettato entro l'ultimo mese

Categoria del cliente			Statistic	Bootstrap[a]			
				Distorsione	Deviazione standard Errore	Intervallo di confidenza 95%	
						Inferiore	Superiore
Servizio base	N	Validi	266	0	0	266	266
		Mancanti	0	0	0	0	0
	Media		.31	.00	.03	.26	.37
E-service	N	Validi	217	0	0	217	217
		Mancanti	0	0	0	0	0
	Media		.27	.00	.03	.21	.34
Servizio plus	N	Validi	281	0	0	281	281
		Mancanti	0	0	0	0	0
	Media		.16	.00	.02	.12	.20
Servizio completo	N	Validi	236	0	0	236	236
		Mancanti	0	0	0	0	0
	Media		.37	.00	.03	.31	.44

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

La tabella Statistiche mostra, per ogni livello di *Categoria del cliente*, il valore medio di *Disdettato entro l'ultimo mese*. Poiché *Disdettato entro l'ultimo mese* accetta solo i valori 0 e 1, dove 1 indica un cliente che ha dato disdetta, la media è uguale alla proporzione dei clienti che hanno dato disdetta. La colonna Statistiche mostra i valori normalmente prodotti da Frequenze con l'insieme di dati originale. Le colonne Bootstrap vengono prodotte dagli algoritmi di bootstrap.

- Distorsione è la differenza tra il valore medio di tali statistiche nei campioni di bootstrap e il valore nella colonna Statistiche. In questo caso, il valore medio di *Disdettato entro l'ultimo mese* viene calcolato per tutti i 1000 campioni di bootstrap, dopo di che viene calcolata la medie di tali medie.
- Errore std. è l'errore standard del valore medio di *Disdettato entro l'ultimo mese* per i 1000 campioni di bootstrap.
- Il limite inferiore dell'intervallo di confidenza di bootstrap del 95% è un'interpolazione del 25esimo e del 26esimo valore medio di *Disdettato entro l'ultimo mese*, se i 1000 campioni di bootstrap sono elencati in ordine crescente. Il limite superiore è un'interpolazione del 975esimo e del 976esimo valore medio.

I risultati nella tabella indicano che il tasso di disdetta differisce in base ai tipi di clienti. In particolare, l'intervallo di confidenza per i clienti *Servizio Plus* non si sovrappone ad altri, il che indica che questi clienti hanno in media meno probabilità di disdetta.

Quando si utilizzano variabili categoriali con due soli valori, questi intervalli di confidenza sono alternativi rispetto a quelli prodotti dalla procedura Test non parametrici a campione singolo o dalla procedura Test T per un campione.

Tabella di frequenza (Analisi delle corrispondenze)

Figura 3-7
Tabella di frequenza con intervallo di confidenza di bootstrap per la proporzione

Categoria del cliente			Frequenza	Percentuale	Percentuale valida	Percentuale cumulata	Bootstrap per Percentuale[a]			
							Distorsione	Deviazione standard Errore	Intervallo di confidenza 95% Inferiore	Intervallo di confidenza 95% Superiore
Servizio base	Validi	No	183	68.8	68.8	68.8	.0	2.8	63.2	74.4
		Sì	83	31.2	31.2	100.0	.0	2.8	25.6	36.8
		Totale	266	100.0	100.0		.0	.0	100.0	100.0
E-service	Validi	No	158	72.8	72.8	72.8	.1	3.1	66.4	78.8
		Sì	59	27.2	27.2	100.0	-.1	3.1	21.2	33.6
		Totale	217	100.0	100.0		.0	.0	100.0	100.0
Servizio plus	Validi	No	237	84.3	84.3	84.3	.0	2.1	80.1	88.3
		Sì	44	15.7	15.7	100.0	.0	2.1	11.7	19.9
		Totale	281	100.0	100.0		.0	.0	100.0	100.0
Servizio completo	Validi	No	148	62.7	62.7	62.7	.0	3.2	56.4	69.1
		Sì	88	37.3	37.3	100.0	.0	3.2	30.9	43.6
		Totale	236	100.0	100.0		.0	.0	100.0	100.0

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

La tabella di frequenza mostra gli intervalli di confidenza per le percentuali (proporzione × 100%) per ogni categoria ed è quindi disponibile per tutte le variabili categoriali. Non sono disponibili intervalli di confidenza confrontabili in nessun'altra area del prodotto.

Utilizzo del bootstrap per ottenere intervalli di confidenza per mediane

Nel corso di una revisione dei record dei dipendenti, la direzione desidera maggiori dettagli sulle loro precedenti esperienze lavorative. L'esperienza lavorativa è asimmetrica verso destra, caratteristica che rende la mediana preferibile alla media come stima della precedente esperienza lavorativa "tipica" tra i dipendenti. Tuttavia, senza il bootstrap, non sono generalmente disponibili intervalli di confidenza per la mediana nelle procedure statistiche del prodotto.

Tali informazioni vengono raccolte nel file *Employee data.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 31. Utilizzo del bootstrap per ottenere intervalli di confidenza per la mediana.

Nota: questo esempio utilizza la procedura Esplora e richiede il modulo Statistics Base.

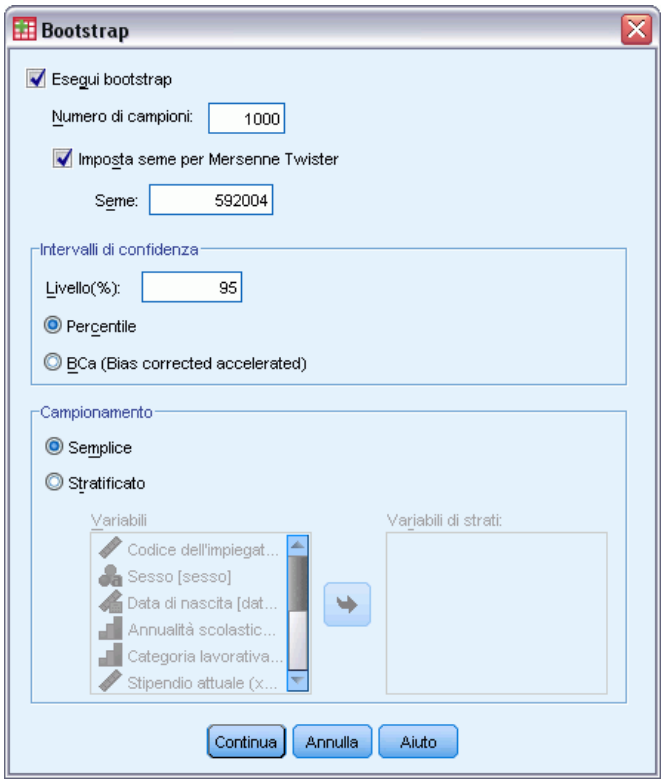
Esecuzione dell'analisi

► Per ottenere gli intervalli di confidenza di bootstrap per la mediana, dai menu scegliere:
Analizza > Statistiche descrittive > Esplora...

Figura 3-8
Finestra di dialogo principale Esplora

- Selezionare *Mesi di lavoro precedenti [prevexp]* come variabile dipendente.
- Selezionare Statistiche nel gruppo Visualizza.
- Fare clic su Bootstrap.

Figura 3-9
Finestra di dialogo Bootstrap



- Selezionare Esegui bootstrap.
- Per replicare esattamente i risultati di questo esempio, selezionare Imposta seme per Mersenne Twister e digitare 592004 come seme.
- Per ottenere intervalli più accurati (anche se è necessario un tempo di elaborazione più lungo), selezionare BCa (Bias corrected accelerated).
- Fare clic su Continua.
- Nella finestra di dialogo Esplora, fare clic su OK.

