IBM SPSS Bootstrapping 19

Note: Before using this information and the product it supports, read the general information under Notices el p. 41.

Prefacio

IBM® SPSS® Statistics es un sistema global para el análisis de datos. El módulo adicional opcional Muestreo autodocimante proporciona las técnicas de análisis adicionales que se describen en este manual. El módulo adicional Muestreo autodocimante se debe utilizar con el sistema básico de SPSS Statistics y está completamente integrado en dicho sistema.

Acerca de SPSS Inc., an IBM Company

SPSS Inc., an IBM Company, es uno de los principales proveedores globales de software y soluciones de análisis predictivo. La gama completa de productos de la empresa (recopilación de datos, análisis estadístico, modelado y distribución) capta las actitudes y opiniones de las personas, predice los resultados de las interacciones futuras con los clientes y, a continuación, actúa basándose en esta información incorporando el análisis en los procesos comerciales. Las soluciones de SPSS Inc. tratan los objetivos comerciales interconectados en toda una organización centrándose en la convergencia del análisis, la arquitectura de TI y los procesos comerciales. Los clientes comerciales, gubernamentales y académicos de todo el mundo confían en la tecnología de SPSS Inc. como ventaja ante la competencia para atraer, retener y hacer crecer los clientes, reduciendo al mismo tiempo el fraude y mitigando los riesgos. SPSS Inc. fue adquirida por IBM en octubre de 2009. Para obtener más información, visite *http://www.spss.com*.

Asistencia técnica

El servicio de asistencia técnica está a disposición de todos los clientes de mantenimiento. Los clientes podrán ponerse en contacto con este servicio de asistencia técnica si desean recibir ayuda sobre la utilización de los productos de SPSS Inc. o sobre la instalación en alguno de los entornos de hardware admitidos. Para ponerse en contacto con el servicio de asistencia técnica, consulte el sitio web de SPSS Inc. en *http://support.spss.com* o encuentre a su representante local a través del sitio web *http://support.spss.com/default.asp?refpage=contactus.asp*. Tenga a mano su identificación, la de su organización y su contrato de asistencia cuando solicite ayuda.

Servicio de atención al cliente

Si tiene cualquier duda referente a la forma de envío o pago, póngase en contacto con su oficina local, que encontrará en el sitio Web en *http://www.spss.com/worldwide*. Recuerde tener preparado su número de serie para identificarse.

Cursos de preparación

SPSS Inc. ofrece cursos de preparación, tanto públicos como in situ. Todos los cursos incluyen talleres prácticos. Los cursos tendrán lugar periódicamente en las principales ciudades. Si desea obtener más información sobre estos cursos, póngase en contacto con su oficina local que encontrará en el sitio Web en *http://www.spss.com/worldwide*.

Publicaciones adicionales

Los documentos *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion* y *SPSS Statistics: Advanced Statistical Procedures Companion*, escritos por Marija Norušis y publicados por Prentice Hall, están disponibles y se recomiendan como material adicional. Estas publicaciones cubren los procedimientos estadísticos del módulo SPSS Statistics Base, el módulo Advanced Statistics y el módulo Regression. Tanto si da sus primeros pasos en el análisis de datos como si ya está preparado para las aplicaciones más avanzadas, estos libros le ayudarán a aprovechar al máximo las funciones ofrecidas por IBM® SPSS® Statistics. Si desea información adicional sobre el contenido de la publicación o muestras de capítulos, consulte el sitio web de la autora: *http://www.norusis.com*

Contenido

Apéndices

Parte I:
Manual del usuario

Capítulo
1

Introducción al muestreo autodocimante

Cuando recopila datos suele estar interesado en las propiedades de la población de la que ha tomado la muestra. Hace inferencias acerca de los parámetros de la población con estimaciones calculadas de la muestra. Por ejemplo, si el conjunto de datos *Employee data.sav* que se incluye con el producto es una muestra aleatoria de una población mayor de empleados, la media de la muestra de 34.419,57 dólares como *Salario actual* es una estimación de la media del salario actual de la población de los empleados. Además, esta estimación tiene un error típico de 784,311 dólares para una muestra de un tamaño de 474; y un intervalo de confianza del 95% para la media del salario actual de la población de los empleados es de 32.878,40 dólares a 35.960,73 dólares. Pero, ¿cuál es el nivel de fiabilidad de estos estimadores? Para algunas poblaciones "conocidas" y parámetros de buen comportamiento, sabemos algo acerca de las propiedades de las estimaciones de la muestra y podemos confiar en estos resultados. El muestreo autodocimante busca más información acerca de las propiedades de los estimadores de poblaciones "desconocidas" y parámetros de mal comportamiento.

Figura 1-1
Realización de inferencias paramétricas acerca de la media de la población

			Estadístico	Error típ.
Salario actual	Media		$34,419.57	$784.311
	Intervalo de confianza 95%	Inferior	$32,878.40	
		Superior	$35,960.73	
	Mediana		$28,875.00	

Funcionamiento del muestreo autodocimante

En su forma más simple, para un conjunto de datos con un tamaño de muestra de *N*, tomará *B* muestras "autodocimantes" de un tamaño *N* sustituyendo del conjunto de datos original y calcular el estimador de cada uno de estas *B* muestras autodocimantes. Estas *B* estimaciones de muestras autodocimantes son una muestra de un tamaño *B* de la que podrá realizar inferencias acerca del estimador. Por ejemplo, si toma 1.000 muestras autodocimantes del conjunto de datos *Employee data.sav* , el error típico de muestras autodocimantes estimado de 776,91 dólares para la media de la muestra de *Salario actual* es una alternativa a la estimación de 784,311 dólares.

Además, el muestreo autodocimante proporciona un error estándar y un intervalo de confianza para la mediana, cuyas estimaciones paramétricas no están disponibles.

Figura 1-2
Realización de inferencias autodocimantes acerca de la media de muestra

			Estadístico	Error típ.	Bootstrap[a]			
					Sesgo	Error típ.	Intervalo de confianza 95%	
							Inferior	Superior
Salario actual	Media		$34,419.57	$784.311	$14.66	$776.91	$32,990.38	$36,026.06
	Intervalo de confianza 95%	Inferior	$32,878.40					
		Superior	$35,960.73					
	Median		$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

Compatibilidad del muestreo autodocimante en el producto

El muestreo autodocimante se incorpora como un cuadro de diálogo subordinado en procedimientos que admiten el muestreo autodocimante. Consulte Procedimientos que admiten el muestreo autodocimante si desea obtener información acerca de los procedimientos que admiten el muestreo autodocimante.

Si se requiere muestreo autodocimante en los cuadros de diálogo, se pega un nuevo comando `BOOTSTRAP` independiente, además de la sintaxis normal que genera el cuadro de diálogo. El comando `BOOTSTRAP` crea las muestras autodocimantes en función de sus especificaciones. Internamente, el producto trata estas muestras autodocimantes como segmentaciones, incluso si no se muestran de forma explícita en el Editor de datos. Significa que, de forma interna, son efectivamente *B***N* casos, de forma que el recuento de casos en la barra de estado contará desde 1 a *B***N* cuando se procesen los datos durante el muestreo autodocimante. El Sistema de gestión de resultados (OMS) se utiliza para recopilar los resultados de la ejecución del análisis en cada "segmentación autodocimante". Estos resultados se combinan y los resultados autodocimantes combinados se muestran en el Visor con el resto del resultado normal que genera el procedimiento. En algunos casos, podrá ver una referencia a "segmentación autodocimante 0"; es el conjunto de datos original.

Capítulo
2

Muestreo autodocimante

Bootstrapping es un método para derivar estimaciones robustas de errores típicos e intervalos de confianza para estimaciones como la media, mediana, proporción, razón de las ventajas, coeficientes de correlación o coeficientes de regresión. También se puede utilizar para crear pruebas hipotéticas. Bootstrapping es más útil como alternativa a estimaciones paramétricas en caso de que los supuestos de esos métodos sean dudosos (como en el caso de modelos de regresión con residuos heteroscedástico se ajusten a muestras pequeñas), o si la inferencia paramétrica no es posible o requiere fórmulas muy complicadas para el cálculo de errores típicos (como en el caso de cálculo de intervalos de confianza de mediana, cuartiles y otros percentiles).

Ejemplos. Una empresa de telecomunicaciones pierde alrededor del 27% de sus clientes por abandono cada mes. Para reducir el porcentaje de abandono, los directivos quieren saber si este porcentaje varía en diferentes grupos de clientes predefinidos. Mediante el muestreo autodocimante, puede determinar si un porcentaje concreto de abandonos describe de forma adecuada los cuatro tipos principales de clientes. Si desea obtener más información, consulte el tema Uso de muestreo autodocimante para obtener intervalos de confianza para proporciones en el capítulo 3 en *IBM SPSS Bootstrapping 19*.

En una revisión de los registros de empleados, los directivos están interesados en las experiencias anteriores de los empleados. La experiencia laboral es asimétrica, lo que hace que la media sea una estimación menos deseable de la experiencia laboral "habitual" entre los empleados que la mediana. Sin embargo, los intervalos de confianza no están disponibles para la mediana en el producto. Si desea obtener más información, consulte el tema Uso de muestreo autodocimante para obtener intervalos de confianza de medianas en el capítulo 3 en *IBM SPSS Bootstrapping 19*.

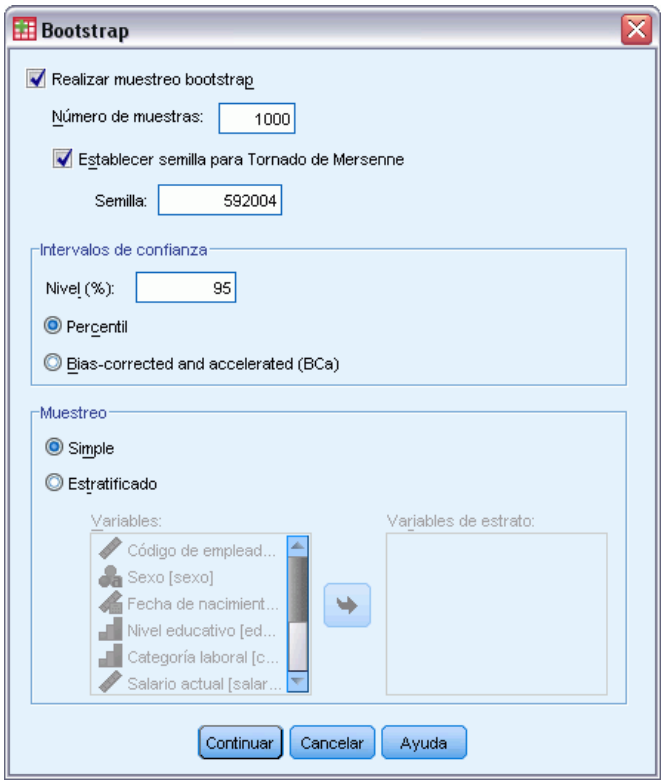
Los directivos también están interesados en determinar los factores que están asociados con los aumentos de salarios de los empleados mediante la definición de un modelo lineal de la diferencia entre el salario inicial y el actual. Al realizar un muestreo autodocimante de un modelo lineal, puede utilizar métodos de muestreado sucesivo especiales (muestreo autodocimante residual y wild) para obtener resultados más precisos. Si desea obtener más información, consulte el tema Uso de muestreo autodocimante para seleccionar mejores predictores en el capítulo 3 en *IBM SPSS Bootstrapping 19*.

Muchos procedimientos admiten el muestreo autodocimante y la combinación de resultados a partir del análisis de muestras autodocimantes. Los controles para especificar análisis autodocimantes se integran directamente como un cuadro de diálogo subordinado común en procedimientos que admiten el muestreo autodocimante. La configuración del cuadro de diálogo de muestras autodocimantes permanece en los procedimientos de forma que, si ejecuta un análisis de frecuencias con muestreo autodocimante en los cuadros de diálogo, el muestreo autodocimante se activará por defecto para otros procedimientos que lo admitan.

Para obtener un análisis autodocimante

▶ En los menús seleccione un procedimiento que admita el muestreo autodocimante y pulse en Autodocimante.

