

# Cuadernos Metodológicos

---

---

**45**

**2.<sup>a</sup> edición**

## Análisis de datos con Stata

**Modesto Escobar Mercado  
Enrique Fernández Macías  
Fabrizio Bernardi**

Stata es uno de los paquetes estadísticos de referencia en las comunidades científicas de muy diversas ramas, como la economía, la ciencia política y la sociología. En este Cuaderno Metodológico se enseñan los rudimentos de su uso mediante aplicaciones prácticas y explicaciones sustantivas de análisis de datos. Los contenidos de esta obra abordan con nivel básico e intermedio las técnicas más utilizadas en la investigación social (tablas de contingencia, comparación de medias, análisis gráfico, regresión lineal, análisis logístico, historia de acontecimientos y ponderaciones). El hecho de que todas las explicaciones estén guiadas con ejemplos reales facilita la comprensión de la técnica y su aplicación práctica en las ciencias sociales. El libro propone numerosos ejercicios con investigaciones reales, cuyos datos e instrucciones están disponibles en [www.cis.es/publicaciones/CM/](http://www.cis.es/publicaciones/CM/). Esta segunda edición se ha actualizado a la versión 12 del programa y la revisión del texto se ha seguido beneficiando del StataCorp's Author Support Program.

**CIS**

---

Centro de Investigaciones Sociológicas

# **Cuadernos Metodológicos**

---

---

**45**

**Análisis de datos  
con Stata**

**2.<sup>a</sup> edición revisada**

**Modesto Escobar Mercado  
Enrique Fernández Macías  
Fabrizio Bernardi**

**CIS**

---

Centro de Investigaciones Sociológicas

Consejo Editorial de la colección Cuadernos Metodológicos

DIRECTOR

Félix Requena Santos, *Presidente del CIS*

CONSEJEROS

Franciso Alvira Martín, *Universidad Complutense de Madrid.*

M.<sup>a</sup> Ángeles Cea D'Ancona, *Universidad Complutense de Madrid.*

Jesús M. de Miguel Rodríguez, *Universidad de Barcelona.*

Modesto Escobar Mercado, *Universidad de Salamanca.*

J. Sebastián Fernández Prados, *Universidad de Almería.*

Juan Ignacio Martínez Pastor, *Universidad Nacional de Educación a Distancia.*

SECRETARIA

M.<sup>a</sup> Paz Cristina Rodríguez Vela, *Directora del Departamento de Publicaciones y Fomento de la Investigación, CIS*

Las normas editoriales y las instrucciones para los autores pueden consultarse en:  
<http://www.cis.es/publicaciones/CM/>

Todos los derechos reservados. Prohibida la reproducción total o parcial de esta obra por cualquier procedimiento (ya sea gráfico, electrónico, óptico, químico, mecánico, fotocopia, etc.) y el almacenamiento o transmisión de sus contenidos en soportes magnéticos, sonoros, visuales o de cualquier otro tipo sin permiso expreso del editor.

COLECCIÓN «CUADERNOS METODOLÓGICOS», NÚM. 45

Catálogo de Publicaciones de la Administración General del Estado  
<http://publicacionesoficiales.boe.es>

Primera edición, diciembre de 2009

Segunda edición, mayo de 2012

© CENTRO DE INVESTIGACIONES SOCIOLÓGICAS  
Montalbán, 8. 28014 Madrid

© Modesto Escobar Mercado.  
© Enrique Fernández Macías.  
© Fabrizio Bernardi.

DERECHOS RESERVADOS CONFORME A LA LEY

Impreso y hecho en España  
*Printed and made in Spain*

NIPO: 004-12-004-0

ISBN: 978-84-7476-588-5

Depósito legal: M. 19.130-2012

Fotocomposición e impresión:  
Efca, S. A.  
Parque Industrial "Las Monjas" Verano, 28  
28850 Torrejón de Ardoz (Madrid)



El papel utilizado para la impresión de este libro es 100% reciclado y totalmente libre de cloro.

# Índice

	<i>Págs.</i>
1. INTRODUCCIÓN .....	9
2. PRIMEROS PASOS CON STATA .....	17
2.1. La información en los archivos de Stata .....	17
2.2. La interfaz de Stata .....	20
2.3. Las ventanas de Stata .....	25
2.4. Modos de trabajo en Stata .....	33
2.5. El fichero de resultados .....	40
2.6. Las variables de la matriz de datos .....	43
2.7. Ejercicios .....	56
3. INTRODUCCIÓN DE DATOS .....	59
3.1. Introducción manual de datos .....	59
3.2. Lectura de datos con Stata .....	64
3.3. Fusión de ficheros .....	78
3.4. Ejercicios .....	82
4. ESTADÍSTICAS DE UNA SOLA VARIABLE .....	85
4.1. Clasificación de variables .....	85
4.2. La tabla de distribución de frecuencias .....	87
4.3. Estadísticos resúmenes de distribuciones.....	90
4.4. Obtención de las medidas características de una distribución .....	96
4.5. La ponderación de los datos .....	99
4.6. El error típico.....	105
4.7. Ejercicios .....	114
5. MANIPULACIÓN Y MODIFICACIÓN DE DATOS .....	117
5.1. Manipulación de datos .....	117

	<i>Págs.</i>
5.2. Generación y modificación de variables .....	128
5.3. Características e instrucciones especiales .....	141
5.4. Ejercicios.....	147
6. GRÁFICOS CON STATA.....	149
6.1. Características de los gráficos de Stata .....	150
6.2. Gráficos unidimensionales .....	153
6.3. Gráficos bidimensionales.....	169
6.4. Componentes de los gráficos.....	184
6.5. Esquemas .....	186
6.6. El editor de gráficos.....	190
6.7. Ejercicios.....	194
7. LA PRUEBA ESTADÍSTICA Y LAS COMPARACIONES.....	195
7.1. Pruebas de una sola variable.....	197
7.2. Comparación de dos variables .....	204
7.3. Comparaciones de dos muestras (independientes) .....	213
7.4. Comparaciones de $k$ muestras independientes .....	219
7.5. Comparaciones de $k$ muestras dependientes .....	229
7.6. Ejercicios.....	236
8. CONFECIÓN Y ANÁLISIS DE TABLAS CON STATA.....	237
8.1. Tablas de contingencia de dos variables.....	238
8.2. Más de dos variables .....	255
8.3. Otras tablas especiales .....	258
8.4. Las tablas de respuesta múltiple.....	265
8.5. Ejercicios.....	274
9. LA REGRESIÓN.....	277
9.1. Nube de puntos, varianza y correlación entre dos variables..	278
9.2. La regresión simple .....	283
9.3. Bondad del ajuste de la regresión.....	289
9.4. Inferencias en la regresión simple .....	293
9.5. Regresión múltiple .....	297
9.6. Regresión con variables ficticias.....	304
9.7. Regresiones con interacción.....	311
9.8. Otras relaciones funcionales de la regresión .....	318
9.9. Ejercicios.....	328

	<i>Págs.</i>
10. DIAGNÓSTICO DE LA REGRESIÓN .....	331
10.1. Supuestos de la regresión lineal.....	331
10.2. Análisis de los casos en la regresión.....	344
10.3. Regresiones especiales .....	353
10.4. Regresión robusta.....	361
10.5. Regresión de cuantiles .....	368
10.6. Regresión por bandas .....	372
10.7. Ejercicios.....	373
11. LA REGRESIÓN LOGÍSTICA .....	375
11.1. El modelo estadístico .....	375
11.2. Estimación del modelo .....	382
11.3. Diagnóstico del modelo.....	388
11.4. Comparación de modelos .....	400
11.5. Interpretación del modelo .....	406
11.6. Ejercicios.....	420
12. REGRESIÓN LOGÍSTICA PARA VARIABLE ORDINAL Y MULTINOMIAL .....	421
12.1. El modelo estadístico del logit ordinal.....	421
12.2. Estimación e interpretación del modelo .....	425
12.3. El supuesto de regresiones paralelas o razones proporcionales .....	430
12.4. Regresión logística para variable dependiente nominal .....	433
12.5. Estimación e interpretación del modelo .....	435
12.6. El supuesto de independencia de alternativas irrelevantes .....	440
12.7. Ejercicios.....	442
13. EL ANÁLISIS DE LA HISTORIA DE ACONTECIMIENTOS CON STATA.....	445
13.1. Qué es y cómo funciona el AHA .....	445
13.2. El AHA con Stata: instrucciones para definir los datos...	451
13.3. La función de supervivencia.....	456
13.4. Modelos de la tasa de transición con tiempo continuo ...	458
13.5. Ejercicios.....	467

	<i>Págs.</i>
14. ANÁLISIS DE DATOS DE ENCUESTA CON STATA .....	469
14.1. Ajustes en el análisis de muestras complejas.....	470
14.2. Ponderaciones, estratos y conglomerados.....	471
14.3. Un ejemplo práctico con Stata. Las órdenes <i>svy</i> .....	476
14.4. Ejercicios.....	486
15. BIBLIOGRAFÍA COMENTADA.....	487

*A María José Echeverría,  
Judit Balbás y Marta Fraile*



# 1

## Introducción

Es innegable que la estadística se ha convertido en una herramienta fundamental para la investigación en las ciencias sociales. Aunque nadie niegue tampoco que puedan realizarse estudios sobre el mundo humano que recojan y analicen datos sin necesidad de operaciones matemáticas, una parte considerable de análisis necesita aplicar conocimientos de esta rama del saber —aun siendo sólo de modo básico para contar ocurrencias o para extraer los resultados de una muestra al conjunto de elementos que se desea investigar.

Hace cincuenta años todos los instrumentos que se disponían para las operaciones estadísticas eran el papel, el lápiz o bolígrafo y, en el mejor de los casos, una calculadora que había que enchufar a la red eléctrica y, sólo en los modelos más exclusivos, capaz de calcular raíces cuadradas. Desde entonces, dos desarrollos casi paralelos han cambiado las posibilidades de aplicación de la estadística a la investigación. Por un lado, el desarrollo de la informática, que ha puesto a disposición del bolsillo de los particulares la adquisición de un ordenador con capacidades de cálculo que antaño sólo estaban a disposición de multinacionales y organismos públicos, y, por el otro, la aparición de programas especializados en tareas estadísticas, que han permitido la ejecución de tareas de enorme complejidad a personas con escasos conocimientos matemáticos.

De acuerdo con este panorama, este libro pretende ser un manual que permita a quien lo trabaje un uso aplicado y racional de las herramientas estadísticas usadas en la investigación social y, por extensión, a la investigación biosanitaria o epidemiológica. Aunque el objetivo central de estas páginas es enseñar a utilizar un programa estadístico determinado —no muy distinto de otros que existen en el mercado—, esta obra también explica cuáles son los requerimientos, los procedimientos y, ante todo, la interpretación de los resultados de aplicar técnicas estadísticas a un conjunto de datos. Por ello, los autores han pretendido conjuntar lo que sería una introducción a la estadística aplicada con un manual de iniciación a Stata.

Stata es una aplicación estadística nacida en el año 1985 en el entorno Unix, e inmediatamente trasladada al sistema operativo DOS, Windows y,

posteriormente, al OS de Apple. Este programa ha tenido tres importantes precursores: por orden de antigüedad destacan el BMDP (*Biomedical Program*), el SAS (*Statistical Analysis System*) y el SPSS (*Statistical Programs for the Social Sciences*). Estos nacieron concebidos en entornos de grandes ordenadores, evitando la programación en Fortran para la resolución de los problemas estadísticos, para pasar a ejecutarse también en ordenadores personales en los años ochenta. Entre estos tres y Stata también cabe destacar la aparición de otras aplicaciones estadísticas (SYSTAT y StatGraphics, por ejemplo) que se implementaron en el entorno Windows con una filosofía mucho más interactiva que los iniciales paquetes, más pensados para procesos por lotes que para instrucciones instantáneas presentes en un menú a disposición del usuario.

Tres son las características más sobresalientes que han permitido que Stata obtenga una posición destacada entre las aplicaciones estadísticas: en primer lugar, el empleo de instrucciones con un lenguaje fácil de modo interactivo. Frente al primer acercamiento de los programas clásicos a través de instrucciones escritas en conjunto en un fichero, o al más moderno estilo de dar órdenes a través de menús, Stata ofrecía un modelo en el que se escribía una instrucción e inmediatamente se veían los resultados, siempre y cuando fuera bien escrita. En segundo lugar, Stata se especializó en el análisis de regresiones. Frente a otros programas estadísticos que prácticamente abarcaban un amplio elenco de análisis, Stata —aunque no de modo exclusivo— se concentró especialmente en los diversos análisis de regresión, ofreciendo una amplia variedad de procedimientos que van desde la regresión simple hasta los modelos de ecuaciones estructurales. Final y principalmente, no sólo se podían empaquetar todas las instrucciones en un fichero para su empleo contrastado y repetido, sino que también, por la propia naturaleza de las instrucciones analíticas, combinadas con las funciones y las órdenes de flujo, era posible para un técnico experto la confección de nuevas utilidades, distintas de las implementadas, pero compatibles en una variedad de situaciones por un conjunto de investigadores con el único requisito común de disponer de este programa, abierto a nuevas programaciones.

Por tanto, podrían destacarse, frente a otros programas de estadística, las siguientes características: en primer lugar sobresale por su facilidad de uso. Especialmente tras su versión 8 (enero de 2003), en la que se incorpora un sistema de menú que prácticamente integra todas las instrucciones disponibles, Stata es un paquete asequible al autoaprendizaje, si además se consulta la detallada documentación en formato pdf que le acompaña. También destaca Stata por una amplia gama de tareas. Bien es cierto que donde sobresale frente a otros programas estadísticos es en las regresiones, pero también se destaca en análisis de muestras complejas, en series temporales, en datos de panel, en análisis de sucesos históricos, en imputaciones de casos perdidos y, más recientemente, en gráficos de contornos y de marginales, así como en el análisis de modelos de ecuaciones estructurales. También

es un punto fuerte de esta aplicación su carácter abierto, pues no sólo da al usuario experto la posibilidad de generar o modificar programas, sino que también permite a los usuarios menos hábiles la importación de esas nuevas herramientas a su sistema<sup>1</sup>. Esto, junto a la extensión de las comunicaciones informáticas, ha generado una comunidad científica que comparte problemas y soluciones, ampliando a su vez las posibilidades de análisis con este paquete estadístico.

Este libro va dirigido a principiantes, especialmente a los que se están iniciando en el uso de Stata. Su curva de aprendizaje es especialmente plana en los comienzos, es decir, en el inicio su uso no resulta fácil; después, se aprende con menor dificultad hasta un nivel en el que ya sólo pueden avanzar aquellos que sobresalgan en estadística o en programación. El objetivo de las páginas que siguen a continuación es hacer menos costoso el aprendizaje inicial de esta herramienta de trabajo. Tiene dos posibles usuarios: en primer lugar aquellos académicos, especialmente del campo de la sociología y de la ciencia política, que nada o poco sepan de estadística y quieran hacerlo de la mano de uno de los programas especializados con mayor prestigio en el mundo universitario; en segundo lugar, para aquellos expertos en técnicas de análisis que quieran aprender a manejar un primer o un segundo programa de tareas estadísticas. Aunque se contengan determinadas explicaciones estadísticas, estas están dirigidas más bien a un público neófito o a veces están redactadas como recordatorio de cosas que se supone son conocidas. No es la intención de estas páginas enseñar más estadística a aquel que ya la conoce y mucho menos está dirigido a quien domina Stata y quiere aprender más técnicas de análisis de datos. Para este último público habría que escribir un nuevo libro donde este termina.

En cualquier caso, debe subrayarse que este no es sólo un manual de un programa de estadística. A la vez que se explica el uso de las órdenes, también se dice para qué sirven, en qué condiciones han de usarse y, sobre todo, cómo han de interpretarse los resultados. Se ha escrito con la intención de que aquellos profesionales o académicos que quieran adentrarse en el empleo de la estadística aplicada no se encuentren sólo con un listado de instrucciones, sino con una guía que les ayude a saber emplearlas y sacarles su jugo. Este programa viene acompañado de un manual tan extenso (veintiún volúmenes en su versión 12, aparecida en julio de 2011) que tiende a desanimar a quien se enfrenta por primera vez a él. Bien es cierto que entre estos volúmenes hay una guía de inicio y otra para el usuario; pero la primera es

---

<sup>1</sup> Como muestra, sirvan los ejemplos de *Panelwhiz* (<http://www.panelwhiz.eu>), que permite trabajar con bases de datos provenientes de algunas instituciones de diferentes países (Alemania, Australia, Reino Unido, EE UU y Europa) y DASP (Distributive Analysis Stata Package), que contiene herramientas estadísticas para el análisis de la distribución de la riqueza (<http://dasp.ecn.ulaval.ca/index.html>), y las rutinas SPost de Long y Freese (2006) para regresiones de variable dependiente nominal, así como la instrucción para la tabulación múltiple *mrtab* de Benn Jann (2005), que serán explicadas más adelante en este libro.

más bien insuficiente y la segunda es desigual en la complejidad de los temas tratados; por lo que se estima que es mejor la secuencia de aprendizaje presentada en esta obra. Además, mientras se planificaban y escribían las páginas presentes, se ha pretendido cubrir los contenidos de un curso de estadística intermedio, con lo que, además de un manual de un programa, también puede ser considerado como un medio para aprender a procesar y analizar datos mediante una aplicación de ordenador.

Por todo ello, este libro presenta una estructura secuencial: de lo simple a lo complejo. Desde los rudimentos básicos a los planteamientos más avanzados. Y no sólo entre capítulos, sino también en el interior de los mismos se ha pretendido ir de lo sencillo a lo complicado. En consecuencia, para muchos lectores las páginas que siguen a continuación no se tienen que leer o estudiar una por una. Se considera muy conveniente que cuando se llegue a un apartado complejo, se pase a un capítulo siguiente, para volver a él, cuando se esté más familiarizado con el programa. Ejemplo claro de ello es el último apartado del primer capítulo, que versa sobre los formatos de las variables, tema complejo donde los haya, que tiene más que ver con la presentación que con el contenido de un análisis y con más interés informático que estadístico, aunque lamentablemente de crucial necesidad para cuando se trabaja con variables de tipo.

El próximo capítulo está dedicado a los elementos básicos del programa. Comienza con la interfaz, que constituye el modo de comunicación de la máquina con el usuario. Se analizan cada una de las ventanas y menús a través de los que el investigador puede solicitar o contemplar resultados. Se explican los tres principales modos de trabajo con el programa, es decir, con menús, con instrucciones o con ficheros, y finalmente se clasifican y describen los distintos ficheros en los que se guardan los datos, los resultados, las órdenes o las ayudas del programa.

A continuación, se dedica un capítulo a la introducción de datos: primero, a la entrada manual de información; después a la lectura automática de otros ficheros en formato texto y a la conversión de archivos escritos en otros programas o aplicaciones, como Excel y SPSS, al formato propio de Stata. Finalmente, termina este apartado con una serie de instrucciones que permiten la manipulación del fichero propio de la aplicación, sea para añadirle casos o para adjuntarle variables procedentes de otros archivos.

El cuarto capítulo es el primero que se dedica al análisis estadístico propiamente dicho. Aborda la estadística univariable descriptiva, comprendiendo las medidas de tendencia central, las de posición, las de dispersión, las de simetría y las de apuntamiento, con el fin de estudiar las distribuciones y su comportamiento. A continuación se exponen los procedimientos más fáciles para la ponderación con Stata, reservando para el último capítulo los procedimientos complejos de ponderación de muestras. Y, al final del capítulo, se realiza una introducción a la estadística inferencial explicando el

error típico y los intervalos de confianza, imprescindibles para la estimación de los parámetros de la población.

Tras este primer capítulo de análisis, se incluye el tema del tratamiento y modificación de datos. Stata contiene una serie de instrucciones que permiten ordenar y seleccionar los casos sin que nada quede alterado. A este proceso se le denomina tratamiento y puede ser útil en una muy amplia variedad de casos, como cuando se desee realizar un análisis específico de jóvenes o de mujeres. Por otro lado, la modificación de datos incluye tanto la recodificación de los valores de las variables del fichero como la generación de nuevas variables mediante transformaciones algebraicas de otras ya existentes, porque muy a menudo no interesa analizar los datos tal como fueron recogidos, sino tras aplicarles algunos cambios que mejoren su presentación, como puede ser el caso de presentar una tabla de la edad con sus valores recodificados.

La versión 8 de Stata modificó radicalmente las instrucciones para la confección de gráficos y, como no podía ser de otra manera, se dedica un capítulo a la realización de estos. Aunque los menús pueden facilitar esta tarea, la petición de un gráfico no es tarea fácil, cuando se desea algo distinto de lo estándar. Debido a esta dificultad, a partir de la versión 10 aparece un editor de gráficos con el que se simplifica en gran medida el empleo de cambios sin tener que aprender las opciones o sin tener que navegar por un enmarañado sistema de menús y solapas. El orden de presentación de los distintos tipos de gráficos es funcional, por su uso. Primero se describen los gráficos unidimensionales que representan una o varias variables en una única escala y, a continuación, se enumeran los que contienen al menos dos escalas distintas. Finalmente, al acabar el capítulo, se explica brevemente el editor de gráficos.

En el séptimo capítulo se tratan las pruebas estadísticas más simples contenidas en Stata. Se comienza con una introducción sobre la prueba estadística aplicada a una sola variable. Se explican las pruebas paramétricas de proporciones y medias y la prueba de los signos. Con ellas, los investigadores pueden comprobar si sus hipótesis descriptivas son congruentes con sus datos. Pero también en este capítulo se explican las pruebas estadísticas que sirven para las hipótesis comparativas. Hay pruebas para dos o más muestras independientes (una variable medida en grupos distintos) o para dos o más dependientes (dos o más variables obtenidas en un único grupo).

Una de las operaciones más empleadas en el análisis de cuestionarios son las tablas de contingencia. A ellas se le dedica todo un capítulo. Se estudian los distintos tipos de porcentajes que se pueden aplicar, los residuos, las pruebas estadísticas de significación y los coeficientes de asociación. Todo ello, principalmente, para estudiar la fuerza de la relación entre dos variables. Complementariamente, se termina el capítulo con una consideración sobre las tablas de más de dos dimensiones en lo que puede considerarse una introducción al análisis multivariante y con la expli-

ción de un programa externo que permite la tabulación de preguntas multi-respuesta.

Los cuatro siguientes capítulos están dedicados a lo que son los procedimientos más notables de Stata: las regresiones. Se comienza en el noveno con la representación de dos variables en la nube de puntos para explicar el concepto de covarianza y el de correlación; se explica el método de mínimos cuadrados para la extracción de una recta que pase lo más cerca posible del conjunto de puntos representados y se abordan los temas más espinosos de la estimación de los parámetros poblacionales. Después de explicar en un primer momento la regresión simple (con una sola variable independiente), se pasa a la regresión múltiple (más de una variable independiente), se aborda posteriormente el uso e interpretación de determinados tipos de variables como las dicotómicas y, finalmente, se presentan modelos con una relación funcional distinta de la lineal (regresiones cuadráticas, cúbicas, exponenciales, logarítmicas, inversas...).

El décimo capítulo se dedica a analizar los supuestos de la regresión y las posibles soluciones a sus anomalías. Se presta atención a las medidas para la detección de casos anómalos que desvirtúan la obtención de la recta y se acaba con un conjunto de regresiones especiales que evitan los problemas generados por el no cumplimiento de las asunciones de este análisis. Se explican, en consecuencia, las regresiones con ponderación, las robustas y las realizadas por bandas o cuantiles.

El tercero de los capítulos dedicados a la regresión se centra en la logística binaria, para los casos en los que se desee efectuar una regresión con variable dependiente dicotómica. En el caso de las ciencias sociales abundan las variables nominales, por lo que esta alternativa a la regresión común puede aplicarse evitando que las predicciones se salgan de los límites propios de este tipo de variables. El último capítulo de regresiones, el duodécimo, versa sobre otras variables de respuesta no cuantitativa. En particular, se centra sobre las regresiones ordinales y multinomiales, para abordar aquellos casos en los que es insuficiente un tratamiento dicotómico de la variable dependiente. Para todas estas técnicas estadísticas se explican no sólo las órdenes del programa, sino también otras disponibles en Internet creadas por Long y Freese (2006).

Tras las regresiones se dedica un capítulo a una técnica en la que Stata posee muy amplios recursos como es el análisis de la historia de acontecimientos, especialmente útil para el análisis dinámico de los fenómenos naturales o sociales.

En último lugar, se cierra este manual introductorio con un capítulo dedicado a las ponderaciones de muestras especiales muy útiles en el procesamiento de encuestas, puesto que en estas rara es la ocasión en la que se resuelven mediante muestreos aleatorios simples.

Obviamente no se trata gran parte de los análisis que están disponibles en Stata. Entre otros, son de especial mención por su importancia las regre-

siones multinivel y condicionales, las series temporales, el análisis factorial o el de conglomerados, los datos dispuestos en panel, la imputación de casos perdidos, o los recién incorporados modelos de ecuaciones estructurales. Su inclusión implicaría doblar las páginas de este manual, cuya pretensión es introductoria. Posiblemente debería escribirse un segundo volumen dedicado a análisis más complejos para cubrir todas estas lagunas. En cualquier caso, siempre están los excelentes manuales de Stata para el que desee ir más allá.

Lo que ha presidido en la redacción de este libro es ante todo la simplicidad. Por ello, en las explicaciones se tiende a ir de lo simple a lo complejo y se insiste en el aprendizaje a través de los ejemplos. Todos los resultados de análisis mostrados están acompañados previamente de la instrucción que los genera y, posteriormente, del comentario pertinente. No se olvida incluir las fórmulas de las operaciones, con el fin de que no sólo sea un manual de un programa, sino también una introducción a la estadística y, como complemento para navegar en el inmenso caudal de órdenes, opciones y subopciones de este programa, se ofrece un índice de instrucciones para que el lector sepa en qué lugar del libro se encuentra su explicación. También son importantes los ejercicios incluidos al final de cada capítulo, ya que sólo con la práctica se conseguirán dominar las dificultades propias de su buen uso. Sin embargo, difícilmente podrán leerse las páginas que siguen a continuación secuencialmente. Para quienes se inician en Estadística, se sugiere que comiencen con los capítulos 4, 6, 7, 8 y 9, para continuar con el 2, el 3 y el 5. Quienes quieran aprender sólo Stata, en cambio, deberían empezar con estos y después proseguir con los capítulos que versen sobre lo que realiza más frecuentemente o con aquellos que le resulten más desconocidos, según su destreza sea poca o mucha. Finalmente, para los iniciados tanto en el programa como en el conocimiento estadístico, puede recomendarse una lectura a partir del décimo o undécimo capítulo, desde donde se abordan las técnicas más complejas.

La autoría de este libro se ha distribuido del siguiente modo: el capítulo 13 ha sido confeccionado por Fabrizio Bernardi; los capítulos 11, 12 y 14, por Enrique Fernández, que además es el autor principal de los capítulos 3, 5 y 8, y el resto de los textos, por Modesto Escobar, quien es también responsable de la actualización de todos los gráficos, ilustraciones e instrucciones para conformarlos a la última versión de Stata, pues en esta segunda edición de la obra se ha querido actualizar el contenido con algunas de las múltiples novedades incorporadas en la versión 12.

Queda, finalmente, agradecer todas aquellas contribuciones que han ayudado a que este producto haya visto la luz en su forma actual. La Universidad de Salamanca y el Instituto Juan March de Estudios e Investigaciones son las instituciones que más han contribuido a que los autores hayan podido dedicar sus esfuerzos en este empeño didáctico. Estudiantes de una y otra institución docente han recibido nuestras enseñanzas en estas mate-

rias, y no hay duda de que gracias a ellas se han reformulado el esquema, la orientación y la didáctica del presente texto. Además, la buena recepción que ha tenido la primera tirada de este manual nos ha motivado a sacar lo antes posible una nueva edición que incorporara los cambios de la última versión del programa. Debe mencionarse, por cierto, la excelente política de la empresa que lo produce, StataCorp LP, de asesoramiento a los autores de libros que emplean su programa. Esta obra se ha acogido a ella y se ha beneficiado de múltiples comentarios críticos de Gustavo Sánchez, especialmente útiles para que los ejemplos de las instrucciones contuvieran los menos errores posibles. Asimismo, Rubén Ruiz leyó una abundante selección del texto e hizo sugerencias muy útiles a los autores. Tampoco sería justo omitir la profesionalidad y paciencia de los editores, incorporando decenas de correcciones, así como la excelente tarea acometida por los dos evaluadores anónimos, al emitir inteligentes aportaciones que hemos intentado reflejar en esta obra. Ni que decir tiene que los errores que sigan presentes en el texto son sólo responsabilidad de quienes lo firmamos.

# 2

## Primeros pasos con Stata

### 2.1. La información en los archivos de Stata

Cualquier programa estadístico trabaja con información en muy diversos formatos almacenada en distintos tipos de fichero. Por ello, en este apartado se van a describir los principales archivos con los que trabaja Stata. Los seis tipos que se verán a continuación pueden dividirse en tres grandes grupos: los que guardan información sin procesar, los que conservan la información procesada y los que permiten o ayudan a transformar la información. En el primer grupo se incluyen los ficheros que contienen los datos individuales tal y como son introducidos en el ordenador después del trabajo de campo; en el segundo se consideran los archivos donde se guardan los estadísticos o gráficos que se generan con el análisis del programa y, finalmente, se consideran del tercer grupo los ficheros donde convenientemente se almacenan las instrucciones necesarias para realizar las tareas.

La base de trabajo es la matriz de datos, que consiste en una disposición ordenada de información, poco o nada procesada. Generalmente, el modo como esta se organiza para su tratamiento es de tal forma que los casos se encuentren expuestos en filas y las variables en columnas. Un ejemplo simple puede bastar para la comprensión de la estructura. Suponiendo que hubiera que analizar a dos personas, una mujer y un hombre de edades respectivas de 21 y 20 años, se pueden distinguir tres conceptos primordiales:

En primer lugar, el concepto de *caso*, esto es, cualquier unidad de la que se recoge información. En los datos anteriores existen dos, las dos personas de las que se saben sus características sociodemográficas. En segundo lugar, el concepto de *variable*, es decir, las características susceptibles de adquirir distintas modalidades. En el ejemplo presente, las dos variables disponibles son sexo y edad. Una de ellas es de naturaleza cualitativa, mientras la otra se presenta como cuantitativa. Cada una de las modalidades, cualitativas o cuantitativas, de estas variables recibe el nombre de *valor*. En este caso, son valores 20 y 21 años. También son valores “mujer” y “hombre”, aunque por no ser de naturaleza numérica, también pueden denominarse *atributos*, *categorías* o, en conjunto, un *factor*.

**ILUSTRACIÓN 2.1. Matriz literal de datos**

Hombre 20
Mujer 21

Esta matriz de datos podría condensarse aún más si se representan los atributos con una serie de códigos. Así puede reducirse *Hombre*, poniendo a todos los casos con esta característica un símbolo que lo represente, que puede ser Ø, H o preferiblemente un dígito, para que la introducción de la información se pueda hacer del modo más rápido y, por costumbre, el 1 para las personas de género masculino y el 2 para las mujeres. De esta forma, la matriz de datos original presenta una estructura como la siguiente:

**ILUSTRACIÓN 2.2. Matriz codificada de datos**

1 20
2 21

Esta matriz o conjunto de datos, para que pueda ser tratada informáticamente más de una vez, ha de ser guardada en un *fichero de datos*. En principio, cualquier archivo que contenga información ordenada puede ser leído directa o indirectamente por Stata. Pero sólo pueden ser utilizados desde el interior del programa media docena de formatos: ASCII o Unicode, XML, ODBC, SAS XPORT y Excel. Para el resto de casos, existen otros programas que transforman los ficheros generados por aplicaciones como hojas de cálculo, bases de datos o incluso otros programas estadísticos en ficheros de trabajo aptos para Stata. Entre ellos, uno de los más conocidos es *Stat-Transfer*<sup>1</sup>, cuyo uso y utilidad se verá en la sección 3.2.3.

Sin embargo, para el trabajo estadístico no basta con tener la matriz de datos bruta. Hay que añadirle al menos los nombres de las variables para que cuando se solicite una determinada tarea el programa sepa qué información se desea tratar. No es lo mismo solicitar una media del sexo que de la edad. Se podría indicar que se desea sólo una media de la segunda variable; pero es mucho más cómodo solicitarla llamándola edad. Por ello, una de las operaciones imprescindibles en todo programa estadístico es la de convertir el fichero de datos brutos en otro con la matriz de datos ampliada con las definiciones y transformaciones de la información original que el usuario considere conveniente.

---

<sup>1</sup> Stat-Transfer no es un producto de la casa *StataCorp.*, sino de *Circle System*, aunque fuera de Estados Unidos lo suelen comercializar las mismas empresas que venden Stata.

Estos específicos ficheros con información bruta, definiciones y transformaciones se denominan *ficheros de trabajo*. Sólo pueden construirse con el programa Stata o con otros pocos programas estadísticos que incluyen la posibilidad de guardar los datos en este formato. Generalmente se les reconoce por tener la extensión *.dta*<sup>2</sup>. En ellos están almacenados los datos de las variables originales y de las creadas posteriormente por el usuario, junto con sus correspondientes nombres, etiquetas y formatos.

Para que puedan comprobarse los ejemplos de su manual, Stata permite acceder a todos los ficheros empleados. Se puede obtener una relación de los ficheros de datos incorporados en la instalación del programa mediante la instrucción *sysuse dir*:

### ILUSTRACIÓN 2.3. Directorio de los ficheros de datos en el sistema

artificial.dta	census.dta	network1.dta	sp500.dta	voter.dta
auto.dta	citytemp.dta	networkla.dta	surface.dta	xtline1.dta
autord.dta	citytemp4.dta	nls88.dta	tsline1.dta	
bplong.dta	educ99gdp.dta	nlswide1.dta	tsline2.dta	
bpwide.dta	gnp96.dta	pop2000.dta	uslifeexp.dta	
cancer.dta	lifeexp.dta	sandstone.dta	uslifeexp2.dta	

Otros archivos de interés en el trabajo con Stata son los *ficheros de resultados* (con extensión *smcl* o *log*): siempre que así se le indique, los resultados de las órdenes dadas al programa son archivados en un fichero para que puedan quedar disponibles permanentemente, sin tener que volver a procesar de nuevo los datos mediante las instrucciones pertinentes. Stata dispone de un formato específico de grabación de los resultados en un fichero (*formatted log*) al que incorpora la extensión *smcl*, que consta de todos los elementos adicionales necesarios para una presentación idónea de las tablas estadísticas. Pero en ocasiones<sup>3</sup> es útil que los resultados se generen en un formato tratable universalmente, como es el caso de los ficheros en código ASCII. Por ello, también existe la posibilidad de grabar los resultados sin formato en archivos generados en Stata con la extensión *log*.

En cualquier caso, en ninguno de estos dos tipos de ficheros se incorporan los gráficos, pues cada uno de estos se guarda en un fichero independiente y específico para este tipo de representación de datos. Como se verá en el capítulo 6, cuando Stata genera un gráfico, el programa abre una ventana especial donde lo ubica y caso de que quiera conservarse, ha de grabarse como un *fichero gráfico*. Stata dispone de un formato propio (*gph*); pero, para que otros usuarios que no usen Stata lo puedan contemplar, también

<sup>2</sup> Determinados ficheros de datos creados para las funciones de impulso/respuesta asociadas a modelos VAR y VEC se guardan con la extensión *irf*, en lugar de *dta*.

<sup>3</sup> Por ejemplo, cuando se desea trasladar los resultados a otro programa, como puede ser un procesador de textos, o cuando se quiere que sean leídos en algún ordenador que no disponga del programa Stata.

permite grabarlo en otros formatos tales como metaarchivo de Windows (*wmf*); metaarchivo mejorado (*emf*); portable de red (*png*); postscript (*ps*); postscript encapsulado (*eps*); formato de documento portátil (*pdf*) y el formato de fichero de imagen etiquetada (*tif*).

Finalmente, es importante cerrar la lista de ficheros de Stata con los denominados ficheros de programa (*do* y *ado*), que contienen conjuntos de instrucciones de Stata que pueden ejecutarse automáticamente sin necesidad de tenerlas que volver a introducir interactivamente. Los hay básicamente de dos tipos: los primeros permiten repetir los mismos análisis o transformaciones de datos cuantas veces se desee a los mismos datos; mientras que los segundos se utilizan para aplicar un tratamiento común a datos diferentes, como si fuera una instrucción más del programa; pues se incorporan automáticamente al arrancarlo. Hay cientos de ellos que pueden obtenerse en las páginas oficiales de Stata; pero un usuario avanzado puede construirlos para su propio uso y puede ponerlos a disposición de la comunidad científica. Además, tienen su complemento en los ficheros de ayuda, distinguibles tanto en Stata como en otros programas por su extensión *sthlp* (o *hlp* en las primeras versiones), donde se incluyen explicaciones concretas de cómo pueden usarse las instrucciones programadas.

Como resumen, puede confeccionarse el siguiente esquema de los seis tipos de ficheros acabados de describir:

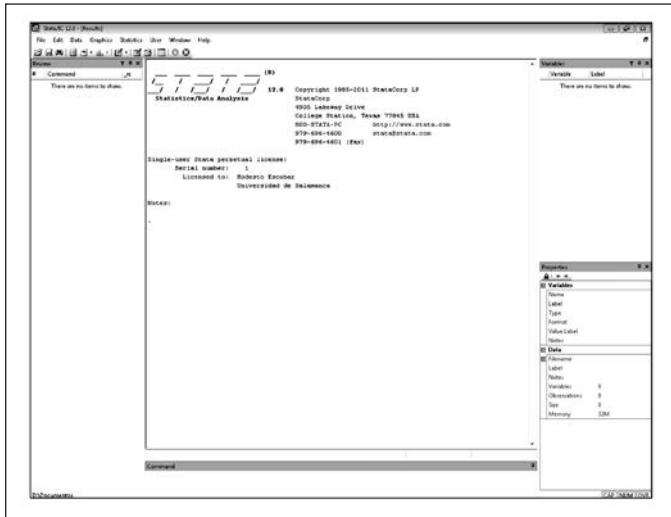
**CUADRO 2.1. Tipos de ficheros en Stata**

Tipo	Subtipo	Extensiones
Datos	Brutos	.dat, .txt y otras
	Ampliados (Trabajo)	.dta e .irf
Resultados	Textuales	.smcl y .log
	Gráficos	.wmf, .emf, .png, .pdf, .ps y .eps
Programas	Programa	.do y .ado
	Ayuda	.sthlp y .hlp

## 2.2. La interfaz de Stata

Al ejecutarse Stata, se muestra una pantalla compuesta por una serie de elementos cuyos usos y funciones se dan a continuación. Lo primero que hay que tener en cuenta son las cinco franjas horizontales que presenta la interfaz del programa. Todas estas divisiones, salvo la cuarta, que es la mayor y está compuesta por un conjunto de ventanas, presentan una sola línea de extensión vertical:

### ILUSTRACIÓN 2.4. Primera pantalla de Stata 12



La primera de las zonas presenta el color que por defecto le adjudique el sistema operativo a los programas que con él se ejecutan. En ella están indicados el nombre y la versión del programa que se ha puesto en marcha, el nombre del fichero de trabajo que estuviera abierto, así como la ruta o directorio del ordenador en el que se encuentra.

A continuación, en la segunda zona horizontal, aparece la franja del menú, compuesto por nueve apartados, que son los siguientes:

- 1) *File*: Este ítem del menú permite realizar la apertura, grabación e impresión de los distintos ficheros de trabajo analizados en el apartado anterior.
- 2) El segundo apartado del menú es *Edit*. Sirve para copiar y pegar fragmentos de texto. El uso más común que se da a esta instrucción es la de trasladar los resultados del análisis a otra aplicación como pueda ser un procesador de texto la mayor parte de veces, una hoja de cálculo o un programa de gráficos. También puede utilizarse para cortar y pegar determinados fragmentos de instrucciones de un lugar a otro. Las dos opciones principales de este menú son *copy* (copiar) y *paste* (pegar). Como en la mayor parte de los programas que se ejecutan con *Windows*, ambas pueden ser sustituidas respectivamente por la combinación de teclas *Ctrl+c* y *Ctrl+v*. También en este apartado se encuentra la opción de las preferencias (*Preferences*). Permite el cambio de determinados aspectos de las ventanas del programa. Dos son los principales apartados que pueden cambiarse. El relativo a los textos o ventanas y el relacionado con los gráficos. En relación con el primero, en las antiguas versiones de

este programa estaba asociada la imagen de los resultados con una pantalla de fondo negra en la que las instrucciones aparecían en blanco, los resultados estadísticos en amarillo, el texto complementario en verde y los errores en rojo. Todos los elementos de esta combinación pueden cambiarse tanto en la pantalla activa de resultados (*Results Colors*) como en el visor de otros ficheros (*Viewer Colors*) de modo independiente. En relación con los gráficos, puede cambiarse el esquema (véase la sección 6.5), la fuente de sus textos y algunos aspectos de la impresión o de su exportación directa<sup>4</sup> a otros programas. Finalmente, la disposición de las ventanas de *Stata* puede cambiarse si el usuario cambia manualmente el tamaño o la posición de estas y guarda su opción mediante *Save Preference Set*. A partir de ese momento, el programa se presentará de esa forma incluso después de salir al arrancar de nuevo. Por su lado, hay en la versión 12.0 incorporados seis modelos de disposición: para obtener la configuración mostrada en la ilustración 2.4 hay que optar por el *WideScreen Layout*, mientras que la disposición clásica de otras versiones se obtiene mediante la opción *Combined Layout*.

- 3) En los tres apartados siguientes del menú (*Data*, *Graphics* y *Statistics*) se despliegan las múltiples operaciones estadísticas de la que es capaz *Stata* a través de cuadros de diálogos. En el primero (*Data*) se incluyen aquellas instrucciones que sirven para describir los datos, transformarlos o hacer manipulaciones al fichero donde están contenidos. Una parte sustancial de estas órdenes están explicadas en la presente obra a lo largo de los capítulos 3 y 5. El segundo de los tres apartados en cuestión (*Graphics*) está reservado a las instrucciones gráficas. Las más importantes están contempladas en el capítulo 6. Y, bajo el rótulo de *Statistics*, se dispone la casi totalidad de operaciones estadísticas de la que es capaz este programa. Hay que tener en cuenta que esta posibilidad de obtener resultados estadísticos mediante menús y cuadros de diálogo sólo se ha incorporado a *Stata* a partir de su versión 8. A un usuario novel de *Stata* le resultará mucho más cómodo el empleo de estas guías. Sin embargo, un usuario experimentado preferirá escribir directamente las instrucciones una a una o recopilarlas en un fichero para ejecutarlas en serie.
- 4) La denominación *User*, situada en la sexta posición de la franja del menú, sirve para que un programador inserte allí sus propias utilidades. Por tanto, nada será dicho sobre este apartado en este libro introductorio.
- 5) El apartado *Windows* permite acceder a diez de los once tipos de ventanas que componen la estructura interna del programa *Stata* en su duodécima versión: instrucciones, resultados, historia (*review*),

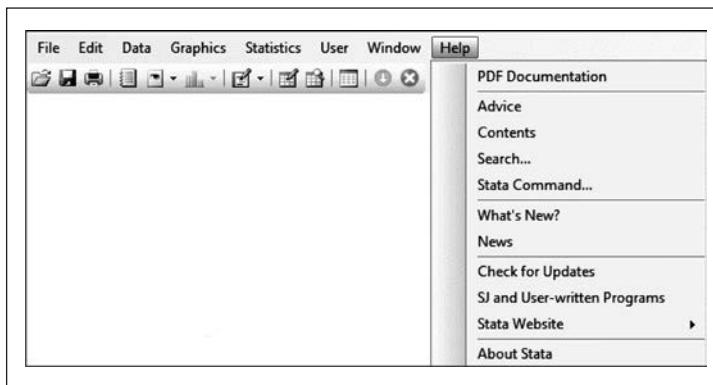
---

<sup>4</sup> Se entiende por exportación directa, cuando se utiliza el portapapeles de *Windows* para pasar un objeto de una aplicación a otra. Otro modo de traspasar un gráfico a otra aplicación es grabándolo en un fichero que sea capaz de ser leído por el susodicho programa.

variables, propiedades, gráficos, visor de ficheros, editor de datos, editor de programas y gestor de variables. Por su especial importancia, se dedicará el próximo apartado a su descripción.

- 6) Finalmente, no falta en el menú el ítem correspondiente a la ayuda (*Help*). En él se distinguen cinco partes diferenciadas: en la primera, sólo existe una línea que remite al manual completo de Stata, dividiendo en cada uno de sus volúmenes<sup>5</sup>; en la segunda, se ofrece toda la ayuda interna disponible del programa que se muestra en las ventanas de ayuda con un formato especial dotado de hipertextualidad, remitiendo tanto a otros contenidos de la misma ventana como a secciones concretas del manual en *pdf*; en la tercera, se ofrecen noticias del programa y detalles sobre el contenido de las actualizaciones desde la instalación inicial del software hasta la última versión instalada; en la cuarta aparecen posibles actualizaciones y extensiones del programa así como la ayuda ofrecida en la red, que será mostrada en el explorador de páginas web por defecto que se disponga, y la quinta ofrece el logotipo y dirección de la empresa, la cantidad de memoria física y la disponible en el ordenador, la versión de Stata y la información sobre la licencia en uso. Estas cinco partes se componen de once líneas distintas tal como se muestra en la ilustración 2.5:

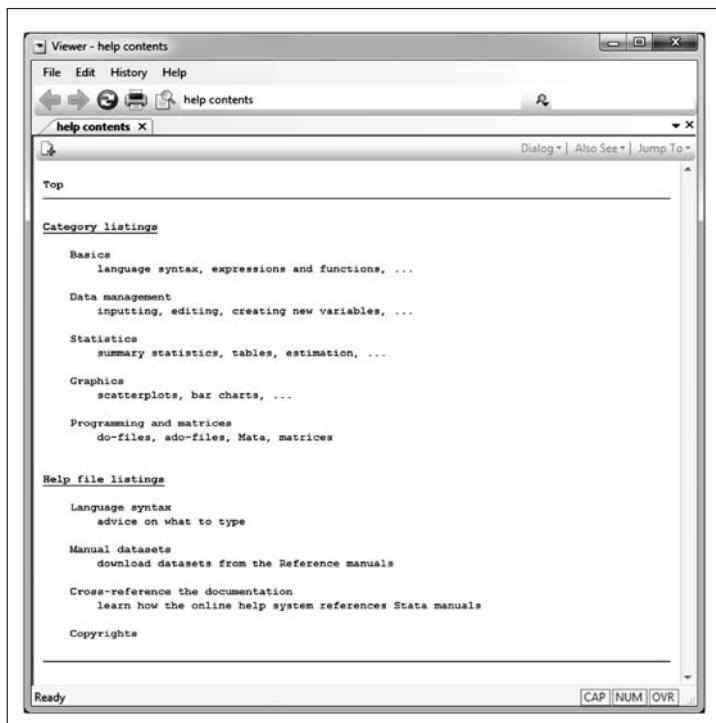
#### ILUSTRACIÓN 2.5. Menú de ayuda



<sup>5</sup> En la versión 12 se cuentan dieciocho unidades en el índice conjunto del fichero contenido en formato *pdf*: Contenidos, Guía de instalación [IG], Primeros pasos [GS] (una breve introducción con los aspectos básicos de Stata), Guía del usuario [U] (desarrollo de los elementos comunes más importantes de Stata: sintaxis, formatos, variables, funciones, macros y elementos básicos de programación), Gestión de datos [D], Gráficos [G], Imputación múltiple [MI], Estadísticas multivariantes [MV], Referencia básica [R] (listado alfabético de las instrucciones de Stata no contempladas en el resto de manuales), Modelos de ecuaciones estructurales [SEM], Análisis de supervivencia [ST], Datos de encuesta [SVY], Series temporales [TS], Datos longitudinales [XT], Programación [P], Mata [M] (lenguaje específico de programación para trabajar con matrices), Índice [I] y Tabla de contenidos.

Tras la documentación en *pdf*, hay cinco líneas correspondientes a la ayuda interna, que ofrece una serie de consejos para la obtención de distintos tipos de ayuda (*Advice*), un índice temático de los contenidos (*Contents*), un buscador de términos (*Search*), una referencia de todas las instrucciones del programa (*Stata Command*), un repertorio de novedades de la última versión instalada (*What's New*) y un noticiero relacionado con el programa (*News*). En cualquiera de los seis casos, aparece el visor de ayuda donde se expone lo solicitado a modo de hipertexto, de modo tal que se puede navegar por la ingente cantidad de información disponible tanto en el formato propio de Stata como en el formato *pdf*, que contiene la documentación completa del programa<sup>6</sup>. Como botón de muestra, la ilustración 2.6 contiene la pantalla obtenida al solicitar el índice temático de contenidos:

#### ILUSTRACIÓN 2.6. Índice temático de contenidos



<sup>6</sup> La ayuda que se muestra en la ventana del visor de Stata incluye prácticamente toda la documentación de los manuales con excepción de los ejemplos aislados del texto, las notas técnicas y las referencias bibliográficas.

Las tres siguientes líneas de la ayuda son las últimas actualizaciones disponibles (*Check for Updates*); programas divulgados en los boletines y revistas del programa (*SJ and User-written Programs*), que pueden incorporarse gratuitamente a los incorporados en el programa, y la página web oficial de la corporación Stata (*Stata Website*), subdividida por su parte en página principal, soporte al usuario, FAQ (preguntas respondidas frecuentemente), blog, revista y editorial.

La tercera franja horizontal de la interfaz de Stata es la barra de herramientas, que está constituida en las versiones 11 y 12 por doce iconos con operaciones útiles y frecuentes del programa. Estas son de izquierda a derecha las siguientes: apertura de un fichero de datos, grabación del fichero de datos activo, impresión de los resultados, apertura (visión o cierre) del fichero de resultados, apertura del visor de ayuda, activación de pantalla de gráficos, edición de programas, editor de datos, visor de datos, gestor de variables, botón de continuación en pantalla de resultados y botón de interrupción de resultados.

**ILUSTRACIÓN 2.7. Barra de herramientas de Stata 12**



En la cuarta franja de la pantalla se ubican entre dos (la de órdenes y la de resultados son inevitables) y cinco ventanas de Stata (las otras tres, optativas y flotantes, son la de variables, la del historial y la de propiedades) que serán descritas con más detalle en el próximo apartado.

Finalmente, en la franja inferior, con el mismo color de fondo que las líneas de menús e iconos, se encuentra la línea de estado, en la que se expone el nombre del directorio de trabajo donde se guardarán y leerán los distintos ficheros, a menos que se especifique un directorio distinto, además de los pilotos que indican en el extremo derecho si se encuentran pulsadas las teclas de fijación de mayúsculas (*CAP*), teclado numérico (*NUM*) e inserción de caracteres (*OVR*).

### 2.3. Las ventanas de Stata

Ya se ha dicho en el apartado anterior que Stata trabaja con once ventanas distintas, cinco internas y seis externas. No todas son igual de importantes, ni todas están presentes al mismo tiempo. De hecho, al empezar una sesión con Stata 12 sólo aparecen las cinco internas. Entre ellas, las más centrales para el trabajo son las de resultados y las de órdenes.

La ventana de órdenes (*Stata Command*), ubicada por defecto (*Wide-screen layout*) en la parte inferior de la cuarta franja del programa, es un

recuadro en blanco donde deben escribirse las instrucciones u órdenes de Stata. Una instrucción básica para empezar es *dir*, para saber los ficheros contenidos en el directorio de trabajo. Si se escribe en la ventana en cuestión esta palabra seguida por la tecla de retorno como final de la orden, inmediatamente aparecerá un texto en la *ventana de resultados*.

dir

En este caso, aparecerá un texto similar al de la ilustración 2.8:

#### **ILUSTRACIÓN 2.8.** Resultado de la instrucción *dir*

The screenshot shows the DataRCA 2.0 application window. The main area displays a file tree under the root directory 'C:\'. The tree includes nodes for 'dir', 'file', 'file2', and 'file3'. A status bar at the bottom indicates the path 'C:\documents\trial\trial\test\data\'. On the right side, there are two panels: 'Variables' and 'Properties'. The 'Variables' panel is currently empty, showing the message 'There are no items to show.' The 'Properties' panel shows details for the selected item 'file3', which is a file located at 'C:\documents\trial\trial\test\data\file3.dat'. The properties listed are: Name = file3, Type = File, Label = , Value = 0, ValueLabel = 0, Notes = , Variables = 0, Observations = 0, and Memory = 10M.

La ventana de órdenes, donde se escribió la instrucción, se queda en blanco después de ejecutarla. Sin embargo, la orden queda guardada en otra pantalla, mostrada aquí en la parte superior derecha de la ilustración, en la llamada *ventana de historia (Review)*. Además, en la ventana de resultados (*Stata Results*) aparece el producto de la primera instrucción, esto es, un listado con todos los ficheros ubicados en el directorio por defecto. Si hay más ficheros que líneas permite el tamaño de la pantalla, aparece el texto —*more*— en color diferente al del resto. Ante este mensaje, hay tres posibilidades: la primera es apretar la tecla *l* o *Intro (Enter)*, en cuyo caso, en la pantalla de resultados aparecerá una línea más. La segunda opción es apretar otra tecla distinta de las dos anteriores o el penúltimo ícono (*Clear-more--condition*). De este modo, el texto, en lugar de avanzar una línea,

avanzará toda una pantalla. Finalmente, si se desea interrumpir la salida de resultados, en el caso de que no se haya obtenido lo deseado, se puede pulsar *q*, la combinación de teclas *Ctrl+k*, o el último ícono de la barra de herramientas. Así se detendrá la orden, no aparecerán más líneas en pantalla y se estará en condiciones de escribir una nueva instrucción. Una línea con un solo punto en la pantalla de resultados indica que el sistema está listo para recibir otra orden.

Para ver algún contenido en la *ventana de variables (Variables)* es preciso crear o recuperar un conjunto de datos. Como la segunda de estas dos tareas es más fácil, se pondrá como primer ejemplo la recuperación de una base de datos incorporada en el programa y se deja la creación de un fichero de datos para un capítulo posterior. Con el fin de recuperarla, en la ventana de órdenes hay que escribir *sysuse auto*, siendo esta última palabra el nombre con el que es conocida esta base de datos, que consiste en un listado de automóviles comercializados en los años setenta en Estados Unidos acompañado con una serie de variables cuyo contenido son las características de los modelos.

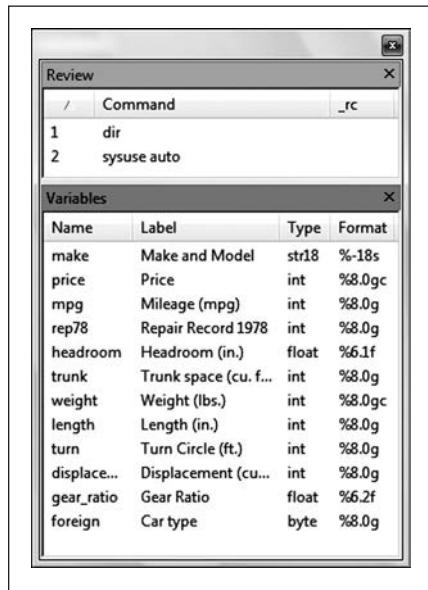
```
sysuse auto
```

Desde el momento en que se introduce esta instrucción, aparece en la ventana de variables la lista de ellas que están incluidas en el fichero *auto*. Es fácil advertir que cada línea corresponde a una variable y está dividida en cuatro columnas: la primera refleja su propio nombre, la segunda contiene su etiqueta, es decir un texto que la acompaña y que proporciona una descripción más extensa de su contenido, la tercera informa del tipo, mientras que la cuarta refleja su formato<sup>7</sup>. Estas propiedades de las variables serán descritas con más detalle en las secciones 2.6.1, 2.6.2 y 2.6.3.

Tanto la ventana de historia como la de variables permiten trasladar su contenido a la pantalla de órdenes. De este modo, si se lleva el cursor a la línea *dir* de la primera de las ventanas, aparecerá el texto en la ventana de órdenes y, si se pulsa *Intro* en esta, o si se pulsa un par de veces el texto en la ventana de historia, la instrucción mencionada será ejecutada de nuevo. En cambio, si se desea eliminar de la ventana de órdenes por haberla incluido por error, puede hacerse mediante la tecla *Esc*. En el caso de las variables, el funcionamiento es similar, salvo en que el doble *clic*, en lugar de ejecutar la instrucción, traslada el nombre de la variable a la ventana de órdenes.

<sup>7</sup> La presentación de la ventana de variables con cuatro columnas (ilustración 2.9) no se obtiene por defecto. Para que aparezcan las dos últimas (*type* y *format*) se ha de colocar el cursor en la barra de contenido (Variable|Label), hacer clic posteriormente en el botón derecho del ratón y, finalmente, marcar la(s) columna(s) deseada(s).

### ILUSTRACIÓN 2.9. Ventanas de historia y variables



Otra ventana de frecuente uso en Stata, ya mostrada en la ilustración 2.6, es el *visor de ayuda*. Con la instrucción *help* orden (*dir*, por ejemplo) se obtiene la información correspondiente en una ventana independiente. También, en lugar de escribir la instrucción, se puede solicitar ayuda mediante el ítem *Help/Stata Command* de la barra de menús. De este modo, aparece un cuadro de diálogo que pide al usuario una orden de Stata y muestra de ella prácticamente toda la información contenida en el manual en una ventana independiente. Ésta posee además varios botones e iconos que realizan operaciones como búsqueda de otras órdenes, exploración de contenidos (*search*) y búsqueda de cadenas en el interior del visor.

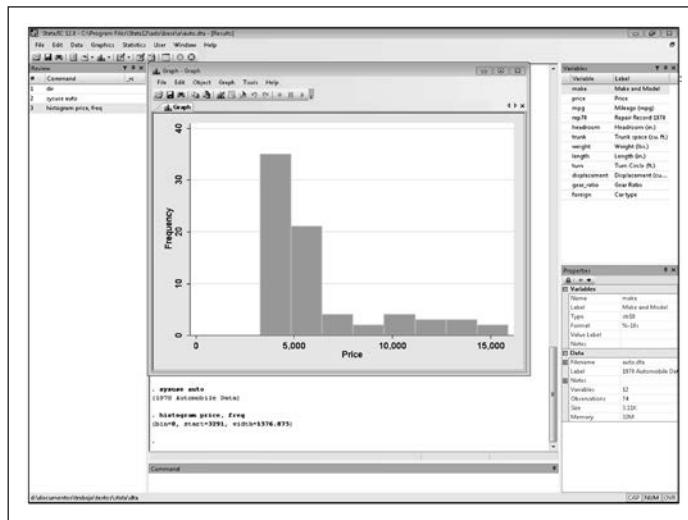
La ventana visor no sólo se emplea para visualizar la ayuda, también puede utilizarse para ver un fichero en ASCII o en formato *smcl*, propio de los resultados grabados de Stata, como se explica con más detalle en la sección 2.5.

Hasta el momento, los textos generados por las instrucciones introducidas en la ventana de órdenes han aparecido en la ventana mayor de Stata, esto es, en la de resultados, o en el visor. Además, hay otro tipo de instrucciones, las gráficas, que muestran su resultado en una ventana diferente. De este modo, si se escribe la siguiente instrucción:

```
histogram price, freq
```

aparece una nueva ventana por encima de la de resultados. Es la *ventana gráfica* de Stata, que se superpone a la anterior. Los resultados pueden ser vueltos a poner en primer plano, pulsando su primera franja. Alternativamente, el gráfico puede aparecer de nuevo pulsando el sexto ícono de la barra de herramientas de la pantalla principal del programa.

**ILUSTRACIÓN 2.10. Ventana de resultados gráficos**



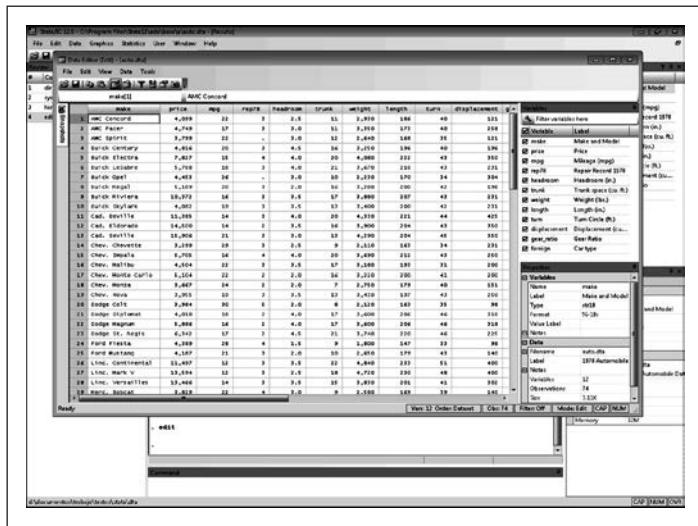
También pueden ser considerados otro tipo de ventanas de Stata los *cuadros de diálogos* que aparecen al solicitar cualquier tarea mediante el menú. Las ilustraciones 2.13 y 2.15 muestran un par de ejemplos de este tipo de ventanas, que tienen la propiedad exclusiva de generar instrucciones de Stata que se acumulan en la ventana de historia y, una vez ejecutadas, muestran su producto en la ventana de resultados. Casos especiales de este tipo de ventanas, presentes sólo a partir de la versiones 11 y 12, son el gestor de variables y el módulo de propiedades, cuyo uso se verá al final de este capítulo.

La utilidad que en Stata crea o modifica la información analizable, conocida como *editor de datos* (*Data Editor*), o la que los inspecciona (*Browser*), generan el octavo tipo de ventana. Su función es mostrar y permitir hacer modificaciones (esta última función sólo en la primera opción) de los datos cargados en la memoria. Como puede apreciarse en la ilustración 2.11, se trata de una ventana, similar a la de una hoja de cálculo, en la que los casos se representan en las líneas y las variables en las columnas. Así, los tres primeros casos corresponden a los modelos *Concord*, *Pacer* y *Spirit* de la casa de automóviles AMC, que tenían en 1979 precios

respectivos de 4.099, 4.749 y 3.799 dólares. En el caso de que se quiera realizar algún cambio, basta llevar el cursor a la casilla correspondiente y reemplazar el valor antiguo con uno nuevo. Tras realizar los cambios deseados, se puede cerrar la ventana con el botón situado en su extremo superior izquierdo que tiene una figura de aspa o, si se prefiere, mantenerla abierta.

edit

**ILUSTRACIÓN 2.11. Ventana del editor**



Varios son los caminos para acceder a la ventana del editor. El más rápido es pulsar el octavo botón de la barra de herramientas (*Data Editor - Edit*). También puede hacerse escribiendo la orden *edit* en la ventana de instrucciones, mediante menú, seleccionando *Data/Data Editor/Edit*, o pulsando la combinación de teclas Ctrl+8. En los cuatro casos anteriores, se permite al usuario realizar modificaciones. Ahora bien, si el propósito es sólo contemplar los casos, sin realizar ningún cambio, es preferible entrar al editor mediante la orden *browse*, pulsando el noveno botón (*Data Editor - Brower*) o seleccionando del menú la entrada *Data/Data editor/Brower*.

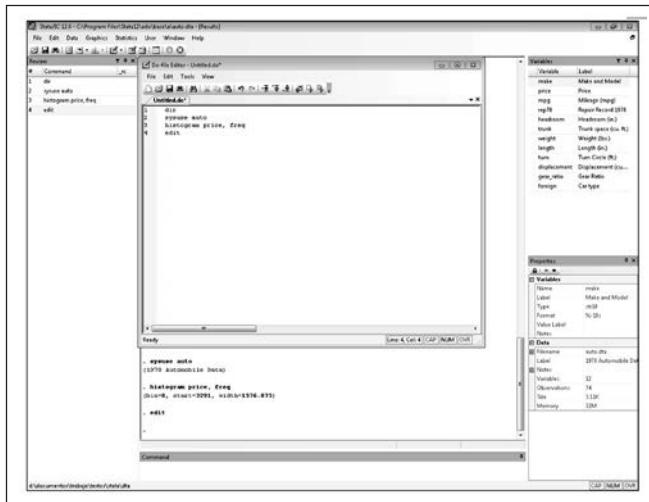
Para finalizar la descripción de *ventanas* de Stata, es preciso explicar de antemano qué es un *programa*, ya que la última que se contempla en este apartado es la de su editor. Un programa de Stata se compone de un conjunto de instrucciones reunidas en un fichero. La forma más cómoda de gene-

rarlo es convirtiendo la historia de instrucciones en un fichero que las contenga. Siguiendo con este primer ejemplo, es decir, las cuatro órdenes para explicar las distintas ventanas de este paquete estadístico, es fácil convertirlas a un fichero de programas haciendo clic con el botón derecho del ratón en la ventana de la historia y señalando la opción *Save All* o *Save Selected*. Tras ello, sale el menú de grabación de ficheros propio del sistema operativo con el que se trabaje y se puede poner el nombre que se desee. Automáticamente Stata le pondrá la extensión *do*, que es con la que se reconoce a este tipo de archivos. Después de grabado, un fichero de programa puede revisarse o ejecutarse cuantas veces se deseé. Otra opción más práctica es mandar este contenido de la ventana de la historia al editor de ficheros de programas de Stata mediante la línea del menú contextual *Send to Do-file editor*.

El editor de estos ficheros también puede ponerse en marcha abriendo su ventana, bien escribiendo en la ventana de órdenes la palabra *doedit*, bien a través de menú (*Window/Do-file Editor*), bien haciendo clic en el séptimo ícono de la barra de herramientas (*Do-file Editor*) o pulsando la combinación de teclas Ctrl+9.

doedit

**ILUSTRACIÓN 2.12. Ventana del editor con el contenido del "Primer programa.do"**



Una vez ejecutado el programa y abierta su correspondiente ventana, hay que abrir el fichero ya guardado (o empezar a escribir uno nuevo)

mediante el menú (*File/Open*), el segundo ícono de las herramientas (*Open*) o bien la combinación de teclas Ctrl+o. Después se selecciona el fichero con extensión *do* deseado en el directorio donde se encuentre y todas las instrucciones de las que se componen aparecerán en la nueva pantalla, de tal modo que aparecen remarcados con diferentes colores, entre otros elementos, las instrucciones, las cadenas, los operadores, las funciones y los comentarios<sup>8</sup>.

Dentro de esta ventana, pueden hacerse cuantas modificaciones se consideren oportunas escribiendo, borrando, copiando, cortando y pegando como en cualquier editor, y ejecutarlas cuantas veces se deseé.

Para esto último hay dos modos: el primero es *do*, en cuyo caso aparecen las órdenes en la pantalla de resultados, y el segundo es *run*, se ejecutan las órdenes pero su contenido y resultados se ocultan. Ambas se encuentran bajo el rótulo del menú *Tools*, y son respectivamente el último (*Execute (do)*) y el penúltimo (*Execute quietly (run)*) ícono de la barra de herramientas propias del editor, cuyo aspecto el lector atento habrá notado diferente del que aparece en la pantalla general de Stata. También es posible realizar la misma operación con las respectivas combinaciones de teclas Ctrl+d o Ctrl+r. Ambos modos funcionan con el conjunto del fichero o con una selección parcial de las órdenes que se consideren más apropiadas para una determinada tarea. Además, con el fin de que todas las modificaciones queden guardadas para uso posterior, también se permite en esta ventana la grabación de su contenido, sea mediante menú (*File/Save*), ícono (el tercero, *Save*), o teclas (Ctrl+s) o (Ctrl+May+s) en el caso de que se quiera dar un nombre distinto al fichero que se graba<sup>9</sup>.

Resumiendo el contenido de este apartado, son once los tipos de ventanas de Stata. Cinco de ellas son internas y aparecen directamente al iniciar el programa: la de *órdenes*, donde el usuario puede ir escribiendo una a una cuantas instrucciones considere relevantes; la de *resultados*, donde aparecerá la ejecución de la instrucción; la de *historia*, en la que se acumularán todas las instrucciones ejecutadas desde el comienzo de la puesta en marcha del programa; la de *variables*, donde se muestra la lista de ellas del fichero de datos que en cada momento se encuentre cargado en memoria y la de *propiedades*, cuyo uso se verá en la sección 2.6.4. Los otros seis tipos de ventanas, las externas, aparecen cuando se realiza una operación que las necesita. En esta categoría se encuentran: el *visor*, que es capaz de mostrar

<sup>8</sup> En la pestaña *Syntax Color* del cuadro de diálogo que aparece con el menú *Edit/Preferencias* del editor de programas aparecen todos los elementos remarcables y se permite cambiar su color y otras propiedades de su fuente.

<sup>9</sup> Otra posibilidad importante para la elaboración, comparación y ejecución de programas en la capacidad de disponer de varios ficheros a la vez en distintas pestañas de la misma ventana. Para lograr abrir más ficheros, se puede pulsar Ctrl+o, el segundo ícono u obtenerlo mediante menú (*File/Open*).

ayuda del programa y resultados grabados; la pantalla de *gráficos*, donde se muestran resultados que no son representables mediante caracteres de texto; el *editor de datos*, para ver o modificar los datos cargados en el programa; los *cuadros de diálogo*, para escribir instrucciones con más facilidad; el *editor de programas*, para la confección, grabación y ejecución de una secuencia de instrucciones que permita resolver peticiones complejas y el *gestor de variables*, que permite asignar o modificar las propiedades de las variables.