Tali selezioni generano la sintassi seguente:

```
PRESERVE.
SET RNG=MT MTINDEX=592004.
SHOW RNG.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES TARGET=prevexp
  /CRITERIA CILEVEL=95 CITYPE=BCA  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
EXAMINE VARIABLES=prevexp
  /PLOT NONE
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

```
RESTORE.
```

- I comandi `PRESERVE` e `RESTORE` memorizzano lo stato corrente del generatore di numeri casuali e ripristinano tale stato del sistema al termine del bootstrap.
- Il comando `SET` imposta il generatore di numeri casuali sul Mersenne Twister e l'indice su 592004, consentendo la replica esatta dei risultati del bootstrap. Il comando `SHOW` visualizza l'indice nell'output per riferimento.
- Il comando `BOOTSTRAP` richiede 1000 campioni di bootstrap con un semplice ricampionamento.
- Il sottocomando `VARIABLES` specifica che viene utilizzata la variabile *prevexp* per determinare la base di casi per il ricampionamento. I record con valori mancanti in base a questa variabile vengono eliminati dall'analisi.
- Il sottocomando `CRITERIA`, oltre a richiedere il numero di campioni di bootstrap, richiede gli intervalli di confidenza di bootstrap BCa (Bias corrected and accelerated) anziché gli intervalli di percentile predefiniti.
- La procedura `EXAMINE` successiva a `BOOTSTRAP` viene eseguita su ognuno dei campioni di bootstrap.
- Il sottocomando `PLOT` disattiva l'output dei grafici.
- Tutte le altre opzioni sono impostate sui valori predefiniti.

Descrittive

Figura 3-10
Tabella Descrittive con intervalli di confidenza di bootstrap

					Bootstrap[a]			
							Intervallo di confidenza 95%	
			Statistica	Errore std.	Distorsione	Errore std.	Inferiore	Superiore
Mesi di lavoro precedenti	Media		95.86	4.804	-.01	4.86	86.29	105.41
	Intervallo di confidenza per la media al 95%	Limite inferiore	86.42					
		Limite superiore	105.30					
	Media 5% trim		84.64		.02	4.94	75.25	94.59
	Mediana		55.00		-.11	3.66	48.00	64.00
	Varianza		10938.281		18.783	977.081	8981.729	12908.885
	Deviazione std.		104.586		-.015	4.689	94.772	113.617
	Minimo		0					
	Massimo		476					
	Intervallo		476					
	Distanza interquartilica		121		-1	10	101	142
	Asimmetria		1.510	.112	.006	.110	1.290	1.751
	Curtosi		1.696	.224	.040	.463	.881	2.774

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

La tabella Descrittive contiene varie statistiche e gli intervalli di confidenza di bootstrap per tali statistiche. L'intervallo di confidenza di bootstrap per la media (86,39, 105,20) è simile all'intervallo di confidenza parametrico (86,42, 105,30) e indica che il dipendente "tipico" ha circa 7-9 anni di esperienza precedente. Tuttavia, *Mesi di lavoro precedenti* presenta una distribuzione asimmetrica, che rende la mediana preferibile alla media come indicatore dello stipendio attuale "tipico". L'intervallo di confidenza di bootstrap per la mediana (50,00, 60,00) è più stretto e

inferiore in valore dell'intervallo di confidenza della media e indica che il dipendente "tipico" ha circa 4-5 anni di esperienza precedente. L'utilizzo del bootstrap ha reso possibile ottenere un intervallo di valori che meglio rappresentano l'esperienza precedente tipica.

Utilizzo del bootstrap per scegliere predittori migliori

