IBM SPSS - Tablas personalizadas 26

Nota

Antes de utilizar esta información y el producto al que da soporte, lea la información del apartado "Avisos" en la página 19.

Información sobre el producto

Esta edición se aplica a la versión 26, release 0, modificación 0 de IBM® SPSS Statistics y a todos los releases y modificaciones posteriores hasta que se indique lo contrario en nuevas ediciones.

Contenido

Tablas personalizadas

Se han incluido las características de tablas personalizadas siguientes en SPSS Statistics Standard Edition o la opción Tablas personalizadas.

Interfaz Tablas personalizadas

Interfaz de generador de tablas

Tablas personalizadas utiliza una interfaz de generador de tablas mediante las acciones simples de arrastrar y soltar que le permite obtener una vista previa de su tabla cuando selecciona variables y opciones. También ofrece un nivel de flexibilidad que no se encuentra en un cuadro de diálogo típico, incluida la capacidad de cambiar el tamaño de la ventana y de los paneles de la ventana.

Generación de tablas

Seleccione las variables y medidas de resumen que aparecerán en las tablas de la interfaz Tablas personalizadas.

Analizar > **Tablas** > **Tablas personalizadas**

Lista de variables. Las variables del archivo de datos se muestran en el panel izquierdo del diálogo. Tablas personalizadas distingue entre dos niveles de medición diferentes para las variables y las trata de forma diferente en función del nivel de medición:

Categóricas. Datos con un número limitado de valores o categorías distintas (por ejemplo, sexo o religión). Las variables categóricas pueden ser variables de cadena (alfanuméricas) o variables numéricas que utilizan códigos numéricos para representar a categorías (por ejemplo, 0 = *hombre* y 1 = *mujer*). También se hace referencia a estos datos como datos cualitativos. Las variables categóricas pueden ser **nominales** u **ordinales**.

- *Nominal*. Una variable puede ser tratada como nominal cuando sus valores representan categorías que no obedecen a una clasificación intrínseca. Por ejemplo, el departamento de la compañía en el que trabaja un empleado. Algunos ejemplos de variables nominales son: región, código postal o confesión religiosa.
- *Ordinal*. Una variable puede ser tratada como ordinal cuando sus valores representan categorías con alguna clasificación intrínseca. Por ejemplo, los niveles de satisfacción con un servicio, que abarquen desde muy insatisfecho hasta muy satisfecho. Entre los ejemplos de variables ordinales se incluyen escalas de actitud que representan el grado de satisfacción o confianza y las puntuaciones de evaluación de las preferencias.

Las variables categóricas definen las categorías (filas, columnas y niveles) en la tabla y el estadístico de resumen predeterminado es el recuento (número de casos en cada categoría). Por ejemplo, una tabla predeterminada de una variable de género categórica simplemente mostrará el número de hombres y el de mujeres.

Escalas. Datos medidos en una escala de intervalo o de razón en los que los valores de los datos indican el orden de los valores y la distancia entre ellos. Por ejemplo, un salario de 72.195€ es superior a un salario de 52.398€ y la distancia entre ambos valores es 19.797€. También se hace referencia a estos datos como datos cuantitativos o continuos.

© Copyright IBM Corp. 1989, 2019

Las variables de escala se resumen normalmente dentro de categorías de variables categóricas y el estadístico de resumen predeterminado es la media. Por ejemplo, una tabla predeterminada de ingresos dentro de categorías de género mostrará los ingresos medios de hombres y los ingresos medios de mujeres.

También puede resumir las variables de escala sin utilizar una variable categórica para definir grupos. Es principalmente útil para los resúmenes de **apilado** de las variables de escala múltiple.

Conjuntos de respuestas múltiples

Tablas personalizadas también admite un tipo especial de variable llamada **conjunto de respuestas múltiples**. Los conjuntos de respuestas múltiples no son variables realmente en el sentido normal. No puede verlos en el Editor de datos y otros procedimientos no los reconocen. Los conjuntos de respuestas múltiples utilizan variables para registrar respuestas a preguntas donde el encuestado puede ofrecer más de una respuesta. Los conjuntos de respuestas múltiples se tratan como variables categóricas y la mayoría de las acciones que puede realizar con las variables categóricas, también las puede realizar con los conjuntos de respuestas múltiples.

Un icono junto a cada variable en la lista de variables identifica el tipo de variable.

Categorías. Cuando selecciona una variable categórica en la lista de variables, las categorías definidas por la variable se muestran en el panel Información de variable. Estas categorías también se mostrarán en el panel de lienzo cuando utilice la variable en una tabla. Si la variable no tiene categorías definidas, el panel Información de variable y el panel del lienzo mostrarán dos categorías de marcador: *Categoría 1* y *Categoría 2*.

