IBM SPSS Missing Values 19

Note: Before using this information and the product it supports, read the general information under Notices el p. 94.

Prefacio

IBM® SPSS® Statistics es un sistema global para el análisis de datos. El módulo adicional opcional Valores perdidos proporciona las técnicas de análisis adicionales que se describen en este manual. El módulo adicional Valores perdidos 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

Parte I:
Manual del usuario

Capítulo
1

Introducción a valores perdidos

Los casos con valores perdidos representan un reto importante, ya que los procedimientos de modelado tradicionales simplemente descartan estos casos para el análisis. Cuando hay pocos valores perdidos (aproximadamente, menos del 5 % del número total de casos) y dichos valores pueden considerarse perdidos de forma aleatoria (es decir, que la pérdida de un valor no depende de otros valores), entonces el método tradicional de eliminación según la lista es relativamente "seguro". La opción Valores perdidos puede ayudarle a determinar si la eliminación según la lista es suficiente; asimismo, proporciona métodos para gestionar los valores perdidos cuando no lo sea.

Análisis de valores perdidos frente a procedimientos de imputación múltiple

La opción Valores perdidos proporciona dos conjuntos de procedimientos para gestionar los valores perdidos:

- Los procedimientos de Imputación múltiple proporcionan un análisis de los patrones de datos perdidos, dirigidos a una imputación múltiple de valores perdidos. Esto es, se producen versiones múltiples del conjunto de datos, cada una con su propio conjunto de valores imputados. Cuando se realizan análisis estadísticos, se combinan las estimaciones de los parámetros de todos los conjuntos de datos imputados, con lo que se ofrecen estimaciones generalmente más precisas de lo que serían con sólo una imputación.
- Análisis de valores perdidos proporciona un conjunto ligeramente diferente de herramientas descriptivas para analizar los datos perdidos (en especial, la prueba MCAR de Little) e incluye una variedad de métodos de imputación individual. Tenga en cuenta que por lo general la imputación múltiple suele considerarse superior a la imputación individual.

Tareas de valores perdidos

Puede empezar con el análisis de valores perdidos siguiendo estos pasos básicos:

- **Examinar la ausencia.** Utilice Análisis de valores perdidos y Analizar patrones para explorar patrones de valores perdidos en sus datos y determinar si es necesario recurrir a la imputación múltiple.
- **Imputar valores perdidos.** Utilice Imputar valores perdidos para imputar de forma múltiple los valores perdidos.
- **Analizar datos "completos".** Utilice cualquier procedimiento que admita datos de imputación múltiple. Consulte Análisis de datos de imputación múltiple el p. 28 para obtener información sobre el análisis de conjuntos de datos de imputación múltiple y una lista de procedimientos que admiten estos datos.

Capítulo
2

Análisis de valores perdidos

El procedimiento Análisis de valores perdidos realiza tres funciones principales:

- Describe el patrón de los datos perdidos. ¿Dónde se encuentran los valores perdidos? ¿Con qué frecuencia aparecen? ¿Hay pares de variables que tienden a tener valores perdidos en varios casos? ¿Son los valores de los datos extremos? ¿Están los valores perdidos de forma aleatoria?
- Estimar las medias, desviaciones típicas, covarianzas y correlaciones para los diferentes métodos de valores perdidos: por lista, por parejas, regresión o EM (maximización esperada). El método por parejas muestra, además, recuentos de los casos completos por parejas.
- Rellena (imputa) los valores perdidos con valores estimados utilizando el método EM o el de regresión; sin embargo, por lo general se considera que la imputación múltiple proporciona resultados más precisos.

El análisis de valores perdidos ayuda a resolver varios problemas ocasionados por los datos incompletos. Si los casos con valores perdidos son sistemáticamente diferentes de los casos sin valores perdidos, los resultados pueden ser equívocos. Además, los datos perdidos pueden reducir la precisión de los estadísticos calculados, porque no se dispone de tanta información como originalmente se pensaba. Otro problema radica en que los supuestos subyacentes a muchos procedimientos estadísticos se basan en casos completos y los valores perdidos pueden complicar la teoría exigida.

Ejemplo. En la evaluación de un tratamiento contra la leucemia se miden diversas variables. Sin embargo, no todas las medidas se encuentran disponibles para todos los pacientes. Los patrones de los datos perdidos se inspeccionan, se tabulan y se consideran aleatorios. Se utiliza un análisis EM para estimar las medias, las correlaciones y las covarianzas. También se utiliza para determinar que los datos están perdidos completamente al azar. A continuación, los valores perdidos se reemplazan por los valores imputados y se guardan en un nuevo archivo de datos para análisis posteriores.

Estadísticos. Estadísticos univariados, incluido el número de valores no perdidos, media, desviación típica, número de valores perdidos y número de valores extremos. Medias estimadas, matriz de covarianza y matriz de correlaciones, utilizando los métodos de regresión, EM, por lista o por parejas. Prueba MCAR de Little con resultados EM. Resumen de medias a través de varios métodos. Para los grupos definidos por valores perdidos frente a valores no perdidos: pruebas *t*. Para todas las variables: los patrones de valores perdidos representados como casos respecto a variables.

Consideraciones de los datos

Datos. Los datos pueden ser categóricos o cuantitativos (de escala o continuos). Sin embargo, puede estimar los estadísticos e imputar los datos perdidos únicamente en el caso de variables cuantitativas. Para cada variable, los valores perdidos que no están codificados como valores perdidos del sistema deben definirse como valores definidos como perdidos por el usuario. Por ejemplo, si un elemento del cuestionario tiene la respuesta *No sabe* codificada como 5 y desea tratarlo como valor perdido, el elemento debería tener el 5 codificado como valor definido como perdido por el usuario.

Ponderaciones de frecuencia. Este procedimiento respeta las ponderaciones de frecuencia (replicación). Los casos de ponderaciones con valor negativo o cero de replicación se ignoran. Las ponderaciones no enteras se truncan.

Supuestos. La estimación por lista, por parejas y mediante regresión depende del supuesto de que el patrón de valores perdidos no depende de los valores de los datos (esta condición se conoce como perdidos completamente al azar o MCAR). (Esta condición se conoce como **perdida completamente al azar** o MCAR). Por tanto, todos los métodos (incluido el método EM) de estimación ofrecen estimaciones coherentes y no sesgadas de las correlaciones y las covarianzas cuando los datos son MCAR. El incumplimiento del supuesto MCAR puede dar lugar a estimaciones sesgadas producidas por los métodos de regresión, por lista o por parejas. Si los datos no son MCAR, es necesario utilizar la estimación EM.

La estimación EM depende del supuesto de que el patrón de los datos perdidos está relacionado únicamente con los datos observados. (Esta condición se denomina **perdidos al azar** o MAR.) Este supuesto permite ajustar las estimaciones utilizando la información disponible. Por ejemplo, en un estudio sobre la educación y los ingresos, los sujetos con un menor nivel educativo pueden tener más valores perdidos de ingresos. En este caso, los datos son MAR, no MCAR. Es decir, para MAR, la probabilidad de que se registren los ingresos depende del nivel educativo del sujeto. La probabilidad puede variar según el nivel educativo pero no según los ingresos *dentro de ese nivel educativo.* Si la probabilidad de que se registre el ingreso también depende del valor de los ingresos dentro de cada nivel educativo (por ejemplo, las personas con ingresos elevados no los declara), los datos no serán ni MCAR ni MAR. Se trata de una situación poco habitual y, si se produce, no hay ningún método adecuado.

Procedimientos relacionados. Muchos procedimientos permiten utilizar la estimación por lista o por parejas. Regresión lineal y Análisis factorial permiten reemplazar los valores perdidos por los valores de las medias. El módulo adicional Predicciones ofrece varios métodos para reemplazar los valores perdidos en las series temporales.

Para obtener un análisis de valores perdidos

- Seleccione en los menús:
  Analizar > Análisis de valores perdidos...

Figura 2-1
Cuadro de diálogo Análisis de valores perdidos

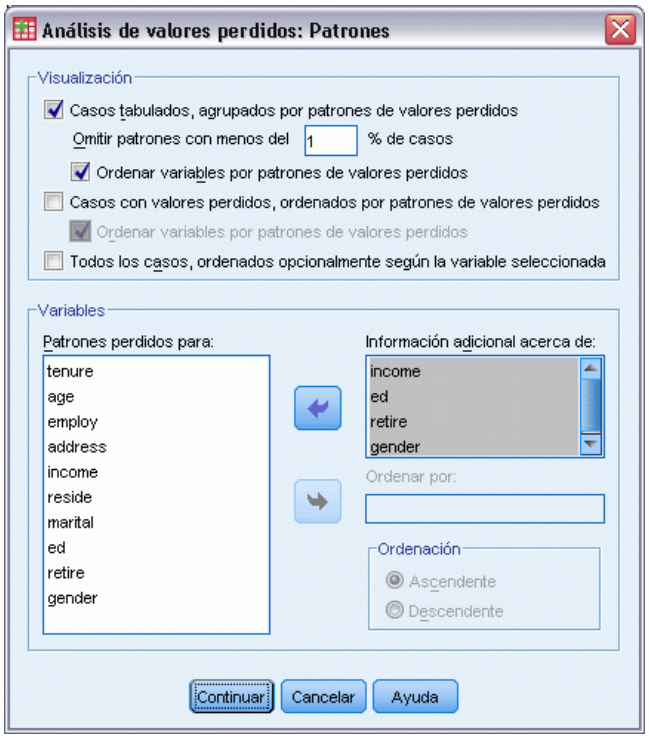
▶ Seleccione al menos una variable cuantitativa (de escala) para estimar los estadísticos y, si lo desea, imputar los valores perdidos.

Si lo desea, puede:

- Seleccionar variables categóricas (numéricas o de cadena) y establecer un límite para el número de categorías (N.º máximo de categorías).
- Pulse en Patrones para tabular los patrones de los datos perdidos. Si desea obtener más información, consulte el tema Visualización de los patrones de los valores perdidos el p. 5.
- Pulse en Descriptivos para mostrar los estadísticos descriptivos de los valores perdidos. Si desea obtener más información, consulte el tema Visualización de los estadísticos descriptivos de los valores perdidos el p. 6.
- Seleccione un método para estimar los estadísticos (medias, covarianzas y correlaciones) y posiblemente imputar los valores perdidos. Si desea obtener más información, consulte el tema Estimación de los estadísticos e imputación de los valores perdidos el p. 8.
- Si selecciona EM o Regresión, pulsar en Variables para especificar el subconjunto que se va a utilizar para la estimación. Si desea obtener más información, consulte el tema Variables pronosticadas y predictoras el p. 11.
- Seleccione una variable de etiqueta de caso. Esta variable se utiliza para etiquetar los casos en las tablas de patrones que muestran los casos individuales.

Visualización de los patrones de los valores perdidos

Figura 2-2
Cuadro de diálogo Análisis de valores perdidos: Patrones



Si lo desea, puede consultar varias tablas que muestran los patrones y el impacto de los datos perdidos. Estas tablas pueden ayudarle a identificar:

- Dónde se encuentran los valores perdidos
- Si hay pares de variables que tienden a tener valores perdidos en casos individuales
- Si los valores de los datos son extremos

Representación

Hay tres tipos de tablas disponibles para ver los patrones de los datos perdidos.

Casos tabulados. Se tabulan los patrones de los valores perdidos en las variables de análisis y se muestran las frecuencias de cada patrón. Utilice Ordenar variables según patrón de valores perdidos para especificar si los recuentos y las variables se ordenan según la similaridad de los patrones. Utilice Omitir patrones con menos del n % de los casos para eliminar los patrones que aparecen con poca frecuencia.

Casos con valores perdidos. Cada caso con un valor perdido o extremo se tabula para cada variable de análisis. Utilice Ordenar variables según patrón de valores perdidos para especificar si los recuentos y las variables se ordenan según la similaridad de los patrones.

Todos los casos. Se tabula cada caso y se indican los valores perdidos y extremos para cada variable. Los casos se enumeran en el orden en que aparecen en el archivo de datos, a menos que se especifique una variable en Ordenar por.

En las tablas que muestran los casos individuales, se utilizan los siguientes símbolos:

+	Valor extremadamente alto
-	Valor extremadamente bajo
S	Valor perdido del sistema
A	Primer tipo de valor definido como perdido por el usuario
B	Segundo tipo de valor definido como perdido por el usuario
C	Tercer tipo de valor definido como perdido por el usuario

Variables

Puede mostrar información adicional acerca de las variables que se incluyen en el análisis. Las variables que se añadan a Información adicional acerca de aparecerán individualmente en la tabla de patrones perdidos. Para las variables cuantitativas (de escala), se muestra la media; para las variables categóricas, se muestra el número de casos que presentan el patrón en cada categoría.

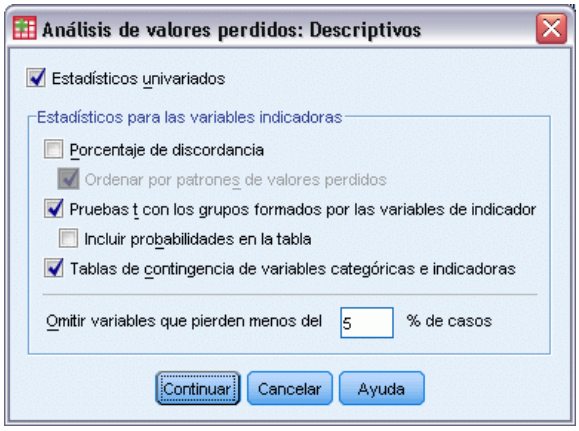
- **Ordenar por.** Los casos se listan según el orden ascendente o descendente de los valores de la variable especificada. Esta opción está disponible sólo si se selecciona Todos los casos.

Para mostrar los patrones de los valores perdidos

- En el cuadro de diálogo principal Análisis de valores perdidos, seleccione las variables cuyos patrones de valores perdidos desea ver.
- Pulse en Patrones.
- Seleccione las tablas de patrones que desea ver.

Visualización de los estadísticos descriptivos de los valores perdidos

Figura 2-3
Cuadro de diálogo Análisis de valores perdidos: Descriptivos



Estadísticos univariantes

Los estadísticos univariados pueden ayudarle a identificar el impacto general de los datos perdidos. Para cada variable, se muestran los siguientes datos:

- Número de valores no perdidos
- Número y porcentaje de valores perdidos

Para las variables cuantitativas (de escala), también se muestran los siguientes datos:

- Media
- Desviación típica
- Número de valores extremadamente altos o bajos

Estadísticos para las variables indicadoras

Para cada variable, se crea una variable de indicador. Esta variable categórica indica si la variable está presente o perdida en un determinado caso. Las variables de indicador se utilizan para crear la discordancia, la prueba *t* y las tablas de frecuencia.

Porcentaje de discordancia. Para cada par de variables muestra el porcentaje de casos en los que una variable tiene un valor perdido y la otra variable tiene un valor no perdido. Cada elemento diagonal de la tabla contiene el porcentaje de valores perdidos para una sola variable.

Pruebas t con los grupos formados por las variables de indicador. Se comparan las medias de los dos grupos para cada variable cuantitativa, utilizando el estadístico *t* de Student. Los grupos especifican si una variable está presente o perdida. Se muestra el estadístico *t*, los grados de libertad, los recuentos de valores perdidos y no perdidos y las medias de los dos grupos. También se pueden mostrar todas las probabilidades bilaterales asociadas con el estadístico *t*. Si el análisis genera más de una prueba, no utilice estas probabilidades para contrastar la significación. Estas probabilidades sólo son adecuadas cuando se calcula una única prueba.

Tablas de contingencia de variables categóricas y de indicador. Para cada variable categórica se muestra una tabla. Para cada categoría, la tabla muestra la frecuencia y el porcentaje de los valores no perdidos para las demás variables. También se muestran los porcentajes de cada tipo de valor perdido.

Omitir variables con menos valores perdidos que el n % de los casos. Para reducir el tamaño de la tabla puede omitir los estadísticos que se calculen sólo para un pequeño número de casos.

Para mostrar los estadísticos descriptivos

▶ En el cuadro de diálogo principal Análisis de valores perdidos, seleccione las variables cuyos estadísticos descriptivos de los valores perdidos desea ver.

▶ Pulse en Descriptivos.

▶ Elija los estadísticos descriptivos que desea que aparezcan.

Estimación de los estadísticos e imputación de los valores perdidos

Puede elegir que se estimen las medias, desviaciones típicas, covarianzas y correlaciones utilizando un método por lista (sólo casos completos), por parejas, EM (maximización esperada) y/o de regresión. También puede elegir imputar los valores perdidos (estimar los valores de sustitución). Tenga en cuenta que por lo general la Imputación múltiple suele considerarse superior a la imputación individual para solucionar el problema de los valores perdidos. La prueba MCAR de Little sigue siendo útil para determinar si la imputación es necesaria.

Método por lista

Este método únicamente utiliza los casos completos. Si alguna de las variables de análisis tiene valores perdidos, se omite dicho caso de los cálculos.

Método por parejas

Este método examina las parejas de variables del análisis y utiliza un caso únicamente si tiene valores no perdidos para ambas variables. Las frecuencias, medias y desviaciones típicas se calculan por separado para cada pareja. Como se ignoran los demás valores perdidos del caso, las correlaciones y las covarianzas de las dos variables no dependen de los valores perdidos de ninguna otra variable.

Método EM

Este método supone que los datos parcialmente perdidos siguen una distribución determinada y basa las inferencias en la probabilidad según dicha distribución. Cada iteración se compone de un paso E y un paso M. El paso E determina la esperanza condicional de los datos “perdidos”, teniendo en cuenta los valores observados y las estimaciones actuales de los parámetros. A continuación, se sustituyen estas esperanzas por los datos “perdidos”. En el paso M, se calculan las estimaciones de máxima verosimilitud de los parámetros como si se hubieran rellenado los datos perdidos. Se especifica “perdidos” entre comillas ya que los valores perdidos no se rellenan directamente, sino que en su lugar se utilizan funciones de ellos en el log verosimilitud.

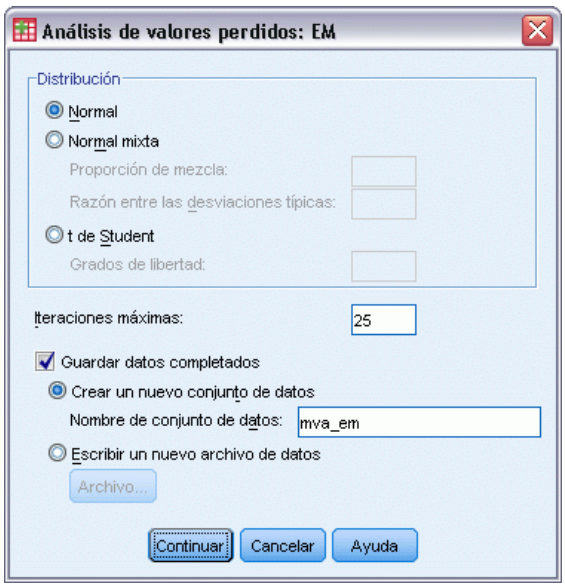
El estadístico de chi cuadrado de Roderick J. A. Little para contrastar si los valores están perdidos completamente al azar (MCAR) se imprime como nota al pie de las matrices de EM. Para este contraste, la hipótesis nula es que los datos están perdidos completamente al azar y el valor *p* es significativo al nivel 0,05. Si el valor es inferior a 0,05, los datos no están perdidos completamente al azar. Los datos pueden estar perdidos al azar (MAR) o no perdidos al azar (NMAR). No se puede suponer la situación en la que se encuentran los datos perdidos, por lo que es necesario analizar los datos para determinar de qué manera están perdidos.

Método de regresión

Este método calcula las estimaciones de regresión lineal múltiple y ofrece opciones que permiten incrementar las estimaciones con componentes aleatorios. Para cada valor pronosticado, el procedimiento puede añadir un residuo de un caso completo seleccionado de manera aleatoria, una desviación normal aleatoria o una desviación aleatoria (escalada por la raíz cuadrada del residuo cuadrático promedio) de la distribución *t*.

Opciones de estimación EM

Figura 2-4
Cuadro de diálogo Análisis de valores perdidos: EM



Utilizando un proceso iterativo, el método EM estima las medias, la matriz de covarianzas y la correlación de las variables cuantitativas (de escala) con los valores perdidos.

Distribución. EM realiza las inferencias basándose en la verosimilitud según la distribución especificada. Por defecto, se supone una distribución normal. Si sabe que las colas de la distribución son más largas que las de una distribución normal, puede solicitar que el procedimiento construya la función de verosimilitud a partir de una distribución *t* de Student con *n* grados de libertad. La distribución normal mixta también proporciona una distribución con colas más largas. Especifique la razón de las desviaciones típicas de la distribución normal mixta y la proporción de mezcla de las dos distribuciones. La distribución normal mixta supone que únicamente difieren las desviaciones típicas de las distribuciones. Las medias deben ser iguales.

Número máximo de iteraciones. Establece el número máximo de iteraciones para estimar la covarianza auténtica. El procedimiento se detiene cuando se alcanza el número de iteraciones, incluso si no han convergido las estimaciones.

Guardar datos completados. Puede guardar un conjunto de datos con los valores imputados en el lugar de los valores perdidos. No obstante, tenga en cuenta que los estadísticos basados en la covarianza que utilicen los valores imputados estimarán valores de los parámetros menores que los reales. El grado en que esta estimación es inferior a la real es proporcional al número de casos que no se observaron conjuntamente.

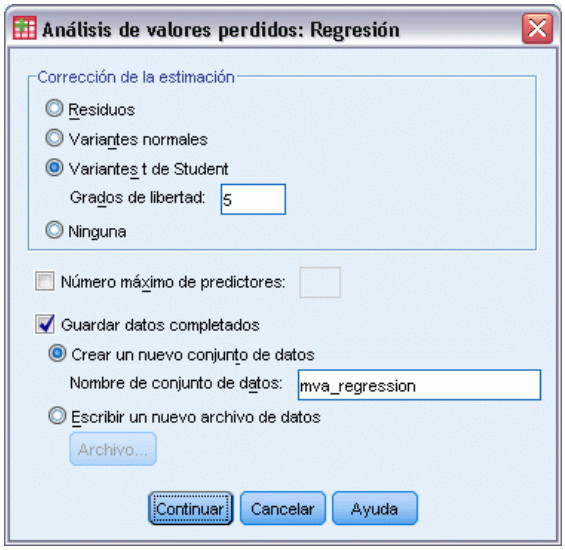
Para especificar las opciones de EM

- En el cuadro de diálogo principal Análisis de valores perdidos, seleccione las variables cuyos valores perdidos desea estimar utilizando el método EM.
- Seleccione EM en el grupo Estimación.

- Para especificar las variables predictoras y pronosticadas, pulse en Variables. Si desea obtener más información, consulte el tema Variables pronosticadas y predictoras el p. 11.
- Pulse en EM.
- Seleccione las opciones de EM que desee.

Opciones de estimación de regresión

Figura 2-5
Cuadro de diálogo Análisis de valores perdidos: Regresión



El método de regresión estima los valores perdidos utilizando la regresión lineal múltiple. Se muestran las medias, la matriz de covarianza y la matriz de correlaciones de las variables pronosticadas.

Corrección de la estimación. El método de regresión puede añadir un componente aleatorio a las estimaciones de regresión. Puede seleccionar residuos, variantes normales, variantes *t* de Student o sin corrección.

- **Residuo.** Los términos de error se eligen al azar de entre los residuos observados en los casos completos, para añadirlos a las estimaciones de regresión.
- **Variantes normales.** Los términos de error se escogen al azar de una distribución con valor esperado 0 y desviación típica igual a la raíz cuadrada del termino error cuadrático medio de la regresión.
- **Variantes de Student.** Los términos de error se escogen al azar de una distribución t con los grados de libertad especificados y se escalan según la raíz del error cuadrático medio (RMSE).

Número máximo de predictores. Establece un límite máximo para el número de variables predictoras (independientes) utilizadas en el proceso de estimación.