## 2.4. Modos de trabajo en Stata

Hay tres formas distintas de proporcionar las instrucciones a la aplicación Stata para obtener los resultados deseados. En este apartado se contempla cómo se emplea cada una de ellas. Aunque prácticamente todo pueda realizarse con estos tres modos de trabajo, la elección de cuál usar dependerá de la tarea que se haga y de las preferencias del usuario. Los tres modos son el de *instrucción*, cuando se introducen literalmente una a una cada orden; el de *menú*, cuando se utiliza un cuadro de diálogo para efectuar una petición, y el de *programación*, en el caso de querer ejecutar automáticamente un conjunto de instrucciones.

### 2.4.1. Modo instrucción

Este modo de trabajo se basa en la inserción manual de instrucciones en la ventana de órdenes. Se caracteriza por ser interactivo ya que cada línea introducida por el usuario genera un resultado y, hasta que este no se complete, no se puede introducir la siguiente orden.

Toda instrucción de Stata está compuesta al menos por una palabra, que es la *orden* propiamente dicha, a veces precedida por una *preinstrucción*, de la que se separa por dos puntos; seguida generalmente por unas *especificaciones*; matizada, si procede, por unos *calificadores*, y ampliada, si cabe, con una serie de *opciones* propias de cada orden, que deben separarse del resto de la instrucción mediante una coma.

La estructura, por tanto, de toda instrucción presenta el siguiente esquema:

[preinstrucción:] orden [especificaciones] [calificadores] [,opciones]
--

Puesto que lo expuesto en corchetes es optativo, por la sintaxis empleada cabe deducir que lo único obligatorio en cada instrucción es la orden. Ahora bien según sea esta, las especificaciones serán obligato-

rias u optativas. Por ejemplo, puede darse la orden *help* sin ninguna especificación, pero no puede emplearse *histogram* seguida de ningún nombre, ya que al menos requiere que se le incluya el de una y sólo una variable.

Fijándose en las cuatro últimas instrucciones que quedan presentes en la ventana de historia de órdenes, éstas eran:

```
help
sysuse auto
histogram price, freq
edit
```

Como fácilmente puede apreciarse, la primera y la última sólo constan de órdenes, mientras que las dos centrales tienen especificaciones. En la segunda la especificación es el nombre del fichero; en la tercera el nombre de una variable. Y la tercera instrucción contiene también una opción *frequency*, que ha sido abreviada<sup>10</sup> con sus cuatro primeras letras. Es fundamental retener que todas las opciones han de figurar detrás de la coma, separadas entre sí al menos por un espacio en blanco.

Además de la orden, sus especificaciones y opciones, la mayor parte de ellas pueden incorporar preinstrucciones, que modifican el funcionamiento de la instrucción, como, por ejemplo, aplicándola a distintas submuestras, y calificadores que restringen el uso de la instrucción a casos con una determinada característica. Todas estas posibilidades son tan importantes que serán tratadas con algo más de detenimiento en el capítulo 5.

#### 2.4.2. Modo menú

El segundo modo de proporcionar instrucciones al programa es mediante los menús. Cuando se habló de la interfaz de Stata y se mencionó la primera zona horizontal de su ventana, se dijo que desde el tercer al quinto de sus apartados (*Data*, *Graphics* y *Statistics*) se podían encontrar prácticamente la totalidad de las instrucciones propias del paquete, mientras que aquellas

<sup>10</sup> Las abreviaturas en Stata pueden emplearse en la inmensa mayor parte de las ocasiones, siempre que no produzcan ambigüedad. De este modo, pueden abreviarse nombres de variables y opciones. Las órdenes sólo pueden recordarse de la forma que se indica en el manual o en la ayuda del programa. Por ejemplo, el programa entiende *hist*; pero interpreta como error *histo* o *histogra*. Sin embargo, en el fichero *auto*, la variable *price* puede ser escrita como *p*, *pr*, *pric* y *price*, ya que ninguna otra variable comienza por *p*. De la misma forma, la opción *frequency* puede ser abreviada, al menos, con *freq*. También entendería bien el programa *frecuenc*.

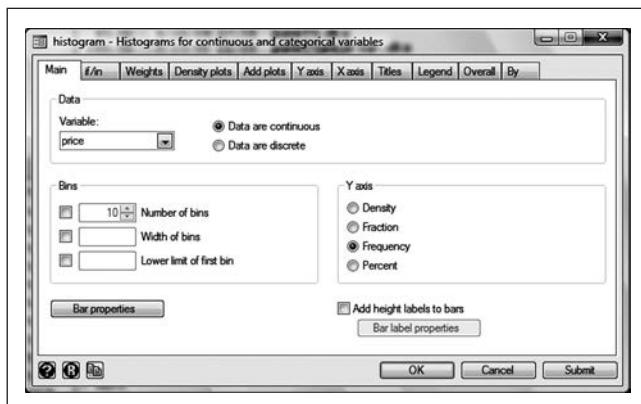
adicionales que el usuario considerara convenientes podían ser añadidas en el siguiente elemento (*User*).

Este modo de trabajo, que facilita la producción de instrucciones a quienes no conocen la sintaxis del programa, fue incorporado a partir de la versión 8, por lo que muchos usuarios anteriores de esta aplicación prefieren seguir utilizando el anterior modo de trabajo, porque es más rápido de usar si se conoce bien. También tratan de evitar este modo de trabajo quienes programan, ya que necesitan recordar continuamente las distintas palabras claves que hay que utilizar en la confección de los programas.

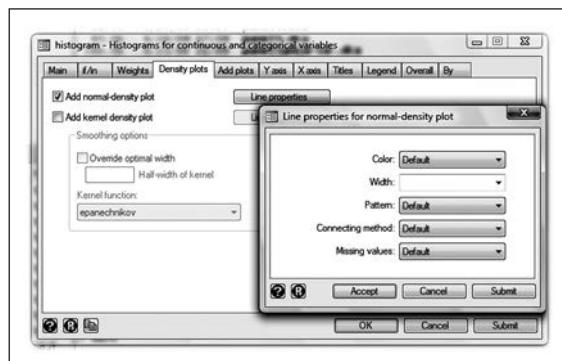
Con el sistema de los menús se han trasladado todas las posibilidades de una orden a un cuadro de diálogo, tanto más complejo cuanto más lo es la instrucción, que se obtiene bien presionando la línea correspondiente del menú, bien escribiendo una orden que lo ponga en funcionamiento.

Por ejemplo, si se desea un histograma de la variable *price*, habrá que optar por especificar *Graphics/Histogram* trasladando el cursor mediante el ratón a los correspondientes elementos que lo componen (primero a *Graphics*, a continuación a *Histogram*) y haciendo clic, una vez obtenido este último. El resultado es un cuadro de diálogo ubicado en una nueva ventana:

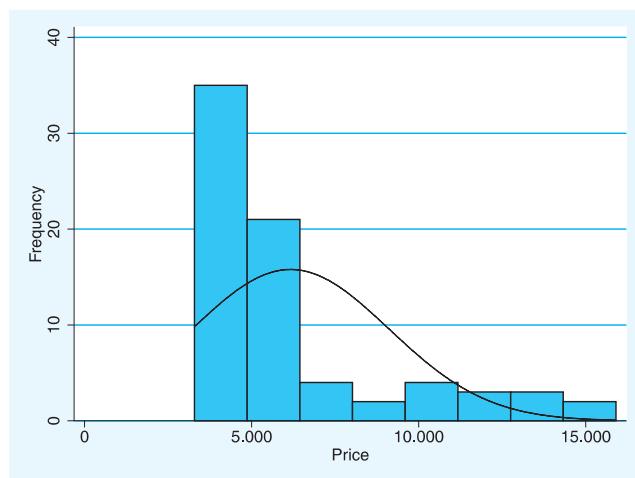
**ILUSTRACIÓN 2.13. Cuadro de diálogo de la orden *histogram***



En la ilustración 2.13 aparece el cuadro de diálogo en el que se ha insertado el nombre de la variable en el correspondiente recuadro (*Variable*) y se ha cambiado la opción *Y-axis*, marcando *Frequency*, en lugar de *Density*, que es con la que opera por omisión. La orden *histogram* posee más posibilidades que las que se muestran en el cuadro de diálogo principal (*main*). Por ello en la línea superior hay otras pestañas referentes de otros conjuntos de opciones. Sólo a modo de ejemplo se muestra a continuación el cuadro de diálogo de la pestaña *Density Plots*:

**ILUSTRACIÓN 2.14. Cuadro de diálogo de la pestaña Density Plots**


Puede apreciarse asimismo que en la línea inferior de cualquier cuadro de diálogo existen seis botones. Tres pequeños en la parte izquierda, el primero (?), para obtener ayuda; el segundo (R), para limpiar el contenido de todos los campos del cuadro y dejarlos en sus opciones por defecto. El tercero es para copiar en el portapapeles el texto de la instrucción que se está construyendo mediante el menú. En la parte derecha, son tres los botones rectangulares con fondo claro. El primero y el último mandan la instrucción: uno, *OK*, cerrando el cuadro de diálogo, el otro, *Submit*, manteniéndolo abierto. El del medio, *Cancel*, sirve para cerrar el cuadro de diálogo sin ningún efecto. Así, pues, tanto el primer como el tercer botón de la parte derecha sirven para obtener un histograma como el mostrado en el gráfico 2.1.

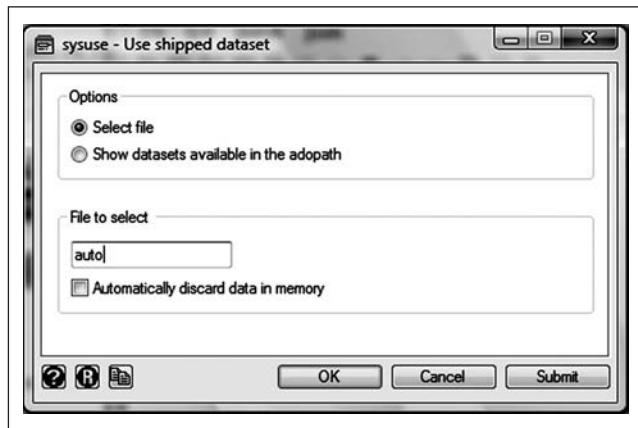
**GRÁFICO 2.1. Histograma del precio de los coches**


Lo más común es utilizar los cuadros de diálogos como se acaba de explicar, es decir, buscándolos a través del menú. Sin embargo, usuarios experimentados de Stata podrían hacerlo de otro modo, empezándolos desde la línea de instrucciones, sobre todo, en el caso en el que las órdenes sean muy complejas y no se recuerden todas sus modalidades y opciones. Para ello hay que escribir *db* seguido de la orden, cuyo cuadro de diálogo se desea obtener. De esta forma, si se quiere que aparezca el cuadro de diálogo de la instrucción *sysuse*, se escribirá una línea como sigue:

```
db sysuse
```

Inmediatamente aparecerá el cuadro de la ilustración 2.15, mucho más simple que el anterior, en la medida de que se trata de una orden con menos parámetros y opciones que la de *histogram*. En este cuadro de diálogo se ofrecen dos posibilidades: abrir un fichero (*Select File*), cuyo nombre hay que escribir en el cuadrado *File to select*, o mostrar todos los ficheros de datos incorporados en el programa disponibles para el usuario (*Show datasets available in the adoptath*). Por último, puede marcarse la opción de descartar datos previamente cargados en memoria (*Automatically discard data in memory*), ya que si existieran unos datos previos modificados con el programa, no podría abrirse el nuevo fichero, a menos que se grabaran las modificaciones o se especificara la opción en cuestión<sup>11</sup>.

**ILUSTRACIÓN 2.15. Cuadro de diálogo de la orden *sysuse***



<sup>11</sup> Stata no permite trabajar con dos ficheros al mismo tiempo. En su lugar, se puede ejecutar el programa varias veces con archivos de datos distintos. Obviamente, no habría interacciones ni intercambios en el trabajo entre ellos.

### 2.4.3. Modo programación

La tercera posibilidad de trabajo con Stata es el modo de programación, que consiste en escribir una serie de instrucciones necesarias para llevar a cabo una tarea, grabarlas en un fichero y desde este ejecutarlas cuantas veces se desee con o sin cambios en las órdenes que lo necesiten. En el apartado anterior, cuando se describió la ventana del editor de programas, se explicó que hay dos formas de ejecución de estos ficheros: uno, mediante la instrucción *run*, en cuyo caso no se muestran las líneas de instrucción; el otro, mediante la instrucción *do*, para obtener el mismo resultado, pero con las órdenes incluidas. Existe un tercer modo de ejecutar un programa escrito con el lenguaje de Stata. Se trata de hacerlo desde el sistema operativo. Cualquier fichero con extensión *do*, con sólo aplicarle un doble clic, es capaz de cargarse con Stata y ejecutarse.

Si además se incluye en el fichero la instrucción *set more off*, el usuario puede desentenderse del proceso y el programa ejecutará sin interrupción todas las instrucciones incluidas. También pueden añadirse comentarios: a) poniendo un asterisco al inicio de cualquier línea, b) colocando los comentarios entre /\* y \*/, c) después de espacio seguido de dos barras hasta el final de la línea y d) del mismo modo que en c) pero con tres barras, en cuyo caso la línea siguiente será considerada de la misma orden. Esta última opción es un modo óptimo en los programas para disponer en varias líneas las instrucciones muy largas.

Por ejemplo, si con el editor preferido<sup>12</sup> se escribe un fichero con las siguientes instrucciones llamado *listauto.do*:

#### ILUSTRACIÓN 2.16. Contenido del fichero *listauto.do*

```
*****
* M. Escobar, E. Fernández, F. Bernardi
* Análisis de datos con Stata
* Madrid. CIS. 2009
* Ejemplo de primer programa (listauto.do)
*****  

set more off //Esta instrucción sirve para que no se pare la pantalla.
sysuse auto, clear /*Lee uno de los ficheros ejemplos de Stata*/
list make /// Con tres barras entiende que la orden continúa en la siguiente línea.
    price
set more on //Vuelve a parar la pantalla de resultados cuando se llene.
```

<sup>12</sup> Incluso puede emplearse un procesador de texto, siempre y cuando a la hora de grabar se tenga la precaución de grabar el fichero en formato ASCII y se le ponga la extensión .do.

Se obtendrá un listado de las dos variables expresadas (*make* y *price*) para todos los casos del fichero *auto* del que a continuación se ofrece un extracto, escribiendo la instrucción:

```
do listauto
```

### ILUSTRACIÓN 2.17. Listado de casos

	make	price
1.	AMC Concord	4,099
2.	AMC Pacer	4,749
3.	AMC Spirit	3,799
4.	Buick Century	4,816
5.	Buick Electra	7,827
6.	Buick LeSabre	5,788
7.	Buick Opel	4,453
8.	Buick Regal	5,189
9.	Buick Riviera	10,372
10.	Buick Skylark	4,082
11.	Cad. Deville	11,385
...		
66.	Subaru	3,798
67.	Toyota Celica	5,899
68.	Toyota Corolla	3,748
69.	Toyota Corona	5,719
70.	VW Dasher	7,140
71.	VW Diesel	5,397
72.	VW Rabbit	4,697
73.	VW Scirocco	6,850
74.	Volvo 260	11,995

Este modo de programación puede hacerse tan flexible como se quiera, tanto por la posibilidad de intercambiar parámetros (enviar al programa información, que luego es devuelta para su uso en la ventana principal) como por la de incluir instrucciones de control de flujo, que permite poner en manos del usuario la posibilidad de escribir con un lenguaje sencillo sus propias rutinas, superando de este modo la rigidez que imponen otras aplicaciones estadísticas que no permiten obtener estadísticos distintos de los que ya vienen preprogramados en el paquete<sup>13</sup>.

---

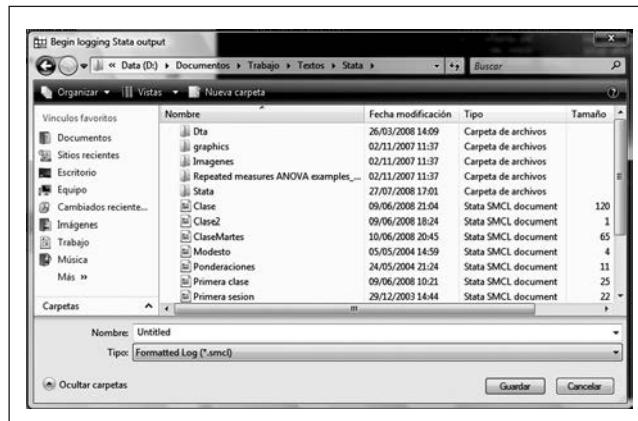
<sup>13</sup> Quienes estén interesados en este tipo de procesos pueden consultar los capítulos 16-18 de la guía del usuario (Stata Corporation, 2011c) y el volumen del manual de Stata dedicado a la programación (Stata Corporation, 2011i).

## 2.5. El fichero de resultados

Quien haya trabajado una larga sesión con Stata y haya querido volver a mirar los resultados de los primeros análisis habrá advertido que la pantalla de resultados tiene una capacidad limitada, pues no es capaz de almacenar más allá de una determinada cantidad de caracteres (200.000 por defecto en la versión 12 de Stata), aunque puede ser ampliada en *Edit/Preferences/General Preferences/Windowing* hasta 2 millones. En trabajos largos esto es un enorme inconveniente. Además, cualquiera que sea la longitud de los resultados, Stata no es capaz de grabarla mediante una instrucción. Si se desea guardar total o parcialmente su contenido, es preciso marcar el bloque deseado, y optar mediante menú de cabecera (*Edit/Copy Text*) o menú contextual (*Copy Text*) su traslado a otro programa, un procesador de texto, por ejemplo<sup>14</sup>.

Sin embargo, el modo en que Stata ha previsto que no se tenga que realizar esta tarea de cortar y pegar, cada vez que se genera un resultado que se quiera guardar, es mediante la grabación de la pantalla de resultados en un fichero. Esta operación no es automática y ha de ser el usuario quien inicie el proceso, lo detenga, lo continúe o lo cierre.

**ILUSTRACIÓN 2.18. Pantalla de inicio de ficheros de resultados**



Como otras operaciones frecuentes de Stata, la creación de un fichero de resultados se puede realizar de cuatro modos: mediante instrucción inte-

<sup>14</sup> Caso de que se copie un resultado de Stata a un procesador de texto, es imprescindible darle una fuente con tipo de letra de espacios fijos (Courier o Lucida), ya que las fuentes proporcionales (Times, Arial, Century, entre otras muchas) producen textos de tamaño variable y, por tanto, no generan textos o números alineados verticalmente.

ractiva o programada (*log using*), con icono (el cuarto de la barra de herramientas, *Log begin*), teclado (Ctrl+L) o mediante menú (*File/Log/Begin*). Con las tres últimas modalidades, aparece una ventana (ilustración 2.18) con un listado de ficheros con extensión *smcl*, que son aquellos en los que Stata guarda sus resultados con un formato propio, en todo momento convertibles a ficheros con formato plano en ASCII. En la mencionada ventana, debe escribirse en la casilla *Nombre* el título que se quiera dar al fichero donde a partir de ese momento se grabarán todos los resultados. También puede elegirse el formato de este fichero. Aunque, si nada se indica, Stata utiliza su formato propio (*smcl*); se puede cambiar desde el principio de la grabación, optando en la casilla *Tipo* por la extensión *log*.

Esta operación también puede hacerse mediante instrucción, sea en la pantalla de órdenes, sea en un programa. Por ejemplo, si se desea generar un fichero de resultados llamado *primero.smcl*<sup>15</sup>, habrá que escribir la siguiente línea<sup>16</sup>:

```
log using primero
```

Si no existe ya ese fichero en el directorio actual de trabajo y si no se ha abierto con anterioridad algún otro fichero de resultados, aparecerá en la ventana de resultados un texto que advierte la operación realizada:

#### ILUSTRACIÓN 2.19. Cabecera de la apertura de un fichero de resultados

```
log: C:\Documents and Settings\...\Mis documentos\stata\primero.smcl
-----
log type: smcl
opened on: Jan 2004, 12:37:07
```

A partir de este momento, todo lo que aparece en la ventana de resultados, salvo la ayuda, será grabado en el fichero, directorio y disco del ordenador especificado. La grabación puede ser revisada, suspendida o finalizada. Si se intenta hacer cualquiera de estas operaciones mediante menú (*File/Log*), mediante icono de la barra de herramientas (*Close/Suspend*), o con el teclado (Ctrl+L), aparece un cuadro de diálogo para que el usuario opte por la fórmula deseada.

<sup>15</sup> Caso de que el fichero contenga espacios en blancos, es obligatorio que su nombre sea escrito entre comillas.

<sup>16</sup> Desde la versión 12 se puede añadir como opción un nombre interno al fichero de resultados [, *name(nombre\_interno)*]. De este modo se puede tener abierto más de uno al mismo tiempo. El control del uso sería con la instrucción *log [offon] nombre\_interno*. También son útiles las opciones *replace* y *append*, que sirven respectivamente para regenerar un fichero ya existente o para añadirle los nuevos resultados.

**ILUSTRACIÓN 2.20. Cuadro de diálogo para un fichero de resultados ya abierto**



Todas esas operaciones también pueden realizarse mediante instrucciones en la ventana de órdenes. Estas son:

```
view nombrefichero.smcl
log close
log off
```

Al igual que se puede grabar un fichero de resultados, también puede hacerse algo similar con todas las instrucciones de una sesión de Stata. Ya se ha visto cómo puede hacerse *a posteriori*, haciendo aparecer el menú de contexto en la ventana de historia. Pero también puede realizarse *a priori*, mediante la instrucción *cmdlog using* nombrefichero<sup>17</sup>. Si se quiere crear un fichero llamado primeras instrucciones, con extensión *do* habrá que escribir la siguiente orden.

```
cmdlog using "primeras instrucciones.do"
```

Y para suspender, reanudar o terminar la grabación, ha de usarse la instrucción *cmdlog* acompañada de *off*, *on* o *close* respectivamente.

Tanto en ficheros de resultados como en ficheros de instrucciones, otro aspecto que ha de tenerse en cuenta es que, en el caso de que se quiera dar un nombre de fichero ya existente, si se intenta con una línea de instrucción, dará un error, a menos que se añada la opción *append*, si se quiere

---

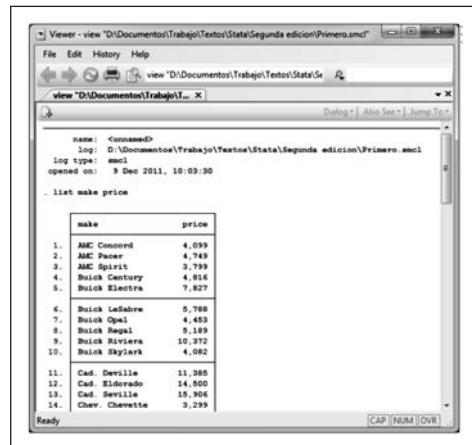
<sup>17</sup> Si no se especifica ninguna extensión al fichero, el programa le pondrá .txt. Si se desea repetir una sesión, conviene cambiarle la extensión por la de .do. De este modo podrán ejecutarse de nuevo.

añadir lo nuevo a lo existente, o *replace*, si se prefiere suplantar el antiguo contenido por el nuevo.

En cualquier momento, por otro lado, se puede visualizar cualquier fichero de resultados, comandos e incluso de ayuda con la instrucción *view* nombrefichero. Esta posibilidad se encuentra también en el menú *File/View*. Si, a continuación, se aprieta en el botón *Browse*, aparece el explorador de ficheros, desde el que puede seleccionarse cualquier archivo escrito en ASCII o en SMCL para su visualización<sup>18</sup>. En cualquier caso, también puede llamarse directamente al fichero si en la instrucción *view* se especifica su nombre y extensión.

```
view primero.smcl
```

**ILUSTRACIÓN 2.21.** Visor de un fichero de resultados



## 2.6. Las variables de la matriz de datos

Como la mayoría de programas informáticos, Stata es básicamente una herramienta para manipular datos: datos estadísticos en este caso. El funcionamiento de Stata consiste en manipular, modificar y realizar operaciones matemáticas sobre una matriz de datos que se almacena en la memoria del ordenador. Para que Stata pueda trabajar con estos datos, tienen que estar en el formato adecuado, en el formato de Stata. En el próximo capítulo, se verá cómo construir o traducir los ficheros propios de este programa, bien

<sup>18</sup> Ha de tenerse en cuenta que se permite la visión de varios ficheros al mismo tiempo, tanto en nuevas ventanas como en nuevas pestañas (*Tab*).

sea introduciendo los datos manualmente o traduciéndolos desde ficheros construidos por otros programas (como SPSS o Excel) al formato de Stata. Pero, primero, es conveniente explicar operaciones más fáciles como su uso, el formato y la disposición de datos que en ellos se contiene.

Al igual que la mayor parte de programas estadísticos, Stata trabaja con un fichero de datos estructurado por casos y variables. Este fichero ha de residir en un determinado directorio de una unidad del ordenador. Para hacer uso de él y aplicarle cuantas operaciones estadísticas se deseé, ha de cargarse en la memoria central de la máquina. Ello se consigue abriendo el fichero mediante el menú *File/Open*, el primer ícono de la barra de herramientas (*Open*), la combinación de teclas Ctrl+O, o la instrucción *use* nombrfichero. Anteriormente se ha utilizado la instrucción *sysuse*, pero esta sólo sirve para ficheros suministrados por el programa para mostrar ejemplos incorporados en el manual. Si se quiere, en consecuencia, abrir un fichero previamente creado por el usuario, como puede ser el fichero con información de países, al que se tituló con el nombre “mundo”, la instrucción debería incluir la opción *clear*, si se desea descartar<sup>19</sup> posibles modificaciones realizadas en un conjunto de datos cargados en memoria.

```
use mundo, clear
```

A partir de ese momento, si el programa ha localizado correctamente el fichero, la matriz de datos se carga en la memoria del ordenador y, como prueba de ello, aparecen las variables en su ventana correspondiente, siempre que se encuentre abierta. Como se ha visto anteriormente, en la pantalla del editor de datos, esta matriz se representa habitualmente como una tabla de datos en la que las filas son casos y las columnas variables (así se representa por ejemplo en el editor de datos de Stata, como se verá con atención en el siguiente apartado). Los casos son los individuos de los que se dispone información, y las variables son las categorías en las que se organiza esa información. En este ejemplo, al disponer de un conjunto de indicadores de países, los casos serán cada uno de los países incluidos, y las variables cada uno de los indicadores disponibles como la superficie, la población, el PIB, etc.

Para conocer las características de las variables que están contenidas en un fichero, Stata posee una instrucción que construye una lista de todas ellas, salvo que se especifique tras la instrucción un subconjunto

<sup>19</sup> Si se usa una versión de Stata anterior a la 12, podría ocurrir que el fichero fuera demasiado grande para que cupiera en la memoria que Stata reserva por defecto (10 Megabytes) al iniciarse el programa. Con la orden *set mem #M*, el usuario puede cambiar la capacidad reservada. A partir de dicha versión, el usuario no tiene que preocuparse de la ampliación de la gestión de la memoria, puesto que el programa la gestiona automáticamente.

de variables. Esta orden es *describe*. Usada sin argumento ni opción, proporciona de cada variable información sobre el tipo de almacenamiento, el formato de presentación y las etiquetas. Aplicada sobre el fichero *mundo* del presente ejemplo, muestra el resultado presentado en la ilustración 2.22.

Se advierte en primer lugar que el fichero consta de 213 observaciones, cada una de ellas corresponde a un país. También se indica que está compuesto por 17 variables y que el tamaño que ocupa en disco (y actualmente en memoria) es de 20.661 bytes.

En la ilustración 2.22 se lista el conjunto de variables disponibles en el fichero, ahora cargadas en la memoria del programa. Cada una de ellas aparece, junto con su nombre, con su tipo de almacenamiento, su formato de presentación y sus etiquetas, conceptos todos ellos que se explican en los tres próximos apartados. De estas propiedades de las variables, la más sencilla, útil, necesaria y empleada es el etiquetaje. Por ello, se aborda en primer lugar. Las otras dos, el tipo y el formato, son más complejas y no tan necesarias, por lo que si no se entienden en un primer momento, el lector puede continuar con los siguientes capítulos de este libro sin temor a perderse algo imprescindible.

### ILUSTRACIÓN 2.22. Descripción de variables

Contains data from mundo.dta				Indicadores de los países. Mundo (2002)
obs:	213	Fuente: The World Bank		
vars:	17			13 Aug 2009 20:02
size:	20,661 (99.9% of memory free)			
variable	storage	display	value	label
name	type	format	label	variable label
pais	str24	%-24s		Pais
capital	str19	%19s		Capital
continente	byte	%7.0f	conti	Continente
ocde	byte	%2.0f	perte	Pertenece a la OCDE
fiocde	float	%d..		Fecha de ingreso en la OCDE
ue	byte	%5.0f	perte	Pertenece a la UE
fiue	float	%d..		Fecha de ingreso en la Unión Europea
superficie	double	%12.0fc		Superficie
poblacion	float	%9.3fc		Población
densidad	float	%8.0fc		* Densidad
evn	float	%3.0f		Esperanza de vida al nacer
tmi	int	%8.0g		Tasa de mortalidad infantil
anal	byte	%8.0g		Tasa de analfabetismo
tascrec	float	%6.2f		Tasa de crecimiento
pib	long	%12.0fc		Producto interior bruto (mil \$)
rnbpc	long	%9.0fc		Renta per cápita (\$)
rnbppa	long	%9.0fc		Renta per cápita (poder de compra)
				* indicated variables have notes
Sorted by:				

### 2.6.1. Etiquetas de variables y de valores

En Stata pueden asignarse etiquetas a la base de datos, a las variables y a los valores. Estas etiquetas harán más fácil la comprensión de los análisis estadísticos, por lo que es conveniente ponerlas.

La instrucción general para etiquetas es *label*, tras la cual se especifica qué es lo que se quiere etiquetar y la etiqueta. Para poner una etiqueta a la base de datos, ha de escribirse *label data* y la etiqueta que se deseé:

```
label data "Indicadores de los países. Mundo (2000)"
```

Las etiquetas de las variables se ponen con la instrucción compuesta *label variable*:

```
label variable pib "Producto interior bruto"
```

Como puede deducirse, la etiqueta debe ir entrecomillada, obligatoriamente si tiene espacios en blanco.

Un poco más complejo es poner etiquetas a los valores. Las etiquetas de valores se definen por listas, y luego se asignan a las variables deseadas. Esto permite que se asigne una misma lista de etiquetas de valores a varias variables con iguales respuestas. Por ejemplo, es posible asignar al mismo tiempo etiquetas a los valores de un conjunto de preguntas que tengan las mismas posibilidades de respuesta, como muy de acuerdo, de acuerdo, en desacuerdo y muy en desacuerdo.

Lo primero que debe hacerse es definir una lista de etiquetas de valores, al que se denominará *conti* (por continente). Esto ha de hacerse del siguiente modo:

```
label define conti 1 "Europa" 2 "Asia" 3 "África" 4 "América" 5 "Oceanía"
```

Como puede apreciarse, tras la orden *label* va la especificación *define*, el nombre de la lista de etiquetas de valores y luego los valores seguidos por sus respectivas etiquetas. Una vez que se introduzca esta instrucción, la lista de etiquetas de valores quedará en memoria junto con los datos. Si se guardan estos, las listas de etiquetas quedarán también grabadas, de tal modo que se recuperan en sesiones subsiguientes de Stata.

Una vez definida una lista de etiquetas, se puede asignar a tantas variables como se deseé, en cualquier momento. En este caso concreto, la instrucción es:

```
label values continente conti
```

Este procedimiento tiene importantes ventajas cuando hay varias variables con el mismo tipo de etiquetas, como sucede con las variables “ocde” y “ue”, ambas relacionadas con la pertenencia o no a estos organismos. Para etiquetar sus valores, primero se definen las etiquetas y después se asignan a cada una de las variables de este modo:

```
label define perte 0 "No" 1 "Sí"
label values ocde perte
label values ue perte
```

En cualquier momento, el usuario puede ver las listas de etiquetas que están definidas en un determinado conjunto disponible de datos. Para ello hay que utilizar la orden *labelbook*, o para un resultado más escueto *label list*, que sólo muestra códigos y etiquetas para cada lista.

```
label list
```

### ILUSTRACIÓN 2.23. Lista de etiquetas

```
contí:
      1 Europa
      2 Asia
      3 África
      4 América
      5 Oceanía
perte:
      0 No
      1 Sí
```

Otras instrucciones útiles para trabajar con etiquetas son *label drop* (que elimina las listas de etiquetas que se declaran expresamente) y *label save* (que guarda la definición de las etiquetas en el archivo .do que se indique). Tecleando *label values* seguido sólo por un nombre de variable, se quitarán las asignaciones que tuvieran sus valores a una lista de etiquetas.

```
label values ocde
```

Mediante la anterior instrucción, la etiqueta *perte* dejará de estar asignada a la variable *ocde*. La lista de etiquetas *perte* no se borra de la base de datos, sólo deja de estar asignada a *ocde* (seguirá asignada a la otra variable asignada, a *ue*). A menos que se escriba la orden específica para hacerlo (con *label drop*), si una etiqueta no está asignada a ninguna variable no desaparece del archivo, con lo que puede ser usada siempre que se requiera.

### 2.6.2. *Tipos de almacenamiento de las variables*

En Stata, cada variable tiene un formato según el tipo de datos que contenga. Los valores de las variables pueden componerse de: a) una cadena de caracteres (string), b) números o c) fechas. Estos son los tres tipos principales de variables en Stata.

Las variables numéricas y de cadena no sólo contienen la información de su tipo, sino también el tamaño máximo de dígitos de la variable. En las variables de cadena la norma y el procedimiento son sencillos: automáticamente Stata asignará a cada variable el tipo *str* y el número de caracteres que contenga el conjunto de caracteres más largo. Por ejemplo, en una variable que incluyera las provincias españolas, la cadena de caracteres más larga sería Santa Cruz de Tenerife, que se escribe con 22 caracteres, por lo que Stata asignaría a la variable provincia la extensión de 22: el tipo sería *str22*.

Para las variables numéricas es algo más complicado, puesto que el tipo no depende directamente del número de dígitos sino del valor máximo: desde *byte*, que puede almacenar desde el valor -127 hasta el 126; hasta *double*, que puede almacenar desde el número  $-9.0^{38}$  hasta el  $9.0^{307}$ . En el cuadro 2.2 pueden verse los distintos nombres y características de los tipos de variables numéricas que utiliza Stata.

**CUADRO 2.2. Tipos de almacenamiento de variables numéricas<sup>20</sup>**

Tipo variable	Valor mínimo	Valor máximo	Valor más cercano a 0 (sin ser 0)	Bytes
byte	-127	100	+/- 1	1
int	-32.767	32.740	+/- 1	2
long	-2.147.483.647	2.147.483.620	+/- 1	4
float	$-1,7 \times 10^{38}$	$1,7 \times 10^{38}$	+/- $10^{-38}$	4
double	$-9,0 \times 10^{307}$	$9,0 \times 10^{307}$	+/- $10^{-323}$	8

Fuente: Stata Corporation (2011c: 110).

En principio, no es necesario preocuparse por el tipo de las variables, porque Stata asigna automáticamente el tipo adecuado, e incluso lo cambia si es necesario (si se introduce un valor mayor que el máximo). Por ejemplo, si se añade un caso con el valor 101 en una variable *byte* (que puede alma-

<sup>20</sup> Los valores mínimos y máximos de las variables flotantes y dobles han sido reducidos a un decimal para simplificar su exposición. El valor exacto puede consultarse en la sección 12.2.3 de la guía del usuario.

cenar hasta el valor 100, véase el cuadro 2.2), automáticamente Stata cambia el formato de la variable y la convierte en *int*, que sí puede almacenar un valor mayor que la centena. Pero, aunque no sea necesario asignar directamente el formato a las variables, porque Stata ya lo hace automáticamente, puede ocurrir que los formatos sean demasiado grandes para los datos, por lo que estos ocupen demasiada memoria. Por ejemplo, en el supuesto de que a una variable de edad, para la que en principio el formato *byte* (que puede almacenar de -127 a 100) es más que suficiente, por un error en la introducción de datos, se introdujera el valor 195; automáticamente, Stata cambiaría el tipo de la variable a *int*. Si luego se advierte el error y se corrige (cambiando el valor a 19), el tipo de la variable seguirá siendo *int*, aunque los datos que almacena la variable edad no superen en ningún caso la centena (y por tanto el tipo *byte* es suficiente). Esto es así porque Stata modifica el formato al alza pero no a la baja: cuando se introduce un valor superior al máximo permitido por el tipo de variable, cambia el formato para que se pueda almacenar el valor correctamente, pero a menos que se especifique nunca cambia el formato, si se reduce el valor máximo. Esto es importante porque el formato de la variable determina cuánta memoria ocupan los datos. Si los formatos de los datos son mayores de lo necesario, puede que la matriz ocupe tanto que no quepa en la memoria de trabajo y, en consecuencia, su proceso se ralentiza por la necesidad de usar el disco duro como memoria virtual.

La instrucción *compress* está específicamente diseñada para este problema. Cuando se introduce esta orden, Stata comprueba uno a uno los formatos de todas las variables de la matriz y asigna a cada una de ellas el formato más pequeño posible. Es una instrucción que nunca modifica los contenidos de la matriz, sólo el tipo de las variables. Caso de que este ya sea tan pequeño como posible, no modificará nada. Pero, en muchos casos, este comando puede hacer mejorar ostensiblemente el funcionamiento de Stata, al reducir el tamaño que ocupan los datos en memoria.

Finalmente hay que conocer el peculiar modo con que Stata trata las fechas. Estas pueden ser una variable de texto: "21 Mar 1952", un conjunto de tres variables numéricas, 21 para el día, 3 para el mes y 1952 para el año, o una sola variable numérica, en cuyo caso se necesita una referencia, una fecha de partida que represente el valor 0, que en Stata es el 1 de enero de 1960. En cualquier caso, para que este programa las trate como variable de fecha, especialmente en los análisis de series temporales, sólo es válida la última forma de almacenamiento.

Resulta evidente que el usuario no va a introducir la variable de fecha según los días que hayan transcurrido desde comienzos del año 1960. Para la conversión se dispone de muchas funciones que permiten tanto pasar del formato usual al modo de trabajo como a la inversa.

Una de las operaciones más empleadas en este sentido es la de proporcionar un determinado formato de presentación a una variable. De este

modo, una variable con el valor numérico temporal de 2 puede aparecer literalmente como “3 january 1960”, si se le indica dicho formato tal como se señala en el próximo apartado.

### 2.6.3. *Formatos de presentación de las variables*

Como en la mayor parte de las aplicaciones informáticas, hay que distinguir entre el modo en el que son almacenados los valores de las variables que presentan los casos y el formato en el que son presentados en la pantalla. El primero está determinado por el tipo de almacenamiento, mientras que el segundo es el que es denominado formato de presentación.

Una instrucción para la que es importante la utilización de los formatos de presentación es *list*. Como ya se ha visto, su función es la de mostrar los valores que tienen los casos en unas determinadas variables. El modo de cambiar la presentación de las variables es mediante otra orden anterior a la mencionada. Se trata de *format*, que ha de presentar la siguiente estructura:

```
format listaveriables %formato
```

Donde aparece listaveriables, ha de figurar una o varias variables mediante las convenciones propias del programa y donde aparece %formato se especifica mediante claves el aspecto con el que se desean mostrar las variables. Desde el punto de vista del formato también es útil distinguir los tres tipos de variables: las numéricas, las textuales y las de fecha.

- 1) Las *variables numéricas* pueden presentarse a su vez de acuerdo a una de las siguientes modalidades:  
 %p.dg, para mostrar todo tipo de formatos.  
 %p.df, para mostrar formatos de un número determinado de decimales.  
 %p.de, para mostrar los números en notación científica.

donde p significa el número de posiciones que se desean obtener de un determinado número y d expresa el número de decimales que se quieren mostrar. A todas ellas se le puede añadir una c, si se desea que se añada una coma cada tres dígitos para mejorar la legibilidad de las cifras largas. Por defecto, Stata emplea los puntos para expresar los decimales y las comas para los millares. Esto puede cambiarse con la instrucción set dp comma.

En realidad, sólo existen dos tipos de formatos para los números, el decimal (*f*) y el científico (*e*). El formato (*g*) hace que sea el mismo programa quien se encargue de seleccionar la prestación más adecuada según las características del número mostrado.

Para que quede mejor aclarado es imprescindible un buen ejemplo con distintos tipos de formato. De la base de datos de los países del mundo se han seleccionado las variables *superficie*, *pib*, *tmi* y *tascrec*. El formato respectivo de cada una de ellas es el siguiente: tanto en *superficie* como en *pib* se han puesto doce caracteres sin ningún decimal. En ambas también se ha añadido el carácter “c” a fin de mejorar la legibilidad. En la tasa de mortalidad se ha fijado un formato de 3 posiciones sin ningún decimal (está medida en tantos por mil). Y, finalmente, la tasa de crecimiento se expresa en formato de seis posiciones y dos decimales. Toda esta información está guardada en el fichero, por lo que no es necesaria su introducción, a menos que se desee cambiar.

Un listado de los diez primeros países de las variables con los formatos antedichos puede solicitarse con la siguiente instrucción:

```
list pais superficie pib tmi tascrec in 1/10
```

Nótese que a la orden *list* se le ha añadido la lista de variables y la partícula *in* seguida de 1/10, lo que significa desde el caso primero hasta el décimo.

#### ILUSTRACIÓN 2.24. Listado parcial de casos

	país	superfici~e	pib	tmi	tascrec
1.	Afganistán	652,090	4,100	163	2.60
2.	Albania	27,400	4,114	20	0.40
3.	Alemania	349,300	1,873,854	4	0.30
4.	Andorra	500	950	.	.
5.	Angola	1,246,700	9,471	128	3.10
6.	Antigua y Barbuda	400	640	16	1.40
7.	Antillas Holandesas	800	2,360	13	0.86
8.	Arabia Saudí	2,149,690	173,287	18	2.80
9.	Argelia	2,381,700	53,009	33	1.90
10.	Argentina	2,736,700	268,773	17	1.30

En cambio, si antes de realizar ese mismo listado se le hubiera cambiado el formato de un modo similar a este:

```
format %8.0g superficie pib tmi tascrec
```

el resultado hubiera sido diferente en las variables *superficie*, *pib* y *tascrec*.

**ILUSTRACIÓN 2.25. Listado formateado de casos (I)**

	país	superficie	pib	tmi	tasrec
1.	Afganistán	652090	4100	163	2.6
2.	Albania	27400	4114	20	.4
3.	Alemania	349300	1.9e+06	4	.3
4.	Andorra	500	950	.	.
5.	Angola	1.2e+06	9471	128	3.1
6.	Antigua y Barbuda	400	640	16	1.4
7.	Antillas Holandesas	800	2360	13	.86
8.	Arabia Saudí	2.1e+06	173287	18	2.8
9.	Argelia	2.4e+06	53009	33	1.9
10.	Argentina	2.7e+06	268773	17	1.3

En estos resultados, en los que se ha aplicado a todas las variables numéricas el formato `%8.0g`<sup>21</sup>, se puede ver que hay casos de las variables *superficie* y *pib* —Angola en la primera y Alemania en la segunda, entre otros— que son mostrados en notación científica, pues de otro modo no cabrían en los 8 espacios. Por otro lado, en la variable de la tasa de crecimiento se detecta otra de las peculiaridades del tratamiento del formato que hace Stata. Es de notar cómo, a pesar de que estén puestas en formato de cero decimales, Stata los muestra todos siempre y cuando estén así almacenados, sin efectuar operación de redondeo como lo hubiera hecho en el caso de haber utilizado el formato `%p.df` de presentación. Por ello, en el caso de las Antillas Holandesas aparecen dos decimales. En el caso de que hubiera algún país sin decimales, obviamente no los mostraría.

- 2) Los formatos de las *variables textuales* también pueden ser cambiados. Pero, en lugar de utilizar los caracteres *f*, *g* o *e*, hay que emplear *s*, abreviatura de *string*. Obviamente, en este caso, no han de aparecer cifras decimales; en cambio, es muy útil utilizar la opción del alineamiento a la izquierda, que se logra mediante el signo menos delante de la cifra que indica el número de posiciones necesarias para la presentación del texto. De este modo, con las dos siguientes instrucciones, se mostraría los cinco primeros países acompañados de su correspondiente tasa de mortalidad infantil:

```
format %-24s país
list país tmi in 1/5, clean
```

<sup>21</sup> Por defecto Stata asigna el formato `%8.0g` a todas las variables almacenadas como *byte* o *integer*, con `%9.0g` a las variables *float*, `%10.0g` a las *double* y `%12.0g` a las *long*.

Es preciso notar cómo en los resultados de la ilustración 2.26, el tamaño de la columna de la variable *país* no contiene 24 columnas. A menos que se le especifique la opción *fast*, el programa examina la longitud de los casos que va a mostrar y automáticamente ajusta el tamaño de la columna al máximo de caracteres. Otra opción interesante de la orden *list* es *clean*, que hace que en el listado los casos no queden separados por líneas horizontales.

#### ILUSTRACIÓN 2.26. Listado formateado de casos (II)

país	tmi
1. Afganistán	163
2. Albania	20
3. Alemania	4
4. Andorra	.
5. Angola	128

- 3) Finalmente, hay que referirse a los complejos formatos de *variables de fecha (date)*. Como se dijo anteriormente, Stata almacena los datos relacionados con fechas como un número de tal modo que el 0 representa el 1 de enero de 1960. Un valor negativo es una fecha anterior a la mencionada y todo positivo, en correspondencia, posterior. Así, como dicho año fue bisiesto, el número 366 equivale al 1 de enero de 1961 y el número -365 representa el primer día del año 1959.

Todo ello es fácilmente apreciable utilizando la orden *display*, que muestra el contenido de una variable o constante con la posibilidad de aplicarle un formato temporal. De este modo, si se escribe la siguiente línea en la ventana de órdenes:

```
display %d -365, %d 0, %d 366
```

se mostrará en la pantalla de resultados tres fechas consecutivas correspondientes a los primeros días de los años 1959, 1960 y 1961.

#### ILUSTRACIÓN 2.27. Exposición de fechas (I)

```
01jan1959 01jan1960 01jan1961
```

Como puede apreciarse, las fechas se muestran en inglés con dos dígitos para el día, seguidos por tres caracteres para el mes y cuatro dígitos para el año. Esto también puede ser cambiado al especificar el formato *%d* seguido con una lista formada por una combinación de las siguientes convenciones.

### CUADRO 2.3. Formatos de fecha

c / C	Muestra el siglo sin/con ceros a la izquierda
y / Y	Muestra los dos dígitos del año sin/con ceros a la izquierda
m / M	Muestra el mes en mayúscula abreviado con 3 letras/sin abreviar
l / L	Muestra el mes en minúscula abreviado con 3 letras/sin abreviar
n / N	Muestra el mes numéricamente sin/con ceros a la izquierda
d / D	Muestra el día del mes sin/con ceros a la izquierda
j / J	Muestra el día del año (1/366) sin/con ceros a la izquierda
w / W	Muestra la semana (1/52) del año sin/con ceros a la izquierda
.,:-'	Caracteres permitidos directamente en el formato de fecha
!	Prefijo para introducir cualquier otro carácter

La misma instrucción anterior puede ser empleada con otros formatos para que la presentación de cada una de las fechas sea totalmente diferente. El próximo ejemplo muestra tres formatos distintos en los que pueden presentarse los segundos días de los años 1959, 1960 y 1961<sup>22</sup>.

```
display %dd_M_cY -364, %dM/d/Y 1, %dd-n-CY 367
```

### ILUSTRACIÓN 2.28. Exposición de fechas (II)

2 January 1959	January/2/60	2-1-1961
----------------	--------------	----------

#### 2.6.4. *El gestor de variables*

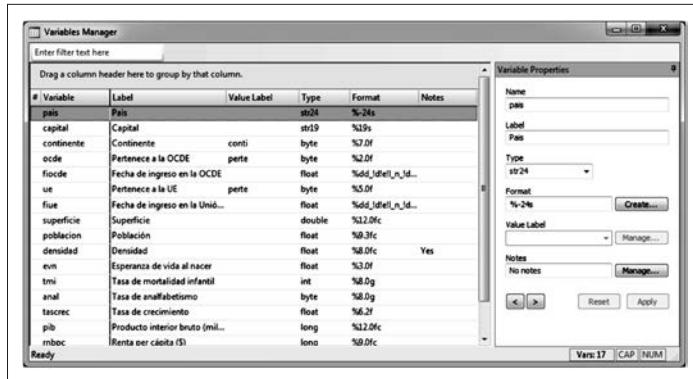
Se podrían sintetizar las tres últimas secciones, explicando una de las novedades introducidas en la versión 11 de Stata. Se trata del gestor de variables, que puede ser llamado mediante menú (*Data/Variables manager*), ícono (el décimo, Variables manager) e incluso desde la ventana de órdenes, mediante la instrucción *varmanage*.

Esta instrucción genera un cuadro de diálogo que contiene tantas líneas como variables se encuentran en la matriz de datos y seis columnas correspondientes al nombre de la variable, su etiqueta y las de sus valores, su tipo, su formato y las notas que el usuario desee incorporar a cada una de ellas.

---

<sup>22</sup> Es preciso añadir que Stata también puede considerar las fechas semanal, mensual, trimestral, semestralmente e incluso fechas con hora incluida. En cualquier caso, siempre el punto de referencia es el 1 de enero de 1960 y la variable queda guardada como numérica; pero en estos casos el número, en lugar de días, significa semanas, meses, trimestres, semestres o milisegundos. Sin embargo, como este libro no trata de series temporales, se considera que no es útil explicar su uso. Se sugiere pedir ayuda en Stata mediante la instrucción *help dates*.

### ILUSTRACIÓN 2.29. Ventana del gestor de variables



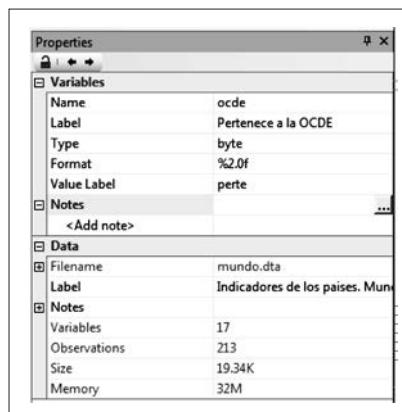
Las variables de esta lista pueden ser ordenadas por su posición en la matriz (#), por su nombre, el de la etiqueta o el de la lista de etiquetas de los valores. Del mismo modo, esto es, apretando en su cabecera, pueden quedar ordenadas por su tipo, formato o la posesión o no de notas. También las variables pueden ser clasificadas por cualquiera de sus características menos el nombre y su orden, siempre y cuando se arrastre su encabezamiento a la franja situada por encima de las cabeceras. Incluso, en el caso de contar con muchas variables, puede realizarse una selección de ellas, escribiendo los caracteres deseados en la casilla situada arriba a la izquierda inmediatamente debajo del marco de la ventana. Todos estos cambios mencionados en este párrafo afectan sólo al gestor, ya que la cantidad de variables y el orden en la matriz quedarán inalterados<sup>23</sup>.

Otra característica sobresaliente del gestor de variables es la posibilidad de trasladar el nombre o la lista de un subconjunto de variables a la ventana de órdenes, al editor de programas o donde se requieran. Apretando la tecla mayúscula o control al tiempo que se pulsa el botón izquierdo del ratón en las respectivas líneas de las variables, quedará activada más de una variable. Una vez que se han seleccionado las variables pertinentes, apretando el botón derecho del ratón aparecerá un menú contextual, cuya última línea hace que todos los nombres de las variables seleccionadas aparezcan en la ventana de órdenes. Alternativamente, si se selecciona la línea *Copy varlist* o se pulsan las teclas Ctrl+c, se almacenarán en el portapapeles y, de este modo, podrán pegarse allá donde se requiera con la combinación Ctrl+v.

<sup>23</sup> Si se desea borrar variables de la matriz podrá hacerse mediante las opciones *keep* y *drop* del menú contextual que se obtiene pulsando el botón derecho del ratón sobre una selección de variables. Ambas opciones son también instrucciones que pueden ejecutarse en la ventana de órdenes para mantener o borrar una lista de variables.

El gestor de variables permite también introducir y editar las características de las variables. En la parte derecha, se encuentra una ventana interna que contiene nombre, etiqueta, tipo, formato, etiqueta de valores y notas de la variable seleccionada. Todas estas características pueden ser modificadas por el usuario, del mismo modo que también pueden efectuarse cambios desde la ventana interna de las propiedades de las variables, que aparece en la parte inferior izquierda de la pantalla por defecto de la versión 12 de Stata (véase la ilustración 2.4), sin necesidad de entrar en el gestor de variables, siempre y cuando aparezca abierto el candado situado debajo del título (*Properties*) de la ventana.

**ILUSTRACIÓN 2.30.** Ventana interna de las propiedades de las variables



## 2.7. Ejercicios

- 1) Familiarízate con los ficheros de ejemplo con los que cuenta Stata mediante la instrucción *sysuse*. Mira qué variables contiene, el modo en que están grabadas, su formato de presentación y las etiquetas que contiene. Finalmente haz un listado de los diez primeros casos. (Ficheros propuestos: census, citytemp, educ99gdp, gnp96, lifeexp, pop2000, uslifeexp, voter).
- 2) Copia los ficheros que se proporcionan con este libro a un directorio. Arranca Stata desde ese directorio (o al menos, una vez dentro de Stata, escribe la instrucción *cd "directorio"*). Lista todos los ficheros que sean del tipo *\*.dta*. Ábrelos e inspecciona también las variables, su formato y etiquetas para terminar haciendo examen de ellos con la instrucción *browse*.
- 3) En el fichero que prefieras de los dos ejercicios anteriores, cambia las etiquetas del fichero, de las variables y de los valores. Por ejem-

plo, todos los ficheros del primer ejercicio tienen etiquetas en inglés, coge uno de ellos y reemplázalas por otro idioma. Si quieras conservar los cambios, no olvides terminar con la instrucción *save, replace*, que se explica en el capítulo 3.



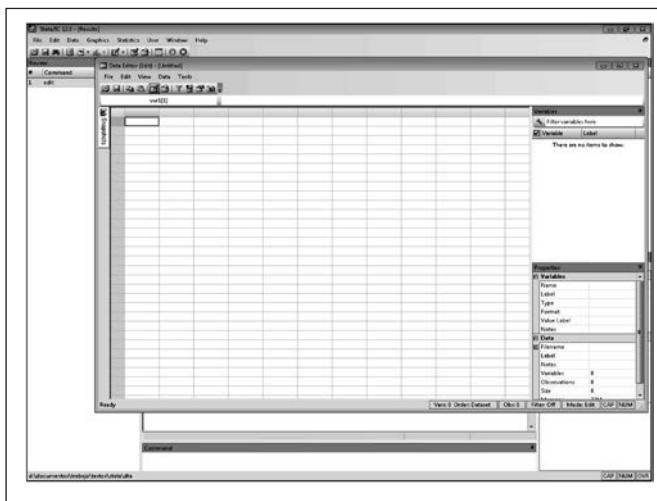
# 3

## Introducción de datos

### 3.1. Introducción manual de datos

Rara vez se introduce información manualmente en Stata. Habitualmente se utilizan datos previamente preparados en otros programas, por lo que es importante saber pasar al formato de Stata los datos grabados en otros formatos (en formato SPSS, Excel, o ASCII, por ejemplo). Eso se contemplará en el siguiente apartado. A continuación, sólo de modo somero, se indica cómo se introducen los datos manualmente, uno a uno, en Stata.

**ILUSTRACIÓN 3.1. Ventana de introducción manual de datos**

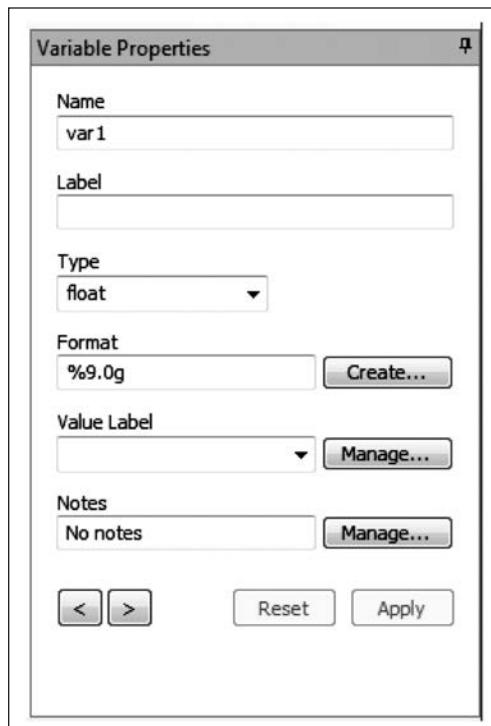


Si se utiliza Stata para Windows, la mejor alternativa es mediante la orden *edit*<sup>1</sup>. Esta hace aparecer un editor de datos tipo hoja de cálculo (como

<sup>1</sup> Esta operación también puede realizarse a través de menú (*Window/Data Editor*) o a través del teclado pulsando la combinación de teclas Ctrl+8 (Ctrl+7 en la versión 11).

Excel, por ejemplo, o la vista de datos de SPSS), con el que puede introducirse y modificarse la matriz de datos (véase la ilustración 3.1). En *Data Editor*, cada fila es un caso y cada columna una variable. Introducir datos es tan simple como teclear en la casilla correspondiente el valor que sea y presionar la tecla *Intro*. Stata crea automáticamente nombres para las variables, que aparecen en la parte de arriba de cada columna: *var1*, *var2*, etc. Obviamente, pueden cambiarse estos nombres para darles más sentido. Para ello, ha de pulsarse el noveno ícono (*Properties*) de la barra de herramientas del editor, o acceder a la ventana de propiedades a través del menú del editor (*View/Properties*), o directamente haciendo clic en el recuadro situado a la derecha de *Name* en el apartado de *Variables* de la ventana llamada *Properties*, situada por defecto en la parte inferior derecha de la ventana de edición. Otra opción es hacer los cambios con el gestor de variables, que aparece mediante el menú *Tools/Variables Manager*. En esta nueva ventana, se encuentra a la derecha un cuadro de diálogo (véase la ilustración 3.2) en el que se puede cambiar el nombre de cualquier variable, ponerle una etiqueta, cambiar su tipo, modificar el formato de visualización de los datos o añadirle notas.

**ILUSTRACIÓN 3.2. Cuadros de diálogos en *Variable Properties* del gestor de variables de Stata**



El nombre de la variable puede tener hasta 32 caracteres desde la versión 7 de Stata (anteriormente el nombre no podía tener más de 8 caracteres), lo que implica que se puedan poner nombres a las variables razonablemente comprensibles. Stata permite que se haga referencia a las variables de forma abreviada, siempre y cuando no haya otra variable que comience con los mismos caracteres: por ejemplo, puede llamarse a una variable *numero\_de\_hijos*, pero utilizarse habitualmente *num* para hacerle referencia, si no existe ninguna otra variable que empiece por estos tres caracteres. Esto es muy conveniente, puesto que permite usar nombres muy descriptivos (en lugar de los típicos nombres incomprensibles) sin necesidad de teclear demasiado cuando haya que nombrarlos para solicitar algún análisis<sup>2</sup>.

Otra manera de cambiar el nombre de las variables es tecleando la instrucción *rename* en la ventana donde se introducen las órdenes. Simplemente se introduce la palabra *rename* seguida del nombre de la variable que se quiere cambiar y después del nuevo nombre que se le desea asignar. Por ejemplo, si se quiere poner un nombre más comprensible a la variable *var1* generada por defecto al escribir el primer dato en la primera fila y columna, puede hacerse mediante la instrucción:

```
rename var1 sexo
```

Es preciso advertir tres aspectos que hay que cuidar al poner nombre a las variables: en primer lugar, que los nombres no pueden empezar con un carácter que sea numérico. De este modo, *1pregunta* no es un nombre válido para Stata. En segundo lugar, hay que tener en cuenta que no se admiten espacios en blanco en los nombres de las variables, ni pueden usarse signos especiales como la coma y el punto. Sí puede incluirse el guión bajo “\_” y este puede ser utilizado como un sustitutivo del espacio en blanco. En la instrucción anterior hay un buen ejemplo de ello con el nombre *numero\_de\_hijos*. Finalmente, es preciso notar que Stata es sensible a la diferencia entre mayúsculas y minúsculas. De este modo, si se opta por poner el nombre *Numero*, si posteriormente se escribe *numero*, no reconocerá el nombre anterior.

En el cuadro de diálogo que se aprecia en la ilustración 3.2 puede también asignarse una etiqueta a la variable para identificarla mejor. Las etiquetas tienen menos restricciones que los nombres en cuanto a los caracteres que pueden incluir (por ejemplo, pueden contener espacios). Cuando se soliciten tablas o cualquier tipo de análisis de datos, Stata

---

<sup>2</sup> Pese a su conveniencia, en este libro se emplearán nombres cortos en las variables para simplificar la lectura tanto en los ejemplos de órdenes que aparecen en el texto como en los párrafos que la explican.

mostrará las etiquetas en vez del nombre, pero en las instrucciones siempre habrá que referirse a las variables por su denominación y no por su etiqueta<sup>3</sup>.

Puede también cambiarse en este cuadro de diálogo el tipo de variable y el formato de visualización de datos. Este último no afecta nada al modo en el que están almacenados los datos, sino a cómo son mostrados en pantalla. El formato de visualización de datos se usa sobre todo para controlar cómo se muestran los decimales. Por ejemplo, el formato de visualización que aparece en la ilustración 3.2 (%9.0g) significa nueve espacios en conjunto sin ningún decimal. El signo de porcentaje que aparece al principio es la convención usada por Stata para designar formatos de visualización de datos. El número que viene después especifica el conjunto de espacios que se van a emplear, y el número que se muestra tras el punto especifica cuántos decimales se van a exponer. La *g* al final indica que el formato es el general, lo que quiere decir que el propio programa decide qué mostrar en función de los datos. Como se vio en la sección 2.6.3, hay dos tipos más que permiten especificar el formato de cantidades numéricas: *científico* (para notación científica, *e*) y *fijo* (que siempre mostrará exactamente el número de decimales especificado, aunque sean innecesarios o demasiado imprecisos, *f*). En este ejemplo, al tener el formato *g*, si un caso tuviera el valor 2,2556, podría aparecer en pantalla como 2,26 (Stata redondea un determinado número de decimales en función del lugar que disponga para mostrar la cantidad en cuestión). Pero lo que se almacena realmente es 2,2556, lo que hace que los cálculos estadísticos sean mucho más precisos.

En la parte superior de la ventana del editor de Stata se encuentra tanto un menú como una barra de herramientas diferentes a la de la ventana principal. Lo más característico en el primero es el ítem *Tools*, en este como en la barra de herramientas destacan dos operaciones peculiares del editor: el filtro de observaciones, mediante el cual se pueden seleccionar las filas de la matriz que cumplan con determinadas características, y el gestor de variables, descrito al final del capítulo anterior. En el último ícono de la barra de herramientas se encuentra el creador de instantáneas (*Snapshots*), que permite hacer grabaciones temporales de los datos, por si hubiera que recuperar datos en un determinado momento por haber cometido algún error en la edición o modificación de variables, o por trabajar indistintamente con distintos conjuntos de datos<sup>4</sup>. Una vez estén introducidos los datos, o si se ha

<sup>3</sup> En el capítulo anterior se detalló cómo poner las etiquetas mediante la instrucción *label variable*.

<sup>4</sup> Pueden hacerse tantas grabaciones parciales de datos como se quiera, incluyendo los procedentes de distintos ficheros, numerándose automáticamente a medida que se producen. Estas instantáneas se mantendrán en tanto que no se salga de Stata o se eliminen intencionalmente por el usuario. Si se quiere preservar permanentemente algún cambio realizado, es preciso realizarlo mediante la orden *save, replace*. Para recuperar las instantáneas, basta con hacer

realizado alguna modificación, al cerrar el editor, se guardan automáticamente los datos tal y como estén en ese momento, no pudiéndose recuperar el estado inicial, a menos que se haya hecho alguna instantánea con anterioridad. Aunque los datos estén en el editor, se puede trabajar con ellos. Pero hay que tener en cuenta que hasta que no se dé la orden a Stata de guardar los cambios en el disco, los datos sólo estarán en la memoria interna del ordenador, mientras Stata esté en funcionamiento<sup>5</sup>. Para almacenar los datos en disco, de modo que se puedan recuperar en otra sesión de Stata, es preciso dar la instrucción *save*, seguida por el nombre que se desee dar al archivo:

```
save "nombrefichero"
```

De este modo, se guardarán los datos en el disco duro del ordenador, en el archivo nombrado entre comillas. La extensión .dta es asignada automáticamente por Stata a sus archivos, para identificarlos como tales. El nombre del archivo en la instrucción *save* sólo es obligatorio expresarlo la primera vez que lo guardemos. Una vez que el archivo ha sido creado en el disco duro, si se modifica, es decir, si se introducen nuevos casos, variables, etc., y se quiere guardar los cambios, ha de añadirse la opción *replace*. Caso de que no se indique el nombre del fichero, Stata le da el que tenía antes de ser modificado. Y así se guardan los cambios sobre el archivo ya existente en el disco duro.

```
save [, replace]
```

Un aspecto importante a tener en cuenta cuando se graba, o cuando se lee, un fichero es el directorio por defecto donde Stata realiza sus operaciones. Se puede saber fácilmente, mediante la orden *cd* y mediante esta también puede cambiarse la carpeta de trabajo. De este modo, si el lector quiere ubicar todos sus ficheros de trabajo con Stata en un directorio llamado “d:\documentos\datos Stata”, al comienzo de la sesión debería verificar que se

---

un doble clic en la instantánea deseada entre las presentes en el listado que se obtiene al apretar el mismo icono que sirve para generarlas. Las instrucciones, caso de que se deseen realizar en la consola de Stata, son *snapshot save* y *snapshot restore*, respectivamente.

<sup>5</sup> Esta es una de las peculiaridades de Stata frente a otros programas de estadística, como el SPSS, que trabajan con los datos en el disco duro. Stata almacena y procesa los datos en la memoria central del ordenador. Sólo utiliza el disco duro para recuperar los datos y para guardarlos. La ventaja es que trabaja mucho más rápido (la velocidad de procesamiento de datos en memoria es mucho mayor que en disco duro). El inconveniente, que requiere mucha más memoria en el ordenador que otros paquetes estadísticos; sin embargo, hoy en día cuando las capacidades de los ordenadores superan los tres dígitos de Megas, casi cualquier base de datos puede almacenarse en memoria RAM. Por otro lado, si esta es insuficiente, el programa la genera dinámicamente en el disco, pero de este modo los procesos se ralentizan considerablemente.

encuentra en ese directorio<sup>6</sup> y, caso de no estarlo, cambiarlo mediante la siguiente orden<sup>7</sup>:

```
cd "d:\documento\datos Stata"
```

### 3.2. Lectura de datos con Stata

En la mayor parte de las ocasiones, los datos con los que se trabaja no son introducidos directamente por el analista, sino que provienen de institutos u organismos dedicados a la realización de encuestas. En estos casos, los datos pueden estar en dos formatos:

1. Formato ASCII: se trata de archivos de texto en el que los datos están almacenados siguiendo alguna pauta, que normalmente se proporciona aparte en el llamado libro de códigos, junto con el cuestionario. Para poder trabajar con estos datos, habrá que transformarlos al formato de Stata, lo que se puede hacer en el propio programa, como se verá más adelante con más detenimiento.
2. Formato binario o mixto: los datos también pueden encontrarse en el formato de algún programa de estadística, de hoja de cálculo o de base de datos, como SPSS, *Excel* o *Acces*. En estos casos, para leer los datos hay varias alternativas no siempre posibles. La más directa es que sea el mismo Stata quien se encargue de la lectura, solución sólo posible en el caso de ficheros Excel, ODBC, XML, SAS o una base de datos de *Haver Analytics*. La segunda solución es que el propio programa sea capaz de traducir su propia base de datos en otra legible por Stata. Este es el caso más directo para el usuario en el caso de disponer de una base de datos cargada en una versión de SPSS superior a la 17.0. Finalmente, para casos no contemplados anteriormente, habrá que utilizar algún programa especializado en conversión, como Stat/Transfer. También más adelante se describe brevemente cómo utilizar este programa complementario.

---

<sup>6</sup> Al instalar Stata puede indicarse el fichero de trabajo con los datos. Una vez instalado puede cambiarse el directorio por defecto en las propiedades del acceso directo del programa, propias del sistema operativo. Si se pulsa el botón derecho en el ícono con el que se comienza el programa y en su menú contextual se opta por propiedades, en la pestaña *Acceso directo* puede indicarse la ruta de comienzo del programa en la casilla *Iniciar en*.

<sup>7</sup> Esta orden es crucial, sobre todo, cuando pasan los datos de un ordenador a otro, porque no siempre la estructura de las carpetas de los ficheros es idéntica en uno u otro, o porque las opciones por defecto de Stata sean diferentes. Para que los ejercicios incluidos funcionen bien, el usuario tendrá que especificar en qué directorio ha colocado los ficheros de trabajo. Esto sólo lo tendrá que hacer una vez en cada fichero, puesto que se emplea el recurso de los macros globales, para evitar tener que realizarlo repetidamente.

### 3.2.1. Leer datos en formato ASCII con Stata

Dependiendo de cómo estén almacenados los datos en el archivo del que se disponga, habrá que utilizar una u otra instrucción de Stata. En el cuadro 3.1 pueden verse unas indicaciones fundamentales para saber qué orden ha de utilizarse en función del formato de la base de datos disponible. Hay cuatro posibilidades principales, de creciente complejidad y versatilidad: desde la instrucción *insheet* (relativamente sencilla de usar) hasta *infile* con diccionario (bastante complicada pero mucho más potente).