Nel corso di una revisione dei record dei dipendenti, la direzione desidera determinare quali fattori sono associati agli aumenti di stipendio dei dipendenti adattando un modello lineare alla differenza tra stipendio attuale e iniziale. Nel bootstrap di un modello lineare è possibile utilizzare speciali metodi di ricampionamento (bootstrap residuale e casuale) per ottenere risultati più accurati.

Tali informazioni vengono raccolte nel file *Employee data.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 31.

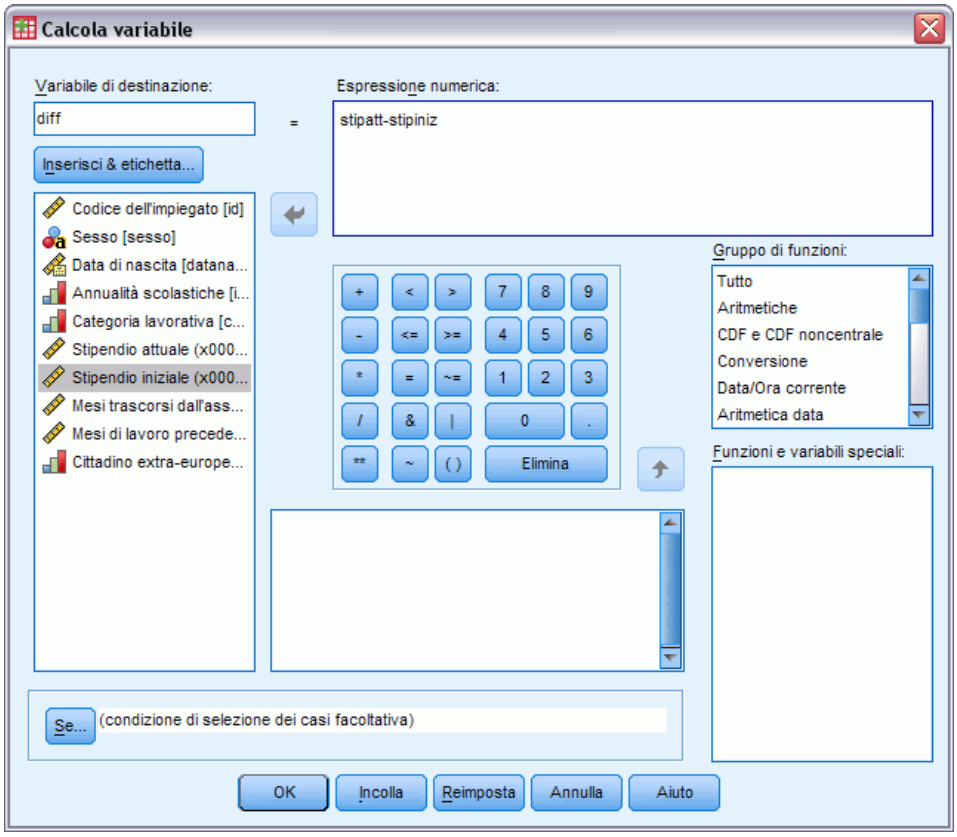
Nota: questo esempio utilizza la procedura GLM univariato e richiede il modulo Statistics Base.

Preparazione dei dati

Prima di tutto è necessario calcolare la differenza tra Stipendio attuale e Stipendio iniziale.

- Dai menu, scegliere:
  Trasforma > Calcola variabile...

Figura 3-11
Finestra di dialogo Calcola variabile



- Digitare diff come variabile di destinazione.
- Digitare salary-salbegin come espressione numerica.
- Fare clic su OK.

Esecuzione dell'analisi

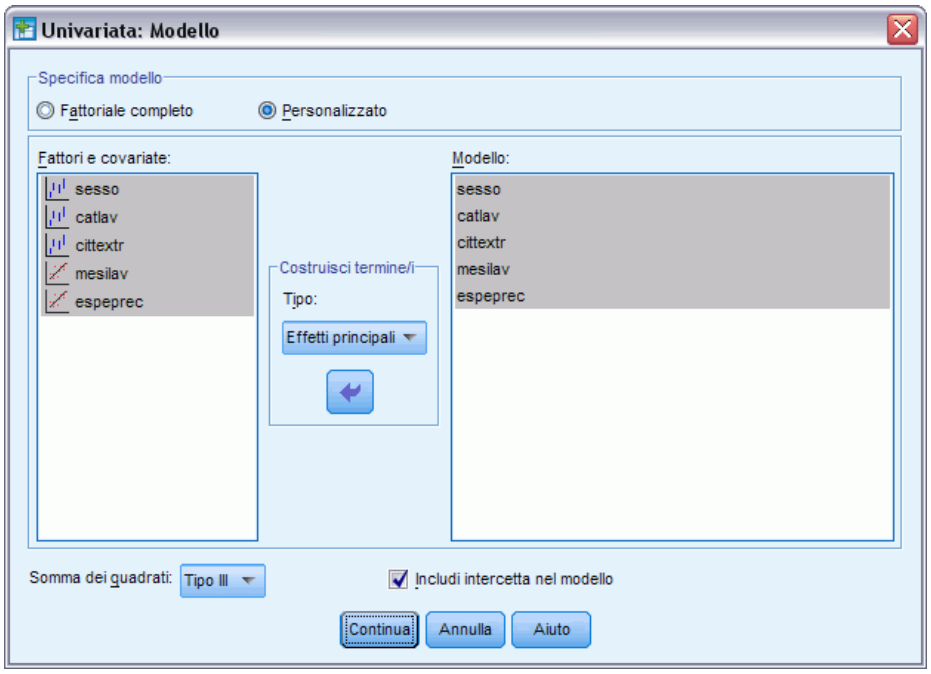
Per eseguire GLM univariato con il bootstrap casuale e residuale, occorre innanzitutto creare residui.

- Dai menu, scegliere:
  Analizza > Modello lineare generalizzato > Univariata...

Figura 3-12
Finestra di dialogo principale GLM univariato

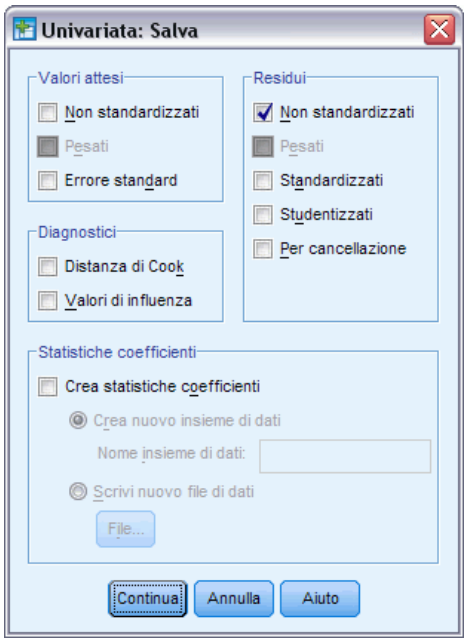
- Selezionare *diff* come variabile dipendente.
- Selezionare *Sesso [gender]*, *Categoria lavorativa [jobcat]* e *Cittadino extra-europeo [minority]* come fattori fissi.
- Selezionare *Mesi di servizio [jobtime]* e *Mesi di lavoro precedente [prevexp]* come covariate.
- Fare clic su Modello.

Figura 3-13
Finestra di dialogo Modello



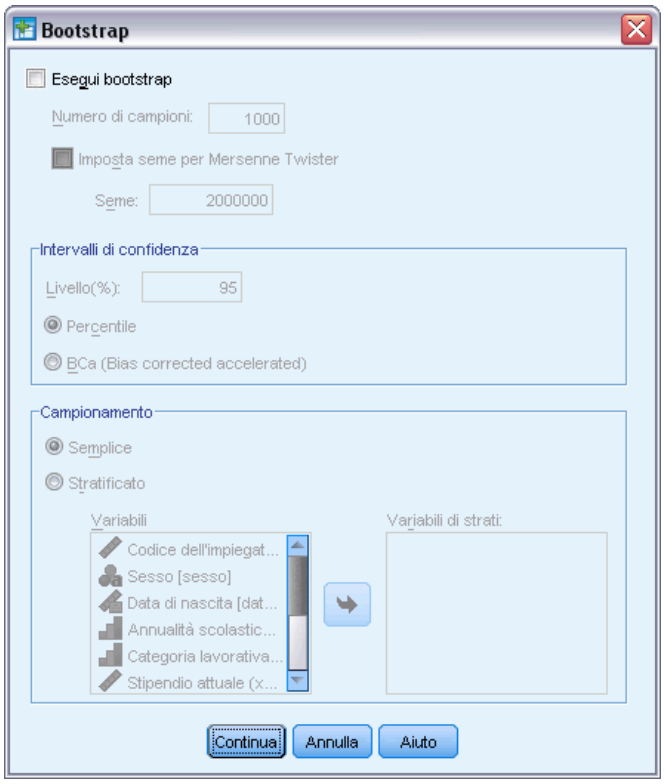
- Selezionare Personalizzato, quindi selezionare Effetti principali dall'elenco a discesa Costruisci termini.
- Selezionare da *gender* a *prevexp* come termini del modello.
- Fare clic su Continua.
- Nella finestra di dialogo GLM univariato fare clic su Salva.

Figura 3-14
Salva



- Selezionare Non standardizzati nel gruppo Residui.
- Fare clic su Continua.
- Nella finestra di dialogo GLM univariato fare clic su Bootstrap.

Figura 3-15
Finestra di dialogo Bootstrap

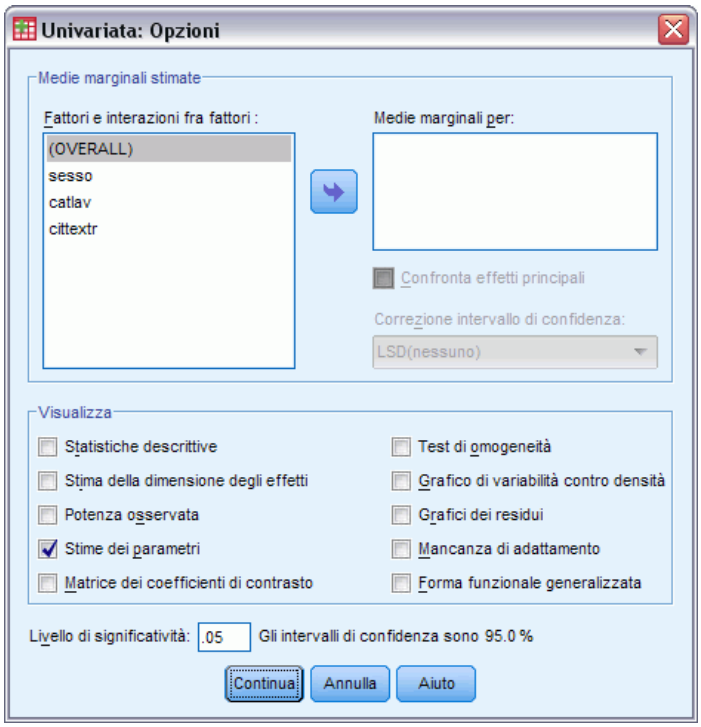


Le impostazioni del bootstrap sono costanti in tutte le finestre di dialogo che supportano il bootstrap. Il salvataggio di nuove variabili nell'insieme di dati non è supportato finché il bootstrap è attivo e, pertanto, è necessario verificare che sia disattivato.

- Se necessario, deselezionare Esegui bootstrap.
- Nella finestra di dialogo GLM univariato fare clic su OK. L'insieme di dati contiene ora una nuova variabile, *RES_1*, che contiene i residui non standardizzati di questo modello.
- Richiamare la finestra di dialogo GLM univariato fare clic su Salva.

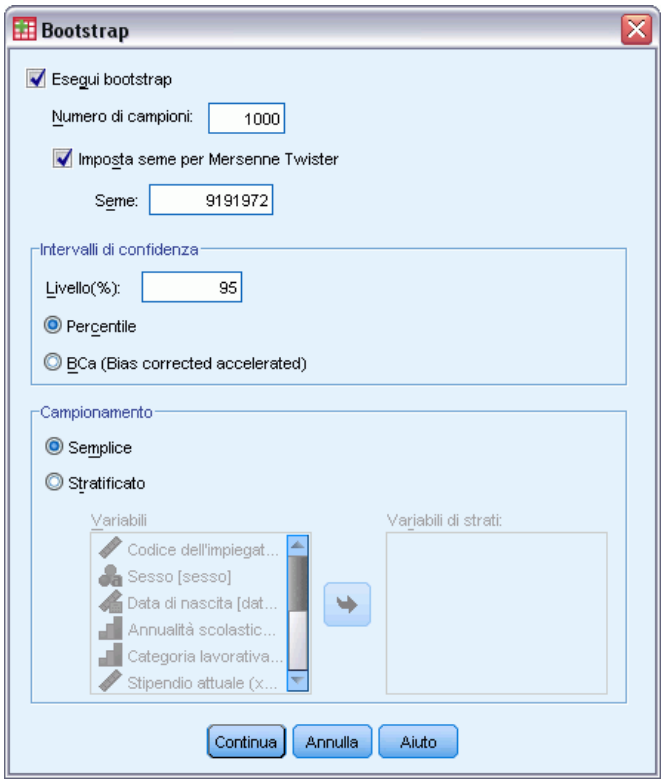
- Deselezionare Non standardizzati, quindi fare clic su Continua e su Opzioni nella finestra di dialogo GLM univariato.

Figura 3-16
Finestra di dialogo Opzioni



- Selezionare Stime dei parametri nel gruppo Visualizza.
- Fare clic su Continua.
- Nella finestra di dialogo GLM univariato fare clic su Bootstrap.

Figura 3-17
Finestra di dialogo Bootstrap



- Selezionare Esegui bootstrap.
- Per replicare esattamente i risultati di questo esempio, selezionare Imposta seme per Mersenne Twister e digitare 9191972 come seme.
- Poiché non sono disponibili opzioni per l'esecuzione del bootstrap casuale nelle finestra di dialogo, fare clic su Continua, quindi su Incolla nella finestra di dialogo GLM univariato.

Tali selezioni generano la sintassi seguente:

```
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
 /SAMPLING METHOD=SIMPLE
 /VARIABLES TARGET=diff INPUT=gender jobcat minority jobtime prevexp
 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
 /MISSING USERMISSING=EXCLUDE.
UNIANOVA diff BY gender jobcat minority WITH jobtime prevexp
 /METHOD=SSTYPE(3)
 /INTERCEPT=INCLUDE
 /PRINT=PARAMETER
 /CRITERIA=ALPHA(.05)
 /DESIGN=gender jobcat minority jobtime prevexp.
```

```
RESTORE.
```

Per eseguire il campionamento bootstrap casuale, modificare la parola chiave `METHOD` del sottocomando `SAMPLING` in modo da specificare `METHOD=WILD(RESIDUALS=RES_1)`.

L'insieme "finale" della sintassi dei comandi ha un aspetto simile al seguente:

```
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
 /SAMPLING METHOD=WILD(RESIDUALS=RES_1)
 /VARIABLES TARGET=diff INPUT=gender jobcat minority jobtime prevexp
 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
 /MISSING USERMISSING=EXCLUDE.
UNIANOVA diff BY gender jobcat minority WITH jobtime prevexp
 /METHOD=SSTYPE(3)
 /INTERCEPT=INCLUDE
 /PRINT=PARAMETER
 /CRITERIA=ALPHA(.05)
 /DESIGN=gender jobcat minority jobtime prevexp.
RESTORE.
```

- I comandi `PRESERVE` e `RESTORE` memorizzano lo stato corrente del generatore di numeri casuali e ripristinano tale stato del sistema al termine del bootstrap.
- Il comando `SET` imposta il generatore di numeri casuali sul Mersenne Twister e l'indice su 9191972, consentendo la replica esatta dei risultati del bootstrap. Il comando `SHOW` visualizza l'indice nell'output per riferimento.
- Il comando `BOOTSTRAP` richiede 1000 campioni di bootstrap utilizzando il campionamento casuale e *RES_1* come variabile contenente i residui.
- Il sottocomando `VARIABLES` specifica che *diff* è la variabile di destinazione nel modello lineare; questa e le variabili *gender*, *jobcat*, *minority*, *jobtime* e *prevexp* vengono utilizzate per determinare la base di casi per il ricampionamento. I record con valori mancanti in base a queste variabili vengono eliminati dall'analisi.
- Il sottocomando `CRITERIA`, oltre a richiedere il numero di campioni di bootstrap, richiede gli intervalli di confidenza di bootstrap BCa (Bias corrected and accelerated) anziché gli intervalli di percentile predefiniti.
- La procedura `UNIANOVA` successiva a `BOOTSTRAP` viene eseguita su ognuno dei campioni di bootstrap e produce stime dei parametri per i dati originali. Inoltre, vengono prodotte statistiche raggruppate per i coefficienti del modello.

Stime di parametri

Figura 3-18
Stime dei parametri

Variabile dipendente:diff

Parametro	B	Deviazione standard Errore	t	Sig.	Intervallo di confidenza 95%	
					Limite inferiore	Limite superiore
Intercetta	22789.014	2920.700	7.803	.000	17049.673	28528.355
[sesso=f]	-4085.253	726.416	-5.624	.000	-5512.701	-2657.804
[sesso=m]	0[a]	.	.	.	.	.
[catlav=1]	-17717.706	939.798	-18.853	.000	-19564.463	-15870.949
[catlav=2]	-13101.918	1780.683	-7.358	.000	-16601.061	-9602.776
[catlav=3]	0[a]	.	.	.	.	.
[cittextr=0]	1332.363	819.349	1.626	.105	-277.705	2942.431
[cittextr=1]	0[a]	.	.	.	.	.
mesilav	145.539	32.586	4.466	.000	81.505	209.572
espeprec	-21.423	3.575	-5.993	.000	-28.447	-14.398

a.Questo parametro viene messo a zero perché è ridondante.

La tabella Stime dei parametri mostra le usuali stime non bootstrap dei parametri per i termini del modello. Il valore di significatività di 0,105 per *[minority=0]* è maggiore di 0,05, a indicare che *Cittadino extra-europeo* non ha effetto sugli aumenti di stipendio.

Figura 3-19
Stime dei parametri bootstrap

Variabile dipendente:diff

Parametro	B	Bootstrap[a]				
		Distorsione	Deviazione standard Errore	Sign. (a due code)	Intervallo di confidenza 95%	
					Inferiore	Superiore
Intercetta	22789.014	-95.084	3280.762	.001	16079.630	28835.063
[sesso=f]	-4085.253	32.480	622.971	.001	-5365.321	-2892.131
[sesso=m]	0	0	0		0	0
[catlav=1]	-17717.706	46.324	1454.230	.001	-20671.451	-14889.507
[catlav=2]	-13101.918	47.958	1753.311	.001	-16658.596	-9671.891
[catlav=3]	0	0	0		0	0
[cittextr=0]	1332.363	-10.592	651.144	.012	57.831	2642.534
[cittextr=1]	0	0	0		0	0
mesilav	145.539	.707	35.285	.001	79.081	217.761
espeprec	-21.423	-.065	2.859	.001	-27.533	-16.055

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Si esamini ora la tabella Bootstrap per stime dei parametri. Nella colonna Errore std. gli errori standard parametrici per alcuni coefficienti, come l'intercetta, sono troppo piccoli rispetto alle stime bootstrap e quindi gli intervalli di confidenza sono più ampi. Per alcuni coefficienti, come *[minority=0]*, gli errori standard parametrici sono troppo grandi mentre il valore di significatività di 0,006 riportato nei risultati di bootstrap, minore di 0,05, mostra che la differenza osservata negli aumenti di stipendio tra dipendenti che sono e non sono cittadini extra-europei non è casuale. L'esecutivo aziendale adesso sa che tale differenza richiede ulteriore analisi per determinare le possibili cause.

Letture consigliate

Per ulteriori informazioni sul bootstrap, consultare i seguenti testi:

Davison, A. C., e D. V. Hinkley. 2006. *Bootstrap Methods and their Application*. : Cambridge University Press.

Shao, J., e D. Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer.

Appendice

A

File di esempio

Il file di esempio installato con il prodotto si trova nella sottodirectory *Samples* della directory di installazione. La sottodirectory Samples contiene cartelle separate per ciascuna delle seguenti lingue: Inglese, Francese, Tedesco, Italiano, Giapponese, Coreano, Polacco, Russo, Cinese semplificato, Spagnolo e Cinese tradizionale.

Non tutti i file di esempio sono disponibili in tutte le lingue. Se un file di esempio non è disponibile in una lingua, la cartella di tale lingua contiene una versione inglese del file.

Descrizioni

Questa sezione contiene brevi descrizioni dei file di esempio utilizzati negli esempi riportati in tutta la documentazione.

- **accidents.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nello studio dei fattori di rischio correlati all'età e al sesso per gli incidenti automobilistici che si verificano in una determinata regione. Ciascun caso corrisponde a una classificazione incrociata della categoria relativa età e del sesso.
- **adl.sav.** File di dati ipotetici che prende in esame l'impegno richiesto per determinare i vantaggi di un tipo di terapia proposto per i pazienti con problemi di cuore. I medici hanno assegnato in modo casuale i pazienti con problemi di cuore di sesso femminile a uno di due gruppi. Al primo gruppo è stata assegnata la terapia fisica standard; al secondo gruppo, un'ulteriore terapia di supporto psicologico. Dopo tre mesi di trattamenti, a ciascuna capacità dei pazienti che consente di riprendere le normali attività giornaliere è stato assegnato un punteggio come variabile ordinale.
- **advert.sav.** File di dati ipotetici che prende in esame l'impegno di un rivenditore al dettaglio che desidera esaminare la relazione tra il denaro speso per la pubblicità e le vendite risultanti. Finora sono stati raccolti i dati delle vendite precedenti e i relativi costi pubblicitari.
- **aflatoxin.sav.** File di dati ipotetici che prende in esame il test di raccolti di mais con presenza di Aflatossina, un veleno la cui concentrazione varia notevolmente nei raccolti. Una macchina per la lavorazione dei cereali ha ricevuto 16 campioni da ciascuno degli otto raccolti di mais e ha misurato i livelli di Aflatossina in parti per miliardo (PPB).
- **anorectic.sav.** Per trovare una sintomatologia standardizzata del comportamento anoressico/bulimico, i ricercatori (Van der Ham, Meulman, Van Strien, e Van Engeland, 1997) hanno condotto uno studio basato su 55 adolescenti affetti da disordini alimentari conosciuti. Ogni paziente è stato visitato quattro volte in quattro anni, per un totale di 220 visite. Durante ogni visita, ai pazienti sono stati assegnati punteggi per ciascuno dei 16 sintomi. I punteggi relativi ai sintomi sono assenti per il paziente 71 alla visita 2, il paziente 76 alla visita 2 e il paziente 47 alla visita 3, con 217 osservazioni valide.