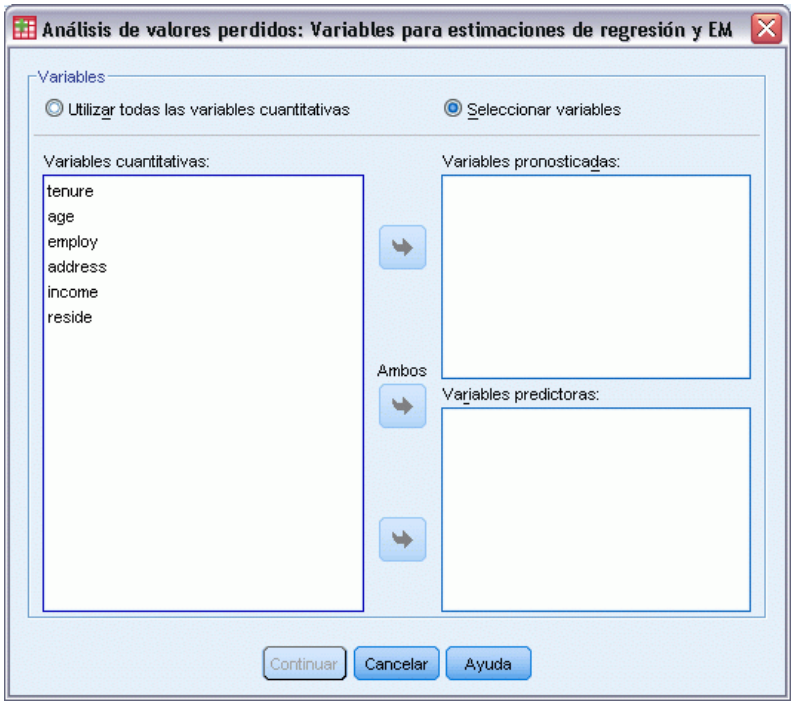
Guardar datos completados. Escribe un conjunto de datos en la sesión actual o en un archivo de datos externo con formato IBM® SPSS® Statistics, reemplazando los valores perdidos por los valores estimados mediante el método de regresión.

Para especificar las opciones de regresión

- En el cuadro de diálogo principal Análisis de valores perdidos, seleccione las variables cuyos valores perdidos desea estimar utilizando el método de regresión.
- Seleccione Regresión en el grupo Estimación.
- Para especificar las variables predictoras y pronosticadas, pulse en Variables. Si desea obtener más información, consulte el tema Variables pronosticadas y predictoras el p. 11.
- Pulse en Regresión.
- Seleccione las opciones de regresión deseadas.

Variables pronosticadas y predictoras

Figura 2-6
Cuadro de diálogo Análisis de valores perdidos: Variables para estimaciones de regresión y EM



Por defecto, se utilizan todas las variables cuantitativas para la estimación de regresión y EM. Si es necesario, puede especificar que determinadas variables se utilicen como variables pronosticadas o variables predictoras en las estimaciones. Una determinada variable puede aparecer en ambas listas, pero hay situaciones en las que quizá quiera restringir el uso de una variable. Por ejemplo, a algunos analistas no les resulta cómodo estimar los valores de las variables de resultados. También es posible que quiera utilizar variables diferentes en estimaciones distintas y ejecutar el procedimiento varias veces. Por ejemplo, si tiene un conjunto de elementos que son valoraciones de enfermeras y otro conjunto que son valoraciones de médicos, tal vez quiera ejecutar el procedimiento una vez utilizando el elemento de las enfermeras para estimar los elementos de las enfermeras y otra vez para estimar los elementos de los médicos.

También hay que hacer otra consideración al utilizar el método de regresión. En la regresión múltiple, el uso de un subconjunto grande de variables independientes puede generar valores pronosticados de peor calidad que los que generaría un subconjunto más pequeño. Por tanto, para que se utilice una variable, debe alcanzar un límite de *F* para entrar de 4,0. Este límite se puede cambiar utilizando la sintaxis.

Para especificar las variables pronosticadas y predictoras

- En el cuadro de diálogo principal Análisis de valores perdidos, seleccione las variables cuyos valores perdidos desea estimar utilizando el método de regresión.
- Seleccione EM o Regresión en el grupo Estimación.
- Pulse en Variables.
- Si desea utilizar determinadas variables, en vez de todas, como variables pronosticadas y variables predictoras, elija Seleccionar variables y mueva las variables a las listas adecuadas.

Funciones adicionales del comando MVA

La sintaxis de comandos también le permite:

- Especificar distintas variables descriptivas para los patrones de valores perdidos, los patrones de los datos y los patrones tabulados, mediante la palabra clave `DESCRIBE` en los subcomandos `MPATTERN`, `DPATTERN` o `TPATTERN`.
- Especificar más de una variable de ordenación para la tabla de patrones de los datos, utilizando el subcomando `DPATTERN`.
- Especificar más de una variable de ordenación para los patrones de los datos, utilizando el subcomando `DPATTERN`.
- Especificar la tolerancia y la convergencia mediante el subcomando `EM`.
- Especifique la tolerancia y la *F* para entrar mediante el subcomando `REGRESSION`.
- Especificar diferentes listas de variables para EM y para Regresión, con los subcomandos `EM` y `REGRESSION`.
- Especificar diferentes porcentajes para suprimir los casos mostrados para `TTESTS`, `TABULATE` y `MISMATCH`.

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

Capítulo
3

Imputación múltiple

El objetivo de la imputación múltiple es generar valores posibles para los valores perdidos, creando así varios conjuntos de datos "completos". Los procedimientos analíticos que trabajan con conjuntos de datos de imputación múltiple producen resultados para cada conjunto de datos "completo", además de resultados combinados que estiman cuáles habrían sido los resultados si el conjunto de datos original no tuviera valores perdidos. Estos resultados combinados suelen ser más precisos que los proporcionados por métodos de imputación individual.

Variables de análisis. Las variables de análisis pueden ser:

- **Nominal.** Una variable se puede tratar como nominal si sus valores representan categorías que no obedecen a una ordenación intrínseca (por ejemplo, el departamento de la empresa en el que trabaja un empleado). Algunos ejemplos de variables nominales son: región, código postal o confesión religiosa.
- **Ordinal.** Una variable puede tratarse como ordinal cuando sus valores representan categorías con alguna ordenación intrínseca (por ejemplo, los niveles de satisfacción con un servicio, que vayan 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.
- **Escala.** Una variable puede tratarse como escala (continua) cuando sus valores representan categorías ordenadas con una métrica con significado, por lo que son adecuadas las comparaciones de distancia entre valores. Son ejemplos de variables de escala: la edad en años y los ingresos en dólares.

  El procedimiento supone que se ha asignado el nivel de medida adecuado a todas las variables. No obstante, puede cambiar temporalmente el nivel de medida para una variable pulsando con el botón derecho en la variable en la lista de variables de origen y seleccionar un nivel de medida en el menú contextual.

Un icono situado junto a cada variable de la lista de variables identifica el nivel de medida y el tipo de datos.

Nivel de medida	**Tipo de datos**			
	Numérico	Cadena	Fecha	Hora
Escala (Continuo)		n/a		

Ordinal				
Nominal				

Ponderaciones de frecuencia. Este procedimiento respeta las ponderaciones de frecuencia (replicación). Los casos de ponderaciones con valor negativo o cero de replicación se ignoran. Las ponderaciones no enteras se redondean al número entero más cercano.

Ponderación de análisis. Las ponderaciones de análisis (regresión o muestreo) se incorporan en resúmenes de valores perdidos y en modelos de imputación que se ajusten. Los casos de ponderaciones de análisis con valor negativo o cero se excluirán.

Muestras complejas. El procedimiento de Imputación múltiple no trata explícitamente los estratos, agrupaciones u otras estructuras de muestreo complejas, aunque puede aceptar ponderaciones de muestreo finales en la forma del análisis de la variable de ponderación. Tenga también en cuenta que los procedimientos de muestreos complejos actualmente no analizan de forma automática varios conjuntos de datos imputados. Para obtener una lista completa de procedimientos que admiten la combinación, consulte Análisis de datos de imputación múltiple el p. 28.

Valores perdidos. Los valores perdidos tanto por el usuario como por el sistema se consideran valores no válidos; es decir, ambos tipos de valores perdidos se sustituyen cuando se imputan los valores y los dos se consideran valores no válidos de variables utilizadas como predictores de modelos de imputación. Los valores perdidos por el usuario y por el sistema también se consideran perdidos en los análisis de valores perdidos.

Replicación de los resultados (Imputar valores perdidos). Si desea replicar exactamente los resultados de imputación, utilice el mismo valor de inicialización para el generador de números aleatorios, el mismo orden de datos y el mismo orden de variables, además de utilizar la misma configuración del procedimiento.

- **Generación de números aleatorios.** El procedimiento utiliza la generación de números aleatorios durante el cálculo de valores imputados. Para reproducir los mismos resultados aleatorios en el futuro, utilice el mismo valor de inicialización para el generador de números aleatorios antes de cada ejecución del procedimiento Imputar valores perdidos.
- **Orden de casos.** Los valores se imputan en el orden de casos.
- **Orden de las variables.** El método de imputación de especificación totalmente condicional imputa los valores en el orden especificado en la lista Variables de análisis.

Existen dos cuadros de diálogo dedicados a la imputación múltiple.

- Analizar patrones proporciona medidas descriptivas de los patrones de valores perdidos en los datos y puede resultar útil como paso exploratorio antes de la imputación.
- Imputar valores perdidos se utiliza para generar imputaciones múltiples. Los conjuntos de datos completos pueden analizarse con procedimientos que admiten conjuntos de datos de imputación múltiple. Consulte Análisis de datos de imputación múltiple el p. 28 para obtener información sobre el análisis de conjuntos de datos de imputación múltiple y una lista de procedimientos que admiten estos datos.

Analizar patrones

Analizar patrones proporciona medidas descriptivas de los patrones de valores perdidos en los datos y puede resultar útil como paso exploratorio antes de la imputación.

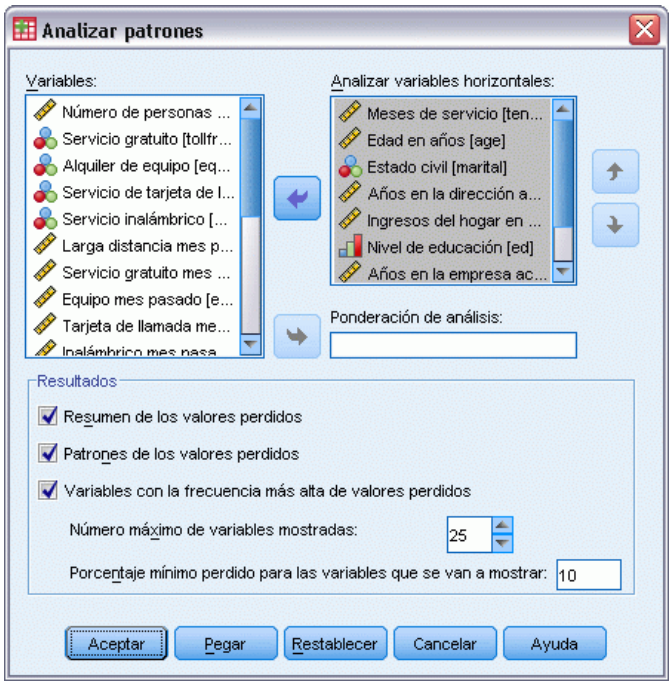
Ejemplo. Un proveedor de telecomunicaciones desea comprender mejor los patrones de uso de servicio en su base de datos de clientes. Tienen datos completos de los servicios utilizados por sus clientes, pero la información demográfica recopilada por la empresa tiene diferentes valores perdidos. El análisis de patrones de valores perdidos puede ayudar a determinar los siguientes pasos que se imputarán. Si desea obtener más información, consulte el tema Uso de imputación múltiple para completar y analizar un conjunto de datos en el capítulo 5 el p. 50.

Para analizar patrones de datos perdidos

Seleccione en los menús:

Analizar > Imputación múltiple > Analizar patrones...

Figura 3-1
Cuadro de diálogo Analizar patrones



- Seleccione al menos dos variables de análisis. El procedimiento analiza patrones de datos perdidos en estas variables.

Configuración opcional

Ponderación de análisis. Esta variable contiene ponderaciones de análisis (regresión o muestra). El procedimiento incorpora ponderaciones de análisis en resúmenes de valores perdidos. Los casos de ponderaciones de análisis con valor negativo o cero se excluirán.

Resultado. Los siguientes resultados opcionales están disponibles:

- **Resumen de valores perdidos.** Esto muestra un gráfico de sectores con paneles que indica el número y el porcentaje de variables de análisis, casos o datos individuales que tengan uno o más valores perdidos.
- **Patrones de valores perdidos.** Esto muestra patrones tabulados de valores perdidos. Cada patrón se corresponde con un grupo de casos con el mismo patrón de datos completos e incompletos sobre variables de análisis. Puede utilizar este resultado para determinar si puede utilizar el método de imputación monotónica para sus datos o, si no, en qué medida se aproximan sus datos a un patrón monotónico. El procedimiento ordena las variables de análisis para revelar o aproximarse a un patrón monotónico. Si no hay patrones que no sean monotónicos después de la reordenación, puede llegar a la conclusión de que los datos tienen un patrón monotónico cuando las variables de análisis se ordenan de tal forma.
- **Variables con la mayor frecuencia de valores perdidos.** Esto muestra una tabla de variables de análisis ordenadas por el porcentaje de valores perdidos en orden descendente. La tabla incluye estadísticos descriptivos (media y desviación típica) para variables de escala.

  Puede controlar el número máximo de variables que se mostrará y el porcentaje de ausencia mínimo de una variable para que se incluya en la visualización. Se muestra el conjunto de variables que cumplen ambos criterios. Por ejemplo, si establece el número máximo de variables como 50 y el porcentaje de ausencia mínimo como 25, hará que la tabla muestre un máximo de 50 variables que tengan un mínimo del 25 % de valores perdidos. Si hay 60 variables de análisis pero sólo 15 tienen un porcentaje igual o mayor al 25 % de valores perdidos, el resultado sólo incluirá 15 variables.

Imputar valores perdidos

Imputar valores perdidos se utiliza para generar imputaciones múltiples. Los conjuntos de datos completos pueden analizarse con procedimientos que admiten conjuntos de datos de imputación múltiple. Consulte Análisis de datos de imputación múltiple el p. 28 para obtener información sobre el análisis de conjuntos de datos de imputación múltiple y una lista de procedimientos que admiten estos datos.

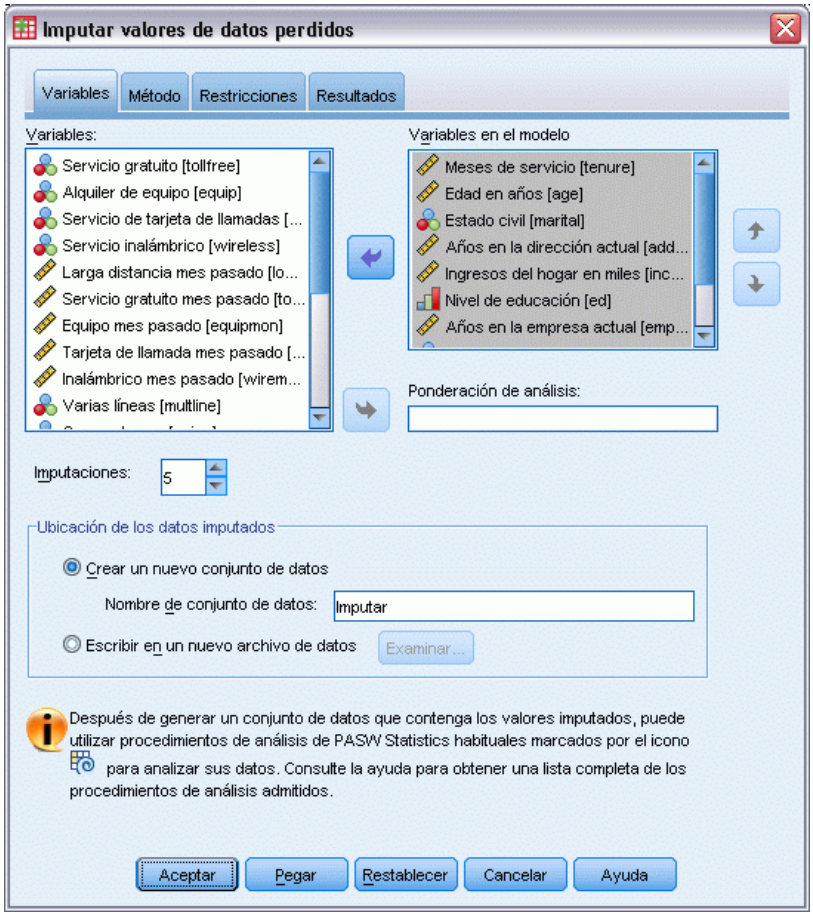
Ejemplo. Un proveedor de telecomunicaciones desea comprender mejor los patrones de uso de servicio en su base de datos de clientes. Tienen datos completos de los servicios utilizados por sus clientes, pero la información demográfica recopilada por la empresa tiene diferentes valores perdidos. Además, estos valores no están perdidos de forma aleatoria, por lo que se utilizará la imputación múltiple para completar el conjunto de datos. Si desea obtener más información, consulte el tema Uso de imputación múltiple para completar y analizar un conjunto de datos en el capítulo 5 el p. 50.

Para imputar valores perdidos

Seleccione en los menús:

Analizar > Imputación múltiple > Imputar valores de datos perdidos...

Figura 3-2
Pestaña Variables, Imputar valores perdidos



- Elija al menos dos variables en el modelo de imputación. El procedimiento imputa valores múltiples para los datos perdidos de estas variables.
- Especifique el número de imputaciones que deben calcularse. Este valor es 5 por defecto.
- Especifique un conjunto de datos o archivo de datos con formato IBM® SPSS® Statistics en el que se escribirán los datos imputados.

  El conjunto de datos de salida consiste en los datos de casos originales con datos perdidos más un conjunto de casos con valores imputados para cada imputación. Por ejemplo, si el conjunto de datos original tiene 100 casos y usted tiene cinco imputaciones, el conjunto de datos de salida contendrá 600 casos. Todas las variables del conjunto de datos de entrada se incluyen en el conjunto de datos de salida. Las propiedades de diccionario (nombres, etiquetas, etc.) de las variables existentes se copian en el nuevo conjunto de datos. El archivo también contiene una nueva variable, *Imputation_*, una variable numérica que indica la imputación (0 para datos originales o 1..*n* para casos con valores imputados).

El procedimiento define automáticamente la variable *Imputation_* como una variable de segmentación cuando se crea el conjunto de datos de salida. Si las divisiones están activadas cuando se ejecuta el procedimiento, el conjunto de datos de salida incluye un conjunto de imputaciones por cada combinación de valores de variables de segmentación.

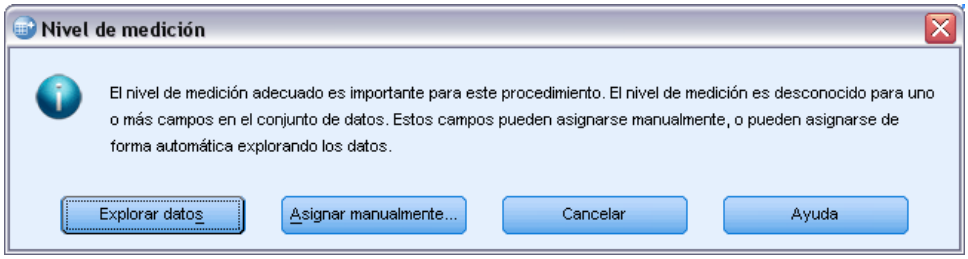
Configuración opcional

Ponderación de análisis. Esta variable contiene ponderaciones de análisis (regresión o muestra). El procedimiento incorpora ponderaciones de análisis en modelos de regresión y clasificación utilizados para imputar valores perdidos. Las ponderaciones de análisis también se utilizan en resúmenes de valores imputados; por ejemplo, media, desviación típica y error típico. Los casos de ponderaciones de análisis con valor negativo o cero se excluirán.

Campos con un nivel de medición desconocido

La alerta de nivel de medición se muestra si el nivel de medición de una o más variables (campos) del conjunto de datos es desconocido. Como el nivel de medición afecta al cálculo de los resultados de este procedimiento, todas las variables deben tener un nivel de medición definido.

Figura 3-3
Alerta de nivel de medición

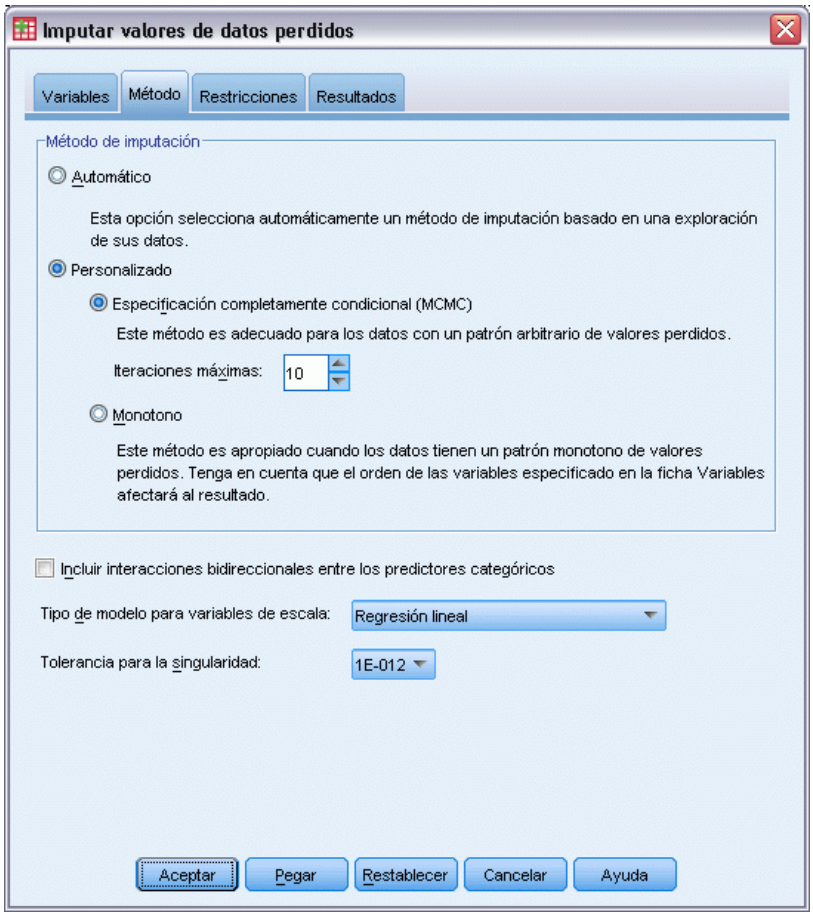


- **Explorar datos.** Lee los datos del conjunto de datos activo y asigna el nivel de medición predefinido en cualquier campo con un nivel de medición desconocido. Si el conjunto de datos es grande, puede llevar algún tiempo.
- **Asignar manualmente.** Abre un cuadro de diálogo que contiene todos los campos con un nivel de medición desconocido. Puede utilizar este cuadro de diálogo para asignar el nivel de medición a esos campos. También puede asignar un nivel de medición en la Vista de variables del Editor de datos.

Como el nivel de medición es importante para este procedimiento, no puede acceder al cuadro de diálogo para ejecutar este procedimiento hasta que se hayan definido todos los campos en el nivel de medición.

Método

Figura 3-4
Pestaña Método, Imputar valores perdidos



La pestaña Método especifica la forma en la que se imputarán los valores perdidos, incluidos los tipos de modelos utilizados. Los predictores categóricos están codificados con indicadores (dummy).

Método de imputación. El método Automático explora los datos y utiliza el método monotónico si los datos muestran un patrón monotónico de valores perdidos; de lo contrario, se utiliza la especificación totalmente condicional. Si está seguro de qué método desea utilizar, puede especificarlo como un método Personalizado.

- **Especificación totalmente condicional.** Éste es un método de Monte Carlo y cadenas de Markov (MCMC) iterativo que puede utilizarse cuando el patrón de datos perdidos es arbitrario (monotónico o no monotónico).

  El método de especificación totalmente condicional (FCS) ajusta un modelo univariante (variable dependiente simple) para cada iteración y variable en el orden especificado en la lista de variables utilizando como predictores todas las demás variables disponibles en el modelo para luego imputar los valores perdidos de las variables que se están ajustando. El método

continua hasta que se alcanza el número máximo de iteraciones y los valores imputados en la máxima iteración se guardan en el conjunto de datos imputado.