**CUADRO 3.1. Diferentes instrucciones para la lectura de datos en Stata**

	<b>Separación de variables</b>	<b>Requiere el nombre de las variables</b>	<b>Requiere comillas en variables de texto</b>	<b>Necesita un diccionario (.Dct)</b>	<b>Requiere que cada fila represente un caso</b>
Insheet	Tabuladores o comas	No	No	No	Sí
Infile formato libre	Espacios o comas	Sí	Sí	No (opcional)	Sí
Infix	Ninguna (ancho fijo)	Sí	No	No (opcional)	No
Infile ancho fijo	Ninguna (ancho fijo)	Sí	No	Sí	No

En todos los casos, una vez que se hayan leído los datos en memoria, podrán guardarse en el disco en el formato propio de Stata, esto es, en un fichero con extensión *.dta*. Es conveniente que antes de guardar datos nuevos se introduzca la orden *compress* para intentar ocupar el menor tamaño posible de memoria y disco y para trabajar mejor con los datos. Como ya se ha indicado anteriormente, para guardar los datos en el disco, basta con acompañar el nombre del fichero a la instrucción *save*:

A partir de ese momento, se podrán recuperar los datos siempre y cuando se estime conveniente, simplemente abriéndolos en Stata.

- a) *Insheet*: De las cuatro posibilidades, *insheet* es la más sencilla, pero también la que impone más restricciones a la base de datos que se quiere leer. Cada línea del archivo de texto debe representar un caso, y los valores de los individuos en las variables deben estar separados por tabuladores o comas o cualquier otro carácter especificado en la opción *delimiter()*. Si un archivo cumple estas restricciones, se podrá leer con *insheet*, del siguiente modo:

```
insheet [listavar] using "nombrefichero" [, clear tab|comma|delimiter("carácter")]
```

La opción *clear* permite cargar en memoria un nuevo fichero, sin perjuicio de que esté cargado algún otro. En *insheet*, puede especificarse también los nombres de las variables que han de leerse, pero no es necesario. En el caso de que la primera línea del archivo represente los nombres de las variables separadas por tabuladores o comas, *insheet* así lo entenderá y dará esos nombres a las variables. Por ejemplo, las tres líneas del archivo *fichero.dat*, que incluye la matriz de datos de la ilustración 2.2, han de ser las siguientes:

**ILUSTRACIÓN 3.3. Fichero de datos con formato de hoja de cálculo  
(*fichero.dat*)**

sex, edad
1, 21
2, 20

La primera línea del archivo *fichero.dat* contiene los nombres de las variables separados por comas. La segunda línea es el primer caso, un hombre de veintiún años, que como puede comprobarse tiene los valores de las variables separados por comas. De este modo, la instrucción:

insheet using "fichero.dat", clear
------------------------------------

leerá los datos del archivo *fichero.dat*<sup>8</sup> en Stata poniendo a las variables los nombres que aparecen en la primera línea.

¿Qué sucede si no existe una primera línea con los nombres de las variables, como el fichero representado en la ilustración 3.3? En ese caso pueden ocurrir dos cosas. Si no se especifican los nombres de variable en la instrucción, como se hizo anteriormente, Stata asigna automáticamente nombres estándares a las variables: la primera será *v1*, la segunda *v2* y así sucesivamente. Si interesa, por el contrario, que las variables tengan un nombre más lógico, puede dárseles nombre en la misma instrucción, del siguiente modo:

insheet sexo edad using "ficheroa.dat", clear
---

Con lo cual Stata leerá las variables del archivo asignándole los nombres que se especifican en la instrucción. Ahora bien, si se explícita el nombre de las variables, hay que nombrar todas las que haya en el archivo, porque si no, Stata dará un mensaje de error.

---

<sup>8</sup> Conviene recordar la distinción entre ficheros brutos de datos (con extensión *dat* o *txt*) y los ficheros de datos y definiciones grabados por Stata (con extensión *dta*).

- b) *Infile con formato libre*: en este caso, los valores de las variables deben estar separados por espacios<sup>9</sup> en el fichero. Es preciso especificar los nombres de las variables, bien sea en el propio comando o en un archivo de diccionario. El funcionamiento de los archivos de diccionario es bastante complejo, y se explicará en los últimos párrafos de esta sección. Realmente, para utilizar *infile* con archivos de datos de formato libre no es necesario crear un archivo de diccionario, se le pueden dar todas las especificaciones en la instrucción, que es como se explica a continuación.

En este caso, es necesario que se especifique el nombre de las variables que contiene el archivo que se desea leer. Como en *insheet*, habrá que especificar todas las variables que haya en el archivo de datos, aunque en este caso si no se realiza de modo completo, no dará un mensaje de error, sino que simplemente leerá mal los datos.

Si no hay variables tipo cadena (*string*, que contienen texto), este comando es casi tan sencillo como *insheet*. Su fórmula general es:

```
infile listavar using "nombrefichero"
```

O sea, hay que escribir la orden *infile* seguida por los nombres de variables y, tras la suborden *using*, el nombre del archivo donde están los datos.

En el caso de que haya variables de tipo cadena, como por ejemplo si se añade al fichero anterior el nombre de las personas, el archivo de datos de origen deberá tener los valores alfabéticos entrecomillados obligatoriamente si incluyen espacios. Esto es lógico porque Stata necesita diferenciar el espacio del valor del espacio que separa los distintos valores de la variable. Por ejemplo, en el archivo *ficherob.dat*, existe una tercera variable.

#### ILUSTRACIÓN 3.4. Fichero de datos con formato fijo y variable cadena (*ficherob.dat*)

1 21 Juan
2 20 "María José"

Como puede verse, los valores están separados por espacios. Stata, al leer los datos con *infile*, pondrá el primer valor en la primera variable, el que está tras el primer espacio en la segunda, etc. En cambio, la línea de la mujer de 20 años tiene un espacio entre “María” y “José”, por lo que todo el nombre debe estar entrecomillado. Así Stata sabe que todo lo que está entre las comillas, independientemente de los espacios, va en la tercera variable.

Por tanto, si en el archivo de datos hay variables alfanuméricas, la orden cambia ligeramente. En este caso, habrá que especificar antes

---

<sup>9</sup> Este comando también se puede utilizar si los valores están separados por comas, pero en ese caso es más sencillo usar *insheet* en la mayor parte de los casos.

del nombre de la variable que la variable es *string* y el número de caracteres máximo (siguiendo la fórmula vista en la sección 2.6.2). En el ejemplo actual, para leer el archivo *ficherob.dat*, ha de aparecer una instrucción similar a la siguiente:

```
infile sexo edad str10 nombre using "ficherob.dat", clear
```

Puede verse cómo la tercera variable (*nombre*) va precedida de *str10*, lo que quiere decir que es una variable tipo alfanumérico de tamaño 10 (el número máximo de caracteres que contiene es 10). Para las variables numéricas no hace falta especificar el tipo, pues Stata lo asigna automáticamente.

- c) *Infix*: esta instrucción está específicamente diseñada para leer archivos de datos con formato de ancho fijo. Esto puede implicar que no haya ninguna separación entre los valores de las variables. Para asignar los valores a las variables correctamente, es preciso saber el número exacto de columnas que ocupa cada variable, e incluirlo en la orden para que Stata pueda leerlas correctamente. En el ejemplo anterior de dos personas, el archivo de datos puede presentarse del modo siguiente:

**ILUSTRACIÓN 3.5. Fichero de datos con formato fijo y una sola línea por caso (*ficheroc.dat*)**

121Juan
220María José

Se observa cómo no existe ningún tipo de separación entre los valores. Los datos están almacenados por columnas (o caracteres): la primera variable ocupa la primera columna, la segunda, las dos siguientes, la tercera, las 10 últimas. En la orden de lectura, hay que indicar a Stata precisamente eso: todo lo que está en la primera columna hay que asignarlo a la primera variable, lo que esté en las 2 siguientes en la segunda, el resto a una variable cadena. La forma concreta para este ejemplo del comando sería:

```
infix sexo 1 edad 2-3 str nombre 4-13 using "ficheroc.dat", clear
```

Como en la orden *infile* de formato libre, también hay que especificar el formato de las variables de texto, aunque aquí no haga falta poner su tamaño porque está implícito en el ancho de la variable. En este caso, se ha puesto antes de la variable *nombre* la palabra *str*, para que Stata la identifique correctamente como variable de texto. Esto no es necesario para las variables numéricas. Tras el nombre de cada variable, ha de especificarse su ancho, indicando de qué columna a qué

columna van los datos que le corresponden. En este caso, los nombres de los sujetos están almacenados desde la columna 4 a la columna 13.

Tanto en *infix* como en *infile* con ancho fijo pueden leerse menos variables que las que realmente hay en el archivo si así se desea. Simplemente, saltando unas determinadas columnas del archivo en la secuencia de la instrucción, esas columnas no serán leídas ni incluidas en ninguna variable. Por ejemplo, si en el último ejemplo que se ha citado se quitara *edad 2-3*, el archivo se leería perfectamente, pero sin esta variable.

**ILUSTRACIÓN 3.6. Fichero de datos con formato fijo y más de una fila por caso (*ficherod.dat*)**

121
Juan
220
María José

Esta instrucción también es útil, si cada caso ocupa más de una línea. En el supuesto de que los datos de los dos sujetos estuviesen almacenados en formato de ancho fijo, pero cada persona no ocupara una sola línea, sino dos como aparece en la ilustración 3.6, sería preciso escribir la instrucción *infix* como sigue:

infix 2 lines 1: sexo 1 edad 2-3 2: str nombre 1-10 using "ficherod.dat", clear
---

Tras el comando *infix*, se especifica el número de líneas (# *lines*) y luego se precede cada nueva línea por su número y dos puntos (#:).

Realmente, para archivos complicados y con muchas más variables es más conveniente utilizar un diccionario en lugar de dar órdenes de lectura en el propio comando. *Infix* puede utilizarse con un diccionario, pero su verdadera utilidad es como un modo sencillo de leer archivos con formato de ancho fijo. Para leer archivos verdaderamente complicados, el comando más potente es *infile* con un diccionario, cuyo formato se explica a continuación.

- d) *Infile de ancho fijo (con diccionario)*: para archivos de datos complicados y con muchas variables, lo más conveniente es utilizar esta instrucción. Al utilizarla de esta manera, todas las especificaciones y órdenes para que Stata lea el archivo de datos están en un archivo aparte, llamado diccionario. De este modo, la instrucción en sí quedaría expresada de la siguiente simple manera:

<b>infile using "ficheroddiccionario"</b>
---

Basta con indicar el nombre del archivo de diccionario que contendrá las órdenes de lectura de los datos. El archivo de diccionario deberá llevar la extensión .dct, y seguir las siguientes pautas:

Comienza con la orden *infile dictionary using "nombrefichero"* (esto es, el archivo donde están los datos a leer), seguido por un corchete (*{*) que marca el comienzo de las órdenes de lectura, y que se cerrará (*}*) cuando finalicen las especificaciones.

Entre los corchetes estarán las instrucciones para leer el archivo. Conviene utilizar una línea de especificaciones para cada variable, del siguiente modo:

1. Primero, se expone la posición de los datos de la variable en el archivo.
2. Segundo, el tipo de datos (opcional).
3. Tercero, el nombre de la variable.
4. Cuarto, el formato de visualización (opcional).
5. Por último, se puede escribir la etiqueta de la variable entre comillas.

Véase un par de ejemplos para entenderlo mejor. En primer lugar, para leer el fichero de datos reflejado en la ilustración 3.6, el contenido del fichero diccionario debería ser como sigue:

**ILUSTRACIÓN 3.7. Contenido de un fichero diccionario de la instrucción  
*infile (diccionario.dct)***

```
infile dictionary using "ficherod.dat" {
    _lines(2)
    _column(1) byte sexo %1f "Sexo"
    _column(2) byte edad %2f "Edad"
    _line(2)
    _column(1) str10 nombre %10s "Nombre de pila"
}
```

Es preciso notar que además de las órdenes correspondientes a las variables, compuestas por posición (*\_column(#)*), tipo (*byte, str#*), nombre (*sexo, edad y nombre*), formato (*%#f o %#s*), etiqueta ("Sexo", "Edad" y "Nombre de pila"), hay dos instrucciones imprescindibles para la lectura de ficheros con más de una línea por caso: Con *\_lines(#)*, se ordena la lectura de # líneas por individuo, 2 en este caso. Y con *\_line(#)* se indica que las columnas posteriores corresponden o están situadas en la línea indicada en el número #.

Una vez construido y grabado el fichero diccionario, es imprescindible utilizarlo mediante la orden *infile using* seguida del nombre que se le ha asignado al fichero, *diccionario.dct*, en este caso:

```
infile using "diccionario", clear
```

En el segundo ejemplo se trata de leer en Stata las dos primeras preguntas del cuestionario 2384 del CIS<sup>10</sup>, la encuesta postelectoral de marzo de 2000.

El archivo de datos se encuentra en formato de ancho fijo, y en el cuestionario se especifican los anchos y las posiciones de cada variable en el archivo. Se dispone, por consiguiente, de una primera pregunta con una sola variable, que ocupa la posición 28 de la primera fila y de una segunda pregunta con cuatro variables, que ocupan respectivamente una posición desde la 29 hasta la 32. El mismo cuestionario que proporciona el CIS da información acerca de cómo fueron hechas las preguntas, cómo están codificados los valores de las variables y también indica en qué posición fue grabada cada variable, a través de un número entre paréntesis que representa en qué columna de la línea del archivo de datos está la variable. Así, se sabe que la pregunta 1 está en la columna 28, y las cuatro siguientes variables de la segunda pregunta en las 29, 30, 31 y 32. Con esa información, pueden leerse los datos en Stata con la orden *infile* y un diccionario.

### ILUSTRACIÓN 3.8. Dos primeras preguntas del estudio del CIS 2384 (2000)

P.1 Para empezar, ¿podría Ud. decirme si recuerda, cuando era niño o adolescente, con qué frecuencia solía hablarse de política en su casa: con mucha frecuencia, de vez en cuando, pocas veces o prácticamente nunca?						
- Con mucha frecuencia .....	1					
- De vez en cuando .....	2					
- Pocas veces .....	3 (28)					
- Prácticamente nunca .....	4					
- No recuerda .....	8					
- N.C. .....	9					
P.2 Indíqueme, por favor, ¿hasta qué punto está Ud. muy de acuerdo, de acuerdo, en desacuerdo o muy en desacuerdo con cada una de las siguientes frases?						
	Muy de ac. De ac.	En desac.	Muy en desac.	NS	NC	
- Por lo general, la política es tan complicada que la gente como yo no puede entender lo que pasa .....	1	2	3	4	8	9 (29)
- A través del voto, la gente como yo puede influir en la política .....	1	2	3	4	8	9 (30)
- Los políticos no se preocupan mucho de lo que piensa la gente como yo ...	1	2	3	4	8	9 (31)
- Esté quien esté en el poder, siempre busca sus intereses personales .....	1	2	3	4	8	9 (32)

<sup>10</sup> Las características de este estudio se encuentran en <http://www.cis.es>.

Ha de crearse un archivo al que puede denominarse cis2384.dct (dct es la extensión de los archivos de diccionario), que contenga lo siguiente:

**ILUSTRACIÓN 3.9. Contenido de un fichero diccionario de la instrucción *infile* (cis2384.dct)**

```
infile dictionary using cis2384.dat {
    _column(28) byte polinf %1f "P1.- Recuerdo conversaciones políticas"
    _column(29) byte pol1 %1f "P2a.- Política complicada"
    _column(30) byte pol2 %1f "P2b.- Gente influyente"
    _column(31) byte pol3 %1f "P2c.- Políticos despreocupados"
    _column(32) byte pol4 %1f "P2d.- Intereses personales del poder"
}
```

Puede apreciarse que el contenido de este fichero comienza con la instrucción *infile dictionary*, seguida por *using* y el nombre del archivo donde se encuentran los datos que hay que leer. Tras el corchete, se han especificado varias líneas en las que se denominan las variables que hay que leer (*polinf*, *pol1*, *pol2*, *pol3* y *pol4*) y cómo. Primero se coloca *\_column* y la columna donde se localizan los valores de cada variable entre paréntesis. En el caso de la primera variable, que se denomina arbitrariamente *polinf*, se encuentran en la columna 28 del archivo de datos. Luego se indica el tipo de datos: en este caso, dado que los valores sólo van de 1 a 8, puede emplearse el tipo *byte*, que es el que menos memoria ocupa. Después, el nombre de la variable, y por último, se expresa el formato de visualización de los datos, que hace que Stata presente los datos a gusto del usuario. En el presente ejemplo, se ha puesto el tipo *%1f*, que quiere decir tipo fijo con una cifra y sin decimales (véase la sección 2.6.3). Tras todo lo anterior puede ponerse, entrecomilladas, etiquetas para las variables que así se deseé<sup>11</sup>.

Con las palabras claves que se acaban de explicar, se pueden leer la gran mayoría de los archivos de datos que habitualmente se utilizan. Pero puede haber casos más específicos o archivos de datos más complejos. Hay muchas otras posibilidades en el comando *infile* con diccionario que no se explican en este manual por brevedad. Caso de que sea necesario, las demás posibilidades de este comando se pueden consultar en los manuales de referencia de Stata o, de modo más rápido, utilizando la orden *help infile*<sup>12</sup>.

<sup>11</sup> En el supuesto de que se dispusiera de más de una línea por caso, habría que especificar cuándo empieza una nueva línea la palabra *\_newline*.

<sup>12</sup> Se utiliza *help infile1* para la orden de lectura con datos con formato libre o con variables separadas por comas. Se encuentra una detallada explicación con numerosos ejemplos en el manual de gestión de datos en las entradas *infile*, *infix* e *insheet* (Stata 2011d: 328-376).

No hay que olvidar que los diccionarios no se introducen en las ventanas de órdenes de Stata, ni tan siquiera en un fichero de instrucciones (*do*). Deben ir en un fichero independiente y deben ser leídos con la orden *infile using*. De este modo, se puede ver el resultado de la lectura de los diez primeros casos del ejemplo mediante las dos líneas siguientes:

```
infile using "cis2384"
list in 1/10, clean
```

- e) *Outfile*: Stata también incorpora la orden *outfile* para el caso de que se desee hacer lo contrario a lo que se ha explicado anteriormente: guardar datos que se encuentren abiertos en Stata en formato ASCII, de tal modo que puedan luego ser leídos por cualquier programa de estadística o base de datos. Con la siguiente instrucción:

```
outfile using "nombrefichero"
```

Stata guardará los datos separados por espacios en un archivo con extensión raw. Es conveniente utilizar la opción *dictionary* para que Stata guarde con los datos un diccionario que luego haga más fácil la lectura de los datos.

```
outfile using "nombrefichero"[, dictionary]
```

También puede resultar útil la opción *comma*, que separa las variables con comas en lugar de con espacios, especialmente en determinados programas que requieran de las variables separadas por este delimitador.

```
outfile using "nombrefichero"[, comma]
```

### 3.2.2. Lectura y escritura de datos en formato Excel

Desde sus últimas versiones, Stata es capaz de leer y traducir ficheros que no estén escritos en formato ASCII. Desde la 12, incorpora la posibilidad de incorporar al formato de ficheros de Stata los datos guardados con formato de matriz<sup>13</sup> en Excel.

---

<sup>13</sup> Se entiende por formato de matriz aquel en el que los casos son colocados en las filas y las variables, con sus correspondientes nombres en la primera fila, en las columnas.

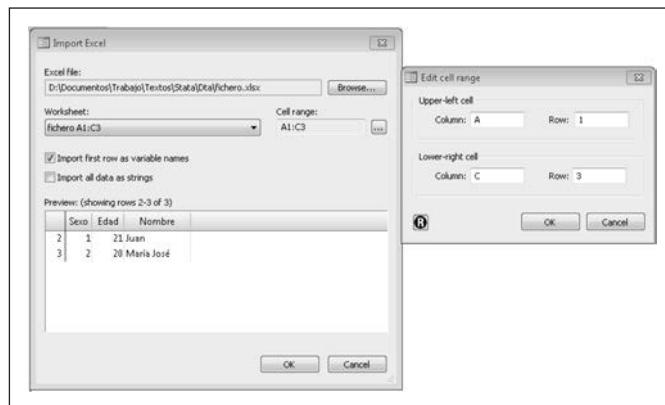
Para llevar a cabo este cometido basta con escribir la siguiente instrucción:

```
import excel"nombrefichero"[, sheet("nombrehoja")] [firstrow [cellrange([rango])]]
```

Esta operación puede realizarse también a través de menú. En ese caso, se debe marcar *File/Import/Excel spreadsheet*. A continuación, aparecerá un cuadro de diálogo donde se le deberá indicar el nombre del fichero, a través del botón *Browse*, el nombre de la hoja en el menú despegable *Worksheet* y, si procede, indicarle el rango de celdas (*Cell range*), si la primera fila contiene los nombres de las variables, ya que si no estuvieran en la hoja de cálculo, Stata se los generaría automáticamente.

En la ilustración 3.10 aparece el cuadro de diálogo correspondiente a la importación de ficheros Excel, en el momento en el que se despliega un cuadro secundario para insertar el rango de la hoja que quiere ser trasladado al formato de Stata. Puede observarse que, en la parte inferior del cuadro principal, aparece una vista previa de los resultados a fin de cerciorarse de que los datos se importan correctamente.

#### ILUSTRACIÓN 3.10. Cuadros de diálogo de la importación de ficheros Excel a Stata



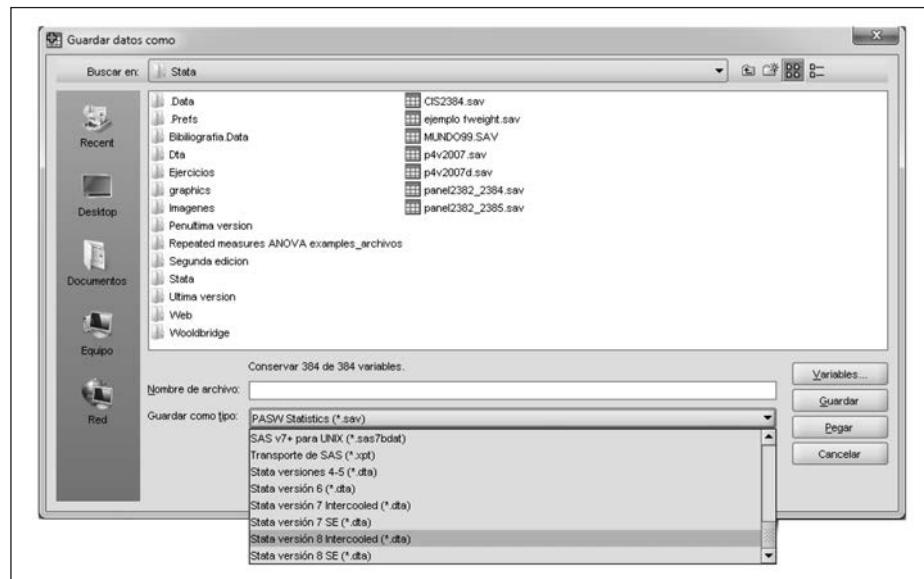
#### 3.2.3. Ficheros procedentes de SPSS

Stata no es capaz de leer ficheros binarios de SPSS. Sin embargo, SPSS puede leer sin dificultad los ficheros de Stata, siempre y cuando se hayan guardado en formato antiguo (*saveold*), pues los ficheros de las últimas versiones tienen un formato distinto no incorporado en las rutinas de lec-

tura<sup>14</sup>. Para ello, al abrir datos mediante el menú SPSS (Archivo/Abrir/Datos), en la pestaña en el menú extensible del tipo de archivos del correspondiente cuadro de diálogo, aparece la opción “Stata(\*.dta)”, que una vez marcada nos mostrará los ficheros con esa extensión en el directorio por defecto presente<sup>15</sup>.

Del mismo modo, SPSS es capaz de traducir sus propios datos en un fichero legible por Stata. Ello quiere decir que todos aquellos que tuvieran sus datos preparados en aquel programa, los podrán leer cómodamente en Stata conservando las etiquetas de variables y valores, así como sus formatos. Para acometer esa transformación, se puede realizar tanto con menú (Archivo/Guardar datos como), tras lo cual hay que cambiar el tipo de archivo con su correspondiente menú extensible, como se indica en la ilustración 3.11. Alternativamente, se puede realizar con código, empleando la orden de SPSS *save translate outfile “nombre de fichero” /type=stata*.

### ILUSTRACIÓN 3.11. Cuadro de diálogo para exportar a Stata datos de SPSS



<sup>14</sup> Desde la versión 7 de Stata existen dos tipos de versiones que generan ficheros ligeramente distintos. Los de la versión especial (SE) tienen una mayor capacidad en el almacenamiento de variables. Asimismo, no hay compatibilidad entre los ficheros de la versión 10, ni de la versión 12 de Stata, con los programas de versiones anteriores. Para solventar este último problema, desde una versión reciente pueden grabarse los ficheros para que puedan ser leídos en versiones anteriores mediante la orden *saveold* nombrefichero.

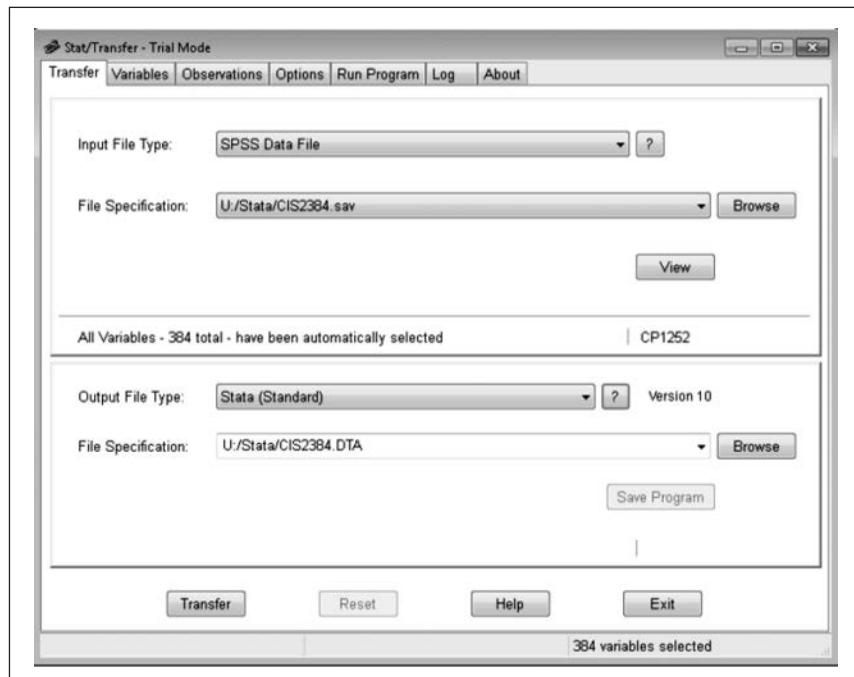
<sup>15</sup> Esta operación también puede hacerse mediante sintaxis, empleando la instrucción de SPSS *get stata file “nombrefichero”*.

### 3.2.4. Lectura de datos en formato binario: Stat/Transfer

Si una determinada base de datos está en un formato nativo de otro programa estadístico, o aplicación, su información no podrá ser leída a no ser que se emplee una utilidad especializada en conversión de datos estadísticos, como es Stat/Transfer, del que se explicará a continuación muy someramente su uso. Alternativamente, si se dispone de este programa, pueden guardarse los datos en formato ASCII, siempre que exista alguna instrucción de exportación, y luego leerlos con Stata. Por ejemplo, si desea convertir datos de SPSS a Stata y no se dispone del Stat/Transfer, caso de disponer de una versión de SPSS anterior a la 16, se pueden grabar los datos como archivo ASCII, en formato fijo, para poder recuperarlos después en Stata con *insheet*, *infix* o *infile*, según se ha explicado en el apartado anterior. Afortunadamente, desde la versión 16 del SPSS, ya es posible guardar los ficheros con la estructura de Stata.

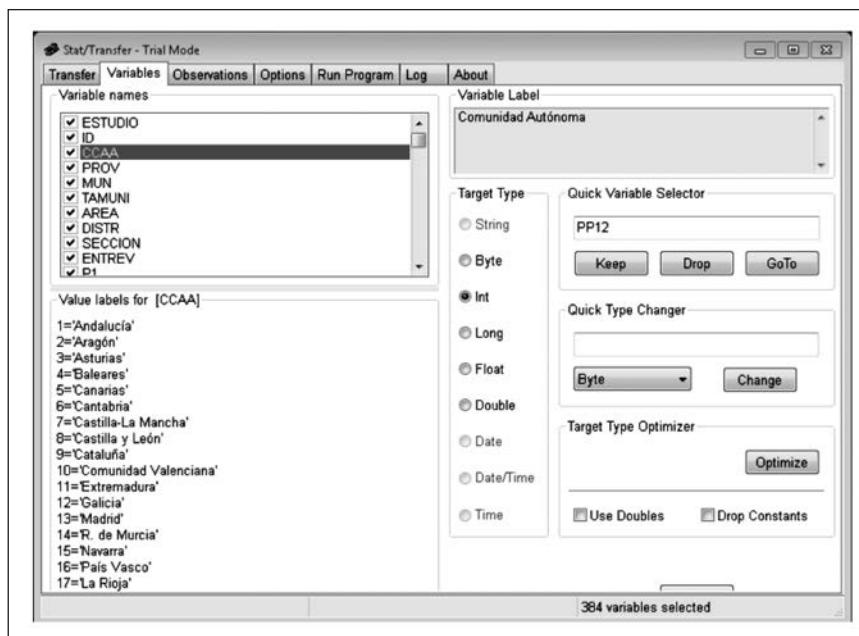
Pero si no se dispone del programa mencionado y, en cambio, se cuenta con un fichero grabado en formato *sav*, habrá que utilizar Stat/Transfer, para transformarlo en *dta* y poderlo utilizar en Stata. Basta un ejemplo para obtener una idea de cómo se utiliza este programa.

**ILUSTRACIÓN 3.12. Cuadro de diálogo principal de Stat/Transfer**



Supóngase que se tiene un archivo, guardado en formato SPSS, tal como lo ofrece el Centro de Investigaciones Sociológicas, que contiene los datos de la encuesta postelectoral de marzo de 2000 (cis2384.sav), y se desea pasar al formato de Stata para trabajar con este otro programa. En la ilustración 3.12 puede observarse el aspecto de la ventana principal de Stat/Transfer. Esta se divide en dos partes, la de arriba para el archivo de origen (el que ha de convertirse a Stata) y la de abajo para el archivo de destino (el nombre que se quiera dar al nuevo archivo convertido). Se elige, en primer lugar, el archivo de datos de origen, el que se desea convertir. Stat/Transfer es capaz de determinar automáticamente su tipo, no hace falta ponérselo. En este caso, el archivo de origen es cis2384.sav, que se ha seleccionado a través del cuadro de diálogo *Seleccionar archivo* que aparece al pulsar *Browse*. Una vez elegido el archivo de origen, hay que seleccionar el tipo de archivo en el que quiere convertirse, donde pone *Output File Type*. En este caso, se ha elegido "Stata version 10". Automáticamente Stat/Transfer pone al archivo de destino el mismo nombre que al de origen, aunque con diferente extensión: la extensión del formato que lleve, en este caso *dta*, la extensión de los archivos de Stata. Una vez especificados el fichero de entrada y el tipo de fichero de salida, se puede pulsar el botón *Transfer* que está en la esquina izquierda inferior para que Stat/Transfer cree un nuevo archivo con toda la información del archivo de origen pero en el formato propio de Stata, de modo que ya pueda trabajarse con el nuevo fichero sin problemas.

### ILUSTRACIÓN 3.13. Cuadro de diálogo de variables en Stat/Transfer



Este programa de conversión tiene muchas diferentes opciones con las que puede controlarse cómo se convierten los archivos de un tipo a otro. Las opciones más comunes son las que se encuentran en la pestaña *Variables* de la ventana de Stat/Transfer (véase la ilustración 3.13). En la parte de la izquierda, pueden seleccionarse las variables que se quieran traducir y las que no (por omisión, están todas seleccionadas) y especificar sus tipos en *Target Type*. En la parte derecha, arriba, se encuentran las opciones de *Quick Variable Selector*, que permite seleccionar (mantener o eliminar) variables en función de determinadas condiciones<sup>16</sup>. Pero lo que es especialmente útil es el botón que está justo debajo, en el recuadro de *Target Type Optimizer*. Stat/Transfer, en principio, asigna a casi todas las variables el tipo float, y esto ocupa mucha memoria. Si se selecciona el botón *Optimize*, se analizan los datos del archivo de origen y se determina para cada variable el tipo más pequeño posible<sup>17</sup>. Esto es muy importante pues permite crear archivos pequeños con más posibilidad de que quepan en memoria y, en consecuencia, con los que Stata trabaja con mucha más rapidez.

### 3.3. Fusión de ficheros

Una vez que se dispone de un fichero, se le puede añadir información similar de otros ficheros, tanto de casos, operación a la que se denominará *unión*, como de variables y en este supuesto se hablará de *combinación*.

#### 3.3.1. Unión de ficheros

En la primera de las opciones, es preciso que las variables (todas o parte) sean iguales; mientras que en la segunda, es necesario que los casos (todos o un subconjunto) sean idénticos.

Aunque esto parezca complejo, con un par de ejemplos, se comprenderá con facilidad. Para ello, se tiene, por un lado, la matriz de datos que figura en la ilustración 3.3 grabada en un fichero de datos Stata con el nombre de *fichero.dta*. Por el otro, se dispone de otra matriz almacenada en *ficheroy.dta*, que contiene dos casos más: un hombre de 20 años y una mujer de 19. Para poner uno a continuación de otro son precisas dos operaciones. La primera es cargar en memoria el primer fichero; la segunda, añadir los casos necesarios, en tanto que ambos ficheros tengan alguna variable en común: *sexo* y *edad*, en este ejemplo.

<sup>16</sup> Aquí pueden emplearse los caracteres comodines (\*) y (?) con el mismo significado que se utilizan en el Sistema Operativo. Por ejemplo P20?, incluiría variables como P20a, P201 o P209, y P20\* también consideraría en la inclusión o exclusión, además de las anteriores, variables como P2000 y P20ab.

<sup>17</sup> Este botón realiza exactamente la misma función que el comando *compress* de Stata.

La orden que permite realizar la primera operación es *use*; la que es necesaria para añadir los casos es *append*, que utiliza la siguiente sintaxis:

```
append using "nombrefichero" [, nolabel keep(listavariables)]
```

En consecuencia, para acoplar, espacialmente hablando, un fichero debajo del otro, habrá que escribir estas instrucciones:

```
use fichero, clear  
append using fichero  
list, clean
```

De este modo, el listado de los cuatro casos que proceden de la unión de ambos ficheros es el siguiente:

#### ILUSTRACIÓN 3.14. Listado de casos de los ficheros unidos

	sexo	edad
1.	1	21
2.	2	20
3.	1	20
4.	2	19

#### 3.3.2. Combinación de ficheros

Si, en lugar de añadir casos, se desea adjuntar variables, se hablará de combinación de ficheros y, en lugar de emplear la instrucción *append*, tendrá que utilizarse *merge*, cuya sintaxis más elemental es la siguiente:

```
merge 1:1 _n using nombrefichero [, opciones]
```

Para ejemplificar esta operación, se utilizará como base el fichero anterior de cuatro casos y dos variables. A este se le combinará el archivo denominado *ficherox.dta*, que contiene en el mismo orden los mismos cuatro casos del otro archivo, con una variable de texto denominada *nombre*<sup>18</sup>.

---

<sup>18</sup> Para que el fichero en memoria se fusione con el guardado en un fichero, este último ha de tener el mismo orden y número de casos. En el capítulo 5 se explica con detalle cómo pueden ordenarse los ficheros, mediante la instrucción *sort*. De todos modos, la instrucción *merge*, desde la versión 11, posee otras muchas posibilidades, como la de emplear una variable para la correcta combinación de los casos y la de fusionar un registro con otros múltiples o viceversa. Se sugiere a quienes necesiten una u otra funcionalidad que consulten la ayuda disponible en el programa.

Para conseguir esta fusión, es preciso emplear la orden *merge* de la que conviene ver su resultado añadiéndole la instrucción *list*.

```
merge 1:1 _n using "ficherox"
list, clean
```

En el listado puede apreciarse una nueva variable (*\_merge*), en este ejemplo con todos los casos con un valor de 3, porque contienen valores procedentes tanto del fichero base como del combinado. Si el primero hubiera tenido más casos que el segundo, los últimos habrían tenido en esta variable el valor de 1 (sin datos combinados); mientras que si es el segundo el más completo, la nueva variable adoptaría el valor de 2 (sin datos en el fichero de base) en los casos adicionales que sólo tienen valores en las variables añadidas.

### ILUSTRACIÓN 3.15. Listado de casos de los ficheros combinados

	sexo	edad	nombre	_merge
1.	1	21	Juan	3
2.	2	20	Maria José	3
3.	1	20	Alfredo	3
4.	2	19	Isabel	3

Existe otra posibilidad de fusión de ficheros, de naturaleza similar a la combinación de ficheros. Se trata de juntar los dos archivos no por el orden en el que están situados los casos, sino por la coincidencia de una o un conjunto de variables. La instrucción correspondiente es *joinby* con la siguiente sintaxis.

```
joinby [varlist] using "nombrefichero"[, unmatched(none | both | master |using)
                                            _merge(nombrevar) update]
```

El ejemplo de esta instrucción usa como base el fichero listado en la ilustración 3.15, al que se le combina en función de dos variables la siguiente matriz de datos guardada en el archivo ficherou.dta, donde además del nombre (único<sup>19</sup>), se encuentra una tercera variable denominada region. La matriz de datos de este nuevo fichero presenta esta disposición:

---

<sup>19</sup> Para un correcto funcionamiento de la combinación de ficheros el conjunto de valores de las variables que siguen al *joinby* ha de ser único con el fin de que se produzca una combinación de caso por caso. El caso más frecuente para ello es una variable de identificación; pero también puede utilizarse un par de variables como en el ejemplo, o en el caso de que disponga de dos bases de datos temporales por países. En este último supuesto, la utilización de las variables *pais* y *año* pueden generar identificaciones únicas que permitan un correcto apareamiento de los datos de uno y otro fichero.

**ILUSTRACIÓN 3.16. Matriz de datos del archivo *ficherou.dta***

nombre	region
Juan	Sur
Alfredo	Norte
Maria José	Sur
Isabel	Norte
Carmen	Sur
José	Norte
Teresa	Norte
Pedro	Sur

Una manera de combinar adecuadamente ambos ficheros es la siguiente:

```
joinby sexo edad using ficherou, _merge(combina)
list, clean
```

Se le ha añadido la opción *\_merge*, para que no se superponga la variable *merge* generada en el anterior ejemplo. De este modo, tras una orden de listado se obtiene el resultado plasmado en la ilustración 3.17.

**ILUSTRACIÓN 3.17. Listado de casos de la conjunción de dos ficheros  
(opción por defecto)**

	sexo	edad	nombre	_merge	combina	region
1.	Hombre	21	Juan	3	both in master and using data	Sur
2.	Mujer	20	Maria José	3	both in master and using data	Sur
3.	Hombre	20	Alfredo	3	both in master and using data	Norte
4.	Mujer	19	Isabel	3	both in master and using data	Norte

Es preciso notar que sólo aparecen cuatro casos, pues cuando no se explicita la opción *unmatched*, Stata adopta su modalidad *none*; por lo que el fichero sólo contiene los individuos que están presentes en los dos ficheros conjuntados. También hay que advertir que además de la variable *\_merge*, aparece la denominada *combina*, etiquetada por el programa indicando que estos cuatro casos se encuentran tanto en el fichero en uso (*master*) como en el conjuntado (*using data*). Finalmente, es fácil comprobar que los casos —sin haberle indicado ninguna orden para ello— aparecen ordenados por las dos variables que se emplean para la conjunción.

El resultado sería diferente —incluiría los ocho casos del fichero conjuntado— si se optara por la opción la modalidad *both* o *using* de la opción *unmatched*, tal como se recoge en la próxima instrucción.

```
joinby sexo edad using ficherou, _merge(combina2) unmatched(both) update
list nombre sexo edad region combina2, clean
```

Así aparecen ocho casos, tantos como hay en uno u otro archivo. Y la nueva variable *combina2* indica en qué circunstancia combinatoria está cada uno de ellos. Se advierte que las filas 5, 6, 7 y 8 corresponden a los casos ausentes en el fichero maestro<sup>20</sup>.

**ILUSTRACIÓN 3.18. Listado de casos de la conjunción de dos ficheros (opción *both*)**

	nombre	sexo	edad	region	combina2
1.	Juan	Hombre	21	Sur	in both, master agrees with using data
2.	María	José	20	Sur	in both, master agrees with using data
3.	Alfredo	Hombre	20	Norte	in both, master agrees with using data
4.	Isabel	Mujer	19	Norte	in both, master agrees with using data
5.	Carmen	Mujer	22	Sur	only in using data
6.	Pedro	Hombre	.	Sur	only in using data
7.	José	Hombre	23	Norte	only in using data
8.	Teresa	Mujer	24	Norte	only in using data

### 3.4. Ejercicios

- 1) Crea con Stata una base de datos sencilla con los datos de sexo, edad y nombre de pila de su familia o de un grupo de amigos.
- 2) Entra en la página del INE ([www.ine.es](http://www.ine.es)), consulta el censo del 2001 (2011) y busca, por ejemplo, las cifras de los habitantes de las capitales de provincia. Descárgalas como un fichero *csv* y, a continuación, léelo con Stata. Nota que el INE pone encabezados y pie a los datos y, antes de ser leídos con el comando del programa estadístico, debería editarse el fichero bajado con un editor de textos (notepad, por ejemplo) a fin de eliminar las primeras y las últimas líneas. Otra cuestión a tener en cuenta es que el INE termina los registros con una coma, como si hubiera una última variable sin dato. Por ello, en el caso de que se lea con nombres de variables, habría que añadir una, que posteriormente puede eliminarse.
- 3) Haz lo mismo con la población por provincias. Después intenta combinar (*merge*) ambos ficheros.
- 4) Descarga un barómetro de la página del CIS, por ejemplo el estudio 2794 de marzo de 2009. Generalmente los estudios del CIS se componen de siete ficheros: la ficha técnica (Ft####), el cuestionario (cues####), el libro de códigos (codigo####), las tarjetas (tarjetas####), el programa en SPSS (ES####), en SAS (Sas####) y los datos en formato ASCII (DA####). Para los ejercicios de los siguientes capítulos, se va a utilizar el barómetro de marzo. Construye un fichero *do* con *infile*. Optativa-

<sup>20</sup> Como se ha expresado la opción *update*, incorpora en la variable *ideología* los valores del fichero conjuntado, en lugar del que está en uso, con sólo cuatro casos fruto de la anterior conjunción.

mente también coloca las etiquetas de variables y valores con las instrucciones aprendidas en el capítulo anterior. Como sugerencia, utiliza como guías el programa de SPSS (para *infile* y *label define*) y el de SAS (para *label variable*).

- 5) Realiza la misma operación que en el ejercicio anterior con otro barómetro, como el de abril de 2009 (estudio 2798). Une los dos estudios mediante la instrucción *append*, manteniendo sólo las variables que sean comunes a ambos.



# 4

## Estadísticas de una sola variable<sup>1</sup>

### 4.1. Clasificación de variables

Aunque informáticamente las variables se distingan por su longitud y su codificación textual o numérica (véase sección 2.6.2), desde un punto de vista estadístico, la mejor clasificación se fija en las características intrínsecas que tengan los valores, por más que estos suelan codificarse numéricamente, independientemente de sus propiedades. En el primer ejemplo de matriz de datos (ilustración 3.3) aparecieron dos variables con códigos numéricos, que en el fondo son diferentes. En la primera, sus códigos (“1” y “2”), aun de naturaleza cuantitativa, representaban cualidades: “Varón” y “Mujer”. La segunda presentaba valores con significado propiamente numérico: “19”, “20”, y “21”. No cabe la menor duda de que el tratamiento que se puede aplicar a una y otra variable ha de ser muy distinto. Una primera clasificación simple es la que se acaba de mencionar entre las variables cuyos valores son cualidades o categorías, también llamadas atributos, y aquellas cuyos valores son números con propiedades aritméticas. El sexo y la edad son ejemplos claros y respectivos de ambos tipos de variable. Pero también lo son la clase social (con sus distintas categorías) y los ingresos (expresados en dólares, pesetas o euros, pero, en todo caso, cantidades).

Entre las *variables cualitativas* se distinguen las *nominales*, cuyos valores sólo poseen la propiedad de la identidad (cualquier valor es igual a sí mismo y diferente del resto), y las *ordinales*, en las que puede establecerse una jerarquía completa entre valores, de manera que, si un valor llamado *a* está situa-

---

<sup>1</sup> No son centenares, sino miles, los libros y manuales de estadística básica que se han escrito desde la segunda mitad del siglo xx. En este contexto se recomienda a quienes empiezan a aproximarse a la estadística que complementen este capítulo y el de gráficos (6) con un buen manual de la materia. Entre ellos, se sugieren como clásicos Blalock (1966) y Spiegel (1970). También son buenas introducciones García Ferrando (1999), Peña y Romo (2003), así como Cuadras (1996). En una línea muy similar a estos capítulos, con programa distinto, se encuentra Escobar (1999). Útiles también son Neter *et al.* (1993), así como Hamilton (2009). Este último contiene, además, como este libro, las órdenes de Stata.

do antes de un segundo denominado  $b$ , a su vez, este precede a un tercero, al que se conocerá con  $c$ , necesariamente el primero ha de estar ubicado por delante del tercero. Ambas propiedades pueden formularse como sigue:

Principio de identidad:

$$\begin{aligned} a &= a \\ a &\neq b \end{aligned} \tag{4.1}$$

Propiedad ordinal de los valores:

$$a > b \wedge b > c \Rightarrow a > c \tag{4.2}$$

Por su lado, las *variables cuantitativas* pueden clasificarse en variables de *intervalo* o de *razón*, según carezcan o tengan un valor 0 que represente la ausencia total de la calidad que están representando. El cociente intelectual sólo puede ser clasificado de variable de intervalo, pues el valor 0 es arbitrario y no equivale a la carencia absoluta de inteligencia; en cambio, puede catalogarse como variable de razón a los ingresos medidos, por ejemplo, en euros, ya que en este caso el 0 indica la ausencia total de lo que expresa la variable. No se trata, como a veces suele confundirse, de que la variable tenga o no el valor 0 para catalogarla de una u otra forma, sino del significado que tiene este valor.

Otra clasificación útil para *variables cuantitativas* es la que separa a las *variables discretas* de las *variables continuas*. Teóricamente, las primeras son aquellas con limitado número de valores, de modo que entre dos valores contiguos es imposible encontrar empíricamente un tercero con un valor intermedio. Una persona puede tener dos o tres hermanos, pero no dos hermanos y medio. En cambio, en las *variables continuas* siempre será posible imaginar valores intermedios, pues el número de ellos es infinito. Así, entre una persona que pesa 60 kg y otra que pesa 61 kg, es posible encontrar otra con 60,5 kg; la única limitación estaría en la precisión de los instrumentos de medida.

Stata contiene una orden en la que se muestra un resumen de los valores que presentan todas las variables de una matriz (o fichero) o un conjunto de variables especificadas. Se trata de la instrucción *codebook*.

```
use fichero4a
codebook sexo edad
```

Aplicada a los datos mostrados como primer ejemplo en este capítulo, muestra los distintos valores que presentan las variables y sus correspondientes frecuencias, esto es, las veces que se repiten entre las unidades que componen la matriz.

### ILUSTRACIÓN 4.1. Libro de códigos de las variables *sexo* y *edad*

<pre>sexo   (unlabeled)</pre> <hr/> <pre>type: numeric (byte)</pre> <hr/> <pre>range: [1,2]                               units: 1 unique values: 2                           missing .: 0/4</pre> <hr/> <pre>tabulation: Freq.   Numeric   Label             2         1   Hombre             2         2   Mujer</pre>	<pre>edad   (unlabeled)</pre> <hr/> <pre>type: numeric (byte)</pre> <hr/> <pre>range: [19,21]                             units: 1 unique values: 3                           missing .: 0/4</pre> <hr/> <pre>tabulation: Freq.   Value             1     19             2     20             1     21</pre>
--	--

Es preciso insistir en que la variable *sexo*, aunque sea cualitativa, tiene sus valores guardados en formato numérico (1 y 2). Por ello, se recurre a etiquetarlos, el primero con “Hombre” y el segundo con “Mujer”. Ambos valores tienen una frecuencia de dos casos. Por otro lado, a las variables literalmente cuantitativas no procede ponerles etiquetas a los valores. Tampoco se les ha puesto en este ejemplo a las variables, puesto que su nombre (*sexo* y *edad*) son lo suficientemente aclaratorios como para que no requieran un título más explícito.

## 4.2. La tabla de distribución de frecuencias

La forma más elemental de resumir la información de un conjunto de datos es la tabla de distribución de frecuencias, que consiste en presentar para cada valor de una —y sólo una— variable el número (frecuencia) de casos que lo comparte. Siguiendo el ejemplo de la ilustración 3.3, de los cuatro casos presentes en la matriz de datos, dos son varones y dos mujeres. De igual modo, en la variable *edad* existen dos casos con el mismo valor (20 años), pero hay otros dos con valores únicos (19 y 21).

La disposición típica de una tabla de distribución de frecuencias consiste en:

- a) Exponer como encabezamiento el nombre de la variable.
- b) Listar en la primera columna el repertorio de los distintos valores que presenta la variable entre los sujetos en estudio.

- c) Mostrar en la segunda columna la frecuencia ( $f_i$ ) correspondiente a cada valor. Esta segunda columna se finaliza con la suma de todas las frecuencias, lo que equivale a expresar el número total de casos analizados.
- d) Crear una tercera columna con las proporciones o frecuencias relativas ( $p_i$ ), que consisten en el cociente entre las frecuencias simples y el número total de casos.

$$p_i = \frac{f_i}{\sum_{i=1}^I f_i} = \frac{f_i}{n} \quad (4.3)$$

Más útil aún es transformarlas en porcentajes, pues de esta forma son de más fácil interpretación y la comunicación con el lector u oyente resulta favorecida (ilustración 4.2).

- e) Además, para variables ordinales o cuantitativas, también resulta útil añadir una columna con los porcentajes acumulados ( $P_i$ ), que consisten en la suma progresiva de los porcentajes simples de la anterior columna.

$$P_i = \sum_{i=1}^i p_i = \frac{\sum_{i=1}^i f_i}{\sum_{i=1}^I f_i} \quad (4.4)$$

Para que Stata elabore la tabla de distribución de frecuencia de una sola variable hay que utilizar la instrucción *tabulate nombre\_de\_variable*. Si de desea con una sola instrucción solicitar más de una variable, hay que utilizar la orden *tab1 nombres\_de\_variables*, en lugar de la original *tabulate*. Así, para obtener las frecuencias absolutas, relativas y acumuladas de las variables *sexo* y *edad*, del actual ejemplo, habrá que escribir la orden:

```
tab1 sexo edad
```

El resultado muestra una variable seguida de la otra.

### ILUSTRACIÓN 4.2. Tablas de distribución de frecuencias de sexo y edad

-> tabulation of sexo				
sexo	Freq.	Percent	Cum.	
Hombre	2	50.00	50.00	
Mujer	2	50.00	100.00	
Total	4	100.00		

-> tabulation of edad				
edad	Freq.	Percent	Cum.	
19	1	25.00	25.00	
20	2	50.00	75.00	
21	1	25.00	100.00	
Total	4	100.00		

La ilustración 4.2 contiene las dos variables. De los cuatro sujetos en estudio, el 50% son hombres y el 50% mujeres. En relación con la edad, un 25% tienen 19; otro 25%, 21, y un 50% han cumplido 20 años. También puede decirse que el 75% de los sujetos tienen 20 años o menos, si de interpretar un porcentaje acumulado se trata.

En este tipo de tablas la notación que se emplea para designar a los valores es  $x_i$ , con  $f_i$  se denominan las frecuencias absolutas, las frecuencias relativas se reconocen por  $p_i$  y el número de casos se expresa bien con  $n$  si los datos corresponden a una muestra, o con  $N$  si se trabaja con los datos de una población. Por último,  $I$  denota el número de valores distintos que posee la variable. Cuando los valores de una tabla son exhaustivos y mutuamente excluyentes, son evidentes las siguientes igualdades:

$$\sum_{i=1}^I f_i = n$$

$$\sum_{i=1}^I p_i = 1 \quad (4.5)$$

Poco frecuentemente se realiza un estudio estadístico con tan sólo cuatro casos. A veces, la estadística ha sido definida como la ciencia de los grandes números, porque generalmente trata de describir grandes conjuntos, aunque para ello no necesite disponer de los datos de todos y cada uno de sus elementos. Se denomina *población* a ese gran conjunto del que se desea obtener una información, mientras que recibe el nombre de *muestra* un subconjunto de esa población extraído con unas determinadas condiciones que aseguren que el análisis que se efectúe con sus datos no difiera excesivamente del que se hubiese realizado teniendo la información de toda la población. El tamaño que han de tener las muestras depende principalmente de cuán homogénea u heterogénea sea la población y, en menor medida, del tamaño de esta última.

La matriz de datos, a partir de la que se obtienen las tablas de distribución de frecuencias, contiene tantas filas como casos tenga la muestra y tantas columnas como variables haya en la investigación. Tampoco es usual organizar una investigación con sólo dos variables, a menos que sean muy difíciles de medir. Por regla general, un estudio comprende un mínimo de diez variables y un máximo, en ocasiones escasas, de varios miles.

### 4.3. Estadísticos resúmenes de distribuciones

Las distribuciones son un resumen de los datos disponibles de las muestras generalmente, pues pocas veces se cuenta con los datos de la población. Se puede condensar aún más la información con la ayuda de los *estadísticos*, datos calculables en la distribución que dan cuenta de alguna característica notable. Cinco son las principales características que pueden resumirse en una distribución: la tendencia central, la posición, la dispersión, la simetría y el apuntamiento.

#### 4.3.1. *Medidas de tendencia central*

Por tendencia central se entiende un valor que representa al conjunto de valores de la distribución de una variable. En el caso extremo de una distribución en la que todos los sujetos tuvieran el mismo valor, ese dato daría cuenta de todos ellos. Pero, como su propio nombre indica, las variables no se caracterizan por presentar valores únicos. Por ello, hay diversos procedimientos para obtener una medida de tendencia central. Las más conocidas y empleadas son:

- a) La *moda*: valor que posee la mayor frecuencia de una distribución. Si en un grupo de cinco personas, tres son varones y dos mujeres, la moda es ser hombre. En la primera distribución de la ilustración 4.2, donde hay cuatro casos, no existe moda porque los dos valores poseen la misma frecuencia. En cambio, en la segunda distribución, en la de la edad, la moda es tener 20 años. Para que haya moda ha de existir un valor con mayor frecuencia que el resto.
- b) La *mediana* es el valor que ocupa la posición central de una distribución ordenada por sus valores. En consecuencia, no tiene sentido su cálculo en el caso de variables nominales. Para obtenerla hay que buscar en una tabla de distribución de frecuencias el primer valor cuya frecuencia relativa acumulada supere el 50%. Así, si se dispone de tres valores [4, 7, 6], la mediana es 6, pues previamente ordenados, es el que ocupa el medio de la distribución y es el primero cuyo porcentaje acumulado (66,6%) está por encima del 50%. En la variable *edad* de la ilustración 4.2, la mediana corresponde a dos valores, pues posee un número par de casos. Por convención, se adopta que la mediana sea la semisuma de los dos valores centrales. En este caso  $(20+20)/2$ , es decir, 20. Por tanto, para obtener la me-

diana cuando un determinado valor posea una frecuencia acumulada igual al 50%, es preciso calcular la semisuma con el siguiente valor de la tabla. En el caso de variables nominales es improcedente tanto el cálculo de la mediana como el de la media aritmética.

- c) La tercera medida de tendencia central es la *media aritmética*, que es un promedio de los valores de la distribución obtenido mediante la división de la suma de todos los valores por el número de casos. La cantidad ofrecida por la media es, utilizando un aforismo, el valor que tendrían todos los valores en el supuesto de que todos los valores tuvieran el mismo valor. Si en un grupo humano una persona tiene un hermano, otra dos y la tercera tres, poseen en total seis hermanos, que si se distribuyeran equitativamente corresponderían a dos por persona. La obtención de este estadístico responde a la siguiente fórmula:

$$\bar{x} = \frac{\sum_{i=1}^I x_i f_i}{\sum_{i=1}^I f_i} \quad (4.6)$$

Así, la media de edad en el grupo del ejemplo considerado sería de 20 años, que es el cociente entre la suma de las edades (80) de las cuatro personas y el número de miembros que la componen (4).

#### 4.3.2. *Medidas de localización*

Son medidas de localización aquellas que indican el valor que ocupa una determinada posición en una distribución ordenada. Las medidas más simples de localización son los valores mínimo y máximo, aquellos que se ubican en la primera y última posición de la tabla. En el caso de la edad, estos valores corresponden al 19 y al 21, respectivamente. Otra medida de localización es la mediana, también medida de tendencia central, pues es el valor que ocupa la posición del centro de la distribución o, dicho de otro modo, el 50% de las observaciones de la distribución tienen valores menores o iguales al de ella y el otro 50% tiene valores mayores o iguales. La mediana también puede ser concebida como aquel valor que divide a la distribución en dos partes iguales.

Otras medidas de localización son los cuartiles, que pueden ser definidos como tres valores que dividen a la distribución en cuatro partes iguales. Así, el primer cuartil tiene un 25% los de casos por debajo de dicho valor; el segundo cuartil coincide con la mediana y el tercero presenta un 25% de los casos con valores superiores. Para obtenerlos, se calculan, en primer lugar, las posiciones de los cuartiles — $O(Q_1)$  y  $O(Q_3)$ — y a partir de ellas se extraen los valores correspondientes. Las posiciones respectivas del primer, segundo (es igual a la mediana) y tercer cuartil son:

$$O(Q_1) = \frac{n+1}{4}; O(Q_2) = \frac{n+1}{2}; O(Q_3) = \frac{3(n+1)}{4} \quad (4.7)$$

Una vez obtenidas las posiciones, se buscan los valores que las ocupan. En el caso de que  $O(Qx)$  dé un valor decimal, se obtiene la semisuma de los valores que ocupan la parte entera de la posición y el que ocupa la siguiente. Así, en el ejemplo de la edad en la ilustración 4.2, dado que son cuatro casos, al primer cuartil le correspondería la posición 1,25 y al tercero la 3,75. En consecuencia, los valores del primer y tercer cuartil serían, respectivamente, de 19,5 y 20,5.

De la misma familia son los deciles y percentiles. En el primer caso, son nueve los valores que dividen la distribución en diez partes iguales y, en el segundo, 99 los que parten los datos en 100 subconjuntos del mismo tamaño. Para hallar lo n-tiles se procede de modo similar a cuando se obtienen los cuartiles. Se busca la posición correspondiente al n-ntil y si esta es decimal, se suman los dos valores contiguos y se dividen entre 2. En general, la posición de un n-ntil ( $T_x$ ) se ajusta a la siguiente fórmula:

$$O(T_x) = \frac{x(n+1)}{T} \quad (4.8)$$

De este modo, el quinto sextil de una distribución con 35 casos ocuparía la posición trigésima:  $5(35+1)/6$ .

#### 4.3.3. *Medidas de dispersión*

El tercer tipo de medidas son las llamadas medidas de dispersión. Indican cuán alejados están los valores de la distribución del valor que la representa, generalmente una medida de tendencia central. Los estadísticos de dispersión más utilizados son:

- a) La *dispersión modal* es la proporción (o porcentaje) de sujetos de una distribución que no tienen el valor modal. Este simple estadístico es uno de los escasos que se pueden utilizar para estudiar la dispersión en variables nominales u ordinales. Su fórmula se representa del siguiente modo:

$$D_{mo} = 1 - p_{mo} \quad (4.9)$$

Así, basta restar a 1 la proporción de casos que tienen la moda. En el ya conocido ejemplo del grupo de cuatro personas, la dispersión modal de la edad sería de 0,5 o, expresada como es común en porcentajes, del 50%, pues esta es la proporción de personas que no tienen 20 años, que es la moda. Stata no calcula este estadístico, pero es fácil de obtener con el cálculo de la proporción o porcentaje complementario.

- b) El *rango* es la diferencia entre los valores extremos de una variable. En el caso de la variable *edad* en el grupo de cuatro miembros que sirve de ejemplo, el rango toma el valor de 2 años, pues es la diferencia entre la edad (21) del mayor y la del menor (19).

$$R = x_{\max} - x_{\min} \quad (4.10)$$

Esta medida puede estar muy condicionada por un solo valor extremo poco representativo de lo que se estudia. Imagínese un grupo de 200 personas de edades comprendidas entre 17 y 18 años, salvo una que tiene 60. En este caso decir que el rango es de 43 años daría una imagen sesgada de este agregado. Por ello se utiliza frecuentemente el llamado rango intercuartílico, que es la diferencia entre los valores correspondientes al tercer y primer cuartil. Así, en el caso del grupo pequeño del ejemplo, sería de 1 año, y en el de los dos centenares de personas el rango intercuartílico sería también de 1 año.

$$R_I = Q_3 - Q_1 \quad (4.11)$$

- c) La *desviación media* es un promedio de los valores absolutos de las desviaciones de los valores con respecto a la media aritmética. Ha de advertirse que se trata de promedio de valores absolutos, pues si no se prescindiese del signo de las desviaciones, por una importante propiedad de la media aritmética, siempre arrojaría el valor de 0. En la distribución de la edad de los miembros del grupo hay dos desviaciones sobre la media (20 años): el más joven se desvía menos 1 año de la media, el mayor más 1 año, mientras que los otros dos tienen la misma edad que la media, por lo que no se desvían nada. La suma de estas cuatro desviaciones es 0, a menos que se añadan los valores sin considerar el signo que les precede, en cuyo caso la suma es de 2 años. De ahí se obtiene el promedio con la división de esta cantidad entre las cuatro personas que componen las observaciones realizadas, 0,5, que representa lo que se desvía en promedio cada caso de la media aritmética. Ello es obvio, pues dos casos se alejan de la media en 1 año y otros dos en ninguno. La fórmula para su cálculo en valores agregados es:

$$D = \frac{\sum_{i=1}^I |x_i - \bar{x}| f_i}{\sum_{i=1}^I f_i} \quad (4.12)$$

- d) La *varianza* es una media aritmética de las desviaciones al cuadrado de los valores con respecto a la media. En lugar de promediar los valores absolutos de las desviaciones, estas se elevan al cuadrado para

que su suma no sea 0 y, de este modo, se penalizan las desviaciones más alejadas de la media. Así, el cuadrado de una unidad de desviación sigue siendo 1; el de dos desviaciones es 4; el de tres, 9; el de 10, 100, y así, sucesivamente, va aumentando en progresión geométrica a medida que las desviaciones se hacen mayores. En la distribución del grupo de jóvenes, tanto el mayor como el menor se desvían 1 año al cuadrado de la media, mientras que los dos restantes no se desvían nada. En consecuencia, el promedio de años al cuadrado que se desvían estos cuatro sujetos de la media de 20 es de 0,5. Aquí iguala el resultado de la desviación media, porque para los valores de 0, de 1 y de -1 el valor al cuadrado es exactamente igual al valor al cuadrado. Pero lo normal es que la varianza sea mayor que la desviación media, salvo que las distancias de los valores al promedio sean menores que la unidad, pues sólo en esos casos el cuadrado es menor que el valor absoluto. Esta operación se formula del siguiente modo:

$$s^2 = \frac{\sum_{i=1}^I (x_i - \bar{x})^2 f_i}{\sum_{i=1}^I f_i} \quad (4.13)$$

- e) La *desviación típica* es la raíz cuadrada de la varianza. Se utiliza para devolver el valor de la varianza a sus unidades originales. Como acaba de verse, la varianza de 0,5 está referida en años cuadrados. Para poder hablar en términos de años, hay que hallar la raíz cuadrada de este valor, resultando ser de 0,7. Su cálculo se obtiene mediante la expresión:

$$s = \sqrt{\frac{\sum_{i=1}^I (x_i - \bar{x})^2 f_i}{\sum_{i=1}^I f_i}} \quad (4.14)$$

- f) El *coeficiente de variación* es una medida de dispersión relativa. Es el cociente entre la desviación típica y el valor absoluto de su correspondiente media aritmética. Al ser una razón o cociente, carece de unidades y, en consecuencia, se utiliza para comparar la dispersión entre variables que tengan distintas unidades de medida. Como la varianza y la desviación típica son siempre positivas, este coeficiente tampoco tiene sentido que sea negativo, aunque la media lo sea. Su valor es 0, como el de las tres medidas de dispersión precedentes, en el caso de que todos los valores de la variable sean idénticos y, salvo distribuciones muy dispersas, su valor suele ser inferior a 1.

$$CV = \frac{s}{\bar{x}} \quad (4.15)$$

#### 4.3.4. Medidas de simetría

Existen otras medidas cuyo propósito es expresar a través de un número la forma de la distribución. Estas se clasifican, a su vez, en dos tipos, las de simetría (que atienden a la forma horizontal de la distribución: si la izquierda de la distribución es semejante a su derecha) y las de apuntamiento (que indican la distribución vertical de los valores: si las frecuencias de los valores centrales son mayores que las de los valores extremos).

Para variables continuas existe un patrón o modelo de distribución de la estadística llamado *distribución normal* que, a primera vista<sup>2</sup>, se caracteriza por: a) tener idéntica la media, la moda y la mediana; b) ser simétrica, es decir, la distribución de los valores por debajo de la media se refleja como en un espejo en la distribución de los valores por encima del promedio (o viceversa), y c) poseer un alto número de casos en los valores centrales e ir descendiendo esta frecuencia a medida que los valores se van alejando del centro de la distribución, esto es, de la mediana.

Las dos primeras propiedades están muy ligadas entre sí, pues en toda distribución simétrica unimodal los tres estadísticos de tendencia central tienen los mismos valores. Para estimar la simetría de una distribución, se calcula el momento de orden 3 con respecto a la media, esto es, el promedio del cubo de las desviaciones de los valores con respecto a la media de la variable:

$$m_{\bar{x}}^3 = \frac{\sum_{i=1}^I (x_i - \bar{x})^3 f_i}{\sum_{i=1}^I f_i} \quad (4.16)$$

La fórmula del momento es de tal naturaleza que si hay predominio de valores por debajo (a la izquierda) de la media, sale negativo, y si hay predominio de valores por encima, resulta positivo. Para obtener un coeficiente de simetría estándar con el que poder hacer comparaciones entre variables se divide este momento de orden 3, cuyas unidades son cúbicas, por la desviación típica al cubo:

$$A = \frac{m_{\bar{x}}^3}{s^3} \quad (4.17)$$

---

<sup>2</sup> Véase la ilustración 4.18 como ejemplo gráfico de la distribución normal.

#### 4.3.5. *Medidas de apuntamiento*

La otra medida sobre la forma de la distribución es el *apuntamiento*, que indica cuán centradas o dispersas están las frecuencias de los valores en relación con el punto medio de la distribución. Si las frecuencias están concentradas en el centro, entonces la distribución se llamará *leptocúrtica*; si las frecuencias mayores se ubican en los extremos de la distribución, la distribución será *platicúrtica*, y, en el caso intermedio, sería una distribución *mesocúrtica*.

Para calcular el apuntamiento de una distribución, también denominado *curtosis*, se utiliza el momento de orden 4 con respecto a la media dividido, para que quede desprovisto de unidades, por la desviación típica a la cuarta. En algunos programas y manuales a este cociente se le restan tres unidades para que este estadístico arroje un valor de 0 en el caso de que se trate de una distribución normal. Sin embargo, en Stata el resultado se calcula sin la sustracción, de este modo:

$$K = \frac{m_{\bar{x}}^4}{s^4} = \frac{\sum_{i=1}^I (x_i - \bar{x})^4 f_i}{\sum_{i=1}^I f_i}$$
(4.18)

Tanto la medida anterior (asimetría) como esta, la curtosis, son útiles porque proporcionan claros indicios de cuándo la distribución de una variable cuantitativa es normal. Para que lo sea, la asimetría debe ser 0 y el apuntamiento igual a 3. Si alguna de estas medidas en una determinada variable no se ajusta a este patrón numérico, no cabrá duda de que no está distribuida normalmente.

### 4.4. Obtención de las medidas características de una distribución

Las medidas más importantes entre las que se acaban de enumerar en el apartado precedente pueden ser obtenidas mediante la orden *summarize*. Su sintaxis elemental consiste en acompañarla de las variables de las que se desea obtener los estadísticos en cuestión, pero, en el caso de que no se especifique ninguna de ellas, se sobreentiende que se pide los de todas.