- **bankloan.sav.** File di dati ipotetici che prende in esame l'impegno di una banca nel tentativo di ridurre il tasso di inadempienza nel rimborso di un prestito. Il file contiene informazioni finanziarie e demografiche su 850 vecchi e potenziali clienti. I primi 700 casi riguardano i

clienti a cui sono stati concessi dei prestiti precedentemente. Gli ultimi 150 casi riguardano i potenziali clienti che la banca deve classificare come rischi di credito positivi o negativi.

- **bankloan_binning.sav.** File di dati ipotetici che contiene informazioni finanziarie e demografiche su 5000 vecchi clienti.
- **behavior.sav.** In un classico esempio (Prezzo e Bouffard, 1974), è stato chiesto a 52 studenti di classificare una combinazione di 15 situazioni e 15 comportamenti utilizzando una scala da 0="molto appropriato" a 9="molto inadeguato". I valori medi riferiti ai partecipanti sono stati considerati dissimilarità.
- **behavior_ini.sav.** Questo file di dati contiene la configurazione iniziale di una soluzione a due dimensioni per *behavior.sav*.
- **brakes.sav.** File di dati ipotetici che prende in esame il controllo di qualità di un'industria che produce freni a disco per automobili con elevate prestazioni. Il file di dati contiene le misurazioni del diametro di 16 dischi da ciascuna delle otto macchine di produzione. L'obiettivo finale è ottenere un diametro dei dischi pari a 322 millimetri.
- **breakfast.sav.** In uno studio classico (Green e Rao, 1972), è stato chiesto a 21 studenti MBA della Wharton School e ai loro consorti di classificare 15 cibi da colazione in ordine di preferenza, dove il valore 1 corrispondeva all'alimento preferito in assoluto e il valore 15 a quello meno preferito. Le loro preferenze sono state registrate per sei diversi scenari, che comprendevano tutti gli scenari compresi tra "Preferenza generale" e "Solo snack con bibita".
- **breakfast-overall.sav.** Questo file contiene le preferenze degli alimenti della colazione solo per il primo scenario, "Preferenza generale".
- **broadband_1.sav.** File di dati ipotetici che contiene il numero di sottoscrittori, per area, di un provider di servizi a banda larga nazionale. Il file di dati contiene il numero dei sottoscrittori mensili di 85 aree in un periodo di quattro anni.
- **broadband_2.sav.** Questo file è identico al file *broadband_1.sav*, ma contiene i dati per ulteriori tre mesi.
- **car_insurance_claims.sav.** Un insieme di dati presentato e analizzato altrove (McCullagh e Nelder, 1989) riguarda le richieste di risarcimento auto. La quantità media di richieste di risarcimento può essere adattata come avente una distribuzione gamma, utilizzando una funzione di collegamento inverso per correlare la media della variabile dipendente a una combinazione lineare di età del contraente della polizza e tipo e anni del veicolo. Il numero delle richieste di risarcimento specificato può essere utilizzato come peso scalato.
- **car_sales.sav.** Questo file di dati ipotetici contiene le stime sulle vendite, i prezzi di listino e le specifiche fisiche di numerose marche e modelli di veicoli. I prezzi di listino e le specifiche fisiche sono state ottenute dal sito *edmunds.com* e dai siti dei produttori.
- **car_sales_uprepared.sav.** Questa è una versione modificata di *car_sales.sav* che non comprende versioni trasformate dei campi.
- **carpet.sav.** Come esempio tipico (Green e Wind, 1973), un'azienda interessata alla commercializzazione di un nuovo battitappeto desidera esaminare l'influenza di cinque fattori sulle preferenze del consumatore, ovvero design della confezione, marca, prezzo, la presenza di un *marchio di qualità* e una garanzia "Soddisfatti o rimborsati". Esistono tre livelli di fattore per il design della confezione, che differiscono per la posizione della spazzola dell'applicatore; tre marchi (*K2R*, *Glory* e *Bissell*); tre livelli di prezzo e due livelli (no o sì) per ciascuno degli ultimi due fattori. Dieci consumatori sono classificati in 22 profili definiti

da questi fattori. La variabile *Preferenza* include il rango delle classificazioni medie per ogni profilo. Classificazioni basse corrispondono a una preferenza elevata. La variabile riflette una misura globale della preferenza per ogni profilo.

- **carpet_prefs.sav.** Questo file di dati si basa sullo stesso esempio del file *carpet.sav*, ma contiene le classificazioni effettive raccolte da ciascuno dei 10 clienti. Ai clienti è stato chiesto di classificare 22 profili di prodotti in ordine di preferenza. Le variabili da *PREF1* a *PREF22* contengono gli ID dei profili associati, come definito nel file *carpet_plan.sav*.
- **catalog.sav.** File di dati ipotetico che contiene le cifre sulle vendite mensili di tre prodotti venduti da una società di vendita per corrispondenza. Il file include anche i dati di cinque possibili variabili predittore.
- **catalog_seasfac.sav.** Questo file di dati è uguale al file *catalog.sav* con l'eccezione che contiene un insieme di fattori stagionali calcolati dalla procedura Decomposizionale stagionale insieme a variabili di dati.
- **cellular.sav.** File di dati ipotetici che prende in esame l'impegno di un'azienda di telefonia cellulare nel tentativo di ridurre il churn, ovvero l'abbandono dei clienti. Agli account vengono applicati i punteggi relativi alla propensione al churn, con valori compresi tra 0 e 100. Gli account con punteggio pari a 50 o superiore è probabile che stiano cercando nuovi provider.
- **ceramics.sav.** File di dati ipotetici che prende in esame l'impegno di un produttore che desidera stabilire se una nuova lega premium ha una maggiore resistenza al calore rispetto alla lega standard. Ciascun caso rappresenta il test separato di una delle leghe. È indicata la temperatura massima alla quale può essere sottoposto il cuscinetto.
- **cereal.sav.** File di dati ipotetici che prende in esame le preferenze relative agli alimenti della colazione di un campione di 880 persone. Il file riporta anche l'età, il sesso e lo stato civile del campione e se le persone conducono uno stile di vita attivo (in base a un'attività sportiva con frequenza di due volte alla settimana). Ogni caso rappresenta un rispondente separato.
- **clothing_defects.sav.** File di dati ipotetici che prende in esame il processo di controllo di qualità di un'industria di abbigliamento. Per ciascun lotto prodotto nella fabbrica, gli ispettori prelevano un campione di abiti per contare il numero dei capi che non sono accettabili per la vendita.
- **coffee.sav.** Questo file di dati contiene informazioni sulle immagini percepite di sei marche di caffè freddo (Kennedy, Riquier, e Sharp, 1996). Per ciascuno dei 23 attributi dell'immagine del caffè freddo, sono state selezionate tutte le marche descritte da tale attributo. Le sei marche sono indicate dalle sigle AA, BB, CC, DD, EE e FF per tutelare la confidenzialità dei dati.
- **contacts.sav.** File di dati ipotetici che prende in esame l'elenco dei contatti di un gruppo di rappresentanti di vendita di computer aziendali. Ciascun contatto è classificato in base al reparto della società in cui lavora e dalle relative categorie aziendali. Il file riporta anche l'importo dell'ultima vendita effettuata, il tempo trascorso dall'ultima vendita e le dimensioni della società del contatto.
- **creditpromo.sav.** File di dati ipotetici che prende in esame l'impegno di un grande magazzino nel tentativo di valutare l'efficacia di una recente promozione con carta di credito. A tale scopo, sono stati selezionati 500 titolari di carta in modo casuale. Alla metà di questi è stato inviato un annuncio promozionale che comunica la riduzione del tasso d'interesse nel caso di acquisti effettuati entro i tre mesi successivi. All'altra metà è stato inviato un annuncio stagionale standard.

- **customer_dbase.sav.** File di dati ipotetico che prende in esame l'impegno di una società nel tentativo di utilizzare le informazioni contenute nel proprio database dei dati per creare offerte speciali per i clienti che più probabilmente risponderanno all'offerta. È stato selezionato in modo casuale un sottoinsieme della base dei clienti a cui è stata inviata l'offerta speciale e sono state registrate le risposte ricevute.