Número máximo de iteraciones. Esto especifica el número de iteraciones, o "pasos", realizadas por la cadena de Markov utilizada por el método de especificación totalmente condicional. Si el método de especificación totalmente condicional se seleccionó automáticamente, utilizará el número predeterminado de 10 iteraciones. Cuando selecciona la especificación totalmente condicional de manera explícita, puede especificar un número personalizado de iteraciones. Puede que deba aumentar el número de iteraciones si la cadena de Markov no ha convergido. En la pestaña Resultados, puede guardar los datos de historial de iteraciones de especificación totalmente condicional y realizar un gráfico de los mismos para evaluar la convergencia.

- **Monotónico.** Éste es un método no iterativo que sólo puede utilizarse cuando los datos tienen un patrón monotónico de valores perdidos. Existe un patrón monotónico cuando puede ordenar las variables de tal forma que, si una variable tiene un valor no perdidos, todas las variables precedentes también tienen valores no perdidos. Al especificar que se trata de un método Personalizado, asegúrese de especificar las variables en la lista y ordenar que muestre un patrón monotónico.

  El método monotónico ajusta un modelo univariante (variable dependiente simple) para cada variable del orden monotónico utilizando como predictores todas las variables anteriores, para luego imputar los valores perdidos de las variables que se están ajustando. Estos valores imputados se guardan en el conjunto de datos imputado.

Incluir interacciones dobles. Cuando el método de imputación se selecciona automáticamente, el modelo de imputación de cada variable incluye un término constante y los efectos principales de las variables predictoras. Al seleccionar un método específico, puede incluir opcionalmente todas las interacciones dobles posibles entre las variables predictoras categóricas.

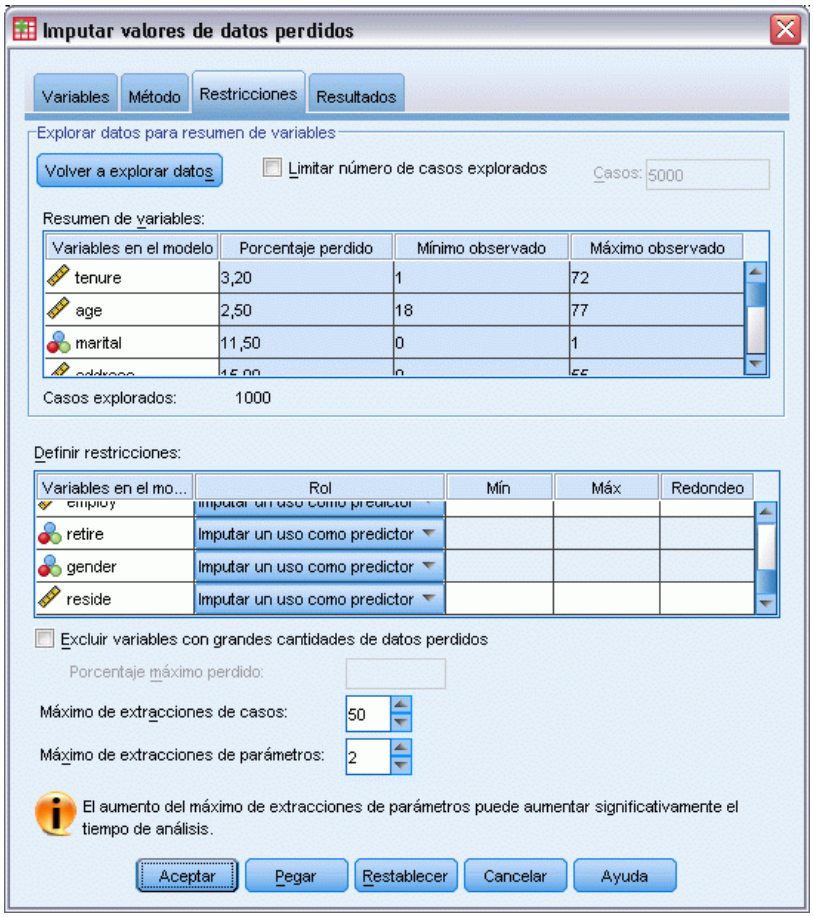
Tipo de modelo para variables de escala. Cuando el método de imputación se selecciona automáticamente, la regresión lineal se utiliza como modelo univariante para variables de escala. Al seleccionar un método específico, también puede seleccionar alternativamente equivalencia de media predictiva como modelo para variables de escala. La equivalencia de media predictiva es una variante de la regresión lineal que iguala los valores imputados calculados por el modelo de regresión con el valor observado más cercano.

La regresión logística siempre se utiliza como modelo univariante para variables categóricas. Los predictores categóricos están codificados con indicadores (dummy), independientemente del tipo de modelo.

Tolerancia para la singularidad. Las matrices singulares (que no se pueden invertir) tienen columnas linealmente dependientes, lo que causar graves problemas al algoritmo de estimación. Incluso las matrices casi singulares pueden generar resultados deficientes, por lo que el procedimiento tratará una matriz cuyo determinante es menor que la tolerancia como singular. Especifique un valor positivo.

Restricciones

Figura 3-5
Pestaña Restricciones, Imputar valores perdidos



La pestaña Restricciones le permite restringir el papel de una variable durante la imputación y restringir el rango de valores imputados de una variable de escala de modo que sean convincentes. Además, puede restringir el análisis a variables con menos de un porcentaje máximo de valores perdidos.

Exploración de datos para resumen de variables. Al pulsar en Explorar datos la lista muestra variables de análisis y el porcentaje observado de ausencia, mínimo y máximo para cada una. Los resúmenes pueden basarse en todos los casos o limitarse a una exploración de los primeros *n* casos, como aparece especificado en el cuadro de texto Casos. Puede actualizar los resúmenes de distribución al pulsar en Volver a explorar datos.

Defina las restricciones

- **Papel.** Esto le permite personalizar el conjunto de variables que deben imputarse y/o tratarse como predictores. Normalmente, cada variable de análisis se considera tanto dependiente como predictora en el modelo de imputación. El Papel puede utilizarse para desactivar la imputación de variables que desee Utilizar sólo como predictor o para excluir el uso de variables

como predictores (Sólo imputar) y, por lo tanto, hacer que el modelo de predicción sea más compacto. Ésta es la única restricción que puede especificarse para variables categóricas o para variables que se utilizan sólo como predictores.

- **Mín. y Máx.** Estas columnas le permiten especificar los valores mínimos y máximos imputados que se permiten para las variables de escala. Si un valor imputado se sale del rango, el procedimiento extrae otro valor hasta que encuentra uno dentro del rango o se alcanza al número máximo de extracciones (consulte Máximo de extracciones a continuación). Estas columnas sólo están disponibles si se selecciona Regresión lineal como el tipo de modelo de variable de escala en la pestaña Método.
- **Redondeo.** Algunas variables se pueden utilizar como escala, pero tienen valores que están restringidos de forma natural, por ejemplo, el número de miembros de una familia deben ser un número entero y la cantidad gastada durante una visita a una tienda de alimentación no puede tener decimales. Esta columna le permite especificar la menor denominación que se puede aceptar. Por ejemplo, para obtener valores enteros, especifique 1 como la denominación de redondeo; para obtener valores redondeados hacia el decimal siguiente, especifique 0,01. En general, los valores se redondean hacia el múltiplo entero más cercano a la denominación de redondeo. La siguiente tabla muestra cómo actúan los diferentes valores de redondeo sobre un valor imputado de 6.64823 (antes del redondeo).

Denominación de redondeo	Valor al que se redondea 6,64832
10	10
1	7
0.25	6.75
0.1	6.6
0.01	6.65

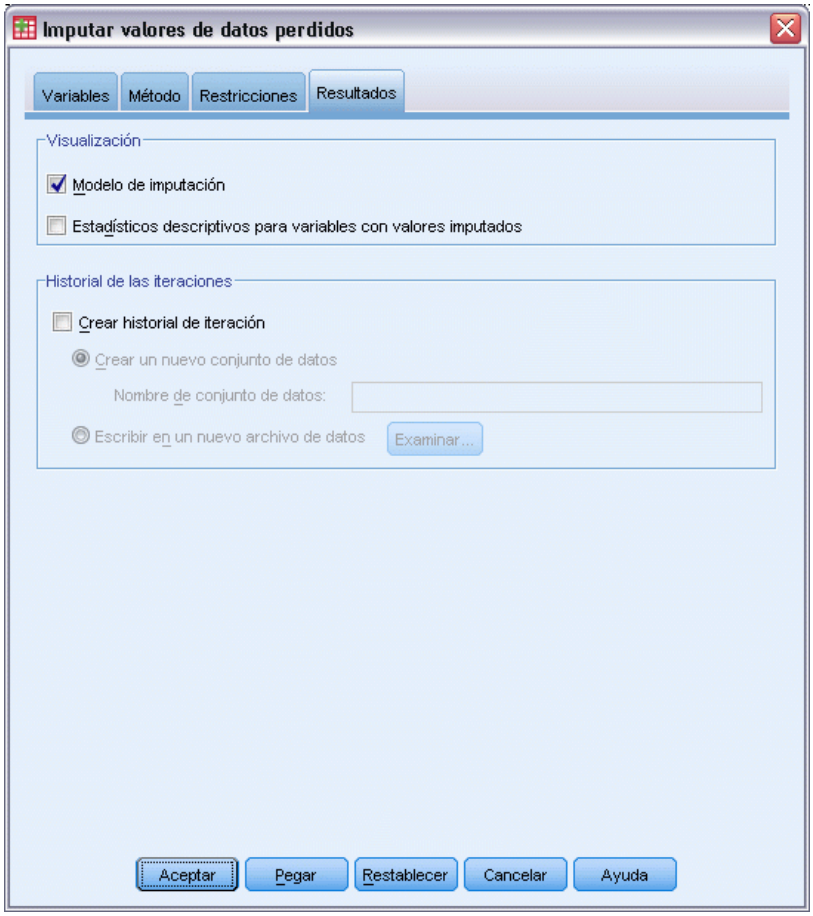
Exclusión de variables con grandes cantidades de datos perdidos. Normalmente, las variables de análisis se imputan y utilizan como predictores independientemente de cuántos valores perdidos tengan, siempre que tengan los suficientes datos para calcular un modelo de imputación. Puede decidir excluir variables con un alto porcentaje de valores perdidos. Por ejemplo, si especifica 50 como Porcentaje máximo de valores perdidos, las variables de análisis con más del 50 % de valores perdidos no se imputarán, ni se utilizarán como predictores en modelos de imputación.

Máximo de extracciones. Si se especifican valores mínimos o máximos para valores imputados de variables de escala (consulte Mín. y Máx. anteriormente), el procedimiento intentará extraer valores para un caso hasta que encuentre un conjunto de valores que se encuentren dentro de los rangos especificados. Si no se obtiene un conjunto de valores dentro del número especificado de extracciones por caso, el procedimiento extraerá otro conjunto de parámetros de modelo y repetirá el proceso de extracción de casos. Se producirá un error si no se obtiene un conjunto de valores que se halle entre los rangos dentro del número especificado de extracciones de casos y parámetros.

Tenga en cuenta que el incremento de estos valores puede incrementar el tiempo de procesamiento. Si el procedimiento tarda demasiado tiempo o si no puede encontrar extracciones adecuadas, compruebe los valores mínimos y máximos especificados para asegurarse de que sean correctos.

Resultados

Figura 3-6
Pestaña Resultados, Imputar valores perdidos



Representación. Controla la visualización de resultados. Siempre se muestra un resumen de imputación general, que incluye tablas que relacionan las especificaciones de imputación, iteraciones (para el método de especificación totalmente condicional), variables dependientes imputadas, variables dependientes excluidas de la imputación y la secuencia de imputación. Si se especifica, también se muestran las restricciones para variables de análisis.

- **Modelo de imputación.** Esto muestra el modelo de imputación para las variables dependientes y los predictores e incluye el tipo de modelo univariante, efectos de modelo y el número de valores imputados.
- **Estadísticos descriptivos.** Esto muestra estadísticos descriptivos para variables dependientes para los que se imputan valores. En el caso de las variables de escala, los estadísticos descriptivos incluyen media, recuento, desviación típica, mín. y máx. de los datos de entrada originales (antes de la imputación), valores imputados (mediante imputación) y datos completos (valores originales e imputados juntos mediante imputación). En el caso de las variables categóricas, los estadísticos descriptivos incluyen recuento y porcentaje por categoría de los datos de entrada originales (antes de la imputación), valores imputados (mediante imputación) y datos completos (valores originales e imputados juntos mediante imputación).

Historial de iteraciones. Cuando se utiliza el método de imputación de especificación totalmente condicional, puede solicitar un conjunto de datos que contenga datos del historial de iteraciones para la imputación de especificación totalmente condicional. El conjunto de datos contiene medias y desviaciones típicas mediante iteración e imputación por cada variable dependiente de escala para la que se imputan valores. Puede realizar un gráfico de los datos para ayudar a evaluar la convergencia de modelo. Si desea obtener más información, consulte el tema Comprobación de la convergencia de FCS en el capítulo 5 el p. 69.

Funciones adicionales del comando MULTIPLE IMPUTATION

La sintaxis de comandos también le permite:

- Especificar un subconjunto de variables para los que se muestran estadísticos descriptivos (subcomando `IMPUTATIONSUMMARIES`).
- Especificar tanto un análisis de patrones perdidos como la imputación en una única ejecución del procedimiento.
- Especifique el número máximo de parámetros de modelo permitido al imputar cualquier variable (palabra clave `MAXMODELPARAM`).

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

Trabajo con datos de imputación múltiple

Cuando se crea un conjunto de datos de imputación múltiple, se añade una variable llamada, *Imputation_* con etiqueta variable *Número de imputación*, y el conjunto de datos se ordena según el mismo en orden ascendente. Los casos del conjunto de datos original tienen el valor 0. Los casos de valores imputados se numeran del 1 al *M*, donde *M* es el número de imputaciones.

Cuando abre un conjunto de datos, la presencia de *Imputation_* identifica el conjunto de datos como un posible conjunto de datos de imputación múltiple.

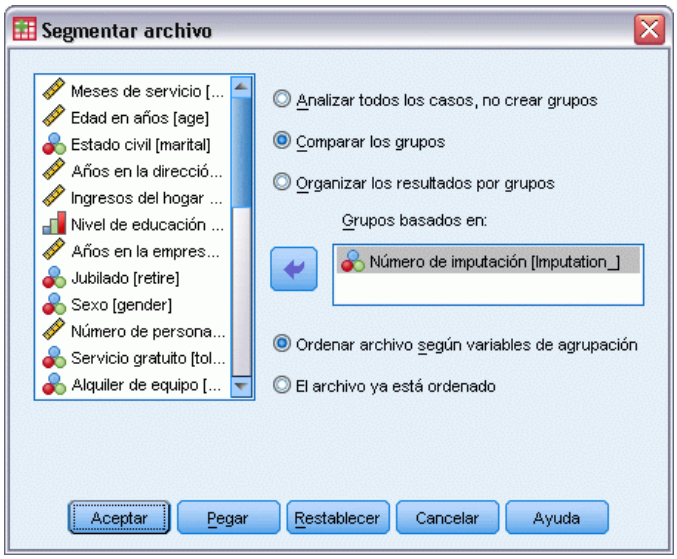
Activación de un conjunto de datos de imputación múltiple para su análisis

El conjunto de datos debe segmentarse utilizando la opción Comparar los grupos, con *Imputation_* como variable de agrupación para que se considere un conjunto de datos de imputación múltiple en los análisis. También puede definir segmentaciones en otras variables.

Seleccione en los menús:

Datos > Segmentar archivo...

Figura 3-7
Cuadro de diálogo Segmentar archivo



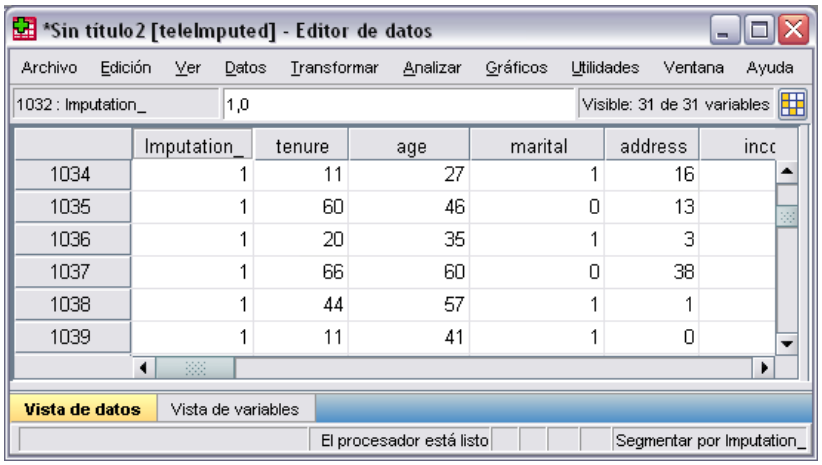
- Seleccione Comparar los grupos.
- Seleccione *Número de imputación [Imputation_]* como variable en la que agrupar los casos.

Asimismo, cuando activa las marcas (consulte a continuación), el archivo se segmenta en el *Número de imputación [Imputation_)]*.

Distinción entre valores imputados y valores observados

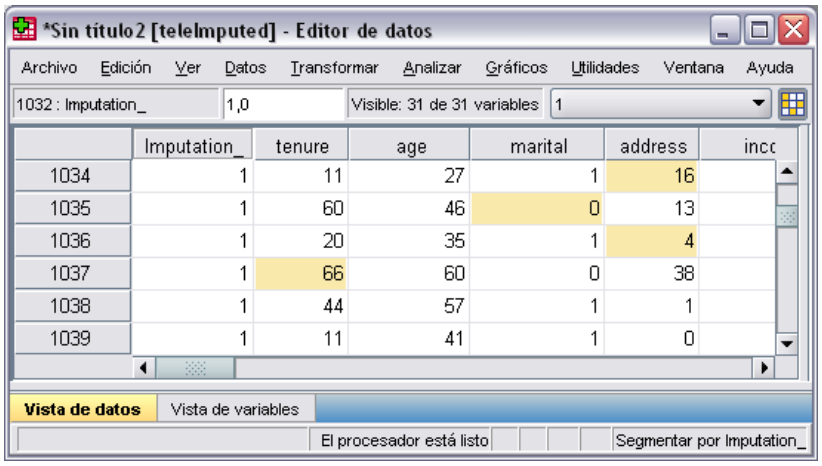
Puede distinguir entre los valores imputados y los observados según el color de fondo de las casillas, la fuente y la negrita (en el caso de valores imputados). Para obtener más detalles sobre qué marcas están activadas, consulte Opciones de imputación múltiple el p. 33. Cuando cree un conjunto de datos nuevo en la sesión actual con el procedimiento Imputar valores perdidos, las marcas se activan por defecto. Cuando abra un archivo de datos guardado que incluye imputaciones, las marcas se desactivan.

Figura 3-8
Editor de datos con marcas de imputación desactivadas



Para activar las marcas, elija en los menús del Editor de datos:
Ver > Marcar datos imputados...

Figura 3-9
Editor de datos con marcas de imputación activadas



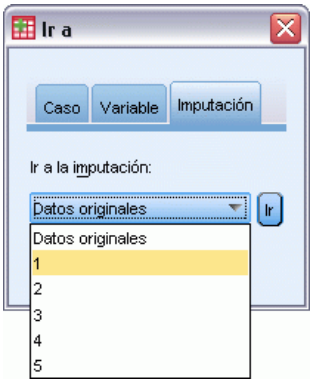
También puede activar las marcas pulsando el botón de activación/desactivación de marcas de imputación situado en el borde derecho de la barra de edición en Vista de datos del Editor de datos.

Desplazamiento entre imputaciones

- Elija en los menús:
  Editar > Ir a la imputación...

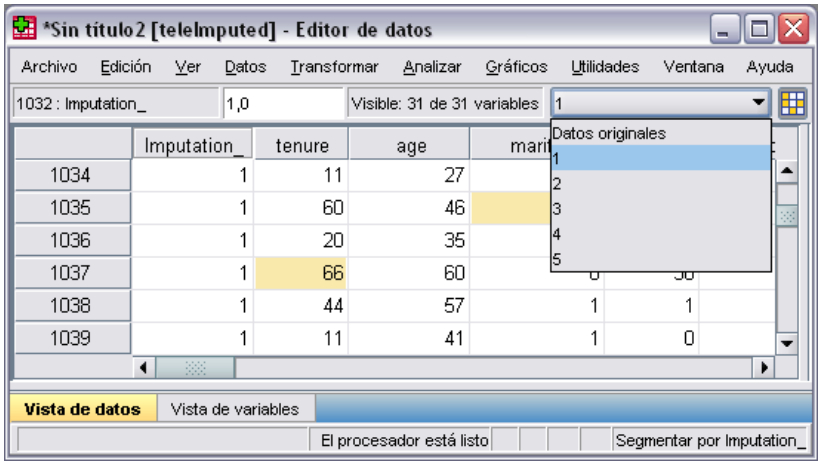
- Seleccione la imputación (o datos originales) en la lista desplegable.

Figura 3-10
Cuadro de diálogo Ir a



También puede seleccionar la imputación en la lista desplegable de la barra de edición en Vista de datos del Editor de datos.

Figura 3-11
Editor de datos con marcas de imputación activadas



La posición relativa de caso se mantiene al seleccionar imputaciones. Por ejemplo, si hay 1.000 casos en el conjunto de datos original, el caso 1.034, el 34º caso de la primera imputación, aparece en la parte superior de la cuadrícula. Si selecciona la imputación 2 en la lista desplegable, el caso 2034, el 34º caso de la segunda imputación, aparecerá en la parte superior de la cuadrícula. Si selecciona Datos originales en la lista desplegable, el caso 34 aparecerá en la parte superior de la cuadrícula. La posición de columna también se mantiene al desplazarse entre imputaciones, de modo que es fácil comparar valores entre imputaciones.

Transformación y edición de valores imputados

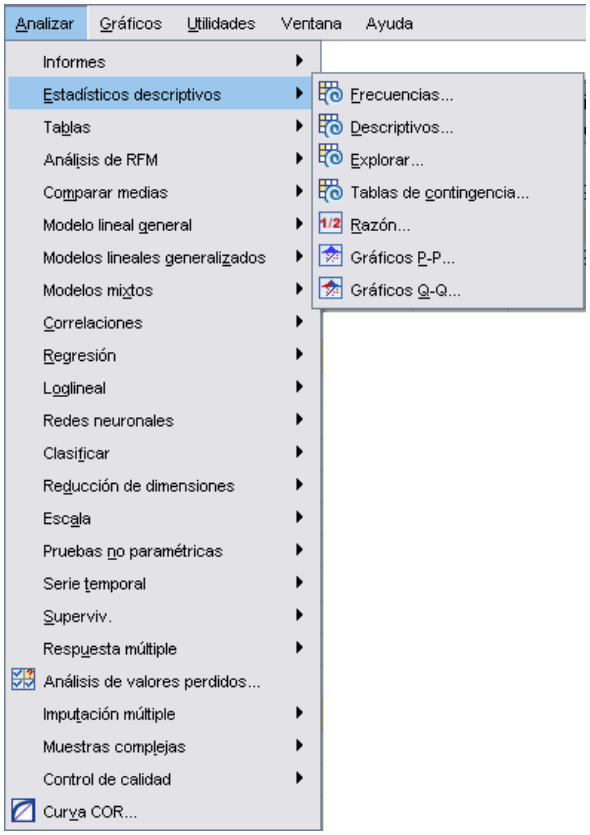
A veces deberá realizar transformaciones en datos imputados. Por ejemplo, puede que desee tomar el registro de todos los valores de una variable de salario y guardar el resultado en una nueva variable. Un valor calculado mediante datos imputados se considerará imputado si difiere del valor calculado utilizando los datos originales.

Si edita un valor imputado en una casilla del Editor de datos, dicha casilla se seguirá considerando imputada. No se recomienda editar valores imputados de esta forma.

Análisis de datos de imputación múltiple

Muchos procedimientos admiten la combinación de resultados a partir del análisis de conjuntos de datos de imputación múltiple. Cuando las marcas de imputación están activadas, aparece un icono especial junto a los procedimientos que admiten la combinación. Por ejemplo, en el submenú Estadísticos descriptivos del menú Analizar, Frecuencias, Descriptivos, Explorar y Tablas de contingencia admiten la combinación, mientras que Cociente, Gráficos P-P y Gráficos Q-Q no lo hacen.

Figura 3-12
Menú Analizar con marcas de imputación activadas



Tanto los resultados tabulares como el modelo PMML pueden combinarse. No hay ningún procedimiento nuevo para solicitar resultados combinados; en su lugar, una nueva pestaña del cuadro de diálogo Opciones le permite tener un control global sobre los resultados de imputación múltiple.