**summarize** listaveriables [, opciones]

Para ver su funcionamiento se va a utilizar una parte de la base de datos de los países. En concreto, se utiliza la versión reducida, consistente en la inclusión de sólo los 15 países que forman parte de la UE a principios de

2004. Antes que nada es necesario, una vez puesto en funcionamiento el programa, abrir el fichero, y, antes de pedir los estadísticos, en la medida en que son pocos casos, también se solicita un listado de un subconjunto de variables del fichero:

```
use europa, clear
list pais superficie poblacion evn
```

El conjunto de los quince países tiene los siguientes valores en las tres variables solicitadas:

#### ILUSTRACIÓN 4.3. Listado de tres variables en quince países

	pais	superf~e	poblac~n	evn
1.	Alemania	349,300	82.200	77
2.	Austria	82,700	8.100	78
3.	Bélgica	33,200	10.300	78
4.	Dinamarca	42,400	5.400	76
5.	ESPAÑA	499,400	39.500	78
6.	Finlandia	304,600	5.200	77
7.	Francia	550,100	59.200	79
8.	Grecia	128,900	10.600	78
9.	Holanda	33,900	16.000	78
10.	Irlanda	70,283	3.800	76
11.	Italia	294,100	57.700	79
12.	Luxemburgo	2,586	0.438	77
13.	Portugal	92,082	10.200	76
14.	Reino Unido	241,600	59.900	77
15.	Suecia	449,964	8.900	80

La solicitud de los principales estadísticos se logra con la ya mencionada instrucción *summarize*:

```
summarize superficie poblacion evn
```

El resultado obtenido muestra una línea para cada variable:

#### ILUSTRACIÓN 4.4. Características de la distribución de tres variables

Variable	Obs	Mean	Std. Dev.	Min	Max
poblacion	15	25.16253	26.77402	.438	82.2
superficie	15	211674.3	185736.7	2586	550100
evn	15	77.6	1.183216	76	80

Los principales estadísticos que aparecen para cada variable son la media y la desviación típica, pero también se muestra el número de observaciones de las que se dispone, el valor mínimo y el valor máximo.

Como puede apreciarse en la ilustración 4.4, la media poblacional de los países de la Unión Europea es de 25,2 millones (la variable está introducida en estas unidades); el tamaño medio es de 211.674 km<sup>2</sup>, y la esperanza de vida al nacer promedio es de 77,6. Por su lado, las desviaciones típicas informan de que por término medio los países se alejan de la media de la población en 26,8 millones, de la de la superficie unos 186.000 km<sup>2</sup> y de la esperanza de vida al nacer 1,2 años.

La opción más utilizable en la instrucción *summarize* es *detail*, que sirve para aumentar el número de estadísticos mostrados como resultados. Así pueden obtenerse estadísticos adicionales de la variable *poblacion* añadiéndola.

```
summarize poblacion, detail
```

En consecuencia, debería aparecer un listado del siguiente tenor:

#### **ILUSTRACIÓN 4.5. Características de la distribución de una variable**

poblacion				
Percentiles				
1%	.438	Smallest	.438	
5%	.438		3.8	
10%	3.8		5.2	Obs 15
25%	5.4		5.4	Sum of Wgt. 15
50%	10.3		Mean 25.16253	
		Largest	Std. Dev. 26.77402	
75%	57.7			
90%	59.9		Variance 716.8482	
95%	82.2		Skewness .9399792	
99%	82.2		Kurtosis 2.344765	

En la primera columna aparece la serie de los nombres de los percentiles que se calculan de la variable; en la siguiente aparecen los valores obtenidos de los mencionados percentiles; en la tercera columna se listan tanto los cuatro valores menores como los cuatro mayores. Y en la última columna se muestran —además del número de observaciones—, la media y la desviación típica, la varianza, la asimetría y la curtosis.

Las estadísticas mostradas pueden leerse como sigue: los países de la Unión Europea con menos población tienen 438.000, 3.800.000, 5.200.000 y 5.400.000 habitantes. Los cuatro países más poblados tienen desde 57.700.000 habitantes hasta 82.200.000. El primer cuartil se encuentra en los 5.400.000 habitantes; el tercer cuartil, en los 57.700.000. La mediana está representada en 10.300.000 habitantes. Sin embargo, la media es bastante más alta: más de 25 millones de habitantes, y el promedio de las desviaciones asciende por encima de los 26 millones. Se trata, por tanto, de una variable muy dispersa (la desviación típica es mayor incluso que la media). Por otro lado, se trata de una distribución asimétrica a la derecha (el coeficiente de asimetría es positivo, cercano a 1), pues son más numerosos los países por debajo de la media que los que están

por encima de ella, y platicúrtica (la curtosis está por debajo de 3), pues no existe abundancia de países con población en torno a la media.

## 4.5. La ponderación de los datos

Por *ponderación estadística* se entiende la modificación del peso igualitario que originalmente poseen las observaciones en el conjunto de datos. Con un ejemplo sencillo se puede entender este procedimiento. Sea un examen que consta de cinco preguntas cortas y dos preguntas largas. Cada una de ellas está puntuada de 0 a 1. Si a todas estas preguntas se les da el mismo peso, la suma de las preguntas nos dará 7 puntos. Para que el resultado se encuentre en un rango de 0 a 10, pueden encontrarse múltiples soluciones. Las más simples son: ponderar igual cada respuesta, para ello habría que multiplicar la puntuación de cada una por la constante 10/7. Otro sistema sería que se diera más peso a las preguntas largas. Por ejemplo, la mitad de la nota para las dos preguntas largas y la otra mitad para las preguntas cortas. Si se opta por esta solución desigual para preguntas cortas y largas, las cinco primeras preguntas cortas tendrían cada una un peso de 1 punto, mientras que a las dos preguntas largas habría que otorgarles un peso de 2,5 puntos. De este modo, cada observación, en este caso cada pregunta, habría que transformarla multiplicándola por su peso del siguiente modo:

$$x'_i = x_i w_i \quad (4.19)$$

Siendo  $w_i$  el peso de cada observación, la nota final obtenida en el examen con las siete preguntas se obtendría con la siguiente fórmula:

$$P = \sum_{i=1}^7 x_i w_i \quad (4.20)$$

En este ejemplo los cinco primeros pesos (desde  $w_1$  a  $w_5$ ) tendrían un valor unitario, mientras que los dos últimos ( $w_6$  y  $w_7$ ) serían igual a 2,5. La suma de todas las ponderaciones es igual a 10, por lo que en el supuesto de que un sujeto puntúe con 1 las siete preguntas, el resultado  $P$  también sería igual a 10.

Con la misma lógica, es de frecuente aplicación estadística la denominada media ponderada, que consiste en obtener los promedios multiplicando cada valor, además de por su frecuencia, por su ponderación. Aparte de ello, también hay que incluir en el denominador de este promedio las ponderaciones a fin de equilibrar las frecuencias:

$$\bar{x} = \frac{\sum_{i=1}^I x_i f_i w_i}{\sum_{i=1}^I f_i w_i} \quad (4.21)$$

Stata cuenta con cinco modos distintos de ponderar los datos. Cuatro de ellos, que se expondrán en tres apartados<sup>3</sup>, son simples y se verán a continuación, mientras que el otro responde a la lógica de diseños muestrales complejos y se explicará en el último capítulo de esta obra. Es preciso tener en cuenta que no todos los procedimientos de ponderación son posibles en las instrucciones de Stata, por lo que para su uso es a menudo conveniente solicitar la ayuda de cada orden, pues en su contenido se indica qué posibilidades de pesos permite<sup>4</sup>.

1. El procedimiento de ponderación más admitido por las órdenes es *fweight*, que en realidad es un multiplicador de los casos por una constante. Este tipo de ponderación ha de ser, por tanto, entera y positiva.

La ocasión más frecuente y oportuna para el uso de esta instrucción es para cuando se dispone de datos ya tabulados, que se desean introducir en el ordenador para el cálculo de determinados estadísticos. Por ejemplo, se sabe que en una clase cinco alumnos han obtenido un 3; diez, un 4; 28, un 5; 19, un 6; 15, un 7; diez, un 8; dos, un 9, y uno, un 10. De este modo, la matriz de partida, en lugar de contener en cada fila un individuo, dispone de un valor distinto de una o varias variables, y una de las columnas, en lugar de ser una variable propiamente dicha, es el peso, o frecuencia, de los valores mencionados. Con la orden *list* se muestra la estructura de esta matriz de partida:

```
use calificaciones, clear
list, clean
```

Se puede apreciar en la ilustración 4.6 la columna donde aparece la variable (*nota*) y la que presenta sus correspondientes pesos (*frecue~a*).

#### **ILUSTRACIÓN 4.6. Matriz de datos agregados**

	nota	frecue~a
1.	3	5
2.	4	10
3.	5	28
4.	6	19
5.	7	15
6.	8	10
7.	9	2
8.	10	1

Sólo consta de ocho casos, que se corresponden con las ocho distintas calificaciones otorgadas (desde el 3 hasta el 10), y estas están acom-

<sup>3</sup> La razón de reducirlos a tres es porque dos de ellos, los analíticos y los muestrales, son tan parecidos que explicar sus diferencias está por encima del nivel de esta introducción. Además, para las órdenes hasta ahora analizadas en este manual no se permite el uso de *pweight* y hay que utilizar, por tanto, como recurso el otro medio de ponderación, que es *aweight*.

<sup>4</sup> Por ejemplo, *tabulate* sólo permite las ponderaciones de frecuencia, de importancia y analítica; en cambio, la orden *regress* permite además la ponderación probabilística.

pañadas de sus correspondientes frecuencias. Para que estas funcionen como variable ponderadora, ha de añadirse a la instrucción entre corchetes la palabra clave *fweight*, seguida del signo igual y del nombre de la variable que contiene el peso, denominada en este ejemplo *frecuencia*.

```
tab1 nota [fweight=frecuencia]
```

De este modo, la instrucción *tab1* se ejecuta con la variable *nota* ponderada con la llamada *frecuencia*. Por ello, en lugar de ocho casos, aparecen en el total 90.

#### ILUSTRACIÓN 4.7. Tabla de distribución de frecuencias de datos agregados

nota	Freq.	Percent	Cum.
3	5	5.56	5.56
4	10	11.11	16.67
5	28	31.11	47.78
6	19	21.11	68.89
7	15	16.67	85.56
8	10	11.11	96.67
9	2	2.22	98.89
10	1	1.11	100.00
Total	90	100.00	

La mayor parte de las órdenes que producen resultados estadísticos permiten utilizar la ponderación *fweight*. Siguiendo con el ejemplo anterior para solicitar los estadísticos de la distribución con todo detalle, la instrucción debería escribirse como sigue:

```
summarize nota [fweight=frecuencia], detail
```

Salvo los valores mínimos y máximos, para cuyo cálculo no se tiene en cuenta la ponderación, el resto de los estadísticos, desde los percentiles hasta la curtosis, se obtienen con los pesos otorgados:

#### ILUSTRACIÓN 4.8. Características de los datos agrupados

nota				
Percentiles		Smallest		
1%	3		3	
5%	3		4	
10%	4		5	Obs 90
25%	5		6	Sum of Wgt. 90
50%	6			Mean 5.8
		Largest		Std. Dev. 1.493055
75%	7		7	
90%	8		8	Variance 2.229213
95%	8		9	Skewness .3047134
99%	10		10	Kurtosis 2.766698

2. La segunda posibilidad de ponderación es *pweight* o *aweight*. Esta ha de expresar la inversa de la probabilidad de un sujeto de ser extraído en la muestra, o bien esta cantidad dividida por  $n$ .

Con un ejemplo, como en el primer tipo de ponderación, se ve más claramente el proceso y el resultado de esta operación.

Supóngase que se ha realizado un muestreo aleatorio: sobre una población de 100 personas, se han seleccionado diez. En la muestra, sin embargo, han salido cuatro hombres y seis mujeres, pese a que la proporción en el universo es del 50%. Para devolver a la muestra el peso que tienen ambos sexos en la población puede ponderarse por el coeficiente de elevación de cada una de las submuestras. Este coeficiente de elevación se obtiene mediante el cociente entre el tamaño de la población de un determinado estrato ( $N_k$ ) y el de la muestra ( $n_k$ ):

$$w_k = \frac{N_k}{n_k} \quad (4.22)$$

En la ilustración siguiente se presenta el listado de estos diez casos con sus correspondientes elevaciones.

#### ILUSTRACIÓN 4.9. Listado de elevaciones por caso

	sexo	elevac~n
1.	Hombre	12.5
2.	Hombre	12.5
3.	Hombre	12.5
4.	Hombre	12.5
5.	Mujer	8.333
6.	Mujer	8.333
7.	Mujer	8.333
8.	Mujer	8.333
9.	Mujer	8.333
10.	Mujer	8.333

Las correspondientes a los hombres son el resultado de dividir el número de hombres en la población (50) entre los cuatro de la muestra, mientras que la elevación de las mujeres es el cociente entre las 50 del universo y las seis seleccionadas.

Si se demandan frecuencias y media de la variable *sexo*, con la ponderación analítica se obtienen las frecuencias de la muestra y la media ponderada por este coeficiente de ponderación:

```
use fichero4b, clear
tabulate sexo [aweight=elevacion]
summarize sexo [aweight=elevacion]
```

Como puede comprobarse a continuación, la media de la variable *sexo* muestra la proporción de hombres en la población; al tiempo se conserva el tamaño de la muestra, aunque los pesos sumen una cantidad cercana a 100. Esto ocurre así porque con este procedimiento se *normalizan* los pesos a fin de que el total coincida con el tamaño de la muestra.

**ILUSTRACIÓN 4.10. Tabla de frecuencias y estadísticos ponderados analíticamente**

sexo	Freq.	Percent	Cum.
Mujer	4.999999	50.00	50.00
Hombre	5.000001	50.00	100.00
Total	10	100.00	
Variable	Obs	Weight	Mean Std. Dev.
sexo	10	99.999981	.50001 .5270463
			Min Max
			0 1

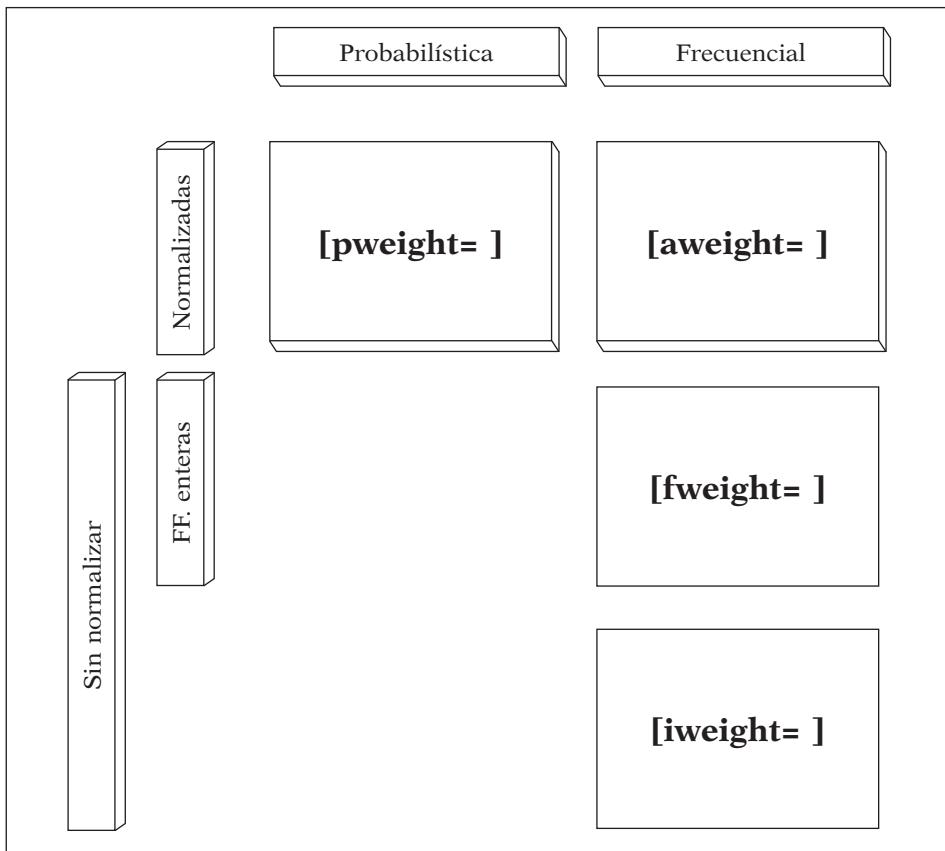
Es preciso advertir de que ni la orden *tabulate* ni la instrucción *summarize* permiten la ponderación probabilística (*pweight*). Stata recomienda utilizar en los casos en que se desee una mejor estimación los procedimientos propios de encuesta, que, aun siendo más complejos, proporcionan un cálculo más robusto de las desviaciones y los errores típicos. Para más detalles de este tipo de ponderaciones véanse las órdenes *svy* en el capítulo 14 de este libro<sup>5</sup>.

- Finalmente se explica a continuación el procedimiento *iweight* por el que a cada dato se le otorga una importancia discrecional. En el supuesto de que se emplee en este modo de ponderación el coeficiente de elevación, la frecuencia total mostrada es la de la población y no la de la muestra, como ocurría en los procedimientos *aweight* o *pweight*. Las órdenes para este tipo de ponderación son idénticas a las precedentes, sólo con el cambio de la palabra clave situada entre corchetes. De este modo, para obtener la tabla y los estadísticos de la ilustración anterior con el nuevo método, habría que dictar las siguientes órdenes:

```
tabulate sexo [iweight=elevacion]
summarize sexo [iweight=elevacion]
```

---

<sup>5</sup> En realidad, el ejemplo que se acaba de exponer debería haberse realizado con la opción *pweight*, porque se ha empleado como variable de ponderación el coeficiente de elevación. Stata recomienda el uso de *aweight* en aquellos casos en los que los datos de los que se dispone son medias o sumas de un conjunto de observaciones, como sucede si se dispone de información agregada de países, en cuyo caso el número de casos de cada medida es el que ha de emplearse como criterio de ponderación.

**CUADRO 4.1.** Procedimientos simples de ponderación en Stata

El resultado muestra un total de 100 casos en la tabla y una desviación típica algo menor en este caso, porque para obtenerla está dividiendo por 99 casos, en lugar de nueve, como en el anterior (véase *infra* [4.31] el motivo de restar 1 al número de casos).

**ILUSTRACIÓN 4.11.** Tabla de frecuencias con ponderación discrecional

sex	Freq.	Percent	Cum.			
Mujer	49.999981	50.00	50.00			
Hombre	50	50.00	100.00			
Total	99.999981	100.00				
Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
sex	10	99.999981	.50001	.5025189	0	1

Se pueden resumir estas ponderaciones afirmando que, por un lado, se encuentra la ponderación probabilística (*pweight*) y, por otro lado, las frecuenciales (el resto). También hay que señalar que tanto *pweight* como *aweight* son ponderaciones normalizadas, de modo que el resultado final queda con el mismo número de casos que la muestra, mientras que las otras dos (*fweight* e *iweight*) transforman el *n* del análisis. Y, finalmente, estas dos últimas se distinguen porque la primera necesita ponderaciones enteras para poderse llevar a cabo. El cuadro 4.1 muestra estas diferencias esquemáticamente.

## 4.6. El error típico

En todo este capítulo se ha tomado como referencia el análisis descriptivo de las variables en las muestras. Sin embargo, en la casi totalidad de las veces en que se trabaja con muestras, los datos que interesan no son los de estas, sino los de la población de la que proceden.

Generalmente, los cálculos que se extraen de la población reciben el nombre de *parámetros* y se les suele representar con una letra griega. De este modo, una media calculada con los datos de toda la población será considerada como un parámetro y se la notará como ( $\mu$ ). Del mismo modo, es también un parámetro la desviación típica ( $\sigma$ ), siempre y cuando se obtuviera con todos los sujetos de una población. Por el contrario, si, como suele ser usual, estas medidas se calculan con datos muestrales, reciben el nombre de *estadísticos* y se les reconocerá con los símbolos  $\bar{x}$  y  $s$ , respectivamente.

Es lógico que el resultado de un estadístico no coincida exactamente con el valor correcto del parámetro. A la diferencia entre uno y otro se le denomina *error muestral*. Todo estadístico tiene su correspondiente error, que se denominará  $e$ , acompañado del subíndice correspondiente. A modo de ejemplo, a continuación se exponen las fórmulas de los errores empíricos de la media, de una proporción y de la desviación típica, sin duda, los tres estadísticos más empleados en el análisis estadístico univariable:

$$\begin{aligned} e_{\bar{x}} &= \bar{x} - \mu \\ e_p &= p - \pi \\ e_s &= s - \sigma \end{aligned} \tag{4.23}$$

Caso de que se realicen muestreos aleatorios simples, seleccionando los elementos de las muestras uno a uno del conjunto de la población, puede procederse de dos modos: con reposición y sin reposición. En el primer método, los sujetos de la población que han sido seleccionados para formar parte de la muestra pueden volver a ser elegidos, formándose de este modo subconjuntos

con elementos repetidos. En cambio, en los muestreos sin reposición, una vez seleccionado un caso, no puede volverse a escoger y, en consecuencia, los elementos de la muestra son únicos e irrepetibles. Tanto por el sentido como por el menor error muestral que generan, son mucho más útiles y empleadas las muestras sin reposición que las que se realizan con reposición. Aquí, por ello, sólo se hará un análisis de los errores de las muestras sin reposición.

Un concepto imprescindible para abordar el problema del error muestral desde un punto de vista probabilístico, en lugar de empírico, es el de *distribución muestral*. Consiste en el comportamiento de un determinado estadístico en el conjunto de muestras de un determinado tamaño que puede extraerse de una población dada. Se comprende mucho mejor a través de un pequeño ejemplo.

Se supone una población de sólo cuatro sujetos y se desea a partir de ella obtener una muestra de dos personas. El número posible de muestras viene determinado por el número combinatorio siguiente:

$$\binom{N}{n} \quad (4.24)$$

En consecuencia, de una población de cuatro elementos pueden extraerse seis muestras diferentes. Más concretamente, a continuación se considera el conjunto de cuatro sujetos presentados en la matriz de la ilustración 4.12:

#### **ILUSTRACIÓN 4.12. Matriz de una población con cuatro elementos**

	sexo	edad
1.	1	21
2.	2	20
3.	1	20
4.	2	19

Si este conjunto es considerado una población y hubiera que realizar todas las muestras posibles de tamaño dos sin reposición, las seis posibilidades serían las siguientes:

#### **ILUSTRACIÓN 4.13. Distribución muestral sin reposición de la población anterior**

Hombre de 21	con	Mujer de 20
Hombre de 21	con	Hombre de 20
Hombre de 21	con	Mujer de 19
Mujer de 20	con	Hombre de 20
Mujer de 20	con	Mujer de 19
Hombre de 20	con	Mujer de 19

En cada una de estas muestras pueden calcularse una serie de estadísticos. Para mayor concreción, a partir de la variable *sexo* puede obtenerse para cada muestra la proporción de hombres (o mujeres) presentes en ellas:

**ILUSTRACIÓN 4.14. Distribución muestral del estadístico *p***

.5
1.0
.5
.5
.0
.5

A partir de estas posibles muestras de tamaño dos con sus respectivos porcentajes de hombres, puede construirse su correspondiente distribución<sup>6</sup>:

```
use "distribucion muestral"
tabulate phombres
```

El resultado proporciona las frecuencias de muestras en las que sale un 0%, 50% y 100% de hombres. Obviamente, con muestras de tamaño dos, no puede salir otro resultado.

**ILUSTRACIÓN 4.15. Distribución de probabilidad del estadístico *p***

phombres	Freq.	Percent	Cum.
0	1	16.67	16.67
.5	4	66.67	83.33
1	1	16.67	100.00
Total	6	100.00	

Finalmente, también es útil, además de contemplar su distribución, calcular sus características

**ILUSTRACIÓN 4.16. Características del estadístico *p* en la distribución muestral**

phombres			
Percentiles		Smallest	
1%	0	0	
5%	0	.5	
10%	0	.5	Obs 6
25%	.5	.5	Sum of Wgt. 6
50%	.5		Mean .5
		Largest .5	Std. Dev. .3162278
75%	.5	.5	Variance .1
90%	1	.5	Skewness 0
95%	1	.5	Kurtosis 3
99%	1	1	

<sup>6</sup> No hay que confundir la distribución muestral con la distribución de una muestra. Esta última es la distribución de una variable empírica en la muestra, mientras que la distribución muestral es una distribución probabilística de una variable aleatoria de los estadísticos calculables en el conjunto de muestras de un determinado tamaño que se puede extraer de una población dada.

Como puede apreciarse, de las seis posibles muestras, una —el 16,6%— presentaría un 0% de hombres; otra —otro 16,6%— mostraría el 100% de varones, y finalmente cuatro muestras —las dos terceras partes— tendrían un 50% de personas masculinas. De estos datos se deduce que con un 66,67% de probabilidad un muestreo de dos personas sobre una población de cuatro en la que la mitad posee una determinada característica mostraría un error nulo en la proporción de hombres, puesto que cuatro de las seis muestras posibles tienen un 50% de ellos, cantidad idéntica a la de la población.

Además, se comprueba empíricamente en este ejemplo que la esperanza matemática (el promedio o *mean*) de la distribución muestral es igual al parámetro de la población. Esto mismo expresado algebraicamente presenta la siguiente equivalencia:

$$E(p) = \Pi \quad (4.25)$$

También puede conocerse con exactitud, a partir de los datos de la población, no sólo la esperanza matemática de la distribución muestral, sino su varianza (*Variance*) y, en consecuencia, su desviación típica (*Std Dev.*).

$$Var(p) = \frac{\Pi(1 - \Pi)}{n} \frac{N - n}{N - 1} \quad (4.26)$$

Precisamente, la raíz cuadrada de la fórmula anterior es la desviación típica de la distribución muestral del estadístico  $p$ , que también recibe el nombre de *error típico*.

Este error típico tiene una importancia central en la estadística inferencial, puesto que es la herramienta imprescindible para el cálculo de los errores muestrales probabilísticos para las estimaciones por intervalo de los parámetros y para la realización de pruebas estadísticas.

Para incidir en su comprensión se expone a continuación un nuevo ejemplo con la misma población, pero en esta ocasión, en lugar de con una variable nominal (cualitativa), con una variable de razón como la edad.

Sabiendo que los valores en la población de cuatro sujetos son de 19, 20, 20 y 21, la distribución muestral de las muestras de tamaño dos presenta la siguiente disposición:

#### ILUSTRACIÓN 4.17. Distribución de probabilidad de la media

medad	Freq.	Percent	Cum.
<hr/>			
19.5	2	33.33	33.33
20	2	33.33	66.67
20.5	2	33.33	100.00
<hr/>			
Total	6	100.00	

En este supuesto, la probabilidad de que la muestra tenga un error de 0 es del 33,3%. Este porcentaje recibe el nombre de *nivel de confianza* y siempre ha de estar relacionado con un margen de error. De este modo, con los datos de la distribución muestral de la misma tabla, puede decirse que con un 100% de confianza o seguridad el error muestral se mantiene en el intervalo de  $\pm 0,5$ .

En cualquier caso, se mantienen una serie de características en la distribución muestral similares a las que se han expuesto con respecto a las proporciones o porcentajes. Estas son:

- a) La esperanza matemática del estadístico en la distribución muestral es igual al parámetro de la población.

$$E(\bar{x}) = \mu \quad (4.27)$$

- b) La desviación típica (o error típico) del estadístico en la distribución muestral es igual a la desviación típica de la población dividida por la raíz cuadrada del tamaño de las muestras, multiplicada por un factor de corrección<sup>7</sup> en el caso de que la muestra se realice sin reposición.

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (4.28)$$

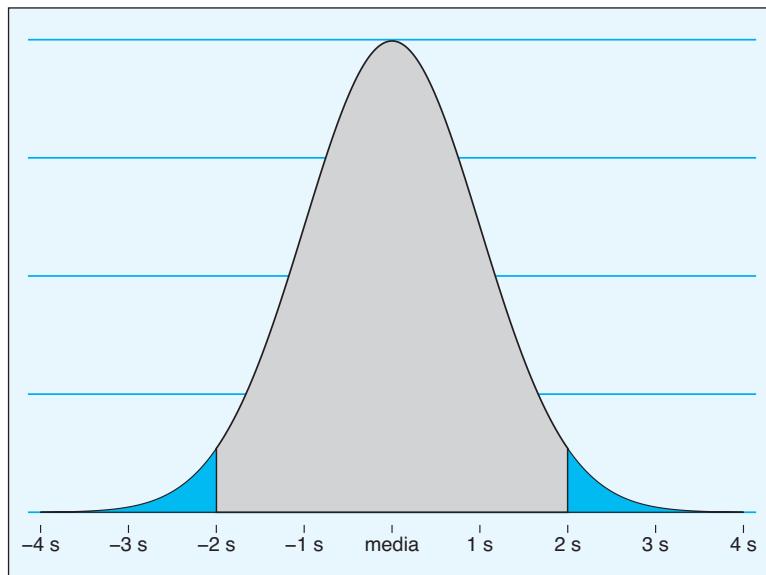
A estas características se suma que, de acuerdo con el teorema del límite central, para casi todas las poblaciones, la distribución muestral de una media (y las proporciones o porcentajes también pueden considerarse como medias) es aproximadamente normal cuando el tamaño de la muestra aleatoria simple es lo suficientemente grande.

Debido a esta distribución muestral que presentan la casi totalidad de las muestras aleatorias practicadas en la investigación, puede averiguarse con el único conocimiento de la varianza o desviación típica de la población el margen de error que presenta una potencial muestra con un determinado nivel de confianza. Esto es así porque se conoce que en una distribución normal existe un determinado porcentaje de casos con valores comprendidos entre un determinado rango de desviaciones típicas. El ejemplo más empleado es el correspondiente a  $\pm 2$  desviaciones típicas, puesto que el 95,5% de las unidades en una distribución se encuentran situadas en una posición no más alejada de dos desviaciones típicas por arriba o por debajo de la media, tal como están representadas en la zona central más oscura del gráfico 4.1. En este caso, las unidades serían muestras. El 0 en el eje de abscisas

<sup>7</sup> Para la varianza, la fórmula de este coeficiente corrector es  $\frac{N-n}{N-1}$

representaría las muestras cuyos estadísticos coinciden totalmente con el parámetro de la población, los valores positivos indican un estadístico mayor que el parámetro: una, dos o tres desviaciones típicas (o valores intermedios e incluso superiores) por encima de él. Por el contrario, los valores negativos se refieren a muestras con estadístico más bajo que el parámetro de la población. Como se trata de una distribución normal, es muy poco probable que se encuentre una muestra con tres desviaciones típicas (errores típicos) por debajo del valor correcto del parámetro.

**GRÁFICO 4.1. Zona central ( $\pm 2\sigma$ ) de la distribución normal**



En el supuesto de disponer de una población infinita<sup>8</sup> en la que la media de una variable de valoración a un personaje público fuera 6 y su desviación típica tuviera un valor de 2, al realizar muestras de tamaño cien, se genera una distribución muestral cuya media sería también 6, pero su desviación típica sería de 0,2 ( $2/\sqrt{100}$ ). En consecuencia, el 95,5% de las muestras presentarían una media comprendida entre 5,6 y 6,4, esto es,  $6 \pm (2 \times 0,2)$ . Esto puede expresarse formalmente del siguiente modo:

$$\Pr(\mu - z_c \sigma_{\bar{x}} \leq \bar{x} \leq \mu + z_c \sigma_{\bar{x}}) = 0,955 \quad (4.29)$$

<sup>8</sup> En las poblaciones infinitas (con más de 100.000 sujetos en la práctica) el coeficiente corrector comentado en la nota anterior se convierte en un número muy próximo a la unidad, con lo que no tiene ninguna incidencia en el error típico.

#### 4.6.1. Estimación e intervalos de confianza

En el apartado anterior se ha explicado el error muestral y el de su correspondiente nivel de confianza partiendo desde la población. Sin embargo, en el trabajo de análisis de datos, generalmente, no se dispone de la información del universo, sino de la muestra. A partir de esta, se pueden predecir los verdaderos parámetros de la población. Esta operación recibe el nombre de *estimación*, que a su vez puede realizarse de dos modos: puntualmente y por intervalos.

- a) La *estimación puntual* consiste en proporcionar un solo valor para el parámetro en cuestión. En los estadísticos más simples, la estimación puntual más certera es el estadístico con la misma denominación que el parámetro de la población. De este modo, el mejor estimador de  $\Pi$  (la proporción en la población) es  $p$  (la proporción en la muestra), y el mejor estimador de la media en la población ( $\mu$ ) es la media de la muestra ( $\bar{x}$ ). Sin embargo, esto no es así en el caso de la varianza, ni en el de la desviación típica, porque las ecuaciones (4.25) y (4.27) no se aplican a estos estadísticos. En cambio, con una adecuada demostración (Peña, 1989a: 274) puede comprobarse que...

$$E(s^2) = \sigma^2 \frac{n - 1}{n} \quad (4.30)$$

De este modo, en las muestras, en lugar de  $s^2$ , se calcula el estadístico, cuya fórmula, similar a la de la varianza (4.13), viene dada por:

$$\hat{s}^2 = \frac{\sum_{i=1}^I (x_i - \bar{x})^2 f_i}{n - 1} \quad (4.31)$$

En este caso, se cumple efectivamente la igualdad siguiente, lo que implica por definición que es un estimador no sesgado de  $\sigma^2$ .

$$E(\hat{s}^2) = \sigma^2 \quad (4.32)$$

Adicionalmente, la *estimación por intervalos* consiste en proporcionar un rango de valores en el que con una determinada probabilidad (el nivel de confianza) se encontrará el valor de la población. La obtención de estos intervalos se realiza sumando y restando al estadístico de la muestra su correspondiente error muestral:

$$\begin{aligned}
 p - \varepsilon_p &\leq \Pi \leq p + \varepsilon_p \\
 \bar{x} - \varepsilon_{\bar{x}} &\leq \mu \leq \bar{x} + \varepsilon_{\bar{x}} \\
 \hat{s}^2 - \varepsilon_{\hat{s}^2} &\leq \sigma^2 \leq \hat{s}^2 + \varepsilon_{\hat{s}^2}
 \end{aligned} \tag{4.33}$$

Stata permite construir los intervalos de confianza para proporciones y medias a partir de una orden de empleo inmediato. Se trata de la orden *ci*, con opciones *binomial*, para aplicarla a proporciones, y *level(#)*, para expresar el nivel de confianza con el que se desea contar.

Algunos ejemplos pueden aclarar el empleo y la interpretación de esta orden. Los dos primeros emplearán la fórmula inmediata de la orden *ci*. Esta consiste en un modo de proporcionar datos sin necesidad de que estos estén en un fichero. Una parte considerable de las instrucciones de Stata permiten esta posibilidad. Entre ellas está la orden que obtiene los intervalos de confianza.

Entre las órdenes inmediatas de intervalos de confianza, la más simple es la que se refiere a los intervalos correspondientes a una variable binomial. Tan sólo hay que proporcionar el número de casos y el de favorables, es decir, aquellos que cumplen una determinada característica. El caso más común de aplicación sería el de un simple juego de azar, como el lanzamiento de una moneda. Sea que de 100 lanzamientos se obtengan 40 caras, en la orden han de figurar en primer lugar las veces que se realiza el experimento (el número de casos, en el supuesto de una muestra) y posteriormente el número de resultados favorables, número de caras en este contexto.

ci 100 40

Entonces, el intervalo de confianza con un 95% de probabilidades estará comprendido, como señala la ilustración 4.18, entre el 30% y el 50%.

#### ILUSTRACIÓN 4.18. Cálculo directo de los intervalos de confianza de una media

Variable	Obs	Mean	Std. Err.	-- Binomial Exact --	
				[95% Conf. Interval]	
	100	.4	.0489898	.3032948	.5027908

Es preciso notar en este ejemplo que no utiliza la distribución normal<sup>9</sup>, sino la binomial. Esto ha de hacerse así por ser una variable dicotómica de naturaleza cualitativa. Sin embargo, dado el número de elementos, el resultado con la distribución normal sería muy similar.

<sup>9</sup> Para muestras pequeñas, y disponiendo del error típico con datos muestrales, en lugar de datos poblacionales, es más apropiado emplear la distribución *t* de Student que la normal, y así lo hace Stata.

Si se desea obtener la estimación por intervalos partiendo del modelo normal hay que proporcionar a la orden directa tres parámetros, en lugar de dos: número de casos, media y desviación típica de la población a ser posible. Los mismos datos anteriores con el modelo normal serían 100 casos, la media sería de 0,4, en lugar de 40, ya que hay que partir esta cantidad por el número de unidades de la muestra y la desviación típica sería la raíz cuadrada de  $p(1-p)$ , en este caso de  $0,40 \times 0,60$ :

```
cii 100 .4 sqrt(.40*.60)
```

Los resultados de esta última orden son muy semejantes a la anterior, como se puede comprobar comparando las ilustraciones pertinentes:

**ILUSTRACIÓN 4.19. Cálculo directo de los intervalos de confianza de una proporción**

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
	100	.4	.0489898	.3027936 .4972064

En el intervalo inferior las diferencias son de 1 milésima, mientras que en el superior apenas llegan a las 5 milésimas.

Pero, con datos de un fichero, debe usarse la orden principal, que es *ci*, en lugar de la del cálculo inmediato (*cii*).

Como ejemplo se utiliza en esta ocasión la muestra postelectoral del CIS, de la que se dispone de más de 5.000 sujetos. En primer lugar, se hace la estimación por intervalos de una variable cuantitativa, la edad. En este caso, basta con seguir la instrucción *ci* del nombre de la variable de la que se desean obtener los intervalos.

```
ci edad
```

Y con los propios datos muestrales se calcula el error típico y los correspondientes intervalos de confianza:

**ILUSTRACIÓN 4.20. Cálculo de los intervalos de confianza de una media**

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
Edad	5283	45.96498	.2529001	45.46919 46.46077

Descartando más de un decimal, puede decirse a la luz del resultado que con un 95% de seguridad la edad de la población estará comprendida entre 45,5 y 46,5 años.

En el caso de variables nominales, hay que hacer la estimación para cada uno de los valores o categorías. De este modo, no puede hacerse la estimación para la variable *sexo*, sino que habría que convertir esta variable en dos dicotómicas (con valores 1 para una categoría y 0 para el resto): *hombre*, por un lado; *mujer*, por el otro<sup>10</sup>. Y, con ellas, se aplica la instrucción *ci*, con la opción *binomial*.

```
use panel4
ci hombre mujer, binomial
```

De este modo se generan dos estimaciones por intervalos basadas en el modelo binomial, una para cada una de las dos categorías convertidas en variables dicotómicas:

#### **ILUSTRACIÓN 4.21. Cálculo de los intervalos de confianza de una proporción**

Variable	Obs	Mean	Std. Err.	-- Binomial Exact --	
				[95% Conf. Interval]	
hombre	5283	.4768124	.0068717	.4632655	.490385
mujer	5283	.5231876	.0068717	.509615	.5367345

Este resultado implica que en la población, con una confianza del 95%, el porcentaje de hombres estará comprendido entre el 46% y el 49%, mientras que el de mujeres lo estará entre el 51% y el 54%. Como puede fácilmente apreciarse, los unos son complementarios de los otros.

## **4.7. Ejercicios**

1. Pon ejemplos de variables que puedan considerarse nominales, ordinales o de razón del cuestionario del barómetro de marzo de 2009 del CIS (cis2794).
2. Obtén del conjunto de datos anterior la distribución de frecuencias de la comunidad autónoma y del tamaño del municipio. ¿En qué comunidades hay mayor número de entrevistados? ¿En qué estrato se concentra la mayor parte de los entrevistados? ¿Cuál sería la mediana del hábitat?
3. En el barómetro de marzo de 2009, dedicado a los trámites administrativos de los ciudadanos, como en la mayor parte de las encuestas, existe una mínima presencia de variables cuantitativas. De este tipo, en concreto, en el cuestionario del mencionado estudio sólo existen estrictamente cuatro variables, ubicadas en dos preguntas: la P.28 y la P.33. Las

<sup>10</sup> Cómo hacer esta operación se explica en el próximo capítulo, dedicado a la transformación de variables.

tres variables de la pregunta 28 son horas y minutos diarios dedicados al tiempo libre y su alternativa horas de ocio a la semana. La variable de la pregunta 33 (*edad*) se encuentra prácticamente en todos los cuestionarios dirigidos a personas. Calcula e interpreta todos los estadísticos convenientes a las mencionadas variables que han sido abordados en este capítulo.

4. En ese fichero también existe una variable, denominada *peso*, cuya función es equilibrar (ponderar) la muestra por sexo y edad. Construye una tabla de frecuencias para el sexo y un sumario de estadísticas para la edad, con y sin ponderación, observando las diferencias. Hazlo también para otra variable nominal, otra ordinal y otra de intervalo. Razona la selección del tipo de ponderación.
5. En ese mismo barómetro de marzo de 2009 (fichero: cis2794.), calcula los intervalos de confianza para la variable *edad* (P.32).
6. Del mismo fichero obtén los intervalos de confianza binomiales para la posesión de cada uno de los aparatos domésticos contenidos en la pregunta 42.



# 5

## Manipulación y modificación de datos

Este capítulo explica las transformaciones que pueden realizarse con Stata sobre la matriz de datos. Se distinguirán dos grandes tipos de transformaciones:

1. Manipulación de datos: no se produce ningún cambio en los datos de la matriz, sino que los datos son reordenados, reagrupados o seleccionados para realizar alguna operación sobre ellos.
2. Modificación de datos: los datos de la matriz son modificados. Hay varios tipos posibles de modificación: generación de nuevas variables, modificación de las ya existentes y modificaciones directas sobre la matriz.

### 5.1. Manipulación de datos

En numerosas ocasiones es preciso reordenar, reagrupar o seleccionar los datos para realizar determinadas operaciones con ellos. Stata incorpora varias instrucciones específicas para estas operaciones.

#### 5.1.1. Ordenación de casos

La instrucción específica de Stata para ordenar los casos es *sort*. Según se vio en un capítulo anterior, la matriz de datos se puede representar por una rejilla en la que los casos son filas y las variables, columnas. Generalmente, el orden de los casos no importa: es igual que un caso aparezca el primero en la matriz o el último, lo importante son los valores de las variables en todos y cada uno de los casos. Pero, para algunos procedimientos estadísticos, es necesario que los casos estén dispuestos de una determinada manera. Por ejemplo, es necesario ordenar los casos, según sus valores en una determinada variable, para realizar análisis paralelos de distintos segmentos de la muestra con la preinstrucción *by*, cuyo funcionamiento se verá más adelante, en la sección 5.1.3.

Para ordenar los casos de menor a mayor según los valores de una variable, sólo hay que teclear *sort* seguido del nombre de la variable en cuestión. Por ejemplo, si se dispone de la última matriz de datos del capítulo 3:

### ILUSTRACIÓN 5.1. Matriz de datos con ocho casos

	sexo	edad
1.	1	20
2.	1	21
3.	1	23
4.	1	.
5.	2	19
6.	2	20
7.	2	22
8.	2	24

Y si se desea ordenar por edad, hay que introducir las siguientes órdenes:

```
use fichero5, clear
sort edad
list sexo edad, clean
```

Con lo que los datos quedarán ordenados del siguiente modo:

### ILUSTRACIÓN 5.2. Resultados de la instrucción *list* después de ordenación por edad

	sexo	edad
1.	Mujer	19
2.	Mujer	20
3.	Hombre	20
4.	Hombre	21
5.	Mujer	22
6.	Hombre	23
7.	Mujer	24
8.	Hombre	.

Como se puede ver, los casos quedan ordenados de menor a mayor, en este ejemplo, del individuo más joven al de mayor edad en la matriz de datos. Nótese que hay un valor perdido, que Stata ha colocado el último en la matriz. Esto es así porque el programa asigna internamente los valores superiores a los valores perdidos, por lo que los coloca al final en la ordenación, aunque nunca los utilice para los procedimientos estadísticos<sup>1</sup>.

A la instrucción *sort* se le puede especificar más de una variable para su ordenación. En ese caso, ordenará primero los casos según su valor en la primera variable, y después según su valor en la segunda variable, etc. Por ejemplo, si se escribe...

```
sort sexo edad
list sexo edad, clean
```

---

<sup>1</sup> La instrucción *sort* es irreversible. Una vez ordenados por un criterio no se puede volver al anterior. Si se necesitara, habría que abrir de nuevo el conjunto de datos mediante *use* u

... el orden en que quedarán los casos será el siguiente:

**ILUSTRACIÓN 5.3. Resultado de la instrucción *list* tras ordenación por dos variables**

	sexo	edad
1.	Hombre	20
2.	Hombre	21
3.	Hombre	23
4.	Hombre	.
5.	Mujer	19
6.	Mujer	20
7.	Mujer	22
8.	Mujer	24

Los valores quedan ordenados en función del sexo primero, y dentro de cada sexo, en función de la edad.

La instrucción *sort* sólo puede ordenar los casos en orden ascendente (de menor a mayor). Si se requiere ordenar los casos de mayor a menor, ha de recurrirse a la instrucción *gsort*, que es una extensión de *sort* con más opciones. Siguiendo con el ejemplo anterior...

```
gsort -edad
```

... ordenaría de mayor a menor los casos según la edad. Por otro lado, *gsort +edad* produce exactamente el mismo resultado que *sort edad*.

La orden *gsort* incorpora un par de opciones que pueden resultar útiles. La opción *generate* (*nombre\_de\_variable*) crea una nueva variable con los valores 1, 2, 3... para cada caso según su orden en la variable por la que ha sido ordenada. Por ejemplo:

```
gsort edad, generate (orden_edad)
list sexo edad orden_edad, clean
```

---

ordenar de nuevo por la variable con la que estuvieran clasificados anteriormente, siempre y cuando esta existiera, como el número de caso, por ejemplo. Más adelante se expone cómo realizar esta operación. Otra posibilidad es la de crear una instantánea de los datos antes de ordenarlos. Véase sección 3.1.

Produce el siguiente resultado:

**ILUSTRACIÓN 5.4. Resultado de la instrucción *list* con variable generada con *gsort***

	sexo	edad	orden_~d
1.	Mujer	19	1
2.	Mujer	20	2
3.	Hombre	20	2
4.	Hombre	21	3
5.	Mujer	22	4
6.	Hombre	23	5
7.	Mujer	24	6
8.	Hombre	.	7

Otra opción que incorpora *gsort* es *mfirst*, que coloca los valores perdidos en primer lugar, en vez de en los últimos de la ordenación.

### 5.1.2. Selección de casos

Para este fin, Stata incorpora dos subinstrucciones, *in* e *if*. Reciben este nombre porque no pueden aparecer aisladas, sino siempre en conjunción con alguna otra orden. De este modo, las cláusulas *in* o *if* hacen que la instrucción a la que acompañen sólo se realice para aquellos casos que se especifiquen en ella. Tienen, pues, efectos temporales porque sólo seleccionan casos para la instrucción donde se introducen<sup>2</sup>.

La subinstrucción *in* se utiliza del siguiente modo:

instrucción **in** rango

Donde *instrucción* indica la orden que se desea que sólo se ejecute en el rango de casos determinado. El rango se declara utilizando los números de caso. Por ejemplo, si en la matriz de sexo y edad se quiere mostrar la edad media de los cinco primeros casos, puede escribirse la siguiente orden:

summarize edad in 1/5

De este modo, se obtendrán los siguientes estadísticos:

---

<sup>2</sup> En el caso de que se desee una eliminación permanente de casos, ha de utilizarse la instrucción *drop*, seguida de las subinstrucciones de selección pertinentes. Por ejemplo, *drop in 1* borra de la matriz de datos el primer caso. Y si se guarda el fichero bajo el mismo nombre con la instrucción *save, replace*, este caso quedará eliminado para siempre, a menos que se disponga de una copia de respaldo.

**ILUSTRACIÓN 5.5. Sumario de una variable con selección numerada de casos**

Variable	Obs	Mean	Std. Dev.	Min	Max
edad	5	20.4	1.140175	19	22

Como puede fácilmente apreciarse, el número de observaciones es 5, y sólo sobre ellas se han realizado los cálculos de la instrucción *summarize*.

Hay que tener en cuenta que la selección de casos con la subinstrucción *in* depende, obviamente, del orden que tengan los casos. Los cinco primeros casos de la matriz serán diferentes según la variable con la que los casos estén ordenados (con *sort* o *gsort*) y según el orden (ascendente o descendente) de los casos. En este caso, como los casos están ordenados por edad y de menor a mayor (orden *sort edad*), lo que muestra la ilustración 5.5 está referido a la edad de los cinco individuos más jóvenes.

Como se ha visto, el rango de casos de *in* se especifica con el número menor del rango, seguido de una barra (/), y del número mayor (por ejemplo, para seleccionar los cinco primeros casos, 1/5). Las letras *f* y *l* se pueden utilizar para hacer referencia a los casos primero (*f*) y último (*l*). Los valores con signos negativos serán interpretados por Stata como distancias desde el final de la matriz. Por ejemplo, si se escribe...

```
list sexo edad in -5/l, clean
```

... se obtiene un listado de los últimos cinco casos con sus correspondientes valores de sexo y edad.

**ILUSTRACIÓN 5.6. Listado de casos con selección numerada final**

	sexo	edad
4.	Hombre	21
5.	Mujer	22
6.	Hombre	23
7.	Mujer	24
8.	Hombre	.

Por otro lado, la subinstrucción *if* permite especificar los casos que se van a utilizar para ejecutar una orden en función de si cumplen una determinada condición lógica. La forma general de uso es:

```
instrucción if expresión
```

*Instrucción* es la orden que se solicita, y *expresión* es la condición que han de cumplir los casos para que les afecte el comando.

Las expresiones contenidas en la cláusula *if* pueden ser numéricas o lógicas. Las expresiones numéricas pueden ser tan simples como una constante o tan complejas como un entramado sucesivo de variables o constantes conectadas con funciones. En el extremo más sencillo, una expresión numérica es en realidad un número que si adopta el valor de 0 excluye del análisis e incluye en el resto de los supuestos. Ello implica que si se escribe *if 0*, la instrucción no se ejecuta con ningún caso, mientras que si se especifica *if -1* o *if 100*, la orden se cumple para todos los casos. En consecuencia, la expresión más elemental detrás del *if* consta de una constante o una variable numérica, y excluye del análisis a todos los casos con el valor 0 en la expresión, e incluye a todos los que tengan distinta circunstancia, también situaciones en las que la expresión numérica especificada tenga un valor perdido.

La expresión siguiente en términos de simplicidad es aquella representada mediante una expresión aritmética, que se compone de variables o valores conectados mediante operadores o funciones matemáticas que dan lugar a un solo valor numérico. De este modo, si se escribiera la expresión *list if variable-variable*, el resultado es que sólo se mostrarían aquellos casos con valor perdido en la variable especificada, ya que al restar su contenido de sí misma, el resultado es siempre 0, a menos que contenga un valor perdido.

En cualquier caso, lo más común es que la expresión que siga a *if* sea lógica, en lugar de numérica. Se considera expresión lógica aquella que puede de mostrar dos valores, falso o cierto, como resultado de emplear un operador de relación. Las más simples expresiones lógicas propiamente dichas, son aquellas que se componen de dos expresiones aritméticas unidas por un operador de relación.

Los operadores de relación posibles son los siguientes:

- `==` Igual que...<sup>3</sup>
- `>` Mayor que...
- `<` Menor que...
- `>=` Mayor o igual que...
- `<=` Menor o igual que...
- `!=` No es igual que

---

<sup>3</sup> Nótese que se ha puesto el signo igual dos veces seguidas. Como puede comprobarse más adelante, esto no es un error. Para Stata no es lo mismo el signo igual, empleado en operaciones matemáticas,  $2+2$  es igual a 4, por ejemplo, que los dos signos igual seguidos, utilizados como operador lógico. Si se quiere producir un resultado, hay que emplear un signo, si se desea hacer una comparación, se emplean los dos signos seguidos. Como regla, en caso de duda, se puede pensar si el signo igual puede ser reemplazado por el símbolo `>`, en cuyo caso, se deben incluir los dos signos iguales.

De este modo se podrían generar expresiones para seleccionar casos que cumplan determinadas condiciones en una variable o en una combinación aritmética de ellas. Por ejemplo, si se desea seleccionar sólo a los hombres, habría que escribir `sexo==1`. Si se quiere un análisis estadístico de los mayores de 20 años, la expresión debería ser `edad>20`. Y también puede indicarse `pibse>pibag*2` para listar, por ejemplo, los países con doble proporción de producción de servicios que de bienes agrarios.

Finalmente, las expresiones lógicas se pueden modificar o conectar entre ellas mediante operadores lógicos. Estos son tres: la conjunción (`&`) y la disyunción (`|`), que tienen por misión conectar otras expresiones, y un terce-  
ro (`!`), la negación, cuyo cometido es el de invertir la veracidad o falsedad de la expresión a la que antecede.

Por ejemplo, si se quiere hacer un análisis que incluya sólo a las mujeres con 20 o menos años, la instrucción en cuestión debe ir acompañada de la expresión `if sexo==2 & edad<=20`. Así, si se desea hacer un resumen estadístico de la edad de las mujeres jóvenes de la muestra, habrá que escribir una instrucción similar a la siguiente:

```
summarize edad if sexo == 2 & edad<=20
```

El resultado se referirá al subgrupo especificado, pero se mostrará como si de la muestra total se tratara. Sólo por el número de casos se podría deducir que se ha efectuado la operación con un filtro.

#### **ILUSTRACIÓN 5.7. Sumario de una variable con selección condicional de casos**

Variable	Obs	Mean	Std. Dev.	Min	Max
edad	2	19.5	.7071068	19	20

El orden de ejecución de los distintos operadores es el siguiente: en primer lugar, la negación lógica, expresada sea con `!`, sea con `~`; después, las funciones; a continuación, la negación aritmética (`-`); seguidamente, las operaciones aritméticas `/` (división) y `*` (multiplicación); después la `-` (resta) y `+` (suma); luego, los operadores relacionales (`!=, >, <, <=, >=` y `=`), y finalmente los operadores lógicos `&` y `|`, por este orden.

Un ejemplo de la importancia del orden de colocación de los operadores se encuentra cuando se han de seleccionar personas con más de 20 años, sean del norte (1) o del sur (2). Si, por ejemplo, la condición se expresa del siguiente modo...

```
tabulate sexo if region ==1 | region== 2 & edad >20
```

... se producirá un resultado erróneo, ya que elegirá tanto a los sureños con más de 20 años, por un lado, como, por el otro, a todos los norteños. Es decir, se ejecuta primero la conjunción (`&`) y después la disyunción (`l`). Para que pueda efectuar la operación adecuadamente se deben utilizar paréntesis que fuercen a realizar con anterioridad las operaciones en ellos incluidos.

```
tabulate sexo if (region ==1 | region == 2) & edad>20
```

De este modo, primero se estima si el caso es del norte o del sur y después se juzga si además es mayor de 20 años, para seleccionar sólo a personas de estas dos regiones mayores de esa edad.

#### **ILUSTRACIÓN 5.8. Tabulación de una variable con selección condicional de casos**

Sexo	Freq.	Percent	Cum.
Hombre	3	60.00	60.00
Mujer	2	40.00	100.00
Total	5	100.00	

El operador contrario a `==` es `!=`, que significa desigual. Es útil para descartar en un análisis o en un listado a los sujetos que no posean un determinado valor en una variable. Así, si sólo se quiere mostrar el sexo de quienes tenemos el nombre, habrá que escribir la instrucción con la cláusula `if nombre!=""`, donde el par de comillas seguidas indica un valor vacío en una variable de texto. En cambio, si se hubiera empleado como filtro una variable numérica, habría que haber usado el signo punto (.), para indicar un caso sin valor o perdido.

```
list sexo nombre if nombre!=""
tabulate sexo if edad!=.
```

#### *5.1.3. Agrupación de casos*

Existe una preinstrucción especial (`by`) que permite agrupar los casos según sus valores en una o más variables y hacer que la instrucción a la que acompañan se ejecute por separado en cada uno de los grupos. Para ello, `by` se especifica al principio, seguido por el nombre de la variable por la cual se desea segmentar el análisis y dos puntos (:). Tras esa expresión, se escribe la instrucción pertinente. Es necesario

ordenar previamente los casos según los valores de la variable que conforma los grupos, para que funcione adecuadamente el propósito de la preinstrucción.

En el caso, por ejemplo, de que se quiera obtener un resumen de la variable *edad* para cada uno de los sexos, por separado, hay que efectuar en primer lugar una ordenación de los casos por sexo:

```
sort sexo
```

Los datos quedan así ordenados por los valores de sexo: primero, los hombres; después, las mujeres (al estar codificados, respectivamente, con los valores 1 y 2). A continuación, habrá que escribir la orden precedida por la preinstrucción *by*:

```
by sexo: summarize edad
```

De esta forma, el programa se encarga de repetir automáticamente la instrucción para hombres y mujeres:

### **ILUSTRACIÓN 5.9. Sumario de una variable por grupos**

-> sexo = Hombre					
Variable	Obs	Mean	Std. Dev.	Min	Max
edad	3	21.33333	1.527525	20	23
-----					
-> sexo = Mujer					
Variable	Obs	Mean	Std. Dev.	Min	Max
edad	4	21.25	2.217356	19	24

Como puede apreciarse en la ilustración 5.9, lo que hace Stata es ejecutar el comando *summarize* dos veces, una para hombres y otra para mujeres, de manera totalmente independiente en cada caso. El prefijo *by* variable se puede usar con casi cualquier instrucción, y siempre realiza independientemente la orden solicitada sobre cada uno de los grupos definidos por la variable especificada. Por otro lado, también conviene saber que puede especificarse más de una variable para obtener los correspondientes análisis cruzados.

```
sort sexo region
by sexo region: summarize edad
```

El resultado obtenido presenta tantos análisis como el producto del número de valores de cada una de las variables implicadas.

#### **ILUSTRACIÓN 5.10. Sumario de una variable por grupos conformados por dos variables**

-> sexo = Hombre, region = Norte					
Variable	Obs	Mean	Std. Dev.	Min	Max
edad	2	21.5	2.12132	20	23

-> sexo = Hombre, region = Sur					
Variable	Obs	Mean	Std. Dev.	Min	Max
edad	1	21	.	21	21

-> sexo = Mujer, region = Norte					
Variable	Obs	Mean	Std. Dev.	Min	Max
edad	2	21.5	3.535534	19	24

-> sexo = Mujer, region = Sur					
Variable	Obs	Mean	Std. Dev.	Min	Max
edad	2	21	1.414214	20	22

Al tener sexo dos valores (*hombre* y *mujer*) y región también (*norte* y *sur*), con *by* sexo región se crean cuatro grupos, para cada uno de los cuales se ejecuta la instrucción *summarize edad*.

Para no tener que escribir la instrucción *sort* antes, puede utilizarse *bysort*, en lugar de *by*, con el mismo efecto y función. Así, las dos líneas anteriores podrían haberse escrito en una sola de este modo:

```
bysort sexo region: summarize edad
```

#### *5.1.4. Ficheros anchos y alargados*

Hasta el momento sólo se ha tratado con ficheros anchos. Son llamados así aquellos en los que todas las variables pertenecientes a un sujeto se encuentran en el mismo registro o línea. El ejemplo quizás más frecuente y fácil de comprender para su transformación a un fichero largo es el de un conjunto de países de los que se dispone de la información de una serie de años. En el formato ancho cada uno de los años se expresa en diferentes columnas.

Sean tres países con información en una variable a lo largo de tres años. En el sencillo ejemplo que se emplea se utilizan como países España, Francia e Italia, los años 2005, 2006 y 2007, y una variable como el número de estudiantes.

```
use paisesanchos, clear
list, clean abbreviate(15)
```

Mediante este listado<sup>4</sup> puede observarse que se trata de un conjunto de datos con tres casos y cuatro variables (los tres años más el nombre del país).

### ILUSTRACIÓN 5.11. Listado de un fichero ancho

	país	estudiantes2005	estudiantes2006	estudiantes2007
1.	España	7537	7529.2	7555.7
2.	Francia	12315.4	12320.5	12296
3.	Italia	9408.9	9464.4	9500.2

Convertirlo al formato alargado supondría tener nueve casos (los tres países por los tres años) reduciendo las tres variables anuales a una sola. Stata llama *i()* a las variables cuyo único valor denota una observación o caso, los países en este ejemplo; *j()*, en cambio, es la variable que denota las subobservaciones repetidas, los años en esta ocasión.

La sintaxis para convertir un fichero ancho en otro largo sería:

```
reshape long nueavar, i(observación) j(subobservación)
```

Hay que tener en cuenta que la variable denominada como *observación* debe existir literalmente en el fichero y la subobservación ha de ser la variable que en el nuevo fichero contenga los distintos años (del 2005 al 2007, en este caso, por tanto conviene denominarla así: *año*). La variable expresada como *nueavar* (estudiantes) se creará automáticamente en la nueva matriz con el nombre que tenía en la original, aunque sin las cifras del año. La nueva estructura de datos contendrá el número de casos de partida multiplicado por el número de variables temporales (tres en este ejemplo, una por cada año).

Para aplicarla a los datos que se acaban de mencionar, las instrucciones que realizan la transformación y la muestran son:

```
reshape long estudiantes, i(pais) j(año)
list, clean
```

---

<sup>4</sup> Se ha añadido la opción *abbreviate(15)* para que Stata no abrevie los nombres de las variables y queden expresados de modo completo. La orden *list* los recorta por defecto a ocho caracteres.

Como puede apreciarse a continuación, se han creado dos nuevas variables: *estudiantes* y *año*, siendo esta última la que triplica el tamaño del fichero, por cuanto que por cada país hay información de tres años.

### ILUSTRACIÓN 5.12. Conversión de un fichero ancho a otro alargado

```
(note: j = 2005 2006 2007)
Data                                         wide    ->   long
-----
Number of obs.                               3      ->     9
Number of variables                         4      ->     3
j variable (3 values)                      ->     año
xij variables:
estudiantes2005 estudiantes2006 estudiantes2007->estudiantes
-----
          país     año    estudiantes
1. España  2005      7537
2. España  2006     7529.2
3. España  2007     7555.7
4. Francia 2005    12315.4
5. Francia 2006    12320.5
6. Francia 2007    12296
7. Italia   2005    9408.9
8. Italia   2006    9464.4
9. Italia   2007    9500.2
```

Del mismo modo que se ha cambiado un fichero ancho a otro largo, se puede realizar la operación inversa: la de convertir una matriz larga en otra ancha. Como las variables cumplen la misma función, observación, subobservación y contenido, la instrucción sólo cambia el adjetivo de *long* por el de *wide*.

```
reshape wide estudiantes, i(pais) j(año)
```

Convertiría el fichero largo a otro ancho. En este caso concreto, el resultado de la instrucción sería la vuelta al fichero inicial.

## 5.2. Generación y modificación de variables

En muchas ocasiones han de transformarse los datos porque el análisis requiere que se trabaje con ellos de forma distinta a como se encuentran registrados en el ordenador. Las razones pueden ser muy diversas. Baste aquí con citar sólo algunas de las más frecuentes.

En primer lugar, es un caso común que los datos estén mal grabados y que se encuentre un código que no existe. Por ejemplo, si al solicitar una tabla de distribución de frecuencias de la variable *sexo*, aparecen 580 casos como hombres, 619 como mujeres y un caso con un valor no etiquetado igual a 5, es obvio que se trata de un error de grabación y debe ser depurado. En otras ocasiones, es preciso cambiar la escala de una variable, como

es el caso de que se tenga el PNB de una serie de países expresados en dólares y se prefiera que aparezcan en euros. También puede suceder que se desee trabajar con una escala logarítmica, en lugar de la aritmética original, por lo que debe transformarse esta última. Otro caso de transformación es cuando se desea trabajar con variables estandarizadas, en lugar de las variables originales, o cuando se quieren presentar los datos de una variable cuantitativa recodificados en intervalos. O para proseguir con una lista interminable de razones para la transformación de variables, puede también citarse el caso en que se desea construir una variable con una combinación de varias, como cuando se genera la clase social en función de la relación con la actividad, la profesión y los estudios, o cuando debe obtenerse la puntuación de una escala mediante la suma de la serie de ítems de los que se compone.

Desde un punto de vista instrumental pueden clasificarse las transformaciones de las variables en algebraicas, de equivalencias y lógicas. En las primeras se obtienen los nuevos valores de las variables mediante la aplicación de una o varias funciones matemáticas o estadísticas; en las segundas, las reglas del cambio se producen mediante una serie de igualdades entre los valores antiguos y los valores nuevos, y en las lógicas, los cambios de valores (sean transformados por una operación algebraica o por una recodificación) se producen si y sólo si se cumplen determinadas condiciones. Un ejemplo fácil de las primeras sería cuando se posee información del año de nacimiento de una serie de personas y se desea transformar en edades. En tales circunstancias basta con restarle al año en que se tomaron los datos el de la fecha de nacimiento. Si un estudio se hizo en el año 2000, es obvio que las personas que nacieron en 1970 cumplieron 30 años a lo largo de dicho año. Si faltara la variable *edad*, podría disponerse de una aproximación mediante esta operación. Otro ejemplo de cambio, esta vez de transformación lógica, sería en el caso de que la variable PNB estuviera grabada en la moneda de cada país y se deseara pasar a una única moneda. En esas circunstancias, habría que multiplicar prácticamente cada país por un número distinto, y por ello, antes de la operación aritmética requerida, hay que exponer la condición que ha de aplicarse. Por ejemplo, en el caso de que el país tenga como moneda el euro, hay que multiplicar su valor del PNB por 0,9 para expresarlo en términos de la moneda americana.

### 5.2.1. Transformaciones algebraicas

Los comandos más importantes de Stata para generar y modificar variables mediante operaciones algebraicas son *generate* y *replace*. El funcionamiento de ambos es básicamente el mismo, sólo que el primero crea una nueva variable y le asigna valores, y el segundo reemplaza los valores de una variable existente.

Para crear una nueva variable ha de utilizarse *generate* del siguiente modo:

```
generate nueavar = expresión
```

Como es fácil inferir, aquí se usa sólo un signo de igual (=), no dos (==) como en las comparaciones lógicas, porque en esta ocasión se trata de asignar, en lugar de comparar, un valor a una variable. En Stata, un signo de igual se utiliza para asignar valores a una variable, y dos signos de igual significan “es igual que”.

A la hora de escribir la instrucción, en lugar de expresión, hay que introducir cualquier fórmula matemática, desde una constante (un número que será igual para todos los casos) hasta una función o varias funciones, pasando por las operaciones aritméticas básicas, como son la suma (+), la resta (-), la multiplicación (\*) y la división (/). Véanse algunos ejemplos:

```
use panel5, clear  
generate total = 1
```

Crea una variable o, mejor dicho, una constante asignada a todos los casos con el valor 1. En cambio, la siguiente instrucción...

```
generate edad=2000-añonacimiento
```

... genera una variable llamada *edad*, que expresa la diferencia entre el año en el que se recoge la información y el año de nacimiento de cada individuo, con lo que se obtienen los años que se cumplen en alguno de los doce meses de 2000.

La instrucción *replace* funciona exactamente igual que *generate*, pero debe emplearse con variables ya existentes. Como ya está definida la variable *edad*, para transformarla a fin de expresar el año adecuado incluso para los que no celebraron aún su cumpleaños en el mes de la encuesta, habría que emplear la orden *replace*, en lugar de *generate*:

```
replace edad = 2000-añonacimiento-1 if mesnacimiento>2  
list mesnacimiento añonacimiento edad in 1/5, clean
```

En este último ejemplo conviene advertir que tanto *generate* como, más frecuentemente, *replace* pueden ir acompañadas por la cláusula *if*, realizán-

dose de este modo la operación sólo en aquellos casos que cumplan la condición expresada<sup>5</sup>. Con la aquí expresada, el año de nacimiento disminuye en una unidad para aquellos que hayan nacido después de febrero (*mesnacimiento>2*), porque habiéndose hecho la encuesta a finales de dicho mes, aún no han cumplido años durante el año de la encuesta. En otras palabras, alguien que el primero de marzo dijera que nació en 1960, tendría 40 años, si ya hubiera celebrado su cumpleaños, es decir, si hubiera nacido en enero o febrero, pero si no hubiera celebrado aún su cumpleaños, tendría aún 39, pues cumpliría los 40 en el tiempo que restara del año. De los cinco casos listados en la ilustración 5.13, los dos últimos se encuentran en el primer supuesto, mientras que los tres primeros aún no cumplieron sus años en 2000, de ahí que la suma del año y la edad no sumen 2000, sino 1999.

**ILUSTRACIÓN 5.13. Listado de las variables *mes*, *añonacimiento* y *edad* en los cinco primeros casos**

	mesnac~o	añonac~o	edad
1.	12	1965	34
2.	5	1962	37
3.	5	1980	19
4.	2	1940	60
5.	2	1973	27

Otros ejemplos útiles son las transformaciones de potencia de las variables cuantitativas con el fin de dar cuenta de relaciones no lineales entre los datos. Las más frecuentes en este sentido son el cuadrado y el logaritmo. En Stata, transformar de estos modos una variable es tan fácil como escribir las siguientes instrucciones.

```
generate edadc=edad^2
generate lnedad=ln(edad)
list edad edadc lnedad in -3/l, clean
```

En este caso se puede ver el resultado de estas dos transformaciones en los tres últimos casos del fichero.

**ILUSTRACIÓN 5.14. Listado parcial de las transformaciones cuadrática y logarítmica de la edad**

	edad	edadc	lnedad
5281.	46	2116	3.828641
5282.	28	784	3.332205
5283.	58	3364	4.060443

<sup>5</sup> Más detalles de este uso del *if* en instrucciones *generate* y *replace* se presentan en la sección 5.2.3.

Las expresiones matemáticas que se pueden utilizar en *generate* y *replace* pueden complicarse tanto como se quiera. Siempre que se requiera hay que utilizar paréntesis para que las operaciones se realicen en el orden deseado. Como es regla matemática e informática habitual, en toda expresión, primero se calculan los paréntesis, luego las potencias (^), después las funciones —*ln()*, por ejemplo—, luego multiplicaciones o divisiones y, finalmente, las sumas o restas. Y, en caso de que haya operaciones en el mismo nivel, se realizan primero las que se encuentren a la izquierda.