- **customer_information.sav.** File di dati ipotetici contenente le informazioni postali del cliente, ad esempio il nome e l'indirizzo.
- **customer_subset.sav.** Un sottoinsieme di 80 casi da *customer_dbase.sav*.
- **debate.sav.** File di dati ipotetici che prende in esame le risposte appaiate a un'indagine da parte dei partecipanti a un dibattito politico prima e dopo il dibattito. Ogni caso rappresenta un rispondente separato.
- **debate_aggregate.sav.** File di dati ipotetici che aggrega le risposte contenute nel file *debate.sav*. Ciascun caso corrisponde a una classificazione incrociata della preferenza prima e dopo il dibattito.
- **demo.sav.** File di dati ipotetici che prende in esame un database di clienti che hanno fatto acquisti al fine di inviare offerte mensili tramite il metodo del direct mailing. Viene registrata la risposta dei clienti, sia che abbiano aderito all'offerta o meno, insieme a diverse informazioni demografiche.
- **demo_cs_1.sav.** File di dati ipotetici che prende in esame il primo passo che una società intraprende per compilare un database con informazioni ricavate dai sondaggi. Ogni caso rappresenta una diversa città. Sono registrate anche le informazioni sulla regione, provincia, distretto e città.
- **demo_cs_2.sav.** File di dati ipotetici che prende in esame il secondo passo che una società intraprende per compilare un database con informazioni ricavate dai sondaggi. Ogni caso rappresenta una diversa unità di abitazione, ricavata dalle città selezionate nel primo passo. Sono registrate anche le informazioni sulla regione, provincia, distretto, città, suddivisione e unità. Il file include inoltre informazioni sul campionamento ottenute dai primi due stadi del disegno.
- **demo_cs.sav.** File di dati ipotetici che contiene informazioni sulle indagini raccolte utilizzando un disegno di campionamento complesso. Ogni caso rappresenta una diversa unità di abitazione. Sono registrate diverse informazioni demografiche e sul campionamento.
- **dmdata.sav.** File di dati ipotetici che contiene informazioni demografiche e di acquisto di una società di direct marketing. *dmdata2.sav* contiene informazioni su un sottoinsieme di contatti che hanno ricevuto un mailing di prova e *dmdata3.sav* contiene informazioni sui contatti rimanenti che non hanno ricevuto il mailing di prova.
- **dietstudy.sav.** File di dati ipotetici che contiene il risultato di uno studio ipotetico sulla dieta chiamato "Stillman diet" (Rickman, Mitchell, Dingman, e Dalen, 1974). Ogni caso rappresenta un diverso soggetto e ne riporta il peso prima e dopo la dieta in libbre e i livelli dei trigliceridi in mg/100 ml.
- **dvdplayer.sav.** File di dati ipotetici che prende in esame lo sviluppo di un nuovo lettore DVD. Utilizzando un prototipo, il personale addetto al marketing ha raccolto dati sui gruppi di interesse. Ogni caso rappresenta un diverso utente che è stato sottoposto all'indagine e include informazioni demografiche personali dell'utente e sulle risposte che ha fornito riguardo al prototipo.

- **german_credit.sav.** Questo file di dati contiene informazioni ricavate dall'insieme di dati "German Credit" del Repository of Machine Learning Databases (Blake e Merz, 1998) presso la University of California, Irvine.
- **grocery_1month.sav.** Questo file di dati ipotetici corrisponde al file di dati *grocery_coupons.sav* con gli acquisti settimanali organizzati in modo che ogni caso corrisponda a un cliente separato. Alcune delle variabili che cambiano settimanalmente non vengono riportate nei risultati; l'importo speso registrato corrisponde ora alla somma degli importi spesi durante le quattro settimane dello studio.
- **grocery_coupons.sav.** File di dati ipotetici che contiene i dati sui sondaggi raccolti da una catena di drogherie interessata alle abitudini di acquisto dei suoi clienti. Ciascun cliente viene seguito per quattro settimane e ciascun caso corrisponde a una settimana per cliente con informazioni sul luogo degli acquisti e i tipi di acquisti, incluso l'importo speso nelle drogherie durante la settimana.
- **guttman.sav.** Bell (Bell, 1961) ha presentato una tabella per illustrare i possibili gruppi sociali. Guttman (Guttman, 1968) ha utilizzato una parte di tale tabella, in cui cinque variabili che descrivono elementi come l'interazione sociale, i sentimenti di appartenenza a un gruppo, la vicinanza fisica dei membri e il grado di formalità della relazione, sono state incrociate con cinque gruppi sociali teorici, compresi folla (ad esempio, le persone presenti a una partita di calcio), uditorio (ad esempio, di uno spettacolo teatrale o di una lezione universitaria), pubblico (ad esempio televisivo), calca (come una folla, ma con un'interazione molto maggiore), gruppi primari (intimi), gruppi secondari (volontari) e la comunità moderna (unione non stretta derivante da una vicinanza fisica elevata e dall'esigenza di servizi specializzati).
- **health_funding.sav.** File di dati ipotetici che contiene i dati sui fondi di assistenza sanitaria (importo per 100 persone), sui tassi di malattie (tasso per 10.000 persone) e sulle visite ai fornitori di assistenza sanitaria (tasso per 10.000 persone). Ogni caso rappresenta una diversa città.
- **hivassay.sav.** File di dati ipotetici che prende in esame l'impegno di un'industria farmaceutica nel tentativo di sviluppare un'analisi che riesca a rilevare in tempi brevi l'infezione da virus HIV. I risultati dell'analisi sono otto sfumature di colore rosso sempre più intenso; le sfumature più intense indicano la maggiore probabilità di infezione. Un esperimento di laboratorio è stato condotto su 2000 campioni di sangue. La metà di questi è risultata infetta al virus HIV, l'altra metà non è risultata infetta.
- **hourlywagedata.sav.** File di dati ipotetici che prende in esame la paga oraria degli infermieri occupati presso uffici e ospedali e in base ai diversi livelli di esperienza.
- **insurance_claims.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nella creazione di un modello per contrassegnare le richieste di risarcimento sospette e potenzialmente fraudolente. Ogni caso rappresenta una richiesta di risarcimento separata.
- **insure.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nello studio dei fattori di rischio, che indicano l'eventualità che un cliente presenti una domanda di indennizzo in un contratto assicurativo sulla vita della durata di dieci anni. Ogni caso nel file di dati rappresenta una coppia di contratti. In un contratto sono contenute informazioni su una richiesta di risarcimento, l'altro sull'età e sul sesso.

- **judges.sav.** File di dati ipotetici che prende in esame il punteggio assegnato, da giurie qualificate (più un appassionato) a 300 prestazioni sportive. Ciascuna riga rappresenta una diversa prestazione; i giudici hanno esaminato le stesse prestazioni.
- **kinship_dat.sav.** Rosenberg e Kim (Rosenberg e Kim, 1975) si prefiggono di analizzare 15 termini indicanti parentela (zia, fratello, cugino, padre, nipote femmina, di nonni, nonno, nonna, nipote maschio di nonni, madre, nipote maschio di zii), nipote femmina di zii, sorella, figlio, zio). Hanno richiesto a quattro gruppi di studenti universitari (due composti da femmine e due da maschi) di ordinare questi termini in base alla similiarità. A due gruppi (uno femminile e uno maschile) è stato richiesto di effettuare l'ordinamento due volte, con il secondo ordinamento basato su un criterio diverso rispetto al primo. Di conseguenza, sono state ottenute sei "sorgenti" in totale. Ogni sorgente corrisponde a una matrice di prossimità 15×15, le cui celle sono uguali al numero delle persone in una sorgente meno il numero di volte in cui gli oggetti sono stati ripartiti insieme nella sorgente.
- **kinship_ini.sav.** Questo file di dati contiene la configurazione iniziale di una soluzione a tre dimensioni per *kinship_dat.sav*.
- **kinship_var.sav.** Questo file di dati contiene variabili indipendenti relative a *sesso*, *generazione* e *grado* di separazione che possono essere utilizzate per interpretare le dimensioni di una soluzione per *kinship_dat.sav*. In modo specifico, tali variabili possono essere utilizzate per limitare lo spazio della soluzione a una combinazione lineare di tali variabili.