- **Combinación de resultados tabulares.** De manera predeterminada, cuando ejecuta un procedimiento compatible en un conjunto de datos de imputación múltiple, se producen resultados automáticamente para cada imputación, los datos originales (no imputados) y los

resultados combinados (finales) que tienen en cuenta la variación entre las imputaciones. Los estadísticos combinados varían según el procedimiento.

- **Combinación de PMML.** También puede obtener la combinación de PMML a partir de procedimientos compatibles que exporten PMML. El modelo PMML combinado se solicita de la misma forma que el no combinado y se guarda en su lugar.

Los procedimientos incompatibles no producen resultados combinados ni archivos PMML combinados.

Niveles de combinación

Los resultados se combinan utilizando uno de los dos niveles siguientes:

- **Combinación Naive.** Sólo está disponible el parámetro combinado.
- **Combinación univariante.** El parámetro combinado, su error típico, el estadístico de contraste y los grados de libertad eficaces, el valor *p*, el intervalo de confianza y los diagnósticos de combinación (fracción de información perdida, eficacia relativa, aumento relativo de la varianza) se mostrarán cuando estén disponibles.

Los coeficientes (regresión y correlación), quieren decir (diferencias) y los recuentos se combinan típicamente. Si el error típico del estadístico está disponible, se utiliza la combinación univariante; en caso contrario se utiliza la combinación naïve.

Procedimientos que admiten combinación

Los siguientes procedimientos admiten conjuntos de datos de imputación múltiple a los niveles de combinación especificados para cada resultado.

Frecuencias

- La tabla Estadísticos admite Medias en la combinación Univariante (si también se pide E. T. de la media) y N válido y N perdido en la combinación Naive.
- La tabla Frecuencias admite Frecuencia en la combinación Naive.

Descriptivos

- La tabla Estadísticos descriptivos admite Medias en la combinación Univariante (si también se pide E. T. de la media) y N en la combinación Naive.

Tablas de contingencia

- La tabla de contingencia admite Recuento en la combinación Naive.

Medias

- La tabla Informe admite Medias en la combinación Univariante (si también se pide E. T. de la media) y N en la combinación Naive.

Prueba T para una muestra

- La tabla Estadísticos admite Media en la combinación Univariante y N en la combinación Naive.
- La tabla Prueba admite Diferencia de medias en la combinación Naive.

Prueba T para muestras independientes

- La tabla Estadísticos de grupo admite Medias en la combinación Univariante y N en la combinación Naive.
- La tabla Prueba admite Diferencia de medias en la combinación Univariante.

Prueba T para muestras relacionadas

- La tabla Estadísticos admite Medias en la combinación Univariante y N en la combinación Naive.
- La tabla Correlaciones admite correlaciones y N en la combinación Naive.
- La tabla Prueba admite Media en la combinación Univariante.

ANOVA de un factor

- La tabla Estadísticos descriptivos admite Media en la combinación Univariante y N en la combinación Naive.
- La tabla Pruebas de contraste admite Valor de contraste en la combinación Univariante.

MLG Univariante, MLG Multivariante y MLG Repetido

- La tabla Factores inter-sujetos admite N en la combinación Naive.
- La tabla Estadísticos descriptivos admite Media y N en la combinación Naive.
- La tabla Estimaciones de los parámetros admite el coeficiente B en la combinación Univariante.
- Medias marginales estimadas: La tabla Estimaciones admite Media en la combinación Univariante.
- Medias marginales estimadas: La tabla Comparaciones por parejas admite Diferencia de medias en la combinación Univariante.

Modelos lineales mixtos

- La tabla Estadísticos descriptivos admite Media y N en la combinación Naive.
- La tabla Estimaciones de efectos fijos admite Estimación en la combinación Univariante.
- La tabla Estimaciones de parámetros de covarianzas admite Estimación en la combinación Univariante.
- Medias marginales estimadas: La tabla Estimaciones admite Media en la combinación Univariante.
- Medias marginales estimadas: La tabla Comparaciones por parejas admite Diferencia de medias en la combinación Univariante.

Modelos lineales generalizados y Ecuaciones de estimación generalizadas. Estos procedimientos admiten la combinación de PMML.

- La tabla Información sobre la variable categórica admite N y Porcentajes en la combinación Naive.
- La tabla Información sobre la variable continua admite N y Media en la combinación Naive.
- La tabla Estimaciones de los parámetros admite el coeficiente B en la combinación Univariante.
- Medias marginales estimadas: La tabla Coeficientes de estimación admite Correlaciones en la combinación Naïve.
- Medias marginales estimadas: La tabla Estimaciones admite Media en la combinación Univariante.
- Medias marginales estimadas: La tabla Comparaciones por parejas admite Diferencia de medias en la combinación Univariante.

Correlaciones bivariadas

- La tabla Estadísticos descriptivos admite Media y N en la combinación Naive.
- La tabla Correlaciones admite correlaciones y N en la combinación Naive.

Correlaciones parciales

- La tabla Estadísticos descriptivos admite Media y N en la combinación Naive.
- La tabla Correlaciones admite correlaciones en la combinación Naive.

Regresión lineal. Este procedimiento admite la combinación de PMML.

- La tabla Estadísticos descriptivos admite Media y N en la combinación Naive.
- La tabla Correlaciones admite correlaciones y N en la combinación Naive.
- La tabla Coeficientes admite B en la combinación Univariante y Correlaciones en la combinación Naive.
- La tabla Coeficientes de correlación admite Correlaciones en la combinación Naive.
- La tabla Estadísticos residuales admite Media y N en la combinación Naive.

Regresión logística binaria. Este procedimiento admite la combinación de PMML.

- La tabla Variables en la ecuación admite B en la combinación Univariante.

Regresión logística multinomial. Este procedimiento admite la combinación de PMML.

- La tabla Estimaciones de los parámetros admite el coeficiente B en la combinación Univariante.

Regresión ordinal

- La tabla Estimaciones de los parámetros admite el coeficiente B en la combinación Univariante.

Análisis discriminante. Este procedimiento admite la combinación del modelo XML.

- La tabla Estadísticos de grupo admite Media y N válido en la combinación Naive.
- La tabla Matrices intra-grupos combinadas admite Correlaciones en la combinación Naive.

- La tabla Coeficientes de funciones discriminantes canónicas admite Coeficientes no tipificados en la combinación Naive.
- La tabla Funciones en centroides de grupo admite Coeficientes no tipificados en la combinación Naive.
- La tabla Coeficientes de función de clasificación admite Coeficientes en la combinación Naive.

Prueba de chi-cuadrado

- La tabla Descriptivos admite Media y N en la combinación Naive.
- La tabla Frecuencias admite N Observado en la combinación Naive.

Prueba binomial

- La tabla Descriptivos admite Medias y N en la combinación Naive.
- La tabla Prueba admite N, Proporción observada y Proporción de prueba en la combinación Naive.

Prueba de rachas

- La tabla Descriptivos admite Medias y N en la combinación Naive.

Prueba Kolmogorov-Smirnov de una muestra

- La tabla Descriptivos admite Medias y N en la combinación Naive.

Pruebas para dos muestras independientes

- La tabla Rangos admite Rango promedio y N en la combinación Naive.
- La tabla Frecuencias admite N en la combinación Naive.

Pruebas para varias muestras independientes

- La tabla Rangos admite Rango promedio y N en la combinación Naive.
- La tabla Frecuencias admite Recuento en la combinación Naive.

Pruebas para dos muestras relacionadas

- La tabla Rangos admite Rango promedio y N en la combinación Naive.
- La tabla Frecuencias admite N en la combinación Naive.

Pruebas para varias muestras relacionadas

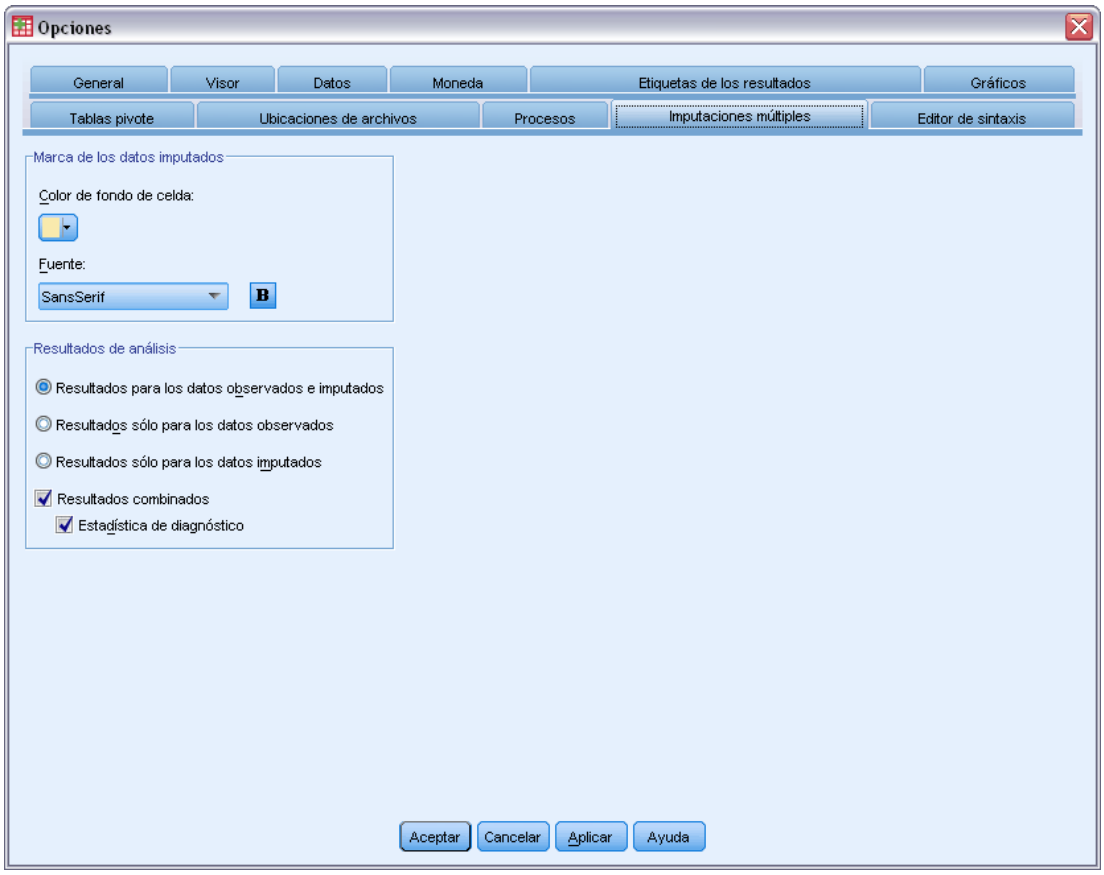
- La tabla Rangos admite Rango promedio en la combinación Naive.

Regresión de Cox. Este procedimiento admite la combinación de PMML.

- La tabla Variables en la ecuación admite B en la combinación Univariante.
- La tabla Medias de covariables admite media en la combinación Naive.

Opciones de imputación múltiple

Figura 3-13
Cuadro de diálogo Opciones: Pestaña Imputaciones múltiples



La pestaña Imputaciones múltiples controla dos tipos de preferencias relacionadas con las imputaciones múltiples:

Aspecto de datos imputados. De manera predeterminada, las casillas que contienen datos imputados tendrán un color de fondo diferente que las casillas con datos no imputados. El aspecto distintivo de los datos imputados debería facilitarle el desplazamiento por un conjunto de datos y la localización de estas casillas. Puede cambiar el color de fondo predeterminado de las casillas, la fuente y hacer que los datos imputados aparezcan en negrita.

Resultados. Este grupo controla el tipo de resultados del Visor producidos cuando se analiza un conjunto de datos imputado de forma múltiple. De manera predeterminada, se producirán resultados para el conjunto de datos originales (de antes de la imputación) y para cada uno de los conjuntos de datos imputados. Además, se generarán resultados combinados finales para los procedimientos que sean compatibles con la combinación de datos imputados. Los diagnósticos de combinación también aparecerán cuando se realice una combinación univariante. Sin embargo, puede suprimir los resultados que no desee ver.

Para establecer opciones de imputación múltiple

Elija en los menús:
Edición > Opciones

Pulse la pestaña Imputaciones múltiples.

Parte II: Ejemplos

Capítulo
4

Missing Value Analysis

Descripción del patrón de los datos perdidos

Un proveedor de telecomunicaciones desea comprender mejor los patrones de uso de servicio en su base de datos de clientes. La compañía quiere asegurarse de que los datos están perdidos completamente al azar antes de llevar a cabo más análisis.

telco_missing.sav contiene una muestra aleatoria de la base de datos de clientes. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Missing Values 19*.

Ejecución del análisis para mostrar estadísticos descriptivos

- Para ejecutar el análisis de valores perdidos, elija en los menús:
  Analizar > Análisis de valores perdidos...

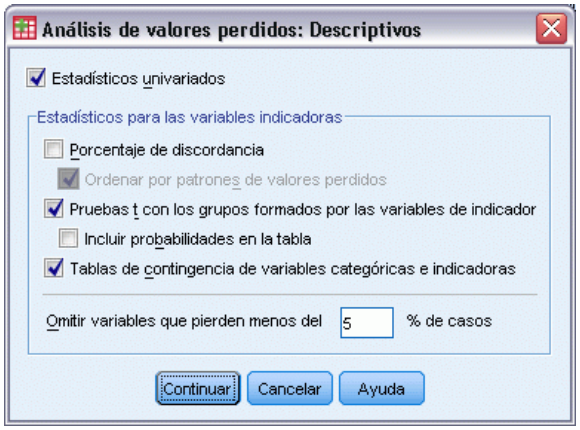
Figura 4-1
Cuadro de diálogo Análisis de valores perdidos

- Seleccione *Estado civil [ecivil]*, *Nivel educativo [ed]*, *Retirado [retire]* y *Sexo [sexo]* como variables categóricas.
- Seleccione desde *Meses de servicio [cargo]* hasta *Número de personas en el hogar [reside]* como variables cuantitativas (escala).

  En este punto, se puede llevar a cabo el procedimiento y obtener estadísticos univariados, pero seleccionaremos estadísticos descriptivos adicionales.

- Pulse en Descriptivos.

Figura 4-2
Cuadro de diálogo Análisis de valores perdidos: Cuadro de diálogo Descriptivos



En el cuadro de diálogo Descriptivos, puede especificar distintos estadísticos descriptivos para mostrarlos en los resultados. Los estadísticos univariados por defecto pueden ayudar a determinar el grado general de los datos perdidos, pero los estadísticos para las variables-indicador ofrecen más información sobre cómo puede afectar el patrón de los datos perdidos de una variable a los valores de otra variable.

- Seleccione Pruebas t con los grupos formados por las variables de indicador.
- Seleccione Tablas de contingencia de variables categóricas y de indicador.
- Pulse en Continuar.
- En el cuadro de diálogo principal Análisis de valores perdidos, pulse en Aceptar.

Evaluación de los estadísticos descriptivos

Para este ejemplo, los resultados incluyen:

- Estadísticos univariantes
- Tabla de pruebas *t* de varianzas separadas, que incluyen medias de subgrupos cuando otra variable está presente o está perdida
- Tablas para cada variable categórica que muestran frecuencias de datos perdidos para cada categoría por cada variable cuantitativa (de escala)

Figura 4-3
Tabla de estadísticos univariados

	N	Media	Desviación típ.	Perdidos		No de extremos[a]	
				Recuento	Porcentaje	Bajos	Altos
MonthsWithService	968	35,56	21,268	32	3,2	0	0
Age	975	41,75	12,573	25	2,5	0	0
YearsAtAddress	850	11,47	9,965	150	15,0	0	9
Income	821	71,1462	83,14424	179	17,9	0	71
PeopleInHousehold	966	2,32	1,431	34	3,4	0	33
YearsWithEmployer	904	11,00	10,113	96	9,6	0	15
EducationalLevel	965			35	3,5		
RetirementStatus	916			84	8,4		
Gender	958			42	4,2		
MaritalStatus	885			115	11,5		

a. Número de casos fuera del rango (C1 - 1.5*AIC, C3 + 1.5*AIC).

Los estadísticos univariados proporcionan una primera idea, variable por variable, acerca del impacto de los datos perdidos. El número de valores no perdidos para cada variable aparece en la columna *N* y el número de valores perdidos aparece en la columna *Recuento de perdidos*. La columna *Porcentaje de perdidos* muestra el porcentaje de casos con valores que faltan y ofrece una buena medida para comparar el grado de datos que faltan entre las variables. *ingres (Ingresos del hogar en miles)* tiene el mayor número de casos con valores que faltan (17,9%), mientras que *edad (Edad en años)* tiene el menor (2,5%). *income* también tiene el mayor número de valores extremos.

Figura 4-4
Tabla de pruebas t de varianzas separadas

		MonthsWithServic	Age	YearsAtAddress	Income	YearsWithEmployer	PeopleInHousehold
address	t	,4	,3	.	3,5	1.4	1.0
	gl	202,2	192,5	.	313,6	191.1	199.5
	#no presente	819	832	850	693	766	824
	#no perdido	149	143	0	128	138	142
	Media(Presentes)	35,68	41,79	11.47	74,0779	11.20	2.34
	Media(Perdidos)	34,91	41,49	.	55,2734	9.86	2.21
income	t	-5,0	-8,3	-3,9	.	-5.9	3.6
	gl	249,5	222,8	191,1	.	203.3	315.2
	#no presente	793	801	693	821	741	792
	#no perdido	175	174	157	0	163	174
	Media(Presentes)	33,93	40,01	10,67	71,1462	9.91	2.39
	Media(Perdidos)	42,97	49,73	14,97	.	15.93	2.02
employ	t	-1,0	-,4	-,7	,5	.	-.3
	gl	110,5	110,2	97,6	114,9	.	110,9
	#no presente	877	881	766	741	904	874
	#no perdido	91	94	84	80	0	92
	Media(Presentes)	35,34	41,69	11,37	71,4953	11,00	2,31
	Media(Perdidos)	37,70	42,27	12,32	67,9125	.	2,37
marital	t	,0	1,8	1,2	-,8	.9	-2.2
	gl	148,1	149,5	138,8	121,2	128.3	134,2
	#no presente	856	862	748	728	805	857
	#no perdido	112	113	102	93	99	109
	Media(Presentes)	35,56	42,00	11,61	70,3887	11,10	2,28
	Media(Perdidos)	35,57	39,85	10,43	77,0753	10,17	2,61
retire	t	-.6	-.4	-.4	.3	.	.2
	gl	95.4	94.4	84.0	93.2	.	99.0
	#no presente	888	893	777	751	904	885
	#no perdido	80	82	73	70	0	81
	Media(Presentes)	35,44	41,70	11,42	71,3356	11,00	2,32
	Media(Perdidos)	36,89	42,29	11,96	69,1143	.	2,30

La tabla de pruebas *t* de varianzas separadas puede ayudar a identificar variables cuyo patrón de valores perdidos puede estar influyendo en las variables cuantitativas (de escala). La prueba *t* se calcula mediante una variable de indicador que especifica si una variable está presente o perdida para un caso individual. Las medias de subgrupo para la variable indicadora también se incluyen en la tabla. Tenga en cuenta que sólo se crea una variable indicadora si una variable tiene valores perdidos en al menos el 5% de los casos.

Parece que los encuestados mayores son menos propensos a informar sobre su nivel de ingresos. Cuando *ingresos* está perdida, la media *edad* es 49,73, comparada con 40,01 cuando *Ingresos* no está perdida. De hecho, la ausencia de *ingresos* parece afectar a las medias de varias variables cuantitativas (de escala). Esto indica que los datos pueden no estar perdidos completamente al azar.

Figura 4-5
Tabla de contingencia de Estado civil [ecivil]

			Total	Unmarried	Married	Perdidos
						Perd. sistema
YearsAtAddress	Presente	Recuento	850	390	358	102
		Porcentaje	85,0	85,5	83,4	88,7
	Perdidos	% perd. sistema	15,0	14,5	16,6	11,3
Income	Presente	Recuento	821	380	348	93
		Porcentaje	82,1	83,3	81,1	80,9
	Perdidos	% perd. sistema	17,9	16,7	18,9	19,1
YearsWithEmployer	Presente	Recuento	904	418	387	99
		Porcentaje	90,4	91,7	90,2	86,1
	Perdidos	% perd. sistema	9,6	8,3	9,8	13,9
RetirementStatus	Presente	Recuento	916	423	392	101
		Porcentaje	91,6	92,8	91,4	87,8
	Perdidos	% perd. sistema	8,4	7,2	8,6	12,2

Las variables indicador con menos del 5% de sus valores perdidos no se muestran.

Las tablas de contingencia de las variables categóricas respecto a las variables indicadoras muestran información similar a la que se encuentra en la tabla de prueba *t* de varianzas separadas. Las variables indicadoras se vuelven a crear, con la excepción de que esta vez se usan para calcular las frecuencias de cada categoría de cada variable categórica. Los valores pueden ayudarle a determinar si existen diferencias en los valores perdidos entre las categorías.

Si observamos la tabla de *ecivil (Estado civil)*, el número de valores perdidos en las variables indicadoras no parece variar mucho entre las categorías de *ecivil*. El hecho de que alguien esté casado o soltero no parece afectar a si los datos están perdidos en ninguna de las variables cuantitativas (de escala). Por ejemplo, las personas solteras indicaron *dirección (Años en dirección)* el 85,5% de las veces y las personas casadas informaron de la misma variable el 83,4% de las veces. La diferencia es mínima y probablemente se deba al azar.

Figura 4-6
Tabla de contingencia de Nivel educativo [ed]

			Total	Did not complete high school	High school degree	Some college	College degree	Post-undergraduate degree	Perdidos
									Perd. sistema
YearsAtAddress	Presente	Recuento	850	163	240	175	186	56	30
		Porcentaje	85,0	83,2	85,7	88,4	81,9	87,5	85,7
	Perdidos	% perd. sistema	15,0	16,8	14,3	11,6	18,1	12,5	14,3
Income	Presente	Recuento	821	155	229	165	193	50	29
		Porcentaje	82,1	79,1	81,8	83,3	85,0	78,1	82,9
	Perdidos	% perd. sistema	17,9	20,9	18,2	16,7	15,0	21,9	17,1
YearsWithEmployer	Presente	Recuento	904	178	254	178	204	60	30
		Porcentaje	90,4	90,8	90,7	89,9	89,9	93,8	85,7
	Perdidos	% perd. sistema	9,6	9,2	9,3	10,1	10,1	6,2	14,3
RetirementStatus	Presente	Recuento	916	180	259	180	207	60	30
		Porcentaje	91,6	91,8	92,5	90,9	91,2	93,8	85,7
	Perdidos	% perd. sistema	8,4	8,2	7,5	9,1	8,8	6,2	14,3
MaritalStatus	Presente	Recuento	885	193	278	148	184	52	30
		Porcentaje	88,5	98,5	99,3	74,7	81,1	81,2	85,7
	Perdidos	% perd. sistema	11,5	1,5	,7	25,3	18,9	18,8	14,3

Las variables indicador con menos del 5% de sus valores perdidos no se muestran.

Consideremos ahora la tabla de contingencia para *ed (Nivel educativo)*. Si un encuestado tiene cierto grado de educación universitaria, es más probable que la respuesta del estado civil esté perdida. Al menos el 98,5% de los encuestados sin educación universitaria informaron del estado civil. Por otro lado, sólo el 81,1% de los encuestados con un título universitario informaron sobre el estado civil. El número es incluso inferior para los encuestados con cierto grado de educación universitaria pero sin título.

Figura 4-7
Tabla de contingencia de Retirado [retire]

			Total	No	Yes	Perdidos
						Perd. sistema
YearsAtAddress	Presente	Recuento	850	744	33	73
		Porcentaje	85,0	85,0	80,5	86,9
	Perdidos	% perd. sistema	15,0	15,0	19,5	13,1
Income	Presente	Recuento	821	732	19	70
		Porcentaje	82,1	83,7	46,3	83,3
	Perdidos	% perd. sistema	17,9	16,3	53,7	16,7
YearsWithEmployer	Presente	Recuento	904	864	40	0
		Porcentaje	90,4	98,7	97,6	,0
	Perdidos	% perd. sistema	9,6	1,3	2,4	100,0
MaritalStatus	Presente	Recuento	885	777	38	70
		Porcentaje	88,5	88,8	92,7	83,3
	Perdidos	% perd. sistema	11,5	11,2	7,3	16,7

Las variables indicador con menos del 5% de sus valores perdidos no se muestran.