Es obvio que tanto *generate* como *replace* permiten el uso de funciones matemáticas especiales incorporadas por Stata. Para una lista completa de cada una de ellas, se recomienda hacer uso de la ayuda de *functions*.

help functions

En ella aparece una lista de subfunciones de las que se puede volver a pedir ayuda adicional. La lista de estos grupos de operaciones posible es la siguiente: matemáticas (*math functions*), probabilísticas (*density*), aleatorias (*random*), textuales (*string*), programadoras (*programming*), de fecha (*date and time*), de series temporales (*time-series*) y matriciales (*matrix*).

Por ejemplo, una función muy útil es *runiform()*<sup>6</sup>, que sirve para crear variables con valores pseudoaleatorios continuos que varían entre 0 y 1. Ahora bien, los valores creados por esta función pueden ser modificados por medio de una función matemática para cambiar sus intervalos a fin de que se adapten mejor a las necesidades del analista.

Por ejemplo, si se desea generar para cada caso una variable con valores aleatorios enteros entre 1 y 100, pueden utilizarse las dos siguientes instrucciones:

```
set seed 43214
generate naleat = int(runiform()*100)+1
```

La orden *set seed* no se usa para cambiar la semilla de aleatorización que genera la serie de números aleatorios. Sólo es aconsejable introducirla cuando se desee generar el mismo conjunto de números aleatorios en múltiples repeticiones de un programa, ya que, caso de que no se explice, la función de generación aleatoria produce conjuntos diferentes de números aleatorios.

---

<sup>6</sup> Aunque esta función no utilice parámetros, siempre ha de ir seguida de los paréntesis.

Existen muchas otras posibilidades de generación de variables aleatorias. Como ejemplos más usuales pueden citarse la generación de números binomiales discretos mediante *rbinomial* (*n,p*), o la producción de una variable aleatoria continua con distribución normal *rnormal* (media, desviación), útiles para la simulación o la adición a datos empíricos de errores aleatorios con una determinada distribución.

```
generate aleatorio_binomial = int(rbinomial(4,.5)
generate aleatorio_normal=rnormal(5,2)
list naleat aleatorio_binomial aleatorio_normal in 1/3, clean abbreviate(20) noobs
```

Con la primera expresión se crearía una nueva variable con valores discretos comprendidos entre el 0 y el 4, con media 2 ( $np$ ) y desviación típica 1 ( $np(1-p)$ ), mientras que con la segunda se generaría una variable continua normal con media 5 y desviación típica 2<sup>7</sup>.

#### **ILUSTRACIÓN 5.15. Listado parcial de los números aleatorios obtenidos**

naleat	aleatorio_binomial	aleatorio_normal
56	2	1.69696
20	1	3.540512
5	0	5.576273

Una expresión utilizable en los comandos de generación o modificación de variables es *\_n*, que sirve para hacer referencia al número de orden del caso actual en la matriz de datos. Puede usarse, por ejemplo, en la instrucción *generate* para crear una variable que exprese la posición del individuo en el fichero, con el posible fin de devolver el orden inicial, después de una ordenación por otro criterio.

```
list sexo edad in 1/3
generate orden = _n
sort edad
list sexo edad in 1/3
sort orden
list sexo edad in 1/3
```

---

<sup>7</sup> Si el lector desarrolla este ejemplo en su ordenador, advertirá que sólo le coincide la primera columna y no las dos últimas. Esto ocurre porque la semilla de aleatorización sólo afecta a la orden inmediatamente posterior. Si se hubieran deseado números aleatorios estables binomiales y normales, habría que haberlos precedido, respectivamente, de la instrucción *set seed* con o sin idéntica constante. Pruebe a hacerlo de nuevo, repitiendo la instrucción con la misma semilla para obtener los mismos resultados que la ilustración 5.15.

### 5.2.2. Transformaciones de equivalencia

Las modificaciones de equivalencia son aquellas en las que a diferentes conjuntos de valores antiguos de una variable se les hace corresponder distintos valores nuevos. En el fondo, se trata de lo mismo que realiza la instrucción *replace*, pero en lugar de aplicar los cambios con una función, lo hace con una serie de equivalencias entre los valores antiguos (vvaas) y los valores nuevos (nuevovalor). La orden necesaria para hacer esto en Stata es *recode*, cuya sintaxis general es la siguiente:

```
recode listaveriables (vvaas=nuevovalor) (vvaas=nuevovalor), [into(nueavariable)]
```

Esta instrucción puede emplearse, por ello, de dos maneras: una, para alterar una variable existente, y otra, para crear una nueva variable, con la opción *into()*, a partir de los valores de la antigua.

Un ejemplo muy habitual de utilización de esta instrucción es cuando se ha detectado una mala grabación de datos. Hay veces que se encuentran un par de casos con códigos que no corresponden a ninguna de las opciones de respuesta posibles, como cuando en la variable *sexo* —codificada con los valores 1 y 2— se encuentren valores como el 3 y el 5. Esta es buena ocasión para transformar los errores en casos perdidos de modo similar al siguiente:

```
use panel5b, clear  
recode sexo (3/.=.)
```

De este modo<sup>8</sup>, la variable *sexo* sólo queda inalterada en el sentido de que los casos codificados con un valor de 3 o superior pasan a ser considerados como perdidos por el sistema. El resto de los valores permanecen idénticos.

Otro ejemplo frecuente es el cambio de una escala de Likert. Cuando se deseen invertir los ítems de modo tal que en una escala del 1 al 4 este último pase a ser el más bajo y el primero el más alto, entonces es pertinente el empleo de esta orden, que puede usarse al mismo tiempo en un conjunto de variables<sup>9</sup>, siempre y cuando las transformaciones se separen entre paréntesis.

---

<sup>8</sup> Recuérdese que la barra (/) significa “hasta” y que el valor perdido se representa como un punto y es considerado por Stata como el valor más alto posible. Por ello 3/ significa desde el 3 hasta el valor mayor, casos perdidos incluidos. Entre otras palabras clave posibles se encuentran *min* y *max* con significados respectivos de valor mínimo y máximo.

<sup>9</sup> Ha de notarse que al tratarse de una transformación lineal, esta operación también podría hacerse con la instrucción *replace*. Concretamente, mediante la expresión “item1=5-item1”, pero *recode* ofrece la ventaja de poderlo aplicar en la misma línea a un conjunto de variables.

```
recode p201 p203 p204 (1=4)(2=3)(3=2)(4=1)
```

Además de expresar una lista seguida de los valores deseados (1 2 3 4 5=0), también pueden aparecer expresiones en vvaas (valores antiguos) como (1/5=0), que transforma todos los valores entre 1 y 5 en 0. Asimismo, pueden cambiarse los valores perdidos utilizando la palabra clave *missing*; los casos válidos no utilizados en otras transformaciones, si se emplea *nonmissing*; los valores mínimo y máximo, mediante *min* y *max*, respectivamente, y finalmente también puede utilizarse *else* para referirse a todo lo que no ha sido cambiado mediante otras equivalencias.

Por ejemplo, si quiere recodificarse la edad en cuatro grupos, puede escribirse una instrucción similar a la siguiente:

```
recode edad (min/35=1) (36/50=2) (51/65=3) (66/max=4) (else=.), into(edadr)
```

Además, si se desea, pueden ponerse etiquetas de valores a la nueva variable en la misma orden de recodificación. Para realizarlo, la instrucción anterior debería convertirse en esta otra:

```
recode edad (min/35=1 "Hasta 35") (36/50=2 "36-50") (51/65=3 "51-65") ///
(66/max=4 "Mas de 65") (else=.), into(edadr2)
```

De este modo, tras pedir una distribución de frecuencias mediante la instrucción siguiente...

```
tabulate edadr2
```

... se obtendría una tabla con los valores recodificados de la edad del siguiente tenor:

#### **ILUSTRACIÓN 5.16. Tabulación de una variable recodificada**

RECODE of		Freq.	Percent	Cum.
edadr2				
Hasta 35	1,902	36.00	36.00	
36-50	1,264	23.93	59.93	
51-65	1,096	20.75	80.67	
Mas de 65	1,021	19.33	100.00	
Total	5,283	100.00		

### 5.2.3. Transformaciones lógicas

En otras ocasiones, para obtener una determinada transformación de los valores de una o varias variables, son necesarias una o varias operaciones lógicas aplicadas a las instrucciones que se acaban de analizar. Por ello, van a ser analizados aquí todos aquellos cambios en las variables que sólo tienen lugar cuando se cumplen una o una serie determinada de condiciones. Cada una de ellas requiere una instrucción condicionada con una cláusula de selección, que será en la mayor parte de los casos un condicional<sup>10</sup>.

El ejemplo más sencillo de transformación lógica se produce cuando se asigna una constante a una nueva o antigua variable para todos aquellos casos que satisfagan una condición. Imagínese, por ejemplo, que se necesite dar el valor 1 en la variable *joven* a todos aquellos individuos que tengan en la variable *edad* un valor inferior o igual a 30. En dicho caso, bastaría escribir la siguiente instrucción.

```
generate joven=1 if edad<=30
```

Es conveniente advertir de que con la instrucción anterior se crea una nueva variable llamada *joven*, que tiene el valor 1 en todos aquellos casos en los que el valor de la edad sea igual o menor que 30 años; en tanto que tendrá el valor perdido(.) en el resto de los casos, ya que no han sido definidos anteriormente. Prueba de ello es el resultado de aplicarle la instrucción *summarize* a la variable *joven*.

**ILUSTRACIÓN 5.17. Sumario de la variable *joven* tras su creación**

Variable	Obs	Mean	Std. Dev.	Min	Max
joven	1356	1	0	1	1

En la muestra a la que se ha aplicado la instrucción hay 1.356 jóvenes. Por ello, la variable *joven* tiene 1.356 casos con media y único valor igual a 1. Si, a continuación, se precisa otorgar el valor 0 a los que tienen más de 30 años, puede seguirse utilizando la cláusula *if*, pero como ya está creada la variable, ahora hay que aplicar la instrucción *replace*, en lugar de *generate*. Más concretamente, la instrucción debería ser escrita del siguiente modo:

---

<sup>10</sup> En lugar de una cláusula condicional, se puede también hacer uso de un condicionante en función del número de casos (in 1/10). Como quiera que este también puede expresarse como un condicional (if \_n<10, por ejemplo), sólo se hará mención a las del primer tipo.

```
replace joven=0 if edad>30 & edad<
```

La segunda condición añadida a la primera es necesaria para que no atribuya el valor 0 a la variable *joven* en aquellos casos (los perdidos) a los que falta el dato de la variable *edad*.

Con ello, tras las dos instrucciones condicionadas anteriores, un resumen de la nueva variable *joven* aportaría el siguiente resultado:

#### **ILUSTRACIÓN 5.18. Sumario de una variable tras su transformación**

Variable	Obs	Mean	Std. Dev.	Min	Max
joven	5283	.2566723	.4368384	0	1

Al tratarse de una variable ficticia y dicotómica, la media indica la proporción de jóvenes que hay en la muestra y la desviación típica es la raíz cuadrada de  $p(1-p)$ .

Con un conocimiento de la lógica de funcionamiento de las expresiones de Stata, las dos instrucciones anteriores pueden reducirse a una sola, si se genera la nueva variable con una expresión lógica seguida de una cláusula condicional. Quiere ello decir que no se dan los valores 0 ó 1 a la nueva variable, sino una expresión lógica con dos estados posibles (verdadero=1 y falso=0), *edad<=30* en este caso, escrita entre paréntesis para mayor claridad, aunque podrían eliminarse. Es muy conveniente también añadir a la instrucción la cláusula *if* para evitar que ponga el valor 0 a los casos perdidos<sup>11</sup>.

```
generate joven2=(edad<=30) if edad<
```

Cuando en las transformaciones se necesita más de una variable, es inevitable el uso de la condición *if*, como, por ejemplo, en el caso de que se deseé una variable compuesta de *sexo* y *edad* que contenga los siguientes valores: hombres jóvenes, mujeres jóvenes, hombres mayores y mujeres mayores. Para generar una variable con estos cuatro valores y tabularla con las correspondientes etiquetas, serían necesarias estas siete líneas:

---

<sup>11</sup> Otra manera de realizar la misma operación es mediante la instrucción *recode* con tres cambios: uno para los valores 0, otro para los valores 1 y, finalmente, un tercero para valores perdidos:

```
recode edad (18/30=1) (31/98=0) (else=.), into(joven3)
```

```

generate sexedad=1 if sexo==1 & edad<=30
replace sexedad=2 if sexo==2 & edad<=30
replace sexedad=3 if sexo==1 & edad>30 & edad<.
replace sexedad=4 if sexo==2 & edad >30 & edad<.
label define sexedad 1 "Joven hombre" 2 "Joven mujer" ///
3 "Hombre mayor" 4 "Mujer mayor"
label value sexedad sexedad
tabulate sexedad

```

A partir de las cuales se mostraría la tabla siguiente:

#### **ILUSTRACIÓN 5.19. Tabulación de una variable compuesta**

sexedad	Freq.	Percent	Cum.
Joven hombre	690	13.06	13.06
Joven mujer	666	12.61	25.67
Hombre mayor	1,829	34.62	60.29
Mujer mayor	2,098	39.71	100.00
Total	5,283	100.00	

Sin embargo, con algo de práctica e imaginación las primeras cuatro líneas podrían haberse reducido a dos de este modo:

```

generate sexedad2=sexo+2 if edad<.
replace sexedad2=sexo if edad<=30

```

Es obvio que todo lo señalado en el apartado 5.1.2 sobre las expresiones lógicas es aplicable en este contexto, tanto para las instrucciones que generan o reemplazan variables como en aquellas que transforman valores.

Para mostrar la versatilidad del lenguaje, a continuación se realiza la misma operación, de la que ya se han dado dos procedimientos, pero esta vez utilizando la instrucción que recodifica valores.

```

recode sexo 1=3 2=4, into(sexedad3)
recode sexedad3 3=1 4=2 if edad<=30

```

Incluso una sola línea de código podría generar en las mismas condiciones la nueva variable, empleando la función *cond*, que consta de condición y dos expresiones, la primera para cuando la premisa es verdadera y la segunda para cuando es falsa. Es conveniente añadir un *if* para no dar valores válidos a casos de los que no se dispone de información en la variable de la

condición, cuando se emplean los operadores mayor o menor que, aunque en este caso concreto no es necesario porque en todos los casos se posee información de la edad.

```
generate sexedad4=cond(edad<=30,sexo,sexo+2) if edad<.
```

#### 5.2.4. Transformaciones extendidas

Aunque en las páginas anteriores se hayan descrito y explicado las más elementales instrucciones transformadoras, el repertorio de Stata no se agota con estas. En este apartado se aportan unos pocos ejemplos usuales, seleccionados entre otras muchas otras transformaciones factibles mediante la instrucción *egen*, variedad de *generate*, previamente explicada en 5.2.1, que permite el empleo de funciones más complejas de lo que permitía aquella; en contrapartida, sólo puede utilizarse una y sólo una de estas funciones sobre una variable o sobre una lista de variables existentes. Su sintaxis general es la siguiente:

```
egen nuevavariable= func(listavar) [if expresión] [in rango] [, opciones]
```

Existe una amplia variedad de operaciones posibles, funciones (*func*), con esta instrucción<sup>12</sup>. Las más útiles y simples para el análisis son *anycount*, *std* y la familia de operaciones *row*.

La función *anycount* sirve para crear una nueva variable consistente en el recuento en cada sujeto del número de variables en las que ha contestado un determinado subconjunto de valores especificados. Por ejemplo, si se quisiera obtener el número de ítems con los que un sujeto ha manifestado estar de acuerdo (2) o muy de acuerdo (1), debería escribirse una instrucción del siguiente tenor.

```
egen acuerdos=anycount(p201-p204), values (1,2)
```

El resultado arrojado es una variable con valores comprendidos entre 0 y 4 para cada sujeto, en función del número de unos o doses que tenga en las cuatro variables implicadas. Los sujetos con 0 serían aquellos que no se han manifestado favorablemente a ninguna de las cuatro preguntas, mientras que aquellos que hayan mostrado aquiescencia total tendrán el valor de 4.

---

<sup>12</sup> Véanse para ello las ayudas del programa (*help egen*) o las páginas 167-172 de la guía de gestión de datos (Stata, 2011d).

Otra función presente en *egen* es *std*, que permite estandarizar una variable cuantitativa o, lo que es lo mismo, realizarle una transformación lineal para que tenga la media y la desviación típica deseada, siendo, respectivamente, 0 y 1 los valores más comunes y los establecidos por defecto. Su forma es:

```
egen nueavar=std(expresion) [, mean(#) std(#)]
```

De este modo, si se desea estandarizar la variable *edad*, la instrucción que debería escribirse sería algo similar a:

```
egen EdadStandard=std(edad)
summarize EdadStandard
```

La segunda orden permitiría comprobar que la media es cercana a 0 y la desviación típica a 1 y que hay mayores desviaciones por encima de la media (máximo=2,6) que por debajo de ella (mínimo=-1,5):

#### ILUSTRACIÓN 5.20. Sumario de la estandarización de una variable

Variable	Obs	Mean	Std. Dev.	Min	Max
EdadStandard	5283	-3.06e-09	1	-1.521336	2.667577

Finalmente, son también útiles las funciones *row*. Estas calculan estadísticos horizontalmente (entre variables), más que verticalmente, como lo hacen las instrucciones típicas, por ejemplo, la que se acaba de emplear, *summarize*. Junto con *row* pueden emplearse las siguientes operaciones: *first*, *last*, *min*, *max*, *total*, *mean*, *sd*, *miss* y *nonmiss*.

Para ver su utilidad se muestra un ejemplo a partir de las cuatro variables escala sobre la participación política (p201-p204), previamente convertidas para que en todas ellas los valores se encuentren en la misma dirección (por ejemplo, a mayor apatía política, mayor puntuación) y para que los valores 8 (No sabe) y 9 (No contesta) figuren como valores perdidos<sup>13</sup>.

---

<sup>13</sup> Desde la versión 8, Stata admite hasta 27 valores perdidos. Además del punto (.), pueden considerarse valores inválidos aquellos compuestos por un punto seguido de uno de los veintiséis caracteres sencillos en minúscula del alfabeto. De este modo, al «No sabe» y al «No contesta» se les pueden dar dos códigos distinguibles y perdidos al mismo tiempo. Para ver los valores perdidos en una tabulación, se debe añadir la opción *missing* a la instrucción *tabulate*.

```

recode p201 p203 p204 (1=4)(2=3)(3=2)(4=1)
recode p201-p204 (8=.a)(9=.b)
egen total=rowtotal(p201-p204), missing
egen promedio=rowmean(p201-p204)
egen perdidos=rowmiss(p201-p204)
summarize total promedio perdidos

```

En el resultado la variable total indica la suma de los cuatro ítems. Al haber puesto la opción *missing*, los sujetos que no hayan respondido a ninguno de los cuatro enunciados se catalogan como casos perdidos. Sin embargo, hay sujetos que sólo tienen un punto, porque han respondido con ese código sólo a una de las preguntas. Suele ser más útil, por ello, solicitar la media (*rowmean*), porque el total es dividido por el número de respuestas válidas. En consecuencia, en la estadística pertinente, los valores mínimos de la variable promedio están comprendidos entre 1 y 4, en tanto que la media del promedio (2,6) no coincide con la de total (9,7) dividida por 4. Finalmente, la variable *perdidos* (número de no contestaciones por individuo) tiene como límites el 0 y el 4. El primero es asignado a toda persona entrevistada que contestó a las cuatro afirmaciones, y el valor máximo corresponde a quienes no respondieron a ninguna.

#### **ILUSTRACIÓN 5.21. Sumario de variables obtenidas con modalidades row de egen**

Variable	Obs	Mean	Std. Dev.	Min	Max
total	5167	9.680085	2.270354	1	16
promedio	5167	2.575882	.5283285	1	4
perdidos	5283	.3043725	.7732964	0	4

### **5.3. Características e instrucciones especiales**

Se acabará este capítulo explicando escuetamente algunas de las características especiales de Stata, que serán útiles para análisis estadísticos que serán abordados más adelante.

La primera de ellas es la capacidad que tienen algunas instrucciones de guardar parte de sus resultados. Estos se almacenan en listas de valores denominados en tres tipos de resultados: *sencillos*, a los que se les reconoce con *r(nombre)*; *estimadores*, denominados como *e(nombre)*, y los apenas empleados *especiales*, que generan nombres antecedidos por la letra *s*.

Entre las instrucciones que se han visto hasta el momento, las únicas que guardan resultado son *tabulate*, *summarize* y *ci*. Las tres lo hacen en el formato sencillo *r*. Como ejemplo de su uso, se verá su utilidad sólo con el segundo. Entre las variables generadas en las órdenes de modificación

de datos, en este capítulo se ha construido la variable *acuerdos*, que era el número de ítems con que los entrevistados se mostraban de acuerdo con la pregunta dos del cuestionario (véase página 112). Sus principales estadísticos pueden obtenerse mediante la orden *summarize*, de la que puede verse los resultados que guarda, mediante la instrucción *return list*<sup>14</sup>.

```
summarize acuerdos
return list
```

Estas dos instrucciones muestran prácticamente los mismos resultados, aunque con distinto formato.

#### ILUSTRACIÓN 5.22. Lista de resultados grabados en la instrucción *summarize*

Variable	Obs	Mean	Std. Dev.	Min	Max
acuerdos	5283	2.459209	1.094543	0	4
<b>scalars:</b>					
r(N)	= 5283				
r(sum_w)	= 5283				
r(mean)	= 2.45920878288851				
r(Var)	= 1.198023379915273				
r(sd)	= 1.094542543675335				
r(min)	= 0				
r(max)	= 4				
r(sum)	= 12992				

Puede comprobarse que las cantidades expuestas bajo el epígrafe *scalars* corresponden a la línea anterior que comienza con el nombre de *acuerdos*. Todas ellas tienen una denominación entre paréntesis precedida por la letra *r*. Además de los cinco que aparecen con la instrucción *summarize*, aparecen *r(sum\_w)*, igual a *r(N)*, porque no hay ponderaciones especiales; *r(Var)*, que es el cuadrado de la desviación típica, *r(sd)*, y *r(sum)*, o sumatorio de todos los valores, que podría obtenerse también multiplicando *r(N)* por *r(mean)*.

La segunda orden que se va a ver en este apartado está muy relacionada con la anterior y con la que se verá a continuación. Se trata de *display*, que puede ser abreviada con sólo *di*. Su misión es la de mostrar la expresión deseada entre la que puede incluirse un resultado<sup>15</sup>. De este modo, escribiendo...

<sup>14</sup> Caso de que la orden fuera del tipo de estimadores, la instrucción para ver los resultados guardados sería *ereturn list*; algo similar podría decirse para la instrucción *sreturn list*.

<sup>15</sup> En la expresión puede incluirse cualquier operación matemática. De ahí que esta instrucción se conozca también como la “calculadora”. Además, en una misma orden pueden escribirse varias instrucciones separadas por espacios o comas (preferible esta última opción). Puede abreviarse con *di*. Por ejemplo, *di ln(10), sqrt(10)* proporciona inmediatamente el logaritmo neperiano y la raíz cuadrada del número 10.

```
display "La media es" r(mean) "y la desviación típica" r(sd)
```

... el resultado sería este:

```
La media es 2.4592088 y la desviación típica 1.0945425
```

Además de combinar texto y resultados en la instrucción *display* pueden incluirse, entre otras funcionalidades más complejas, expresiones y formatos. Con el siguiente ejemplo se puede mostrar el rango y el coeficiente de variación.

```
display "Rango= " %1.0f r(max)-r(min) "; C. de variación= " %3.1f r(sd)/r(mean)
```

De este modo, el resultado sería en este caso:

```
Rango= 4; C. de variación= 0.4
```

Una de las características de las listas de resultados (*r*, *e* o *s*) es que son suplantadas en el momento en el se ejecuta otra instrucción posterior. Para evitar perder estos datos obtenidos, Stata ofrece la posibilidad de guardarlos en macros<sup>16</sup>.

La orden para dar un valor a una macro es:

```
global nombremacro=expresión
```

De este modo, si se desea guardar el valor de la media de los acuerdos para disponer de ella hasta que se trasplante por otro valor, se borre o se salga del programa, habría que escribir tras la instrucción *summarize* acuerdos, una orden del siguiente tenor:

```
global media_acuerdos=r(mean)
```

Ahora bien, siempre que se quiera volver a recordar el contenido de esta macro, hay que preceder su nombre con el carácter \$. Consiguentemente, si se desea obtener una nueva variable que recoja las diferencias de los valo-

<sup>16</sup> Hay dos tipos de macros: globales y locales. En este libro sólo se tratarán las primeras, pues son de uso más sencillo y suficientes, siempre y cuando no se entre en temas de programación. Los resultados también pueden guardarse en variables con las órdenes *generate* o *replace* o en otros receptáculos de constantes llamados escalares (*scalars*) y matrices con las instrucciones *scalar* y *matrix*. Una explicación de la diferencia entre una macro y un escalar se encuentra en Stata (2009e: 375-376).

res con respecto a la media, se podría realizar sin necesidad de escribir los valores numéricos de esta forma:

```
generate acuerdos0=acuerdos-$media_acuerdos
summarize acuerdos0
return list
display "Nueva media: " %3.1f r(mean), "Media anterior: " %3.1f $media_acuerdos
```

Con las instrucciones anteriores se crea una nueva variable, restando de la antigua (*acuerdos*) el valor almacenado en la macro *\$media\_acuerdos*. Por ello, la media será próxima a 0 y cambia el mínimo y el máximo. Número de observaciones y desviación típica quedan invariables. Ahora, los valores de los resultados *r* son distintos, pues son los correspondientes a *acuerdos0*, en lugar de *acuerdos*.

#### **ILUSTRACIÓN 5.23. Empleo de resultados, macros e instrucción *display***

Variable	Obs	Mean	Std. Dev.	Min	Max
acuerdos0	5283	2.01e-08	1.094543	-2.459209	1.540791

scalars:

```
r (N) = 5283
r (sum_w) = 5283
r (mean) = 2.01164265823e-08
r (Var) = 1.198023372394011
r (sd) = 1.094542540239534
r (min) = -2.459208726882935
r (max) = 1.540791273117065
r (sum) = .0001062750816345
```

Nueva media: 0.0 Media anterior: 2.5

Las macros no sólo sirven para almacenar resultados. También pueden ser empleadas para guardar fragmentos de instrucciones. De este modo, podrían utilizarse para representar una determinada lista de variables o una cláusula, como se muestra en las siguientes líneas de código:

```
global vv="p201-p204"
global cond1="if sexo==1 & edad>35 & edad<=50"
list sexo edad $vv $cond1 in 1/40, clean
```

Con las instrucciones anteriores se crean dos macros *\$vv* y *\$cond1*. La primera contiene una lista de variables, expresada mediante el guión para indicar que no es una y otra, sino desde la primera hasta la segunda referenciada. La siguiente macro es otra cadena que se compone tanto de una cláusula *if* como de una expresión lógica compleja compuesta de tres variables,

tres valores, tres operadores de relación y dos signos lógicos (&) que los vinculan. La última orden pide un listado de dos variables, más las cuatro contenidas en la primera macro, con la condición expresada en la segunda, limitada a los cuarenta primeros casos (in 1/40).

#### ILUSTRACIÓN 5.24. Listado de casos empleando macros

	sexo	edad	p201	p202	p203	p204
7.	Hombre	40	en desacu	de acuerd	de acuerd	.a
14.	Hombre	48	de acuerd	de acuerd	de acuerd	de acuerd
22.	Hombre	36	de acuerd	de acuerd	de acuerd	de acuerd
36.	Hombre	36	de acuerd	de acuerd	de acuerd	de acuerd
37.	Hombre	44	de acuerd	de acuerd	muy de ac	muy de ac

Finalmente, en este apartado, por motivos diversos, se va a explicar una instrucción descatalogada del manual, pero bastante útil para principiantes que desean ir aprendiendo nuevas posibilidades del programa que simplifican el trabajo<sup>17</sup>.

Supóngase que se desean estandarizar las tres variables creadas con la orden *egen*, a partir de los cuatro ítems de la pregunta dos. Las variables que se crearon fueron *total*, con la suma de la puntuaciones; *promedio*, con su suma, y *perdidos*, con el número de no respuestas a la mencionada pregunta. Si se desearan estandarizar estas tres variables, se necesitarían tres líneas. Con la instrucción *for* puede realizarse de una sola vez en una sola línea, que genera un bucle con los distintos elementos que se le especifiquen.

```
for var total-perdidos : egen X_e=std(X)
```

Lo más sorprendente de esta instrucción es la X mayúscula que aparece dos veces. Esta es una macro especial que se va reemplazando automáticamente y secuencialmente con los elementos de la lista aportada delante de los dos puntos, en este caso, con las tres variables comprendidas entre *total* y *perdidos*. De este modo, la orden anterior es equivalente a las tres siguientes:

```
egen total_e=std(total)
egen promedio_e=std(promedio)
egen perdidos_e=std(perdidos)
```

Funciona del siguiente modo. Después de la orden *for* puede especificarse *var*, *numlist* o *any*, dependiendo de si la lista va a ser de variables anti-

<sup>17</sup> Se explica porque, aun descatalogada desde la versión 9, sigue funcionando y es bastante más fácil que aquellas que la han suplantado, esto es, que *foreach* y *forvalues*. En todo caso, al usuario avanzado de Stata se le recomiendan las dos últimas.

guas, una lista de valores u otra lista de cualquier otra cosa (nuevas variables, cadenas o funciones, por ejemplo), seguida del listado correspondiente. Una vez terminada la lista, se insertan los dos puntos y a continuación la X suplantará los elementos de la lista en la instrucción siguiente siempre y cuando aparezca. En el ejemplo anterior aparecía dos veces: una, al final entre paréntesis para expresar la variable existente; la otra, delante del signo igual, seguida de \_e , para distinguir las nuevas variables, indicando que se trata de estandarizaciones.

Otra función útil de esta instrucción es crear variables dicotómicas o indicadores a partir de variables categóricas. En primer lugar, se pone un ejemplo muy fácil. Se va a convertir en una sola línea la variable *sexo* en dos: hombre y mujer.

```
for numlist 1/2 : generate sexoX=(sexo==X) if sexo<.
```

De esta manera, la instrucción se multiplica por los dos valores especificados en la lista y genera dos variables: *sexo1* y *sexo2*, ubicando, respectivamente, un 1 en estas a hombres y a mujeres. Es decir, en *sexo1*, los hombres tendrán 1 y las mujeres 0; mientras que en *sexo2*, las mujeres tendrán la unidad y los hombres el valor nulo. Aunque en estos datos no haya valores perdidos, siempre es precavido terminar la instrucción con la cláusula *if variable<.,* con el fin de transmitirlos a las variables dicotómicas.

A continuación se explica un caso similar, algo más complejo, pues en lugar de poner número a las nuevas variables (*sexo1* y *sexo2*), se pone un nombre distinto, para lo que se necesita una lista adicional que se separa de la anterior mediante una barra invertida (\). La primera lista será utilizada con la macro X, mientras que la segunda se empleará con la macro Y (también mayúscula). Para que quede más claro, se exponen dos instrucciones con variables distintas (*sexo* y *edad* recodificada), a fin de crear indicadores con ellas:

```
for any hombre mujer \ numlist 1/2: generate X=(sexo==Y) if sexo<
for any joven adulto maduro mayor \ numlist 1/4: generate X=(edadr==Y) if edadr<
list sexo hombre mujer edad edadr joven-mayor in 2/6, clean noobs
```

#### **ILUSTRACIÓN 5.25. Listado de indicadores generados a partir de variables categóricas**

sexo	sexol	sexo2	edad	edadr	joven	adulto	maduro	mayor
Mujer	0	1	37	36-50	0	1	0	0
Hombre	1	0	19	Hasta 35	1	0	0	0
Hombre	1	0	60	51-65	0	0	1	0
Hombre	1	0	27	Hasta 35	1	0	0	0
Hombre	1	0	66	Mas de 65	0	0	0	1

## 5.4. Ejercicios

1. Abre el fichero cis2794, ordénalo por el número de cuestionario y lista los diez últimos casos. Comprueba cómo estuvieron planificados en este estudio 2.500 casos pero sólo se dispone de 2475 entrevistas.
2. Obtén con el mismo fichero de dos modos diferentes la edad media y el número de casos de hombres y mujeres. (Empleando en ambas la instrucción *summarize*). Con el procedimiento más cómodo, haz el mismo cálculo de edad media y número de casos para los distintos estratos de tamaño de municipio.
3. Con cuidado de no sumar la no respuesta (99), construye con las tres variables de la pregunta P.28 una nueva que exprese los minutos que una persona dispone de tiempo libre a la semana.
4. Recodifica la edad en intervalos de 10 años, poniéndole etiquetas a los valores. Asimismo, recodifica la variable del ejercicio anterior (minutos que una persona dispone de tiempo libre) en los siguientes intervalos: hasta 1 hora diaria; entre 1 hora diaria y 2; de 2 a 4 horas diarias; de 4 a 8 horas diarias, más de 8 horas diarias.
5. Unifica las preguntas P.34 y P.34a en una sola que sea el nivel de estudios del entrevistado.
6. Estandariza la variable *edad* (P.33) de modo que tenga media 0 y desviación típica 1. Haz lo mismo con la obtenida en el ejercicio 3. Compara los mínimos y máximos de los resultados de ambas variables.
7. Construye una variable con la cantidad de equipamiento que posee el domicilio del entrevistado (P.42).
8. Teniendo cuidado en la recodificación de los valores iniciales, construye un índice de valoración de la situación política que vaya de 0 a 10, compuesto por la valoración presente más la prospectiva (P.2), contando la primera el doble de la segunda. Haz lo mismo con la situación económica (preguntas 3 y 4), guardando en macros la media y desviación típica de ambas.
9. Con una sola línea de código convierte los índices del ejercicio anterior en variables estandarizadas con media 5 y desviación típica 1,5. Lista después los casos que estén fuera del rango de 0 a 10 en una u otra variable.



# 6

## Gráficos con Stata

Una de las capacidades básicas que ha de tener cualquier aplicación estadística es la de ser capaz de generar gráficos. Tan importante es la capacidad de tratamiento de variables y la de generación de estadísticos como la de hacer que se muestren los datos representados mediante una imagen, que en muchas ocasiones dice bastante más que mil números.

Hay muy diversos tipos de gráficos en la representación estadística, pero, con objeto de simplificar la amplia variedad existente, estos pueden ubicarse en dos clasificaciones: por un lado, la del número de dimensiones que representan y, por el otro, el tipo de variables representado. En el primer caso se pueden encontrar gráficos unidimensionales, que representan los valores y frecuencias de cada variable independientemente de las demás, si las hubiere; gráficos bidimensionales, en los que se muestran distribuciones conjuntas de dos variables, y representaciones multidimensionales, donde se muestran distribuciones multivariantes. Es necesario precisar que no siempre coincide el concepto de dimensión con el de variable. Así, en un gráfico unidimensional pueden representarse dos o más variables, en cuyo caso, según se construya el gráfico, se podrá estudiar la asociación existente entre ellas<sup>1</sup> o comparar sus características representadas. Por otro lado, los gráficos también pueden clasificarse según el tipo de variable que quieren representar. Así, hay gráficos que se adecuan especialmente a variables cualitativas, como son el gráfico de sectores o el de barras, mientras que otros, como las nubes de puntos o el histograma, están indicados principalmente para variables cuantitativas.

La aplicación Stata es capaz de producir gráficos de tres modos distintos:

---

<sup>1</sup> Para estudiar la asociación en gráficos unidimensionales es preciso añadir a su representación de única entrada otra dimensión. Esto se logra, como se verá más adelante, mediante dos modos: con *over* la operación se realiza en los mismos ejes del gráfico, con *by* se construye otro gráfico paralelo.

- a) En primer lugar, existe una instrucción que contiene la mayor parte de los gráficos más usuales. Se trata de la instrucción *graph*, que será la única que será abordada en este capítulo<sup>2</sup>.
- b) En segundo lugar, existen otra serie de instrucciones que son capaces de realizar gráficos más específicos. En este caso nos encontramos instrucciones como la de *dotplot*, que realiza histogramas basados en puntos, o *stem*, que realiza un gráfico de tallo y hoja.
- c) Stata también dispone de ciertos procedimientos de operaciones estadísticas que se pueden complementar con algún tipo de gráfico. De este modo, instrucciones gráficas como *greigen*, *rvfplot* o *cluster dendrogram* sólo son posibles tras la realización de análisis estadísticos previos como *factor*, *regress* y *cluster*, respectivamente.

Con la instrucción más específica de gráficos (*graph*) se pueden realizar dos modalidades de representación de variables:

- a) Las univariadas, como son los gráficos de sectores (*pie*), los de barras (*bar*), los de puntos (*dot*) y los de caja (*box*).
- b) Las bivariadas, en gráficos de dos dimensiones (*twoway*) o múltiples (*matrix*).

## 6.1. Características de los gráficos de Stata

Antes de explicar los distintos gráficos que pueden realizarse con Stata, es preciso presentar una breve introducción acerca de cómo esta aplicación los produce, ya que, como se pudo comprobar en el capítulo 2, las imágenes que se producen no se ubican en la pantalla de resultados, sino en una ventana propia aislada de los estadísticos y de otros gráficos. Además, si no se mantienen las precauciones debidas, la producción de un segundo gráfico hace desaparecer al primero. Esto es así porque cada gráfico es guardado en un espacio de la memoria interna del ordenador al que por omisión se le domina *Graph*, que desaparece para siempre al salir de la aplicación, o es suplantado por un nuevo gráfico al solicitarlo sin nombre.

Por tanto, es fundamental saber cómo dar un nombre distintivo a un gráfico y obtener una lista de los gráficos almacenados en un de-

---

<sup>2</sup> A partir de la versión 8, Stata implementó una sintaxis bastante diferente de las anteriores. Sin embargo, aún se permite que los viejos programas puedan ejecutarse. Para ello, ha de cambiarse la instrucción *graph* por *graph7*, o bien, en el interior de un programa, advertir al comienzo de que se está trabajando con una versión anterior a la 8, con la instrucción *version*.

terminado momento. Como de momento sólo se conoce la instrucción *histogram*, mencionada en la sección 2.4, se empleará esta para ejemplificar cómo el nombre de cualquier gráfico se pone mediante la opción *name* seguida, entre paréntesis, por el nombre deseado y una coma para indicar las subopción *replace*, que evita el error en caso de repetición del gráfico.

```
sysuse auto, clear
histogram price
histogram mpg, name(Hist_Mpg, replace)
graph dir
```

De este modo, además de generarse dos pantallas de gráficos en pestanas independientes, aparece la lista de sus nombres internos en la pantalla de resultados, gracias a la última instrucción:

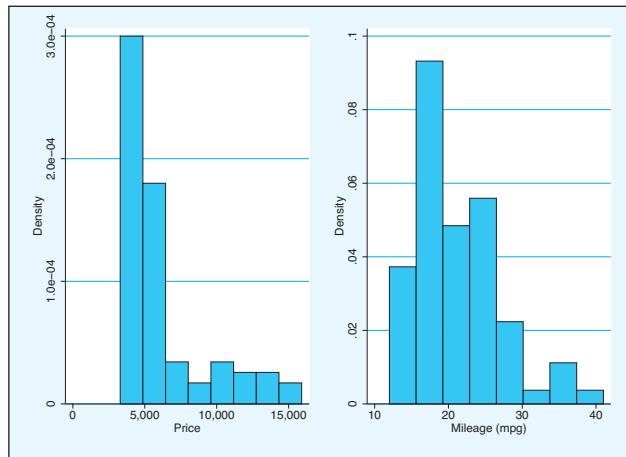
#### **ILUSTRACIÓN 6.1. Listado de gráficos en memoria**

```
Graph Hist_Mpg
```

Una vez que se dispone de una serie de gráficos almacenados en la memoria, además de listarlos, se pueden realizar las siguientes operaciones: a) describirlos (*describe*) con informaciones tales como cuándo fueron creados, con qué datos e instrucción, entre otras; b) renombrarlos (*rename*) para que tengan otro modo de llamarlos; c) revisualizarlos (*display*), para que vuelvan a aparecer en una ventana visible para el investigador; d) copiarlos (*copy*) para tener más de un ejemplar del mismo gráfico; e) borrarlos (*drop*) para que no ocupen espacio en memoria. Con los gráficos anteriores realizados podrían funcionar las siguientes instrucciones, o incluso combinarlos (*combine*) para que se presente en un mismo recuadro:

```
graph describe Graph
graph rename Graph Hist_Price, replace
graph display Hist_Mpg
graph copy Hist_Mpg Mpg, replace
graph drop Hist_Mpg
graph combine Hist_Price Mpg, name(G1, replace)
```

El resultado de la última instrucción, siempre y cuando se hayan construido previamente las figuras *Hist\_Price* y *Mpg*, es el gráfico combinado representado en el gráfico 6.1:

**GRÁFICO 6.1. Combinación de gráficos**

Más importante aún que conocer que los gráficos se mantienen en la memoria RAM del ordenador con un nombre sujeto a las convenciones de las variables es saber que pueden ser grabados en el disco, para poderlos usar siempre que se necesiten tanto con el programa Stata como con cualquier otro que sea capaz de leer y procesar ficheros de texto. Realizar esta última operación es posible de tres formas:

1. Mediante el menú contextual que se obtiene al pulsar el botón derecho del ratón ubicado encima de un gráfico.
2. Al salir del editor de gráficos, pues si se quieren guardar los cambios, el gráfico ha de grabarse en el disco.
3. Mediante la opción *saving(nombrefichero)* añadida a cualquier instrucción gráfica.
4. Mediante la instrucción *graph save*, cuya estructura es la siguiente:

```
graph save [nombregrafico] nombrefichero [, replace asis]
```

Con esta orden el gráfico se graba en un fichero del directorio indicado en su nombre o, si no se indica, en el directorio por defecto con la extensión *.gph*.

Sin embargo, este fichero guardado sólo es legible con Stata. Si se quiere disponer de un fichero que pueda ser tratado con cualquier programa, incluyendo especialmente los procesadores de texto y de gráficos, ha de emplearse la orden *graph export*.

```
graph export nombrefichero.ext [, replace name(nombregrafico) as(pslepslwmf  
lwmflpngltiflpictlpdf)]
```

De este modo, según la extensión que se ponga al nombre del fichero o según la clave que se seleccione a la opción *as*, el fichero se grabará según las normas de los ficheros postScript (*ps*), postScript encapsulado (*eps*), metafichero de Windows (*wms*), gráfico portátil de redes (*png*), *tiff*, formato Macintosh (*pict*) o *pdf*.

También es posible hacer estas operaciones (guardar en formato nativo o exportable) haciendo clic con el botón derecho del ratón encima de un gráfico. Con esta operación sale un menú contextual mediante el que se puede grabar el fichero con el formato deseado (*save graph*); copiarlo al portapapeles (*copy*), para poderlo trasladar a otro programa en código (*wmf* o *emf*, según preferencias) mediante la combinación de teclas Ctrl+v en el programa de destino; imprimirla (*print*) en la impresora que se seleccione, o modificarlo (*Start graph editor*), empleando la utilidad cuyas características principales se verán al final de este capítulo.

Solicitar un gráfico con unas determinadas características es un proceso bastante complejo, que requiere largas y complejas instrucciones en inglés. Afortunadamente, desde la versión 8 de Stata, los menús simplifican mucho la construcción de los gráficos. Asimismo, desde la versión 10, se ha incorporado un editor de gráficos que permite realizar mediante el ratón cuantas modificaciones se consideren oportunas. Ambas posibilidades son tratadas al final de este capítulo, pero, fieles al estilo de este manual, el enfoque principal será la explicación de las órdenes.

## 6.2. Gráficos unidimensionales

### 6.2.1. Gráficos de sectores

Los gráficos de sectores son representaciones de los datos en un círculo cuyos segmentos representan proporcionalmente la frecuencia de los valores contenidos de una o varias variables.

La instrucción mínima para realizar gráficos de sectores es la siguiente:

```
graph pie listadevariables
```

Hay que tener en cuenta que esta instrucción produce un gráfico en el que cada variable es un sector cuya área viene determinada por la suma de los valores de variables.

Esto implica que, para obtener un gráfico de sectores en el que un sector representara a los hombres y el otro a las mujeres, los datos han de disponerse de dos posibles modos:

1. Disponiendo de un fichero con un solo caso y dos variables: *Hombre* y *Mujer*, con valores que representen sus respectivas frecuencias:

**ILUSTRACIÓN 6.2. Disposición de datos para gráfico de sectores**

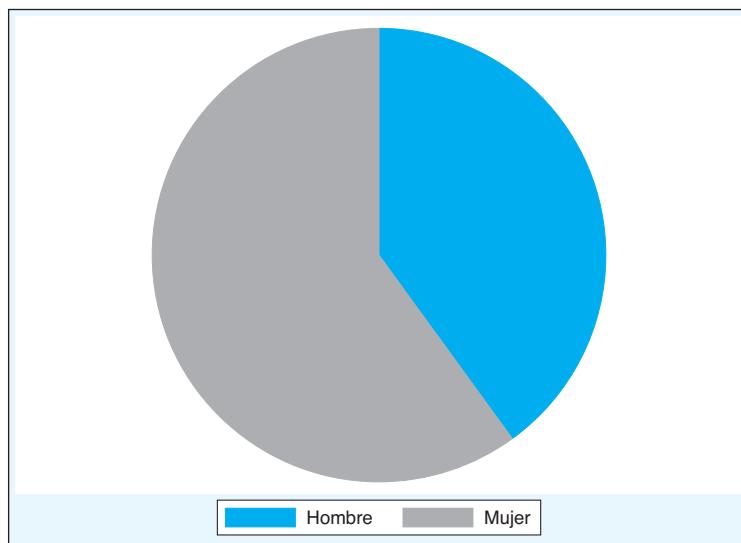
	Hombre	Mujer
1	40	60

A partir de los datos contemplados en la ilustración 6.2, bastaría con indicar la instrucción...

```
use ejemplo6a, clear
graph pie Hombre Mujer, name(G2, replace)
```

... para producir el siguiente gráfico:

**GRÁFICO 6.2. Gráfico de sectores**



2. Sin embargo, lo más común es disponer los datos por individuo en una variable categórica, tal como pueda ser el *sexo*, con 5.000 casos

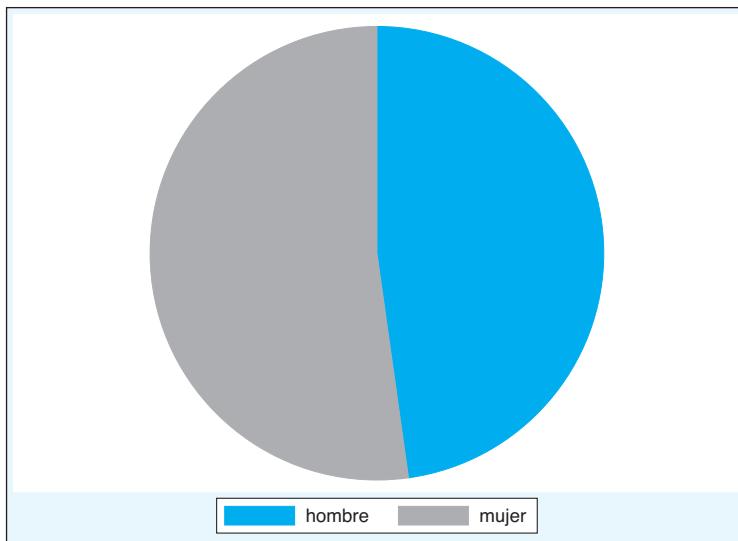
y dos valores *hombre* y *mujer*, en cuyo caso habría que escribir la instrucción del siguiente modo:

```
graph pie, over(sexo), name(G3, replace)
```

... donde *sexo* es la variable que se quiere representar en el gráfico de sectores.

De este modo se genera el gráfico 6.3, donde puede advertirse que el programa pone automáticamente a cada uno de los sectores las etiquetas de los valores que tiene la variable original.

**GRÁFICO 6.3. Gráfico de sectores con la variante *over***

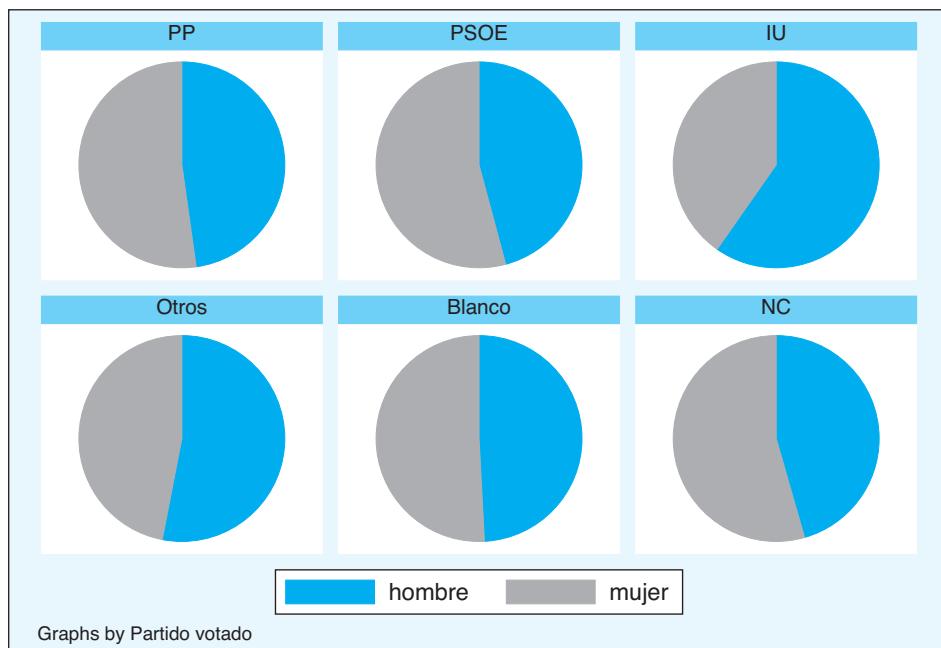


La instrucción *graph pie* admite la posibilidad de introducir una variable categórica para la obtención de tantos gráficos como valores tenga esta. De este modo, si se desean los perfiles de sexo en función de los distintos votantes, hay que especificarlo mediante la opción *by(variable)...*

```
graph pie, over(sexo) by(Voto_2000), name(G4, replace)
```

... que da lugar al siguiente gráfico bidimensional, donde se puede estudiar el perfil de género de los votantes de cada uno de los partidos:

**GRÁFICO 6.4.** Gráficos de sectores según una segunda variable



### 6.2.2. Gráficos de barras

Los gráficos de barras, recomendados en el caso de que se tenga un número mayor de categorías en la variable que se quiere representar, necesitan instrucciones con opciones bastante distintas a las de los gráficos de sectores. Sin embargo, la sintaxis general es muy similar a la anterior:

```
graph bar listadevariables
```

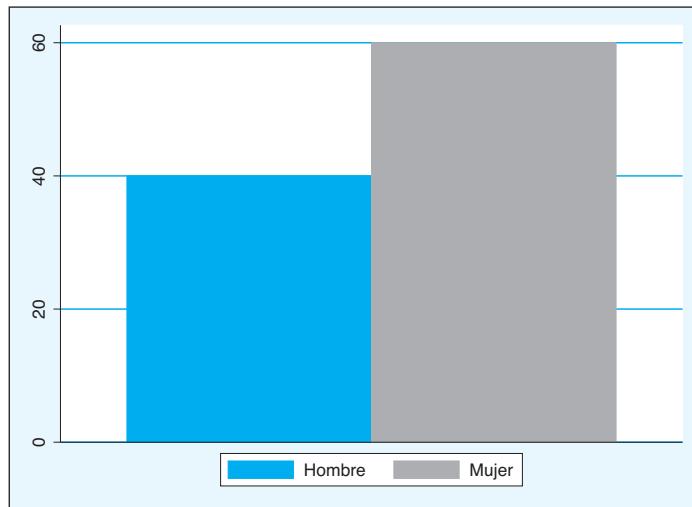
De este modo, la instrucción del primer gráfico realizado con la opción *pie* ahora quedaría del siguiente modo<sup>3</sup>:

```
use ejemplo6, clear
graph bar Hombre Mujer, nolabel name(G5, replace)
```

<sup>3</sup> En esta instrucción se produce la paradoja de que para que aparezcan en la leyenda los nombres de las variables (*Hombre* y *Mujer*), se debe especificar la opción *nolabel*. Si esta no aparece, las etiquetas mostradas son las automáticas del gráfico, es decir, “Mean of Hombre” y “Mean of Mujer”.

... y produciría el siguiente gráfico:

**GRÁFICO 6.5. Gráfico de barras**



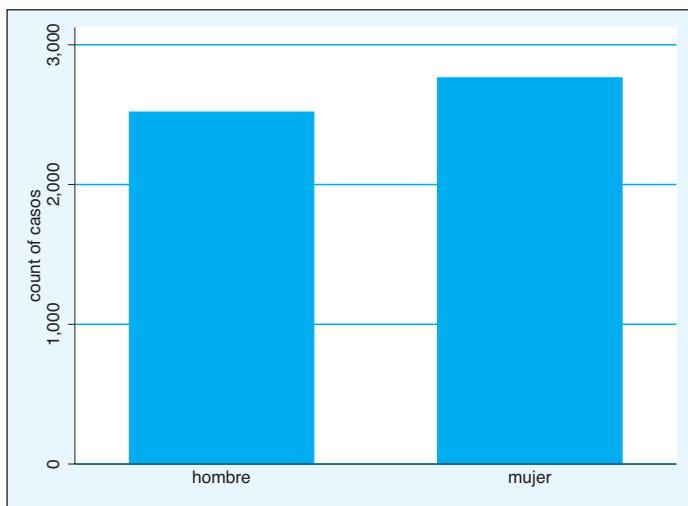
No obstante, es más frecuente disponer los datos en un fichero en el que cada registro representa un caso, en cuya situación, en el caso de los gráficos de barras, no se puede utilizar directamente la opción *over* como se aplicó en la modalidad de sectores. Para poder hacer algo similar, hay que confeccionar el gráfico en dos pasos: en el primero, mediante dos instrucciones, se genera una constante ficticia, equivalente al peso en porcentaje del caso<sup>4</sup>, y en el segundo se pide la representación del recuento de esta<sup>5</sup> cruzada con la variable propiamente dicha, que en el ejemplo siguiente es *sexo*. Y esto es debido a que Stata considera el gráfico de barras más como un caso de variable numérica (de intervalo o de razón) que de variable con atributos (nominal u ordinal)<sup>6</sup>.

```
use panel6
tabulate sexo
generate casos=100/r(N)
graph bar (count) casos, over(sexo) name(G6, replace)
```

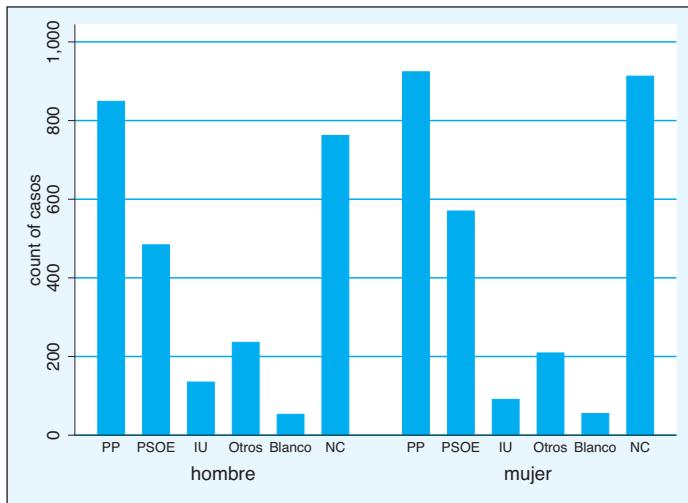
<sup>4</sup> Si se quieren proporciones, en lugar de porcentajes, basta con sustituir el 100 por un 1.

<sup>5</sup> Para que calcule los porcentajes (en vez del número de casos) hay que usar (*sum*) en vez de (*count*) como función de resumen de casos. Si sólo se desea el número de casos, se puede hacer de manera más simple sustituyendo las dos instrucciones anteriores por: gen casos = 1.

<sup>6</sup> Una alternativa al uso de instrucciones de Stata para estos gráficos es el empleo de la instrucción escrita por Cox (2004) llamada *catplot*. Se puede instalar mediante la orden *ssc install catplot*.

**GRÁFICO 6.6.** Gráfico de barras con la variante *over*

En estos gráficos cabe también la posibilidad de realizar un control por una segunda variable para realizar un gráfico bidimensional de barras, que es muy útil para representar gráficamente tablas de contingencia (véase el capítulo 8). En este caso, para cambiar de ejemplo, se utiliza el sexo como independiente y se emplea la intención de voto como variable determinada, para ver la distribución del voto de hombres y mujeres:

**GRÁFICO 6.7.** Gráfico de barras con variable de control

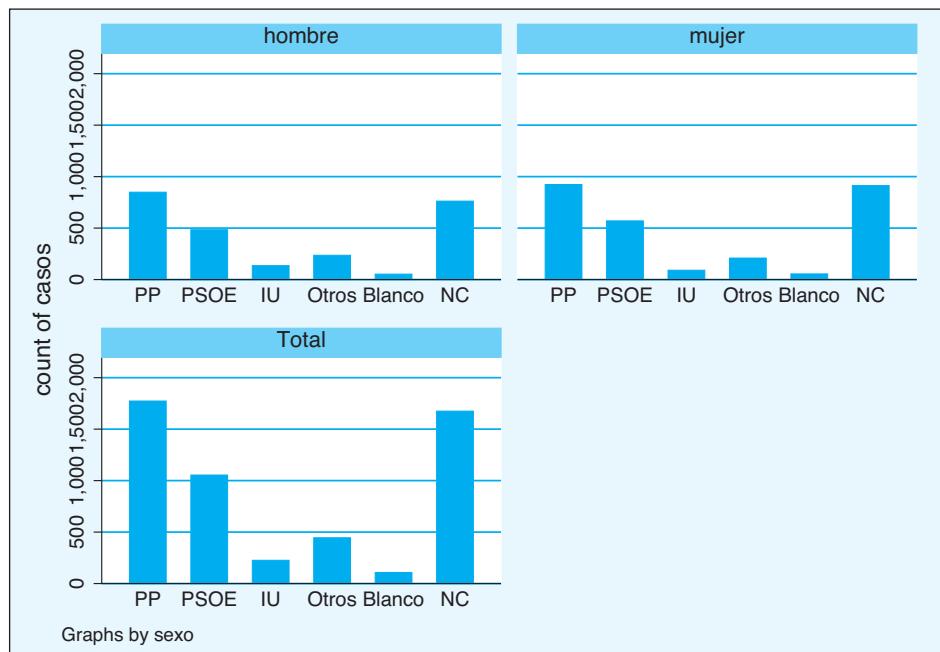
En la figura anterior se han obtenido dos grupos de barras: unas para los hombres y otras para las mujeres. Para obtenerlos se ha tenido que escribir esta instrucción<sup>6</sup>:

```
graph bar (count) casos, over(Voto_2000) over (sexo) name(G7, replace)
```

Hay otro modo de que se produzca un resultado similar al anterior. Se trata de mostrar tantos gráficos como valores tenga la variable que se especifique detrás de la opción *by(variable)*. Incluso, si se desea, puede obtenerse al mismo tiempo el gráfico correspondiente al conjunto de la muestra, si se añade después de la variable la opción *total*:

```
graph bar (count) casos, over(Voto_2000) by(sexo, total ) name(G8, replace)
```

**GRÁFICO 6.8. Gráficos de barras con la opción *by***

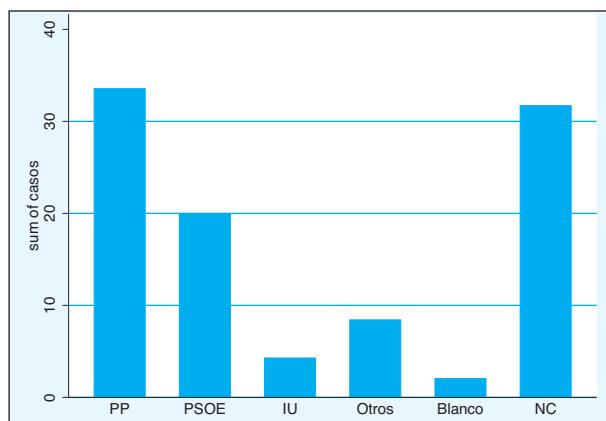


Es obvio que las etiquetas del eje que el programa crea por omisión no es la deseable en el caso de que se quiera publicar el gráfico en castellano. Para arreglarla es preciso leer el apartado del editor de gráficos (6.6).

Especialmente en este gráfico se nota cómo hasta ahora lo que se representan son frecuencias y no porcentajes. Para obtenerlos o para representar proporciones<sup>7</sup>, en lugar de frecuencias, hay que solicitar la estadística (*sum*), en lugar de *count*, que aparecía en los anteriores gráficos.

```
graph bar (sum) casos, over(Voto_2000) name(G9, replace)
```

**GRÁFICO 6.9. Gráfico de barras con frecuencias**

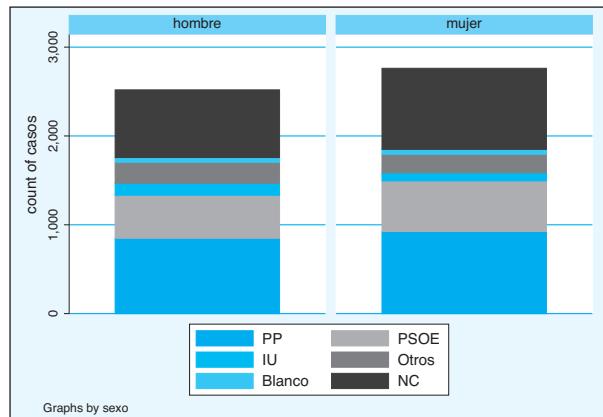


Una variante ineludible del gráfico de barras es la apilada, en la que en lugar de aparecer paralelas las barras correspondientes a las categorías de la variable, aparecen contiguas en la misma columna. Esta alternativa permite, en la mayor parte de los casos, facilitar la comparación entre categorías. Para obtenerla, es necesario añadir a la instrucción dos opciones: la primera es *asyvar*, que trata la variable expresada en *over()* como si fueran valores de distintas variables. Por eso las barras aparecen dibujadas con distintos colores. La segunda opción es *stack*, que, como su propio nombre indica, es la que hace que las barras queden apiladas.

```
graph bar (count) casos, over(Voto_2000) asyvar by(sexo) stack name(G10, replace)
```

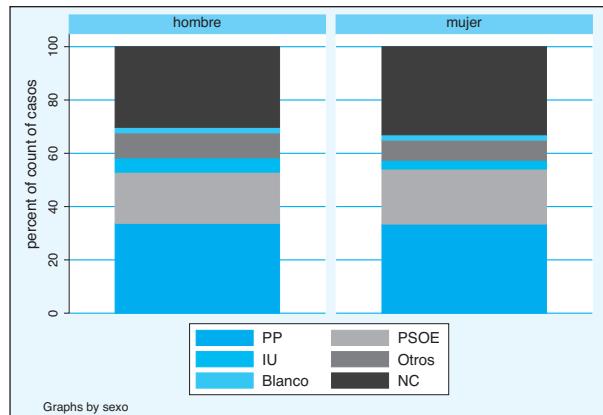
---

<sup>7</sup> Sacará porcentajes o frecuencias según se haya construido la variable ficticia con la que se construyen los gráficos de barras (*casos*, en este ejemplo). Como más arriba se construyó dividiendo 100 por el tamaño de la muestra (*\_N*), entonces se obtienen porcentajes. Si se hubiera utilizado 1, en lugar de 100, se habrían obtenido proporciones.

**GRÁFICO 6.10.** Gráfico de barras apiladas

Como puede fácilmente apreciarse, por el hecho de acumular el número de casos, las alturas no alcanzan el tope y la de las mujeres, más numerosas en la muestra, es más alta que la de los hombres. Para igualar las bases de la comparación, es preciso añadir la opción *percent*, en cuyo caso la escala que representan las frecuencias cambia hasta tener el máximo de 100 y, en consecuencia, todas las barras se igualan.

```
graph bar (count) casos, over(Voto_2000) asyvar by(sexo) stack percent ///
name(G11, replace)
```

**GRÁFICO 6.11.** Gráfico de barras apiladas e igualadas

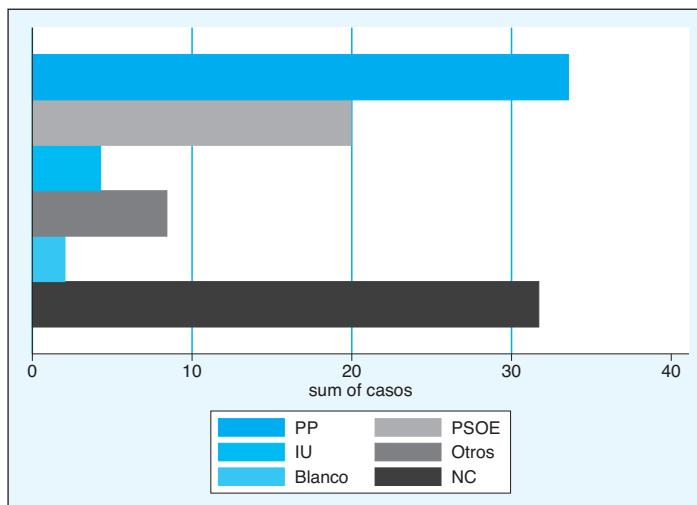
Finalmente, hay que señalar que todos los gráficos de barra aquí expuestos pueden dibujarse horizontalmente. Para ello, sólo es preciso cambiar la segunda palabra de la instrucción por *hbar* en lugar de *bar*.

Por ejemplo, si se desea, dibujar la intención de voto en barras horizontales, se debería escribir la siguiente línea:

```
graph hbar (sum) casos, over(Voto_2000) asyvar name(G12, replace)
```

De este modo, se obtiene el siguiente gráfico con barras de distinto color por haber especificado la opción *asyvar*:

**GRÁFICO 6.12. Gráfico de barras horizontales**



### 6.2.3. Histogramas

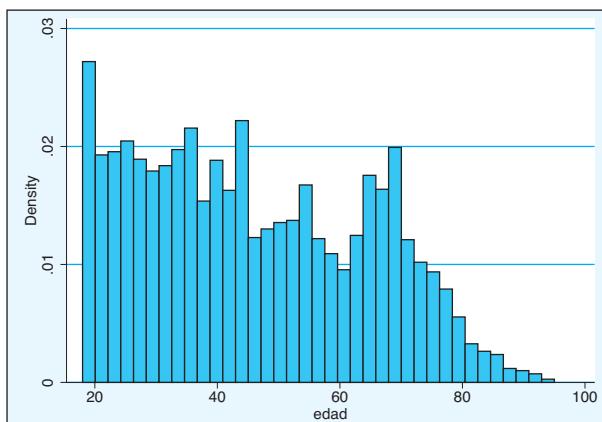
Los histogramas son gráficos que se emplean para la representación de variables cuantitativas continuas. Consisten en dividir los valores en una serie de intervalos y representar cada uno de estos con un área proporcional a su tamaño. Generalmente, los valores se expresan en el eje de abscisas de un gráfico de coordenadas, mientras que, en el caso de que todos los intervalos tengan amplitud constante, en las ordenadas se expresan las frecuencias absolutas o relativas correspondientes a cada grupo de valores.

En Stata basta con dos palabras para generar un gráfico de este tipo: el comando *histogram*<sup>8</sup> seguido del nombre de la variable que se quiere representar:

```
histogram edad, name(G13, replace)
```

Sin ninguna otra especificación añadida, el histograma aparece del siguiente modo:

**GRÁFICO 6.13. Histograma automático**

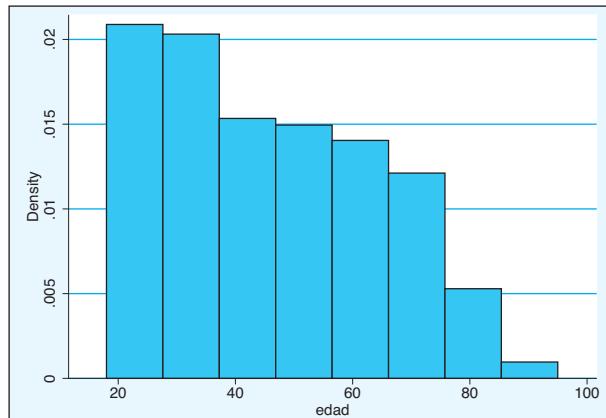


Para obtener un número no automático de intervalos en el histograma, existe la opción *bin(#)*, siendo # el número de ellos que se quiere queden dibujados. De este modo si se desean ocho intervalos, en lugar de los 43 anteriores, debería escribirse:

```
histogram edad, bin(8) name(G14)
```

---

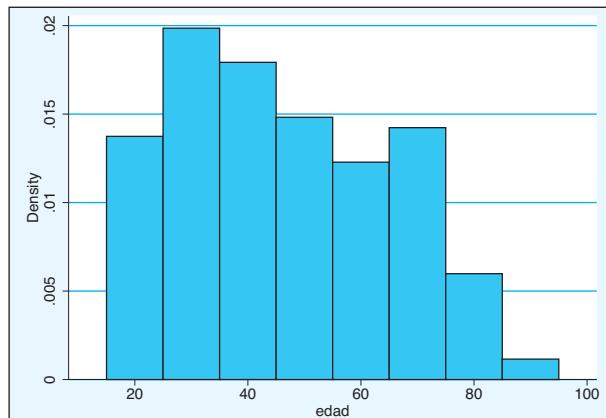
<sup>8</sup> Esta (*histogram*) es una de las instrucciones específicas (diferentes a *graph*) para realizar gráficos. Sin embargo, en este caso se puede obtener el mismo resultado con el siguiente bloque de órdenes: *graph twoway histogram*, especialmente útil cuando se quieren integrar los histogramas con otro tipo de representación bivariada. Por eso, en este contexto donde se están viendo los gráficos de una sola variable, y por razones de brevedad, sólo se señala la primera forma de solicitarlos.