Las categorías definidas que se muestran en el generador de tablas se basan en **etiquetas de valores**, etiquetas descriptivas asignadas a valores de datos diferentes (por ejemplo, valores numéricos de 0 y 1, con etiquetas de valores *hombre* y *mujer*). Puede definir etiquetas de valor en el panel Información de variable en el Editor de datos.

Panel de lienzo. Genera una tabla arrastrando y soltando las variables en las filas y columnas del panel de lienzo. El panel de lienzo muestra una vista previa de la tabla que se creará. El panel de lienzo no muestra los valores de datos reales en las casillas, pero debe proporcionar una vista bastante precisa del diseño de la tabla final. Para variables categóricas, la tabla real puede contener más categorías que la vista previa si el archivo de datos contiene valores exclusivos para los que no se han definido etiquetas de valores.

Reglas básicas y limitaciones para la generación de una tabla

- Para las variables categóricas, los estadísticos de resumen se basan en la variable más interior de la dimensión de origen de estadísticos.
- La dimensión predeterminada de origen de estadísticos (fila o columna) para las variables categóricas se basa en el orden en el que arrastre y suelte las variables en el panel de lienzo. Por ejemplo, si primero arrastra una variable a la bandeja de filas, la dimensión de la fila es la dimensión predeterminada de origen de estadísticos.
- Las variables de escala se pueden resumir sólo dentro de las categorías de la variable más interior en la dimensión de la fila o la columna. (Puede colocar la variable de escala a cualquier nivel de la tabla, pero se resume al nivel más interior.)
- Las variables de escala no se pueden resumir dentro de otras variables de escala. Puede apilar resúmenes de variables de escala múltiple o resumir variables de escala dentro de categorías de variables categóricas. No puede anidar una variable de escala dentro de otra ni poner una variable de escala en la dimensión de la fila y otra variable de escala en la dimensión de la columna.
- Si cualquier variable del conjunto de datos activo contiene más de 12.000 etiquetas de valores definidos, no puede utilizar el generador de tablas para crear tablas. Si no necesita incluir variables que excedan este límite en sus tablas, puede definir y aplicar conjuntos de variables que excluyan dichas

variables. Si necesita incluir cualquier variable con más de 12.000 etiquetas de valores definidos, no puede utilizar la sintaxis de comandos CTABLES para generar tablas.

Creación de una tabla

1. Seleccione en los menús:

   Analizar > **Tablas** > **Tablas personalizadas**
2. Arrastre y suelte una o más variables en las áreas de fila y/o columna del panel de lienzo.
3. Pulse **Crear** para crear la tabla.

Para suprimir una variable del panel de lienzo

1. Seleccione (pulse) una variable en el panel de lienzo.
2. Pulse con el botón derecho y seleccione **Suprimir variable** en el menú desplegable.

Anidación de variables

La anidación, como tabulación de contingencia, puede mostrar la relación entre dos variables categóricas, excepto una variable anidada dentro de otra en la misma dimensión. Por ejemplo, puede anidar *Género* dentro de la categoría *Edad* en la dimensión de la fila, mostrando el número de hombres y mujeres de cada categoría de edad.

También puede anidar una variable de escala dentro de una variable categórica. Por ejemplo, podría anidar *Ingresos* dentro de *Género*, mostrando medias separadas (o medianas u otras medidas de resumen) de los valores de ingresos para hombres y mujeres.

Para anidar variables

1. Arrastre y suelte una variable categórica en el área de fila y/o columna del panel de lienzo.
2. Arrastre y suelte una variable categórica o de escala encima de una variable de columna o fila categórica.
3. Seleccione **Anidar por encima de todas las variables**, **Anidar a la izquierda** o **Anidar a la derecha** en el menú.

Tabla 1. Variables categóricas anidadas

Variable 1	Variable 2	Estadísticos de resumen
Categoría 1	Categoría 1	12
	Categoría 2	34
	Categoría 3	56
Categoría 2	Categoría 1	12
	Categoría 2	34
	Categoría 3	56

Nota: Las tablas personalizadas no tienen en cuenta el procesamiento del archivo segmentado en capas. Para lograr los mismos resultado que los archivos segmentados en capas, coloque las variables de archivo segmentado en las capas más exteriores de la tabla.

Editar estadísticos

El panel Editar estadísticos le permite:

- Añadir y eliminar estadísticos de resumen de una tabla.

Los estadísticos (y otras opciones) disponibles en el panel Editar estadísticos dependen del nivel de medición de la variable de origen de estadísticos. El origen de los estadísticos (la variable en la que se basan los estadísticos) se determina por:

- **Nivel de medición**. Si una tabla (o una sección de la tabla apilada) contiene una variable de escala, los estadísticos se basan en la variable de escala.
- **Orden de selección de variable**. La dimensión predeterminada de origen de estadísticos (fila o columna) para las variables categóricas se basa en el orden en el que arrastre y suelte las variables en el panel de lienzo. Por ejemplo, si primero arrastra una variable al área de filas, la dimensión de la fila es la dimensión predeterminada de origen de estadísticos.
- **Anidación**. Para las variables categóricas, los estadísticos se basan en la variable más interior de la dimensión de origen de estadísticos.