Se puede observar una diferencia más drástica en *retire (Retirado).* Los encuestados que están retirados son mucho menos propensos a informar sobre sus ingresos comparados con los encuestados que no están retirados. Sólo el 46,3% de los clientes retirados informó sobre el nivel de ingresos, mientras que el porcentaje de los que no están retirados e informaron sobre el nivel de ingresos fue del 83,7.

Figura 4-8
Tabla de contingencia para Sexo [sexo]

			Total	Male	Female	Perdidos
						Perd. sistema
YearsAtAddress	Presente	Recuento	850	363	456	31
		Porcentaje	85,0	78,6	91,9	73,8
	Perdidos	% perd. sistema	15,0	21,4	8,1	26,2
Income	Presente	Recuento	821	381	406	34
		Porcentaje	82,1	82,5	81,9	81,0
	Perdidos	% perd. sistema	17,9	17,5	18,1	19,0
YearsWithEmployer	Presente	Recuento	904	412	457	35
		Porcentaje	90,4	89,2	92,1	83,3
	Perdidos	% perd. sistema	9,6	10,8	7,9	16,7
RetirementStatus	Presente	Recuento	916	420	461	35
		Porcentaje	91,6	90,9	92,9	83,3
	Perdidos	% perd. sistema	8,4	9,1	7,1	16,7
MaritalStatus	Presente	Recuento	885	400	445	40
		Porcentaje	88,5	86,6	89,7	95,2
	Perdidos	% perd. sistema	11,5	13,4	10,3	4,8

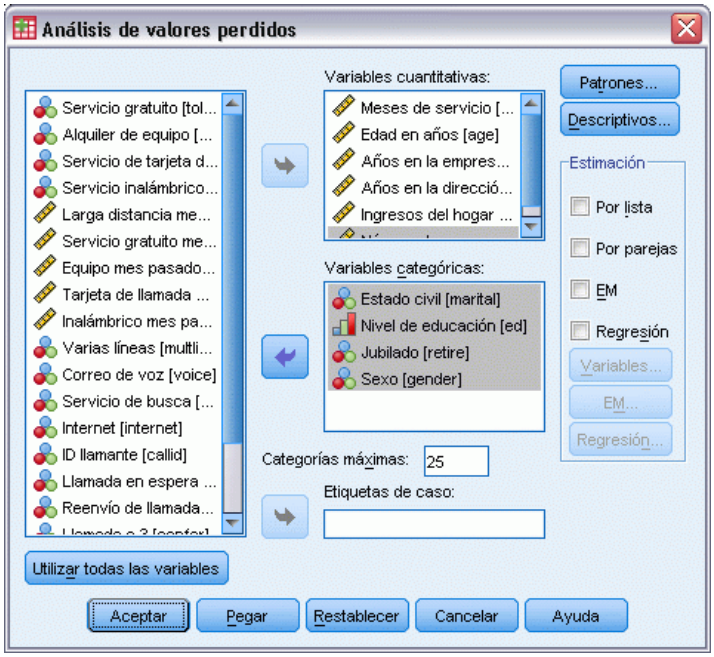
Las variables indicador con menos del 5% de sus valores perdidos no se muestran.

Existe otra discrepancia clara con *sexo (Sexo)*. La información de la dirección falta más a menudo en los hombres que en las mujeres. Aunque estas discrepancias podrían deberse al azar, no parece muy probable. Los datos no parecen estar perdidos completamente al azar.

Observaremos los patrones de los datos perdidos para estudiar más detalles.

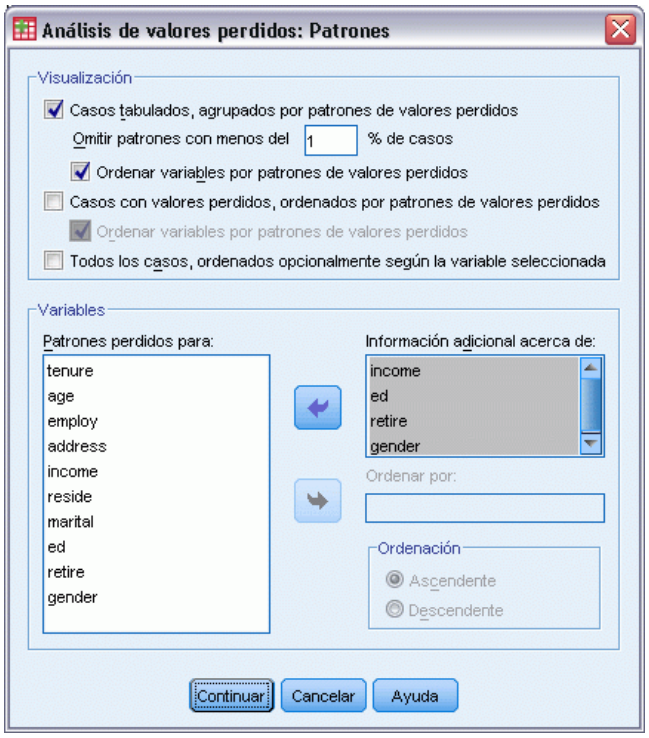
Volver a ejecutar el análisis para mostrar patrones

Figura 4-9
Cuadro de diálogo Análisis de valores perdidos



- Vuelva a mostrar el cuadro de diálogo Análisis de valores perdidos. El cuadro de diálogo recuerda las variables utilizadas en el análisis anterior. No las modifique.
- Pulse en Patrones.

Figura 4-10
Cuadro de diálogo Análisis de valores perdidos: Patrones



En el cuadro de diálogo Patrones, se pueden seleccionar varias tablas de patrones. Mostraremos los patrones tabulados agrupados por patrones de valores perdidos. Dado que los patrones perdidos de *ed (Nivel educativo)*, *retire (Retirado)* y *sexo (Sexo)* parecen influir en los datos, elegiremos mostrar información adicional sobre estas variables. También incluiremos información adicional para *ingres (Ingresos del hogar en miles)* debido al gran número de valores perdidos.

- Seleccione Casos tabulados, agrupados por patrones de valores perdidos.
- Seleccione *ingres*, *ed*, *retire* y *sexo* y añádalos a la lista Información adicional acerca de.
- Pulse en Continuar.
- En el cuadro de diálogo principal Análisis de valores perdidos, pulse en Aceptar.

Evaluación de la tabla de patrones

Figura 4-11
Tabla de patrones tabulados

	Patrones perdidos[a]												EducationalLevel[d]					RetirementStatus[d]		Gender[d]	
Número de casos	Age	PeopleInHousehold	MonthsWithService	EducationalLevel	Gender	RetirementStatus	YearsWithEmployer	MaritalStatus	YearsAtAddress	Income	Completo si...[b]	Income[c]	Did not complete high school	High school degree	Some college	College degree	Post-undergraduate degree	No	Yes	Male	Female
475											475	76,5853	99	157	87	101	31	463	12	201	274
109										X	584	.	27	35	19	17	11	95	14	47	62
16									X	X	687	.	5	9	0	1	1	12	4	12	4
87									X		562	54,4368	21	27	9	24	6	85	2	66	21
13		X									488	56,0000	4	3	2	3	1	13	0	4	9
60								X			535	77,2167	1	2	27	24	6	59	1	35	25
16				X							491	47,8125	0	0	0	0	0	16	0	6	10
17			X								492	76,2353	2	7	3	4	1	17	0	7	10
18					X						493	54,1111	3	7	4	4	0	17	1	0	0
16								X		X	660	.	0	0	7	8	1	14	2	6	10
37						X	X				520	59,4595	9	14	5	8	1	0	0	15	22

Los patrones con menos del 1% de los casos (10 o menos) no se muestran.

a. Las variables se ordenan según los patrones perdidos.

b. Número de casos completos si las variables perdidas en ese patrón (marcado con X) no se utilizan.

c. Medias en cada patrón único

d. Distribución de frecuencias en cada patrón único

La tabla de patrones tabulados muestra si los datos tienden a estar perdidos para varias variables en casos individuales. Es decir, puede ayudarle a determinar si los datos están perdidos conjuntamente.

Existen tres patrones de datos perdidos conjuntamente que se producen en más del 1% de los casos. Las variables *empcat (Años con la empresa actual)* y *retire (Retirado)* están perdidas conjuntamente con más frecuencia que otros pares. Esto no resulta sorprendente, ya que *retire* y *empcat* registran información similar. Si no sabe si un encuestado está retirado, probablemente tampoco conocerá los años que lleva con la empresa actual el encuestado.

La media *ingres (Ingresos del hogar en miles)* parece variar considerablemente dependiendo del patrón de valores perdidos. Concretamente, la media *ingres* es mucho más alta en el 6% (60 de 1000) de los casos, cuando *ecivil (Estado civil)* está perdido. (También es más alta cuando *cargo (Meses con servicio)* está perdido, pero este patrón sólo tiene en cuenta el 1,7% de los casos.) Recuerde que los encuestados con un nivel educativo superior eran menos propensos a responder a la pregunta sobre el estado civil. Esta tendencia se puede observar en las frecuencias mostradas para *ed (Nivel educativo)*. Podemos explicar el aumento de *ingres* si suponemos que los encuestados con un nivel de educación superior ganan más dinero y son menos propensos a informar sobre su estado civil.

Considerando los estadísticos descriptivos y los patrones de datos perdidos, podemos concluir que los datos no están perdidos completamente al azar. Podemos confirmar esta conclusión mediante la prueba MCAR de Little, que se imprime con las estimaciones EM.

Volver a ejecutar el análisis de la prueba MCAR de Little

Figura 4-12
Cuadro de diálogo Análisis de valores perdidos

- Vuelva a mostrar el cuadro de diálogo Análisis de valores perdidos.
- Pulse en EM.
- Pulse en Aceptar.

Figura 4-13
Tabla Medias marginales estimadas

MonthsWithService	Age	YearsAtAddress	Income	YearsWithEmployer	PeopleInHousehold
36,12	41,90	11,58	77,3941	11,22	2,29

a. Prueba MCAR de Little: Chi-cuadrado = 179,836, GL = 107, Sig. = ,000

Los resultados de la prueba MCAR de Little aparecen en las notas al pie de cada tabla de estimaciones EM. La hipótesis nula de la prueba MCAR de Little es que los datos están perdidos completamente al azar (MCAR). Los datos están MCAR cuando el patrón de valores perdidos no depende de los valores de los datos. Dado que el valor de significación es inferior a 0,05 en nuestro ejemplo, podemos concluir que los datos *no* están perdidos completamente al azar. Esto confirma la conclusión que se dedujo de los estadísticos descriptivos y los patrones tabulados.

En este punto, como los datos no están perdidos completamente al azar, no es seguro eliminar según la lista casos con valores perdidos ni imputar valores perdidos individualmente. Sin embargo, puede utilizar imputación múltiple para analizar más este conjunto de datos.

Capítulo
5

Imputación múltiple

Uso de imputación múltiple para completar y analizar un conjunto de datos

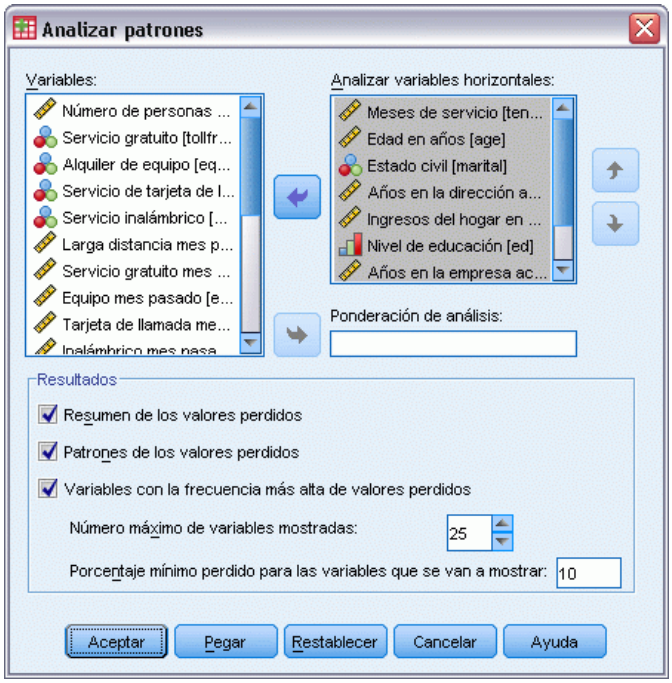
Un proveedor de telecomunicaciones desea comprender mejor los patrones de uso de servicio en su base de datos de clientes. Tienen datos completos de los servicios utilizados por sus clientes, pero la información demográfica recopilada por la empresa tiene diferentes valores perdidos. Además, estos valores no están perdidos de forma aleatoria, por lo que se utilizará la imputación múltiple para completar el conjunto de datos.

telco_missing.sav contiene una muestra aleatoria de la base de datos de clientes. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Missing Values 19*.

Análisis de los patrones de los valores perdidos

- En primer lugar, mire los patrones de datos perdidos. Elija en los menús:
  Analizar > Imputación múltiple > Analizar patrones...

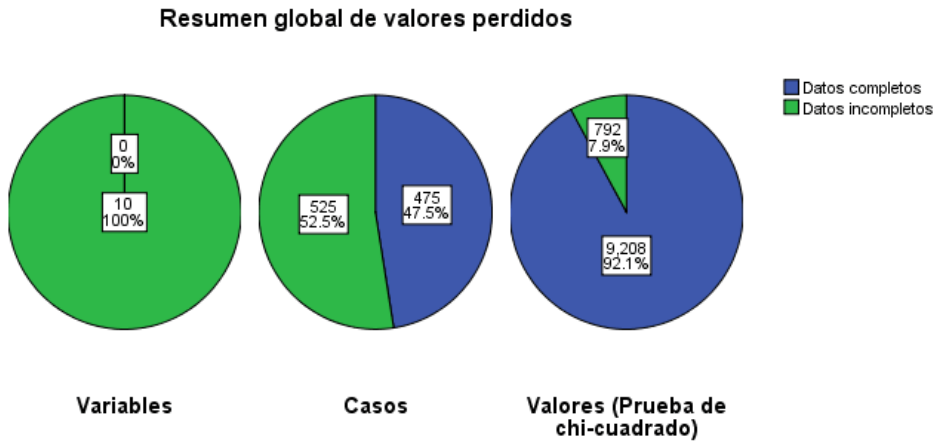
Figura 5-1
Cuadro de diálogo Analizar patrones



- Seleccione desde *Meses de servicio [cargo]* hasta *Número de personas en el hogar [reside]* como variables de análisis.

Resumen global

Figura 5-2
Resumen global de valores perdidos



El resumen global de valores perdidos muestra tres gráficos de sectores que muestran diferentes aspectos de los valores perdidos en los datos.

- El gráfico *Variables* muestra que las 10 variables de análisis tiene al menos un valor perdido en un caso.
- El gráfico *Casos* muestra que 525 de los 1000 casos tienen al menos un valor perdido en una variable.
- El gráfico *Valores* muestra que faltan 792 de los 10.000 valores (casos × variables).

Cada caso con valores perdidos tiene, de media, valores perdidos en aproximadamente 1,5 de las 10 variables. Sugiere que **eliminación por lista** perdería gran parte de la información del conjunto de datos.

Resumen de variables

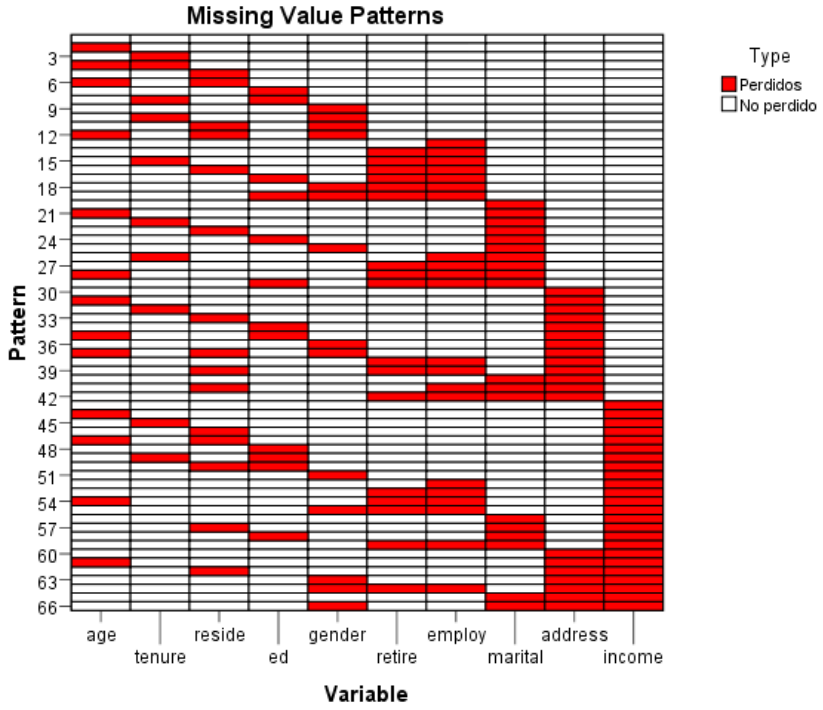
Figura 5-3
Resumen de variables

	Perdidos		N válido	Media	Desviación típica
	N	Porcentaje			
Ingresos del hogar en ...	179	17,9%	821	71,1462	83,14424
Años en la dirección ...	150	15,0%	850	11,47	9,965
Estado civil	115	11,5%	885		

Se muestra el resumen de variable de las variables con al menos el 10% de los valores perdidos y muestra el número y porcentaje de valores perdidos de cada variable de la tabla. También muestra la media y la desviación típica de los valores válidos de variables de escala y el número de valores válidos de todas las variables. *Ingresos del hogar en miles*, *Años en la dirección actual* y *Estado civil* tienen la mayoría de valores perdidos, en ese orden.

Patrones

Figura 5-4
Patrones de valores perdidos

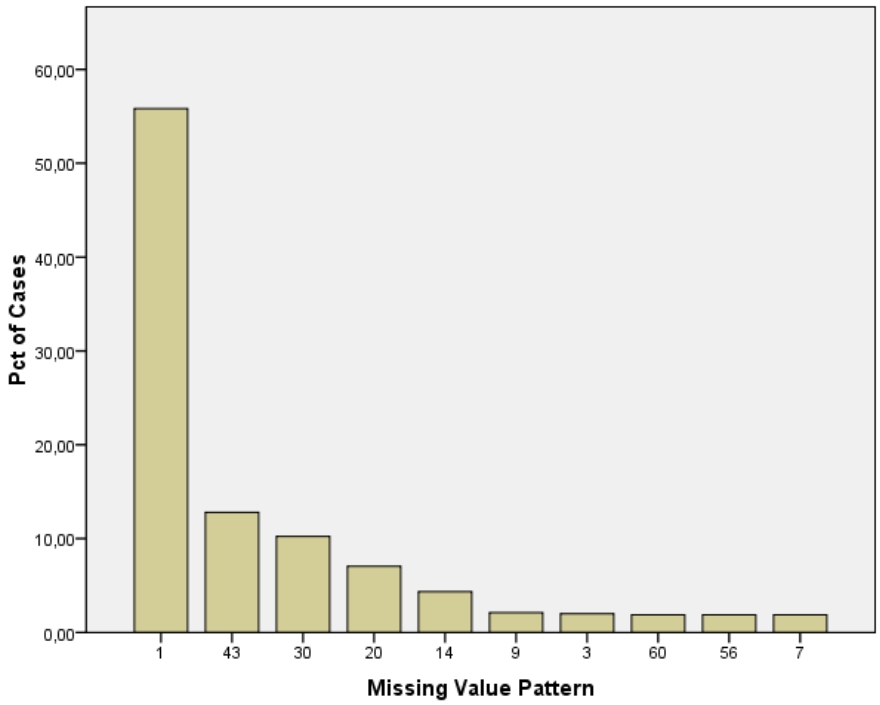


El gráfico de patrones muestra patrones de valores perdidos de las variables de análisis. Cada patrón se corresponde con un grupo de casos con el mismo patrón de datos completos e incompletos. Por ejemplo, el patrón 1 representa casos que no tienen valores perdidos, mientras que el patrón 33 representa los casos que tienen valores perdidos en *residen (Número de personas en el hogar)* y *dirección (Años en la dirección actual)* y el patrón 66 representa los casos que tiene valores perdidos en *sexo (Sexo)*, *marital (Estado civil)*, *dirección* e *ingresos (Ingresos del hogar en miles)*. Un conjunto de datos puede tener 2 patrones de $^{\text{número de variables}}$. Para 10 variables de análisis, es $2^{10} = 1024$; sin embargo, sólo se representan 66 patrones en los 1000 casos del conjunto de datos.

El gráfico ordena las variables y patrones de análisis para revelar las tendencias de monotonía. De forma específica, las variables se ordenan de izquierda a derecha aumentando el orden de valores perdidos. Los patrones se clasifican, en primer lugar, por la última variable (valores no perdidos primero y los valores perdidos después), a continuación por la segunda a la última variable, etcétera, de derecha a izquierda. Revela si el método de imputación monotónica para sus datos o, si no, en qué medida se aproximan sus datos a un patrón monotónico. Si los datos son monótonos, todas las casillas perdidas y no perdidas del gráfico serán contiguas; es decir, no quedarán "islas" de casillas no perdidas en la parte inferior derecha del gráfico y no quedarán "islas" de casillas perdidas en la parte superior izquierda del gráfico.

Este conjunto de datos no es monótono y hay tantos valores que se necesitan imputar para lograr la monotonía.

Figura 5-5
Frecuencias de patrones



Si los patrones se solicitan, un gráfico de barras muestra el porcentaje de casos de cada patrón. Muestra que más de la mitad de los casos del conjunto de datos tienen el patrón 1 y el gráfico de patrones de valores perdidos muestra que es el patrón de los casos sin valores perdidos. El patrón 43 representa casos con un valor perdido de *ingresos*; el patrón 30 representa casos con un valor perdido de *dirección* y el patrón 20 representa casos con un valor perdido de *ecivil*. La gran mayoría de casos, aproximadamente, 4 de 5, se representan en estos cuatro patrones. Los patrones 14, 60 y 56 son los únicos patrones entre los diez patrones más frecuentes para representar casos sin valores perdidos en más de una variable.

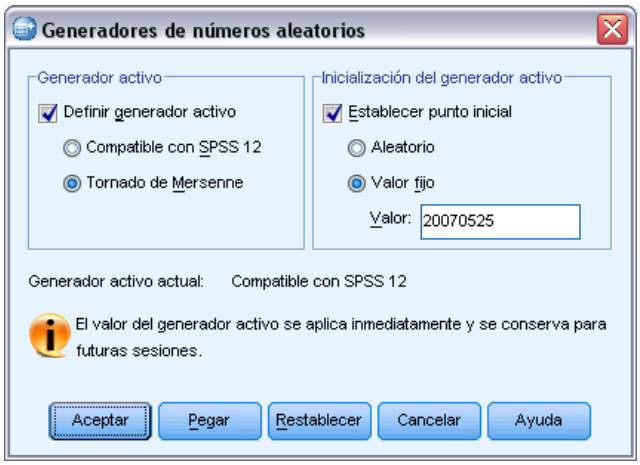
El análisis de patrones perdidos no ha revelado ningún obstáculo concreto en la imputación múltiple, salvo que el uso del método de monotonía no serán viables.

Imputación automática de valores perdidos

Ahora podrá iniciar a imputar valores; comenzaremos con una ejecución con ajustes automáticos, pero después de solicitar imputaciones, solicitaremos la semilla aleatoria. Al establecer la semilla aleatoria, puede replicar el análisis de manera exacta.