- **marketvalues.sav.** File di dati che prende in esame le vendite di abitazioni in un nuovo centro abitato in Algonquin, Ill., durate gli anni 1999–2000. Tali vendite sono una questione di dominio pubblico.
- **nhis2000_subset.sav.** Il National Health Interview Survey (NHIS) è un sondaggio di grandi dimensioni condotto sulla popolazione civile americana. Le interviste vengono realizzate di persona e si basano su un campione rappresentativo di famiglie a livello nazionale. Per ogni membro di una famiglia vengono raccolte osservazioni e informazioni di carattere demografico relative allo stato di salute. Questo file di dati contiene un sottoinsieme delle informazioni ottenute dall'indagine del 2000. National Center for Health Statistics. National Health Interview Survey, 2000. File di dati e documentazione di dominio pubblico. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accesso 2003.
- **ozone.sav** I dati includono 330 osservazioni basate su sei variabili meteorologiche per quantificare la concentrazione dell'ozono dalle variabili rimanenti. I precedenti ricercatori, (Breiman e Friedman, 1985) e (Hastie e Tibshirani, 1990), hanno rilevato non linearità tra queste variabili, che impediscono un approccio di regressione standard.
- **pain_medication.sav.** File di dati ipotetici che contiene i risultati di un test clinico per stabilire la cura antinfiammatoria per il trattamento del dolore generato dall'artrite cronica. Di particolare interesse, il test ha evidenziato il tempo che impiega il farmaco ad avere effetto e il confronto con altri farmaci esistenti.
- **patient_los.sav.** File di dati ipotetici che contiene informazioni sul trattamento dei pazienti ricoverati per sospetto di infarto del miocardio. Ogni caso corrisponde a un diverso paziente e contiene diverse variabili correlate alla degenza nell'ospedale.
- **patlos_sample.sav.** File di dati ipotetici che contiene informazioni sul trattamento di un campione di pazienti curato con trombolitici durante la degenza per infarto del miocardio. Ogni caso corrisponde a un diverso paziente e contiene diverse variabili correlate alla degenza nell'ospedale.

- **poll_cs.sav.** File di dati ipotetici che prende in esame i sondaggi per stabilire il livello di sostegno pubblico nei confronti di un disegno di legge prima che diventi una legge vera e propria. I casi corrispondono ai votanti registrati. Ciascun caso riporta informazioni sulla contea, sul comune e sul quartiere in cui vive il votante.
- **poll_cs_sample.sav.** File di dati ipotetici che contiene un campione dei votanti elencati nel file *poll_cs.sav*. Il campione è stato selezionato in base al disegno specificato nel file di piano *poll.csplan* e questo file di dati contiene le probabilità di inclusione e i pesi del campione. Tuttavia, notare che poiché fa uso del metodo PPS (probability-proportional-to-size, probabilità proporzionale alla dimensione), esiste anche un file contenente le probabilità di selezione congiunte (*poll_jointprob.sav*). Le ulteriori variabili corrispondenti ai dati demografici dei votanti e alla loro opinione sul disegno di legge, sono state raccolte e aggiunte al file di dati dopo aver acquisito il campione.
- **property_assess.sav.** File di dati ipotetici che prende in esame l'impegno di un perito di una contea nel tentativo di mantenere gli accertamenti sui valori delle proprietà aggiornati in base alle risorse limitate. I casi rappresentano le proprietà vendute nella contea nello scorso anno. Ogni caso nel file di dati contiene informazioni sul comune in cui si trova la proprietà, il perito che per ultimo ha visitato la proprietà, il tempo trascorso dall'accertamento, la valutazione fatta in tale momento e il valore di vendita della proprietà.
- **property_assess_cs.sav.** File di dati ipotetici che prende in esame l'impegno di un perito di uno stato nel tentativo di mantenere aggiornati gli accertamenti sui valori delle proprietà in base alle risorse limitate. I casi corrispondono alle proprietà nello stato. Ogni caso nel file di dati include informazioni sulla contea, il comune e il quartiere in cui risiede la proprietà, la data dell'ultimo accertamento e la valutazione fatta in tale data.
- **property_assess_cs_sample.sav.** File di dati ipotetici che contiene un campione delle proprietà elencate nel file *property_assess_cs.sav*. Il campione è stato selezionato in base al disegno specificato nel file di piano *property_assess.csplan* e questo file di dati contiene le probabilità di inclusione e i pesi del campione. L'ulteriore variabile *Valore corrente* è stata raccolta e aggiunta al file di dati dopo aver acquisito il campione.
- **recidivism.sav.** File di dati ipotetici che prende in esame l'impegno delle Forze dell'Ordine nel tentativo di valutare il tasso di recidività nella propria area di giurisdizione. Ogni caso corrisponde a un precedente trasgressore e include le informazioni demografiche, alcuni dettagli sul primo crimine, il tempo trascorso fino al secondo arresto e se tale arresto è avvenuto entro due anni dal primo.
- **recidivism_cs_sample.sav.** File di dati ipotetici che prende in esame l'impegno delle Forze dell'Ordine nel tentativo di valutare il tasso di recidività nella propria area di giurisdizione. Ogni caso corrisponde a un trasgressore precedente, rilasciato dopo il primo arresto durante il mese di giugno del 2003 e registra le relative informazioni demografiche, alcuni dettagli sul primo crimine commesso e i dati del secondo arresto, se si è verificato prima della fine di giugno del 2006. I trasgressori sono stati selezionati dai dipartimenti sottoposti a campione in base al piano di campionamento specificato nel file *recidivism_cs.csplan*. Poiché viene utilizzato un metodo PPS (Probability-Proportional-to-Size, probabilità proporzionale alla dimensione), esiste anche un file contenente le probabilità di selezione congiunte (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** File di dati ipotetici contenente i dati delle transazioni di acquisto, inclusa la data di acquisto, gli articoli acquistati e il valore monetario di ciascuna transazione.

- **salesperformance.sav.** File di dati ipotetici che prende in esame la valutazione di due nuovi corsi di formazione alle vendite. Sessanta dipendenti, divisi in tre gruppi, ricevono tutti la formazione standard. In più, al gruppo 2 viene assegnato un corso di formazione tecnica e al gruppo 3 un'esercitazione pratica. Alla fine del corso di formazione, ciascun dipendente viene sottoposto a un esame e il punteggio conseguito viene registrato. Ciascun caso nel file di dati rappresenta un diverso partecipante. Il file di dati include il gruppo a cui è assegnato il partecipante e il punteggio conseguito all'esame finale.
- **satisf.sav.** File di dati ipotetico che prende in esame un'indagine sulla soddisfazione dei clienti condotta da una società di vendita al dettaglio presso 4 negozi. Sono stati intervistati 582 clienti e ciascun caso rappresenta le risposte ottenute da un singolo cliente.
- **screws.sav.** Questo file di dati contiene informazioni sulle caratteristiche di viti, bulloni, dadi e puntine (Hartigan, 1975).
- **shampoo_ph.sav.** File di dati ipotetici che prende in esame il processo di controllo di qualità di un'industria di prodotti per capelli. A intervalli di tempo regolari, vengono misurati sei diversi lotti prodotti e ne viene registrato il relativo pH. I valori accettati sono compresi tra 4,5 e 5,5.
- **ships.sav.** Ad esempio, un insieme di dati presentato e analizzato altrove (McCullagh et al., 1989) riguarda i danni subiti dalle navi da carico a causa delle onde. I conteggi degli incidenti possono essere presentati con un tasso di Poisson in base al tipo di nave, al periodo di costruzione e al periodo di servizio. I mesi di servizio aggregati di ciascuna cella della tabella generata dalla classificazione incrociata dei fattori fornisce i valori di esposizione al rischio.