Estadísticos de resumen para variables categóricas: Los estadísticos básicos disponibles para las variables categóricas son recuentos y porcentajes. También puede especificar estadísticos de resumen personalizados para los totales y subtotales. Estos estadísticos de resumen personalizados incluyen medidas de tendencia central (como media y mediana) y dispersión (como la desviación estándar) que puede ser adecuada para algunas variables categóricas ordinales.

Recuento. Número de casos en cada casilla de la tabla o número de respuestas para conjuntos de respuestas múltiples. Si la ponderación está en vigor, este valor es el recuento ponderado.
- Si la ponderación está en vigor, el valor es el recuento ponderado.
- El recuento ponderado es el mismo para la ponderación de conjunto de datos global (**Datos** > **Ponderar casos...**).

Recuento no ponderado. Número de casos no ponderado en cada casilla de la tabla. Sólo se diferencia del recuento si está activada la ponderación.

Recuento ajustado. El recuento ajustado que se utiliza en los cálculos de ponderación de base efectiva. Si no utiliza una variable de ponderación de base efectiva, el recuento ajustado es el mismo que el recuento.

Porcentajes de fila. Porcentajes dentro de cada fila. Los porcentajes de cada fila de una subtabla (para porcentajes simples) suman 100 %. Los porcentajes de fila son normalmente útiles sólo si tienen una variable de *columna* categórica.

Porcentajes de Columna. Porcentajes dentro de cada columna. Los porcentajes de cada columna de una subtabla (para porcentajes simples) suman 100 %. Los porcentajes de columna son normalmente útiles sólo si tienen una variable de *fila* categórica.

Porcentajes Subtabla. Los porcentajes de cada casilla se basan en la subtabla. Todos los porcentajes de casillas en la subtabla se basan en el mismo número total de casos y suman 100 % dentro de la subtabla. En las tablas anidadas, la variable que precede al nivel de anidación más interior define las subtablas. Por ejemplo, en una tabla de *Estado civil* dentro de *Género* dentro de la categoría *Edad*, *Género* define las subtablas.

Porcentajes de Tabla. Los porcentajes de cada casilla se basan en toda la tabla. Todos los porcentajes de casillas se basan en el mismo número total de casos y suman 100 % (para porcentajes simples) de toda la tabla.

Intervalos de confianza
- Están disponibles los límites de confianza inferior y superior para recuentos, porcentajes, medias, medianas, percentiles y sumas.
- La cadena de texto "&[Nivel de confianza]" en la etiqueta incluye el nivel de confianza en la etiqueta de columna de la tabla.
- El error estándar está disponible para recuentos, porcentajes, medias y sumas.
- Los intervalos de confianza y el error estándar no están disponibles para conjuntos de respuestas múltiples.

Nivel Nivel de confianza para intervalos de confianza, expresado como porcentaje. El valor debe ser mayor que 0 y menor que 100.

Conjuntos de resp. múltiples

Los conjuntos de respuestas múltiples pueden tener los porcentajes basados en casos, respuestas o recuentos. Consulte el tema "Estadísticos de resumen para conjuntos de respuestas múltiples" para obtener más información.

Base porcentual: Los porcentajes se pueden calcular de tres formas diferentes, determinadas por el tratamiento de los valores perdidos en la base computacional:

Porcentaje Simple. Los porcentajes se basan en el número de casos utilizado en la tabla y siempre suman 100 %. Si una categoría se excluye de la tabla, los casos en dicha categoría se excluyen de la base. Los casos con los valores perdidos del sistema siempre se excluyen de la base. Los casos con los valores perdidos del usuario se excluyen si las categorías perdidas del usuario se excluyen de la tabla (el valor predeterminado) o se incluyen si las categorías perdidas del usuario se han incluido en la tabla. Cualquier porcentaje que no contenga *N válido* o *N total* en su nombre es un porcentaje simple.

Porcentaje de N total. Los casos con los valores perdidos del sistema y del usuario se añaden a la base porcentual simple. Los porcentajes pueden sumar menos de 100 %.

Porcentaje de N válido. Los casos con los valores perdidos del usuario se eliminan de la base porcentual simple incluso si se incluyen en la tabla las categorías perdidas del usuario.

Nota: De la base siempre se excluyen los casos de categorías excluidas manualmente que no sean categorías perdidas del usuario.

Estadísticos de resumen para conjuntos de respuestas múltiples: Los siguientes estadísticos de resumen adicionales están disponibles en los conjuntos de respuestas múltiples.

Porcentaje de respuestas de columna/fila/capa. Porcentaje basado en respuestas.