Figura 2-1
Cuadro de diálogo Autodocimante



▶ Seleccione Ejecutar bootstrapping.

También puede controlar las siguientes opciones:

Número de muestras. Para los intervalos de percentil y BCa producidos, se recomienda utilizar al menos 1000 muestras autodocimantes. Especifique un número entero positivo.

Definir semilla para tornado de Mersenne. Si se establece una semilla es posible replicar análisis. El uso de este control es parecido a establecer el tornado de Mersenne como generador activo y especificar un punto de inicio fijo en el cuadro de diálogo Generadores de números aleatorios, con la importante diferencia de que la definición de la semilla de este cuadro de diálogo mantendrá el estado actual del generador de números aleatorios y restaurará dicho estado cuando haya terminado el análisis.

Intervalos de confianza. Especifique un nivel de confianza mayor que 50 y menor que 100. Los intervalos de percentiles sólo utilizan los valores autodocimantes ordenados correspondientes a los percentiles de intervalo de confianza deseados. Por ejemplo, un intervalo de confianza de percentil del 95% utiliza los percentiles 2,5 y 97,5 de los valores autodocimantes como los límites inferior y superior del intervalo (interpolando los valores autodocimantes si es necesario). Los

intervalos de sesgo corregidos y acelerados (BCa) son intervalos ajustados que son más precisos, pero necesitan más tiempo de cálculo.

Muestreo. El método simple consiste en volver a muestrear los casos reemplazándolos del conjunto de datos original. El método estratificado consiste en volver a muestrear los casos sustituyendo el conjunto de datos original, *en* los estratos definidos por las variables de estratos de clasificación cruzada. El muestreo autodocimante estratificado puede ser muy útil si las unidades de los estratos son relativamente homogéneas aunque las unidades para todos los estratos son muy diferentes.

Procedimientos que admiten el muestreo autodocimante

Los siguientes procedimientos admiten el muestreo autodocimante.

Nota:

- El muestreo autodocimante no funciona con conjuntos de datos de imputación múltiple. Si hay una variable *Imputation_* en el conjunto de datos, el cuadro de diálogo Autodocimante se desactiva.
- El muestreo autodocimante utiliza eliminación por lista para determinar los casos; es decir, los casos con valores perdidos en cualquiera de las variables de análisis se eliminan del análisis, de forma que, cuando el muestreo autodocimante está en efecto, eliminación por lista se activa incluso si el procedimiento de análisis especifica otra forma de gestión de valores perdidos.

Opción Estadísticas básicas

Frecuencias

- La tabla Estadísticos admite estimaciones autodocimantes de media, desviación típica, varianza, mediana, asimetría, curtosis y percentiles.
- La tabla Frecuencias admite estimaciones autodocimantes de porcentaje.

Descriptivos

- La tabla Estadísticos descriptivos admite estimaciones autodocimantes de media, desviación típica, varianza, asimetría y curtosis.

Explorar

- La tabla Descriptivos admite estimaciones autodocimantes de media, media recortada al 5%, desviación típica, varianza, mediana, asimetría, curtosis y amplitud intercuartil.
- La tabla Estimadores-M admite estimaciones autodocimantes de estimador-M de Huber, estimador biponderado de Tukey, estimador-M de Hampel y onda de Andrews.
- La tabla Percentiles admite estimaciones autodocimantes de percentiles.

Tablas de contingencia

- La tabla Medidas direccionales admite estimaciones autodocimantes de Lambda, Goodman y Kruskal Tau, coeficiente de incertidumbre y d de Somers.

- La tabla Medidas simétricas admite estimaciones autodocimantes de Phi, V de Cramer, coeficiente de contingencia, tau-b de Kendall, tau-c de Kendall, Gamma, correlación de Spearman y r de Pearson.
- La tabla Estimación de riesgo admite estimaciones autodocimantes de la razón de las ventajas.
- La tabla de razón de las ventajas común de Mantel-Haenszel admite estimaciones autodocimantes y pruebas de significación de ln(Estimación).

Medias

- La tabla Informe admite estimaciones autodocimantes de media, mediana, mediana agrupada, desviación típica, varianza, curtosis, asimetría, media armónica y media geométrica.

Prueba T para una muestra

- La tabla Estadísticos admite estimaciones autodocimantes de media y desviación típica.
- La tabla Prueba admite estimaciones autodocimantes y pruebas de significación de diferencia de medias.

Prueba T para muestras independientes

- La tabla Estadísticos de grupo admite estimaciones autodocimantes de media y desviación típica.
- La tabla Prueba admite estimaciones autodocimantes y pruebas de significación de diferencia de medias.

Prueba T para muestras relacionadas

- La tabla Estadísticos admite estimaciones autodocimantes de media y desviación típica.
- La tabla Correlaciones admite estimaciones autodocimantes de correlaciones.
- La tabla Prueba admite estimaciones autodocimantes de media.

ANOVA de un factor

- La tabla Estadísticos descriptivos admite estimaciones autodocimantes de media y desviación típica.
- La tabla Comparaciones múltiples admite estimaciones autodocimantes de diferencia de medias.
- La tabla Pruebas de contraste admite estimaciones autodocimantes y pruebas de significación de valor de contraste.

MLG Univariante

- La tabla Estadísticos descriptivos admite estimaciones autodocimantes de media y desviación típica.
- La tabla Estimaciones de los parámetros admite estimaciones autodocimantes y pruebas de significación de coeficiente B.
- La tabla de resultados de contraste admite estimaciones autodocimantes y pruebas de significación de diferencia.
- Medias marginales estimadas: La tabla Estimaciones admite estimaciones autodocimantes de media.

- Medias marginales estimadas: La tabla Comparaciones por parejas admite estimaciones autodocimantes de diferencia de medias.
- Pruebas post hoc: La tabla Comparaciones múltiples admite estimaciones autodocimantes de diferencia de medias.

Correlaciones bivariadas

- La tabla Estadísticos descriptivos admite estimaciones autodocimantes de media y desviación típica.
- La tabla Correlaciones admite estimaciones autodocimantes de correlaciones.

Nota: Si se requieren correlaciones no paramétricas (tau-b de Kendall o Spearman) además de las correlaciones de Pearson, el cuadro de diálogo pega los comandos `CORRELATIONS` y `NONPAR CORR` con un comando `BOOTSTRAP` diferente para cada una. Se utilizarán las mismas muestras autodocimantes para calcular todas las correlaciones.

Correlaciones parciales

- La tabla Estadísticos descriptivos admite estimaciones autodocimantes de media y desviación típica.
- La tabla Correlaciones admite estimaciones autodocimantes de correlaciones.

Regresión lineal

- La tabla Estadísticos descriptivos admite estimaciones autodocimantes de media y desviación típica.
- La tabla Correlaciones admite estimaciones autodocimantes de correlaciones.
- La tabla Resumen de modelo admite estimaciones autodocimantes de Durbin-Watson.
- La tabla Coeficientes admite estimaciones autodocimantes y pruebas de significación de coeficiente B.
- La tabla Coeficientes de correlación admite estimaciones autodocimantes de correlaciones.
- La tabla Estadísticos residuales admite estimaciones autodocimantes de media y desviación típica.

Regresión ordinal

- La tabla Estimaciones de los parámetros admite estimaciones autodocimantes y pruebas de significación de coeficiente B.

Análisis discriminante

- La tabla Coeficientes de funciones discriminantes canónicas tipificados admite estimaciones autodocimantes de coeficientes tipificados.
- La tabla Coeficientes de funciones discriminantes canónicas admite estimaciones autodocimantes de coeficientes no tipificados.
- La tabla Coeficientes de función de clasificación admite estimaciones autodocimantes de coeficientes.

Opción Estadísticas avanzadas

MLG Multivariante

- La tabla Estimaciones de los parámetros admite estimaciones autodocimantes y pruebas de significación de coeficiente B.

Modelos lineales mixtos

- La tabla Estimaciones de efectos fijos admite estimaciones autodocimantes y pruebas de significación de estimación.
- La tabla Estimaciones de parámetros de covarianzas admite estimaciones autodocimantes y pruebas de significación de estimación.

Modelos lineales generalizados

- La tabla Estimaciones de los parámetros admite estimaciones autodocimantes y pruebas de significación de coeficiente B.

Regresión de Cox

- La tabla Variables en la ecuación admite estimaciones autodocimantes y pruebas de significación de coeficiente B.

Opción Regresión

Regresión logística binaria

- La tabla Variables en la ecuación admite estimaciones autodocimantes y pruebas de significación de coeficiente B.

Regresión logística multinomial

- La tabla Estimaciones de los parámetros admite estimaciones autodocimantes y pruebas de significación de coeficiente B.

Funciones adicionales del comando BOOTSTRAP

La sintaxis de comandos también le permite:

- Realice muestreos autodocimantes residuales y wild (subcomando `SAMPLING`)

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Parte II: Ejemplos

Capítulo
3

Muestreo autodocimante

Bootstrapping es un método para derivar estimaciones robustas de errores típicos e intervalos de confianza para estimaciones como la media, mediana, proporción, razón de las ventajas, coeficientes de correlación o coeficientes de regresión. También se puede utilizar para crear pruebas hipotéticas. Bootstrapping es más útil como alternativa a estimaciones paramétricas en caso de que los supuestos de esos métodos sean dudosos (como en el caso de modelos de regresión con residuos heteroscedástico se ajusten a muestras pequeñas), o si la inferencia paramétrica no es posible o requiere fórmulas muy complicadas para el cálculo de errores típicos (como en el caso de cálculo de intervalos de confianza de mediana, cuartiles y otros percentiles).

Uso de muestreo autodocimante para obtener intervalos de confianza para proporciones

Una empresa de telecomunicaciones pierde alrededor del 27% de sus clientes por abandono cada mes. Para reducir el porcentaje de abandono, los directivos quieren saber si este porcentaje varía en diferentes grupos de clientes predefinidos.

Esta información se recoge en el archivo *telco.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A el p. 31. Utilice el muestreo autodocimante para determinar si un porcentaje concreto de abandonos describe de forma adecuada los cuatro tipos principales de clientes.

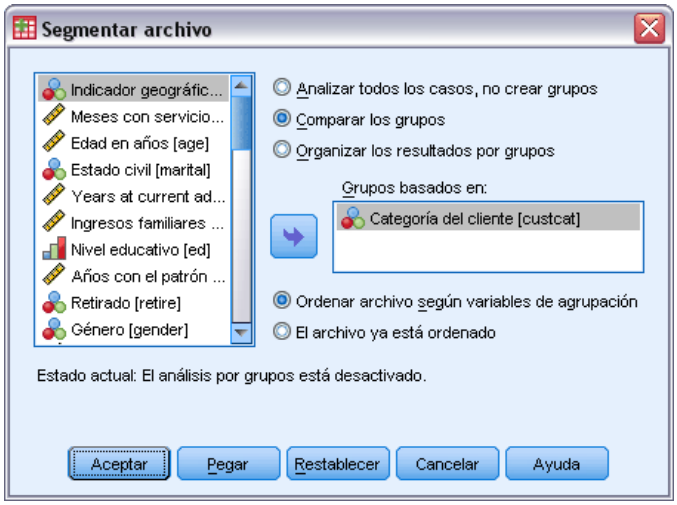
Nota: Este ejemplo utiliza el procedimiento de frecuencias y requiere la opción Statistics Base.

Preparación de datos

En primer lugar debe segmentar el archivo por *Categoría del cliente*.

- Para segmentar el archivo, elija en los menús del Editor de datos:
  Datos > Segmentar archivo...

Figura 3-1
Cuadro de diálogo Segmentar archivo

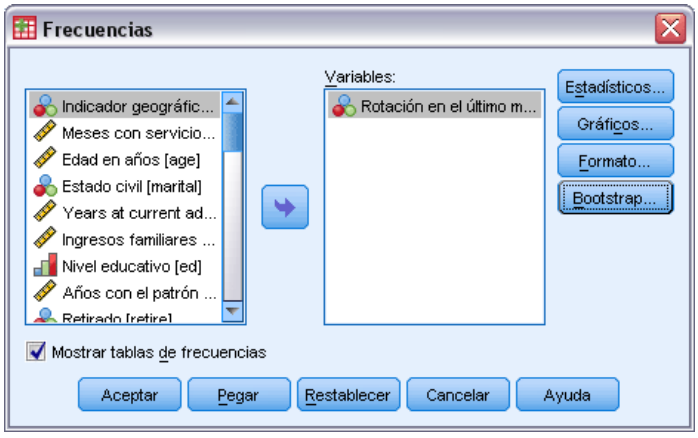


- Seleccione Comparar los grupos.
- Seleccione *Categoría del cliente* como la variable en la que se basan los grupos.
- Pulse en Aceptar.