**GRÁFICO 6.14.** Histograma con ocho intervalos

Pero también es posible especificar, en lugar del número de intervalos, el ancho que se desea tengan las barras a través de la opción *width(#)* e incluso el punto de partida con *start(#)*. Y obvio es que ambas se pueden combinar para obtener un histograma a gusto del usuario:

```
histogram edad, start(15) width(10) name(G15)
```

Con esta última instrucción, el histograma adopta la siguiente forma:

**GRÁFICO 6.15.** Histograma con intervalos constantes

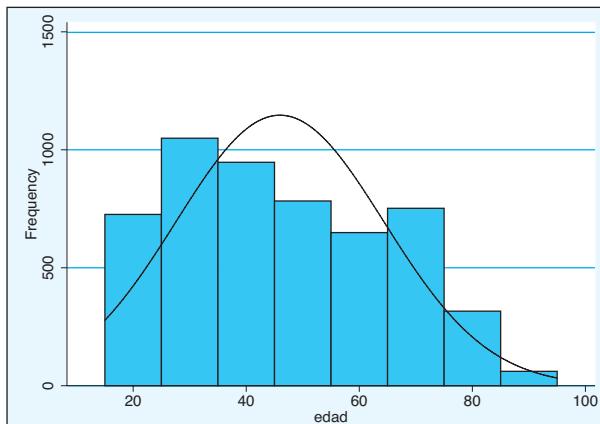
Dos opciones adicionales permiten mejorar la presentación del histograma. Por un lado, *frequency* hace mostrar las frecuencias, en lugar de los porcentajes.

Por el otro, *normal* sobrescribe sobre el histograma la curva de Gauss para que pueda compararse la distribución empírica con la distribución normal.

```
histogram edad, start(15) width(10) frequency normal name(G16)
```

El resultado es más que evidente:

**GRÁFICO 6.16. Histograma con curva normal**



#### 6.2.4. Gráficos de densidad

Una alternativa de los histogramas a la representación de las variables continuas son los gráficos de densidad, que pueden ser considerados como un método de suavización de las frecuencias de una variable.

Así como el histograma divide la distribución en un conjunto de tramos a los que se les representa por una frecuencia atribuida constante, en el caso de los gráficos de densidad también se procede a una división del rango de la variable representada en una serie de intervalos, pero en lugar de asignar una probabilidad constante, atribuye a cada valor un peso con el que se asigna la probabilidad final. El resultado es un polígono de frecuencias suavizado.

Existen muy distintos modos de obtener representaciones de densidad para la misma variable. Básicamente depende de dos parámetros: sobre todo, del ancho de los intervalos, pero también influye el método para calcular los pesos<sup>9</sup>.

Este gráfico unidimensional puede realizarse con Stata de dos modos: uno es mediante una instrucción propia llamada *kdensity*, en la que pueden

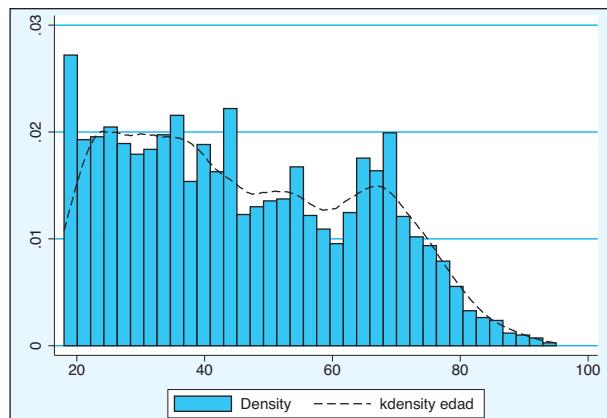
<sup>9</sup> El algoritmo utilizado por defecto es el de *Epanechnikov*, pero también emplea, siempre que se especifique en las opciones, los siguientes: *biweight*, *cosine*, *gaussian*, *parzen*, *rectangular* y *triangular*.

especificarse como opciones el ancho de los intervalos (*width(#)*), el método (véase nota 75), la comparación con una distribución normal (*normal*) o de Student (*student*) e incluso la generación de dos nuevas variables, *generate* (*variable\_con\_valores*, *variable\_con\_frecuencias*), para ver el resultado no sólo gráficamente, sino también numéricamente.

Otra manera de realizarlo es a través de la instrucción *graph twoway*, mediante la cual se pueden combinar en los mismos ejes un histograma y un gráfico de densidad, con objeto de que se aprecie el papel suavizador que tiene la estimación de las frecuencias con el sistema proporcionado por el segundo.

```
graph twoway (histogram tmi) (kdensity tmi), name(G17)
```

**GRÁFICO 6.17. Combinación de histograma y gráfico de densidad**



En el histograma se aprecia cómo las alturas se ven afectadas por la acumulación de casos en una determinada categoría. En este caso, especialmente la primera barra queda suavizada mediante la línea que se genera con la ponderación de Epanechnikov.

### 6.2.5. Gráficos de caja

Los gráficos de caja poseen una peculiar importancia en el análisis exploratorio de datos. Consisten en la representación de los datos en un rectángulo de anchura arbitraria y longitud igual al rango intercuartílico. Esto se logra dibujando uno de los límites del rectángulo en el primer cuartil y el otro en el tercero. Entre el uno y el otro también se dibuja en el rectángulo otra línea que representa la mediana. De cada extremo del rectángulo ha de salir también una línea con longitud nunca superior a vez y media el rango intercuartílico, que llegue

hasta el caso que cumpla esa condición. Finalmente, siempre que haya al menos un caso fuera de esos rangos (casos extremos), se expresa en forma de puntos.

La forma de obtener estos gráficos con Stata es similar a la de los otros gráficos ya contemplados. Cambia, en este caso, la palabra clave que sigue a la instrucción *graph*:

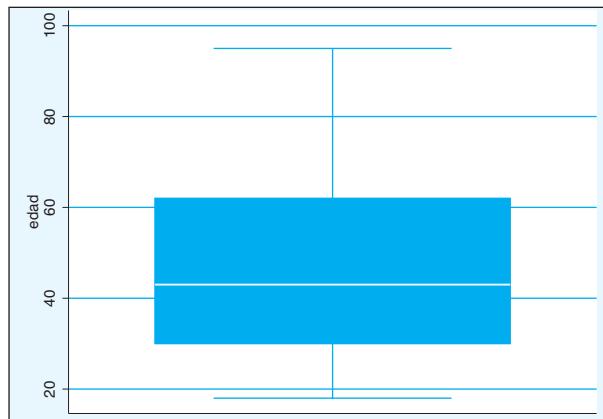
```
graph box listadevariables
```

Así, para obtener la representación de la variable *edad*, basta con escribir la siguiente instrucción.

```
graph box edad, name(G18, replace)
```

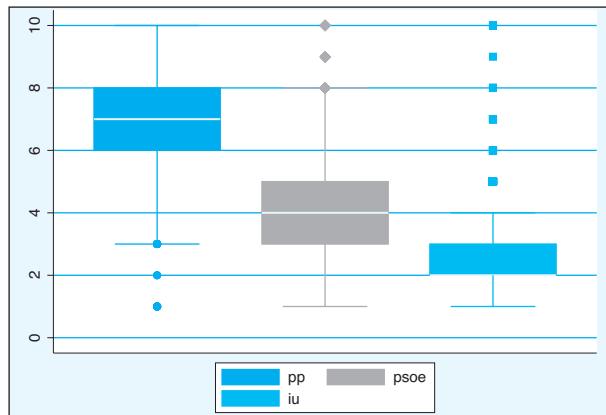
El resultado muestra el mínimo en 18, el máximo en 98, una mediana próxima a 44 y cuartiles respectivos de 30 y 63 años.

**GRÁFICO 6.18. Gráfico de caja**



El número de variables puede ser mayor que uno, en cuyo caso para cada una de ellas se dibuja una caja paralela a fin de que se puedan comparar las distribuciones. Con las reservas propias del carácter ordinal de estas variables, se puede poner como ejemplo comparativo la atribución ideológica que hacen los encuestados a los partidos españoles con representación parlamentaria en el conjunto nacional:

```
tabstat ideopp-ideoiu, statistics(p25 p50 p75)
graph box ideopp-ideoiu, name(G19, replace)
```

**GRÁFICO 6.19.** Gráfico de cajas con varias variables

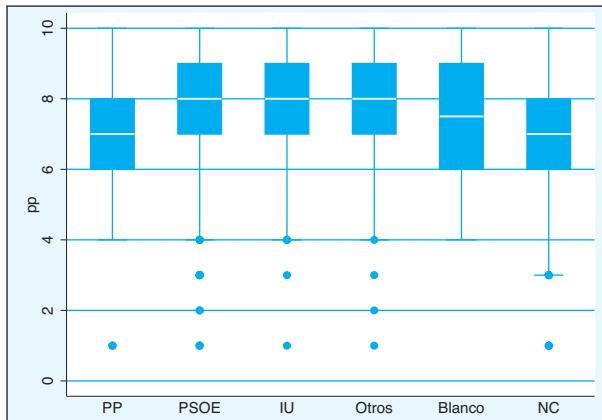
	stats	ideopp	ideoiu	idps_e
p25		2	6	3
p50		2	7	4
p75		3	8	5

En este gráfico se observa cómo el primer rectángulo, correspondiente a Izquierda Unida (*ideoiu*), no tiene línea mediana en el rectángulo, porque este estadístico coincide con el primer cuartil. La línea inferior del rectángulo llega a 1 porque es el valor empírico inferior, pero la superior sólo llega hasta el 4, porque al ser variable discreta no existe empíricamente el supuesto máximo (4,5), esto es, el tercer cuartil (3), más vez y media el rango intercuartílico (1,5). En cambio, hay cuestionarios —no se sabe cuántos por medio del gráfico— que han recogido para esta variable valores desde el 5 hasta el 10.

El rectángulo del medio, el correspondiente al PP (*ideopp*), tiene un rango intercuartílico de dos puntos (entre el 6 y el 8) con mediana en el 7. Por eso la línea de abajo alcanza hasta el 3, esto es, 6 menos vez y media el rango, que tiene en este caso el valor de 2. Y la de arriba llega hasta el máximo valor posible, es decir, el 10, porque parte desde el tercer cuartil.

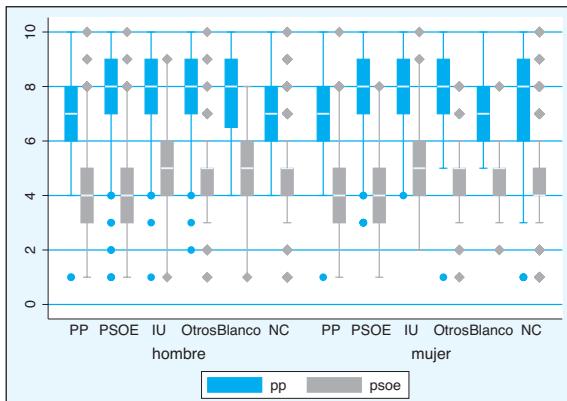
En el siguiente ejemplo, en lugar de representar distintas variables, se dibuja una sola (la ubicación en la escala ideológica del PP, *ideopp*), pero en tantos grupos como valores tenga una variable de control (el partido político al que se votó, *Voto\_2000*):

```
graph box ideopp, over(Voto_2000) name(G20, replace)
```

**GRÁFICO 6.20.** Gráfico de cajas con variable de control

Análogamente puede representarse más de una variable (en este caso, las valoraciones de las ideologías de dos partidos) por una o varias variables de control (en este ejemplo, el voto en las últimas elecciones y el sexo):

```
graph box ideopp ideopsoe, over(Voto_2000) over(sexo) name(G21, replace)
```

**GRÁFICO 6.21.** Gráfico de cajas con dos variables de control

### 6.3. Gráficos bidimensionales

La versión 12 de Stata agrupa bajo la orden *graph twoway* 40 modalidades diferentes de gráficos. Algunas poseen características muy similares, pero otras son extremadamente diferentes e incluso difíciles de considerar como bidimensionales. El programa considera bidimensional todo aquel gráfico en el que los dos ejes o escalas (la X, o eje horizontal, y la Y, o eje vertical) son numéricos. Según

esa definición, un histograma siempre es considerado bidimensional<sup>10</sup>, del mismo modo que a ciertos gráficos de barras y puntos, aunque propiamente sean unidimensionales, el programa los puede tratar como bidimensionales, siempre y cuando estén representándose variables cuantitativas (en un eje se representa el valor de esta variable y en el otro, según sea el caso, su frecuencia o el valor en otra variable). Una característica esencial y versátil de esta instrucción es la de poder combinar en el mismo gráfico distintas representaciones, sean de la misma o de diferente modalidad. Basta para ello separar las órdenes de los distintos gráficos por paréntesis, como ya se hizo en la instrucción que generó el gráfico 6.17.

En general, la instrucción para realizar gráficos bidimensionales presenta la siguiente sintaxis:

```
graph twoway modalidad [lista_de_variables] [weight=variable] [if exp]
[in rango], [opciones_comunes] [opciones_específicas]
```

Las modalidades de gráficos bidimensionales posibles en la versión 11 de Stata pueden ser agrupadas en los siguientes grupos: nubes de puntos, matrices gráficas, gráficos de líneas, gráficos de área, gráficos de ajuste, gráficos de función y gráficos de rangos. Véanse a continuación las características e instrucciones de cada uno de ellos.

### 6.3.1. Nubes de puntos

Las nubes de puntos son los gráficos específicos para el estudio de la relación entre dos variables cuantitativas y continuas. Son ideales cuando existe un número intermedio de casos, aproximadamente entre 30 y 300. Menos casos pueden arrojar poca luz sobre una relación robusta entre los datos y más casos producen superposiciones de puntos de tal naturaleza que no permiten valorar claramente dónde se produce el grueso de la asociación entre las variables.

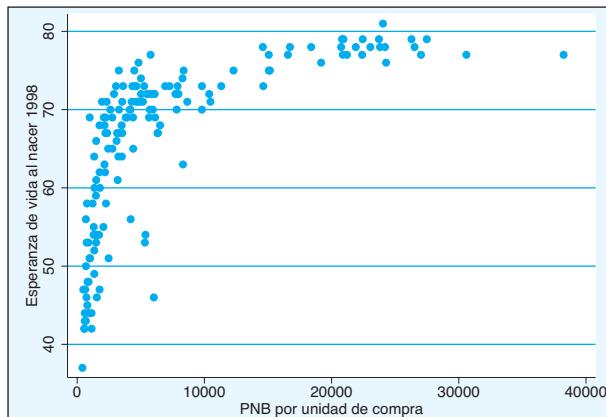
Aunque la sintaxis completa de este tipo de gráficos es *graph twoway scatter*, basta con la última palabra para que Stata reconozca la instrucción y genere inmediatamente una nube de puntos que relaciona dos variables de naturaleza cuantitativa. Así, con la base de datos mundial, se puede representar la relación existente entre el producto nacional bruto y la esperanza de vida al nacer por países. Basta con escribir estas tres palabras para producir la siguiente imagen:

```
use mundo996
scatter evn pnbppa, name(G22)
```

---

<sup>10</sup> A pesar de eso, en este capítulo la modalidad del histograma ha sido considerada entre los gráficos unidimensionales. La orden que se explicó fue *histogram*. Pero, de ahora en adelante, es conveniente saber que esta es una abreviatura de *graph twoway histogram*. Esto es importante porque este tipo de gráficos puede mezclarse con otros de naturaleza propiamente bidimensional.

GRÁFICO 6.22. Nube de puntos



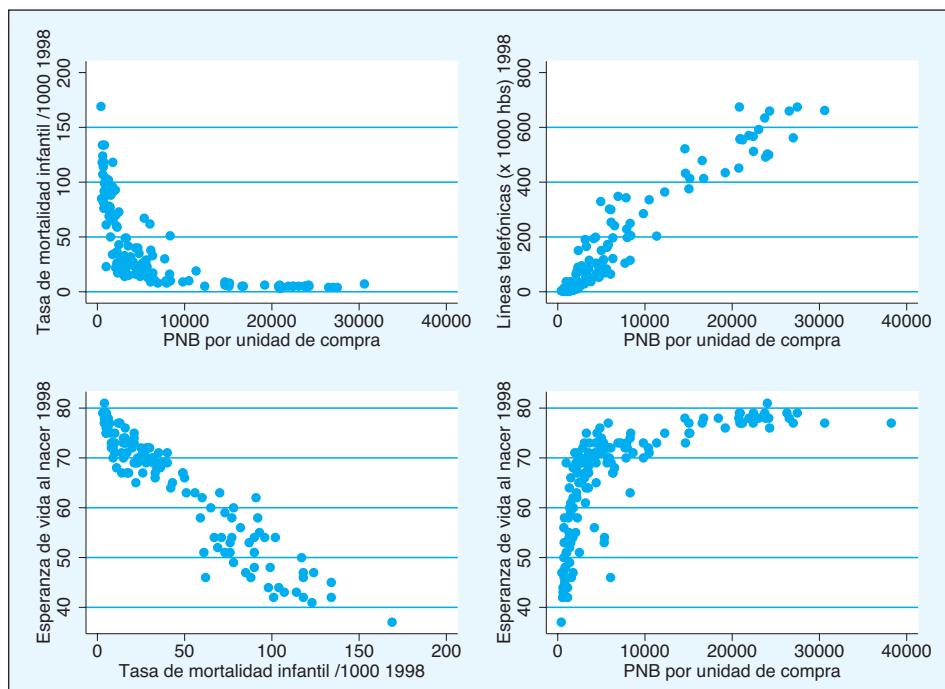
Como la principal utilidad de las nubes de puntos es estudiar la relación y asociación entre variables cuantitativas, mediante un examen de la distribución de los casos en el gráfico puede observarse si dos variables poseen relación, es decir, si son independientes o no una de otra o, dicho con otras palabras, a valores altos de una de ellas no le corresponden valores altos (o bajos) de la otra.

```
scatter tmi pnbppa, name(G23a, replace)
scatter lntfno pnbppa, name(G23b, replace)
scatter evn tmi, name(G23c, replace)
scatter evn pnbppa, name(G23d, replace)
graph combine G23a G23b G23c G23d, name (G23)
```

Puede haber muy distintos tipos de asociaciones. En el gráfico 6.23 se exponen cuatro modelos diferentes y reales de asociación entre variables. En primer lugar, se expone la relación entre el producto nacional bruto (PNB) y la tasa de inflación. Como puede apreciarse, la mayor parte de los países se concentran entre el 0% y el 10%. Sólo unos pocos, pero todos en la franja de renta baja, están por encima o por debajo de estos topes. La distribución bivariante es muy distinta en el gráfico superior derecho. En este se relaciona el PNB con las líneas telefónicas por mil habitantes, y puede verse claramente cómo a valores bajos de la primera variable le corresponden valores también pequeños de la segunda, mientras que los países de alta renta tienen en contrapartida tasas de líneas telefónicas elevadas. En este caso se está ante una asociación *lineal positiva*, puesto que los puntos siguen una pauta recta ascendente. En el tercer gráfico sucede lo contrario. La pauta sigue siendo una línea recta, pero los valores bajos de la tasa de mortalidad infantil están asociados lógicamente con valores altos de esperanza de vida al nacer y, a medida que va aumentando esta tasa, va disminuyendo la

altura en el eje vertical en la que están situados los países que tienen esperanza de vida menor. En esta situación también existe una asociación *lineal*, pero *negativa*. Finalmente, el gráfico inferior derecho muestra una asociación particular en la medida en que fácilmente se aprecia que no es lineal, sino *curvilinea*. También ocurre que las altas esperanzas de vida al nacer se encuentran en países con alta renta y las bajas en los de bajo PNB, pero se aprecia que entre los de bajo nivel económico un ligero ascenso del producto produce un considerable aumento de la esperanza de vida, mientras que, entre los países de alto nivel económico, el enriquecimiento en similares cantidades conlleva muchos menores progresos en el número de años que la gente vive.

**GRÁFICO 6.23. Gráficos de distintos tipos de relaciones**

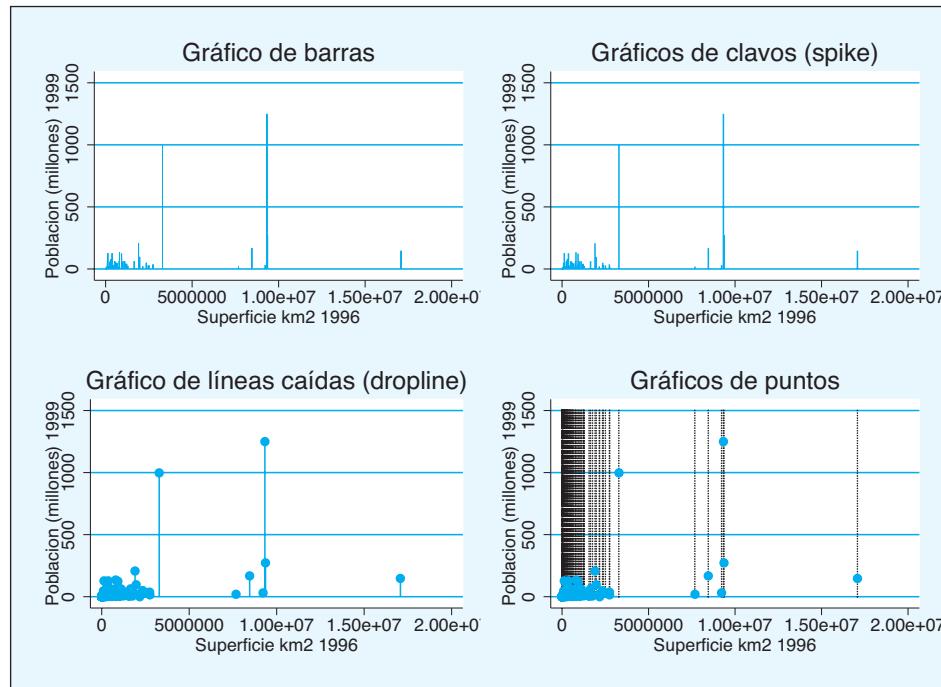


En este apartado se consideran también cuatro gráficos similares a las nubes de puntos, de los que se diferencian en que poseen una guía que une el punto (representado o no) con el eje de abscisas. Por tanto, aunque algunos de ellos se hayan visto en el apartado de gráficos unidimensionales, en el fondo son muy distintos, pues en lugar de representar una variable cualitativa con su frecuencia o con otro estadístico de otra variable, se están representando los valores de dos variables cuantitativas, la mayor parte de las veces siendo la independiente (expresada en el eje horizontal) el tiempo.

Estas cuatro modalidades son barras (*bar*), en el caso de que lo que une al punto sea una columna; líneas con o sin puntos (*dropline* o *spike*), cuando en lugar de una columna se une el punto representado con los ejes mediante una línea recta y puntos guiados, y puntos (*dot*), en el caso de que se quiera que quede como guía todo el eje vertical (incluido el espacio superior al punto). Un mismo ejemplo al que se le aplican las cuatro modalidades muestra la similitud de todos estos tipos de gráficos.

```
graph twoway bar pob supkm2, name(G24a, replace) title("Barras")
graph twoway dropline pob supkm2, name(G24c, replace) title("Clavos")
graph twoway spike pob supkm2, name(G24b, replace) title ("Líneas caídas")
graph twoway dot pob supkm2, name(G24d, replace), title("Puntos")
graph combine G24a G24b G24c G24d, name(G24)
```

**GRÁFICO 6.24. Otros gráficos bidimensionales**



En estas cuatro representaciones de más de 200 países aparece la variable extensión territorial en el eje horizontal, y la altura de los puntos, líneas o barras indica el tamaño de sus respectivas poblaciones. Como en el fondo son iguales, en todos ellos destacan del resto los siete países mayores del planeta. En sentido decreciente, son Rusia, Estados Unidos, China, Canadá, Brasil,

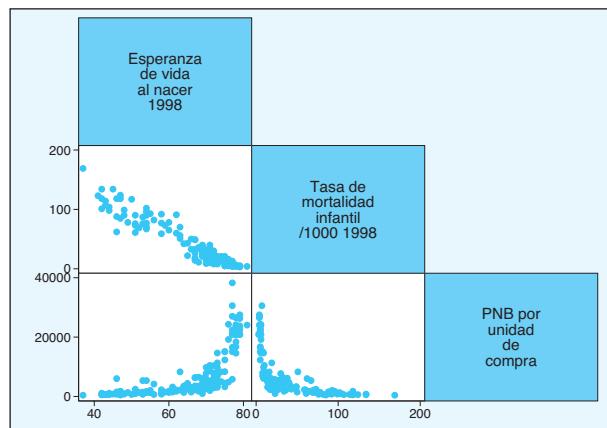
Australia y la India. Este último tiene una población aproximada de 1.000 millones de habitantes, sólo superados por los 1.250 de China. Los cinco restantes países de gran extensión tienen una población mucho más reducida, por debajo de los 300 millones de habitantes. Y, entre los países pequeños, destaca Indonesia por su población por encima de los 200 millones de habitantes.

### 6.3.2. Gráficos de matriz

Para un análisis exploratorio del conjunto de relaciones entre más de dos variables, el programa Stata dispone de la modalidad *matrix* en su programa de gráficos<sup>11</sup>. Esta produce tantos gráficos de dispersión como pares de contrastes se puedan realizar entre una serie de variables. De este modo, si se escriben tres variables, tres son los posibles gráficos no redundantes que se generan en el gráfico 6.25.

```
graph matrix tmi pnbppa evn, half name(G25)
```

**GRÁFICO 6.25. Gráficos de matriz**



Cuando se dispone de una variable dependiente y un conjunto de variables independientes, lo más adecuado es ubicar la primera al final de la lista. De este modo, en la última fila de la matriz de gráficos se dispone del conjunto de cruces de las variables independientes (ubicadas en el eje horizontal

<sup>11</sup> Paradójicamente, aunque represente relaciones bivariadas entre variables, este gráfico no es tratado como bidimensional por Stata. La razón es sencilla, por su propia naturaleza de inclusión de múltiples gráficos bivariados no puede incrustarse con otros gráficos sencillos. Operativamente, la instrucción *graph twoway* sólo es aplicable a gráficos que puedan integrarse entre ellos. Sin embargo, este tipo de gráfico se incluye en este apartado por su alta similitud de contenido y uso con los gráficos de dispersión.

de abscisas) con la variable dependiente (situada en el eje vertical). La opción *half*, utilizada en el reciente ejemplo, sirve para que sólo se reproduzcan los gráficos de la parte inferior de la matriz, pues el resto es redundante.

### 6.3.3. Gráficos de líneas

Los casos dibujados en una nube de puntos pueden conectarse entre sí siguiendo distintas reglas a fin de que mejore la apreciación de la pauta que siguen los puntos o a fin de que se dé una sensación de continuidad en los datos, como puede ser en el caso de datos que representen funciones o en el caso de representación de series temporales.

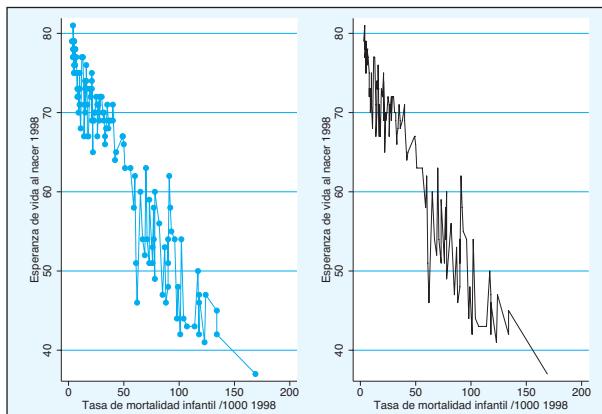
Existen dos instrucciones que permiten desarrollar este tipo de gráficos. La primera (*connected*) dibuja los puntos y los conecta. La segunda (*line*) tan sólo hace la conexión y deja invisibles los puntos. A ambas es recomendable acompañarlas con la opción *sort*, que ordena los casos en función de la variable independiente (en el eje horizontal) para que la conexión se produzca entre casos contiguos y no se produzcan cruces entre las líneas dibujadas.

A continuación se exponen las dos instrucciones que generan los gráficos compuestos representados en la próxima figura:

```
graph twoway connected evn tmi, sort name(G26a, replace)
graph twoway line evn tmi, sort name(G26b, replace)
graph combine G26a G26b, name(G26, replace)
```

Como puede apreciarse, las diferencias entre ambos gráficos están en la presencia o ausencia de los puntos que representan a los casos:

**GRÁFICO 6.26. Gráficos de líneas**



### 6.3.4. Gráficos de área

Son una modalidad de los anteriores, puesto que lo único que los diferencia es que aparece rellena el área existente entre la línea formada por la conexión de los puntos y el eje horizontal. Son idóneos cuando se quiere representar frecuencias o también cuando se representan cantidades, puesto que proporcionan al lector una considerable sensación de volumen.

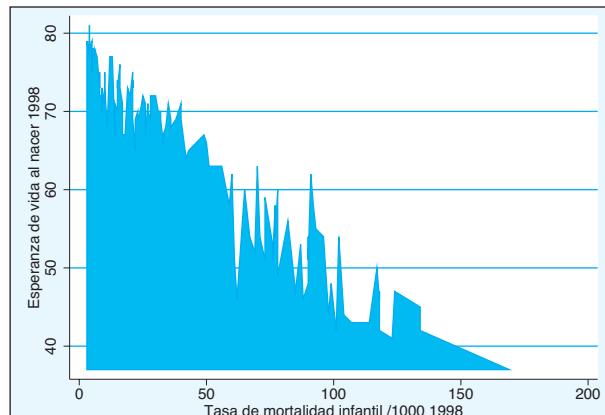
Además de la opción *sort*, siempre recomendable en este tipo de gráficos, tiene otras dos importantes: la primera es *horizontal*, que permite cambiar la orientación del gráfico, poniendo en el eje vertical la segunda variable (la independiente) y en el eje horizontal la primera (la dependiente); la segunda es *base(#)*, que permite indicarle al gráfico el punto de arranque del área.

Como ejemplo de uso, se utilizan los mismos datos de los gráficos de línea para que se aprecien sus semejanzas.

```
graph twoway area evn tmi, sort name(G27, replace)
```

El gráfico de área presenta el siguiente aspecto:

**GRÁFICO 6.27. Gráfico de área**



### 6.3.5. Gráficos de ajuste

En lugar de dibujar líneas quebradas que unan todos los puntos de una distribución bivariada, se puede optar por trazar una línea —recta o curva— que trate de pasar lo más cerca posible de los puntos con el fin

de dar cuenta simplificada de la realidad, esto es, generar un modelo de relación entre las variables que explique de modo simple cómo una variable cambia sus valores, en la medida en que otra variable modifica los suyos.

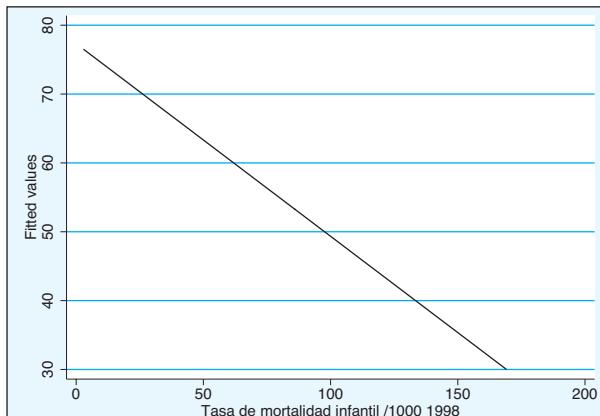
El ajuste más simple y utilizado, aunque no el único, como se verá más adelante, es la línea recta. Y el criterio más común que se utiliza (véase el primer capítulo dedicado a la regresión) es el de mínimos cuadrados, esto es, se traza la recta cuya distancia cuadrática respecto a los puntos empíricos reales sea mínima<sup>12</sup>.

A pesar de la aparente complicación del proceso de ajuste de la recta, mediante el programa gráfico de Stata, el trazado de esta línea es extremadamente simple. Basta con pedir un gráfico bidimensional con la modalidad *lfit* y aportar las variables que han de ubicarse, respectivamente, en el eje vertical y horizontal. Así, escribiendo la siguiente instrucción...

```
graph twoway lfit evn tmi, name(G28, replace)
```

... en lugar de dibujarse los puntos empíricos, se traza la línea que mejor ajusta la distancia cuadrática de estos a la recta. Es preciso notar que en el eje vertical aparecen los valores ajustados de la esperanza de vida al nacer, en lugar de la variable propiamente dicha.

**GRÁFICO 6.28. Gráfico de ajuste lineal**



Mucho más útil que dibujar sólo la recta ajustada es representar juntas con ella los puntos que representan los valores medidos de ambas variables. Como se ha dicho al inicio de los gráficos bidimensionales, la

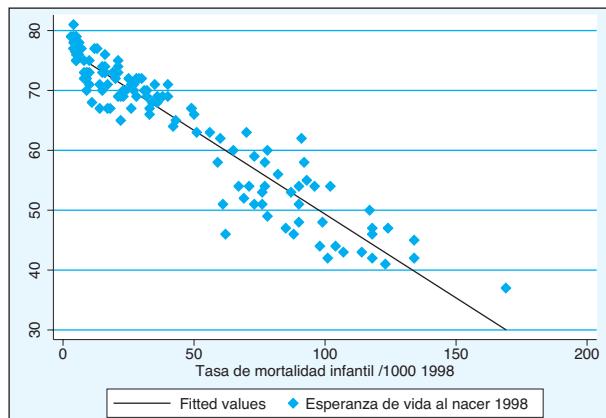
<sup>12</sup> El método de ajuste de líneas es contemplado con más detenimiento en el capítulo destinado a la regresión.

instrucción *graph twoway* posee la facultad de dibujar en los mismos ejes varios gráficos al mismo tiempo con una gran facilidad; basta con expresar los distintos gráficos entre paréntesis o separarlos por dos líneas verticales (||). Por ello, las dos siguientes instrucciones dan el mismo resultado:

```
graph twoway (lfit evn tmi) (scatter evn tmi)
graph twoway lfit evn tmi || scatter evn tmi, name(G29, replace)
```

De esta forma, además de los puntos que representan cada uno de los casos empíricos de los que se disponen datos, aparece la línea recta que mejor ajusta los valores empíricos de la tasa de mortalidad infantil y la esperanza de vida al nacer:

**GRÁFICO 6.29. Nube de puntos y ajuste lineal**



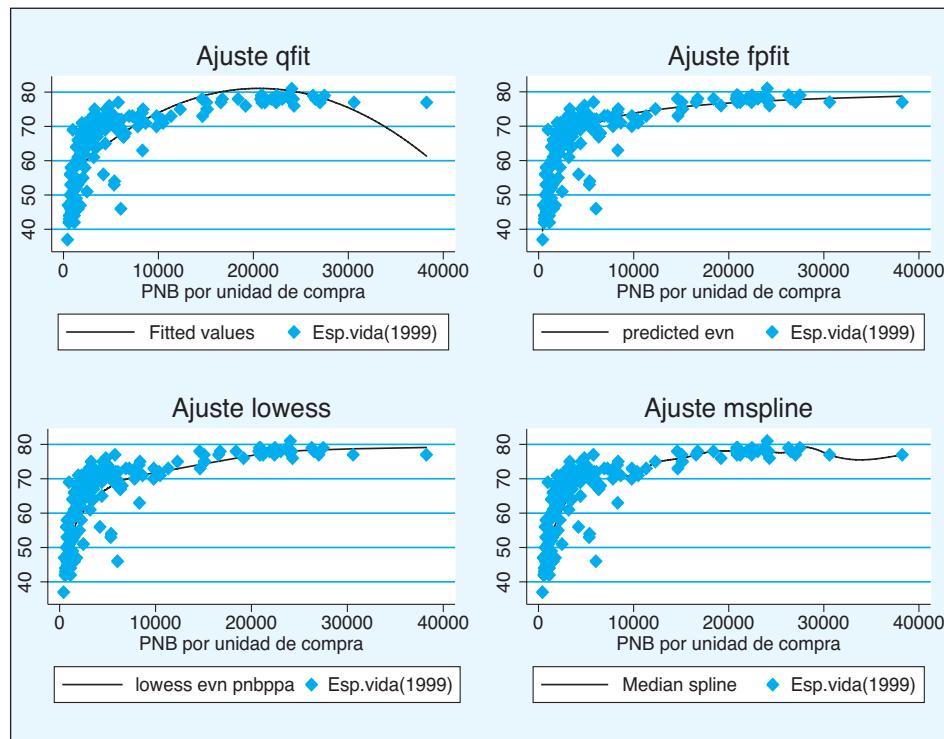
Además del ajuste lineal, la opción gráfica de Stata permite otros ajustes inmediatos. El cuadrático (*qfit*) y el polinómico (*fppfit*)<sup>13</sup>, por un lado, son ajustes en última instancia lineales. El ajuste *lowess* es un suavizado basado en regresiones ponderadas localmente de los valores  $y_i$ . Los ajustes *mband* y *mspline* dividen la distribución de la variable independiente en distintos sectores (bandas) y, a través de la mediana, en cada una de ellas construye un ajuste no suavizado, como en el primer caso, o suavizado, como en el segundo.

<sup>13</sup> Este ajuste implica la realización de una regresión fraccional polinómica en la que el programa busca las mejores potencias sobre la variable independiente para que ajuste los valores de la variable dependiente. Véase para más detalle la instrucción *fracpoly* en el manual de Stata (2009f: 399).

Mediante las cuatro instrucciones siguientes posteriormente combinadas se obtienen los cuatro gráficos de la próxima figura, donde pueden comprobarse las diferentes características de los ajustes expuestos en sus respectivos títulos:

```
twoway (qfit evn pnbppa) (scatter evn pnbppa), name(G30a) title("Ajuste qfit")
twoway (fpfit evn pnbppa) (scatter evn pnbppa), name(G30b) title("Ajuste fpfit")
twoway (lowess evn pnbppa) (scatter evn pnbppa), name(G30c) title("Ajuste lowess")
twoway (mspline evn pnbppa) (scatter evn pnbppa), name(G30d) title("Ajuste mspline")
graph combine G30a G30b G30c G30d, name(G30, replace)
```

**GRÁFICO 6.30. Gráficos de cuatro ajustes distintos**



### 6.3.6. Gráficos de rango

Son aquellos que al mismo tiempo, para cada valor de la variable independiente, representan dos puntos distintos correspondientes a dos valores de

sendas variables dependientes. Hay varias formas de presentación, pero todas ellas se caracterizan por lo que se acaba de definir.

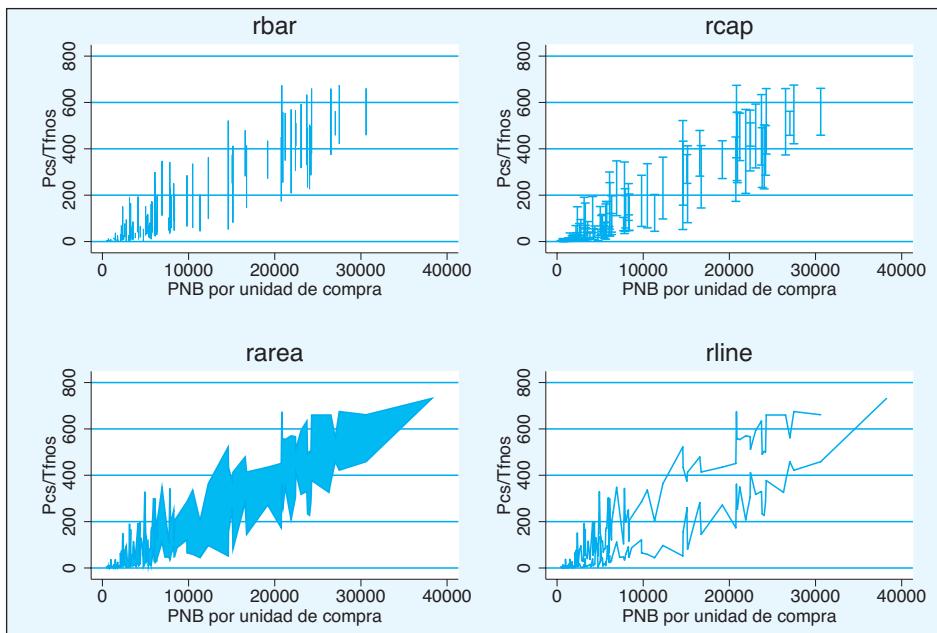
Por un lado, las dos variables representadas pueden estar unidas entre ellas, sea por barras (*rbar*), por líneas verticales (*rspike*), por líneas rematadas (*rcap* o *rcapsim*) o por áreas (*rarea*). Otra posibilidad es que se representen las dos series con dos líneas conectadas, pero paralelas entre sí, insertando o no los puntos que representan los diversos casos (*rconnected* y *rline*).

La sintaxis de este tipo de gráficos comienza con la orden *graph twoway*, después continúa con la modalidad de gráfico deseada y seguidamente han de ponerse en primer lugar las dos variables representadas en el eje vertical y, a continuación, la variable independiente, es decir, la del eje horizontal. El orden de las dos primeras es irrelevante, puesto que mediante la barra o el área se representa la distancia absoluta entre los dos valores.

A continuación se presentan, para mostrar las distintas modalidades de representación de los gráficos de rango, cuatro modelos distintos obtenidos con las siguientes instrucciones:

```
graph twoway rbar lntfno pcx1000 pnbppa, name(G31a) title("rbar")
graph twoway rcap lntfno pcx1000 pnbppa, name(G31b) title("rcap")
graph twoway rarea lntfno pcx1000 pnbppa, sort name(G31c) title("rarea")
graph twoway rline lntfno pcx1000 pnbppa, sort name(G31d) title("rline")
graph combine G31a G31b G31c G31d, name(G31, replace)
```

La combinación de estas cuatro instrucciones da lugar a los siguientes gráficos:

**GRÁFICO 6.31.** Gráficos de rangos

Las variables que definen el rango son el número de líneas telefónicas (máximo) y el número de ordenadores personales (mínimo) por mil habitantes. La variable independiente es el producto nacional bruto per cápita. Los gráficos muestran bajo diversas formas cómo los dos indicadores de desarrollo tecnológico crecen a medida que lo hace el PNB per cápita y dejan entrever que donde más divergencias se da entre teléfonos y ordenadores es en algunos países con renta per cápita media.

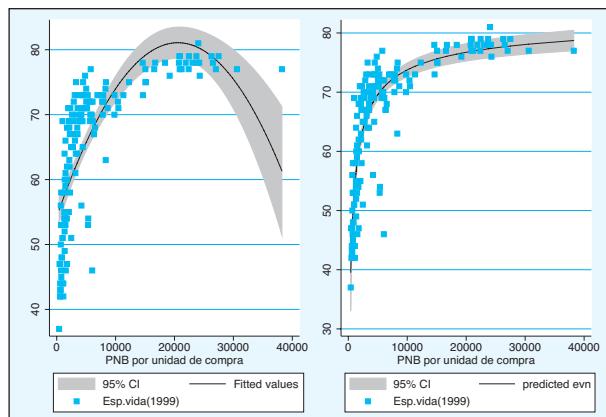
También podrían clasificarse en esta categoría aquellos gráficos que ajustan datos y dan un determinado rango de ocurrencia. Se corresponden con los gráficos *lfit*, *qfit* y *fpfit*, es decir, ajustes lineales, cuadráticos y polinómicos fraccionales, pero, en lugar de aportar una sola curva, muestran dos correspondientes a la probabilidad señalada. En estos casos, la modalidad del gráfico se indica con las palabras claves *lfitci*, *qfitci* y *fpfitci*. Además, en este tipo de gráfico son importantes las opciones *level(#)*, donde se indica el porcentaje de confianza deseado para la representación, y *stdf*, en el caso de que se desee contar con el error típico del pronóstico, en lugar del de la predicción<sup>14</sup>, o la opción *stdr*, si se desea utilizar para el cálculo de los intervalos el error típico de los residuales.

<sup>14</sup> Véase el capítulo de la regresión.

Un par de ejemplos con las opciones por omisión muestran dos gráficos con los intervalos basados en el error típico de la predicción y un 95% de confianza, salvo en el caso de que se modifique este parámetro con la instrucción *set level*.

```
graph twoway (qfitci evn pnbppa) (scatter evn pnbppa), name(G32a, replace)
graph twoway (fpfitci evn pnbppa) (scatter evn pnbppa), name (G32b, replace)
graph combine G32a G32b, name(G32, replace)
```

**GRÁFICO 6.32. Gráficos de ajustes con intervalos de confianza**



### 6.3.7. Gráficos de función

Son aquellos en los que se representa la curva resultante de aplicar una función a una variable de rango establecido (entre los valores de 0 y 1, en caso de que el usuario no lo indique en las opciones).

La sintaxis de estos gráficos es sencilla:

```
graph twoway function var_dep=f(x), opciones
```

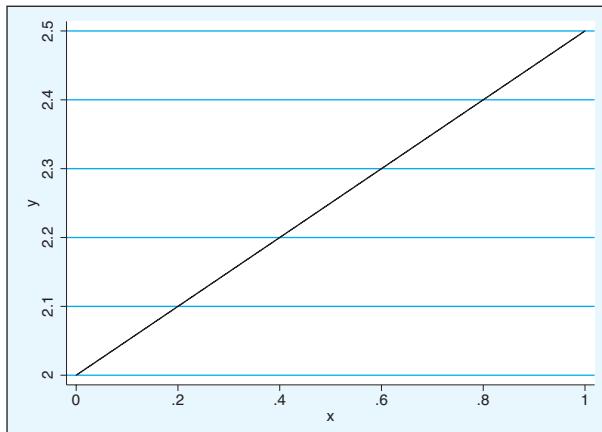
La expresión  $\text{var\_dep}=f(x)^{15}$  es la que representa a la función que se quiere representar. Así, si se desea dibujar una recta con parámetros  $a=2$  y  $b=.5$ , la instrucción siguiente genera la línea deseada.

<sup>15</sup> En este caso,  $x$  representa la variable que va a fluctuar un número determinado de veces (300 por omisión) en un rango dado (entre 0 y 1, si nada se especifica).

```
graph twoway function y=2+.5*x, name(G33, replace)
```

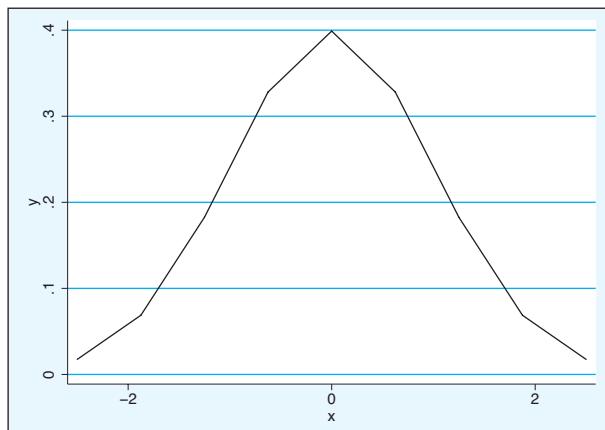
El gráfico muestra el valor en la variable  $y$  al aplicar la expresión tras el igual a 300 valores comprendidos entre el 0 y el 1.

**GRÁFICO 6.33. Gráfico de función**



El usuario, a través de las opciones, puede controlar tanto el número de estimaciones de la función como el rango de la variable  $x$ . Así, si se desea representar la función de probabilidad de la normal sólo a través de nueve valores, la instrucción necesaria es la que se expone a continuación:

```
graph twoway function y=normalden(x), range(-2.5 2.5) n(9), name(G34, replace)
```

**GRÁFICO 6.34.** Gráfico de la función normal

Como puede apreciarse, la curva normal pierde su apariencia de campana curvada por tener tan pocos puntos de referencia.

## 6.4. Componentes de los gráficos

Una vez visto cómo proceder para obtener lo básico de los distintos gráficos que Stata genera con su instrucción *graph*, se van a considerar otros elementos que, aunque auxiliares, son muy importantes para la definición final de los gráficos. Para cualquier tipo de gráficos, independientemente de la instrucción que lo genere o de las características propias de su forma, pueden distinguirse una serie de elementos complementarios, a veces considerados secundarios, pero muy importantes para la presentación adecuada. Sin pretensión de ser exhaustivos, aquí se presenta una lista de ellos<sup>16</sup>:

*Títulos:* Cumplen una doble función: por un lado, aclaran al lector qué es lo que se está representando y, por el otro, el título principal de un gráfico es un elemento esencial para exponerlo en un índice de una publicación donde el número de gráficos sea considerable. En muchas ocasiones, estos títulos deben acompañarse de un subtítulo, consistente en una línea adicional que complementa la información del primero.

<sup>16</sup> No es objeto de un manual introductorio explicar la compleja estructura de órdenes y opciones a través de las cuales se pueden introducir o modificar estos elementos. Para realizar cambios en un gráfico se sugiere el uso del editor de programas, explicado al final de este capítulo. Se recomienda, asimismo, tanto el libro de Mitchell (2008), dedicado exclusivamente a los gráficos, como la página web de la UCLA, <http://www.ats.ucla.edu/stat/stata/library/GraphExamples/default.htm>, donde se exponen una serie de modelos, con las órdenes que hay que escribir para obtenerlos.

*Ejes:* Son escalas donde se ubican los valores o las frecuencias de las variables representadas. En teoría puede haber gráficos sin ejes, como los de sectores, y los puede haber hasta con seis (tres dimensiones con dos ejes cada una de ellas), siempre y cuando no se combinen una serie de gráficos, pero lo más frecuente es que un gráfico sólo tenga uno o dos. Dentro de los ejes pueden considerarse las marcas y las cuadrículas. Las primeras son pequeños signos, generalmente perpendiculares al eje, que especifican dónde se encuentra un determinado valor. Las cuadrículas, en cambio, son líneas que tienen su origen en un determinado eje y llegan hasta el otro extremo del gráfico con el fin de poder ubicar la posición de un determinado elemento dentro del conjunto.

*Elementos:* Son cada uno de los componentes esenciales de un gráfico propiamente dicho, que representan bien un caso o un grupo de casos, bien un valor o conjunto de valores. Son elementos, por ejemplo, los sectores de un gráfico circular, los rectángulos que forman un diagrama de barras, los puntos de una nube de puntos o las líneas que representan una regresión. En general, aun teniendo en cuenta las excepciones de las distintas variedades, los elementos pueden diferenciarse de cuatro maneras distintas. En primer lugar, la *forma*. De este modo, para distinguir distintos tipos de casos, puede utilizarse un círculo, un cuadrado o cualquier otra forma similar, según se quieran expresar los de un tipo u otro. En segundo lugar, el *tamaño* también puede diferenciar unos elementos de otros, aunque en la mayor parte de los gráficos el tamaño suele emplearse para distinguir la frecuencia de unos determinados casos o valores. En tercer lugar, la *posición*, pues en muchas ocasiones un valor no está representado por el tamaño del elemento, sino por lo cercano o alejado que esté del punto de origen de una escala. En cuarto lugar, los gráficos pueden utilizar el *color* para diferenciar los elementos. Así, un valor puede quedar representado con un color y el resto de los valores con otros. Y, finalmente, de modo alternativo o complementario al color, se pueden utilizar distintas *tramas* al dibujar cualquier elemento, como por ejemplo líneas continuas, discontinuas o punteadas, o barras con superficies lisas, rayadas o punteadas.

*Leyendas:* Son el repertorio de símbolos que se utilizan en un gráfico, junto al significado que estos poseen. Sirven para descifrar el significado de las formas, colores o tramas que se emplean para la representación de los datos y son voluntarias aunque altamente recomendables.

*Etiquetas:* Son los textos aclaratorios que acompañan los elementos esenciales del gráfico. Pueden ser textuales para identificar al objeto al que acompañan, o bien numéricas, en cuyo caso indican el valor concreto que posee un determinado símbolo o posición del gráfico.

*Marcos:* Son rectángulos que envuelven al gráfico o a partes de este, por un motivo principalmente estético.

*Notas:* Son textos, normalmente ubicados en la parte inferior del gráfico, que sirven para aclarar, resaltar o precisar algunas de las características peculiares de los datos, especialmente la fuente de donde proceden.

## 6.5. Esquemas

Dada la complejidad de las opciones y subopciones de los gráficos en Stata, esta herramienta estadística ha querido simplificar al usuario la producción de gráficos a través de los esquemas. Los esquemas son conjuntos de opciones con los que los gráficos son representados en la pantalla. Ejemplo de las especificaciones que puede contener un esquema son el tipo y tamaño de letra, los colores de fondo y de los cuadros, los sucesivos colores que incorporan los elementos (sectores, barras, líneas...) de los distintos tipos de gráficos, el grosor y la textura de las líneas, la presencia —y en su caso la forma— o ausencia de marcas, ejes, rejillas, etc. Por omisión, Stata trabaja con uno de la docena de esquemas que tiene disponibles<sup>17</sup>. Para saber los nombres disponibles y cuál está activo en un determinado momento se emplean, respectivamente, las siguientes dos instrucciones:

```
graph query, schemes
query graphics
```

El resultado de ella puede variar de ordenador a ordenador, según los esquemas en él incorporados a través de Internet o de la propia construcción. Un ejemplo de listado es el siguiente:

### ILUSTRACIÓN 6.3. Listado de esquemas gráficos

```
Available schemes are
economist      see help scheme_economist
sicolor         see help scheme_sicolor
simanual        see help scheme_simanual
simono          see help scheme_simono
slrcolor        see help scheme_slrcolor
s2color          see help scheme_s2color
s2colororg       see help scheme_s2colororg
s2manual         see help scheme_s2manual
s2mono           see help scheme_s2mono
sj               see help scheme_sj

Graphics settings
    set graphics      on
    set scheme        sj
    set printcolor    automatic   may be automatic, asis, gs1, gs2, gs3
    set copycolor     automatic   may be automatic, asis, gs1, gs2, gs3
```

En la primera parte de este recuadro aparecen todos los esquemas disponibles en la máquina. En la segunda parte se expresa que el esquema

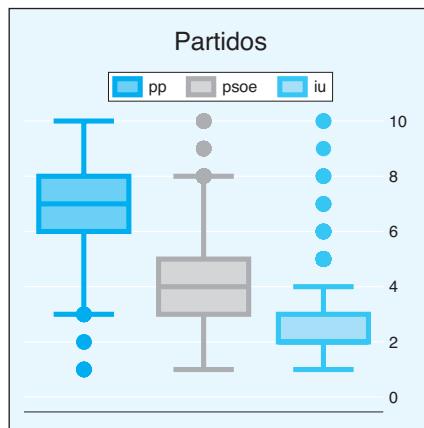
<sup>17</sup> El usuario puede importar nuevos esquemas por Internet, y con un poco de destreza incluso puede construir nuevos esquemas a partir de los existentes, que residen en los directorios de los ficheros .ado.

puesto por defecto (*set scheme*) es el *sj*, que corresponde a los que se han empleado hasta el momento, que es el utilizado en las publicaciones del *Stata Journal*. Como puede apreciarse, además de este y del propio del semanario *The Economist*, aparecen dos esquemas en blanco y negro (*s1mono* y *s2mono*), dos estilos en color (*s1color* y *s2color*) y dos estilos manuales (*s1manual* y *s2manual*).

Para cambiar el esquema del próximo gráfico hay que introducir la instrucción *set scheme nombre\_del estilo*. Haciéndolo así, el gráfico 6.19 se convierte en este otro con el esquema de *The Economist*:

```
use ejemplo6, clear
set scheme economist
graph box ideopp-ideoiu, title("Partidos", position(12)) name(G35, replace)
```

**GRÁFICO 6.35. Gráfico de caja con esquema personalizado**



### 6.5.1. Gráficos con menús

Dado que controlar las múltiples opciones que ofrecen las posibilidades gráficas de Stata es complicado y requiere un conocimiento pormenorizado de opciones y subopciones, resulta de gran utilidad recurrir a los menús que se ofrecen a partir de la versión 8 de este programa estadístico. No obstante, hay que reparar en que —salvo en los gráficos de sectores para los que se dispone de una posibilidad inmediata ubicada en el menú de los gráficos fáciles—, para la representación simple de variables categóricas, no basta con poner esta variable en la casilla correspondiente. Como se vio en el apartado 6.2, para la construcción de

gráficos unidimensionales de variables, hay que generar una nueva con el peso de cada caso, que es la que aparece en el eje de frecuencias, mientras que la que genera los distintos valores de la variable aparece bajo la opción *over*.

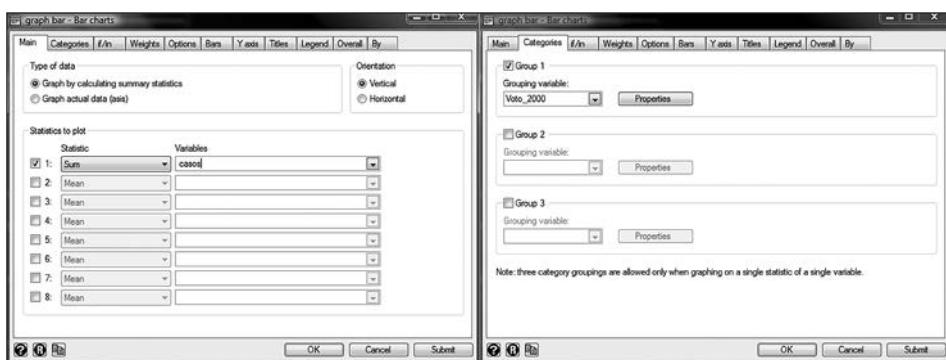
Un ejemplo con un gráfico de barras del sexo mostrado a través de los diversos menús ayudará a realizar la representación de las variables cualitativas.

```
use panel6, clear
tabulate sexo
generate casos=100/r(N)
db graph bar
```

Para no complicar excesivamente el ejemplo, se recurre a la modalidad de gráficos de barras (*Graphics/Bar chart*). Una vez que se han seleccionado desde el menú estas dos opciones, aparece un cuadro de diálogo con once pestañas (*Main*, para exponer las variables del gráfico y su tratamiento; *Categories*, para incluir las variables que marcan los distintos segmentos del gráfico; *if/in*, para seleccionar los casos que se desean exponer en el gráfico; *Weights*, para adjudicar pesos; *Options*, para especificar alguna modalidad del gráfico [barras apiladas, tratamiento de variables]; *Bars*, para controlar el formato de las barras; *Y axis*, para manejar la apariencia de la escala vertical; *Titles*, para poner títulos, subtítulos, aclaraciones y notas adicionales al gráfico; *Legend*, para solicitar una leyenda; *Overall* para dar nombre al gráfico y especificar su tamaño, y *By* con el propósito de sacar varios gráficos en función de una segunda variable).

De ellas las dos primeras son las más importantes para el gráfico deseado y han de ser dispuestas del modo siguiente:

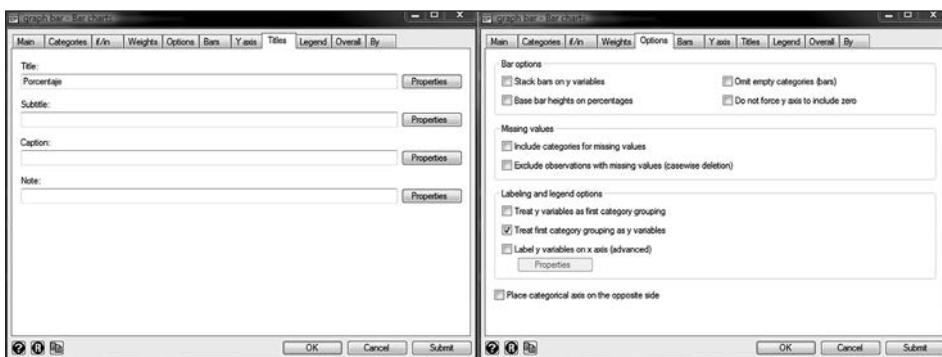
#### ILUSTRACIÓN 6.4. Menús del gráfico de barras



En el cuadro de diálogo de la izquierda aparece en *Statistic* la modalidad *Sum* (*Count nonmissing*, en el caso de que se deseen frecuencias absolutas y no relativas) y en *Variables*, se ha insertado la variable instrumental (*casos*) que se crea a fin de que aparezcan porcentajes o proporciones en lugar de sumas (véase el apartado 6.2.2). En el de la derecha, en la ventana de las variables de cruce, es donde aparece la verdadera variable de la que se desea la representación. El nombre que posee la variable en el fichero es el que aparece en la primera ventanilla y en este menú pueden cambiarse sus etiquetas pulsando en el botón *Properties*.

Con estas dos instrucciones bastaría para confeccionar el gráfico deseado. No obstante, puede ser mejorado sólo con dos detalles. En primer lugar, dando un título distinto al eje vertical que representa en este caso los porcentajes. Esto se logra especificándolo en la casilla *Title* de la pestaña *Y axis*. Y, en segundo lugar, haciendo que el programa trate la variable de cruce como variable principal. Para ello, en la última pestaña, puede marcarse la casilla *Treat first category group as y-variables*. De este modo, cada barra, que representa cada uno de los valores de la variable, será dibujada con un color o tonalidad diferente.

#### ILUSTRACIÓN 6.5. Menús del gráfico de barras (*continuación*)



Una característica interesante que incorporan los menús son los tres botones situados abajo a la izquierda representados con un ícono. El signo de interrogación abre el fichero de ayuda de la instrucción en concreto en cuyo menú se encuentre el usuario. La R lo que hace es despejar todos los campos de los menús para empezar a dar órdenes desde el principio. Por último, el símbolo que representa una cuartilla escrita copia en el portapapeles la sintaxis de la instrucción que se está solicitando a través del menú. Es muy útil para quienes no se contentan con hacer todo mediante menú y desean guardar el resultado para que en futuras ocasiones un gráfico determinado sepa solicitarlo mediante programa.

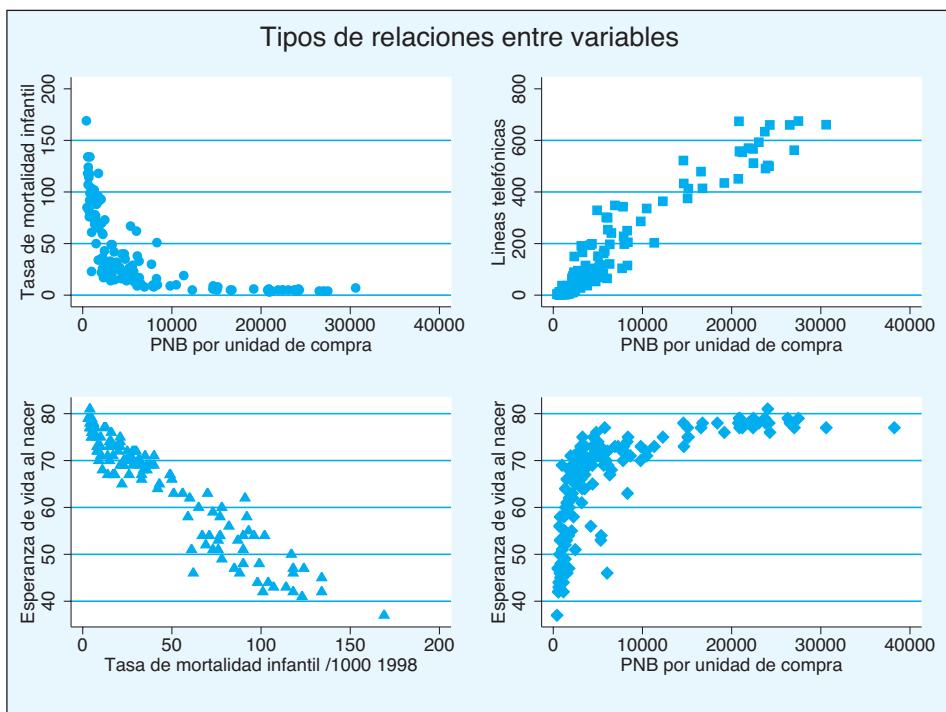
## 6.6. El editor de gráficos

Desde la versión 10, Stata ha incorporado un editor de gráficos. La idea es que, una vez producido un gráfico, el creador pueda transformar el producto sin necesidad de escribir complejas instrucciones.

Su funcionamiento es a la vez intuitivo y potente. Es muy fácil de utilizar, sobre todo, mediante menús contextuales y, gracias a ellos, los componentes del gráfico que pueden cambiarse son prácticamente todos.

Para empezar a emplearlo hay que ir a la ventana gráfica y comenzar el estado de edición. Ello se logra mediante menú (*File/Start Graph Editor*) o mediante el sexto ícono, que tiene el dibujo de un gráfico de barras.

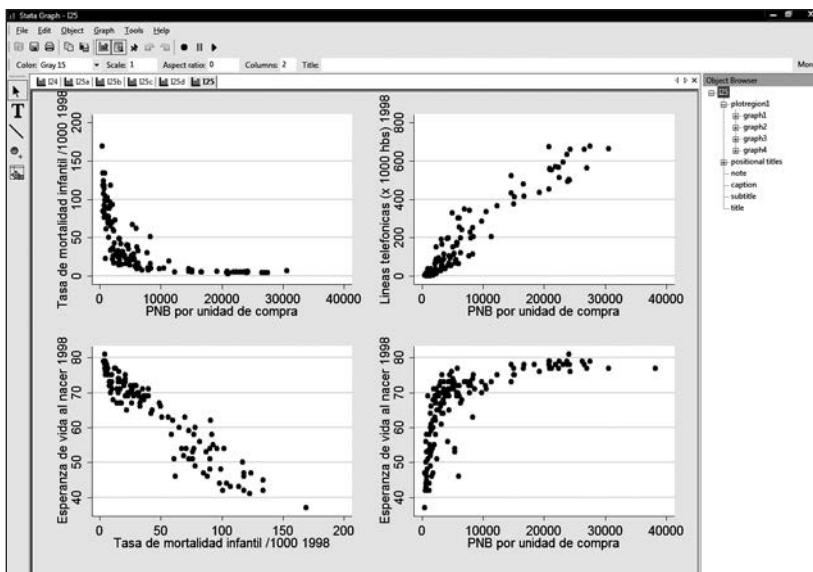
**GRÁFICO 6.36. Resultado de la edición de un gráfico Stata**



Una vez en modo edición, aparecen en la columna izquierda nuevos iconos que representan la flecha para señalar, una letra para escribir texto, una línea para dibujar rectas, un círculo para dibujar puntos y una

reja para editar las secciones. El modo más frecuente para la edición es el primero, puesto que es el que nos permite seleccionar objetos del gráfico que se desean transformar. Los posibles objetos de selección son títulos, marcos, leyendas, ejes, marcas, etiquetas, puntos, líneas, barras, cajas... Al señalar cualquiera de estos, aparece debajo de los iconos horizontales un nuevo menú que indica los aspectos más transformables de los objetos seleccionados. Así, de este modo, si se señala el título de un gráfico, aparecerá el color de la letra, el tamaño, el margen y el contenido del texto. Pero también al final de la barra aparece la palabra *More...* para que puedan cambiarse otros elementos no tan centrales de aquello que se quiere transformar. Pulsar sobre ellos, abre el menú de propiedades de un objeto, donde se dispone en distintas pestañas, todo aquello modificable. Las características de toda caja de texto, títulos incluidos, contienen las pestañas de texto (*text*), caja (*box*), formato (*format*) y avanzado (*advanced*), de tal forma que pueden cambiarse una veintena de aspectos diferentes de los títulos.

#### ILUSTRACIÓN 6.6. Pantalla del editor de gráficos con el explorador de objetos



La estrategia a seguir para editar un gráfico debe ser la siguiente: buscar el elemento que se desea cambiar, señalarlo con el cursor, a partir de lo cual se marca automáticamente en rojo, averiguar si lo que se desea cambiar está contenido en el nuevo menú horizontal que surge debajo de la barra de iconos y, si no lo está, pulsar el botón derecho del ratón, porque gene-

ralmente en la última línea del menú contextual emergente se encuentra el acceso a las propiedades del objeto.

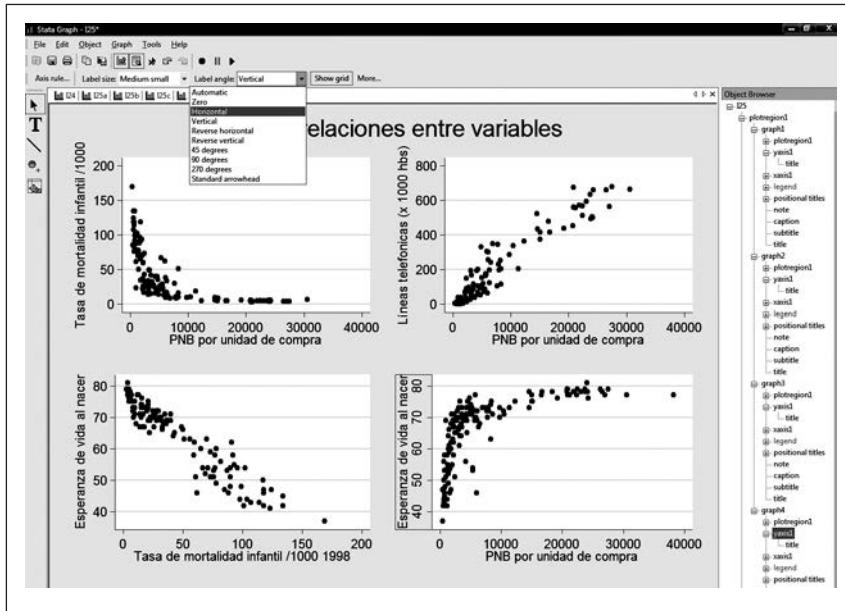
Como son centenares las posibilidades de edición, en adelante sólo se pondrá un ejemplo de edición, basado en el gráfico 6.23 para obtener el resultado del gráfico 6.36.