Porcentaje de respuestas de columna/fila/capa (Base: recuento). Las respuestas son el numerador y el recuento total el denominador.

Porcentaje de recuento de columna/fila/capa (Base: respuestas). El recuento es el numerador y el total de respuestas es el denominador.

Porcentaje de respuestas de columna/fila de capa. Porcentajes en las subtablas. Porcentaje basado en respuestas.

Porcentaje de respuestas de columna/fila de capa (Base: recuento). Porcentajes en las subtablas. Las respuestas son el numerador y el recuento total el denominador.

Porcentaje de respuestas de columna/fila de capa (Base: respuestas). Porcentajes en las subtablas. El recuento es el numerador y el total de respuestas es el denominador.

Respuestas. Recuento de respuestas.

Subtable/Table Responses % [Porcentajes de respuestas de subtabla/tabla]. Porcentaje basado en respuestas.

Porcentaje de respuestas de subtabla/tabla (Base: recuento). Las respuestas son el numerador y el recuento total el denominador.

Porcentaje de recuentos de subtabla/tabla (Base: respuestas). El recuento es el numerador y el total de respuestas es el denominador.

Estadísticos de resumen para variables de escala y totales categóricos personalizados: Además de los recuentos y porcentajes disponibles para las variables categóricas, los siguientes estadísticos de resumen están disponibles para las variables de escala y como resúmenes de totales y subtotales personalizados para las variables categóricas. Estos estadísticos de resumen no están disponibles para conjuntos de respuestas múltiples o variables (alfanuméricas) de cadena.

Media. Media aritmética; la suma dividida por el número de casos.

Mediana. Los valores superior e inferior a los que corresponden la mitad de los casos; el percentil 50.

Modo. Valor más frecuente. Si hay un empate, se muestra el valor más pequeño.

Mínimo. El valor más pequeño (inferior).

Máximo. El valor más grande (superior).

Perdidos. Recuento de valores perdidos (tanto del usuario como del sistema).

Percentil. Puede incluir los porcentiles 5, 25, 75, 95 y/o 99.

Rango. Diferencia entre los valores máximo y mínimo.

Desviación estándar. Una medida de dispersión sobre la media. En una distribución normal, el 68% de los casos entran dentro de una desviación estándar de la media y el 95% en dos desviaciones estándar. Por ejemplo, si la edad media es 45, con una desviación estándar del 10,95 % de los casos sería entre 25 y 65 en una distribución normal (la raíz cuadrada de la varianza).

Suma. Suma de valores.

Porcentaje de suma. Porcentajes basados en las sumas. Disponible para filas y columnas (dentro de subtablas), filas y columnas completas (en subtablas), capas, subtablas y tablas completas.

N total. Recuento de valores no perdidos, perdidos del usuario y perdidos del sistema. No incluye casos en categorías excluidas manualmente distintas de las categorías perdidas del usuario.

N total ajustado. El N total ajustado utilizado en cálculos de ponderación de base efectiva. Si no utiliza una variable de ponderación de base efectiva (pestaña Opciones), el N total ajustado es el mismo que el N total.

N válidos. Recuento de valores no perdidos. No incluye casos en categorías excluidas manualmente distintas de las categorías perdidas del usuario.

N válidos ajustados. Los N válidos ajustados utilizados en cálculos de ponderación de base efectiva. Si no utiliza una variable de ponderación de base efectiva (pestaña Opciones), N válidos ajustados son los mismos que los N válidos. Esta estadística no está disponible para conjuntos de respuestas múltiples.

Varianza. Medida de dispersión sobre la media, igual a la suma de las desviaciones al cuadrado de la media divida por el número de casos menos uno. La varianza se mide en unidades que son el cuadrado de las variables (el cuadrado de la desviación estándar).

Intervalos de confianza

- Están disponibles los límites de confianza inferior y superior para recuentos, porcentajes, medias, medianas, percentiles y sumas.

- La cadena de texto "&[Nivel de confianza]" en la etiqueta incluye el nivel de confianza en la etiqueta de columna de la tabla.
- El error estándar está disponible para recuentos, porcentajes, medias y sumas.
- Los intervalos de confianza y el error estándar no están disponibles para conjuntos de respuestas múltiples.

Nivel Nivel de confianza para intervalos de confianza, expresado como porcentaje. El valor debe ser mayor que 0 y menor que 100.

Tablas apiladas

Cada sección de tabla definida por una variable de apilado se trata como una tabla separada y los estadísticos de resumen se calculan según ésta.

Categorías y totales

Con Tablas personalizadas, puede:

- Reordenar categorías.
- Insertar totales.
- Para las variables con etiquetas de valores no definidos, sólo puede clasificar las categorías e insertar totales.

Para acceder a las categorías y a las opciones de totales

1. Arrastre y suelte una variable categórica o un conjunto de respuestas múltiples en el panel de lienzo.
2. Pulse con el botón derecho del ratón en la variable del panel de lienzo y seleccione una de las categorías u opciones de total en el menú emergente.