Ejecución del análisis

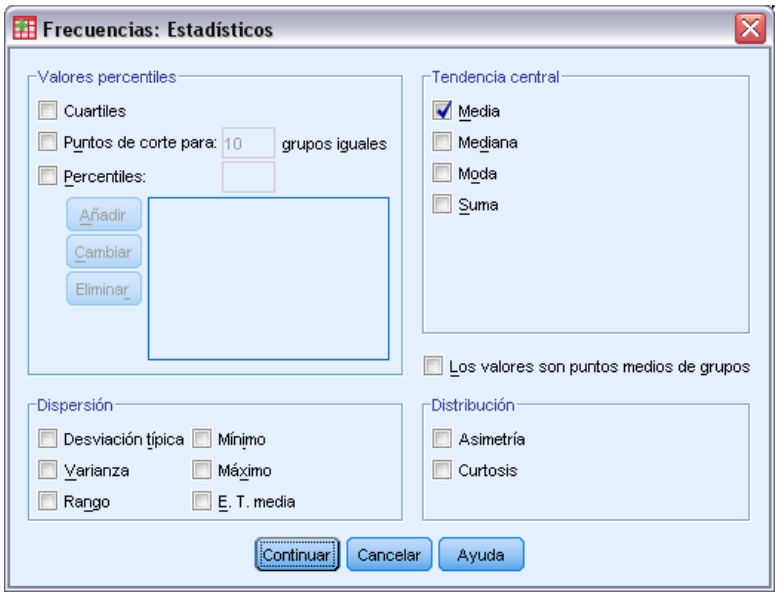
- Para obtener intervalos de confianza autodocimantes para proporciones, seleccione en los menús: Analizar > Estadísticos descriptivos > Frecuencias...

Figura 3-2
Cuadro de diálogo principal Frecuencias



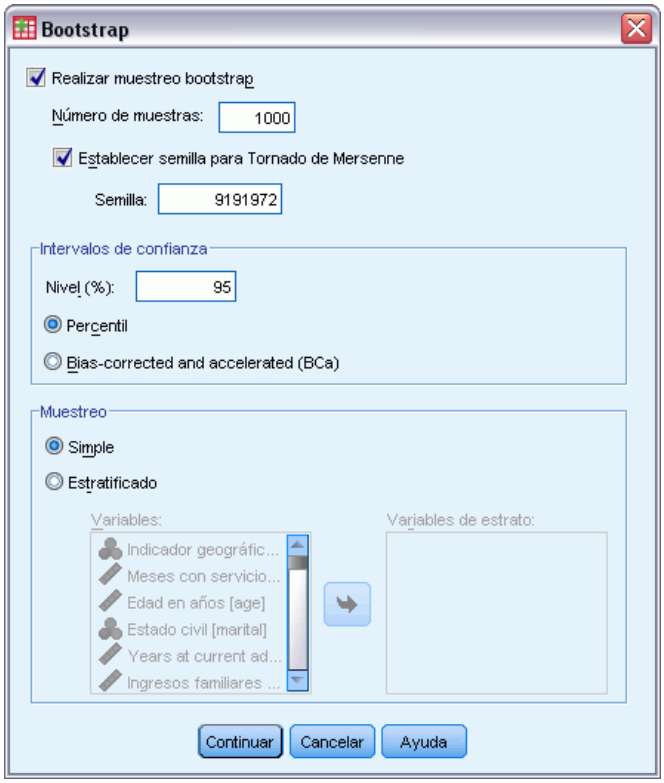
- Seleccione *Abandonaron durante el último mes [abandono]* como una variable en el análisis.
- Pulse en Estadísticos.

Figura 3-3
Cuadro de diálogo Estadísticos



- Seleccione Media en el grupo Tendencia central.
- Pulse en Continuar.
- Pulse en Autodocimante en el cuadro de diálogo Frecuencias.

Figura 3-4
Cuadro de diálogo Autodocimante



- Seleccione Ejecutar bootstrapping.
- Para replicar los resultados de este ejemplo de forma exacta, seleccione Establecer semilla para Tornado de Mersenne e introduzca 9191972 como semilla.
- Pulse en Continuar.
- Pulse en Aceptar en el cuadro de diálogo Frecuencias.

Estas selecciones generan la siguiente sintaxis de comandos:

```
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.
```

- Los comandos `SORT CASES` y `SPLIT FILE` dividen el archivo en la variable *custcat*.

- Los comandos `PRESERVE` y `RESTORE` "recuerdan" el estado actual del generador de números aleatorios y restaurar el sistema al estado posterior a la finalización del método autodocimante.
- El comando `SET` define el generador de números aleatorios a Mersenne Twister y el índice a 9191972, para que los resultados del muestreo autodocimante se puedan replicar exactamente. El comando `SHOW` muestra el índice en el resultado para futura referencia.
- El comando `BOOTSTRAP` solicita 1.000 muestras autodocimantes mediante nuevas muestras simples.
- La variable *churn* se utiliza para determinar las muestras caso a caso. Los registros con valores perdidos en esta variable se eliminan del análisis.
- El procedimiento `FREQUENCIES` posterior a `BOOTSTRAP` se ejecuta en cada una de las muestras autodocimantes.
- El subcomando `STATISTICS` produce la media de la variable *churn* en los datos originales. Además, las estadísticas combinadas se producen para la media y los porcentajes en la tabla de frecuencias.

Especificaciones de Bootstrap

Figura 3-5
Especificaciones de muestreo autodocimante

Método de muestreo	Simple
Número de muestras	1000
Nivel de intervalo de confianza	95.0%
Tipo de intervalo de confianza	Percentil

La tabla de especificaciones de muestreo autodocimante contiene los ajustes utilizados durante las nuevas muestras y es una referencia útil para comprobar si se han completado los análisis previstos.

Estadísticas

Figura 3-6
Tabla de estadísticos con el intervalo de confianza autodocimante para la proporción

Rotación en el último mes

Categoría del cliente			Statistic	Bootstrap[a]			
				Sesgo	Típ. Error	Intervalo de confianza al 95%	
						Inferior	Superior
Servicio básico	N	Válidos	266	0	0	266	266
		Perdidos	0	0	0	0	0
	Media		.31	.00	.03	.26	.37
E-Servicio	N	Válidos	217	0	0	217	217
		Perdidos	0	0	0	0	0
	Media		.27	.00	.03	.21	.34
Servicio Plus	N	Válidos	281	0	0	281	281
		Perdidos	0	0	0	0	0
	Media		.16	.00	.02	.12	.20
Servicio Total	N	Válidos	236	0	0	236	236
		Perdidos	0	0	0	0	0
	Media		.37	.00	.03	.31	.44

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

La tabla de estadísticos muestra, para cada nivel de *Categoría del cliente*, el valor de media de *Abandonaron durante el último mes*. Como *Abandonaron durante el último mes* sólo toma los valores de 0 y 1, con 1 para un cliente que ha abandonado, la media es igual a la proporción de los usuarios que han abandonado. La columna Estadísticos muestra los valores que suele producir Frecuencias, utilizando el conjunto de datos original. Las columnas Autodocimante se producen por los algoritmos de muestreo autodocimante.

- Bias es la diferencia entre el valor promedio de este estadístico entre las muestras autodocimantes y el valor en la columna Estadístico. En este caso, el valor promedio de *Abandonaron durante el último mes* se calcula para las 1000 muestras autodocimantes y posteriormente se calcula el promedio estas medias.
- Desv. El error es el error típico de *Abandonaron durante el último mes* en las 1000 muestras autodocimantes.
- El límite inferior del 95% del intervalo de confianza autodocimante es una interpolación de los valores 25 y 26 de *Abandonaron durante el último mes*, si las 1000 muestras autodocimantes se clasifican en orden ascendente. El límite superior es una interpolación de los valores de las medias 975 y 976.

Los resultados de la tabla sugieren que el índice de abandono es diferente entre tipos de clientes diferentes. En concreto, el intervalo de confianza de los clientes de *Servicio plus* no se superpone con ningún otro, lo que sugiere que de media es menos probable que estos clientes abandonen.

Si trabaja con variables categóricas con sólo dos valores, estos intervalos de confianza son alternativas a los producidos por el procedimiento de Pruebas no paramétricas para una muestra o Prueba T para una muestra.

Tabla de frecuencia

Figura 3-7
Tabla de frecuencias con el intervalo de confianza autodocimante para la proporción

Categoría del cliente			Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado	Bootstrap para Porcentaje[a]			
							Sesgo	Típ. Error	Intervalo de confianza al 95%	
									Inferior	Superior
Servicio básico	Válidos	No	183	68.8	68.8	68.8	.0	2.8	63.2	74.4
		Sí	83	31.2	31.2	100.0	.0	2.8	25.6	36.8
		Total	266	100.0	100.0		.0	.0	100.0	100.0
E-Servicio	Válidos	No	158	72.8	72.8	72.8	.1	3.1	66.4	78.8
		Sí	59	27.2	27.2	100.0	-.1	3.1	21.2	33.6
		Total	217	100.0	100.0		.0	.0	100.0	100.0
Servicio Plus	Válidos	No	237	84.3	84.3	84.3	.0	2.1	80.1	88.3
		Sí	44	15.7	15.7	100.0	.0	2.1	11.7	19.9
		Total	281	100.0	100.0		.0	.0	100.0	100.0
Servicio Total	Válidos	No	148	62.7	62.7	62.7	.0	3.2	56.4	69.1
		Sí	88	37.3	37.3	100.0	.0	3.2	30.9	43.6
		Total	236	100.0	100.0		.0	.0	100.0	100.0

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

La tabla de frecuencias muestra los intervalos de confianza de los porcentajes (proporción × 100%) de cada categoría y están disponibles para todas las variables categóricas. Otras características del producto no tienen intervalos de confianza comparables.

Uso de muestreo autodocimante para obtener intervalos de confianza de medianas

En una revisión de los registros de empleados, los directivos están interesados en las experiencias anteriores de los empleados. La experiencia laboral es asimétrica, lo que hace que la media sea una estimación menos deseable de la experiencia laboral “habitual” entre los empleados que la mediana. Sin embargo, sin muestreo autodocimante, los intervalos de confianza de la mediana no están disponibles de forma general en procedimientos estadísticos del producto.

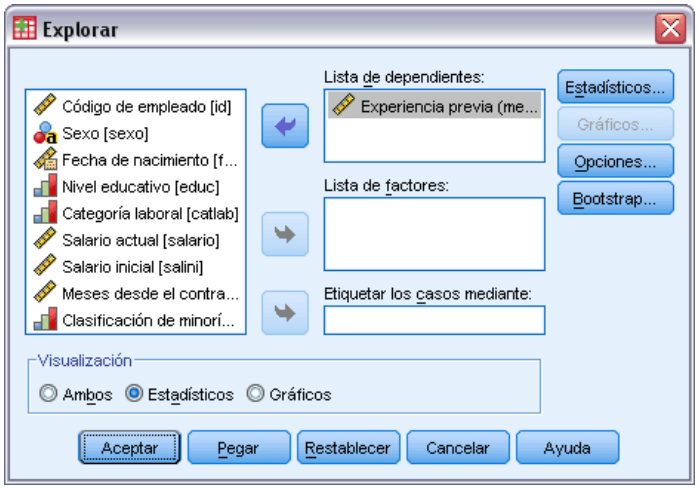
Esta información se recoge en el archivo *Employee data.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A el p. 31. Uso de muestreo autodocimante para obtener intervalos de confianza de la media.

Nota: Este ejemplo utiliza el procedimiento Explorar y requiere la opción Statistics Base.