La primera operación será la introducción del título. En el menú *Graph* del editor se encuentran las cinco propiedades principales del objeto gráfico: títulos (*Title*), tamaño (*Graph Size*), aspecto (*Aspect ratio*), región (*Graph region properties*) y combinación (*By/combine organization*). Mediante la primera se obtiene un cuadro de diálogo, ubicado en la pestaña *Titles*, con cuatro casillas donde escribir los textos del título, subtítulo, pie (*caption*) y nota. Se escribe el título en la primera y se pulsa su botón de propiedades, donde pueden modificarse la posición, el tipo y tamaño de letra y otra docena de aspectos. Aquí interesa cambiar el tamaño (*Size*) y la justificación (*Justification* en *Format*). El primero se pone en mediano y la segunda se centra.

La segunda modificación se realizará sobre las etiquetas del eje Y, ya que salieron muy largas y llegan a alcanzar el título. Para mejorar la presentación se eliminará el año de recogida de la información de la variable presente en la etiqueta. Resulta conveniente hacer aparecer el explorador de objetos. Ello se logra pulsando el ícono que está al lado de aquel usado para entrar en el editor, o bien, si se prefiere, en el menú *Tools/show object browser*. Esta herramienta permite seleccionar con más precisión un determinado elemento del gráfico que se desea modificar y conocer mejor la estructura que tienen los objetos. Así, el título del eje vertical se encuentra en la primera y única región del gráfico (*plotregion1*), en el primer gráfico de los cuatro (*graph1*), en el primer eje vertical (*yaxis1*). Una vez seleccionado el objeto *title*, con doble clic se abre su cuadro de propiedades, texto incluido, donde puede eliminarse el año. Cuando se haga, puede pulsarse el botón *Apply*, en lugar de *OK*, pues de esta forma, si no gusta el resultado o se desean cambiar más propiedades, no es necesario volver a abrir el cuadro. Operación similar tendría que hacerse con las tres etiquetas restantes.

El tercer cambio que se va a realizar es cambiar la orientación de las etiquetas del mismo eje vertical. Para ello se selecciona el eje que se quiere cambiar, se accede a sus propiedades y, bajo las propiedades globales (*Axis properties*), se eligen las etiquetas (*Label properties*). Tras ello aparece, entre otras características modificables, el ángulo. Poniéndolo en horizontal, se obtiene el efecto deseado.

### ILUSTRACIÓN 6.7. Uso del editor para modificar la orientación de las etiquetas



También se puede cambiar el rango de la escala de lo representado. En los gráficos tercero y cuarto del gráfico 6.23, la esperanza de vida al nacer se representaba a partir de los 40. Si se quiere que el gráfico comience la escala en 0, puede indicarse una regla al eje (*Axis rule*). Entre las posibles opciones, puede ser la más idónea la de *Range/Delta*, en la que hay que especificar el valor mínimo (0), el máximo (80) y el incremento (*Delta*), 20 en este caso. Para poder comparar mejor, conviene que esta operación se realice de modo igual en los dos gráficos inferiores.

Finalmente, se va a cambiar la forma de los puntos que representan los casos. Estos se encuentran en el núcleo del gráfico (*Plot*). La operación es similar: se seleccionan, se accede a sus propiedades y se transforma lo deseado. En este caso, el símbolo (*Symbol*), pudiéndose elegir, entre otros, el punto, el cuadrado, el círculo, el triángulo, etc. Junto con la forma, también pueden alterarse en el mismo menú el tamaño (*Size*) y el color de los puntos, así como el ancho (*Outline width*) y el color de los bordes o perímetro del símbolo.

Una vez realizadas todas las operaciones de transformación, se sale del editor mediante el menú (*File/Stop Graph Editor*), el ícono de la barra de herramientas o el menú contextual obtenido en cualquier zona del gráfico. Siempre que se opta por interrumpir la edición, pregunta que si se quiere grabar (en disco) el gráfico. Sin embargo, caso de que se diga que no, los cambios siguen presentes en la memoria del ordenador. Para volver a obte-

ner el gráfico inicial, habría que repetir la instrucción gráfica, introduciendo la orden de nuevo o solicitándolo mediante menú.

## 6.7. Ejercicios

1. Utiliza el fichero cis2794 del barómetro de marzo de 2009 del CIS y representa en tres gráficos diferentes el sexo, la edad y los estudios alcanzados por el entrevistado.
- 2 Con el mismo fichero cis2794 haz un diagrama de barras con los usos que la gente hace de Internet (P.27C1-P.27C10). Finalmente, con el editor, mejora el gráfico para una correcta presentación. (Sugerencia: cambia las etiquetas de la leyenda en *legend/key región/label(#)*).
- 3 Emplea ahora la base de datos mundial (mundo2005). Haz sendos gráficos de cajas con las variables *esperanza de vida al nacer* y *tasa bruta de natalidad* y combinalos en un solo gráfico. Comenta la diferente distribución de ambas. ¿Por qué no aparece ningún punto en las extremidades de ambos gráficos? Cámbiales el aspecto aplicándole un esquema distinto del que tengas por defecto.
- 4 Utilizando la misma base de datos de países representa una nube de puntos con las variables *teléfonos por mil habitantes* y *renta nacional bruta per cápita* en unidades de poder adquisitivo. Dibuja sobre la misma representación un ajuste lineal y otro cuadrático. ¿Cuál de ellos parece ajustarse mejor a los datos?

# 7

## La prueba estadística y las comparaciones

Generalmente, en estadística se trabaja con muestras, y gran parte del propósito de los cálculos de estadísticos es comprobar si con los datos disponibles de una fracción de la población puede deducirse alguna conclusión válida. En otras palabras, los investigadores suelen emitir hipótesis relacionadas con los datos de la población y la muestra aporta pruebas de si las mismas son o no sostenibles.

En este capítulo se van a abordar las pruebas estadísticas de hipótesis más simples y utilizadas en la investigación. Son aquellas relacionadas bien con la distribución, la proporción o la media de una o dos variables. La formulación de hipótesis ha de plantearse por pares: en primer lugar, es preciso emitir una llamada hipótesis nula, en términos de igualdad, a partir de la cual se genera la distribución muestral que se derivaría en el caso de que fuera cierta, para poder obtener la probabilidad de que el dato obtenido en la muestra proceda de esa suposición, porque, en el caso de que sea improbable, la decisión más lógica sería el rechazo de tal igualdad.

El caso más simple se da cuando se dispone de una sola variable cuantitativa y se emite una hipótesis sobre el valor que ha de tener en la población. En este caso, se dice que el valor de un parámetro ( $\mu$ ) de la población, la media, en este ejemplo, ha de asumir el valor  $x$ :

$$h_0 : \mu = x \quad (7.1)$$

Más concretamente, puede enunciarse una hipótesis nula con un enunciado consistente en decir que la evaluación de un determinado líder político en la población alcanza el valor de 5 en una escala con valores entre el 0 y el 10.

Toda hipótesis nula ha de estar acompañada por su correspondiente hipótesis alternativa, aquella que se aceptaría en el caso de que no pueda mantenerse la igualdad inicial. Existen distintas modalidades alternativas. En primer lugar, puede formularse *unidireccionalmente* (sólo se rechaza la nula, si los datos muestrales son mayores o menores que el valor  $x$ ). Y, si así

fuerá, la anterior hipótesis tendría como alternativa unidireccional una de las dos desigualdades siguientes:

$$h_1 : \mu > x \quad (7.2)$$

o

$$h_1 : \mu < x \quad (7.3)$$

Obviamente, aquí la alternativa al ejemplo puesto sería que el mencionado líder alcanza una puntuación inferior (o superior) al valor central del rango de la escala utilizada.

Otra manera de plantear la hipótesis alternativa es haciéndola *bidireccional*, de tal suerte que se rechace la nula, tanto si el dato muestral se aleja significativamente por encima como si lo hace por debajo del valor hipotetizado.

$$h_1 : \mu \neq x \quad (7.4)$$

Siguiendo el ejemplo del líder, se diría que la media que le otorga la población no es igual a cinco<sup>1</sup>.

La hipótesis nula puede rechazarse, en cuyo caso ha de adoptarse la alternativa, o puede aceptarse. Si se rechaza en el caso de que fuera cierta, se cometería un error denominado de tipo I. Si, en cambio, se acepta siendo falsa, se cometería el llamado error de tipo II. Las habituales pruebas estadísticas ejecutadas sobre muestras permiten trabajar con el control del primero de estos errores, ya que sólo puede obtenerse el valor exacto que adopta el segundo tipo, en el supuesto de conocer el valor exacto del parámetro.

En este capítulo, a través de ejemplos de un sondeo preelectoral, se va a aprender tanto a formular como a tomar decisiones sobre hipótesis relacionadas con proporciones<sup>2</sup> y medias. Y se hará tanto con procedimientos

<sup>1</sup> El problema práctico para quien trabaja es determinar si la hipótesis ha de formularse unidireccionalmente o bidireccionalmente. En realidad, si nos interesa acertar con un valor puntual, la alternativa ha de ser una desigualdad. Si lo que interesa comprobar es si se supera (o no se llega) al valor en cuestión, entonces se opta por la unidireccional. En un caso electoral, por ejemplo, si interesa ver si la intención de voto va a ser del 40%, entonces se formularía bidireccionalmente, pero si lo que se desea es averiguar si se va a superar esa cantidad, entonces es preferible la alternativa unidireccional.

<sup>2</sup> En realidad, el caso de las proporciones o porcentajes es una extensión del de las medias, puesto que el promedio de una variable con valores 0 y 1 coincide con la proporción de casos que poseen el valor 1. Así, en una muestra con cuatro casos, tres de ellos casados (1) y uno soltero (0), la media sería  $\frac{3}{4}$  (de cuatro personas tres casadas), esto es, 0,75, que multiplicado por 100, muestra que el 75% del conjunto es casado.

paramétricos, condicionados a que los datos cumplan determinados requisitos; como con los llamados no paramétricos, en los que las condiciones de las distribuciones implicadas pueden ser menos rigurosas.

Se empezará con la hipótesis de una sola variable empleada con proporciones, medias y medianas. Seguidamente, se abordan tesis con dos variables, útiles para la comparación de estadísticos, procedentes de la misma población (muestras dependientes) o de poblaciones diferentes (muestras independientes).

## 7.1. Pruebas de una sola variable

### 7.1.1. Prueba paramétrica de proporciones

Supóngase que se desea predecir el voto de unas elecciones y se sostiene que un partido con más del 35% de los votos sobre el conjunto de la población obtiene en un sistema no proporcional y multipartidista como el español la mayoría parlamentaria. Por tanto, interesa probar que en la población un porcentaje mayor del señalado optará por una determinada opción política (el PP, en las ya celebradas de 2000). Como quiera que, para poder construir la distribución muestral del estadístico, la hipótesis nula siempre ha de formularse en términos de igualdad, en este caso ha de ser la siguiente:

$$h_0 : \Pi = 0,35 \quad (7.5)$$

La hipótesis alternativa en este caso ha de ser unidireccional, porque el interés está centrado sólo en un lado de la distribución. Sólo interesa saber si el mencionado partido obtiene más de la cantidad antes enunciada. Por tanto, la hipótesis alternativa ha de ser expresada de este modo:

$$h_1 : \Pi > 0,35 \quad (7.6)$$

Antes de proceder a la ejecución del programa propio de la prueba de hipótesis, es necesario realizar ciertos ajustes a la variable con la que se está trabajando. Pues, inicialmente, en el cuestionario se trata de una variable nominal con muchos valores, algunos de los cuales no deben ser tenidos en cuenta.

**ILUSTRACIÓN 7.1. Pregunta sobre intención de voto en el estudio del CIS  
número 2384 (2000)**

<b>P.13</b>	Suponiendo que mañana se celebrasen elecciones generales, es decir, al Parlamento español, ¿a qué partido o coalición votaría Ud.?				
- IU .....	01	- UV .....	12	- PR .....	23
- PP .....	02	- IC-V .....	13	- BNV .....	24
- PSOE.....	03	- GIL .....	14	- PSPC.....	25
- EA.....	04	- PAR.....	15	- Otro.....	49
- PNV.....	06	- CHA.....	16	- En blanco .....	96
- CiU .....	07	- PSM-EN .....	17	- No votaría.....	97
- ERC.....	08	- UM.....	18	- No sabe.....	98
- BNG.....	09	- UPL .....	20	- N.C.....	99
- PA.....	10	- TC-PNC .....	21		
- CC .....	11	- CDN .....	22		

En esta pregunta el valor correspondiente al PP, partido sobre el que se va a comprobar la hipótesis es el 2, por ello se debe generar una variable ficticia con valores 0/1, sobre todos aquellos que supuestamente van a votar. Por ello, se consideran como datos perdidos los valores 97 y 99 de la variable P13, y con el resto se construye la nueva dicotomizada, mediante las siguientes instrucciones:

```
use panel7, clear
generate intpp=(intvoto==2) if (intvoto <97 | intvoto==98)
label var intpp "Intención de voto al PP"
```

Para comprobar el resultado de las instrucciones anteriores es útil una tabla que cruce la antigua con la nueva variable:

```
tabulate intvoto intpp, missing
```

En la tabla resultante (ilustración 7.2) se puede ver cómo sólo tienen el valor 1 en la nueva variable aquellos que tenían 2 en la original y, además, sólo son considerados casos no válidos en la nueva variable 4.616 individuos de la muestra que no contestan a la pregunta 13 o que dicen que lo más probable es que no voten en las próximas elecciones.

**ILUSTRACIÓN 7.2. Tabla de distribución de frecuencias del voto**

intvoto	Intención de voto al PP			Total
	0	1	.	
iu	919	0	0	919
pp	0	7350	0	7350
psoe	4437	0	0	4437
ea	58	0	0	58
eh	43	0	0	43
pnv	310	0	0	310
ciu	463	0	0	463
erc	108	0	0	108
bng	199	0	0	199
pa	109	0	0	109
cc	108	0	0	108
uv	27	0	0	27
ic/v	83	0	0	83
gil	37	0	0	37
par	27	0	0	27
cha	46	0	0	46
psm/en	14	0	0	14
um	6	0	0	6
upl	14	0	0	14
tc/pnc	11	0	0	11
cdn	5	0	0	5
pr	5	0	0	5
bvn	18	0	0	18
otros partidos	191	0	0	191
en blanco	674	0	0	674
no votaría	0	0	1917	1917
no sabe todavía	4162	0	0	4162
n.c.	0	0	2699	2699
Total	12074	7350	4616	24040

A partir de este punto, ya se puede realizar la prueba de significación en relación con las hipótesis nula y alternativa. Para ello basta indicar la primera precedida de la orden *prtest*.

```
prtest intpp==.35
```

El resultado muestra, además del test propiamente dicho, los siguientes estadísticos de la muestra: número de casos, media, desviación típica, error típico e intervalos con un nivel de confianza del 95%, por defecto.

### ILUSTRACIÓN 7.3. Prueba de una proporción en una muestra

One-sample test of proportion			intpp: Number of obs = 19424
Variable	Mean	Std. Err.	[95% Conf. Interval]
intpp	.3783979	.0034799	.3715775 .3852183
p = proportion(intpp)			z = 8.2978
Ho: p = 0.35			
Ha: p < 0.35		Ha: p != 0.35	Ha: p > 0.35
Pr(Z < z) = 1.0000		Pr( Z  >  z ) = 0.0000	Pr(Z > z) = 0.0000

La clave de la prueba de hipótesis paramétrica está en el error típico, que se halla dividiendo la desviación típica por la raíz cuadrada del número de casos, o de manera más directa a través de la siguiente fórmula:

$$\sigma_p = \sqrt{\frac{\Pi(1 - \Pi)}{n}} \quad (7.7)$$

Ese error típico (*Std. Err.*), en este caso muy bajo (0,003) por el alto número de entrevistados, representa la desviación típica de la distribución muestral del estadístico  $y$ , por tanto, se utiliza en la construcción de los intervalos de confianza. En la salida del ejemplo, con el 37,8% de los entrevistados<sup>3</sup> que tienen intención de dar su voto al PP en la muestra, se puede pronosticar que en la población ese valor debe estar entre el 37,2% y el 38,5%, que se obtiene sumando y restando al valor de la media 1,96 veces el error típico, ya que se está ante una distribución normal y en esta el 95% de los casos se encuentra entre +1,96 y -1,96 desviaciones típicas.

Recapitulando, en la primera parte de la ilustración 7.3 se encuentra el número de casos y más abajo la media —en este caso proporción— y el error típico junto con el intervalo de confianza de la media.

La segunda parte es la correspondiente al test de hipótesis propiamente dicho. Aparece tanto la hipótesis nula, enunciada en la instrucción, como las tres posibles alternativas. El valor de  $z$  (la media real menos la media de la hipótesis dividida por el error típico) es idéntico en los tres supuestos, esto es 8,3.

$$t = \frac{p - \Pi}{\sigma_p} \quad (7.8)$$

<sup>3</sup> En la salida del programa, al poner el valor 1 a los que muestran su intención de voto al PP, los resultados aparecen en proporciones y no en porcentajes. Basta con multiplicar por 100 para conseguir los datos expresados en porcentajes.

Sin embargo, lo que varía es la significación. Como en este caso la alternativa es unidireccional —se ha establecido que fuera mayor que— se ha de prestar atención a la columna de la derecha. Se obtiene que, en el caso de que la hipótesis nula fuera cierta, la probabilidad de encontrar en una muestra ese valor es ínfima (menor que 0,0000), por tanto, se puede rechazar con bajo riesgo de equivocación.

### 7.1.2. Prueba paramétrica de medias

Esta misma hipótesis con una sola variable también podría funcionar con medias en lugar de proporciones, pero, en lugar de operar con proporciones, se trabaja con promedios y, como no suele disponerse de la varianza de la población, se recurre a la cuasivarianza obtenida con los datos de la muestra. Por ello, la fórmula del error típico presenta notables diferencias con el de las proporciones.

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n-1}} \quad (7.9)$$

Como a partir del error típico el proceso es similar, sólo se procederá a formular las hipótesis y a analizar los resultados. En este caso, en lugar de la variable *estimación de voto*, se va a tomar la de *probabilidad subjetiva de voto*. Y se plantea como hipótesis que la probabilidad del mismo partido es de 5, en una escala de 0 a 10.

En consecuencia, la formulación de las hipótesis nula y alternativa sería como sigue:

$$\begin{cases} h_0 : \mu = 5 \\ h_1 : \mu \neq 5 \end{cases} \quad (7.10)$$

Consecuentemente, a estas hipótesis se debe introducir la siguiente instrucción, a la que se le añade la opción *level*, para que aparezcan los intervalos con un nivel de confianza del 99%, en lugar del 95% que aparece por defecto:

```
ttest pvote==5, level(99)
```

El resultado de la instrucción se muestra en la siguiente ilustración:

### ILUSTRACIÓN 7.4. Prueba de una media en una muestra

One-sample t test						
Variable	Obs	Mean	Std. Err.	Std. Dev.	[99% Conf. Interval]	
pvotopp	20533	4.708956	.0257672	3.692276	4.642578	4.775335
mean =	mean(pvotopp)				t = -11.2951	
Ho: mean =	5				degrees of freedom =	20532
Ha: mean <	5				Ha: mean != 5	
Pr(T < t) =	0.0000				Pr( T  >  t ) = 0.0000	
					Ha: mean > 5	
					Pr(T > t) = 1.0000	

En este ejemplo se ve cómo, aun considerando un riesgo de equivocación inferior al 1% (el complementario del 99% del intervalo de confianza), la hipótesis nula ha de ser rechazada, puesto que lo más probable es que el dato de la población esté comprendido entre 4,6 y 4,8. Esto mismo se hace patente en la columna central del segundo bloque de la salida, donde se ve que con un valor de  $t$  de -11,3, su significación es tan baja que es posible el rechazo de la hipótesis nula prácticamente sin ningún error. Por tanto, puede decirse que en la población la probabilidad de votar al PP no puede ser igual a 5.

#### 7.1.3. El test de los signos

Si se desea utilizar en el caso de una muestra un test no paramétrico, es decir, que no parta de la suposición de que la variable original en la población tiene distribución normal, puede utilizarse el test de los signos, que está sustentado en la probabilidad binomial.

En este caso, la hipótesis no se refiere a la media, sino a la mediana. De este modo, la formulación de la hipótesis nula sería como sigue:

$$h_0 : \mu\varepsilon = x \quad (7.11)$$

Si la mencionada hipótesis fuera cierta, entonces la mitad de los casos de la muestra caerían por debajo del valor  $x$  y la otra mitad por encima. Para ver la distribución ha de emplearse la instrucción *tabulate*.

```
tabulate pvotopp
```

Considerando el ejemplo anterior y partiendo de la distribución de frecuencias de la variable *pvotopp* (probabilidad otorgada de voto al PP), se ve

que, de los 20.533 casos de los que se compone la muestra, 2.871 coinciden con la mediana (*zero sign*), cuyo valor es el de 5.

### ILUSTRACIÓN 7.5. Distribución de frecuencias de la variable a comprobar

pp	Freq.	Percent	Cum.
0	5680	27.66	27.66
1	600	2.92	30.58
2	828	4.03	34.62
3	781	3.80	38.42
4	859	4.18	42.60
5	2871	13.98	56.59
6	1365	6.65	63.23
7	1483	7.22	70.46
8	2028	9.88	80.33
9	876	4.27	84.60
10	3162	15.40	100.00
Total	20533	100.00	

Por debajo del valor mediano (*negative sign*) hay en la distribución empírica de la muestra 8.748 casos ( $5.680 + 600 + 828 + 781 + 859$ ), mientras que por encima (*positive sign*) del 5 se sitúan 8.914 observaciones ( $1.365 + 1.483 + 2.028 + 876 + 3.162$ ). Es obvio que en el caso de que la mediana fuera, como se ha establecido en la hipótesis, igual a 5, entonces se tendría que haber encontrado igual número de casos por debajo que por encima del mencionado valor. El test de los signos averigua cuál es la probabilidad de encontrar un número igual o superior de casos por encima de la mediana (observaciones de signo positivo), para el caso de que la hipótesis alternativa sea unidireccional de signo “mayor que” o por debajo de la mediana (observaciones de signo negativo), en el supuesto de que la alternativa sea de naturaleza “menor que”.

La probabilidad de la prueba con alternativa unidireccional se obtiene aplicando la distribución binomial al número de observaciones positivas (8.914) o negativas (8.748) con un número de casos igual al de los que no siguen la mediana ( $8.914 + 8.748$ , en el ejemplo que se considera) y una probabilidad de 0,5. Por su lado, la prueba bidireccional se obtiene multiplicando por 2 el valor de la probabilidad menor de las anteriores (0,107, en este caso, que se convierte en 0,214).

La interpretación es simple y sigue la norma de todos los test de hipótesis. Siempre y cuando la probabilidad obtenida sea menor de 0,05, se puede rechazar la hipótesis nula con una seguridad mayor del 95%. En el ejemplo contemplado se ve que no puede ser rechazada la hipótesis de que la mediana en la población haya sido igual a 5, en ningún caso, sea cual fuere la alternativa propuesta.

Para obtener del ordenador este test, debe escribirse la orden *signtest* seguida del nombre de la variable, el signo igual y el valor de la hipótesis

nula. Siempre la salida en pantalla muestra el resultado del test en los tres supuestos de hipótesis alternativa.

```
signtest pvotopp=5
```

### ILUSTRACIÓN 7.6. Prueba de los signos con una sola variable

```
Sign test

sign | observed expected
-----+-----
positive |      8914     8831
negative |      8748     8831
zero |      2871     2871
-----+-----
all |      20533    20533

One-sided tests:
Ho: median of pvotopp - 5 = 0 vs.
Ha: median of pvotopp - 5 > 0
Pr(#positive >= 8914) =
Binomial(n = 17662, x >= 8914, p = 0.5) = 0.1072

Ho: median of pvotopp - 5 = 0 vs.
Ha: median of pvotopp - 5 < 0
Pr(#negative >= 8748) =
Binomial(n = 17662, x >= 8748, p = 0.5) = 0.8956

Two-sided test:
Ho: median of pvotopp - 5 = 0 vs.
Ha: median of pvotopp - 5 != 0
Pr(#positive >= 8914 or #negative >= 8914) =
min(1, 2*Binomial(n = 17662, x >= 8914, p = 0.5)) = 0.2144
```

## 7.2. Comparación de dos variables

Cuando se desea comparar dos variables procedentes de la misma población se está ante el caso de pruebas en muestras dependientes. Reciben este nombre porque cada caso posee un par de valores conectados de cada una de las variables en cuestión.

De modo paralelo al de los test anteriores, se van a presentar estos análisis en tres apartados: en el primero se comparan proporciones, en el segundo se equiparan medias y en el tercero se contempla el test no paramétrico de Wilcoxon o prueba de los rangos con signo.

### 7.2.1. Comparación de dos proporciones en muestras dependientes

En lugar de contrastar una proporción con un valor, se trata de comparar dos proporciones obtenidas de la misma base, esto es, con idéntico denominador. En esta ocasión, en lugar de formular la hipótesis con una cantidad, se utilizan dos variables ficticias, puesto que al igual que ocurría en la prueba con una variable, se ha de proceder como si fueran medias de variables con dos valores 0/1 ó 0/100. Matemáticamente, la formulación de la hipótesis nula es la siguiente:

$$h_0 : \Pi_x = \Pi_y \quad (7.12)$$

Supóngase que se desea contrastar con los datos de la muestra si la intención de voto a dos partidos es igual o si sigue habiendo diferencia a favor del que en tiempo pasado era superior (de este modo la alternativa es unidireccional). La variable que representa la proporción del partido previamente superior será llamada  $x$  y la del inferior como  $y$ . Por tanto, la alternativa debe aparecer como:

$$h_1 : \Pi_x > \Pi_y \quad (7.13)$$

Ambas pueden convertirse en igualdades o desigualdades en las que en uno de los dos términos aparezca el valor nulo.

$$\begin{cases} h_0 : \Pi_x - \Pi_y = 0 \\ h_1 : \Pi_x - \Pi_y > 0 \end{cases} \quad (7.14)$$

Para efectuar con Stata el correspondiente test estadístico es preciso utilizar la instrucción *ttest*, seguida de las dos variables-proporción separadas por el signo igual.

```
generate intpsoe=(intvoto==3) if (intvoto<97 | intvoto==98)
prtest intpp=intpsoe
```

El resultado es similar al obtenido en el caso de una sola muestra.

**ILUSTRACIÓN 7.7. Prueba de comparación de dos proporciones  
(muestras dependientes)**

Two-sample test of proportion				intpp: Number of obs =	19424
				intpsoe: Number of obs =	19424
Variable	Mean	Std. Err.	z	P> z	[95% Conf. Interval]
intpp	.3783979	.0034799		.3715775	.3852183
intpsoe	.2284287	.0030123		.2225248	.2343327
diff	.1499691	.0046025		.1409483	.1589899
	under Ho:	.004665	32.15	0.000	
diff = prop(intpp) - prop(intpsoe)				z = 32.1478	
Ho: diff = 0		Ha: diff != 0		Ha: diff > 0	
Pr(Z < z) = 1.0000		Pr( Z  <  z ) = 0.0000		Pr(Z > z) = 0.0000	

La diferencia básica respecto a la ilustración 7.3 es que, en lugar de aparecer una sola línea con la variable, aparecen cuatro: una para cada variable contrastada, una tercera para una nueva variable, que es la diferencia entre ambas, lo que sólo se manifiesta en la media y en los intervalos, ya que el número de observaciones es lógicamente el mismo, y una cuarta que calcula el error típico en el supuesto de que las medias de ambas fueran iguales, lo que genera un pequeño cambio en valor<sup>4</sup>, cuyo cálculo responde a la siguiente fórmula:

$$\sigma_{p_x-p_y} = \sqrt{\frac{2(p(1-p)}{n-1}} \quad (7.15)$$

El PP tiene una intención de voto del 37,8%, mientras que la del PSOE es del 22,8%, la diferencia es cercana al 15%. Mirando los intervalos de confianza para la diferencia se aprecia que están situados entre el 14,1% y el 15,9%. Obviamente, es inasumible la hipótesis nula de que la proporción de intención de voto de ambos partidos pudiera ser idéntica.

A esta misma conclusión se llega con el examen atento del estadístico  $z$  calculado y de su significación. Como la hipótesis alternativa era del tipo “mayor que”, se ha de prestar atención a la columna de la derecha, donde aparece una significación sustancialmente inferior al convencional límite del 0,05. Por tanto, estos datos dicen que en la población, en el momento

<sup>4</sup> El cambio se debe a suponer de partida que las dos medias son iguales, en lugar de asumir en contra de la hipótesis que cada variable tiene una media distinta, tal como ocurre en la muestra. Para satisfacer esta suposición, se obtiene  $p$  como promedio de las proporciones de una y otra variable. En este ejemplo tendría el valor de 0,303.

de realización del estudio, había una diferencia significativa en la intención declarada de voto a favor del PP.

### 7.2.2. Comparación de dos medias en muestras dependientes

El test de comparación de medias sigue exactamente las mismas pautas que el de proporciones, puesto que en realidad este es una adaptación de aquél, que se consigue convirtiendo uno de los valores de una variable cualitativa en otra dicotómica con valores 0 y 1. Para medias, hipótesis nula y alternativa (bidireccional, en el ejemplo, pero también puede formularse unidireccionalmente) se expresan del siguiente modo:

$$\begin{cases} h_0 : \mu_x = \mu_y \\ h_1 : \mu_x \neq \mu_y \end{cases} \quad (7.16)$$

Se trata, por tanto, en este tipo de pruebas de comparar las medias de dos variables distintas, denominadas  $x$  e  $y$ ; aunque, en el fondo, lo que se realiza es construir una nueva variable, denominada  $D$ , que es la sustracción en cada caso de los respectivos valores de  $x$  e  $y$  y verificar la hipótesis de que el valor de la nueva media sea igual a 0. El error típico de esta variable se obtiene mediante la expresión:

$$\sigma_d = \sqrt{\frac{\sum_{i=1}^n (x_i - y_i)^2}{n - 1}} \quad (7.17)$$

Si se toman las probabilidades de 0 a 10 que cada persona se atribuye de votar a dos partidos distintos en unas próximas elecciones, continuando con el ejemplo anterior, PP y PSOE, se puede adoptar una hipótesis claramente unidireccional que dependerá del momento político en el que se planteen los comicios. En el año 2000, el que se está utilizando, es obvio que la alternativa habrá que asumirla en dirección favorable al PP.

Para obtener los resultados estadísticos, se ha de recurrir a la misma instrucción que en el ejemplo anterior, utilizando la instrucción *ttest* para comparar dos variables cuantitativas (*pivotopp* y *pivotopsoe*, en esta ocasión).

```
ttest pivotopp==pivotopsoe
```

Tras lo cual aparece una tabla con similar aspecto al de la diferencia de proporciones:

**ILUSTRACIÓN 7.8. Prueba de comparación de dos medias  
(muestras dependientes)**

Paired t test						
Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
pivotopp	20310	4.693156	.0258863	3.689135	4.642417	4.743895
pivotop~e	20310	4.345987	.0229813	3.275143	4.300942	4.391032
diff	20310	.3471689	.0394482	5.62189	.2698472	.4244906
		mean(diff) = mean(pivotopp - pivotopsoe)			t =	8.8006
Ho: mean(diff) = 0					degrees of freedom =	20309
Ha: mean(diff) < 0		Ha: mean(diff) != 0			Ha: mean(diff) > 0	
Pr(T < t) = 1.0000			Pr( T  >  t ) = 0.0000			Pr(T > t) = 0.0000

Estos resultados muestran cómo la media para el PP (*pivotopp*) es significativamente más alta que la correspondiente al PSOE (*pivotopsoe*). Se diferencian entre ellas 35 centésimas, pero, al tratarse de una muestra de tamaño superior a 20.000 casos, esa pequeña diferencia no puede haberse debido a errores de muestreo, siempre y cuando este haya sido aleatorio. Obsérvese, además, cómo los intervalos de confianza de una y otra variable no se superponen. La puntuación del intervalo inferior para el PP es de 4,6, mientras que la del superior para el PSOE es de 4,4. En consecuencia, los extremos de los intervalos de las diferencias son ambos positivos, por lo que puede rechazarse la hipótesis nula.

### 7.2.3. Comparaciones no paramétricas de dos variables

Para estas circunstancias, el programa Stata ofrece un par de alternativas. Como en el caso de una sola variable, puede utilizarse la prueba de los signos y puede también emplearse una prueba de rangos.

En el primer caso, el procedimiento es similar al que ya se estudió en la prueba paramétrica de una sola variable. Pero en lugar de compararse los datos con un solo valor —el de la hipótesis— como punto de referencia, se compara con el de la otra variable en el mismo caso. Como resultado de la comparación puede obtenerse un empate, caso de que ambos valores sean idénticos, que la primera variable tenga el valor más alto o —por el contrario— que sea la segunda la de mayor valor.

Si se sigue con el ejemplo contemplado en la comparación paramétrica de medias con las variables relativas a la probabilidad personal en

una escala de 0 a 10 de votar a dos partidos, cada caso (individuo) puede ser clasificado en tres tipos: aquellos que dan la misma probabilidad de voto a los dos partidos (*zero*), los que dan mayor probabilidad al primero (*positive*) y, en tercer lugar, los que dan mayor probabilidad al segundo (*negative*). Es evidente que si hubiera equilibrio entre las dos variables, el número de sujetos del segundo y del tercer tipo debería ser similar, si no idéntico.

Para la obtención de este análisis, basta con emplear la instrucción *signtest*, escribiendo a continuación las dos variables que quieren compararse separadas por el signo igual.

```
signtest pvotepp=pvotopsoe
```

El resultado (ilustración 7.9) clasifica de la forma señalada los casos y establece las probabilidades binominales correspondientes a las tres hipótesis alternativas posibles.

Como puede apreciarse, hay 4.534 casos, de los 20.310 que componen la muestra, que comparten el mismo valor en las variables correspondientes a los partidos implicados. Pero hay 8.717 que asignan más probabilidades de voto al primero (PP) que al segundo (PSOE) y sólo 7.059 que —al contrario— dan más al segundo. En consecuencia, se puede descartar de entrada la hipótesis alternativa de que la probabilidad de que se vote al partido de izquierda sea mayor que al partido de la derecha, como se pone de manifiesto en el resultado prácticamente igual a la unidad de la probabilidad de obtener un valor igual o superior a 7.059. En cambio, si se observa la primera de las pruebas realizadas, la de que al PP se le da mayor probabilidad de ser votado, entonces la significación, obtenida a partir de los 8.717 casos que dan mayor valor a este partido, sale inferior al punto crítico del 5%. Asimismo, sale estadísticamente significativa la prueba si se opta por una hipótesis alternativa bidireccional.

### ILUSTRACIÓN 7.9. Prueba de los signos para muestras dependientes

Sign test		
sign	observed	expected
positive	8717	7888
negative	7059	7888
zero	4534	4534
all	20310	20310

One-sided tests:

```

Ho: median of pvotopp - pvotopsoe = 0 vs.
Ha: median of pvotopp - pvotopsoe > 0
Pr(#positive >= 8717) =
Binomial(n = 15776, x >= 8717, p = 0.5) = 0.0000

Ho: median of pvotopp - pvotopsoe = 0 vs.
Ha: median of pvotopp - pvotopsoe < 0
Pr(#negative >= 7059) =
Binomial(n = 15776, x >= 7059, p = 0.5) = 1.0000

```

Two-sided test:

```

Ho: median of pvotopp - pvotopsoe = 0 vs.
Ha: median of pvotopp - pvotopsoe != 0
Pr(#positive >= 8717 or #negative >= 8717) =
min(1, 2*Binomial(n = 15776, x >= 8717, p = 0.5)) = 0.0000

```

En definitiva, a la vista de los resultados mostrados, puede concluirse con tranquilidad que en la población el número de votantes que dan mayor probabilidad al Partido Popular es superior al número de votantes que se la otorgan al Partido Socialista.

Otro test para el mismo tipo de datos que incorpora más información en la medida en que también tiene en cuenta el rango de las diferencias entre las dos variables es el test del signo de los rangos. Con objeto de estudiar este procedimiento, adecuado en el supuesto de que ambas distribuciones sean simétricas, se va a considerar una selección de los diez primeros casos de la muestra.

La siguiente tabla muestra los valores de probabilidad en la escala del 0 al 10 atribuidos al Partido Popular (PP) y al Partido Socialista (PS) de los diez primeros casos numerados. En la columna siguiente (*dif*) aparecen las diferencias entre los valores de ambas variables. Para calcular el rango hay que considerar el valor absoluto de estas diferencias. Es obvio que los casos en los que ambos valores de las variables son idénticos son los que poseen la diferencia absoluta menor, por tanto todos aparecen en la columna (*Rango*) con el valor en cursiva 1. Como son cinco, se les sustituye por el rango promedio (1, 2, 3, 4, 5; esto es, el 3). Además, se encuentran tres diferencias positivas (+), cuyos rangos suman 23 (10, 6 y 7) y dos negativas con un total de rangos 17 (9 y 8).

**ILUSTRACIÓN 7.10. Rangos de las diferencias entre dos variables (PP-PS)**

O	PP	PS	dif	dif	R	Cero	+	-
1	0	0	0	0	1	3		
2	10	0	10	10	10		10	
3	0	0	0	0	1	3		
4	0	8	-8	8	9			9
5	0	0	0	0	1	3		
6	0	0	0	0	1	3		
7	0	5	-5	5	8			8
8	5	5	0	0	1	3		
9	8	5	3	3	6		6	
10	7	3	4	4	7		7	
					15	23	17	

El valor de los rangos tanto positivos como negativos esperados en el supuesto de que fuera cierta la hipótesis nula de que no hubiera diferencias entre una y otra variable se obtiene aplicando la siguiente fórmula:

$$E(R_+) = \frac{n(n+1) - 2S_0}{4} \quad (7.18)$$

... siendo  $R_+$  el número de rangos positivos,  $n$  el número de casos y  $S_0$  la suma de los rangos 0 (es decir,  $n_0(n_0 + 1)/2$ ). En este ejemplo, teniendo en cuenta que hay 10 casos ( $n$ ) y 5 empates ( $n_0$ ), la suma de los rangos correspondientes a la diferencia 0 ( $S_0$ ) es igual a 15 y el valor esperado de los rangos positivos  $E(R_+)$  es de 20. Es obvio que el valor esperado de la suma de rangos negativos ( $R_-$ ) ha de ser igual que la de los positivos ( $R_+$ ) y se cumple la igualdad

$$S = \frac{n(n-1)}{2} = S_0 + E(R_+) + E(R_-) \quad (7.19)$$

Es decir, la suma total de los rangos de  $n$  casos ( $S$ ) es igual a la suma de los rangos empadados, la de los positivos y la de los negativos.

El test se ejecuta en Stata mediante la instrucción *signrank*, seguida por las variables pareadas separadas por el signo igual. Para comprobar la natu-

raleza de esta prueba, se va a realizar en primer lugar la instrucción con los diez primeros casos aparecidos en la anterior tabla:

```
signrank pvotopp=pvotopsoe in 1/10
```

A partir de lo cual aparecen los siguientes resultados aplicados a los casos de la muestra numerados del 1 al 10.

**ILUSTRACIÓN 7.11. Prueba de Wilcoxon para muestras dependientes (diez casos)**

```
Wilcoxon signed-rank test

sign |     obs    sum ranks   expected
-----+
positive |      3        23       20
negative |      2        17       20
zero |      5        15       15
-----+
all |     10        55       55

unadjusted variance      96.25
adjustment for ties      0.00
adjustment for zeros     -13.75
-----+
adjusted variance        82.50

Ho: pvotopp = pvotopsoe
      z =      0.330
Prob > |z| =      0.7412
```

Además de la suma de rangos, el programa calcula su varianza (ajustada por posibles rangos empatados y por los primeros rangos procedentes de variables con el mismo valor) para obtener la variable  $z$ , que se distribuye normalmente. Por eso, en este ejemplo, con tan sólo diez casos, no sería posible rechazar la hipótesis nula, ya que su valor (0,33) tiene una significación (0,74) superior a la considerada como nivel aceptable de comisión de errores de tipo I (0,05).

Pero, si en lugar de pedir el análisis para los diez primeros casos, se solicita para el conjunto de la muestra ( $n = 20.310$ ), la suma de rangos es astronómica, y el valor de  $z$  lo suficientemente alto para rechazar la hipótesis nula con un nivel de significación de 5%.

```
signrank pvotopp=pvotopsoe
```

**ILUSTRACIÓN 7.12. Prueba de Wilcoxon para muestras dependientes (conjunto)**

Wilcoxon signed-rank test			
sign	obs	sum ranks	expected
positive	8717	1.067e+08	97988680
negative	7059	89260195	97988680
zero	4534	10280845	10280845
-----+-----			
all	20310	2.063e+08	2.063e+08
 unadjusted variance    6.982e+11			
adjustment for ties   -1.491e+09			
adjustment for zeros   -7.770e+09			
-----			
adjusted variance       6.889e+11			
 H <sub>0</sub> : pvotopp = pvoтопsoe			
z = 10.516			
Prob >  z  = 0.0000			

### 7.3. Comparaciones de dos muestras (independientes)

Acaban de explicarse las pruebas con muestras dependientes o paralelas, que comparan dos variables procedentes de la misma población en donde cada valor de una variable está ligado al de la otra por pertenecer al mismo caso de estudio. Las situaciones en las que se ha de comparar una medida de los mismos sujetos en dos momentos temporales son las más típicas de este tipo de pruebas, aunque no las únicas, como se ha visto en los ejemplos expuestos. En cambio, si se pretende efectuar la comparación con el mismo estadístico en dos muestras distintas para ver si proceden de poblaciones similares, se está ante las pruebas con muestras independientes, como, por ejemplo, puede ser la comparación de la intención de voto entre mujeres y hombres.

Al igual que se hiciera en el apartado anterior, se va a subdividir este apartado en distintas secciones. En primer lugar, se procederá a la comparación de proporciones; seguidamente, se pasará a la comparación de varianzas, ya que es paso previo para optar por una u otra fórmula de comparación de medias, que será abordada en la tercera parte, y finalmente se abordará la confección de pruebas paramétricas con muestras independientes.

#### 7.3.1. Comparación de dos proporciones (en muestras independientes)

En el fondo, no existe un procedimiento específico de comparación de proporciones, por lo que pueden usarse indistintamente *prtest* y *ttest*, con la precaución de expresar el porcentaje deseado como el valor “uno” de una variable dicotómica. Al igual que en el ejemplo anterior, se va a tomar como

ejemplo la intención de voto, pero en este caso no se necesita la de dos partidos. Basta con una sola comparada en dos grupos distintos, que serán, para simplificar el ejemplo, el de los hombres, por un lado, y el de las mujeres por el otro. O, dicho de otro modo, se trata de ver si para los dos grupos generados por la variable *sexo*, el porcentaje de intención de voto al PP (*intpp*) es similar o significativamente distinto.

$$\begin{cases} h_0 : \Pi_1 = \Pi_2 \\ h_1 : \Pi_1 \neq \Pi_2 \end{cases} \quad (7.20)$$

Para realizar esta prueba estadística ha de emplearse también la orden *ttest* (o *prtest*), puesto que la diferencia de proporciones se ajusta a esta distribución de Student, asimilable a la normal, si los grados de libertad son suficientes, aproximadamente cuando  $n > 30$ .

```
ttest intpp, by (sexo)
```

Aunque en este caso, por la enorme muestra que se está empleando, se esté utilizando en realidad la distribución normal, el resultado del programa sigue presentando la distribución de Student, que con tantos grados de libertad se aproxima a la forma estándar de la campana de Gauss.

### ILUSTRACIÓN 7.13. Prueba de comparación de medias (muestras independientes)

Two-sample t test with equal variances						
Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
hombre	9388	.3857052	.005024	.4867874	.375857	.3955534
mujer	10033	.3715738	.0048245	.4832494	.3621167	.3810309
combined	19421	.3784048	.0034802	.4850018	.3715833	.3852264
diff		.0141313	.0069637		.0004818	.0277809
diff = mean(hombre) - mean(mujer)					t =	2.0293
Ho: diff = 0					degrees of freedom =	19419
Ha: diff < 0						
Pr(T < t) = 0.9788						
Ha: diff != 0						
Pr( T  >  t ) = 0.0424						
Ha: diff > 0						
Pr(T > t) = 0.0212						

La tabla de resultados se presenta de modo similar a la de comparación de muestras dependientes, pues incluye las dos líneas correspondientes a las estadísticas de uno y otro grupo (antes una y otra variable) y otra línea con

las diferencias entre ambas. Pero, a diferencia de la primera, incluye una línea (la tercera de cifras) con los resultados del conjunto (*combined*) de la muestra. El 37,8% de las 19.421 personas que contestaron a esta pregunta en la muestra dicen decantarse por el voto al Partido Popular.

Como antes de realizar el estudio no estaba claro qué grupo de personas iba a tener mayor o menor proporción de preferencias por este partido, es más que razonable que la hipótesis alternativa sea bidireccional. En este caso, el análisis muestra que el 38,6% de hombres tienen intención de votar al partido en consideración y sólo el 37,2% de las mujeres. Y como una vez más la muestra es muy amplia, los errores típicos en cada uno de los grupos conformados por la variable *sexo* son muy pequeños. Por ello, el intervalo de confianza con un 95% de seguridad esta comprendido sólo entre 2 puntos porcentuales: para los primeros entre el 37,6% y el 39,6% y para las segundas entre el 36,2% y el 38,1%.

Pero, sin duda, los datos de mayor interés en esta prueba aparecen en la línea de las diferencias (*diff*), donde aparecen cuatro cifras: la primera es la diferencia de las proporciones correspondientes a los dos grupos (hombres y mujeres), la segunda el error típico, o desviación típica de la distribución muestral de la diferencia de proporciones, mientras que la tercera y la cuarta son los límites inferior y superior correspondientes al intervalo de confianza (por defecto con un 95% de seguridad en muestras aleatorias) de la diferencia de proporciones. Como en este caso, estos límites no incluyen el valor 0, pues ambos son positivos, puede ser rechazada la hipótesis nula con una seguridad mayor del 5%.

Esto mismo se deduce al observar las pruebas de hipótesis efectuadas con el estadístico *t* de Student. Si se observa la columna correspondiente a la hipótesis alternativa bidireccional (*ha !=0*), la probabilidad correspondiente al valor de la *t* empírica (cociente entre la diferencia y su error típico) es menor que el consabido 5%, con el que suelen trabajar los científicos sociales.

### 7.3.2. Comparación de varianzas (*muestras independientes*)

Del igual modo que se comparan proporciones o medias, también existen pruebas estadísticas para determinar si las diferencias encontradas en los valores de la varianza de dos muestras han podido ser debidas o no a errores de muestreo. Pero, si en los primeros casos se utiliza la distribución de la *t* de student, prácticamente normal a partir de 30 grados de libertad, para la comparación de la homogeneidad de dos muestras, hay que utilizar la distribución F de Snedecor.

Puesto que se trabaja con dos grupos, las varianzas o desviaciones típicas de cada uno de ellos se reconoce mediante la inclusión en un subíndice del número 1 o del número 2, correspondientes a las dos muestras que se están comparando. Por tanto, en hipótesis alternativas bidireccionales, la notación de esta prueba es como sigue:

$$\begin{cases} h_0 : \sigma_1 = \sigma_2 \\ h_1 : \sigma_1 \neq \sigma_2 \end{cases} \quad (7.21)$$

La instrucción válida para esta operación en Stata es *sdtest*, que tiene una sintaxis completamente similar a la de *ttest*, salvo en las opciones. Por tanto, para ver si la probabilidad asignada al voto al PP tiene una homogeneidad similar entre hombres y mujeres, se debe escribir la siguiente línea:

```
sdtest pvotepp, by(sexo)
```

Mediante esta instrucción se realiza una prueba de comparación de varianzas o desviaciones típicas mediante la prueba *F*. El resultado será similar al que aparece a continuación:

#### ILUSTRACIÓN 7.14. Prueba de comparación de varianzas (muestras independientes)

Variance ratio test								
Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]			
hombre	10302	4.593671	.0365807	3.712892	4.521966	4.665376		
mujer	10228	4.824892	.0362677	3.66788	4.753801	4.895984		
combined	20530	4.708865	.025769	3.692256	4.658356	4.759374		
					f = 1.0247			
Ho: ratio = 1					degrees of freedom = 10301, 10227			
Ha: ratio < 1		Ha: ratio != 1		Ha: ratio > 1				
Pr(F < f) = 0.8917		2*Pr(F > f) = 0.2166		Pr(F > f) = 0.1083				

Las tres primeras líneas numéricas coinciden con las pruebas de comparación de medias. Sólo se encuentran diferencias a partir del enunciado de la hipótesis nula consistente en que la desviación típica en el primer grupo, el de hombres en este caso, es igual que la propia del segundo grupo, mujeres en este ejemplo. En la muestra se ve que los primeros tienen una desviación típica algo mayor de 3,7, mientras que las mujeres tienen algo menos de dicha cantidad. La diferencia es sólo de 5 centésimas. Pero como las muestras son muy grandes, el test de la *F* indica que podría haberse debido a errores muestrales.

La salida del programa calcula tres probabilidades distintas del valor de *F*, según la hipótesis alternativa sea unidireccional (en los extremos) o bidireccional (en el centro). El primero (el situado a la izquierda) en el caso de que la primera desviación típica sea menor que la segunda; el segundo (en el centro) para la hipótesis alternativa bilateral, y el tercero (a la derecha), cuando la alternativa sea que el primer grupo tiene una heterogeneidad mayor que el segundo.

Si se utiliza este programa como paso anterior para la comparación de medias (que exige comprobar previamente si las varianzas son o no iguales entre los grupos), el valor más indicado es el bilateral. Por tanto, en este ejemplo, aun siendo la muestra bastante grande, no puede rechazarse la hipótesis nula de la homocedasticidad<sup>5</sup> en los dos grupos.

### 7.3.3. Comparación de medias en muestras independientes

Una vez que se ha realizado la comprobación de si las varianzas en la variable que se va a comparar son iguales o diferentes en los dos grupos, se puede proceder a formular el test de comparación de medias de muestras independientes. Dado que ahora se trata de medias, la formulación ha de ser como sigue:

$$\begin{cases} h_0 : \mu_1 = \mu_2 \\ h_1 : \mu_1 \neq \mu_2 \end{cases} \quad (7.22)$$

Según se haya o no rechazado la hipótesis nula de igualdad de varianzas, existen dos fórmulas para solicitar el análisis pertinente. Si se ha admitido la hipótesis alternativa, es decir, caso de que las varianzas sean significativamente diferentes, se deberá incluir la opción *unequal*:

```
ttest variable_dependiente, by(variable_grupal) unequal
```

Pero en el supuesto de que el resultado de la prueba de las varianzas sea no significativo, esto es, sin posibilidad de rechazar la hipótesis nula, tal como sucedió en el ejemplo precedente, la instrucción ha de ser similar a la anterior sin la opción *unequal*:

```
ttest variable_dependiente, by(variable_grupal)
```

De este modo, para ver si la probabilidad de que se vaya a votar al PP es igual o distinta según se sea hombre o mujer, la instrucción literal es como sigue:

```
ttest pvotopp, by (sexo)
```

---

<sup>5</sup> *Homocedasticidad* significa igualdad de varianza o similar homogeneidad. Su antónimo es *heterocedasticidad*.

El resultado es semejante al que se obtiene cuando se comparan dos proporciones:

### **ILUSTRACIÓN 7.15. Prueba de comparación de dos medias independientes (varianzas iguales)**

Two-sample t test with equal variances											
Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]						
hombre	10302	4.593671	.0365807	3.712892	4.521966	4.665376					
mujer	10228	4.824892	.0362677	3.66788	4.753801	4.895984					
combined	20530	4.708865	.025769	3.692256	4.658356	4.759374					
diff		-.2312213	.0515143		-.3321934	-.1302492					
diff = mean(hombre) - mean(mujer)											
Ho: diff = 0											
Ha: diff < 0											
Pr(T < t) = 0.0000		Ha: diff != 0			Ha: diff > 0						
		Pr( T  >  t ) = 0.0000			Pr(T > t) = 1.0000						
t = -4.4885											
degrees of freedom = 20528											

En la ilustración anterior se ve cómo la mujer en una escala del cero al diez da casi tres décimas más de probabilidad al voto al Partido Popular. A pesar de tan reducidas diferencias, salen significativas por estar trabajando con muestras tan considerables. Como puede apreciarse en la línea de las diferencias de medias (*diff*), el intervalo de confianza se mantiene entre -0,33 y -0,13. Como ambos límites, inferior y superior, son negativos, puede rechazarse con un 95% de seguridad la hipótesis nula. Igual conclusión se obtiene si se observa la probabilidad bilateral (tratándose de una hipótesis alternativa bidireccional) del valor empírico de *t*, es decir, de -4,48. Al ser inferior a 0,05, puede rechazarse la hipótesis de que las medias respectivas de hombres y mujeres sean iguales en la población.

#### *7.3.4. Pruebas no paramétricas para muestras independientes*

La prueba estadística no paramétrica más indicada para muestras independientes es la *U* de Mann-Whitney, también conocida como la prueba de las suma de rangos de Wilcoxon. Opera de modo similar a la ya expuesta del signo de los rangos: se agrupan los datos de las dos muestras en un solo grupo, se les asigna el rango correspondiente al valor de cada caso y se intenta comprobar si la suma de los rangos de un grupo es igual o no a la del otro grupo.

En este caso la suma esperada de los rangos del grupo *j* se ajusta a la siguiente expresión:

$$R_j^* = \frac{n_j(n+1)}{2} \quad (7.23)$$

Y el valor  $z$ , con distribución normal, es el resultado de dividir la diferencia entre esta suma de rangos encontrada y la esperada por la desviación típica ajustada.

Para que Stata produzca esta prueba se debe utilizar la instrucción *ranksum* con el mismo formato que la instrucción *ttest*. De este modo, si a los datos anteriores se les quiere aplicar una prueba no paramétrica, el modo adecuado de solicitarlo es mediante la inserción del siguiente comando:

```
ranksum pvotopp, by (sexo)
```

... tras cuya inserción el resultado obtenido sería el siguiente:

#### **ILUSTRACIÓN 7.16. Prueba de la suma de rangos**

Two-sample Wilcoxon rank-sum (Mann-Whitney) test			
sexo	obs	rank sum	expected
hombre	10302	1.039e+08	1.058e+08
mujer	10228	1.068e+08	1.050e+08
combined	20530	2.108e+08	2.108e+08
<hr/>			
unadjusted variance	1.803e+11		
adjustment for ties	-5.315e+09		
<hr/>			
adjusted variance	1.750e+11		
<hr/>			
Ho: pvotopp(sexo==hombre) = pvotopp(sexo==mujer)			
z = -4.366			
Prob >  z  = 0.0000			

Como puede apreciarse aquí, la diferencia en la suma de rangos es significativa al proporcionar un valor normalizado superior a 4. Por tanto, se puede afirmar que la pauta ordinal de la variable *voto al PP* es distinta entre hombres y mujeres; dicho de modo más simplificado, las medianas de estas dos variables son distintas.

## **7.4. Comparaciones de $k$ muestras independientes**

Las pruebas que se han visto hasta ahora sólo podían aplicarse a la comparación de dos entidades. Las que se abordan a continuación permiten comparar más de dos objetos. En un primer momento, se estudiarán las pruebas que permiten averiguar si son iguales o no medias, proporciones o varianzas calculadas en distintas muestras y, posteriormente en el próximo apartado, se analizarán las que implican la comparación de más de dos variables. Y, como en los análisis precedentes, también cabe aquí la aplicación

de técnicas paramétricas, cuando se cumplen una serie de supuestos, o de pruebas más robustas que no necesitan estos requerimientos.

#### *7.4.1. Comparaciones no paramétricas de k muestras*

Para la comparación de más de  $k$  muestras existe una ampliación de la prueba de Mann-Whitney, llamada de Kruskal-Wallis, que utiliza la distribución de  $\chi^2$ . Está basada, como la técnica precedente, en comparar en cada grupo la suma de rangos.

Antes de proceder a las pruebas que comparan más de dos grupos, es conveniente solicitar una tabla que describa número de casos, media y desviación típica de cada grupo. Como ejemplo, se va a tomar la misma variable dependiente, es decir, la probabilidad de voto que cada entrevistado se atribuye al Partido Popular, pero, en lugar de emplear como variable grupal el *sexo*, se utilizará la *edad* recodificada en siete categorías. Recuérdese que esta operación de agrupamiento de valores ha de hacerse mediante la siguiente instrucción:

```
recode edad (18/25=1 "18-25") (26/35=2 "26-35") (36/45=3 "36-45") ///
(46/55=4 "46-55") (56/65=5 "56-65") (66/75=6 "66-75") ///
(76/98=7 "76-98") (99=.,) gen(edadr)
```

La solicitud de los estadísticos para cada grupo puede realizarse mediante la siguiente modalidad de la orden *tabulate*:

```
tabulate edadr, summarize(pvotopp)
```

Mediante ella se obtiene media, desviación típica y número de casos tanto para cada uno de los grupos como para el conjunto de la muestra:

#### **ILUSTRACIÓN 7.17. Tabla de comparación de medias y desviaciones típicas**

edadr	Summary of pp		
	Mean	Std. Dev.	Freq.
18-25	4.2746458	3.5555559	3317
26-35	4.396222	3.6012627	4235
36-45	4.2245061	3.6288486	3746
46-55	4.7932773	3.6942614	2975
56-65	5.5415361	3.7011868	2552
66-75	5.3428237	3.7629557	2564
75+	5.2223199	3.84858	1138
Total	4.7092123	3.6924366	20527

El promedio que da el conjunto de la muestra a la probabilidad de votar al partido en cuestión es de 4,7. Puede observarse, además, cómo los grupos de edad con menos de 45 años están por debajo de este valor y los que tienen más de 56 años otorgan una puntuación sensiblemente más alta, superior a los 5 puntos. Se trata de ver ahora si estas diferencias son significativas. A este fin se aplica la instrucción que realiza la prueba de Kruskal-Wallis:

```
kwallis pvotopp, by(edadr)
```

Nótese que el orden de las variables es el inverso del de la orden *tabulante*. Como en las otras pruebas comparativas de muestras independientes, la variable que ha de ser comparada ha de expresarse en primer lugar y en el último la variable grupal.

#### **ILUSTRACIÓN 7.18. Prueba de Kruskal-Wallis de igualdad de medias en muestras independientes**

```
Test: Equality of populations (Kruskal-Wallis test)

edadr      Obs     RankSum
18-25      3317    31742682.00
26-35      4235    41328124.00
36-45      3746    35568528.00
46-55      2975    30939640.00
56-65      2552    29588968.00
66-75      2564    28905476.00
75+        1138    12615706.00

chi-squared =   365.915 with 6 d.f.
probability =  0.0001

chi-squared with ties =  377.031 with 6 d.f.
probability =  0.0001
```

En los resultados se ofrecen dos cantidades de  $\chi^2$ : la original y la corregida por la presencia de empates en el rango de las puntuaciones. Ambas confirman que puede rechazarse con un nivel de seguridad superior al 95% la hipótesis nula de que el rango medio en cada grupo de edad es similar. O, lo que es lo mismo, puede asegurarse que existen diferencias significativas por edad en la probabilidad de voto al Partido Popular. Este, como se ha visto en la tabla de medias, es probablemente más votado entre las personas mayores.

#### **7.4.2. Comparaciones paramétricas de k medias**

De una población dividida en  $k$  grupos con medias  $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \dots, \alpha_k$ , se extraen  $k$  muestras aleatorias con medias  $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \dots, \bar{x}_k$ . La prueba del

análisis de varianza trata de verificar si con las medias maestrales obtenidas puede sostenerse la hipótesis de igualdad de medias en la población:

$$\left\{ \begin{array}{l} h_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots = \mu_k \\ h_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \dots \neq \mu_k \end{array} \right. \quad (7.24)$$

Los supuestos para realizar comparaciones paramétricas de  $k$  medias son los siguientes:

1. Las muestras se han seleccionado aleatoria e independientemente de las  $k$  poblaciones.
2. Las distribuciones en la población de la variable cuya media se compara son normales en cada uno de los grupos.
3. Las desviaciones típicas de la variable en cada una de las poblaciones son iguales entre sí.

La primera condición se cumple siempre y cuando se hayan extraído muestras independientes y aleatorias de los datos. Es presumible que al aplicar un cuestionario a personas seleccionadas al azar se cumpla este supuesto del análisis de varianza, siempre y cuando los valores de la variable grupal sean mutuamente excluyentes.

Para el segundo supuesto, existen distintas pruebas para comprobar la normalidad de unos datos maestrales. En Stata son utilizables diversos procedimientos para comprobar si una distribución es o no normal. Entre ellos están *swilk*, *sfrancia*, *ksmirnov* y *sktest*.

Entre estos son más recomendables en este contexto los dos primeros, pues son los únicos que permiten realizarse con la opción *by* para obtener una prueba de normalidad de la variable en cuestión para cada una de las muestras independientes extraídas.

Como un ejemplo de análisis de varianza, se considera como variable dependiente la probabilidad de voto al Partido Popular (*pvtopp*) por niveles de estudio. Antes de ejecutar el análisis de varianza, se combinan las variables *escuela* y *estudios*, para aplicarles conjuntamente la prueba de normalidad para cada grupo de la muestra mediante la instrucción *swilk* precedida por *bysort*:

```
replace estudios=1 if (escuela==1 | escuela==2)
recode estudios .=9
label define estudios 1 "Sin estudios", add
bysort estudios: swilk pvtopp
```

Como la variable *estudios* posee, además de los *no contesta*, siete valores, el análisis procede a la realización de ocho pruebas. Como puede comprobarse las seis primeras ofrecen diferencias sustantivas con respecto a la normalidad,

mientras que los dos últimos grupos, los menores en tamaño, podrían provenir de poblaciones en las que la distribución de la variable fuera normal. Sin embargo, a pesar de que en los grupos importantes no se cumpla el supuesto de normalidad, cuando los tamaños grupales son grandes, el requisito no es tan sustancial. Sirve más bien para determinar qué prueba de homocedasticidad es más adecuada aplicar, pues el supuesto de igualdad de varianzas es mucho más importante que el de la normalidad de las poblaciones.

### ILUSTRACIÓN 7.19. Prueba de Shapiro-Wilk de normalidad en los datos

```
-> estudios = Sin estudios
      Shapiro-Wilk W test for normal data
      Variable |   Obs      W      V      z     Prob>z
-----+-----+
      pvtopp | 1603    0.98572   13.849   6.631  0.00000

-> estudios = primarios
      Shapiro-Wilk W test for normal data
      Variable |   Obs      W      V      z     Prob>z
-----+-----+
      pvtopp | 9534    0.98257   83.289  11.824  0.00000

-> estudios = secundarios
      Shapiro-Wilk W test for normal data
      Variable |   Obs      W      V      z     Prob>z
-----+-----+
      pvtopp | 3482    0.98361   32.118   8.998  0.00000

-> estudios = formación profesional
      Shapiro-Wilk W test for normal data
      Variable |   Obs      W      V      z     Prob>z
-----+-----+
      pvtopp | 2533    0.98383   23.761   8.127  0.00000

-> estudios = universitarios de grado medio
      Shapiro-Wilk W test for normal data
      Variable |   Obs      W      V      z     Prob>z
-----+-----+
      pvtopp | 1795    0.98323   18.028   7.327  0.00000

-> estudios = universitarios de grado superior
      Shapiro-Wilk W test for normal data
      Variable |   Obs      W      V      z     Prob>z
-----+-----+
      pvtopp | 1500    0.98038   17.906   7.260  0.00000

-> estudios = otros estudios no reglados
      Shapiro-Wilk W test for normal data
      Variable |   Obs      W      V      z     Prob>z
-----+-----+
      pvtopp | 50      0.96225   1.775    1.224  0.11045

-> estudios = n.c.
      Shapiro-Wilk W test for normal data
      Variable |   Obs      W      V      z     Prob>z
-----+-----+
      pvtopp | 36      0.95404   1.676    1.080  0.14009
```

El tercer supuesto que hay que comprobar es el de la homocedasticidad; para ello se puede utilizar el test de Levene o cualquiera de sus variantes, que se pueden obtener a través de la instrucción *robvar*.

robvar pvtopp, by(estudios)
-----------------------------

Mediante esta instrucción, recomendable siempre que se quiera hacer una comparación de medias mediante el análisis de varianza, las pruebas de homocedasticidad están precedidas por las medias, desviaciones típicas y número de casos de cada uno de los grupos.

#### ILUSTRACIÓN 7.20. Prueba de Levene de igualdad de varianzas

estudios	Summary of pp		
	Mean	Std. Dev.	Freq.
Sin estud	4.6219588	3.7790826	1603
Primarios	4.495794	3.6834252	9534
Secundari	4.6550833	3.650202	3482
Formación	4.1729175	3.6172672	2533
Medios	4.6250696	3.6810303	1795
Superiore	4.3446667	3.7301225	1500
Otros	5.56	4.0866133	50
N.c.	3.75	4.0523362	36
Total	4.7089563	3.6922759	20533
W0 = 3.5717968	df(7, 20525)	Pr > F = .00076178	
W50 = 2.677433	df(7, 20525)	Pr > F = .00905666	
W10 = 3.5717968	df(7, 20525)	Pr > F = .00076178	

Tras el resumen por grupos y total de la variable cuyas diferencias se quiere encontrar, aparecen en tres líneas las tres pruebas incluidas en este procedimiento. En primer lugar aparece el test de Levene (*W0*), en el que se efectúa un análisis de varianza con las diferencias absolutas de las puntuaciones de cada individuo con respecto a su media grupal. Después aparece el *W50*, que hace lo mismo pero efectuando las diferencias con relación a la mediana grupal, y el *W10*, que lo realiza con la media recortada, calculada con el 80% de los casos centrales, es decir, excluyendo al 10% de los casos con puntuaciones menores y al otro 10% con puntuaciones mayores.

En este caso, la conclusión que hay que tomar es que no se dan las condiciones de homocedasticidad, por lo que habría que tomar con mucha precaución el resultado del análisis de varianza. Para que Stata haga este se pueden emplear las instrucciones *oneway* y *anova*. En este apartado sólo se contempla la primera, dejando la segunda para otro próximo, pues posee también la posibilidad de comparar muestras dependientes.

Para obtener un análisis de varianza con el procedimiento *oneway*, basta con expresar detrás de la instrucción, en primer lugar, la variable cuantitativa y en segundo lugar la grupal. En el ejemplo actual, primero ha de aparecer la variable *pvotepp* y después la variable *estudios*.

```
oneway pvotepp estudios
```

El resultado no sólo expresa las sumas y medias cuadráticas de las tres fuentes de variación (total, interna y externa), sino que también incluye otro test de comparación de varianzas, el de Bartlett.

### ILUSTRACIÓN 7.21. Análisis de varianza para muestras independientes

Source	Analysis of Variance			F	Prob > F
	SS	df	MS		
Between groups	1622.11558	7	231.730797	17.09	0.0000
Within groups	278288.607	20525	13.5585192		
Total	279910.723	20532	13.632901		
Bartlett's test for equal variances: chi2(7) = 6.6295 Prob>chi2 = 0.468					

La suma cuadrática total (*Total SS* o *SCT*) representa la suma de las desviaciones al cuadrado de todos los valores con respecto a la media global de la muestra.

$$SCT = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 \quad (7.25)$$

En la fórmula se suman en cada uno de los *J* grupos, las *n<sub>j</sub>* diferencias cuadráticas entre los valores (*x<sub>ij</sub>*) y las medias (*bar{x}*). No es esto otra cosa que el numerador de la varianza; por lo que al dividirla por los grados de libertad (*df*) de la muestra, se obtiene la cuasivarianza, conocida en este contexto como la media cuadrática total (*Total MS* o *MCT*).

$$MCT = \frac{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2}{n - 1} \quad (7.26)$$

Su raíz cuadrada equivale a la cuasidesviación típica de la variable cuyas medias se están comparando en distintas submuestras, en este caso, de la variable que mide la probabilidad subjetiva de voto al PP.