Para ordenar categorías

1. Pulse con el botón derecho del ratón en la variable del panel de lienzo, seleccione **Ordenar categorías** desde el menú emergente y después seleccione el método de clasificación.
   - Por valor
   - Por etiqueta
   - Por recuento
   - Por valor inferior

Totales

1. Pulse con el botón derecho del ratón en una variable del panel de lienzo, seleccione **Mostrar total** en el menú emergente y después seleccione dónde desea que se muestre el total.
   - Por encima de la categoría
   - Por debajo de la categoría

Si la variable seleccionada se anida dentro de otra variable, los totales se insertarán para cada subtabla.

Tablas personalizadas: Estadísticos de prueba

La característica Estadísticos de prueba proporciona pruebas de significación para tablas personalizadas.

Estas pruebas no están disponibles para las tablas en la cuales las etiquetas de categoría se trasladan fuera de su dimensión de tabla predeterminada o para categorías calculadas.

Pruebas de medias de columna y proporciones de columna

Las pruebas de medias de columna están disponibles para las variables de escala. Las pruebas de proporciones de columna están disponibles para las variables categóricas.

Comparar medias de columna

Pruebas por parejas de la igualdad de las medias de columna. La tabla debe tener una variable categórica en las columnas y una variable de escala como el nivel más interno de las filas. La tabla debe incluir la media como estadístico de resumen.

Para variables categóricas ordinarias, la varianza se puede estimar de todas las categorías o sólo de las categorías que se comparan. Para variables de respuestas múltiples, la varianza para la prueba de media se basa siempre sólo en las categorías que se comparan.

Comparar las proporciones de columna

Pruebas por parejas de la igualdad de las proporciones de columna. La tabla debe tener como mínimo una variable categórica en las columnas y las filas. La tabla debe incluir recuentos o porcentajes de columna.

Nivel de significación

Nivel de significación para pruebas de proporciones de columna y medias de columna.

- El valor debe ser mayor que 0 y menor que 1.
- Su especifica dos niveles de significación, se utilizan letras mayúsculas para identificar los valores de significación menores que o iguales al nivel menor. Se utilizan letras minúsculas para identificar valores de significación menores que o iguales al nivel mayor.
- Si selecciona **Utilizar subíndices de estilo APA**, el segundo valor se ignora.

Ajustar valores p para varias comparaciones

La corrección **Bonferroni** ajusta el índice de errores relacionados con la familia (FWER). El método **Benjamini-Hochberg** es un ajuste de índice de descubrimiento falso (FDR). Este método es menos conservador que la corrección Bonferroni.

Identificar diferencias significativas

Para pruebas de medias de columna y proporciones de columna, puede visualizar resultados significativos en una tabla independiente o en la tabla principal.

En una tabla separada

Los resultados de las pruebas de significación se muestran en una tabla independiente. Si dos valores son significativamente diferentes, la casilla correspondiente al valor mayor muestra una clave que identifica la columna con el valor menor.

Mostrar valores de significación

Los valores de significación se visualizan entre paréntesis después de cada valor de clave en la casilla. Esta opción sólo está disponible cuando se visualizan resultados significativos en una tabla independiente.

En la tabla principal

Los resultados de prueba de significación se visualizan en la tabla principal. Cada categoría de columna de la tabla se identifica con una clave alfabética. Para cada par significativo, la clave de la categoría con la proporción o media de columna menor aparece en la categoría con la proporción o media de columna mayor.

- Cuando se pasa el cursor sobre una clave en la casilla de etiqueta de columna en una tabla dinámica, se resaltan todas las casillas de la tabla con esa clave de significación. Para una tabla con varias variables en la dimensión de columna, sólo se resaltan las casillas de la subtabla.
- Para seleccionar todas las casillas de una tabla (o subtabla) que tienen la misma clave de significación, pulse el botón derecho en la casilla de etiqueta de columna y elija **Seleccionar** > **Seleccionar todas las casillas con esta clave de significación**.

Utilizar subíndices de estilo APA

Identificar diferencias significativas con formato de estilo APA que utiliza letras

de subíndice. Si hay dos valores significativamente diferentes, cuyos valores muestran subíndices de letras diferentes. Estos subíndices no son notas al pie. Si esta opción está activada, el estilo de nota al pie definido en el aspecto de tabla actual se sustituye y las notas al pie se muestran como números de superíndice. Para seleccionar todas las casillas de la misma fila con la misma clave de significación, pulse el botón derecho en una casilla que tenga una clave de significación y elija **Seleccionar casillas con una significación similar**

Pruebas de independencia (Chi-cuadrado)
Prueba de chi-cuadrado de independencia para tablas en las que existe como mínimo una variable de categoría en las filas y columnas.