Ejecución del análisis

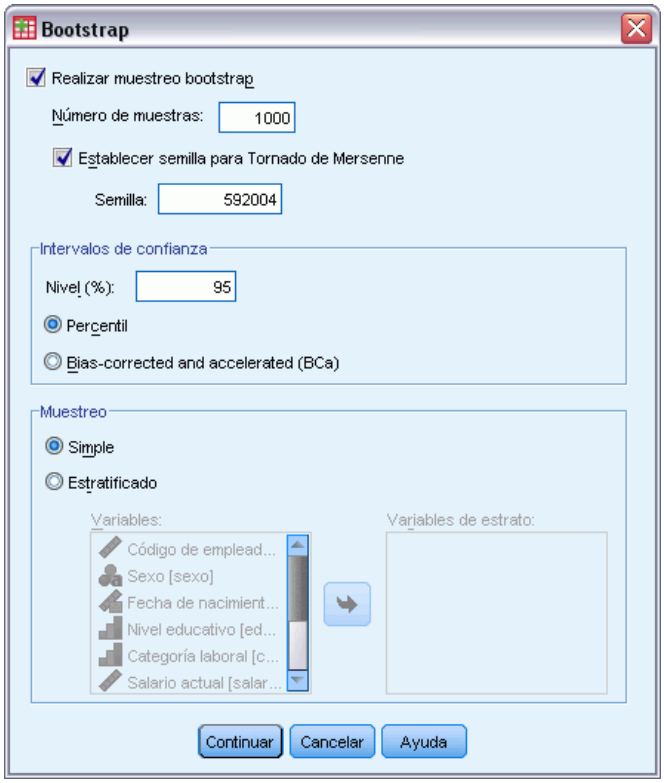
- Para obtener intervalos de confianza autodocimantes de la mediana, seleccione en los menús:
  Analizar > Estadísticos descriptivos > Explorar...

Figura 3-8
Cuadro de diálogo principal Explorar



- Seleccione *Experiencia anterior (meses) [prevexp]* como variable dependiente.
- Seleccione Estadísticos en la sección Mostrar.
- Pulse en Autodocimante.

Figura 3-9
Cuadro de diálogo Autodocimante



- Seleccione Ejecutar bootstrapping.
- Para replicar los resultados de este ejemplo de forma exacta, seleccione Establecer semilla para Tornado de Mersenne e introduzca 592004 como semilla.
- Para obtener resultados más precisos (requiere más tiempo de procesamiento), seleccione Bias corregido acelerado (BCa).
- Pulse en Continuar.
- Pulse en Aceptar en el cuadro de diálogo Explorar.

Estas selecciones generan la siguiente sintaxis de comandos:

```
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.
```

- Los comandos PRESERVE y RESTORE "recuerdan" el estado actual del generador de números aleatorios y restaurar el sistema al estado posterior a la finalización del método bootstrap.
- El comando SET define el generador de números aleatorios a Mersenne Twister y el índice a 592004, para que los resultados del muestreo bootstrap se puedan replicar exactamente. El comando SHOW muestra el índice en el resultado para futura referencia.
- El comando BOOTSTRAP solicita 1000 muestras bootstrap mediante nuevas muestras simples.
- El subcomando VARIABLES especifica que la variable *prevexp* se utiliza para determinar las muestras caso a caso. Los registros con valores perdidos en esta variable se eliminan del análisis.
- El subcomando CRITERIA , además de requerir el número de muestras de bootstrap, requiere intervalos de confianza de bootstrap de sesgo corregidos y acelerados en lugar de los intervalos de percentiles predefinidos.
- El procedimiento EXAMINE posterior a BOOTSTRAP se ejecuta en cada una de las muestras bootstrap.
- El subcomando PLOT desactiva el resultado de la representación.
- El resto de opciones están establecidas en sus valores por defecto.

Descriptivos

Figura 3-10
Tabla Descriptivos con intervalos de confianza autodocimantes

					Bootstrap[a]			
							Intervalo de confianza al 95% de BCa	
			Estadístico	Error típ.	Sesgo	Error típ.	Inferior	Superior
Experiencia previa (meses)	Media		95.86	4.804	-.01	4.86	86.39	105.20
	Intervalo de confianza para la media al 95%	Límite inferior	86.42					
		Límite superior	105.30					
	Media recortada al 5%		84.64		.02	4.94	75.38	94.21
	Mediana		55.00		-.11	3.66	50.00	60.00
	Varianza		10938.281		18.783	977.081	8954.509	13057.229
	Desv. típ.		104.586		-.015	4.689	94.644	114.245
	Mínimo		0					
	Máximo		476					
	Rango		476					
	Amplitud intercuartil		121		-1	10	103	137
	Asimetría		1.510	.112	.006	.110	1.284	1.768
	Curtosis		1.696	.224	.040	.463	.823	2.876

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

La tabla descriptivos contiene un número de intervalos de confianza de estadísticos y autodocimantes de esos estadísticos. El intervalo de confianza autodocimante de la media (86,39, 105,20) es similar al intervalo de confianza paramétrico (86,42, 105,30) y sugiere que el empleado "típico" tiene unos 7-9 años de experiencia previa. Sin embargo, *Experiencia anterior (meses)* tiene una distribución asimétrica, que convierte a la media en un indicador menos deseable del salario actual "típico" que la mediana. El intervalo de confianza autodocimante de la mediana (50,00, 60,00) es más estrecho e inferior que el intervalo de confianza de la media y sugiere que el

empleado “típico” tiene unos 4-5 años de experiencia previa. El uso de muestreo autodocimante ha hecho posible obtener un intervalo de valores que representen mejor la experiencia típica anterior.

Uso de muestreo autodocimante para seleccionar mejores predictores

Durante una revisión de los registros de los empleados, los directivos también están interesados en determinar los factores que están asociados con los aumentos de salarios de los empleados, al definir un modelo lineal de la diferencia entre el salario inicial y el actual. Al realizar un muestreo autodocimante de un modelo lineal, puede utilizar métodos de muestreado sucesivo especiales (muestreo autodocimante residual y wild) para obtener resultados más precisos.

Esta información se recoge en el archivo *Employee data.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A el p. 31.

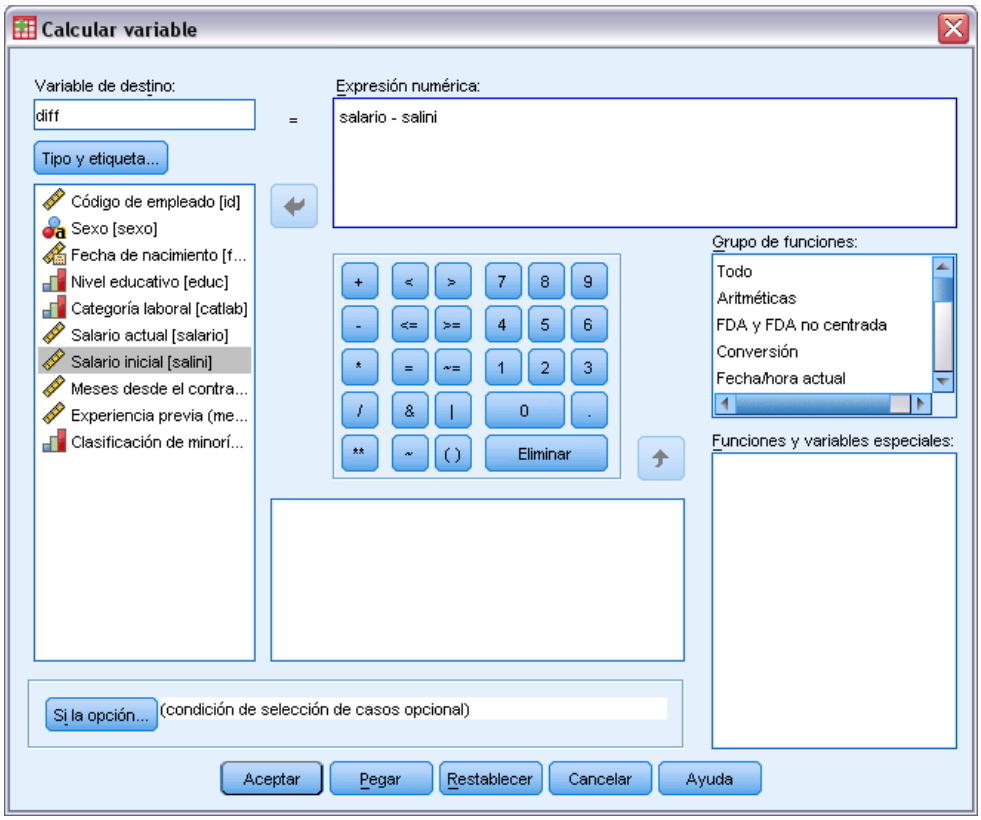
Nota: Este ejemplo utiliza el procedimiento MLG Univariante y requiere la opción Statistics Base.

Preparación de datos

En primer lugar debe calcular la diferencia entre Salario actual y Salario inicial.

- Seleccione en los menús:
  Transformar > Calcular variable...

Figura 3-11
Cuadro de diálogo Calcular variable



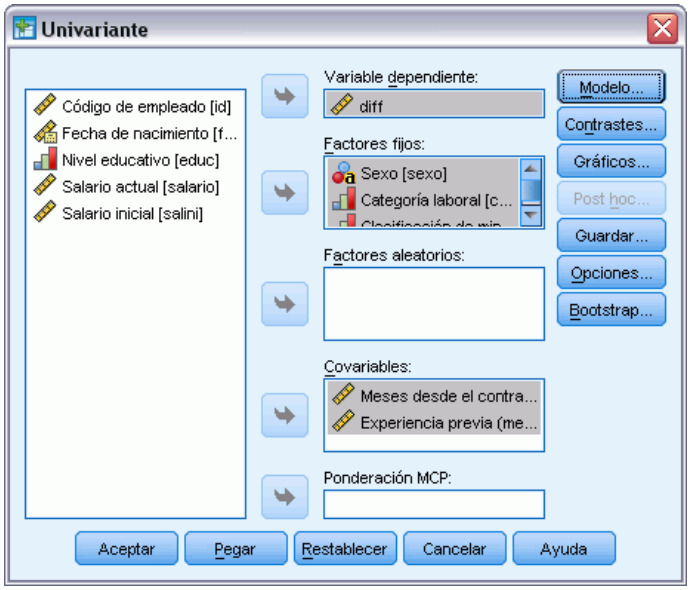
- Escriba diff como variable de destino.
- Escriba salario-iniciosalario como expresión numérica.
- Pulse en Aceptar.

Ejecución del análisis

Para ejecutar MLG Univariante con muestreo autodocimante residual y wild, necesita crear residuos.

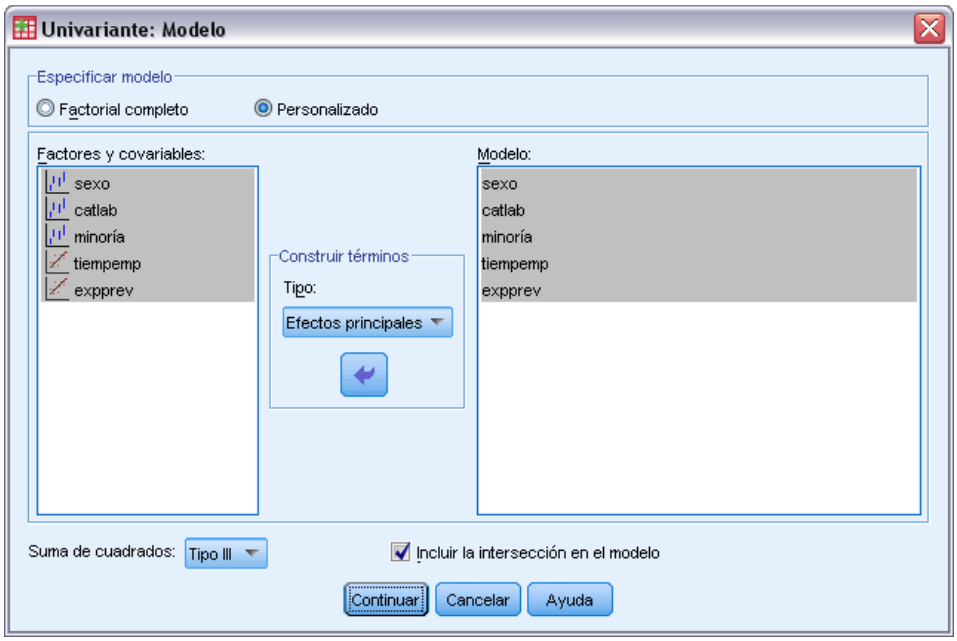
- Seleccione en los menús:
  Analizar > Modelo lineal general > Univariante...