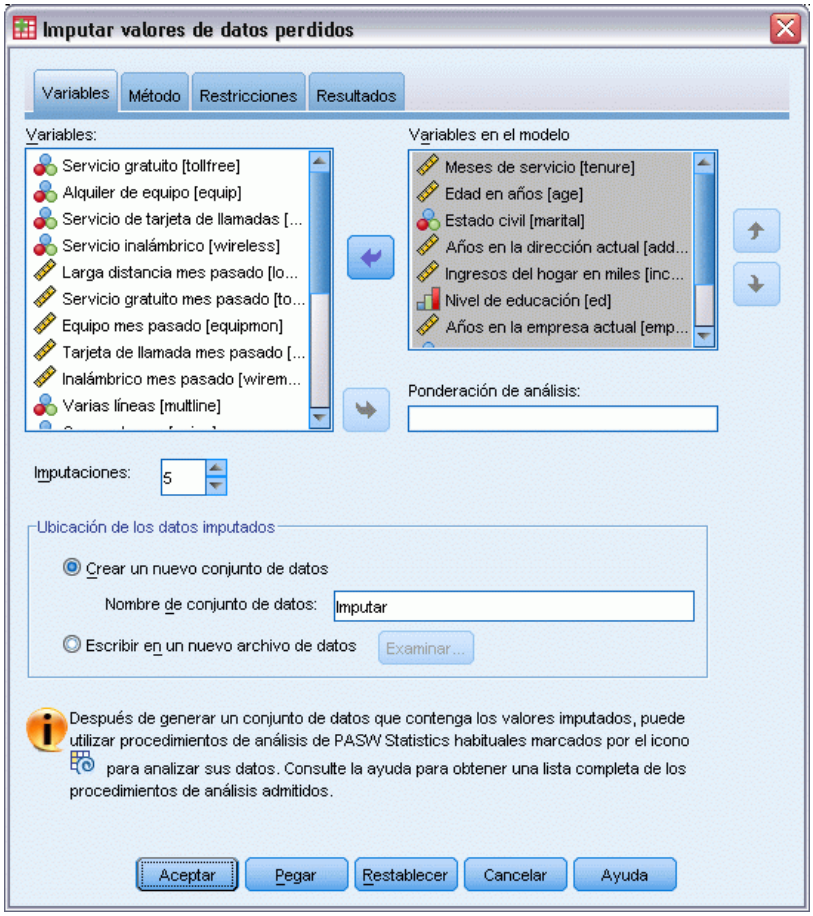
- Para establecer la semilla aleatoria, elija en los menús:
  Transformar > Generadores de números aleatorios...

Figura 5-6
Cuadro de diálogo Generadores de números aleatorios



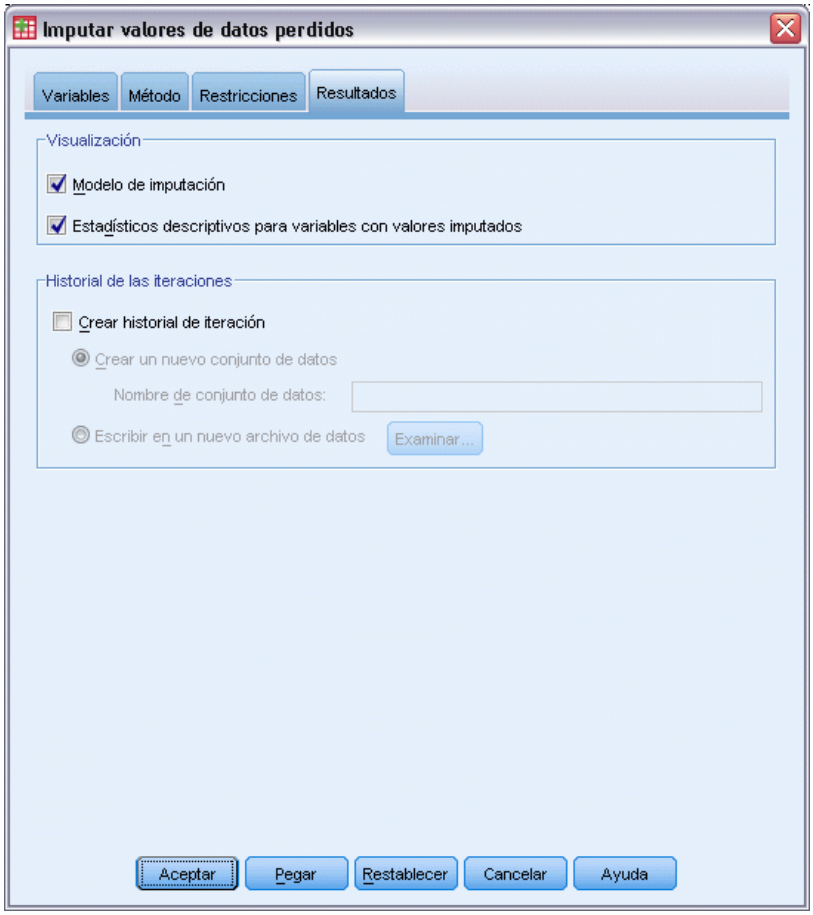
- Seleccione Definir generador activo.
- Seleccione Tornado de Mersenne.
- Seleccione Establecer punto inicial.
- Seleccione Valor fijo y escriba 20070525 como el valor.
- Pulse en Aceptar.
- Para multiplicar los valores de datos perdidos, seleccione en el menú:
  Analizar > Imputación múltiple > Imputar valores perdidos...

Figura 5-7
Cuadro de diálogo Imputar los valores de datos perdidos



- Seleccione desde *Meses de servicio [cargo]* hasta *Número de personas en el hogar [reside]* como variables del modelo de imputación.
- Introduzca telcoImputed como conjunto de datos en el que se guardarán los datos imputados.
- Pulse en la pestaña Resultados.

Figura 5-8
Pestaña Resultados



- Seleccione Estadísticos descriptivos de las variables con valores imputados.
- Pulse en Aceptar.

Especificaciones de imputación

Figura 5-9
Especificaciones de imputación

Método de imputación	Automático
Número de imputaciones	5
Modelo para variables ...	Regresión lineal
Interacciones incluidas...	(ninguna)
Porcentaje máximo de ...	100,0%

La tabla de especificaciones de imputación es una herramienta muy útil de lo que ha solicitado para que pueda confirmar que las especificaciones son correctas.

Resultados de imputación

Figura 5-10
Resultados de imputación

Método de imputación		Especificación condicional completa
Iteraciones de método de especificación totalmente condicional		10
Variables dependientes	Imputado	tenure,age,marital,address,income,ed, employ,retire,gender,reside
	No imputado (demasiados valores perdidos)	
	No imputado (sin valores perdidos)	
Secuencia de imputación		age,tenure,reside,ed,gender,retire,employ, marital,address,income

Los resultados de imputación proporcionan una perspectiva general de lo que ha ocurrido durante el proceso de imputación. En concreto, obsérvese que:

- El método de imputación en la tabla de especificaciones era Automático y el método de selección automática es Especificación totalmente condicional.
- Todas las variables solicitadas se han imputado.
- La secuencia de imputación es el orden en el que las variables aparecen en el eje *x* en el gráfico Patrones de valores perdidos.

Modelo de imputación

Figura 5-11
Modelo de imputación

	Modelo (Fiabilidad)		Valores perdidos	Valores imputados
	Tipo	Efectos		
Edad en años	Regresión lineal	ed,gender,retire,marital,tenure, reside,employ,address,income	25	125
Meses de servicio	Regresión lineal	ed,gender,retire,marital,age,reside, employ,address,income	32	160
Número de personas en el hogar	Regresión lineal	ed,gender,retire,marital,age,tenure, employ,address,income	34	170
Nivel de educación	Regresión logística	gender,retire,marital,age,tenure, reside,employ,address,income	35	175
Sexo	Regresión logística	ed,retire,marital,age,tenure,reside, employ,address,income	42	210
Jubilado	Regresión logística	ed,gender,marital,age,tenure, reside,employ,address,income	84	420
Años en la empresa actual	Regresión lineal	ed,gender,retire,marital,age,tenure, reside,address,income	96	480
Estado civil	Regresión logística	ed,gender,retire,age,tenure,reside, employ,address,income	115	575
Años en la dirección actual	Regresión lineal	ed,gender,retire,marital,age,tenure, reside,employ,income	150	750
Ingresos del hogar en miles	Regresión lineal	ed,gender,retire,marital,age,tenure, reside,employ,address	179	895

La tabla de modelos de imputación proporciona más información acerca de cómo se ha imputado cada variable. En concreto, obsérvese que:

- Las variables se incluyen en el orden de secuencia de imputación.

- Las variables de escala se modelan con una regresión lineal y las variables categóricas con una regresión logística.
- Todos los modelos utilizan el resto de variables como efectos principales.
- Se registra el número de valores perdidos de cada variable, junto con el número total de valores calculados para esa variable (número perdido× número de imputaciones).

Estadísticos descriptivos

Figura 5-12
Estadísticas descriptivas de periodo (Meses de servicio)

Datos	Imputac...	N	Media	Desviación típica	Mínimo	Máximo
Datos originales		968	35,56	21,268	1,00	72,00
Valores imputados	1	32	36,06	24,218	-6,72	90,02
	2	32	37,64	22,229	-,19	88,03
	3	32	30,82	27,245	-40,99	104,77
	4	32	39,97	20,585	1,29	80,50
	5	32	37,87	20,669	3,44	94,21
Datos completos después de la imputación	1	1000	35,58	21,355	-6,72	90,02
	2	1000	35,63	21,291	-,19	88,03
	3	1000	35,41	21,484	-40,99	104,77
	4	1000	35,70	21,251	1,00	80,50
	5	1000	35,63	21,243	1,00	94,21

Las tablas de estadísticas descriptivas muestran resúmenes de las variables con valores imputados. Se crea un modelo diferente para cada variable. Los tipos de estadísticas mostradas dependen de si la variable es de escala o categórica.

Las estadísticas de las variables de escala incluyen el recuento, media, desviación típica, mínima y máxima mostradas para los datos originales, cada conjunto de valores imputados y cada conjunto de datos completo (combinando los datos originales y los valores calculados).

La tabla de estadísticas descriptivas de *periodo (Meses de servicio)* muestra las medias y desviaciones estándar en cada conjunto de valores imputados aproximadamente a los de los datos originales; sin embargo, un problema inmediato se presenta cuando observa el mínimo y ve que los valores negativos de *periodo* se han calculado.

Figura 5-13
Estadísticos descriptivos para ecivil (Estado civil)

Datos	Im...	Ca...	N	Porcentaje
Datos originales		0	456	51,5
		1	429	48,5
Valores imputados	1	0	51	44,3
		1	64	55,7
	2	0	41	35,7
		1	74	64,3
	3	0	49	42,6
		1	66	57,4
	4	0	43	37,4
		1	72	62,6
	5	0	53	46,1
		1	62	53,9
Datos completos después de la imputación	1	0	507	50,7
		1	493	49,3
	2	0	497	49,7
		1	503	50,3
	3	0	505	50,5
		1	495	49,5
	4	0	499	49,9
		1	501	50,1
	5	0	509	50,9
		1	491	49,1

En las variables categóricas, las estadísticas incluyen el recuento y porcentaje por categoría de los datos originales, los valores imputados y todos los datos. La tabla de *ecivil (Estado civil)* muestra un resultado interesante, ya que, para los valores imputados, se calcula que se considera una mayor parte de los casos como casados que en los datos originales. Puede deberse a una variación aleatoria; alternativamente, las posibilidades de perderse pueden deberse al valor de esta variable.

Figura 5-14
Estadísticas descriptivas de ingresos (Ingresos del hogar en miles)

Datos	Imputac...	N	Media	Desviación típica	Mínimo	Máximo
Datos originales		821	71,1462	83,14424	9,0000	944,0000
Valores imputados	1	179	87,6574	91,13179	-189,1959	373,2412
	2	179	101,6724	94,20599	-122,0010	346,4294
	3	179	100,9445	95,00789	-127,8572	342,5208
	4	179	107,0787	90,23638	-113,0959	369,9674
	5	179	101,1043	90,40865	-167,6978	314,2533
Datos completos después de la imputación	1	1000	74,1017	84,81851	-189,1959	944,0000
	2	1000	76,6104	85,98067	-122,0010	944,0000
	3	1000	76,4801	86,10024	-127,8572	944,0000
	4	1000	77,5781	85,52821	-113,0959	944,0000
	5	1000	76,5087	85,22154	-167,6978	944,0000

Al igual que *periodo* y el resto de variables de escala, *ingresos (Ingresos del hogar en miles)* muestra los valores negativos imputados — claramente, necesitaremos ejecutar un modelo personalizado con limitaciones en determinadas variables. Sin embargo, *ingresos* muestra otros problemas potenciales. Los valores de media de cada imputación son considerablemente más elevados que para los datos originales y los valores máximos de cada imputación son

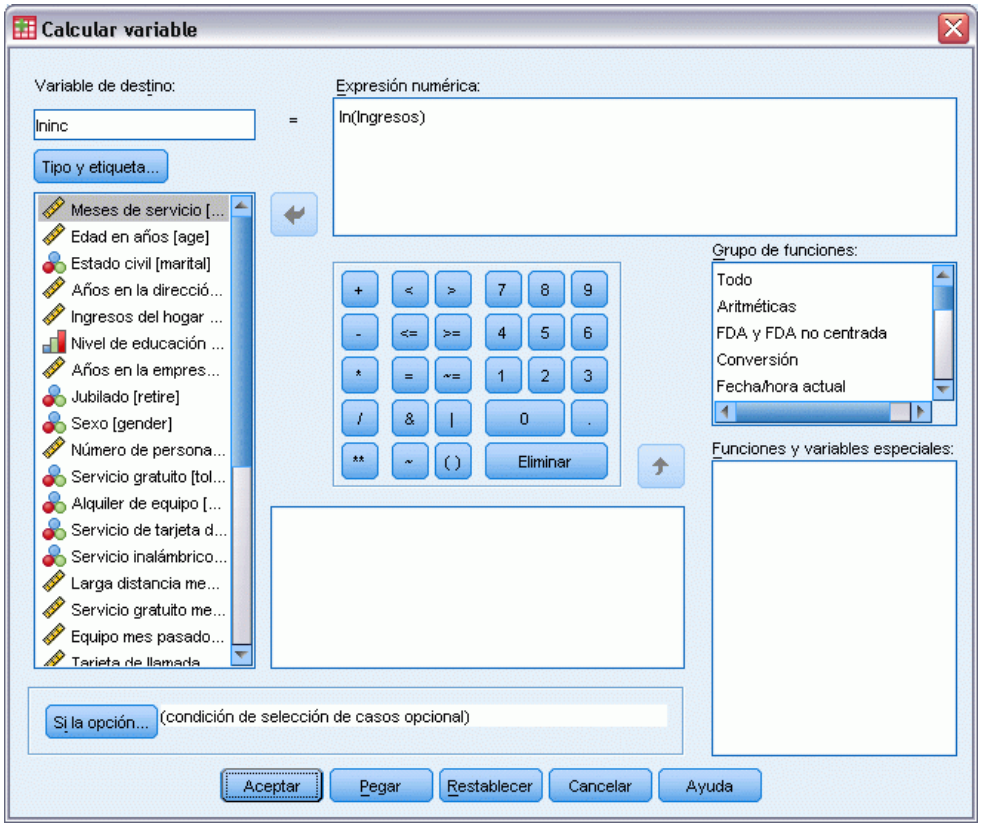
considerablemente inferiores que para los datos originales. La distribución de los ingresos tiene una clara tendencia a la derecha, por lo que puede ser el origen del problema.

Modelo de imputación personalizada

Para evitar que los valores imputados queden fuera del intervalo razonable de los valores de cada variable, especificaremos un modelo de imputación personalizada con limitaciones en las variables. Además, *Ingresos del hogar en miles* tiene una clara tendencia hacia la derecha y otros análisis requerirán el uso del logaritmo de *ingresos*, por lo que parece posible imputar el logaritmo de ingresos directamente.

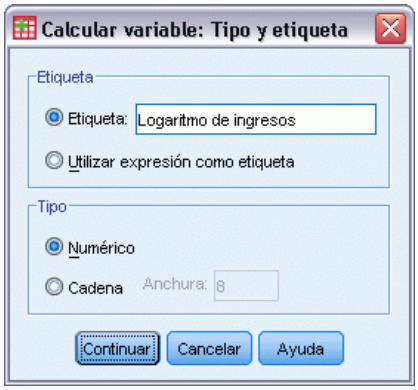
- Asegúrese de que el conjunto de datos original está activo.
- Para crear una variable de logaritmos de ingresos, seleccione en los menús:
  Transformar > Calcular variable...

Figura 5-15
Cuadro de diálogo Calcular variable



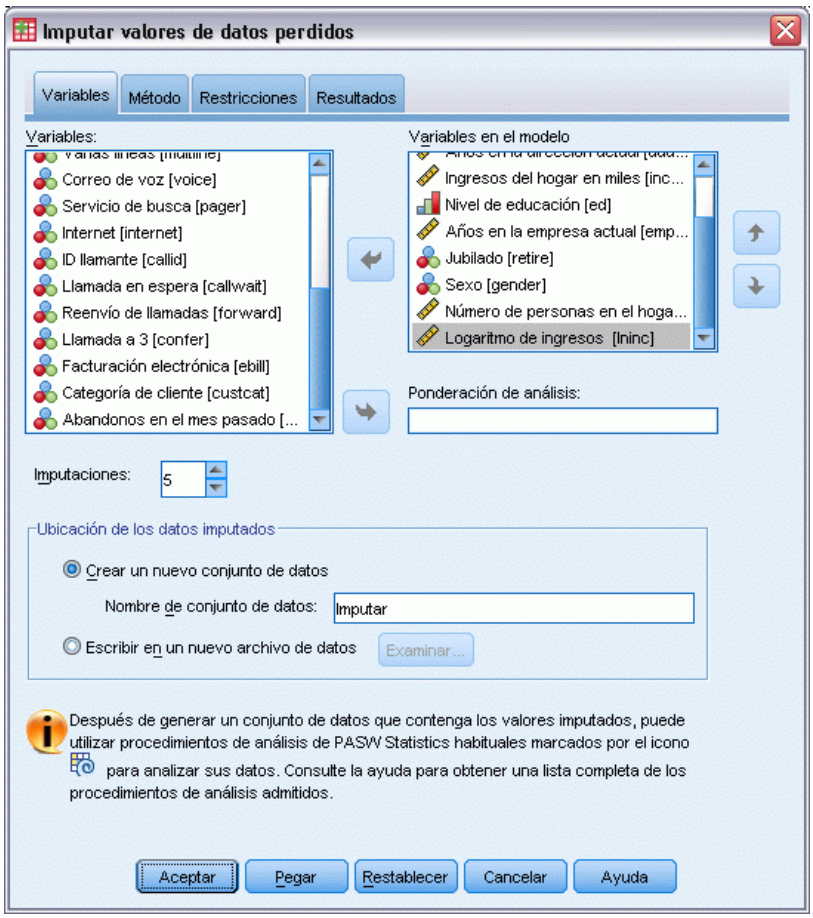
- Introduzca *lninc* como variable de destino.
- Introduzca ln(Ingresos) como expresión numérica.
- Pulse en Tipo & Etiqueta..

Figura 5-16
Cuadro de diálogo Tipo y etiqueta



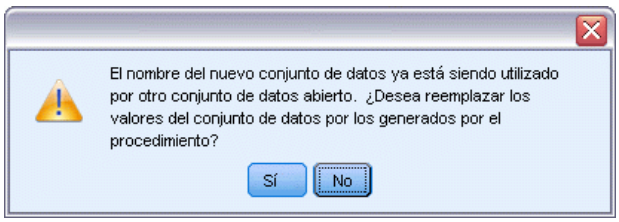
- Introduzca *Logaritmo de ingresos* como etiqueta.
- Pulse en Continuar.
- Pulse Aceptar en el cuadro de diálogo Calcular variable.

Figura 5-17
Pestaña Variables con Logaritmo de ingresos sustituyendo Ingresos del hogar en miles en el modelo de imputación



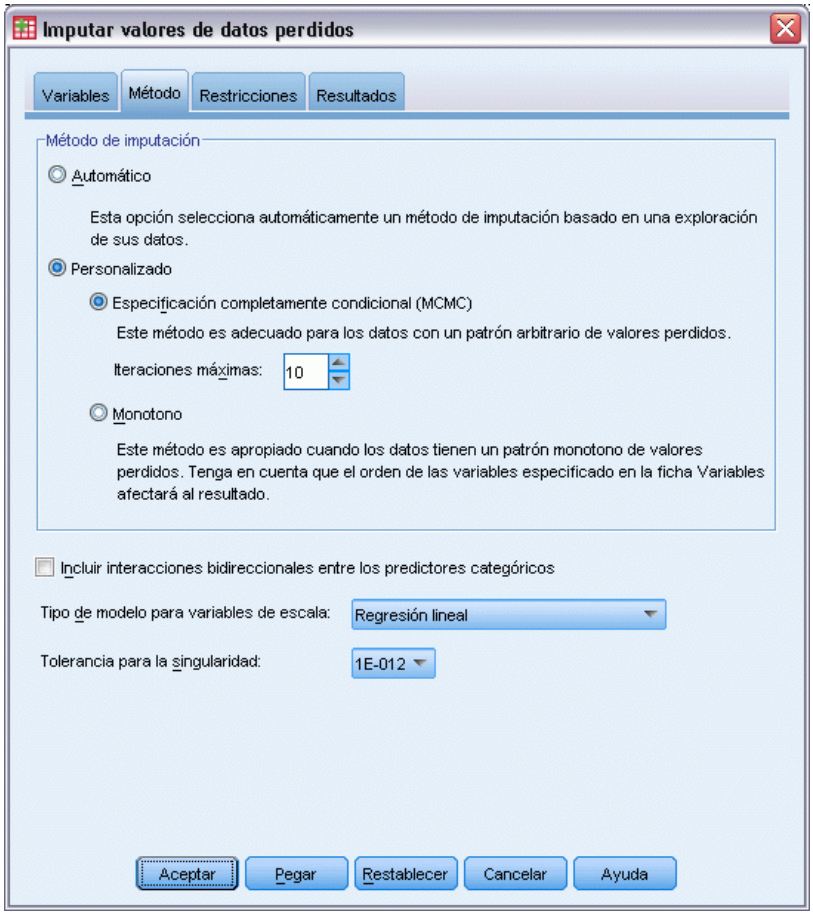
- Abra el cuadro de diálogo Imputar valores perdidos y pulse la pestaña Variables.
- Cancele la selección de *Ingresos del hogar en miles [ingres]* y seleccione *Logaritmo de ingresos [lninc]* como variable del modelo.
- Pulse en la pestaña Método.

Figura 5-18
Alerta para sustituir el conjunto de datos existente



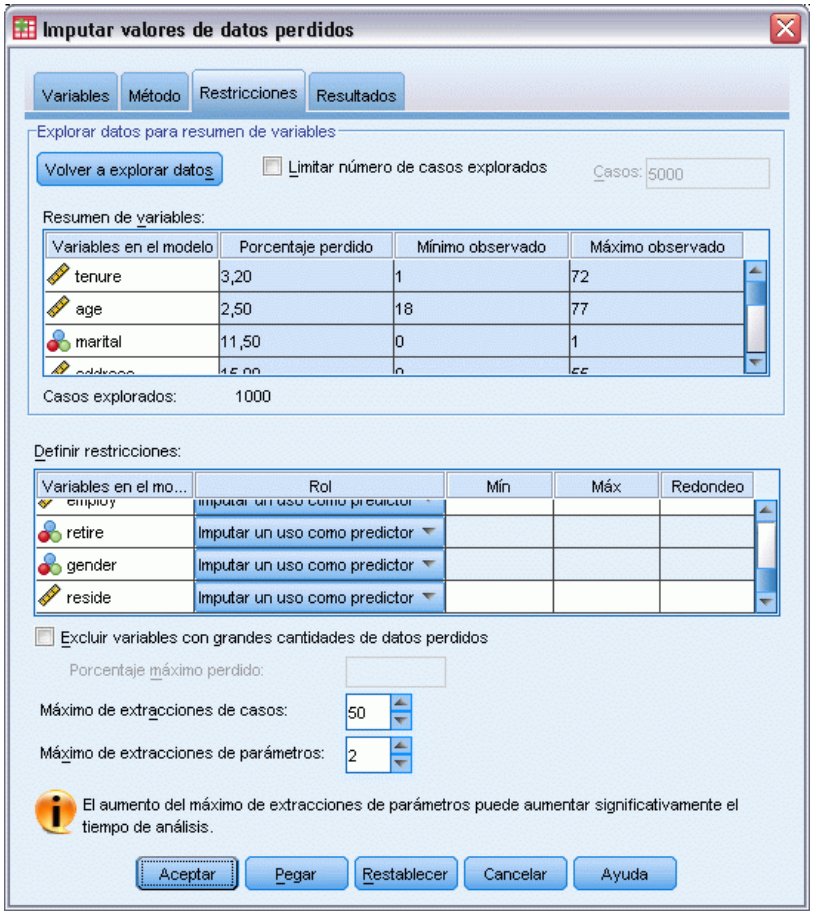
- Pulse en Sí en la alerta que aparece.