Utilizar subtotales en lugar de categorías de los subtotales
Cada subtotal sustituye las categorías para la prueba de significación. En caso contrario, sólo los subtotales cuyas categorías con subtotales están ocultas sustituirán sus categorías en la comprobación.

Incluir variables de respuesta múltiple en las pruebas
Las categorías de conjuntos de respuestas múltiples se incluyen en las pruebas de significación. De lo contrario, no se incluyen conjuntos de respuestas múltiples en las pruebas de significación.

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 por 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 han recopilado 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).
- **anorectic.sav.** Mientras trabajaban en una sintomatología estandarizada del comportamiento anoréxico/bulímico, los investigadores [1] realizaron un estudio de 55 adolescentes con trastornos de la

1. Van der Ham, T., J. J. Meulman, D. C. Van Strien, and H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry,* 170, 363-368.

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.

- **anticonvulsants.sav.** Los investigadores médicos pueden utilizar un modelo lineal mixto generalizado para determinar si un nuevo medicamento antiepiléptico puede reducir la tasa de crisis epilépticas de un paciente. Las mediciones repetidas del mismo paciente normalmente se correlacionan de forma positiva, de modo que sería adecuado utilizar un modelo mixto con algunos efectos aleatorios. El campo objetivo, el número de convulsiones, utiliza valores enteros positivos, por lo que podría ser adecuado utilizar un modelo lineal mixto generalizado con una distribución de Poisson y enlace log.
- **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.
- **bankloan_cs.sav.** Este es un archivo de datos hipotéticos que se ocupa de los esfuerzos de un banco por identificar características que son indicativas de personas que probablemente no pagarán los préstamos y, por lo tanto, utilizan estas características para identificar riesgos crediticios buenos y malos.
- **bankloan_cs_noweights.sav.** Este esun archivo de datos hipotéticos que se ocupa de los esfuerzos de un banco por identificar características que son indicativas de personas que probablemente no pagarán los préstamos y, por lo tanto, utilizan estas características para identificar riesgos crediticios buenos y malos. En el archivo no están incluidas las ponderaciones del muestreo.
- **behavior.sav.** En un ejemplo clásico [2], se pidió a 52 estudiantes que valoraran las combinaciones de 15 situaciones y 15 comportamientos en una escala de 10 puntos que oscila 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 mediciones 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 [3], 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.
- **cable_survey.sav.** Los ejecutivos de un proveedor de servicios de televisión por cable, teléfono e Internet desean saber más sobre sus clientes potenciales. Realizan una encuesta a 2.000 personas en sus

2. Price, R. H., and D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, 579-586.

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

áreas de servicio y les preguntan (1) si no tienen el servicio; (2) si están suscritos al servicio con otros proveedores; o (3) si tienen el servicio con la empresa, para cada uno de los tres servicios. La encuesta también recopila alguna información demográfica como, por ejemplo, género, categoría de edad (4 niveles), categoría de formación (3 niveles), categoría de ingresos (3 niveles), categoría de tipo de residencia (4 niveles), categoría de años en la dirección actual (3 niveles), número de personas en el hogar, etc.

- **car_insurance_claims.sav.** Un conjunto de datos presentados y analizados en otro lugar [4] 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 ponderación de escala.
- **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, [5], 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.

---

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

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

- **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 [6]. 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.
- **credit_card.sav.** Un estudio hipotético del uso de la tarjeta de crédito sigue el gasto mensual de cada sujeto en su tarjeta principal durante dos años, con el gasto desglosado por el tipo de transacción (Alimentación, Minorista, Entretenimiento, Viajes y Otros). Cada registro del conjunto de datos corresponde al mes de gasto determinado y el tipo de transacción, así que los datos recopilados para cada sujeto requiere 2 años × 12 meses al año × 5 tipos de transacciones = 120 registros.
- **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.
- **cross_sell.sav.** Una empresa de pedidos por correo tiene un club de libros y un club de CD. Cada mes, ponen a disposición de los miembros del club ofertas especiales. La empresa desea crear un modelo para el total de compras de ofertas especiales del mes basándose en las compras totales de libros, de CD y el tipo de oferta dado a los miembros del club. La regresión de mínimos cuadrados de 2 etapas es un enfoque apropiado para esta situación porque el dinero que se gasta en ofertas especiales es dinero que no se gasta en libros o CD; por lo tanto, existe una retroalimentación entre la respuesta y estos dos predictores.
- **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*.
- **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

6. Kennedy, R., C. Riquier, and B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, 56-70.

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 del muestreo complejo. Cada caso corresponde a una unidad familiar distinta, y se recopila información demográfica y de muestreo diversa.
- **diabetes_costs.sav.** Es un archivo de datos hipotético que contiene información mantenida por una compañía de seguros acerca de los titulares de póliza que tienen diabetes. Cada caso corresponde a un titular de pólica diferente.
- **dietstudy.sav.** Este archivo de datos hipotéticos contiene los resultados de un estudio sobre la "dieta Stillman" [7]. 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.
- **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 recibieron un envío de correos de prueba, y *dmdata3.sav* contiene información sobre el resto de contactos que no recibieron el envío de prueba.
- **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.
- **Employee data.sav.** Se trata de un archivo de datos hipotéticos que contiene información específica de empleado (nivel de formación, categoría de empleo, salario actual, experiencia anterior, etc.).
- **german_credit.sav.** Este archivo de datos se toma del conjunto de datos "German credit" de las Repository of Machine Learning Databases [8] 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 [9] presentó una tabla para ilustrar posibles grupos sociales. Guttman [10] 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).