Figura 3-12
Cuadro de diálogo principal MLG Univariante



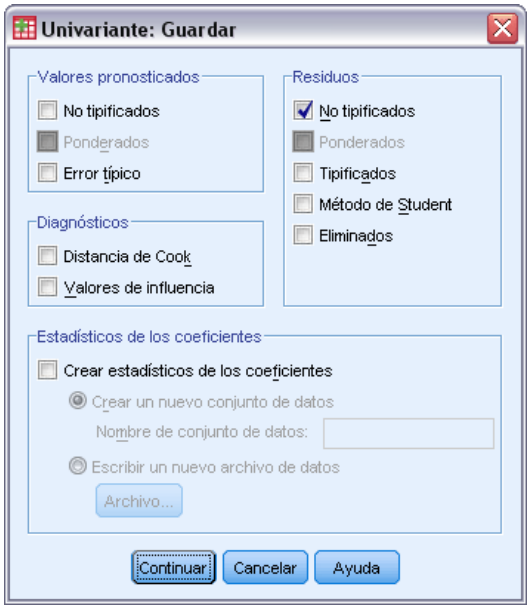
- Seleccione *diff* como la variable dependiente.
- Seleccione *Género [gender]*, *Categoría laboral [gender]* y *Clasificación étnica [minority]* como factores fijos.
- Seleccione *Meses desde el contrato [jobtime]* y *Experiencia anterior (meses) [prevexp]* como covariables.
- Pulse en Modelo.

Figura 3-13
Cuadro de diálogo Modelo



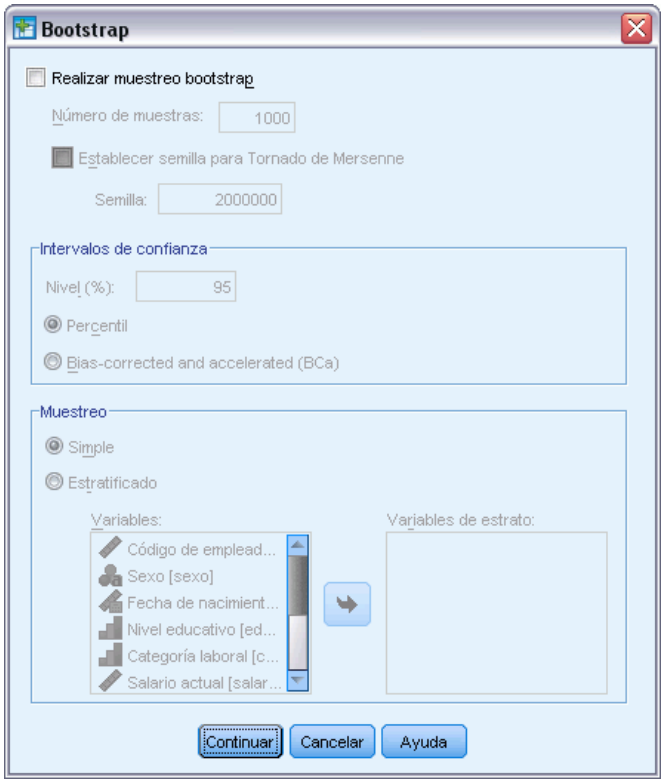
- Seleccione Personalizado y seleccione Efectos principales en la lista desplegable Construir términos.
- Seleccione *gender* hasta *prevexp* como términos de modelo.
- Pulse en Continuar.
- Pulse en Guardar en el cuadro de diálogo MLG Univariante.

Figura 3-14
Cuadro de diálogo Guardar



- Seleccione No tipificados en el grupo Residuos.
- Pulse en Continuar.
- Pulse en Autodocimante en el cuadro de diálogo MLG Univariante.

Figura 3-15
Cuadro de diálogo Autodocimante

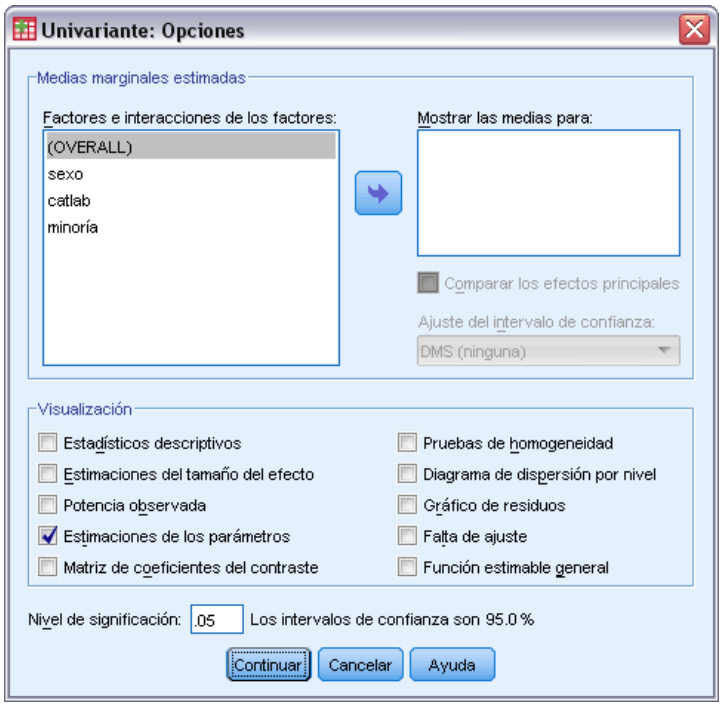


La configuración autodocimante se mantiene en los cuadros de diálogo que admiten el muestreo autodocimante. Mientras el muestreo autodocimante esté activado no se podrán guardar nuevas variables en el conjunto de datos, así que deberá asegurarse de que está desactivado.

- Si es necesario, elimine la selección de Ejecutar bootstrapping.
- Pulse en Aceptar en el cuadro de diálogo MLG Univariante. El conjunto de datos contiene ahora una nueva variable, *RES_1*, que contiene los residuos no tipificados del modelo.
- Active el cuadro de diálogo MLG Univariante y pulse Guardar.

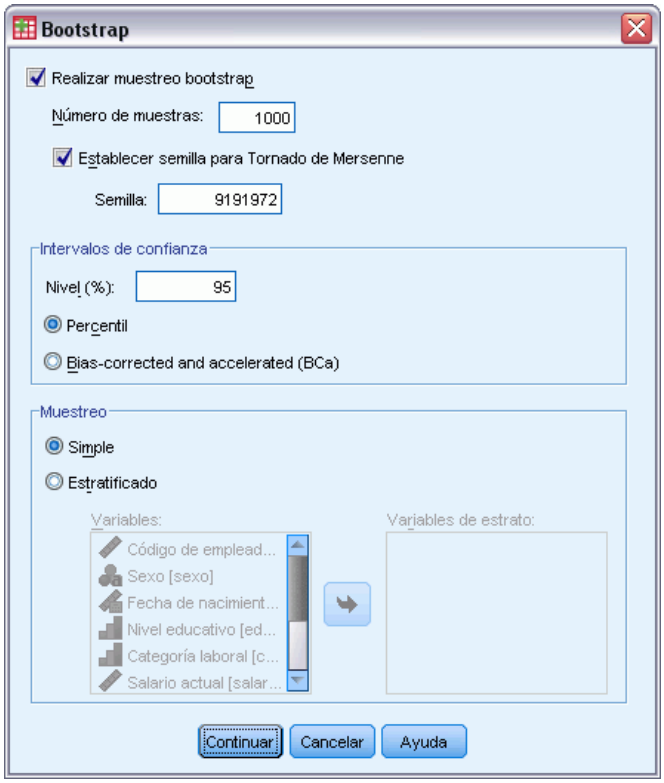
- Cancele la selección de No tipificados y pulse Continuar y Opciones en el cuadro de diálogo MLG Univariante.

Figura 3-16
Cuadro de diálogo Opciones



- Seleccione Estimaciones de los parámetros en la sección Mostrar.
- Pulse en Continuar.
- Pulse en Autodocimante en el cuadro de diálogo MLG Univariante.

Figura 3-17
Cuadro de diálogo Autodocimante



- Seleccione Ejecutar bootstrapping.
- Para replicar los resultados de este ejemplo de forma exacta, seleccione Establecer semilla para Tornado de Mersenne e introduzca 9191972 como semilla.
- No hay opciones para ejecutar muestreo autodocimante wild en los cuadros de diálogo, por lo que tendrá que pulsar Continuar y, a continuación, Pegar en el cuadro de diálogo MLG Univariante.

Estas selecciones generan la siguiente sintaxis de comandos:

```
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.
```

Para ejecutar muestreo autodocimante wild, edite la palabra clave `METHOD` del subcomando `SAMPLING` a `METHOD=WILD(RESIDUALS=RES_1)`.

El conjunto "final" de la sintaxis de comandos tendrá la siguiente apariencia:

```
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.
```

- Los comandos `PRESERVE` y `RESTORE` "recuerdan" el estado actual del generador de números aleatorios y restaurar el sistema al estado posterior a la finalización del método bootstrap.
- El comando `SET` define el generador de números aleatorios a Mersenne Twister y el índice a 9191972, para que los resultados del muestreo bootstrap se puedan replicar exactamente. El comando `SHOW` muestra el índice en el resultado para futura referencia.
- El comando `BOOTSTRAP` requiere 1000 muestras de bootstrap con muestreo wild y *RES_1* como la variable que contiene los residuos.
- El subcomando `VARIABLES` especifica que *diff* es la variable objetivo del modelo lineal. Esta variable y *gender*, *jobcat*, *minority*, *jobtime* y *prevexp* se utilizan para determinar las muestras caso a caso. Los registros con valores perdidos en estas variables se eliminan del análisis.
- El subcomando `CRITERIA` , además de requerir el número de muestras de bootstrap, requiere intervalos de confianza de bootstrap de sesgo corregidos y acelerados en lugar de los intervalos de percentiles predefinidos.
- El procedimiento `UNIANOVA` posterior a `BOOTSTRAP` se ejecuta en cada muestra bootstrap y produce estimaciones de los parámetros para los datos originales. Además, los estadísticos combinados se producen para los coeficientes del modelo.

Estimaciones de los parámetros

Figura 3-18
Estimaciones de los parámetros

Variable dependiente:diff

Parámetro	B	Error típ.	t	Sig.	Intervalo de confianza 95%	
					Límite inferior	Límite superior
Intersección	18703.761	2961.969	6.315	.000	12883.323	24524.199
[sexo=h]	4085.253	726.416	5.624	.000	2657.804	5512.701
[sexo=m]	0[a]	.	.	.	.	.
[catlab=1]	-17717.706	939.798	-18.853	.000	-19564.463	-15870.949
[catlab=2]	-13101.918	1780.683	-7.358	.000	-16601.061	-9602.776
[catlab=3]	0[a]	.	.	.	.	.
[minoría=0]	1332.363	819.349	1.626	.105	-277.705	2942.431
[minoría=1]	0[a]	.	.	.	.	.
tiempemp	145.539	32.586	4.466	.000	81.505	209.572
expprev	-21.423	3.575	-5.993	.000	-28.447	-14.398

a.Al parámetro se le ha asignado el valor cero porque es redundante.

La tabla Estimaciones de los parámetros muestra las estimaciones normales sin muestreo autodocimante de los parámetros de los términos de modelo. El valor de significación de 0,105 para *[minority=0]* es mayor que 0,05, lo que sugiere que *Clasificación étnica* no tiene ningún efecto en los aumentos de los salarios.