Figura 5-19
Pestaña Método



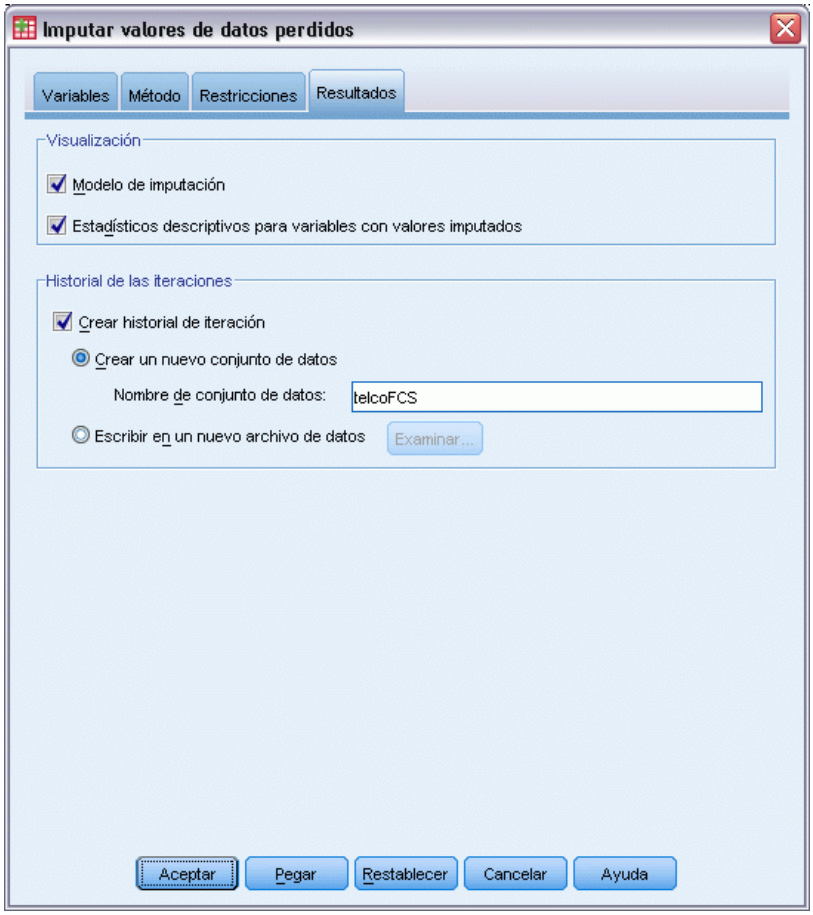
- Seleccione Personalizado y deje Especificación totalmente condicional seleccionada como método de imputación.
- Pulse en la pestaña Restricciones.

Figura 5-20
Pestaña Restricciones



- Pulse en Explorar datos.
- En la casilla Defina las restricciones, introduzca 1 como el valor mínimo de *Meses de servicio [periodo]*.
- Introduzca 18 como el valor mínimo de *edad (Edad en años)*.
- Introduzca 0 como el valor mínimo de *edad (Edad en años)*.
- Introduzca 0 como el valor mínimo de *empleo (Años con empresa actual)*.
- Introduzca 1 como el valor mínimo y 1 como el nivel de redondeo para *residencia (Número de miembros en la familia)*. Tenga en cuenta que muchas del resto de las variables escala se incluyen en los valores enteros, es posible plantear que una persona ha vivido durante 13,8 años en su dirección actual, pero no cabe pensar que 2,2 personas viven allí.
- Introduzca 0 como el valor mínimo de *lninc (Logaritmo de ingresos)*.
- Pulse en la pestaña Resultados.

Figura 5-21
Pestaña Resultados



- Seleccione Crear historial de iteraciones e introduzca telcoFCS como el nombre del nuevo conjunto de datos.
- Pulse en Aceptar.

Restricciones de imputación

Figura 5-22
Restricciones de imputación

	Papel en imputación		Valores imputados		
	Dependiente (Regresión logística)	Predictor	Mínimo	Máximo	Redondeo
Meses de servicio	Sí	Sí	1	(ninguna)	
Edad en años	Sí	Sí	18	(ninguna)	
Estado civil	Sí	Sí			
Años en la dirección ...	Sí	Sí	0	(ninguna)	
Nivel de educación	Sí	Sí			
Años en la empresa ...	Sí	Sí	0	(ninguna)	
Jubilado	Sí	Sí			
Sexo	Sí	Sí			
Número de personas e...	Sí	Sí	1	(ninguna)	Entero
Ingresos del hogar en ...	Sí	Sí	0	(ninguna)	

El modelo de imputación personalizado da como resultado una nueva tabla que revisa las limitaciones aplicadas en el modelo de imputación. Todo parece estar de acuerdo con sus especificaciones.

Estadísticos descriptivos

Figura 5-23
Estadísticas descriptivas de periodo (Meses de servicio)

Datos	Imputac...	N	Media	Desviación típica	Mínimo	Máximo
Datos originales		968	35,56	21,268	1,00	72,00
Valores imputados	1	32	38,73	20,920	7,06	100,30
	2	32	32,81	19,632	3,52	70,63
	3	32	36,53	22,083	5,11	88,25
	4	32	39,50	19,840	9,47	78,51
	5	32	40,66	19,112	1,57	75,08
Datos completos después de la imputación	1	1000	35,66	21,254	1,00	100,30
	2	1000	35,47	21,214	1,00	72,00
	3	1000	35,59	21,284	1,00	88,25
	4	1000	35,69	21,226	1,00	78,51
	5	1000	35,72	21,213	1,00	75,08

La tabla estadísticas descriptivas de *periodo (Meses de servicio)* en el modelo de aplicación personalizado con limitaciones que muestran que el problema de los valores negativos imputados de *periodo* se ha resuelto.

Figura 5-24
Estadísticos descriptivos para ecivil (Estado civil)

Datos	Imputación	Categoría	N	Porcentaje
Datos originales		0	456	51,5
		1	429	48,5
Valores imputados	1	0	48	41,7
		1	67	58,3
	2	0	46	40,0
		1	69	60,0
	3	0	51	44,3
		1	64	55,7
	4	0	53	46,1
		1	62	53,9
	5	0	49	42,6
		1	66	57,4
Datos completos después de la imputación	1	0	504	50,4
		1	496	49,6
	2	0	502	50,2
		1	498	49,8
	3	0	507	50,7
		1	493	49,3
	4	0	509	50,9
		1	491	49,1
	5	0	505	50,5
		1	495	49,5

La tabla de *ecivil (Estado civil)* tiene una imputación (3) cuya distribución está más en la línea de los datos originales, pero la mayoría sigue mostrando una mayor proporción de los casos estimados como casados que en los datos originales. Puede deberse a una variación aleatoria, pero puede requerir un estudio con mayor profundidad de los datos para determinar su estos valores no faltan de forma aleatoria (MAR). Este estudio no se trata aquí.

Figura 5-25
Estadísticos descriptivos de lninc (Logaritmo de ingresos)

Datos	Imputación	N	Mean	Std. Deviation	Minimum	Maximum
Datos originales		821	3,9291	,75305	2,1972	6,8501
Valores imputados	1	179	4,1816	,94574	1,4428	6,6748
	2	179	4,2562	,98346	1,6633	6,8224
	3	179	4,1743	1,01487	1,4443	6,8437
	4	179	4,1774	,82705	2,2532	6,2680
	5	179	4,1894	,96403	1,6667	6,6677
Datos completos después de la imputación	1	1000	3,9743	,79638	1,4428	6,8501
	2	1000	3,9876	,80842	1,6633	6,8501
	3	1000	3,9730	,81107	1,4443	6,8501
	4	1000	3,9735	,77228	2,1972	6,8501
	5	1000	3,9756	,80064	1,6667	6,8501

Como *periodo* y el resto de variables de escala, *lninc (Logaritmo de ingresos)* no muestra los valores negativos calculados. Además, los valores de las medias de las imputaciones están más próximos a la media para los datos originales que en la ejecución de imputación automática — en la escala de *ingresos*, la media de los datos originales de *lninc* es aproximadamente $e^{3,9291}$ = 50,86, mientras el valor de la media típica entre las imputaciones es muy aproximada $e^{4,2}$=

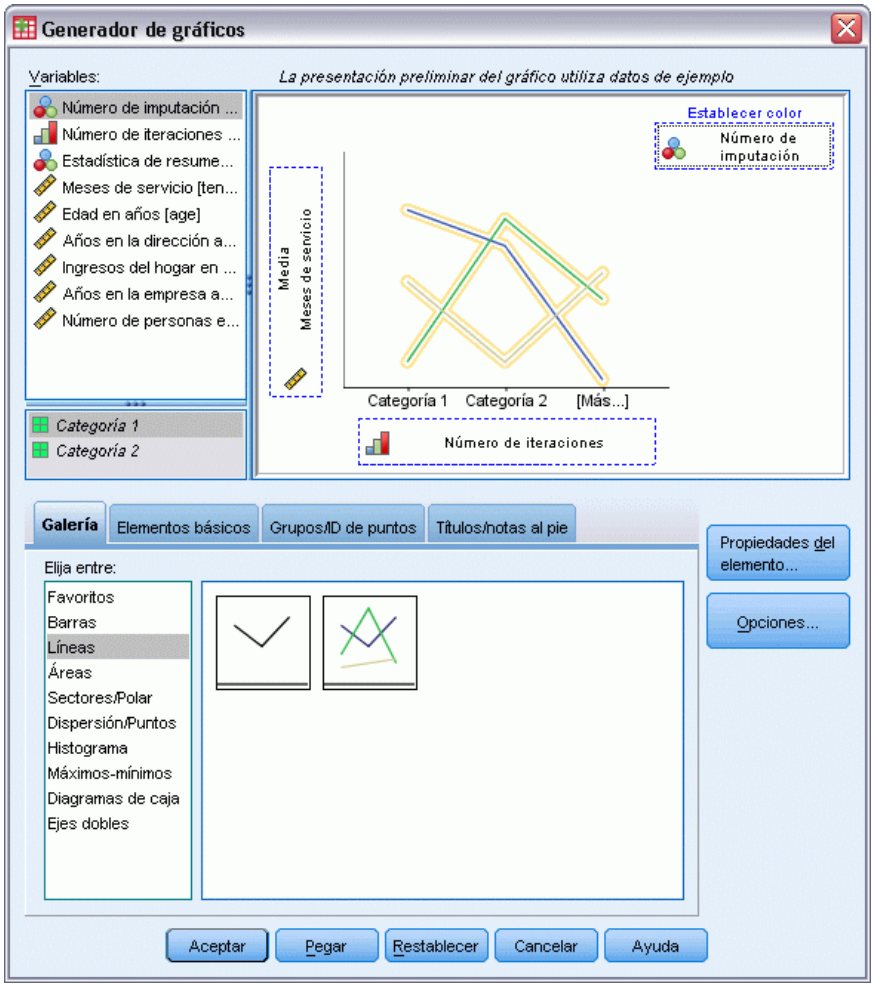
66,69. Además, los valores máximos de cada imputación están más cercanos al valor máximo de los datos originales.

Comprobación de la convergencia de FCS

Cuando se utiliza en método de especificación condicional, es una buena idea comprobar los gráficos de las medias y desviaciones típicas por iteraciones y el cálculo de cada variable de escala dependiente cuyos valores se calculan para ayudar a evaluar la convergencia del modelo.

- Para crear este tipo de gráfico, active el conjunto de datos *telcoFCS* y en el menú seleccione: Gráficos > Generador de gráficos...

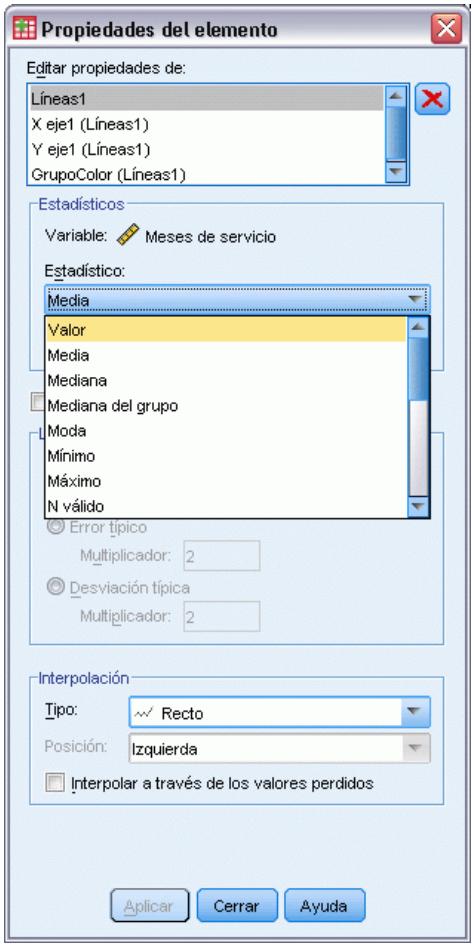
Figura 5-26
Generador de gráficos, gráficos de líneas múltiples



- Seleccione la galería Línea y seleccione Líneas múltiples.
- Seleccione *Meses de servicio [periodo]* como la variable que se trazará en el eje *Y*.

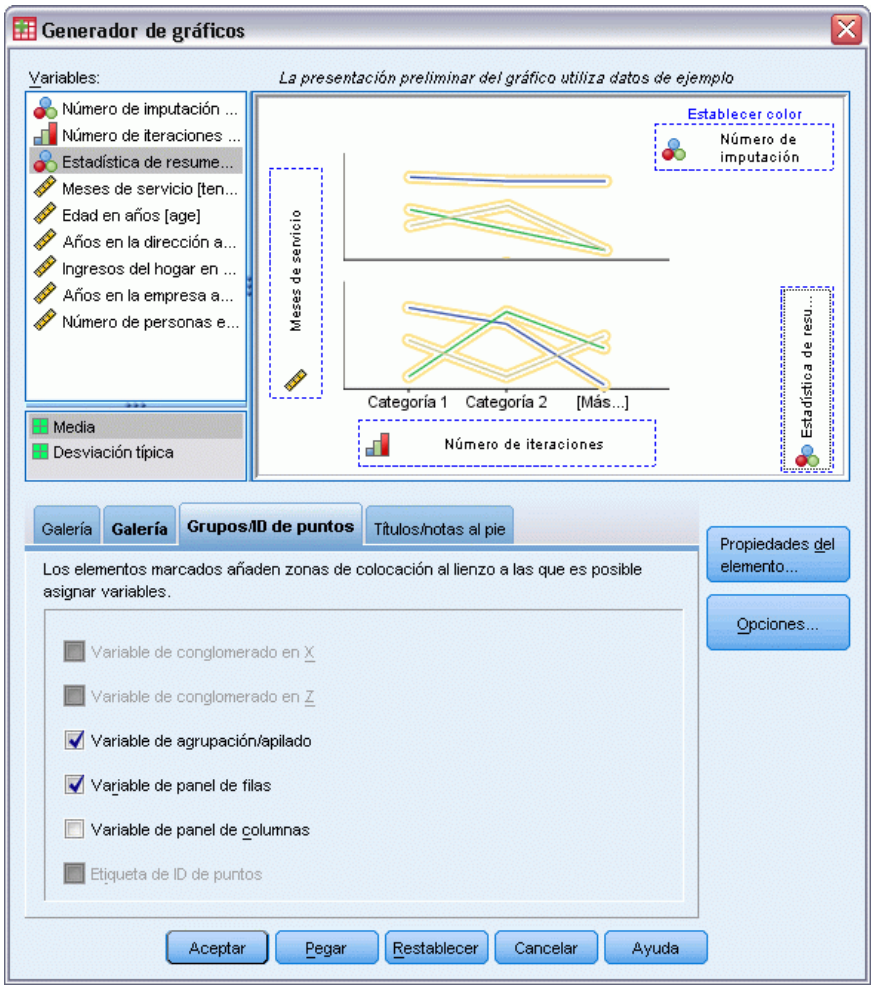
- Seleccione *Número de iteración [Iteration_]* como la variable que se representará en el eje *X*.
- Seleccione *Número de imputación [Imputations_]* como la variable para definir los colores.

Figura 5-27
Generador de gráficos, Propiedades del elemento



- En Propiedades del elemento, seleccione Valor como la estadística que se mostrará.
- Pulse en Aplicar.
- En el Generador de gráficos, seleccione la pestaña Grupos/ID de puntos.

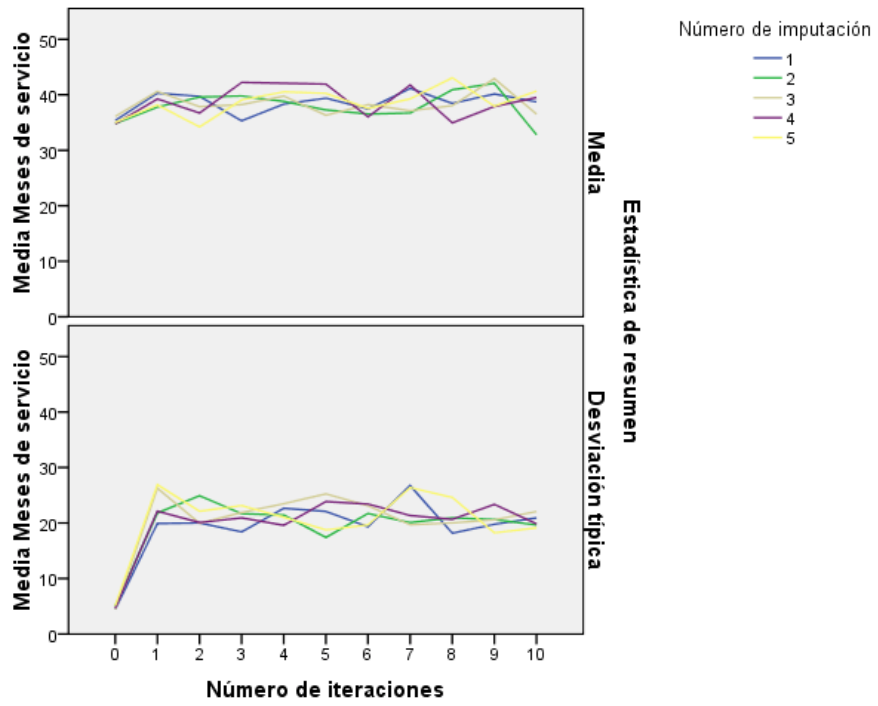
Figura 5-28
Generador de gráficos, pestaña Grupos/ID de puntos



- Seleccione Variable de panel de filas.
- Seleccione *Estadístico de resumen [SummaryStatistic_]* como variable de panel.
- Pulse en Aceptar.

Gráficos de convergencia FCS

Figura 5-29
gráfico de convergencia FCS



Ha creado un par de gráficos de líneas múltiples, que muestran la media y desviación típica de los valores imputados de *Meses de servicio [periodo]* en cada iteración del método de imputación de FCS para cada una de las 5 imputaciones solicitadas. El objeto de este gráfico es observar los patrones de las líneas. No debe haber ningún patrón y las líneas deben parecer “aleatorias”. Puede crear gráficos similares para el resto de variables de escala y tenga en cuenta que estos gráficos tampoco muestran patrones perceptibles.

Analizar datos completos

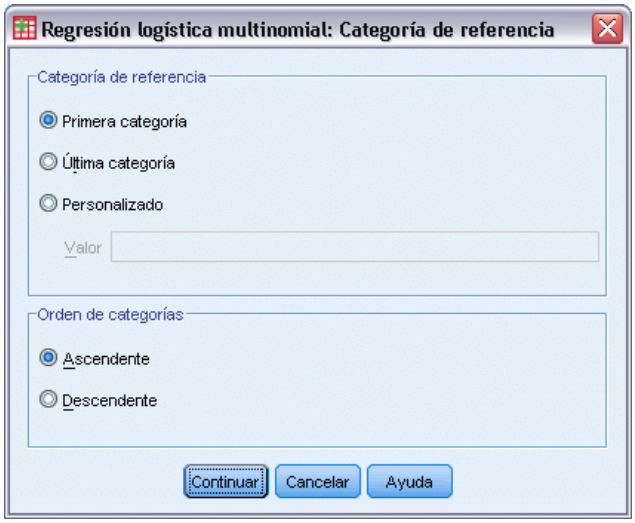
Una vez que los valores imputados parecen ser satisfactorios, puede ejecutar un análisis de los datos completos. El conjunto de datos contiene una variable *Categoría del cliente [custcat]* que divide la base de clientes por patrones de uso de servicio, categorizando los clientes en cuatro grupos. Si puede ajustar un modelo utilizando información demográfica para predecir la pertenencia a un grupo, se pueden personalizar las ofertas para cada uno de los posibles clientes.

- Active el conjunto de datos *telcoImputed* . Para crear un modelo de regresión logística multinomial para los datos completos, selección en el menú:
  Analizar > Regresión > Logística multinomial...

Figura 5-30
Cuadro de diálogo Regresión logística multinomial

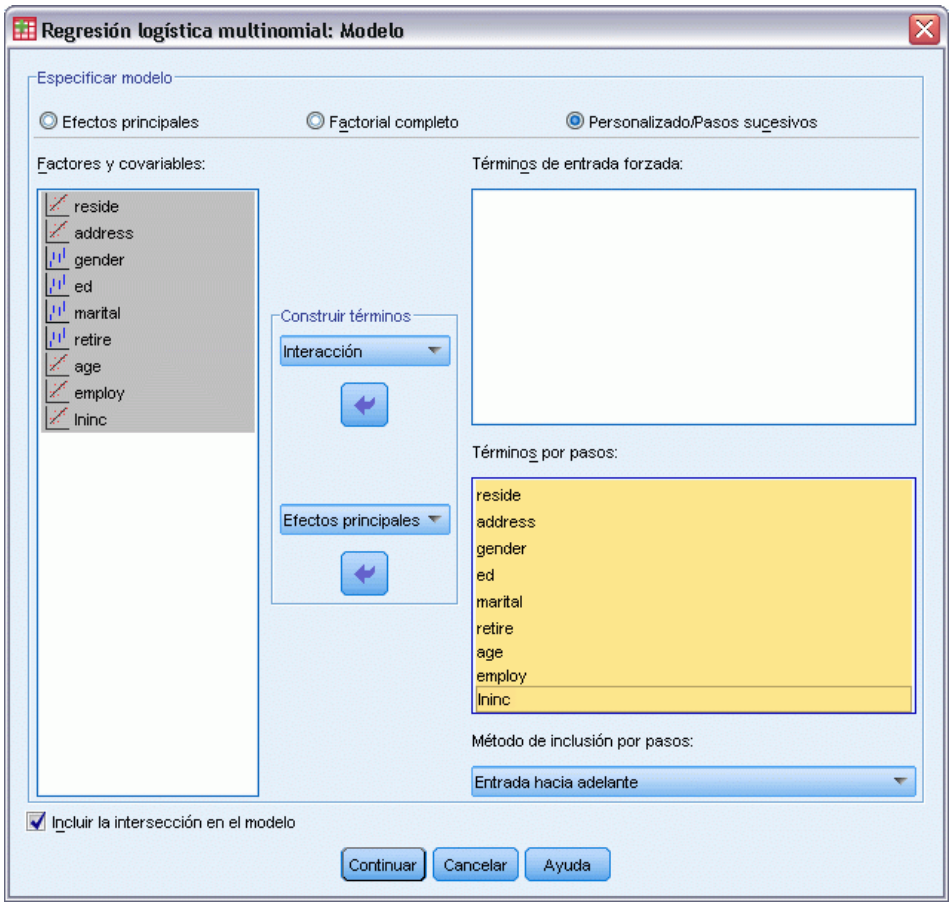
- Seleccione *Categoría de cliente* como la variable dependiente.
- Seleccione *Estado civil*, *Nivel educativo*, *Retirado* y *Sexo* como factores.
- Seleccione *Edad en años*, *Años en la dirección actual*, *Años con empresa actual*, *Número de miembros en la familia* y *Logaritmo de ingresos* como covariables.
- Tal vez quiera comparar otros clientes con los que se han suscrito al servicio básico, para lo que debe seleccionar *Categoría de cliente* y Categoría de referencia.

Figura 5-31
Cuadro de diálogo Categoría de referencia



- Seleccione Primera categoría.
- Pulse en Continuar.
- En el cuadro de diálogo Regresión logística multinomial, pulse en Modelo.