---

7. Rickman, R., N. Mitchell, J. Dingman, and J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association,* 228:, 54-58.

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

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

10. Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika,* 33, 469-506.

- **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 señalar 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 [11] comenzaron a analizar 15 términos de parentesco (tía, hermano, primo, 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 "orígenes". Cada origen se corresponde con una matriz de proximidades de 15 x 15 cuyas casillas son iguales al número de personas de un origen menos el número de veces que se particionaron los objetos en ese origen.
- **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 1999–2000. 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.

11. Rosenberg, S., and M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research,* 10, 489-502.

- **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[12], [13] han encontrado que no hay linealidad entre estas variables, lo que dificulta los métodos de regresión estándar.
- **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.
- **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

12. Breiman, L., and J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580-598.

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

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 [14].
- **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 [15] 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.
- **stocks.sav** Este archivo de datos hipotéticos contiene precios de acciones y volumen de un año.
- **stroke_clean.sav.** Este archivo de datos hipotéticos contiene el estado de una base de datos médica después de haberla limpiado mediante los procedimientos en Statistics Base Edition.
- **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.

---

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

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

- **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.
- **tcm_kpi.sav.** Es un archivo de datos hipotético que contiene valores de indicadores clave de rendimiento semanales para una empresa. También contiene datos semanales para diversas métricas controlables durante el mismo periodo de tiempo.
- **tcm_kpi_upd.sav.** Este archivo de datos es idéntico a *tcm_kpi.sav* pero contiene datos para cuatro semanas adicionales.
- **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 predeterminado de las variables antes de realizar la asignación de nivel de medición 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 [16].
- **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 [17].

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

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

- **verd1985.sav.** Archivo de datos sobre una encuesta [18]. 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 [19]. 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.
- **worldsales.sav** Este archivo de datos hipotéticos contiene ingresos por ventas por continente y producto.

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

19. Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, and B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, 366-374.

Avisos

Esta información se ha desarrollado para productos y servicios ofrecidos en los EE.UU. Este material puede estar disponible en IBM en otros idiomas. Sin embargo, es posible que deba ser propietario de una copia del producto o de la versión del producto en dicho idioma para acceder a él.

Es posible que IBM no ofrezca los productos, servicios o características que se tratan en este documento en otros países. El representante local de IBM le puede informar sobre los productos y servicios que están actualmente disponibles en su localidad. Cualquier referencia a un producto, programa o servicio de IBM no pretende afirmar ni implicar que solamente se pueda utilizar ese producto, programa o servicio de IBM. En su lugar, se puede utilizar cualquier producto, programa o servicio funcionalmente equivalente que no infrinja los derechos de propiedad intelectual de IBM. Sin embargo, es responsabilidad del usuario evaluar y comprobar el funcionamiento de todo producto, programa o servicio que no sea de IBM.

IBM puede tener patentes o solicitudes de patente en tramitación que cubran la materia descrita en este documento. Este documento no le otorga ninguna licencia para estas patentes. Puede enviar preguntas acerca de las licencias, por escrito, a:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
EE.UU.

Para consultas sobre licencias relacionadas con información de doble byte (DBCS), póngase en contacto con el departamento de propiedad intelectual de IBM de su país o envíe sus consultas, por escrito, a:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokio 103-8510, Japón

INTERNATIONAL BUSINESS MACHINES CORPORATION PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL", SIN GARANTÍAS DE NINGUNA CLASE, NI EXPLÍCITAS NI IMPLÍCITAS, INCLUYENDO, PERO SIN LIMITARSE A, LAS GARANTÍAS IMPLÍCITAS DE NO VULNERACIÓN, COMERCIALIZACIÓN O ADECUACIÓN A UN PROPÓSITO DETERMINADO. Algunas jurisdicciones no permiten la renuncia a las garantías explícitas o implícitas en determinadas transacciones; por lo tanto, es posible que esta declaración no sea aplicable a su caso.

Esta información puede incluir imprecisiones técnicas o errores tipográficos. Periódicamente, se efectúan cambios en la información aquí y estos cambios se incorporarán en nuevas ediciones de la publicación. IBM puede realizar en cualquier momento mejoras o cambios en los productos o programas descritos en esta publicación sin previo aviso.