Figura 3-19
Estimaciones de parámetros autodocimantes

Variable dependiente:diff

Parámetro	B	Bootstrap[a]				
		Sesgo	Error típ.	Sig. (bilateral)	Intervalo de confianza 95%	
					Inferior	Superior
Intersección	18703.761	-62.604	3330.877	.001	12141.023	24980.359
[sexo=h]	4085.253	-32.480	622.971	.001	2892.131	5365.321
[sexo=m]	0	0	0		0	0
[catlab=1]	-17717.706	46.324	1454.230	.001	-20671.451	-14889.507
[catlab=2]	-13101.918	47.958	1753.311	.001	-16658.596	-9671.891
[catlab=3]	0	0	0		0	0
[minoría=0]	1332.363	-10.592	651.144	.012	57.831	2642.534
[minoría=1]	0	0	0		0	0
tiempemp	145.539	.707	35.285	.001	79.081	217.761
expprev	-21.423	-.065	2.859	.001	-27.533	-16.055

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

Ahora mire la tabla Estimaciones de parámetros autodocimantes. En la columna Error típico, verá que los errores típicos paramétricos de algunos coeficientes, como intersección, son demasiado pequeños en comparación con las estimaciones autodocimantes y los intervalos de confianza son mayores. En algunos coeficientes, como *[minority=0]*, los errores típicos paramétricos eran demasiado grandes y el valor de significación de 0,006 en los resultados autodocimantes, menor de 0,05, muestra que la diferencia observada en aumentos de salarios entre los empleados

pertenecientes a minorías étnicas o no no obedecen a las posibilidades. Los directivos saben ahora que merece la pena investigar más a fondo esta diferencia para determinar sus posibles causas.

Lecturas recomendadas

Consulte los siguientes textos si desea obtener más información acerca de muestreos autodocimantes:

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

Shao, J., y D. Tu. 1995. *The Jackknife and Bootstrap*. Nueva York: Springer.

Apéndice

A

Archivos muestrales

Los archivos muestrales instalados con el producto se encuentran en el subdirectorio *Samples* del directorio de instalación. Hay una carpeta independiente dentro del subdirectorio Samples para cada uno de los siguientes idiomas: Inglés, francés, alemán, italiano, japonés, coreano, polaco, ruso, chino simplificado, español y chino tradicional.

No todos los archivos muestrales están disponibles en todos los idiomas. Si un archivo muestral no está disponible en un idioma, esa carpeta de idioma contendrá una versión en inglés del archivo muestral.

Descripciones

A continuación, se describen brevemente los archivos muestrales usados en varios ejemplos que aparecen a lo largo de la documentación.

- **accidents.sav.** Archivo de datos hipotéticos sobre una compañía de seguros que estudia los factores de riesgo de edad y género que influyen en los accidentes de automóviles de una región determinada. Cada caso corresponde a una clasificación cruzada de categoría de edad y género.
- **adl.sav.** Archivo de datos hipotéticos relativo a los esfuerzos para determinar las ventajas de un tipo propuesto de tratamiento para pacientes que han sufrido un derrame cerebral. Los médicos dividieron de manera aleatoria a pacientes (mujeres) que habían sufrido un derrame cerebral en dos grupos. El primer grupo recibió el tratamiento físico estándar y el segundo recibió un tratamiento emocional adicional. Tres meses después de los tratamientos, se puntuaron las capacidades de cada paciente para realizar actividades cotidianas como variables ordinales.
- **advert.sav.** Archivo de datos hipotéticos sobre las iniciativas de un minorista para examinar la relación entre el dinero invertido en publicidad y las ventas resultantes. Para ello, se recopilaron las cifras de ventas anteriores y los costes de publicidad asociados.
- **aflatoxin.sav.** Archivo de datos hipotéticos sobre las pruebas realizadas en las cosechas de maíz con relación a la aflatoxina, un veneno cuya concentración varía ampliamente en los rendimientos de cultivo y entre los mismos. Un procesador de grano ha recibido 16 muestras de cada uno de los 8 rendimientos de cultivo y ha medido los niveles de aflatoxinas en partes por millón (PPM).
- **aflatoxin20.sav.** Este archivo de datos contiene las medidas de aflatoxina de cada una de las 16 muestras de los rendimientos 4 y 8 procedentes del archivo de datos *aflatoxin.sav*.
- **anorectic.sav.** Mientras trabajaban en una sintomatología estandarizada del comportamiento anoréxico/bulímico, los investigadores (Van der Ham, Meulman, Van Strien, y Van Engeland, 1997) realizaron un estudio de 55 adolescentes con trastornos de la alimentación conocidos. Cada paciente fue examinado cuatro veces durante cuatro años, lo que representa un total

de 220 observaciones. En cada observación, se puntuó a los pacientes por cada uno de los 16 síntomas. Faltan las puntuaciones de los síntomas para el paciente 71 en el tiempo 2, el paciente 76 en el tiempo 2 y el paciente 47 en el tiempo 3, lo que nos deja 217 observaciones válidas.

- **autoaccidents.sav.** Archivo de datos hipotéticos sobre las iniciativas de un analista de seguros para elaborar un modelo del número de accidentes de automóvil por conductor teniendo en cuenta la edad y el género del conductor. Cada caso representa un conductor diferente y registra el sexo, la edad en años y el número de accidentes de automóvil del conductor en los últimos cinco años.
- **band.sav** Este archivo de datos contiene las cifras de ventas semanales hipotéticas de CD de música de una banda. También se incluyen datos para tres variables predictoras posibles.
- **bankloan.sav.**Archivo de datos hipotéticos sobre las iniciativas de un banco para reducir la tasa de moras de créditos. El archivo contiene información financiera y demográfica de 850 clientes anteriores y posibles clientes. Los primeros 700 casos son clientes a los que anteriormente se les ha concedido un préstamo. Al menos 150 casos son posibles clientes cuyos riesgos de crédito el banco necesita clasificar como positivos o negativos.
- **bankloan_binning.sav.** Archivo de datos hipotéticos que contiene información financiera y demográfica sobre 5.000 clientes anteriores.
- **behavior.sav.** En un ejemplo clásico (Price y Bouffard, 1974), se pidió a 52 estudiantes que valoraran las combinaciones de 15 situaciones y 15 comportamientos en una escala de 10 puntos que oscilaba entre 0 =“extremadamente apropiado” y 9=“extremadamente inapropiado”. Los valores promediados respecto a los individuos se toman como disimilaridades.
- **behavior_ini.sav.** Este archivo de datos contiene una configuración inicial para una solución bidimensional de *behavior.sav*.
- **brakes.sav.** Archivo de datos hipotéticos sobre el control de calidad de una fábrica que produce frenos de disco para automóviles de alto rendimiento. El archivo de datos contiene las medidas del diámetro de 16 discos de cada una de las 8 máquinas de producción. El diámetro objetivo para los frenos es de 322 milímetros.
- **breakfast.sav.**En un estudio clásico (Green y Rao, 1972), se pidió a 21 estudiantes de administración de empresas de la Wharton School y sus cónyuges que ordenaran 15 elementos de desayuno por orden de preferencia, de 1=“más preferido” a 15=“menos preferido”. Sus preferencias se registraron en seis escenarios distintos, de “Preferencia global” a “Aperitivo, con bebida sólo”.
- **breakfast-overall.sav.** Este archivo de datos sólo contiene las preferencias de elementos de desayuno para el primer escenario, “Preferencia global”.
- **broadband_1.sav** Archivo de datos hipotéticos que contiene el número de suscriptores, por región, a un servicio de banda ancha nacional. El archivo de datos contiene números de suscriptores mensuales para 85 regiones durante un período de cuatro años.
- **broadband_2.sav** Este archivo de datos es idéntico a *broadband_1.sav* pero contiene datos para tres meses adicionales.
- **car_insurance_claims.sav.** Un conjunto de datos presentados y analizados en otro lugar (McCullagh y Nelder, 1989) estudia las reclamaciones por daños en vehículos. La cantidad de reclamaciones media se puede modelar como si tuviera una distribución Gamma, mediante

una función de enlace inversa para relacionar la media de la variable dependiente con una combinación lineal de la edad del asegurado, el tipo de vehículo y la antigüedad del vehículo. El número de reclamaciones presentadas se puede utilizar como una ponderación de escalamiento.

- **car_sales.sav.** Este archivo de datos contiene estimaciones de ventas, precios de lista y especificaciones físicas hipotéticas de varias marcas y modelos de vehículos. Los precios de lista y las especificaciones físicas se han obtenido de *edmunds.com* y de sitios de fabricantes.
- **car_sales_uprepared.sav.**Ésta es una versión modificada de *car_sales.sav* que no incluye ninguna versión transformada de los campos.
- **carpet.sav** En un ejemplo muy conocido (Green y Wind, 1973), una compañía interesada en sacar al mercado un nuevo limpiador de alfombras desea examinar la influencia de cinco factores sobre la preferencia del consumidor: diseño del producto, marca comercial, precio, sello de *buen producto para el hogar* y garantía de devolución del importe. Hay tres niveles de factores para el diseño del producto, cada uno con una diferente colocación del cepillo del aplicador; tres nombres comerciales (*K2R*, *Glory* y *Bissell*); tres niveles de precios; y dos niveles (no o sí) para los dos últimos factores. Diez consumidores clasificaron 22 perfiles definidos por estos factores. La variable *Preferencia* contiene el rango de las clasificaciones medias de cada perfil. Las clasificaciones inferiores corresponden a preferencias elevadas. Esta variable refleja una medida global de la preferencia de cada perfil.
- **carpet_prefs.sav** Este archivo de datos se basa en el mismo ejemplo que el descrito para *carpet.sav*, pero contiene las clasificaciones reales recogidas de cada uno de los 10 consumidores. Se pidió a los consumidores que clasificaran los 22 perfiles de los productos empezando por el menos preferido. Las variables desde *PREF1* hasta *PREF22* contienen los ID de los perfiles asociados, como se definen en *carpet_plan.sav*.
- **catalog.sav**Este archivo de datos contiene cifras de ventas mensuales hipotéticas de tres productos vendidos por una compañía de venta por catálogo. También se incluyen datos para cinco variables predictoras posibles.
- **catalog_seasfac.sav**Este archivo de datos es igual que *catalog.sav*, con la excepción de que incluye un conjunto de factores estacionales calculados a partir del procedimiento Descomposición estacional junto con las variables de fecha que lo acompañan.
- **cellular.sav.** Archivo de datos hipotéticos sobre las iniciativas de una compañía de telefonía móvil para reducir el abandono de clientes. Las puntuaciones de propensión al abandono de clientes se aplican a las cuentas, oscilando de 0 a 100. Las cuentas con una puntuación de 50 o superior pueden estar buscando otros proveedores.
- **ceramics.sav.**Archivo de datos hipotéticos sobre las iniciativas de un fabricante para determinar si una nueva aleación de calidad tiene una mayor resistencia al calor que una aleación estándar. Cada caso representa una prueba independiente de una de las aleaciones; la temperatura a la que registró el fallo del rodamiento.
- **cereal.sav.** Archivo de datos hipotéticos sobre una encuesta realizada a 880 personas sobre sus preferencias en el desayuno, teniendo también en cuenta su edad, sexo, estado civil y si tienen un estilo de vida activo o no (en función de si practican ejercicio al menos dos veces a la semana). Cada caso representa un encuestado diferente.
- **clothing_defects.sav.** Archivo de datos hipotéticos sobre el proceso de control de calidad en una fábrica de prendas. Los inspectores toman una muestra de prendas de cada lote producido en la fábrica, y cuentan el número de prendas que no son aceptables.