Figura 5-32
Cuadro de diálogo Modelo



- Seleccione Personalizado/Pasos sucesivos.
- Seleccione Efectos principales en la lista desplegable de construcción de términos de los términos de pasos sucesivos.
- Seleccione desde *lninc* hasta *residen* como términos de pasos sucesivos.
- Pulse en Continuar.
- En el cuadro de diálogo Regresión logística multinomial, pulse en Aceptar.

Resumen de los pasos

Figura 5-33
Resumen de pasos

Número de imputación	Modelo	Acción	Efecto(s)	Criterio de ajuste del modelo	Contrastes de selección de efectos		
				-2 log verosimilitud	Chi-cuadrado[b]	gl	Sig.
Original data	0	Introducido	Intersección	1353,555	.		
	1	Introducido	ed	1260,972	92,583	12	,000
	2	Introducido	employ	1237,664	23,308	3	,000
	3	Introducido	marital	1229,808	7,856	3	,049
1	0	Introducido	Intersección	2762,531	.		
	1	Introducido	ed	2608,189	154,342	12	,000
	2	Introducido	employ	2563,671	44,518	3	,000
	3	Introducido	marital	2549,200	14,470	3	,002
	4	Introducido	address	2541,050	8,151	3	,043
2	0	Introducido	Intersección	2762,531	.		
	1	Introducido	ed	2603,940	158,591	12	,000
	2	Introducido	employ	2563,367	40,573	3	,000
	3	Introducido	marital	2545,743	17,624	3	,001
	4	Introducido	address	2536,532	9,211	3	,027
3	0	Introducido	Intersección	2762,531	.		
	1	Introducido	ed	2600,074	162,457	12	,000
	2	Introducido	employ	2558,560	41,514	3	,000
	3	Introducido	marital	2546,062	12,499	3	,006
	4	Introducido	address	2536,348	9,714	3	,021
4	0	Introducido	Intersección	2762,531	.		
	1	Introducido	ed	2601,616	160,915	12	,000
	2	Introducido	employ	2558,463	43,153	3	,000
	3	Introducido	marital	2543,747	14,716	3	,002
	4	Introducido	address	2533,341	10,406	3	,015
5	0	Introducido	Intersección	2762,531	.		
	1	Introducido	ed	2604,773	157,759	12	,000
	2	Introducido	employ	2561,792	42,980	3	,000
	3	Introducido	marital	2549,096	12,696	3	,005

Método por pasos: Entrada hacia adelante

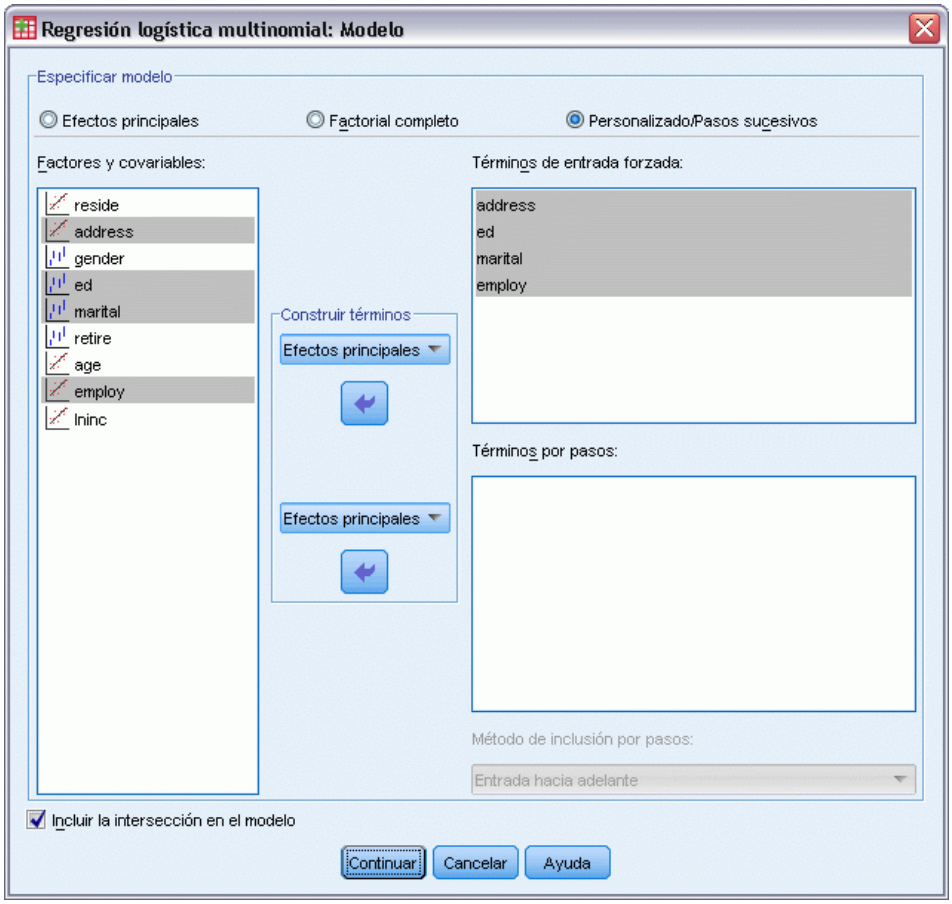
a. El valor de chi-cuadrado para su inclusión se basa en la prueba de la razón de verosimilitudes.

La Regresión logística multinomial admite la combinación de coeficientes de regresión; sin embargo, observará que *todas* las tablas muestran los resultados para cada imputación y los datos originales. Se debe a que el archivo está dividido en *Imputation_*, para que todas las tablas que tienen en cuenta la variable de división presentarán los grupos de archivos en una única tabla.

También verá que la tabla Estimaciones de los parámetros no muestra las estimaciones combinadas; para saber las razones, consulte el Resumen de los pasos. Hemos solicitado la selección de por pasos de los efectos del modelo y el mismo conjunto de efectos no se ha seleccionado para todas las imputaciones, por lo que no se puede realizar la combinación. Sin embargo. proporciona información de gran utilidad porque vemos que *educ (Nivel educativo)*, *empleo (Años con empresa actual)*, *ecivil (Estado civil)* y *dirección (Años en la dirección actual)* se suelen seleccionar por la selección por pasos entre las imputaciones. Ajustaremos otro modelo utilizando estos predictores.

Ejecución del modelo con un subconjunto de predictores

Figura 5-34
Cuadro de diálogo Modelo

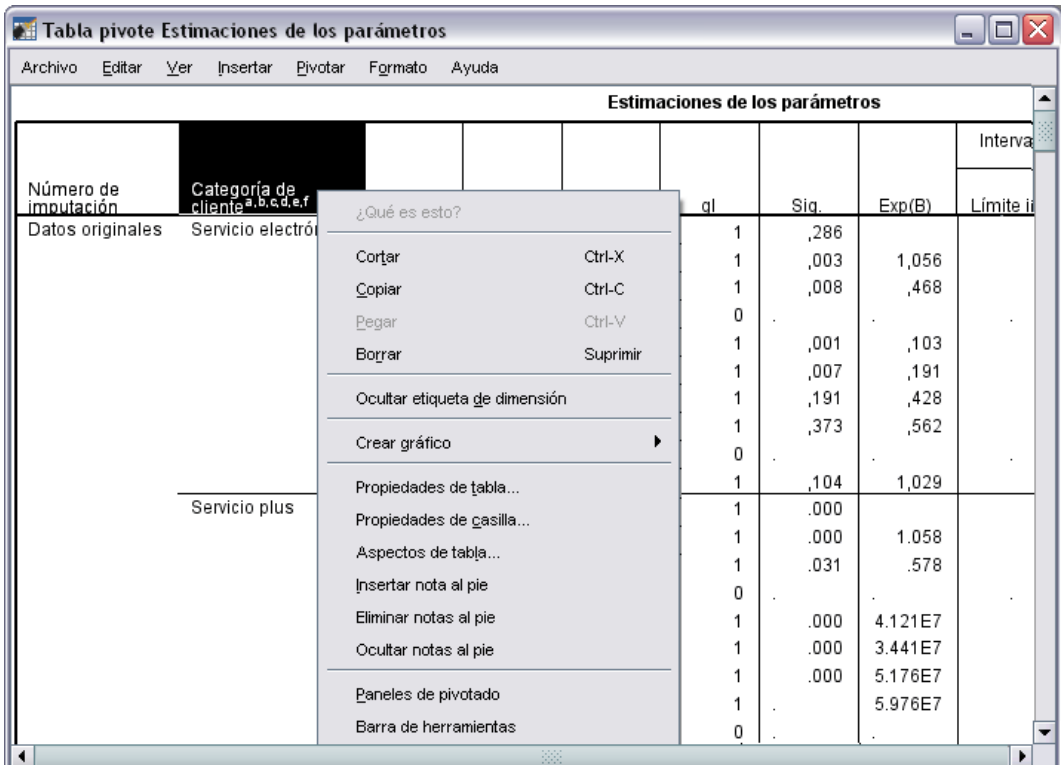


- Abra el cuadro de diálogo Regresión logística multinomial y pulse en Modelo.
- Cancele la selección de las variables de la lista Términos por pasos.
- Seleccione Efectos principales en la lista desplegable de construcción de términos de los términos de entrada forzada.
- Seleccione *empleo*, *ecivil*, *educ* y *dirección* como Términos de entrada forzada.
- Pulse en Continuar.
- En el cuadro de diálogo Regresión logística multinomial, pulse en Aceptar.

Estimaciones combinadas de parámetros

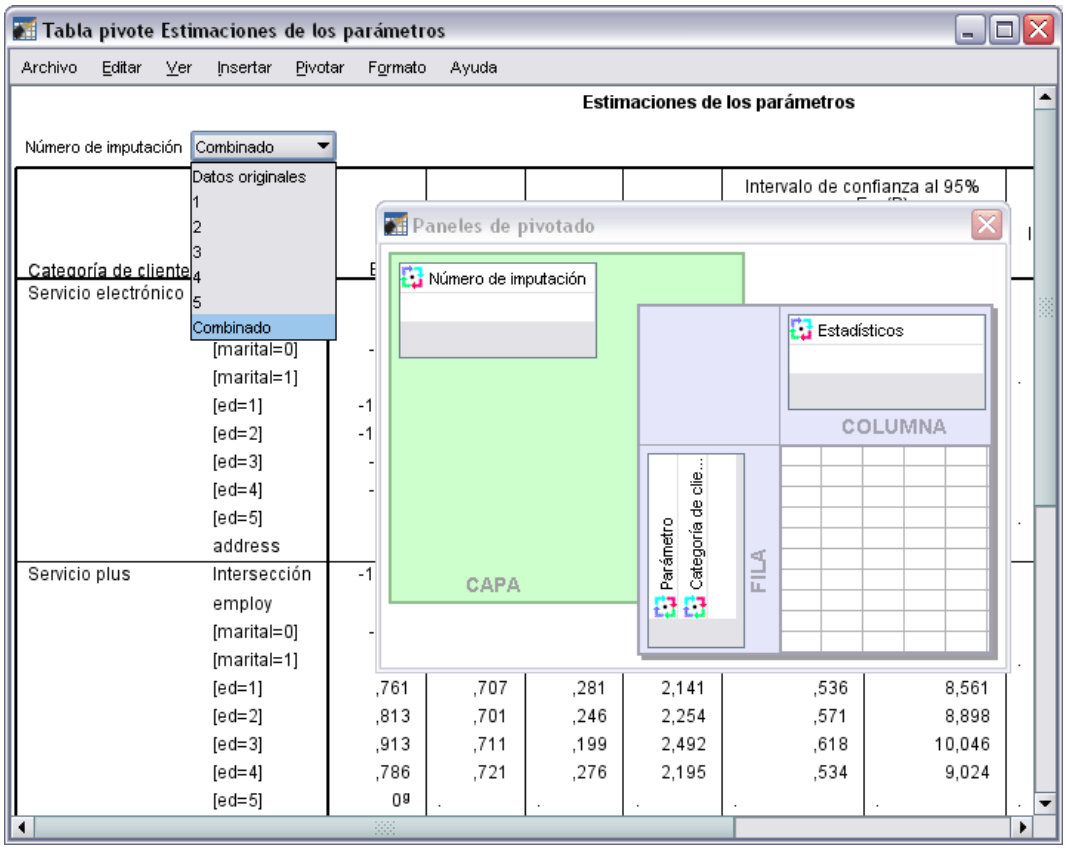
Esta tabla es muy grande, pero la pivotación nos proporcionará un par de vistas diferentes de gran utilidad del resultado.

Figura 5-35
Estimaciones combinadas de parámetros



► Active (pulse dos veces) la tabla y seleccione Paneles de pivotado en el menú contextual.

Figura 5-36
Estimaciones combinadas de parámetros



- Cambie *Número de imputación* de Fila a Capa.
- Seleccione Combinado en la lista desplegable Número de imputación.

Figura 5-37
Estimaciones combinadas de parámetros

Número de imputación=Combinado

Categoría de cliente[a,b,c,d,e,f]		B	Error típ.	Sig.	Exp (B)	Intervalo de confianza al 95% para Exp(B)		Información perdida de fracción.	Varianza de incremento relativa	Eficacia relativa
						Límite inferior	Límite superior			
Servicio electrónico	Intersección	,461	,464	,321				,007	,007	,999
	employ	,036	,015	,015	1,036	1,007	1,066	,122	,132	,976
	[marital=0]	-,691	,210	,001	,501	,332	,756	,008	,008	,998
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	-1,935	,518	,000	,144	,052	,399	,028	,029	,994
	[ed=2]	-1,147	,485	,018	,318	,123	,821	,030	,030	,994
	[ed=3]	-,532	,488	,276	,587	,226	1,529	,003	,003	,999
	[ed=4]	-,446	,488	,361	,640	,246	1,666	,017	,017	,997
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	,030	,016	,061	1,031	,998	1,064	,338	,444	,937
Servicio plus	Intersección	-1,236	,695	,075				,010	,010	,998
	employ	,050	,013	,000	1,052	1,026	1,078	,051	,053	,990
	[marital=0]	-,411	,200	,040	,663	,448	,982	,034	,035	,993
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	,761	,707	,281	2,141	,536	8,561	,014	,014	,997
	[ed=2]	,813	,701	,246	2,254	,571	8,898	,010	,010	,998
	[ed=3]	,913	,711	,199	2,492	,618	10,046	,001	,001	1,000
	[ed=4]	,786	,721	,276	2,195	,534	9,024	,022	,022	,996
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	,015	,013	,240	1,015	,990	1,041	,088	,093	,983
Servicio total	Intersección	1,117	,444	,012				,044	,045	,991
	employ	,045	,014	,002	1,046	1,017	1,076	,006	,006	,999
	[marital=0]	-,664	,223	,003	,515	,332	,798	,110	,118	,978
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	-3,381	,623	,000	,034	,010	,117	,219	,255	,958
	[ed=2]	-1,779	,457	,000	,169	,069	,414	,018	,018	,996
	[ed=3]	-1,229	,467	,009	,293	,117	,731	,014	,014	,997
	[ed=4]	-,512	,454	,260	,599	,246	1,460	,029	,029	,994
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	,007	,015	,634	1,007	,978	1,037	,141	,154	,973

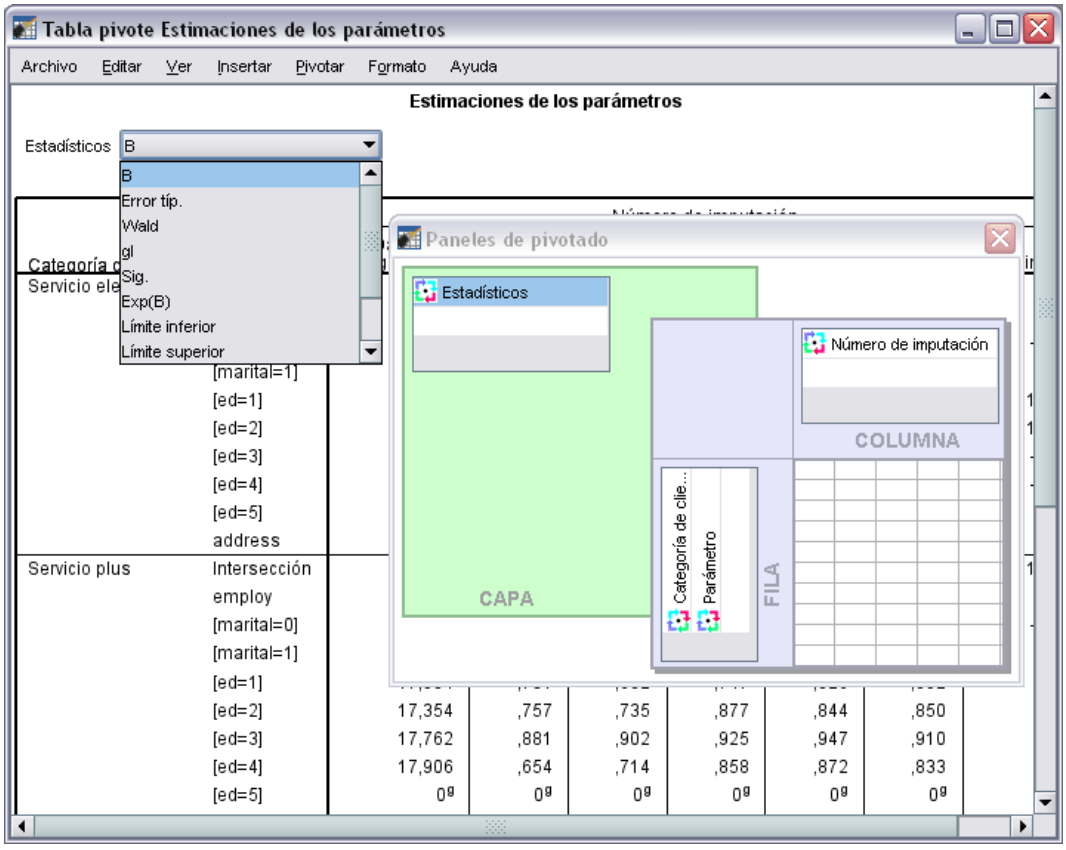
Esta vista muestra todas las estadísticas de los resultados combinados. Puede utilizar e interpretar estos coeficientes de la misma manera que utilizaría esta tabla para un conjunto de datos sin valores perdidos.

La tabla de estimaciones de los parámetros resume el efecto de cada predictor. La razón del coeficiente respecto a su error típico, al cuadrado, equivale al estadístico de Wald. Si el nivel de significación del estadístico de Wald es pequeño (inferior a 0,05) el parámetro es diferente de 0.

- Los parámetros con coeficientes negativos significativos disminuyen la probabilidad de dicha categoría de respuesta respecto a la categoría de referencia.
- Los parámetros con coeficientes positivos aumentan la probabilidad de dicha categoría de respuesta.
- Los parámetros asociados con la última categoría de cada factor son redundantes si se conoce el término de intersección.

Hay tres columnas adicionales en la tabla que proporcionan más información acerca de los resultados combinados. La **fracción de información perdida**es una estimación de la proporción de información perdida para "completar" la información, basada en el **aumento relativo de la varianza** por causa de la ausencia de respuestas, que es un porcentaje (modificado) de los valores entre imputaciones y una media de la varianza en la imputación del coeficiente de regresión. La **eficacia relativa** es una comparación de este cálculo con respecto a un cálculo (teórico) utilizando un número infinito de cálculos. La eficacia relativa está determinada por la fracción de información perdida y el número de imputaciones utilizadas para obtener el resultado combinado; si la fracción de información perdida es grande, se necesitará un gran número de imputaciones para aproximar la eficacia relativa a 1 y el cálculo combinado al cálculo ideal.

Figura 5-38
Estimaciones combinadas de parámetros



- Vuelva a activar (pulse dos veces) la tabla y seleccione Paneles de pivotado en el menú contextual.
- Cambie *Número de imputación* de Capa a Columna.
- Cambie *Estadísticos* de Columna a Capa.
- En la lista desplegable Estadísticos, seleccione B.

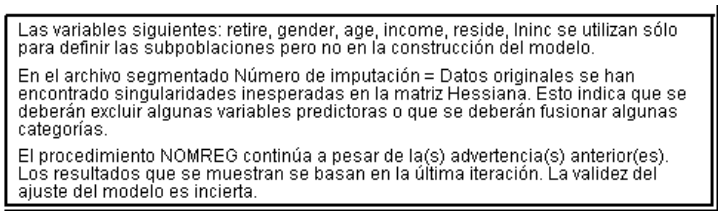
Figura 5-39
Estimaciones combinadas de parámetros, Número de imputación en columnas y Estadísticos en Capa

Estadísticos= B

Categoría de cliente[a,b,c,d,e,f]		Número de imputación						
		Datos originales	1	2	3	4	5	Combinado
Servicio electrónico	Intersección	,637	,519	,438	,437	,447	,465	,461
	employ	,054	,041	,029	,037	,039	,033	,036
	[marital=0]	-,760	-,678	-,707	-,710	-,685	-,673	-,691
	[marital=1]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	[ed=1]	-2,272	-2,010	-2,013	-1,845	-1,940	-1,865	-1,935
	[ed=2]	-1,657	-1,229	-1,221	-1,056	-1,097	-1,133	-1,147
	[ed=3]	-,848	-,537	-,549	-,500	-,513	-,561	-,532
	[ed=4]	-,576	-,499	-,490	-,424	-,358	-,457	-,446
	[ed=5]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	address	,028	,024	,044	,028	,025	,030	,030
Servicio plus	Intersección	-17,912	-1,153	-1,189	-1,268	-1,273	-1,297	-1,236
	employ	,056	,054	,048	,052	,049	,049	,050
	[marital=0]	-,549	-,419	-,405	-,433	-,440	-,355	-,411
	[marital=1]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	[ed=1]	17,534	,701	,682	,747	,826	,852	,761
	[ed=2]	17,354	,757	,735	,877	,844	,850	,813
	[ed=3]	17,762	,881	,902	,925	,947	,910	,913
	[ed=4]	17,906	,654	,714	,858	,872	,833	,786
	[ed=5]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	address	,023	,010	,019	,015	,017	,015	,015
Servicio total	Intersección	1,266	1,229	1,111	1,132	1,121	,992	1,117
	employ	,044	,046	,044	,045	,044	,046	,045
	[marital=0]	-,522	-,650	-,685	-,741	-,684	-,562	-,664
	[marital=1]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	[ed=1]	-3,590	-3,288	-3,676	-3,307	-3,594	-3,040	-3,381
	[ed=2]	-2,133	-1,833	-1,840	-1,724	-1,766	-1,730	-1,779
	[ed=3]	-1,214	-1,302	-1,203	-1,184	-1,262	-1,194	-1,229
	[ed=4]	-,468	-,616	-,530	-,441	-,517	-,454	-,512
	[ed=5]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	address	,012	,001	,014	,004	,009	,007	,007

Esta vista de la tabla es útil para comparar los valores entre imputaciones, para obtener una vista rápida de la variación en el coeficiente de regresión de imputación a imputación, e incluso con respecto a los datos originales. En concreto, cambiar los estadísticos de la capa a error típico permite ver cómo la imputación múltiple ha reducido la variabilidad en las estimaciones de coeficiente con respecto a la eliminación por lista (datos originales).

Figura 5-40
Advertencias



Sin embargo, en este ejemplo, el conjunto de datos original causa un error, que explica las grandes estimaciones de parámetros para los niveles de intersección y no redundantes de *Servicio plus* de *educ (Nivel educativo)* en la columna de datos originales de la tabla.

Resumen

Mediante los procedimientos de imputación múltiple, ha analizado patrones de valores perdidos y ha detectado que perdería la mayoría de esa información si utilizara el método de la eliminación por lista simple. Tras una ejecución automática inicial de imputación múltiple, ha observado que necesitaba limitaciones para mantener los valores en los límites razonables. La ejecución con limitaciones produce buenos resultados y no existen pruebas de que el método FCS no fuera adecuado. Mediante un conjunto de datos “completo” con valores con imputación múltiple, ha ajustado una regresión logística multinomial a los datos y ha obtenido cálculos de regresión combinada y ha descubierto que el modelo final no habría sido posible utilizando el método de la eliminación por lista en los datos originales.

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 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 , 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 , 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 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 , 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 . 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" . 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 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 presentó una tabla para ilustrar posibles grupos sociales. Guttman 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 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, 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 .
- **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 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 .
- **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 .
- **verd1985.sav.** Archivo de datos sobre una encuesta . 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 . 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.

Índice