Las referencias hechas en esta publicación a sitios web que no son de IBM se proporcionan sólo para la comodidad del usuario y no constituyen de modo alguno un aval de esos sitios web. La información de esos sitios web no forma parte de la información de este producto de IBM y la utilización de esos sitios web se realiza bajo la responsabilidad del usuario.

IBM puede utilizar o distribuir la información que se le proporcione del modo que considere adecuado sin incurrir por ello en ninguna obligación con el remitente.

Los titulares de licencias de este programa que deseen tener información sobre el mismo con el fin de permitir: (i) el intercambio de información entre programas creados independientemente y otros programas (incluido este) y (ii) el uso mutuo de la información que se ha intercambiado, deberán ponerse en contacto con:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
EE.UU.

Esta información estará disponible, bajo las condiciones adecuadas, incluyendo en algunos casos el pago de una cuota.

El programa bajo licencia que se describe en este documento y todo el material bajo licencia disponible los proporciona IBM bajo los términos de las Condiciones Generales de IBM, Acuerdo Internacional de Programas Bajo Licencia de IBM o cualquier acuerdo equivalente entre las partes.

Los ejemplos de datos de rendimiento y de clientes citados se presentan solamente a efectos ilustrativos. Los resultados reales de rendimiento pueden variar en función de las configuraciones específicas y condiciones de operación.

La información relacionada con productos no IBM se ha obtenido de los proveedores de esos productos, de sus anuncios publicados o de otras fuentes disponibles públicamente. IBM no ha probado esos productos y no puede confirmar la exactitud del rendimiento, la compatibilidad ni ninguna otra afirmación relacionada con productos no IBM. Las preguntas sobre las posibilidades de productos que no son de IBM deben dirigirse a los proveedores de esos productos.

Las declaraciones sobre el futuro rumbo o intención de IBM están sujetas a cambio o retirada sin previo aviso y representan únicamente metas y objetivos.

Esta información contiene ejemplos de datos e informes utilizados en operaciones comerciales diarias. Para ilustrarlos lo máximo posible, los ejemplos incluyen los nombres de las personas, empresas, marcas y productos. Todos estos nombres son ficticios y cualquier parecido con personas o empresas comerciales reales es pura coincidencia.

LICENCIA DE DERECHOS DE AUTOR:

Esta información contiene programas de aplicación de muestra escritos en lenguaje fuente, los cuales muestran técnicas de programación en diversas plataformas operativas. Puede copiar, modificar y distribuir estos programas de muestra de cualquier modo sin realizar ningún pago a IBM, con el fin de desarrollar, utilizar, comercializar o distribuir programas de aplicación que se ajusten a la interfaz de programación de aplicaciones para la plataforma operativa para la que se han escrito los programas de muestra. Estos ejemplos no se han probado exhaustivamente en todas las condiciones. Por lo tanto, IBM no puede garantizar ni dar por supuesta la fiabilidad, la capacidad de servicio ni la funcionalidad de estos programas. Los programas de muestra se proporcionan "TAL CUAL" sin garantía de ningún tipo. IBM no será responsable de ningún daño derivado del uso de los programas de muestra.

Cada copia o cada parte de los programas de ejemplo o de los trabajos que se deriven de ellos debe incluir un aviso de copyright como se indica a continuación:

© IBM 2019. Algunas partes de este código procede de los programas de ejemplo de IBM Corp.

© Copyright IBM Corp. 1989 - 20019. Reservados todos los derechos.

Marcas comerciales

IBM, el logotipo de IBM e ibm.com son marcas registradas o marcas comerciales de International Business Machines Corp., registradas en muchas jurisdicciones en todo el mundo. Otros nombres de productos y servicios podrían ser marcas registradas de IBM u otras compañías. En Internet hay disponible una lista actualizada de las marcas registradas de IBM, en "Copyright and trademark information", en www.ibm.com/legal/copytrade.shtml.

Adobe, el logotipo Adobe, PostScript y el logotipo PostScript son marcas registradas o marcas comerciales de Adobe Systems Incorporated en Estados Unidos y/o otros países.

Intel, el logotipo de Intel, Intel Inside, el logotipo de Intel Inside, Intel Centrino, el logotipo de Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium y Pentium son marcas comerciales o marcas registradas de Intel Corporation o sus filiales en Estados Unidos y otros países.

Linux es una marca registrada de Linus Torvalds en Estados Unidos, otros países o ambos.

Microsoft, Windows, Windows NT, y el logotipo de Windows son marcas comerciales de Microsoft Corporation en Estados Unidos, otros países o ambos.

UNIX es una marca registrada de The Open Group en Estados Unidos y otros países.

Java y todas las marcas comerciales y los logotipos basados en Java son marcas comerciales o registradas de Oracle y/o sus afiliados.

Índice

Impreso en España