- **coffee.sav.** Este archivo de datos pertenece a las imágenes percibidas de seis marcas de café helado (Kennedy, Riquier, y Sharp, 1996). Para cada uno de los 23 atributos de imagen de café helado, los encuestados seleccionaron todas las marcas que quedaban descritas por el atributo. Las seis marcas se denotan AA, BB, CC, DD, EE y FF para mantener la confidencialidad.
- **contacts.sav.**Archivo de datos hipotéticos sobre las listas de contactos de un grupo de representantes de ventas de ordenadores de empresa. Cada uno de los contactos está categorizado por el departamento de la compañía en el que trabaja y su categoría en la compañía. Además, también se registran los importes de la última venta realizada, el tiempo transcurrido desde la última venta y el tamaño de la compañía del contacto.
- **creditpromo.sav.** Archivo de datos hipotéticos sobre las iniciativas de unos almacenes para evaluar la eficacia de una promoción de tarjetas de crédito reciente. Para este fin, se seleccionaron aleatoriamente 500 titulares. La mitad recibieron un anuncio promocionando una tasa de interés reducida sobre las ventas realizadas en los siguientes tres meses. La otra mitad recibió un anuncio estacional estándar.
- **customer_dbase.sav.** Archivo de datos hipotéticos sobre las iniciativas de una compañía para usar la información de su almacén de datos para realizar ofertas especiales a los clientes con más probabilidades de responder. Se seleccionó un subconjunto de la base de clientes aleatoriamente a quienes se ofrecieron las ofertas especiales y sus respuestas se registraron.
- **customer_information.sav.** Archivo de datos hipotéticos que contiene la información de correo del cliente, como el nombre y la dirección.
- **customer_subset.sav.** Un subconjunto de 80 casos de *customer_dbase.sav*.
- **customers_model.sav.** Este archivo contiene datos hipotéticos sobre los individuos a los que va dirigida una campaña de marketing. Estos datos incluyen información demográfica, un resumen del historial de compras y si cada individuo respondió a la campaña. Cada caso representa un individuo diferente.
- **customers_new.sav.** Este archivo contiene datos hipotéticos sobre los individuos que son candidatos potenciales para una campaña de marketing. Estos datos incluyen información demográfica y un resumen del historial de compras de cada individuo. Cada caso representa un individuo diferente.
- **debate.sav.** Archivos de datos hipotéticos sobre las respuestas emparejadas de una encuesta realizada a los asistentes a un debate político antes y después del debate. Cada caso corresponde a un encuestado diferente.
- **debate_aggregate.sav.** Archivo de datos hipotéticos que agrega las respuestas de *debate.sav*. Cada caso corresponde a una clasificación cruzada de preferencias antes y después del debate.
- **demo.sav.** Archivos de datos hipotéticos sobre una base de datos de clientes adquirida con el fin de enviar por correo ofertas mensuales. Se registra si el cliente respondió a la oferta, junto con información demográfica diversa.
- **demo_cs_1.sav.**Archivo de datos hipotéticos sobre el primer paso de las iniciativas de una compañía para recopilar una base de datos de información de encuestas. Cada caso corresponde a una ciudad diferente, y se registra la identificación de la ciudad, la región, la provincia y el distrito.
- **demo_cs_2.sav.**Archivo de datos hipotéticos sobre el segundo paso de las iniciativas de una compañía para recopilar una base de datos de información de encuestas. Cada caso corresponde a una unidad familiar diferente de las ciudades seleccionadas en el primer paso, y

se registra la identificación de la unidad, la subdivisión, la ciudad, el distrito, la provincia y la región. También se incluye la información de muestreo de las primeras dos etapas del diseño.

- **demo_cs.sav.**Archivo de datos hipotéticos que contiene información de encuestas recopilada mediante un diseño de muestreo complejo. Cada caso corresponde a una unidad familiar distinta, y se recopila información demográfica y de muestreo diversa.
- **dmdata.sav.** Éste es un archivo de datos hipotéticos que contiene información demográfica y de compras para una empresa de marketing directo. *dmdata2.sav* contiene información para un subconjunto de contactos que recibió un envío de prueba, y *dmdata3.sav* contiene información sobre el resto de contactos que no recibieron el envío de prueba.
- **dietstudy.sav.**Este archivo de datos hipotéticos contiene los resultados de un estudio sobre la "dieta Stillman" (Rickman, Mitchell, Dingman, y Dalen, 1974). Cada caso corresponde a un sujeto distinto y registra sus pesos antes y después de la dieta en libras y niveles de triglicéridos en mg/100 ml.
- **dvdplayer.sav.** Archivo de datos hipotéticos sobre el desarrollo de un nuevo reproductor de DVD. El equipo de marketing ha recopilado datos de grupo de enfoque mediante un prototipo. Cada caso corresponde a un usuario encuestado diferente y registra información demográfica sobre los encuestados y sus respuestas a preguntas acerca del prototipo.
- **german_credit.sav.**Este archivo de datos se toma del conjunto de datos "German credit" de las Repository of Machine Learning Databases (Blake y Merz, 1998) de la Universidad de California, Irvine.
- **grocery_1month.sav.** Este archivo de datos hipotéticos es el archivo de datos *grocery_coupons.sav* con las compras semanales "acumuladas" para que cada caso corresponda a un cliente diferente. Algunas de las variables que cambiaban semanalmente desaparecen de los resultados, y la cantidad gastada registrada se convierte ahora en la suma de las cantidades gastadas durante las cuatro semanas del estudio.
- **grocery_coupons.sav.** Archivo de datos hipotéticos que contiene datos de encuestas recopilados por una cadena de tiendas de alimentación interesada en los hábitos de compra de sus clientes. Se sigue a cada cliente durante cuatro semanas, y cada caso corresponde a un cliente-semana distinto y registra información sobre dónde y cómo compran los clientes, incluida la cantidad que invierten en comestibles durante esa semana.
- **guttman.sav.**Bell (Bell, 1961) presentó una tabla para ilustrar posibles grupos sociales. Guttman (Guttman, 1968) utilizó parte de esta tabla, en la que se cruzaron cinco variables que describían elementos como la interacción social, sentimientos de pertenencia a un grupo, proximidad física de los miembros y grado de formalización de la relación con siete grupos sociales teóricos, incluidos multitudes (por ejemplo, las personas que acuden a un partido de fútbol), espectadores (por ejemplo, las personas que acuden a un teatro o de una conferencia), públicos (por ejemplo, los lectores de periódicos o los espectadores de televisión), muchedumbres (como una multitud pero con una interacción mucho más intensa), grupos primarios (íntimos), grupos secundarios (voluntarios) y la comunidad moderna (confederación débil que resulta de la proximidad cercana física y de la necesidad de servicios especializados).
- **health_funding.sav.** Archivo de datos hipotéticos que contiene datos sobre inversión en sanidad (cantidad por 100 personas), tasas de enfermedad (índice por 10.000 personas) y visitas a centros de salud (índice por 10.000 personas). Cada caso representa una ciudad diferente.

- **hivassay.sav.** Archivo de datos hipotéticos sobre las iniciativas de un laboratorio farmacéutico para desarrollar un ensayo rápido para detectar la infección por VIH. Los resultados del ensayo son ocho tonos de rojo con diferentes intensidades, donde los tonos más oscuros indican una mayor probabilidad de infección. Se llevó a cabo una prueba de laboratorio de 2.000 muestras de sangre, de las cuales una mitad estaba infectada con el VIH y la otra estaba limpia.
- **hourlywagedata.sav.** Archivo de datos hipotéticos sobre los salarios por horas de enfermeras de puestos de oficina y hospitales y con niveles distintos de experiencia.
- **insurance_claims.sav.** Éste es un archivo de datos hipotéticos sobre una compañía de seguros que desee generar un modelo para etiquetar las reclamaciones sospechosas y potencialmente fraudulentas. Cada caso representa una reclamación diferente.
- **insure.sav.** Archivo de datos hipotéticos sobre una compañía de seguros que estudia los factores de riesgo que indican si un cliente tendrá que hacer una reclamación a lo largo de un contrato de seguro de vida de 10 años. Cada caso del archivo de datos representa un par de contratos (de los que uno registró una reclamación y el otro no), agrupados por edad y sexo.
- **judges.sav.** Archivo de datos hipotéticos sobre las puntuaciones concedidas por jueces cualificados (y un aficionado) a 300 actuaciones gimnásticas. Cada fila representa una actuación diferente; los jueces vieron las mismas actuaciones.
- **kinship_dat.sav.** Rosenberg y Kim (Rosenberg y Kim, 1975) comenzaron a analizar 15 términos de parentesco [tía, hermano, primos, hija, padre, nieta, abuelo, abuela, nieto, madre, sobrino, sobrina, hermana, hijo, tío]. Le pidieron a cuatro grupos de estudiantes universitarios (dos masculinos y dos femeninos) que ordenaran estos grupos según las similitudes. A dos grupos (uno masculino y otro femenino) se les pidió que realizaran la ordenación dos veces, pero que la segunda ordenación la hicieran según criterios distintos a los de la primera. Así, se obtuvo un total de seis "fuentes". Cada fuente se corresponde con una matriz de proximidades de 15×15 cuyas casillas son iguales al número de personas de una fuente menos el número de veces que se particionaron los objetos en esa fuente.
- **kinship_ini.sav.** Este archivo de datos contiene una configuración inicial para una solución tridimensional de *kinship_dat.sav*.
- **kinship_var.sav.** Este archivo de datos contiene variables independientes *sexo*, *gener*(ación), y *grado* (de separación) que se pueden usar para interpretar las dimensiones de una solución para *kinship_dat.sav*. Concretamente, se pueden usar para restringir el espacio de la solución a una combinación lineal de estas variables.
- **marketvalues.sav.** Archivo de datos sobre las ventas de casas en una nueva urbanización de Algonquin, Ill., durante los años 1999 y 2000. Los datos de estas ventas son públicos.
- **nhis2000_subset.sav.** La National Health Interview Survey (NHIS, encuesta del Centro Nacional de Estadísticas de Salud de EE.UU.) es una encuesta detallada realizada entre la población civil de Estados Unidos. Las encuestas se realizaron en persona a una muestra representativa de las unidades familiares del país. Se recogió tanto la información demográfica como las observaciones acerca del estado y los hábitos de salud de los integrantes de cada unidad familiar. Este archivo de datos contiene un subconjunto de información de la encuesta de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Archivo de datos y documentación de uso público. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Fecha de acceso: 2003.

- **ozono.sav.** Los datos incluyen 330 observaciones de seis variables meteorológicas para pronosticar la concentración de ozono a partir del resto de variables. Los investigadores anteriores(Breiman y Friedman, 1985), (Hastie y Tibshirani, 1990) han encontrado que no hay linealidad entre estas variables, lo que dificulta los métodos de regresión típica.
- **pain_medication.sav.** Este archivo de datos hipotéticos contiene los resultados de una prueba clínica sobre medicación antiinflamatoria para tratar el dolor artrítico crónico. Resulta de particular interés el tiempo que tarda el fármaco en hacer efecto y cómo se compara con una medicación existente.
- **patient_los.sav.** Este archivo de datos hipotéticos contiene los registros de tratamiento de pacientes que fueron admitidos en el hospital ante la posibilidad de sufrir un infarto de miocardio (IM o "ataque al corazón"). Cada caso corresponde a un paciente distinto y registra diversas variables relacionadas con su estancia hospitalaria.
- **patlos_sample.sav.** Este archivo de datos hipotéticos contiene los registros de tratamiento de una muestra de pacientes que recibieron trombolíticos durante el tratamiento del infarto de miocardio (IM o "ataque al corazón"). Cada caso corresponde a un paciente distinto y registra diversas variables relacionadas con su estancia hospitalaria.
- **polishing.sav.** Archivo de datos "Nambeware Polishing Times" (Tiempo de pulido de metal) de la biblioteca de datos e historiales. Contiene datos sobre las iniciativas de un fabricante de cuberterías de metal (Nambe Mills, Santa Fe, N. M.) para planificar su programa de producción. Cada caso representa un artículo distinto de la línea de productos. Se registra el diámetro, el tiempo de pulido, el precio y el tipo de producto de cada artículo.
- **poll_cs.sav.** Archivo de datos hipotéticos sobre las iniciativas de los encuestadores para determinar el nivel de apoyo público a una ley antes de una asamblea legislativa. Los casos corresponden a votantes registrados. Cada caso registra el condado, la población y el vecindario en el que vive el votante.
- **poll_cs_sample.sav.** Este archivo de datos hipotéticos contiene una muestra de los votantes enumerados en *poll_cs.sav*. La muestra se tomó según el diseño especificado en el archivo de plan *poll.csplan* y este archivo de datos registra las probabilidades de inclusión y las ponderaciones muestrales. Sin embargo, tenga en cuenta que debido a que el plan muestral hace uso de un método de probabilidad proporcional al tamaño (PPS), también existe un archivo que contiene las probabilidades de selección conjunta (*poll_jointprob.sav*). Las variables adicionales que corresponden a los datos demográficos de los votantes y sus opiniones sobre la propuesta de ley se recopilaron y añadieron al archivo de datos después de tomar la muestra.
- **property_assess.sav.** Archivo de datos hipotéticos sobre las iniciativas de un asesor del condado para mantener actualizada la evaluación de los valores de las propiedades utilizando recursos limitados. Los casos corresponden a las propiedades vendidas en el condado el año anterior. Cada caso del archivo de datos registra la población en que se encuentra la propiedad, el último asesor que visitó la propiedad, el tiempo transcurrido desde la última evaluación, la valoración realizada en ese momento y el valor de venta de la propiedad.
- **property_assess_cs.sav.** Archivo de datos hipotéticos sobre las iniciativas de un asesor de un estado para mantener actualizada la evaluación de los valores de las propiedades utilizando recursos limitados. Los casos corresponden a propiedades del estado. Cada caso del archivo de datos registra el condado, la población y el vecindario en el que se encuentra la propiedad, el tiempo transcurrido desde la última evaluación y la valoración realizada en ese momento.