- **site.sav.** File di dati ipotetici che prende in esame l'impegno di una società nella scelta di nuovi siti in cui espandere la propria presenza. La società ha incaricato due consulenti separati che, oltre a valutare i siti e presentare un report completo, devono classificarli come potenzialmente "molto adatti", "adatti" o "poco adatti".
- **smokers.sav.** Questo file di dati è un estratto del 1998 National Household Survey of Drug Abuse e rappresenta un campione probabile di famiglie americane. (*http://dx.doi.org/10.3886/ICPSR02934*) Il primo passo nell'analisi di questo file di dati consiste quindi nel pesare i dati per rispecchiare le tendenze della popolazione.
- **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 miocardiaco, dell'ictus ischemico o emorragico e viene registrata l'ora dell'evento. Questo campione viene troncato a sinistra perché include solo i pazienti che sono sopravvissuti fino alla fine del programma riabilitativo post-ictus.
- **stroke_valid.sav.** File di dati ipotetici che riporta lo stato di un database medico dopo il controllo dei valori eseguito con la procedura Convalida i dati. Il database contiene comunque casi potenzialmente anomali.

- **survey_sample.sav.** File di dati che contiene i dati dell'indagine, compresi i dati demografici e varie misure dell'atteggiamento. Si basa su un sottoinsieme di variabili tratte dal 1998 NORC General Social Survey, benché i valori di alcuni dati siano stati modificati e siano state aggiunte variabili fittizie a scopo dimostrativo.
- **telco.sav.** File di dati ipotetici che prende in esame l'impegno di un'azienda di telecomunicazioni nel tentativo di ridurre il churn, ovvero l'abbandono dei propri clienti. Ciascun caso rappresenta un cliente separato e riporta diverse informazioni demografiche e sull'uso del servizio.
- **telco_extra.sav.** Questo file di dati è simile al file *telco.sav*, ma le variabili "tenure" e spesa del cliente trasformata tramite logaritmo sono state sostituite dalle variabili di spesa del cliente trasformata tramite logaritmo standardizzate.
- **telco_missing.sav.** Questo file di dati è un sottoinsieme del file di dati *telco.sav*, ma alcuni dei valori di dati demografici sono stati sostituiti con valori mancanti.
- **testmarket.sav.** File di dati ipotetici che prende in esame i piani di una catena di fast food per aggiungere un nuovo prodotto al proprio menu. Sono previste tre campagne promozionali del nuovo prodotto. Il prodotto viene introdotto in diversi mercati selezionati in modo casuale. Per ogni sede viene utilizzata una promozione differente registrando le vendite settimanali della nuova voce per le prime quattro settimane. Ogni caso rappresenta un luogo e una settimana diversi.
- **testmarket_1month.sav.** Questo file di dati ipotetici corrisponde al file *testmarket.sav* con le vendite settimanali organizzate in modo che ogni caso corrisponda a un luogo separato. Alcune delle variabili che cambiano settimanalmente non vengono riportate nei risultati; le vendite registrate corrispondono ora alla somma delle vendite conseguite durante le quattro settimane dello studio.
- **tree_car.sav.** File di dati ipotetici che contiene dati demografici e sul prezzo di acquisto dei veicoli.
- **tree_credit.sav.** File di dati ipotetici che contiene dati demografici e sulla cronologia dei mutui di una banca.
- **tree_missing_data.sav.** File di dati ipotetici che contiene dati demografici e sulla cronologia dei mutui di una banca con un numero elevato di valori mancanti.
- **tree_score_car.sav.** File di dati ipotetici che contiene dati demografici e sul prezzo di acquisto dei veicoli.
- **tree_textdata.sav.** File di dati semplice con due variabili destinato principalmente per mostrare lo stato predefinito delle variabili prima dell'assegnazione dei livelli di misurazione e delle etichette dei valori.
- **tv-survey.sav.** File di dati ipotetici che prende in esame un sondaggio condotto da una emittente televisiva che deve stabilire se estendere la durata di un programma di successo. A un campione di 906 intervistati è stato chiesto se preferisce guardare il programma con diverse condizioni. Ciascuna riga rappresenta un diverso intervistato e ciascuna colonna una diversa condizione.
- **ulcer_recurrence.sav.** Questo file contiene informazioni parziali su uno studio svolto per mettere a confronto l'efficacia di due terapie preventive per la recidiva delle ulcere. Fornisce un ottimo esempio di dati acquisiti a intervalli ed è stato presentato e analizzato in altri luoghi (Collett, 2003).

- **ulcer_recurrence_recoded.sav.** In questo file sono contenute le informazioni del file *ulcer_recurrence.sav* riorganizzate per consentire di presentare la probabilità degli eventi per ciascun intervallo dello studio, anziché solo alla fine. È stato presentato e analizzato in altri luoghi (Collett et al., 2003).
- **verd1985.sav.** Questo file di dati prende in esame un'indagine (Verdegaal, 1985). Sono state registrate le risposte di quindici soggetti a otto variabili. Le variabili di interesse sono suddivise in tre insiemi. L'insieme 1 include *età* e *statociv*, l'insieme 2 include *andom* e *giornale* e l'insieme 3 include *musica* e *vicinato*. *Andom* viene scalata come nominale multipla ed *età* come ordinale; tutte le altre variabili vengono scalate come nominali singole.
- **virus.sav.** File di dati ipotetici che prende in esame l'impegno di un ISP (Internet Service Provider) nel tentativo di determinare gli effetti che un virus può generare nelle sue reti. Si è tenuta traccia della percentuale (approssimativa) di traffico e-mail infettato da virus sulla rete in un lasso di tempo, dal momento dell'individuazione fino alla soppressione della minaccia.
- **wheeze_steubenville.sav.** Questo file è un sottoinsieme di uno studio longitudinale degli effetti che l'inquinamento provoca sulla salute dei bambini (Ware, Dockery, Spiro III, Speizer, e Ferris Jr., 1984). I dati contengono misure binarie ripetute del livello di asma dei bambini della città di Steubenville, Ohio, di 7, 8, 9 e 10 anni. I dati indicano anche se la mamma dei bambini era fumatrice durante il primo anno dello studio.
- **workprog.sav.** File di dati ipotetici che prende in esame un programma di lavoro governativo il cui obiettivo è fornire attività più adatte alle persone diversamente abili. È stato seguito un campione di potenziali partecipanti al programma, alcuni dei quali sono stati selezionai in modo casuale e altri no. Ogni caso rappresenta un diverso partecipante al programma.
- **worldsales.sav** Questo file di dati ipotetici contiene i ricavi suddivisi per continenti e prodotti.

Appendice

B

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, COMMERCIABILITÀ 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.

Bibliografia

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Blake, C. L., e C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., e J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, .

Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

Davison, A. C., e D. V. Hinkley. 2006. *Bootstrap Methods and their Application*. : Cambridge University Press.

Green, P. E., e V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., e Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, .

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, T., e R. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

Kennedy, R., C. Riquier, e B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement and Analysis for Marketing*, 5, .

McCullagh, P., e J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Prezzo, R. H., e D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Rickman, R., N. Mitchell, J. Dingman, e J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

Rosenberg, S., e M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, .

Shao, J., e D. Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer.

Van der Ham, T., J. J. Meulman, D. C. Van Strien, e H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, .

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (in Dutch)*. Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, e B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, .

Indice