- **property_assess_cs_sample.sav**Este archivo de datos hipotéticos contiene una muestra de las propiedades recogidas en *property_assess_cs.sav*. La muestra se tomó en función del diseño especificado en el archivo de plan *property_assess.csplan*, y este archivo de datos registra las probabilidades de inclusión y las ponderaciones muestrales. La variable adicional *Valor actual* se recopiló y añadió al archivo de datos después de tomar la muestra.
- **recidivism.sav.** Archivo de datos hipotéticos sobre las iniciativas de una agencia de orden público para comprender los índices de reincidencia en su área de jurisdicción. Cada caso corresponde a un infractor anterior y registra su información demográfica, algunos detalles de su primer delito y, a continuación, el tiempo transcurrido desde su segundo arresto, si ocurrió en los dos años posteriores al primer arresto.
- **recidivism_cs_sample.sav.** Archivo de datos hipotéticos sobre las iniciativas de una agencia de orden público para comprender los índices de reincidencia en su área de jurisdicción. Cada caso corresponde a un delincuente anterior, puesto en libertad tras su primer arresto durante el mes de junio de 2003 y registra su información demográfica, algunos detalles de su primer delito y los datos de su segundo arresto, si se produjo antes de finales de junio de 2006. Los delincuentes se seleccionaron de una muestra de departamentos según el plan de muestreo especificado en *recidivism_cs.csplan*. Como este plan utiliza un método de probabilidad proporcional al tamaño (PPS), también existe un archivo que contiene las probabilidades de selección conjunta (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** Archivo de datos hipotéticos que contiene datos de transacciones de compra, incluida la fecha de compra, los artículos adquiridos y el importe de cada transacción.
- **salesperformance.sav.** Archivo de datos hipotéticos sobre la evaluación de dos nuevos cursos de formación de ventas. Sesenta empleados, divididos en tres grupos, reciben formación estándar. Además, el grupo 2 recibe formación técnica; el grupo 3, un tutorial práctico. Cada empleado se sometió a un examen al final del curso de formación y se registró su puntuación. Cada caso del archivo de datos representa a un alumno distinto y registra el grupo al que fue asignado y la puntuación que obtuvo en el examen.
- **satisf.sav.** Archivo de datos hipotéticos sobre una encuesta de satisfacción llevada a cabo por una empresa minorista en cuatro tiendas. Se encuestó a 582 clientes en total y cada caso representa las respuestas de un único cliente.
- **screws.sav** Este archivo de datos contiene información acerca de las características de tornillos, pernos, clavos y tacos (Hartigan, 1975).
- **shampoo_ph.sav.**Archivo de datos hipotéticos sobre el control de calidad en una fábrica de productos para el cabello. Se midieron seis lotes de resultados distintos en intervalos regulares y se registró su pH. El intervalo objetivo es de 4,5 a 5,5.
- **ships.sav.** Un conjunto de datos presentados y analizados en otro lugar (McCullagh et al., 1989) sobre los daños en los cargueros producidos por las olas. Los recuentos de incidentes se pueden modelar como si ocurrieran con una tasa de Poisson dado el tipo de barco, el período de construcción y el período de servicio. Los meses de servicio agregados para cada casilla de la tabla formados por la clasificación cruzada de factores proporcionan valores para la exposición al riesgo.
- **site.sav.**Archivo de datos hipotéticos sobre las iniciativas de una compañía para seleccionar sitios nuevos para sus negocios en expansión. Se ha contratado a dos consultores para evaluar los sitios de forma independiente, quienes, además de un informe completo, han resumido cada sitio como una posibilidad “buena”, “media” o “baja”.

- **smokers.sav.**Este archivo de datos es un resumen de la encuesta sobre toxicomanía 1998 National Household Survey of Drug Abuse y es una muestra de probabilidad de unidades familiares americanas. (*http://dx.doi.org/10.3886/ICPSR02934*) Así, el primer paso de un análisis de este archivo de datos debe ser ponderar los datos para reflejar las tendencias de población.
- **stroke_clean.sav.** Este archivo de datos hipotéticos contiene el estado de una base de datos médica después de haberla limpiado mediante los procedimientos de la opción Preparación de datos.
- **stroke_invalid.sav.** Este archivo de datos hipotéticos contiene el estado inicial de una base de datos médica que incluye contiene varios errores de entrada de datos.
- **stroke_survival.** Este archivo de datos hipotéticos registra los tiempos de supervivencia de los pacientes que finalizan un programa de rehabilitación tras un ataque isquémico. Tras el ataque, la ocurrencia de infarto de miocardio, ataque isquémico o ataque hemorrágico se anotan junto con el momento en el que se produce el evento registrado. La muestra está truncada a la izquierda ya que únicamente incluye a los pacientes que han sobrevivido al final del programa de rehabilitación administrado tras el ataque.
- **stroke_valid.sav.** Este archivo de datos hipotéticos contiene el estado de una base de datos médica después de haber comprobado los valores mediante el procedimiento Validar datos. Sigue conteniendo casos potencialmente anómalos.
- **survey_sample.sav.** Este archivo de datos contiene datos de encuestas, incluyendo datos demográficos y diferentes medidas de actitud. Se basa en un subconjunto de variables de NORC General Social Survey de 1998, aunque algunos valores de datos se han modificado y que existen variables ficticias adicionales se han añadido para demostraciones.
- **telco.sav.** Archivo de datos hipotéticos sobre las iniciativas de una compañía de telecomunicaciones para reducir el abandono de clientes en su base de clientes. Cada caso corresponde a un cliente distinto y registra diversa información demográfica y de uso del servicio.
- **telco_extra.sav.** Este archivo de datos es similar al archivo de datos *telco.sav*, pero las variables de meses con servicio y gasto de clientes transformadas logarítmicamente se han eliminado y sustituido por variables de gasto del cliente transformadas logarítmicamente tipificadas.
- **telco_missing.sav.** Este archivo de datos es un subconjunto del archivo de datos *telco.sav*, pero algunos valores de datos demográficos se han sustituido con valores perdidos.
- **testmarket.sav.** Archivo de datos hipotéticos sobre los planes de una cadena de comida rápida para añadir un nuevo artículo a su menú. Hay tres campañas posibles para promocionar el nuevo producto, por lo que el artículo se presenta en ubicaciones de varios mercados seleccionados aleatoriamente. Se utiliza una promoción diferente en cada ubicación y se registran las ventas semanales del nuevo artículo durante las primeras cuatro semanas. Cada caso corresponde a una ubicación semanal diferente.
- **testmarket_1month.sav.** Este archivo de datos hipotéticos es el archivo de datos *testmarket.sav* con las ventas semanales “acumuladas” para que cada caso corresponda a una ubicación diferente. Como resultado, algunas de las variables que cambiaban semanalmente desaparecen y las ventas registradas se convierten en la suma de las ventas realizadas durante las cuatro semanas del estudio.
- **tree_car.sav.** Archivo de datos hipotéticos que contiene datos demográficos y de precios de compra de vehículos.

- **tree_credit.sav** Archivo de datos hipotéticos que contiene datos demográficos y de historial de créditos bancarios.
- **tree_missing_data.sav** Archivo de datos hipotéticos que contiene datos demográficos y de historial de créditos bancarios con un elevado número de valores perdidos.
- **tree_score_car.sav.** Archivo de datos hipotéticos que contiene datos demográficos y de precios de compra de vehículos.
- **tree_textdata.sav.** Archivo de datos sencillos con dos variables diseñadas principalmente para mostrar el estado por defecto de las variables antes de realizar la asignación de nivel de medida y etiquetas de valor.
- **tv-survey.sav.** Archivo de datos hipotéticos sobre una encuesta dirigida por un estudio de TV que está considerando la posibilidad de ampliar la emisión de un programa de éxito. Se preguntó a 906 encuestados si verían el programa en distintas condiciones. Cada fila representa un encuestado diferente; cada columna es una condición diferente.
- **ulcer_recurrence.sav.** Este archivo contiene información parcial de un estudio diseñado para comparar la eficacia de dos tratamientos para prevenir la reaparición de úlceras. Constituye un buen ejemplo de datos censurados por intervalos y se ha presentado y analizado en otro lugar (Collett, 2003).
- **ulcer_recurrence_recoded.sav.** Este archivo reorganiza la información de *ulcer_recurrence.sav* para permitir modelar la probabilidad de eventos de cada intervalo del estudio en lugar de sólo la probabilidad de eventos al final del estudio. Se ha presentado y analizado en otro lugar (Collett et al., 2003).
- **verd1985.sav.** Archivo de datos sobre una encuesta (Verdegaal, 1985). Se han registrado las respuestas de 15 sujetos a 8 variables. Se han dividido las variables de interés en tres grupos. El conjunto 1 incluye *edad* y *ecivil*, el conjunto 2 incluye *mascota* y *noticia*, mientras que el conjunto 3 incluye *música* y *vivir*. Se escala *mascota* como nominal múltiple y *edad* como ordinal; el resto de variables se escalan como nominal simple.
- **virus.sav.**Archivo de datos hipotéticos sobre las iniciativas de un proveedor de servicios de Internet (ISP) para determinar los efectos de un virus en sus redes. Se ha realizado un seguimiento (aproximado) del porcentaje de tráfico de correos electrónicos infectados en sus redes a lo largo del tiempo, desde el momento en que se descubre hasta que la amenaza se contiene.
- **wheeze_steubenville.sav.** Subconjunto de un estudio longitudinal de los efectos sobre la salud de la polución del aire en los niños (Ware, Dockery, Spiro III, Speizer, y Ferris Jr., 1984). Los datos contienen medidas binarias repetidas del estado de las sibilancias en niños de Steubenville, Ohio, con edades de 7, 8, 9 y 10 años, junto con un registro fijo de si la madre era fumadora durante el primer año del estudio.
- **workprog.sav.** Archivo de datos hipotéticos sobre un programa de obras del gobierno que intenta colocar a personas desfavorecidas en mejores trabajos. Se siguió una muestra de participantes potenciales del programa, algunos de los cuales se seleccionaron aleatoriamente para entrar en el programa, mientras que otros no siguieron esta selección aleatoria. Cada caso representa un participante del programa diferente.

Apéndice
B

Notices

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing,

using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Bibliografía

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

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

Breiman, L., y 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., y D. V. Hinkley. 2006. *Bootstrap Methods and their Application*. : Cambridge University Press.

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

Green, P. E., y 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*. Nueva York: John Wiley and Sons.

Hastie, T., y R. Tibshirani. 1990. *Generalized additive models*. Londres: Chapman and Hall.

Kennedy, R., C. Riquier, y 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., y J. A. Nelder. 1989. *Modelos lineales generalizados*, 2nd ed. Londres: Chapman & Hall.

Price, R. H., y 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, y J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

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

Shao, J., y D. Tu. 1995. *The Jackknife and Bootstrap*. Nueva York: Springer.

Van der Ham, T., J. J. Meulman, D. C. Van Strien, y 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 (en neerlandés)*. Leiden: Department of Data Theory, University of Leiden.

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

Índice