La suma cuadrática se descompone en dos: la externa (intergrupal, *between* o *SCE*) y la interna (intragrupal, *within* o *SCI*). La primera de estas recoge las desviaciones al cuadrado de cada una de las medias de los grupos con respecto a la media global, es decir, refleja las diferencias existentes entre los distintos grupos.

$$SCE = \sum_{j=1}^k (\bar{x}_j - \bar{x})^2 n_j \quad (7.27)$$

En cambio, la suma cuadrática interna representa las desviaciones existentes de los valores con respecto a la media de su grupo, es decir, la variación que existe en el interior de cada una de las submuestras obtenidas.

$$SCI = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \quad (7.28)$$

De las sumas cuadráticas se obtienen las medias cuadráticas, dividiendo las primeras por sus grados de libertad. En el caso de la suma cuadrática externa (*between*), los grados de libertad son iguales al número de grupos menos uno y en el de la suma cuadrática interna (*within*) al número de casos del conjunto menos el número de grupos.

$$\begin{aligned} MCE &= \frac{SCE}{k-1} \\ MCI &= \frac{SCI}{n-k} \end{aligned} \quad (7.29)$$

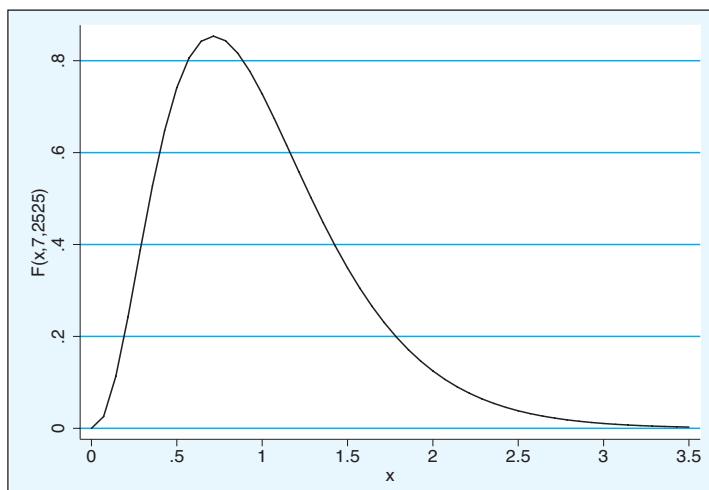
Como la hipótesis nula que se está comprobando es la igualdad de medias de los distintos grupos, es obvio que cuanto mayor sea la suma cuadrática intergrupal, menores serán las garantías de que aquella sea cierta. Se sabe que, en caso de que se cumplan los supuestos mencionados, el cociente de la media cuadrática intergrupal entre la media cuadrática intragrupal sigue la distribución *F*.

$$F = \frac{MCE}{MCI} \quad (7.30)$$

En el ejemplo anteriormente expuesto, la *F* da un valor extremadamente grande (17,09), cuya probabilidad de ocurrencia, en el caso de que la hipó-

tesis nula fuera cierta, sería ínfima (no superior a 0,00005); por ello, sería poco arriesgado el rechazo de esta, esto es, decir que la valoración electoral a un determinado partido es distinta según los diferentes niveles de estudio de los votantes. En el gráfico 7.1 está representada la teórica distribución de la  $F$  en el caso de que la hipótesis nula fuera cierta. Como se puede apreciar, es ínfima la posibilidad de que el valor sea superior incluso a 3,5 ( $p \leq 0,001$ ).

**GRÁFICO 7.1. Distribución F con 7 y 2525 grados de libertad**



En este ejemplo se impone volver al tema del supuesto de homocedasticidad, puesto que ofrece ciertas contrariedades que deben ser tenidas en cuenta. Con el análisis *oneway*, Stata obtiene la prueba de Bartlett de igualdad de varianzas. En el ejemplo de la ilustración 7.21 no parece haber indicios de heterocedasticidad, ya que la probabilidad de la medida es superior a 0,05. Esto se contradice con el anterior test de Levene, con el que se rechazó la hipótesis nula de homocedasticidad. ¿Cuál de estos artilugios, el de Bartlett o el de Levene, es más fiable? En principio, el segundo, pues el primero es más sensible al supuesto de que las subpoblaciones tengan una distribución normal, condición que se comprobó con la prueba de Shapiro-Wilks que no era cierta. Por tanto, a pesar de la prueba de Bartlett, el ejemplo no ofrece garantías de homocedasticidad. No obstante, como la significación del valor  $F$  en la prueba de comparación de medias es tan ínfima, a pesar de no cumplirse los supuestos, seguiría pudiéndose rechazar la hipótesis nula con gran tranquilidad de no equivocarse.

Mediante la opción *tabulate* del programa *oneway* se obtiene una tabla similar a la que producen otras instrucciones, como *tabulate*, *table* o *robvar*,

con medias, desviaciones típicas y frecuencias por grupo, pero, además, también es posible realizar pruebas de comparaciones múltiples con las opciones *scheffe*, *bonferroni* o *sidak*. Estas tres responden con diferentes criterios a corregir el problema que supone realizar muchas pruebas de significación al mismo tiempo. Si de cada cien comparaciones, cinco salen significativas aleatoriamente de cada diez comparaciones, media tendría que salir significativa. Para evitar rechazar aleatoriamente, diferencias que no lo son, se aplican criterios más estrictos que la significación otorgada por la *t* de Student para comparar sólo dos muestras.

En el ejemplo actual, utilizando el criterio de Scheffé, que consiste en dividir el cuadrado de la distribución *t* de Student por el número de grupos menos 1, se obtiene una distribución *F* con  $k-1$  grados de libertad en el numerador y  $n-1$  grados en el denominador. Estas operaciones se obtendrían con la siguiente instrucción, cuyos resultados se presentan en la ilustración 7.22:

oneway pvtopp estudios, noanova scheffe

### ILUSTRACIÓN 7.22. Comparación de medias con correcciones de Scheffé

		Comparison of pp by estudios (Scheffe)						
Row Mean	Col Mean	Sin estu	Primario	Secundar	F.P	Medios	Superior	Otros
Primario		.335981 0.121						
Secundar		.033124 1.000	-.302857 <b>0.016</b>					
F.P		-.449041 0.042	-.785023 0.000	-.482166 0.001				
Medios		.003111 1.000	-.33287 0.090	-.030014 1.000	.452152 <b>0.027</b>			
Superior		-.277292 0.733	-.613273 <b>0.000</b>	-.310417 0.384	.171749 0.957	-.280403 0.692		
Otros		.938041 0.871	.60206 0.988	.904917 0.887	1.38708 0.433	.93493 0.872	1.21533 0.627	
N.c		-.871959 0.961	-1.20794 0.796	-.905083 0.951	-.422917 1.000	-.87507 0.960	-.594667 0.996	-1.81 0.653

De esta matriz de comparaciones —donde aparecen en cada celda dos cifras: la diferencia de medias (de la correspondiente al grupo de fila menos de la correspondiente al de la columna) y la significación corregida— se deduce que a pesar de que los que no contestan al nivel de estudios son los que con menor probabilidad dan su voto al PP (todas las restas de su fila son negativas) y que los que poseen otros estudios son los que otorgan mayor probabilidad (todas las restas de su fila son positivas), ninguna de las dos resultan significativas, porque son categorías con una pequeña cantidad de casos.

En cambio, la categoría F.P, con una media de 4,2, aparece como el grupo con menor probabilidad significativa de votar a favor del PP. Sólo presenta asociaciones no significativas con el grupo de estudios superiores —que también tienen una probabilidad media por debajo de la del conjunto de la muestra— y con los ya mencionados grupos menores de otros estudios y de casos que no contestan. Y, en el otro extremo, se encuentra la categoría “Primarios”, que presenta medias significativamente mayores que —consecuentemente— la categoría “F.P”, pero también con la de estudios secundarios y con la de estudios superiores.

También pueden efectuarse comparaciones de muestras independientes con la instrucción *anova*, pero el uso de esta se verá con más detalle en el siguiente apartado, cuando se vean los análisis comparativos de muestras dependientes.

## 7.5. Comparaciones de *k* muestras dependientes

Algo más complejo es el empleo de pruebas de muestras dependientes con el programa Stata. En primer lugar, no se puede emplear la más simple instrucción *oneway*, sino que ha de utilizarse la más compleja *anova*. Y eso no es todo, porque además han de disponerse los datos en el formato alargado, tal como se explicó en la sección 5.1.4.

**ILUSTRACIÓN 7.23. Estructura ancha de la matriz de datos**

id	ideopp	ideopnv	ideoea
1	7	6	4
2	8	5	3
...	...	...	...

Generalmente, el formato de los ficheros de datos se ajusta al modelo ancho, de forma que en las líneas se encuentran los casos y en las columnas, las variables. Para muestras dependientes esto implicaría que cada variable aparece en una columna distinta. Sin embargo, el tratamiento de la instrucción *anova* requiere que la información esté dispuesta en otro formato, el alargado. Se necesita que haya una variable única con todo lo que se desea comparar (*vardep*, *ideo* en el ejemplo) y al menos otras dos variables: una que sea la fuente de la comparación (*varrep*, *partido*), o el número de repetición de la medida, y otra que indique al sujeto de comparación (*varid*, *id*). Esto, que parece tan complejo, puede entenderse mejor visualmente comparando los mismos datos expresados a lo ancho (ilustración 7.23) y a lo largo (ilustración 7.24):

**ILUSTRACIÓN 7.24. Estructura alargada de la matriz de datos**

	<b>id</b>	<b>partido</b>	<b>ideo</b>
	1	1	7
	1	2	6
	1	3	4
	2	1	8
	2	2	5
	2	3	3
	...	...	...

Con el fichero en formato alargado, se puede aplicar la instrucción idónea para efectuar un análisis de varianza de medidas repetidas, que posee esta fórmula general:

```
anova vardep varid varrep, repeated (varrep)
```

Donde, *vardep* es la variable dependiente (la *ideo* de la ilustración 7.24); *varid*, la identificadora de los individuos o casos (*id*), y *varrep*, la que indica de qué número de variable se trata (*partido*).

Véase todo el proceso con los datos del cuestionario electoral, suponiendo que se tengan que comparar las medias de ubicación en la escala ideológica de tres partidos en el país vasco: PP, PNV y EA.

En primer lugar, se debe proceder al arreglo del fichero para que sea posible el tratamiento. Ya es conocido que Stata trabaja normalmente con archivos *anchos*, esto quiere decir que todas las variables se encuentran en la dimensión vertical de la matriz de datos. Pero, para que se pueda realizar este análisis, las distintas variables han de estar dispuestas en distintas filas como si de casos diferentes se tratara. Esto puede solucionarse fácilmente con la instrucción *reshape long*, pero esta necesita que las variables tengan nombres similares que se distingan únicamente por números consecutivos al final, por ejemplo, *variable1*, *variable2...*, *variable#*. En el ejemplo actual, las variables que se quieren comparar se denominan *ideopp*, *ideoea* y *ideopnv*. Por ello, han de ser transformadas, para disponer de una serie ordinal consecutiva que pueda ser tratada con la instrucción que convierte el formato de la matriz. De ahí que haya que generar o renombrar las variables. Si se tiene espacio y memoria suficiente, es preferible la primera opción. Por ello, habría que crear las variables que se quieren comparar con la instrucción *generate*.

```
use panel7, clear  
generate partido1=ideopp  
generate partido2=ideopnv  
generate partido3=ideoea
```

Hay que tener en cuenta que el programa *anova* de medidas repetidas tiene ciertas limitaciones de cálculo. No puede trabajar con más de 800 casos, pues por las características del programa estos son tratados como si fueran valores distintos de una variable. Por eso, es conveniente eliminar del fichero todos aquellos casos con valores perdidos en las variables tratadas y, como aún eso no es suficiente, dada la gran muestra disponible, habrá que construir una submuestra con la instrucción *sample*, en este caso basta con hacer una que comprenda el 50% de los sujetos disponibles. Finalmente, también es conveniente reducir el tamaño del fichero de trabajo para que el transformado no contenga variables inútiles y para que se aminore el esfuerzo de conversión. Estas tres operaciones de selección han de realizarse mediante las siguientes instrucciones:

```
recode partido1-partido3 98 99=.  
drop if partido1==. | partido2==. | partido3==.  
sample 50  
keep id partido1-partido3
```

Obviamente, con las dos primeras, al eliminar los casos con valores perdidos en estas variables, se seleccionan los casos que han contestado a la evaluación de la posición ideológica de los tres partidos considerados; con la tercera, se seleccionan la mitad de los casos y, finalmente —aunque podría haber sido también ubicada en primer lugar—, sólo se mantienen las cuatro variables con las que se obtendrá la nueva disposición de los datos. Una vez realizadas estas operaciones, procede la transformación de la matriz de datos, mediante la instrucción, que en este ejemplo adoptaría la siguiente forma:

```
reshape long partido, i(id) j(par)
```

Una vez introducida esta instrucción, el fichero se prepara automáticamente para que sea posible realizar el análisis de varianza con el diseño de medidas repetidas. Consecuencia del proceso, en pantalla se muestra lo siguiente:

### ILUSTRACIÓN 7.25. Parámetros de la transformación de matrices

```
(note: j = 1 2 3)

Data      wide    ->      long
Number of obs.      670    ->    2010
Number of variables   4      ->     3
j variable (3 values)           ->     par
xij variables:
partido1  partido2      partido3    ->    partido
```

De los datos dispuestos en formato *ancho* se pasa al formato *largo*. De 670 casos se pasa a 2.010, esto es, se multiplica por tres las líneas del fichero; las variables pasan de 4 a 3, es decir, de tener la identificación y tres variables, ahora se tiene la identificación, las tres variables en una sola (partido) y otra nueva variable nominal (par) con valores del 1 al 3, que indican de qué partido se trata la medición<sup>6</sup>.

Antes de efectuar el análisis de varianza de medidas repetidas conviene poner etiquetas a la nueva variable *par* y solicitar una tabla con *tabstat* de los estadísticos básicos para reconocer las evaluaciones que se han dado a los tres partidos en cuestión:

```
label define partidos 1 "PP" 2 "PNV" 3 "EA"
label values par partidos
tabstat partido, by(par) statistics(n mean sd)
```

De este modo se obtienen los estadísticos correspondientes a las tres variables. Cada una de ellas contiene el *n* total. En ese sentido, el *n* de la fila total es ficticio, puesto que está sumando a cada individuo tres veces para obtener la media conjunta de las tres variables:

### ILUSTRACIÓN 7.26. Tabla de medias y desviaciones típicas de muestras dependientes

Summary for variables: partido by categories of: par			
par	mean	sd	N
PP	7.998507	1.829422	670
PNV	5.291045	1.899668	670
EA	4.425373	1.53376	670
Total	5.904975	2.327455	2010

<sup>6</sup> Por todo ello, conviene cerrar el fichero de trabajo después de estos análisis, ya que los casos se multiplican por tantas veces como medidas “repetidas” se disponga.

Tras estos pasos previos con los datos, hay que incluir la orden *anova* para obtener el análisis:

`anova partido id par, repeated (par)`

El resultado es similar al de la instrucción *oneway* del anterior apartado:

### ILUSTRACIÓN 7.27. Análisis de varianza de muestras dependientes

		Number of obs = 2010		R-squared = 0.7137	
		Root MSE = 1.53364		Adj R-squared = 0.5701	
Source	Partial SS	df	MS	F	Prob > F
Model	7844.54378	671	11.6908253	4.97	0.0000
id	3073.60249	669	4.5943236	1.95	0.0000
par	4770.94129	2	2385.47065	1014.20	0.0000
Residual	3147.05871	1338	2.35206181		
Total	10991.6025	2009	5.47118093		
<hr/>					
Between-subjects error term: id					
Levels: 670                         (669 df)					
Lowest b.s.e. variable: id					
<hr/>					
Repeated variable: par					
Huynh-Feldt epsilon = 0.8979					
Greenhouse-Geisser epsilon = 0.8956					
Box's conservative epsilon = 0.5000					
<hr/>					
Source		df	F	Prob > F	
				Regular	H-F
par		2	1014.20	0.0000	0.0000
Residual		1338		0.0000	0.0000
<hr/>				G-G	Box

La tabla de varianza está dividida en seis columnas. La primera indica la fuente de la variación correspondiente a cada fila. La segunda da cuenta de las sumas cuadráticas. La tercera, de los grados de libertad. La cuarta es el cociente entre la segunda y la tercera, esto es, las medias cuadráticas. La quinta son los valores *F*. La sexta, sus correspondientes probabilidades.

Las fuentes de variación tienen equivalencia con las comparaciones en muestras independientes: la del modelo equivale a la externa (*between*); la residual, a la interna (*within*), y la total es, como en el caso anterior, la suma de todas las diferencias al cuadrado de los valores de las tres variables con respecto a la media global de todas.

La variación total, como puede comprobarse fácilmente, es la suma de la que explica el modelo y de la residual. A su vez, la variación del modelo se

descompone en dos factores: la que es explicada por el hecho de que estén siendo evaluados tres partidos diferentes y la que es explicada por cuanto que hay 670 individuos con evaluaciones diferentes. Como hay tantos sujetos, la suma cuadrática es casi tan alta como la de los partidos, pero, una vez hallada la media cuadrática, se ve que las diferencias existentes entre los partidos son bastante mayores que las existentes entre individuos. Esto es así porque existe cierto consenso social sobre la ubicación en la escala ideológica donde están situados los partidos.

Ahora bien, el principal cometido para el que se ha hecho este análisis es el de comparar las medias que los sujetos dan en la escala ideológica a los partidos. Por ello, la fuente de variación más importante es la de la variable *par*, pues es la que instrumenta las diferencias entre las tres originales. El estadístico central para la comparación de las medias es la *F* calculada con la fuente de variación de la variable *par*, en este caso, 1.014,20. Sin embargo, al tratarse de unas medidas que carecen del supuesto de independencia, ya que están emitidas por la misma persona, necesitan una corrección para que se ajusten a la teórica distribución de la *F* de Snedecor. Existen, entre otras medidas correctivas, denominadas  $\epsilon$ , tres que utiliza el programa Stata, la de Feldt, la de Geisser y la de Box. En los tres casos se trata de un número menor o igual a 1, que reduce el tamaño de la *F*, evitando la comisión de un error de tipo I, esto es, de rechazar la hipótesis nula siendo cierta.

En el ejemplo actual, la *F* obtenida es 1.014,20 y los factores de corrección de 0,90, 0,90 y 0,50, respectivamente. En cualquier caso, se puede con tranquilidad rechazar la hipótesis de que la media en la escala ideológica de los tres partidos vascos sea idéntica en la población.

### *7.5.1. Pruebas no paramétricas de comparación de muestras dependientes*

También el análisis de varianza de muestras dependientes requiere que se cumplan los requisitos de normalidad de los datos poblacionales y, sobre todo, de homocedasticidad. Pero, para el supuesto de que estas asunciones no se cumplan, existen otras pruebas estadísticas que no requieren condiciones tan estrictas. En el caso de que se deseen comparar más de dos variables de una misma muestra —a diferencia del caso en el que se trate de cotejar una sola variable en varias muestras— no se puede utilizar la prueba de Kruskall-Wallis, sino la de Friedman.

El programa Stata no incorpora originalmente ningún procedimiento capaz de obtener este estadístico; sin embargo, entre sus librerías disponibles y adquiribles a través de Internet, se encuentra un procedimiento capaz de proporcionar los cálculos necesarios. Se trata del programa *snp2*. Por tanto, si no se tiene aún instalado, es preciso escribir la siguiente instrucción, conectado a Internet.

```
net install snp2.pkg
```

Una vez que ya se ha incorporado este procedimiento en el disco duro del ordenador con el programa Stata, la instrucción necesaria para producir el estadístico de Friedman y el coeficiente de Kendall es la siguiente:

```
friedman lista_de_variables [in rango] [if exp]
```

Este programa adolece de un pequeño defecto: no puede trabajar con variables que contengan valores perdidos. Por tanto, antes de escribir la instrucción hay que asegurarse de que sólo va a trabajar con los casos válidos. Hay diversos modos de acometer esta operación, pero quizás el más cómodo sea creando una variable ficticia e instrumental con la instrucción *mark*, que da el valor 1 a aquellos casos que no tienen ningún caso perdido en la lista de variables señaladas. Esto se logra para las variables del ejemplo anterior (*ideopp*, *ideoea* y *ideopnv*) con las dos siguientes líneas, que generan la variable ficticia *selecciona*:

```
use panel7, clear
mark selecciona
markout selecciona ideopp ideoea ideopnv
```

A continuación, ya puede utilizarse la instrucción *friedman* con el condicionante correspondiente:

```
friedman ideopp ideoea ideopnv if selecciona
```

El resultado proporciona tres líneas que contienen el estadístico de Friedman, su significación y el coeficiente de Kendall.

#### ILUSTRACIÓN 7.28. Prueba de Friedman para muestras dependientes

```
Friedman = 5.4e+03
Kendall = 0.7490
p-value = 0.0000
```

El estadístico de Friedman posee una distribución  $\chi^2$  con el número de grupos menos uno como grados de libertad. Como en este caso el valor es

tan alto, la significación, indicada a través del *p-value*, es bajísima. Por ello puede rechazarse con toda tranquilidad la hipótesis nula de que el rango ideológico que los sujetos dan a los tres partidos sea idéntico. El coeficiente de concordancia de Kendall es una medida que varía entre 0 y 1: cuanto más cerca esté del 1, indica mayores acuerdos entre personas en el juicio efectuado a un objeto, en este caso, a los partidos.

## 7.6. Ejercicios

1. Con los datos del barómetro de marzo de 2009 (cis2794), prueba la hipótesis de que más del 60% de los ciudadanos prefieren ir personalmente a la Oficina de Administración cuando necesitan información relacionada con la gestión de trámites administrativos (P.15). Comprueba asimismo que más del 50% de la población española ha accedido a Internet en los últimos 5 meses.
2. Verifica que la mayoría de los ciudadanos que han hecho gestiones administrativas han quedado satisfechos con la atención (P.23). Esto podría hacerse tanto con medias como con porcentajes. Presta especial atención a excluir a quienes no han emitido opinión por no haber realizado estos trámites.
3. Empleando cualquier barómetro político del CIS (enero-abril-julio-octubre), averigua si hay diferencias estadísticamente en la valoración de los líderes políticos. Por ejemplo, utilizando el de abril de 2009 (cis2798), compara la evaluación de José Luis Rodríguez Zapatero con la de Cayo Lara, la de Rosa Díez y la de Mariano Rajoy.
4. Evalúa si hay diferencias estadísticamente significativas en las medias de los líderes mencionados en el ejercicio anterior o cualquier otro que selecciones, según la valoración la hagan hombres o mujeres, jóvenes (menores de 45 años) o mayores (más de 45) e izquierda (1-5) o derecha (6-10). Aprovecha para recordar la instrucción *recode* vista en el quinto capítulo. Realiza otra agrupación de los valores de las dos últimas variables (*edad* e *ideología*) para comparar más de dos grupos de edad e ideología al mismo tiempo.
5. Agrupa las categorías conservadora/demócrata cristiana/liberal, por un lado, y socialdemócrata/socialista/comunista, por el otro. ¿Hay diferencias significativas en la muestra de abril de 2009 de la sociedad española entre ideologías clásicas de derechas e izquierdas? ¿Son los jóvenes más de izquierda? ¿Son las mujeres más proclives a las ideologías de derechas? (Para facilitar el ejercicio, cuenta sólo con la primera opción de respuesta de la P.16 del barómetro de abril de 2009, cis2798).
6. Prueba la hipótesis de que todas las valoraciones de los líderes nacionistas (Puigcercós, Quintana y Barkos) son iguales.

# 8

## Confección y análisis de tablas con Stata<sup>1</sup>

Probablemente, el método estadístico más ampliamente utilizado (al menos en sociología) sea la tabla de contingencia. Una tabla que muestra un cruce entre dos o más variables es una manera fácil e intuitiva de estudiar la relación entre dos o más variables. Pero esta facilidad de uso e interpretación, pese a ser una de las principales ventajas del análisis de tablas de contingencia, también supone en muchas ocasiones un problema. A menudo el análisis estadístico con tablas se realiza de manera menos cuidadosa y sólida que con otros métodos, sin prestar atención a problemas de significación estadística o a interrelaciones complejas entre variables. Sin embargo, existen herramientas estadísticas que permiten hacer un análisis cuidadoso de tablas de contingencia, estadísticamente riguroso. En este capítulo se abordan en primer lugar las tablas de contingencia propiamente dichas; seguidamente, se tratan otros tipos de tablas que permiten comparar estadísticos distintos de los porcentajes, cuyo tratamiento inferencial se vio con más profundidad en el capítulo anterior, y se finaliza con las tablas propias de variables con valores que no son mutuamente excluyentes.

El análisis de tablas de contingencia está indicado para el estudio de la relación o asociación entre variables cualitativas (nominales u ordinales), y aunque el modelo básico es para dos variables, se puede extender fácilmente a más de dos. En cualquier caso, en el análisis de tablas de contingencia se deben especificar al menos dos variables en orden indiferente, si bien en muchas ocasiones en el terreno teórico puede pensarse que estas dos pueden tener un estatus diferente: una variable dependiente (cuyo comportamiento se intenta explicar), y una o más variables independientes (que se comprobará si correlacionan y en qué medida con el comportamiento de la variable dependiente). Es, por tanto, una técnica

---

<sup>1</sup> Para ampliar conocimientos de este tema se recomienda el libro básico de Sánchez CarrIÓN (1989) o los de Ruiz Maya (1990 y 1995). En inglés, entre otros, son recomendables Everitt (1977), Andersen (1997) como introductorios y, entre los avanzados, Agresti (2002) y Lawal (2003).

que sirve para estudiar, en principio, la asociación entre variables y, por extensión, el efecto (a través de la dependencia entre valores) de una o varias variables sobre otra.

A lo largo de la primera parte de este capítulo se utilizará un mismo ejemplo para ilustrar las explicaciones. Con datos de la encuesta postelectoral del CIS de 2000<sup>2</sup>, se examina el efecto de los ingresos familiares sobre el voto. La idea de partida es que la situación económica del individuo puede afectar a sus preferencias políticas. Existen numerosos estudios sociológicos, tanto en España como en otros países, que estudian tal relación entre situación económica y voto<sup>3</sup>. Al hilo de este ejemplo, se verán las instrucciones necesarias para hacer análisis de tablas de contingencia con Stata y, al mismo tiempo, se ofrecerán las fórmulas y las interpretaciones de los estadísticos que se requieren para un correcto uso de esta técnica.

## 8.1. Tablas de contingencia de dos variables

Las variables que se van a emplear para el análisis son el recuerdo de voto en las últimas elecciones y los ingresos familiares mensuales declarados por el entrevistado en la encuesta postelectoral del CIS de 2000<sup>4</sup>. Es muy conveniente realizar como paso previo una tabla de frecuencias simple para cada una de las variables por separado. Ello se logra de dos formas: bien repitiendo en distintas líneas la instrucción *tabulate* tantas veces como variables se deseen tabular, bien escribiendo la orden *tab1* y a continuación especificando tantas como variables se deseé, puesto que, si se opta por esta forma, el programa entiende que sólo se requieren tablas unidimensionales, por muchas variables que en ella se listen.

```
tab1 rvoto ingresos
```

<sup>2</sup> Estudio CIS 2384. Muestra de 5.283 casos, representativa de la población española de ambos性os, de 18 y más años, realizada mediante entrevista personal en marzo de 2000.

<sup>3</sup> Véase, por ejemplo, González (1995); también Paramio (2000).

<sup>4</sup> En 2000 aún no era el euro la unidad de cambio monetario en España. Las etiquetas de la variable *ingresos* están expresadas en millares de pesetas. Para obtener su correspondiente aproximado en euros, basta con multiplicar por 6. Así 150 (miles de pesetas) equivale a 900 € y 300 a 1.800 €.

### ILUSTRACIÓN 8.1. Distribuciones univariadas de frecuencias

-> tabulation of rvoto				
Recuerdo de	Freq.	Percent	Cum.	
PP	1,773	33.57	33.57	
PSOE	1,054	19.96	53.53	
IU	226	4.28	57.81	
Nacionalista	350	6.63	64.44	
Otros	95	1.80	66.24	
Blanco	108	2.05	68.28	
No voto	890	16.85	85.14	
NC	785	14.86	100.00	
Total	5,281	100.00		

-> tabulation of ingresos				
Ingresos	Freq.	Percent	Cum.	
familiares				
mensuales				
entrevistado				
entre	Freq.	Percent	Cum.	
<150	1,846	34.94	34.94	
150-300	1,449	27.43	62.37	
>300	452	8.56	70.93	
Ns/Nc	1,536	29.07	100.00	
Total	5,283	100.00		

Como se aprecia en la ilustración 8.1, la primera variable, *rvoto*, es una variable nominal con 5.281 casos, un 14,86% de los cuales “No saben/No contestan”. *Ingresos*, por su parte, es ordinal con tres categorías y un porcentaje de “No sabe/No contesta” aún mayor, de más de un 29%. El primer problema que se plantea es qué hacer con los “No sabe/No contesta”. No se suelen incluir en el análisis a menos que se quieran estudiar específicamente, puesto que, aunque aportan información, no es del tipo que interesa normalmente en función de las hipótesis, lo que también es el presente caso. El “Ns/Nc” a la pregunta de ingresos indica simplemente que no se quieren declarar los ingresos, mientras que el “Ns/Nc” a la pregunta de voto puede indicar o bien lo mismo o falta de competencia (percibida o real) para hablar de temas políticos. No se incluirá en consecuencia el “Ns/Nc” en el análisis, aunque aparecerá en las primeras tablas para explorar qué tipo de sesgos tienen los “Ns/Nc” con respecto a las variables estudiadas (por ejemplo, si la no respuesta se da más en unos ingresos determinados).

El comando de Stata para tablas de contingencia es *tabulate*<sup>5</sup>. Su sintaxis general es:

<sup>5</sup> Es el mismo comando que para tablas de frecuencias. Si se escribe *tabulate* y el nombre de una variable, Stata entenderá que se desea mostrar su tabla de frecuencias y así lo hará. Si se expresa la instrucción *tabulate* seguida por los nombres de dos variables, Stata mostrará una tabla de contingencia.

```
tabulate variable1 variable2 [if expresión] [in rango] [aweight=varpeso]
[, opciones]
```

Tras la palabra *tabulate*<sup>6</sup>, en variable1 se ha de escribir el nombre de la variable dependiente (la que se trata de explicar), que aparecerá en las filas de la tabla; mientras que en variable2 ha de expresarse el nombre de la variable independiente (la que se supone que explica el comportamiento de la anterior), que aparecerá en las columnas. Optativamente, la orden de tabulación se puede acompañar con la expresión de la cláusula *if*, que puede de ser numérica o lógica; el rango tras *in*, que se refiere a un conjunto de casos contiguos definidos mediante el caso inferior y el superior separados por una barra, y se permite la ponderación analítica, frecuencial (*fweight*), analítica (*aweight*) o discrecional (*iweight*), que ha de ir acompañada por la variable por la que se pondrá precedida por el signo igual. Finalmente, tras una coma, pueden incluirse las opciones que se irán explicando en las próximas páginas.

En el ejemplo actual, la variable dependiente es *rvoto* y la independiente *ingresos*, por lo que para generar una tabla de contingencia sencilla sólo hay que escribir...

```
tabulate rvoto ingresos
```

... para que aparezca la siguiente tabla:

### ILUSTRACIÓN 8.2. Distribución bivariada de frecuencias

Recuerdo de voto recodificado	Ingresos familiares mensuales entrevistado				Total
	<150	150-300	>300	Ns/Nc	
PP	613	552	183	425	1,773
PSOE	477	293	79	205	1,054
IU	69	90	26	41	226
Nacionalista	86	108	37	119	350
Otros	16	30	23	26	95
Blanco	31	30	16	31	108
No voto	286	230	61	313	890
NC	268	114	27	376	785
Total	1,846	1,447	452	1,536	5,281

<sup>6</sup> Del mismo modo que se permitía la instrucción *tab1* listavar, también existe *tab2* listavar. Esta hace cruzar todas las variables que se incluyan en la lista de variables. Si se escribe con la opción *firstonly*, sólo se cruza la primera con todas las demás.

Aparece en la ilustración 8.2 una tabla con las frecuencias cruzadas de recuerdo de voto e ingresos familiares. Aunque esta es ya propiamente una tabla de contingencia, los datos mostrados resultan difíciles de interpretar. Además de conocer las frecuencias, es preciso contar con los porcentajes para estudiar adecuadamente la relación entre ambas variables.

¿Pero cómo se calculan los porcentajes? Hay tres maneras de hacerlo. Si se representan las frecuencias de una tabla de contingencia, del modo en el que aparecen en el cuadro 8.1:

**CUADRO 8.1. Notación de las tablas de contingencia**

		Variable independiente (x)			
		columna 1	columna 2	columna 3	Total
Variable dependiente (y)	Fila 1	$f_{11}$	$f_{12}$	$f_{13}$	$f_{1..}$
	Fila 2	$f_{21}$	$f_{22}$	$f_{23}$	$f_{2..}$
	Total	$f_{..1}$	$f_{..2}$	$f_{..3}$	$n = f_{..}$

Las tres modalidades de porcentajes se pueden calcular del siguiente modo:

1. *Porcentaje de columna:* el porcentaje que representa la frecuencia de una celda sobre el total de la columna:

$$p_{i|j} = \frac{f_{ij}}{f_{..j}} \times 100 \quad (8.1)$$

Para que los muestre Stata se especifica la opción *col* tras la coma:

```
tabulate rvoto ingresos, col
```

... con lo que aparecen los datos de la ilustración 8.3.

**ILUSTRACIÓN 8.3. Tabla de contingencia con frecuencias  
y porcentajes verticales**

		Ingresos familiares mensuales entrevistado				
		<150	150-300	>300	Ns/Nc	Total
Recuerdo de voto	recodificado	613	552	183	425	1,773
		33.21	38.15	40.49	27.67	33.57
PSOE		477	293	79	205	1,054
		25.84	20.25	17.48	13.35	19.96
IU		69	90	26	41	226
		3.74	6.22	5.75	2.67	4.28
Nacionalista		86	108	37	119	350
		4.66	7.46	8.19	7.75	6.63
Otros		16	30	23	26	95
		0.87	2.07	5.09	1.69	1.80
Blanco		31	30	16	31	108
		1.68	2.07	3.54	2.02	2.05
No voto		286	230	61	313	890
		15.49	15.89	13.50	20.38	16.85
NC		268	114	27	376	785
		14.52	7.88	5.97	24.48	14.86
Total		1,846	1,447	452	1,536	5,281
		100.00	100.00	100.00	100.00	100.00

Como puede apreciarse, esta tabla expresa no sólo la frecuencia, sino también el porcentaje de casos que hay en cada casilla sobre el total de cada categoría de la variable *ingresos*. Así, el 33,2% de los entrevistados que tenían ingresos inferiores a 150.000 ptas. votaron al PP; el 25,8, al PSOE; el 3,7%, a IU, etc.

2. *Porcentaje de filas*: el porcentaje que representa la frecuencia de cada casilla sobre el total de la fila:

$$p_{j|i} = \frac{f_{ij}}{f_i} \times 100 \quad (8.2)$$

En Stata se obtiene mediante la opción *row* de la instrucción *tabulate*. Como también puede hacerse con los porcentajes verticales, si se desea que no aparezcan las frecuencias absolutas de las casillas, es preciso añadir la opción *nofreq*, como se pone de manifiesto en el siguiente ejemplo.

```
tabulate rvoto ingresos, row nofreq
```

De este modo, se obtiene una tabla más compacta con lectura más fácil de los porcentajes horizontales.

#### ILUSTRACIÓN 8.4. Tabla de contingencia con porcentajes horizontales

Recuerdo de	Ingresos familiares mensuales entrevistado				
recodificado	<150	150-300	>300	Ns/Nc	Total
PP	34.57	31.13	10.32	23.97	100.00
PSOE	45.26	27.80	7.50	19.45	100.00
IU	30.53	39.82	11.50	18.14	100.00
Nacionalista	24.57	30.86	10.57	34.00	100.00
Otros	16.84	31.58	24.21	27.37	100.00
Blanco	28.70	27.78	14.81	28.70	100.00
No voto	32.13	25.84	6.85	35.17	100.00
NC	34.14	14.52	3.44	47.90	100.00
Total	34.96	27.40	8.56	29.09	100.00

En la ilustración 8.4 se lee que el 34,6% de los que votaron al PP tienen ingresos inferiores a 150.000 ptas., el 31,1% tienen ingresos de 150.000 a 300.000, etc.

3. *Porcentaje total:* el porcentaje de la frecuencia de cada casilla sobre el total de los casos de la tabla:

$$p_{ij} = \frac{f_{ij}}{f_{..}} \times 100 \quad (8.3)$$

Su obtención es análoga a las anteriores con la opción *cell*. Por ello en el ejemplo actual la instrucción debería ser...

```
tabulate rvoto ingresos, cell nofreq
```

... para obtener el resultado de la ilustración 8.5.

**ILUSTRACIÓN 8.5. Tabla de contingencia con porcentajes totales**

Recuerdo de voto	Ingresos familiares mensuales entre entrevistado recodificado	<150	150-300	>300	Ns/Nc	Total
PP	11.61	10.45	3.47	8.05	33.57	
PSOE	9.03	5.55	1.50	3.88	19.96	
IU	1.31	1.70	0.49	0.78	4.28	
Nacionalista	1.63	2.05	0.70	2.25	6.63	
Otros	0.30	0.57	0.44	0.49	1.80	
Blanco	0.59	0.57	0.30	0.59	2.05	
No voto	5.42	4.36	1.16	5.93	16.85	
NC	5.07	2.16	0.51	7.12	14.86	
Total	34.96	27.40	8.56	29.09	100.00	

La lectura en este caso es que el 11,6% de todos los encuestados son votantes del PP con ingresos inferiores a 150.000 ptas.

Las tres formas de calcular los porcentajes expresan los mismos datos, pero cada una acentúa un distinto aspecto del cruce y una manera de comparar las distribuciones. El porcentaje de columnas permite comparar el comportamiento de la variable dependiente en las diferentes categorías de la independiente. El de fila muestra la distribución de frecuencias de la dependiente para las categorías de la independiente y el porcentaje sobre el total permite estudiar la distribución conjunta de ambas variables. El porcentaje más utilizado es el de columna, pues permite ver el efecto de la variable independiente sobre el comportamiento de la dependiente. Si los porcentajes de columna son muy distintos en las distintas categorías de la variable independiente, habrá indicios de asociación estadística entre las variables.

De este modo, en este ejemplo, se detecta en la ilustración 8.3 cómo los ingresos parecen tener influencia sobre el voto. Los que dicen tener más ingresos votan más al PP que los que tienen ingresos medios y bajos; los que tienen ingresos inferiores votan más al PSOE que los de ingresos medios y altos, y los que declaran ingresos medios votan más a IU que los que tienen ingresos altos y bajos. Con respecto a los “No sabe/No contesta”, aumentan conforme baja el nivel de ingresos. Más allá de constatar este hecho, no interesa incluirlos en el análisis, por lo que es conveniente generar una nueva tabla que no incluya los “No sabe/No contesta”:

```
tabulate rvoto ingresos if rvoto != 99 & ingresos != 9, col norefq
```

La expresión *if* quita los casos con “No sabe/No contesta” de ambas variables. Ahora se ve más claramente la relación entre ambas variables: el voto al PP y a los nacionalistas aumenta ligeramente con el nivel de ingresos, el voto al PSOE y la abstención aumentan según disminuyen los

ingresos, y el voto a IU es algo mayor en los ingresos medios. Parece, por tanto, que sí existe relación entre situación económica y voto.

**ILUSTRACIÓN 8.6. Tabla de frecuencias con porcentajes verticales sin valores perdidos**

Recuerdo de voto recodificado	Ingresos familiares mensuales entrevistado			Total
	<150	150-300	>300	
PP	38.85	41.41	43.06	40.41
PSOE	30.23	21.98	18.59	25.45
IU	4.37	6.75	6.12	5.55
Nacionalista	5.45	8.10	8.71	6.92
Otros	1.01	2.25	5.41	2.07
Blanco	1.96	2.25	3.76	2.31
No voto	18.12	17.25	14.35	17.30
Total	100.00	100.00	100.00	100.00

Además de comparar los distintos porcentajes entre sí para ver el efecto de la variable independiente sobre la dependiente, es interesante comparar los porcentajes de columna de las categorías con el porcentaje de columna total, que se llama *marginal* (el que aparece en la última columna de la derecha). Si no hubiera relación alguna entre las variables, los porcentajes de columna de las categorías deberían ser iguales o muy parecidos a los marginales de columna. En este caso, puede comprobarse que no es así: los porcentajes de las casillas se distancian sensiblemente de sus marginales.

La comparación entre los porcentajes de columna y los marginales sirve como una primera aproximación al concepto de independencia en tablas de contingencia. Buena parte de los estadísticos de significación y asociación estadística se basan en la comparación entre las *frecuencias observadas* de las casillas y las *frecuencias esperadas* en caso de independencia, esto es, si no hubiera relación entre las variables (la variable dependiente sería —aunque parezca un juego de palabras— independiente de la variable independiente).

¿Cómo se construyen estas *frecuencias esperadas*? Para calcular cuál sería la frecuencia de cada casilla si no hubiera relación entre las variables, se multiplica el número de casos marginal de cada fila por el número de casos marginal de cada columna y se divide por el número de casos total de la tabla. Utilizando la nomenclatura de la tabla 2, la frecuencia esperada se calcula operativamente del siguiente modo:

$$f_{ij}^* = \frac{f_i \cdot f_j}{f_{..}} \quad (8.4)$$

En la ilustración 8.7, la frecuencia esperada de la primera casilla 1-1 (“Votó PP” e “ingresos menores a 150.000 ptas.”) sería:

$$f_{11}^* = \frac{f_{1.} f_{.1}}{f_{..}} = \frac{1348 \times 1578}{3336} = 637,6 \quad (8.5)$$

Es decir, si no hubiera relación entre voto e ingresos, el número de personas que votó al PP con ingresos inferiores a 150.000 ptas. debería ser 637,6. Como el valor observado es distinto (613), hay indicios de cierta relación entre las variables. La diferencia entre el valor observado y el esperado se llama *residuo*, que manifiesta dependencia entre pares de valores de las variables respectivas, siempre y cuando su valor difiera de 0. El residuo de la casilla 1-1 sería: 613-637,6 = -24,6, que indica que la frecuencia de la casilla es menor en 24,6 casos a la esperable en caso de independencia.

La siguiente instrucción construye una tabla de contingencia (ilustración 8.7) con las frecuencias observadas y esperadas:

```
tabulate rvoto ingresos if rvoto != 99 & ingresos != 9, expected
```

**ILUSTRACIÓN 8.7. Tabla de contingencia con frecuencias observadas y esperadas**

		Ingresos familiares mensuales entrevistado			Total
		<150	150-300	>300	
RECODE of p34	Key	frequency	expected frequency		
PP		613 637.6	552 538.6	183 171.7	1,348 1,348.0
PSOE		477 401.6	293 339.2	79 108.2	849 849.0
IU		69 87.5	90 73.9	26 23.6	185 185.0
Nacionalistas		86 109.3	108 92.3	37 29.4	231 231.0
Otros		16 32.6	30 27.6	23 8.8	69 69.0
Blanco		31 36.4	30 30.8	16 9.8	77 77.0
No votó		286 272.9	230 230.6	61 73.5	577 577.0
Total		1,578 1,578.0	1,333 1,333.0	425 425.0	3,336 3,336.0

Como puede apreciarse, el segundo valor de cada casilla es el valor esperado (el de la primera casilla es el mismo que se ha calculado manualmente más arriba). La diferencia entre la frecuencia observada y la esperada indica, como ya se ha señalado, la relación entre las variables. Así, el número

de no votantes de ingresos bajos es superior al esperable si no hubiera relación entre las variables, el de ingresos medios es igual al esperado y el de ingresos altos, inferior al esperado, sugiriendo una relación lineal entre las variables (a más ingresos, menor abstención).

El comando *tabulate* de Stata no permite obtener los residuos. Para trabajar con ellos, ha de instalarse un módulo *.ado* de Stata creado por Nick Cox (1999). Ha de instalarse con conexión a Internet, mediante las siguientes instrucciones:

```
net from http://www.stata.com/users/njc
net install tab_chi
```

De este modo, se instala el programa *tabchi*, especialmente diseñado para trabajar con residuos de tablas de contingencia.

Con el uso de opciones, *tabchi* permite analizar los residuos. La opción *raw* muestra el residuo “crudo” ( $r_{ij}$  frecuencia observada menos frecuencia esperada); *pearson* ( $r_{ij}^*$ ) muestra los residuos de Pearson, a veces llamados residuos estandarizados (residuo neto dividido por la raíz cuadrada de la frecuencia esperada); *cont* muestra la contribución de cada casilla al  $\chi^2$  de la tabla (observada menos esperada al cuadrado dividido por esperada, es decir, el residuo de Pearson al cuadrado), y *adj* muestra los residuos ajustados (residuos de Pearson divididos por su desviación típica).

$$\begin{aligned} r_{ij} &= f_{ij} - f_{ij}^* \\ r_{ij}^s &= \frac{r_{ij}}{\sqrt{f_{ij}^*}} \\ r_{ij}^a &= \frac{r_{ij}^s}{\sqrt{\left(1 - \frac{f_{i.}}{f_{..}}\right) \left(1 - \frac{f_{.j}}{f_{..}}\right)}} \end{aligned} \tag{8.6}$$

Véase ahora una tabla con estas opciones, añadiéndole también *noo* y *noe*, para eliminar las frecuencias observadas y esperadas que ya se han obtenido en la ilustración 8.7. La instrucción se compondría de las siguientes opciones:

```
tabchi rvoto ingresos if rvoto!=99 & ingresos !=9 , noo noe raw pearson adjust
```

En cada casilla de la ilustración 8.8 aparecen todos los estadísticos pedidos. El primero es el residuo, que simplemente expresa la distancia entre el valor observado y el esperado de la casilla. El segundo es el residuo de Pearson, que no tiene mucho interés en este caso. Y, por último, aparece el

residuo ajustado, que es el más interesante para la interpretación. Está estandarizado para poder estudiar la significación estadística de la frecuencia de cada casilla. El valor que toma sigue una distribución normal con media 0 y desviación típica 1 ( $N[0,1]$ ), por lo que puede utilizarse para comprobar en la tabla de probabilidades de la normal si el valor del residuo es significativo o se puede deber a errores de muestreo. En general, si supera 1,96 en términos absolutos (negativo o positivo), puede decirse que la diferencia entre el valor observado y el esperado en caso de no haber relación entre las variables no es debido a errores de muestreo, con un 95% de confianza. En este caso, las casillas de voto al PP no son significativas (no llegan al valor crítico de 1,96), por lo que las diferencias observadas en voto al PP por ingresos pueden ser debidas a errores de muestreo y no deben ser tomadas en consideración. Las diferencias en voto al PSOE, en cambio, sí que son significativas, así como las de IU y nacionalistas para ingresos medios y bajos.

### **ILUSTRACIÓN 8.8. Tabla de contingencia con residuos brutos, estandarizados y ajustados**

	raw residual	Pearson residual	adjusted residual
Recuerdo de voto recodificado	Ingresos familiares mensuales entrevistado		
	<150	150-300	>300
PP	-24.633 -0.976 -1.741	13.366 0.576 0.963	11.267 0.860 1.192
PSOE	75.405 3.763 6.003	-46.244 -2.511 -3.753	-29.161 -2.804 -3.476
IU	-18.509 -1.979 -2.804	16.078 1.870 2.483	2.431 0.501 0.552
Nacionalista	-23.268 -2.226 -3.178	15.697 1.634 2.186	7.571 1.396 1.549
Otros	-16.638 -2.912 -4.054	2.429 0.463 0.603	14.210 4.793 5.184
Blanco	-5.423 -0.899 -1.252	-0.768 -0.138 -0.181	6.190 1.976 2.141
No voto	13.067 0.791 1.198	-0.558 -0.037 -0.052	-12.509 -1.459 -1.717

El análisis de los residuos, por tanto, sugiere que el voto al PSOE, a IU y a los partidos nacionalistas se vio afectado por el nivel de ingresos, mientras que el voto al PP, la abstención y el voto en blanco no presentaron pautas especiales según ingresos de modo significativo.

Además del análisis de la significación de cada casilla, puede interesar comprobar si la relación global entre voto e ingresos es significativa. Para ello se utiliza el estadístico  $\chi^2$ , que se calcula a partir de los residuos. Su fórmula es:

$$\chi^2 = \sum_{j=1}^J \sum_{i=1}^I \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*} \quad (8.7)$$

El valor de  $\chi^2$  será 0 cuando no haya relación alguna entre las dos variables, y aumentará cuanto mayor sea la relación. No obstante, no se suele utilizar como indicador de la fuerza de la asociación, porque no varía entre 0 y 1 (más adelante se estudian estadísticos que sí lo hacen y son por ello más adecuados para estudiar la fuerza de la asociación). De mayor interés que el valor del  $\chi^2$  es su significación. El valor de  $\chi^2$  de una tabla tiene una distribución de probabilidad conocida en función de sus grados de libertad, de modo que conociendo ambos valores ( $\chi^2$  y grados de libertad) puede comprobarse la significación estadística de la diferencia global entre frecuencias observadas y esperadas en una tabla de contingencia, y por tanto comprobar si efectivamente existe relación entre las variables.

Los grados de libertad se calculan siguiendo la fórmula:  $gl = (\text{número de filas} - 1) * (\text{número de columnas} - 1)$ . En el ejemplo actual:  $gl = (7-1) * (3-1) = 12$ . El valor de  $\chi^2$ , que aparece debajo de la tabla generada con la orden *tabchi*, es de 86,73. Un  $\chi^2$  de 86,73 con 12 grados de libertad tiene una significación estadística altísima, con probabilidad de que se deba a errores de muestreo inferior a 0,001 (este valor aparece en la penúltima fila de la ilustración 8.8, a la derecha del valor de  $\chi^2$ ). Se puede, por tanto, descartar la hipótesis nula de que las diferencias observadas se deban a errores de muestreo. La relación entre voto e ingresos es estadísticamente significativa.

Debajo de la prueba de  $\chi^2$  aparece la razón de verosimilitud (*likelihood-ratio*  $\chi^2$ ), que es una corrección del  $\chi^2$  que se usa para modelos log-lineales. En muestras pequeñas difiere del valor de  $\chi^2$ , pero según aumenta el tamaño de la muestra ambos valores tienden a converger. Su fórmula es la siguiente:

$$L^2 = 2 \sum_{j=1}^J \sum_{i=1}^I f_{ij} \ln \frac{f_{ij}}{f_{ij}^*} \quad (8.8)$$

La prueba de  $\chi^2$  y la de razón de verosimilitud permiten asegurar con un alto nivel de confianza que la relación que se observa en las tablas entre voto e ingresos es estadísticamente significativa. Para que Stata calcule ambos estadísticos, habrá que especificar en la instrucción *tabulate* las opciones *chi2* y *lrchi2*. Pero el que una relación sea estadísticamente significativa no implica que sea importante, simplemente que no se debe a errores de muestreo. Para estudiar la fuerza de la relación debe recurrirse a otros estadísticos: los de *asociación*. Para ello, puede también utilizarse la instrucción *tabulate* de Stata, que incorpora tres estadísticos de asociación: la *V* de Cramer, el coeficiente *γ* de Goodman y Kruskall, y la  $\tau_b$  de Kendall. Para que Stata muestre estos estadísticos, han de introducirse, respectivamente, las opciones *V* (en mayúscula), *gamma* y *taub*. También pueden pedirse todos (los de significación y los de asociación) mediante la opción *all* tras la habitual instrucción (ilustración 8.10):

```
tabulate rvoto ingresos if rvoto != 99 & ingresos != 9, nofreq chi2 V
```

Por otro lado, al incluir la opción *nofreq*, no se muestra la tabla de contingencia, sólo aparecen los estadísticos expresamente solicitados, como se puede comprobar en la ilustración 8.9.

#### **ILUSTRACIÓN 8.9. Prueba de $\chi^2$ y V de Cramer de una tabla de contingencia**

Pearson chi2(12) =	86.7344	Pr = 0.000
Cramer's V =	0.1140	

La *V* de Cramer es un estadístico de asociación basado en el  $\chi^2$ , que está especialmente indicado para variables nominales. Varía entre 0 y 1, siendo 0 ninguna asociación y 1 asociación perfecta<sup>7</sup>. Su fórmula es:

$$V = \sqrt{\frac{\chi^2}{n \min(I - 1, J - 1)}} \quad (8.9)$$

La *V* de Cramer lo que hace es convertir el  $\chi^2$  en un indicador de asociación dividiéndolo por el valor máximo que puede tomar (el tamaño de la muestra multiplicado por el mínimo número de filas o columnas menos 1).

<sup>7</sup> Una asociación perfecta significa que todos los valores de la tabla se encuentran en una diagonal: cada valor en la variable independiente se corresponde con un valor único en la independiente (el resto de las casillas será 0). En este caso, el valor de la variable independiente determina de manera absoluta el valor de la variable dependiente.

En este ejemplo,  $V$  toma el valor 0,11, lo que corresponde a un nivel de asociación bastante bajo. En la práctica, casi cualquier estadístico de asociación para datos de encuesta se puede considerar bajo si es menor que 0,15; moderado si está entre 0,15 y 0,30, y alto si es superior a 0,30. Por tanto, según la  $V$  de Cramer, la asociación entre nivel de ingresos y voto es baja, aunque significativa.

Tanto  $\gamma$  como  $\tau_b$  son estadísticos de asociación para variables ordinales. No sólo miden la fuerza de la asociación entre dos variables, sino también su dirección. En este caso no se pueden usar porque *voto* no es una variable ordinal, sino nominal. Pero ya que la asociación entre voto e ingresos no parece ser demasiado importante, sería conveniente y relevante poner como segundo ejemplo la relación entre ideología e ingresos. En este contexto<sup>8</sup>, *ideología* es una variable ordinal si sólo se toman los valores válidos (izquierda-centro-derecha), y por tanto servirá de ejemplo para explicar  $\gamma$  y  $\tau_b$ . Para obtener la tabla, se introduce la siguiente instrucción:

```
tabulate ideologia ingresos if ideologia !=99 & ingresos !=9, col all
```

Y el programa ofrece tanto la tabla como todos los estadísticos implementados.

#### ILUSTRACIÓN 8.10. Estadísticos de la tabla de contingencia

Ideología	Ingresos familiares mensuales				
entrevista	entrevistado				
do	Menos de	De 150.00	Más de 30	Total	
Izquierda	525	514	178	1217	
	38.29	39.75	41.40	39.33	
Centro	607	574	194	1375	
	44.27	44.39	45.12	44.44	
Derecha	239	205	58	502	
	17.43	15.85	13.49	16.22	
Total	1371	1293	430	3094	
	100.00	100.00	100.00	100.00	
 Pearson chi2(4) = 4.2793 Pr = 0.370					
likelihood-ratio chi2(4) = 4.3678 Pr = 0.359					
Cramer's V = 0.0263					
gamma = -0.0462 ASE = 0.027					
Kendall's tau-b = -0.0284 ASE = 0.016					

<sup>8</sup> En los cuestionarios del CIS la ideología suele medirse en una escala del 1 (extrema izquierda) al 10 (extrema derecha) presentada a los individuos: los números no significan sino un orden en el supuesto continuo izquierda/derecha. Para simplificar en este capítulo, se han agrupado las respuestas 1-3 como izquierda, 4-7 como centro y 8-10 como derecha. Obviamente, si se preguntara por la ideología concreta que se profesa (liberal, conservadora, socialista, comunista, anarquista) tendría carácter nominal. De modo discutible, a veces también se usa esta valoración ideológica como variable cuantitativa, calculando medias y desviaciones típicas.

Tanto  $\gamma$  como  $\tau_b$  varían entre -1 y +1. 0 indica que no existe asociación; 1, que existe una asociación perfecta positiva (cuando aumenta el valor de una variable también lo hace el de la otra), y -1, una asociación perfecta negativa (cuando aumenta el valor de una disminuye el de la otra). Es curioso cómo en este caso la asociación es muy diferente a la que aparecía entre ingresos y voto (los porcentajes parecen que indican que a más ingresos, más ideología de izquierdas). De todos modos, hay que fijarse en la prueba de  $\chi^2$ , que muestra que la asociación que aparece en la tabla no es significativa. No obstante, se explicará a continuación qué indica y cómo se calcula  $\gamma$  para entenderlo mejor. El signo negativo de  $\gamma$  expresa que la relación es inversa: cuando se sube en la escala de ingresos, la ideología tiende más hacia la izquierda (valores altos en ingresos se corresponden con valores bajos en ideología, que son los de izquierda). De todos modos, la relación es bajísima (0,04 en *gamma*) y, como ya se ha señalado, no significativa. El coeficiente  $\tau_b$  muestra esencialmente lo mismo, con valores más bajos, pues hace una medición más conservadora de la asociación, como se puede deducir a continuación de la comparación de sus fórmulas.

Ambos coeficientes ordinales proceden del cálculo de los pares posibles de valores. Para entender esto, se va a suponer una tabla de contingencia *ordinal* como la esquematizada en el cuadro 8.2:

**CUADRO 8.2. Croquis de tabla de contingencia ordinal**

		Variable independiente (x)		
		1	2	3
Variable dependiente (y)	1	a	b	c
	2	d	e	f
	3	g	h	i

Todo caso que esté en la casilla (a) es inferior tanto en la variable independiente (en la que vale 1) como en la dependiente (en la que vale 1) a cualquier otro caso situado en las casillas (e), (f), (h) o (i). A este tipo de pares (en los que la relación entre ambos es concordante en ambas variables, menor [o mayor] en *x*, también menor [o mayor] en *y*) se les llama pares *concordantes*. Por el contrario, un caso que esté en la casilla (g) tendrá un valor inferior en la variable independiente pero superior en la dependiente a cualquier caso que esté en las casillas (b), (c), (e) o (f). Dicho con más concisión, un individuo en (g) posee un 3 en la variable dependiente, mientras que las personas situadas en (b) y (c) tienen en la variable dependiente (y) valores inferiores, pues en esta son iguales a 1, y en el caso de (e) y (f) también son inferiores en dicha variable, pues

tienen 2; por el contrario, en la variable independiente (g) vale 1 y (b), (c), (e) y (f) más de 1. Todo par de casos de este tipo se llama par *discordante*.

Pues bien,  $\gamma$  lo que hace es contar, en función del número de casos de cada casilla, el número total de pares de casos concordantes y discordantes para medir la relación entre ambas variables del siguiente modo:

$$\gamma = \frac{P_c - P_d}{P_c + P_d} \quad (8.10)$$

... siendo  $P_c$  los pares concordantes, que se obtienen sumando todos los productos de las frecuencias de cada casilla por la suma de todas las frecuencias de casillas que se encuentren al mismo tiempo debajo y a la derecha. Siguiendo el cuadro 8.2, hay que sumar  $a(e+f+h+i)$ ,  $b(f+i)$ ,  $d(h+i)$  y  $e i$ . Por otro lado,  $P_d$  son los pares discordantes, que se calculan sumando todos los productos de las frecuencias de cada casilla por la suma de todas las frecuencias de casillas que se encuentren al mismo tiempo arriba y a la derecha. Es decir, en este caso hay que sumar  $g(b+c+e+f)$ ,  $h(c+f)$ ,  $d(b+c)$  y ec.

El denominador ( $P_c + P_d$ ) es el número total de pares de casos en los que puede existir relación ordinal entre las variables<sup>9</sup> y el numerador ( $P_c - P_d$ ) es el número de pares en los que existe relación positiva menos el número de pares en los que existe relación negativa. Por tanto, el signo de  $\gamma$  indica el tipo de pares que predomina en la tabla, y su valor expresa qué porcentaje representa este predominio en el total de pares estrictamente ordinales. En este caso, predominan ligeramente los pares de valores con una relación negativa (en los que la variable *ingresos* toma un valor alto e *ideología*, uno bajo, y viceversa), por lo que el numerador es negativo, y por tanto también lo es  $\gamma$ . Por otra parte, el valor de  $\gamma$  es muy bajo porque el número de pares concordantes y discordantes es casi igual, por lo que el numerador es casi 0 al contrarrestarse unos a otros. En efecto, si en una tabla no hubiera relación alguna entre variables, el número de pares concordantes sería igual que el número de pares discordantes, por lo que al restarse el resultado sería 0. Para mayor claridad, a continuación se dan tres ejemplos distintos de relaciones en tabla ordinal que muestran resultados muy diferentes:

---

<sup>9</sup> Estrictamente hablando, el número total de pares que se pueden formar en una tabla de  $n$  casos es  $n(n-1)/2$ , cifra esta siempre mayor, salvo en casos excepcionales, que la suma de pares concordantes y discordantes. Esta última, pares estrictamente ordinales, sólo tiene en cuenta las parejas de casos que no tienen un valor idéntico entre sí en una u otra variable.

### ILUSTRACIÓN 8.11. Tablas para distintos valores de $\gamma$

Ejemplo de Gamma = 0 (ninguna relación entre variables):

	1	2
1	4	4
2	4	4

$$\text{Gamma} = \gamma = \frac{N_c - N_d}{N_c + N_d} = \frac{16 - 16}{16 + 16} = \frac{0}{32} = 0$$

Ejemplo de Gamma = 1 (relación perfecta positiva):

	1	2
1	4	0
2	0	4

$$\text{Gamma} = \gamma = \frac{N_c - N_d}{N_c + N_d} = \frac{16 - 0}{16 + 0} = \frac{16}{16} = 1$$

Ejemplo de Gamma = -1 (relación perfecta negativa):

	1	2
1	0	4
2	4	0

$$\text{Gamma} = \gamma = \frac{N_c - N_d}{N_c + N_d} = \frac{0 - 16}{0 + 16} = \frac{-16}{16} = -1$$

$\tau_b$  se construye igual que  $\gamma$ , sólo que en el denominador incorpora una corrección para los pares empasados en una variable ( $P_y$ ) y en otra ( $P_x$ ). Su fórmula es:

$$\tau_b = \frac{P_c - P_d}{\sqrt{(P_c + P_d + P_x)(P_c + P_d + P_y)}} \quad (8.11)$$

Su valor es siempre menor que el de  $\gamma$  al incorporar más pares en el denominador. Los que incorpora son los pares empasados en  $x$  ( $P_x$ ), que son individuos que, teniendo distinto valor en la variable dependiente, sin embargo tienen el mismo en la independiente; mientras que los empasados en  $y$  ( $P_y$ ) son los que poseen valores desiguales en la independiente pero iguales en la dependiente. Es decir, tanto unos como otros podrían haber sido congruentes o incongruentes, pero se quedaron en mitad del camino porque en una variable presentan valores idénticos<sup>10</sup>.

---

<sup>10</sup> Dicho con otras palabras, son pares empasados en una variable aquellos que tienen un mismo valor en esa variable y un valor superior o inferior en la otra. Por ejemplo, volviendo al cuadro 8.2, el par formado por un caso que estuviera en la casilla (a) y otro que estuviera en la casilla (d) sería un par empasado en la variable independiente ( $Tx$ ). O sea, que ambos casos valen 1 en  $x$ , pero (a) es inferior en  $y$  que (d).

## 8.2. Más de dos variables

En muchas ocasiones, los resultados de una tabla de contingencia de dos variables están mediados por el efecto de una tercera variable, o de varias más, desdibujándose o mostrándose así de forma incorrecta la relación entre las variables de interés. En estos casos se debe tratar de encontrar estas tercera variables e incluirlas en las tablas, controlando así su efecto. En este caso, se hablará de tablas multivariadas de contingencia. En el ejemplo anterior se ha visto que el efecto de los ingresos familiares sobre el voto es débil, aunque existe. Dado que la variable de ingresos está referida al conjunto de los obtenidos en un hogar, dos personas con un mismo nivel de ingresos familiares pueden variar en su situación económica según su edad, puesto que los ingresos no se distribuyen de manera absolutamente equitativa dentro de las familias. También dos personas con un mismo nivel de ingresos pero distinta edad pueden tener distinta actitud política porque el ciclo de vida puede determinar la percepción de responsabilidad, por ejemplo. En cualquier caso, el posible efecto de la edad sobre la relación entre clase y voto es una hipótesis que se debería contrastar haciendo análisis de tablas multivariadas de contingencia con Stata.

La manera de hacerlo es utilizando la preinstrucción *bysort* seguida por la(s) variable(s) de control, dos puntos y la orden específica de la tabla. Como ya se vio en capítulo anterior, *bysort* permite ejecutar una misma orden para diferentes categorías de una o más variables especificadas al tiempo que ordena los datos para poderlo realizar. Como puede apreciarse en el ejemplo más adelante, el comando en sí que se emplea para generar la tabla es *tabulate*, el mismo que se emplea para tablas de dos variables, pero precedido por *bysort*, el nombre de la variable de control y los dos puntos. Stata muestra así una tabla de contingencia de voto por ingresos separada *para cada una de las categorías de edad*. Por tanto, como la edad ha sido recodificada por simplificar el ejemplo en dos categorías, Stata muestra sólo dos tablas, con sus correspondientes estadísticos.

```
recode edad (18/49=1 "18/49") (50/98=2 "50/98"), into(edadr)
label variable edadr "Edad"
bysort edadr: tabulate rvoto ingresos if rvoto != 99 & ingresos != 9, col norefq all
```

**ILUSTRACIÓN 8.12. Tabla de contingencia tridimensional**

		Ingresos familiares mensuales entrevistado			
Recuerdo de voto		<150	150-300	>300	Total
PP		29.49	37.32	39.05	34.92
PSOE		29.34	21.34	19.37	23.76
IU		6.44	7.73	6.67	7.12
Nacionalistas		4.79	7.84	8.25	6.86
Otros		1.95	2.68	6.35	3.02
Blanco		2.99	2.89	3.17	2.97
No votó		25.00	20.21	17.14	21.35
Total		100.00	100.00	100.00	100.00
Pearson chi2(12) = 51.8490 Pr = 0.000 Cramér's V = 0.1152					

		Ingresos familiares mensuales entrevistado			
Recuerdo de voto		<150	150-300	>300	Total
PP		45.71	52.34	54.55	48.16
PSOE		30.88	23.69	16.36	27.84
IU		2.86	4.13	4.55	3.33
Nacionalistas		5.93	8.82	10.00	7.01
Otros		0.33	1.10	2.73	0.72
Blanco		1.21	0.55	5.45	1.37
No votó		13.08	9.37	6.36	11.57
Total		100.00	100.00	100.00	100.00
Pearson chi2(12) = 50.2101 Pr = 0.000 Cramér's V = 0.1347					

¿De qué modo se interpretan estas tablas multivariadas? Para interpretar este tipo de tablas, se debe comparar cada una de las tablas multivariadas con la tabla bivariada original (en nuestro caso, con la ilustración 8.6), y las tablas multivariadas entre sí. En esta comparación se pueden producir las siguientes cuatro situaciones diferentes:

- La relación observada entre variable independiente y dependiente se debilita o desaparece al introducir la variable de control:* en este caso, toda o parte de la relación entre la variable independiente y dependiente se debía al efecto oculto de la variable de control, por lo que al introducir esta, la relación desaparece. El ejemplo clásico es el de la relación entre el número de cigüeñas y la tasa de natalidad. Si se cruza tasa de natalidad por número de cigüeñas en el municipio, es probable que la relación sea significativa: la natalidad es mayor en los municipios en los que hay más cigüeñas. ¿Quiere esto decir que la causa de la mayor natalidad son las cigüeñas?

Evidentemente, no: esta relación es espuria, y se debe al efecto de una tercera variable, tamaño del hábitat (cuanto más pequeño es el hábitat, mayor es el número de hijos). Si se controla por tamaño de hábitat, se comprueba que toda la relación entre número de cigüeñas y tasa de natalidad desaparece (en municipios del mismo tamaño, no existe relación alguna entre número de cigüeñas y número medio de hijos).

- b) *La relación entre las variables se mantiene más o menos igual en cada categoría de la variable de control:* en tal caso la variable de control no afecta a la relación. Por ejemplo, en nuestro caso, si las tablas en los tres grupos de edad fueran prácticamente iguales, el efecto de los ingresos sobre el voto sería independiente de la edad, por lo que no sería necesario incluir la edad como variable de control.
- c) *La relación en las variables se incrementa al controlar por una tercera variable:* esto implicaría que sí que existe una relación significativa entre la variable independiente y la dependiente, que se muestra debilitada si no se considera la variable de control. Esto es lo que parece que sucede en el ejemplo propuesto. La relación entre ingresos y voto es más clara dentro de cada grupo de edad que en todas las edades consideradas conjuntamente (compárese la ilustración 8.12 con la ilustración 8.6). Para una misma edad, conforme aumentan los ingresos, aumenta el voto al PP y disminuye el voto al PSOE. Las dos tablas controladas por edad son significativas por  $\chi^2$ , y las V de Cramer son superiores a las que aparecieron en la ilustración 8.9 para la relación sin control entre ingresos y voto.
- d) *La relación entre variable independiente y dependiente cambia de forma al incluir la variable de control:* en este caso, se trataría de una interacción de la variable independiente y la variable de control. En el ejemplo en cuestión parece que existe una ligera interacción entre ingresos y edad, porque la relación entre ingresos y voto es menor para los más jóvenes que para los mayores (véase ilustración 8.12): los ingresos familiares afectan más al voto según aumenta la edad del entrevistado. Para analizar este tipo de interacciones en tablas multivariadas es aconsejable recurrir a técnicas más avanzadas, como el análisis log-lineal.

Para incluir más de una variable de control, sólo hay que añadir la en la preinstrucción *bysort*. Por ejemplo, si se quisiera controlar por sexo y edad en este ejemplo, las instrucciones deberían ser las siguientes:

```
bysort sexo edadr: tabulate rvoto ingresos if rvoto != 99 & ingresos != 9, col all
```

Con lo que Stata mostraría cuatro tablas: ingresos por edad para hombres menores de 50 años, para hombres con más de 50, para mu-

jeres jóvenes y para mujeres mayores. El análisis de estas tablas sería igual al que se ha considerado para el caso de una sola variable de control, sólo que ligeramente más complicado.

### 8.3. Otras tablas especiales

Además de las instrucciones *summarize* y *tabulate*, a partir de la versión 8, Stata incluye otras tres que permiten representar en tablas una serie de estadísticos. Si la primera de las mencionadas puede mostrar los estadísticos propios de una única variable cuantitativa y la segunda pone en relación las frecuencias (absolutas, relativas y condicionales) de dos o más variables cualitativas, las instrucciones que se muestran en el siguiente apartado tienen por cometido cruzar información de variables cuantitativas con variables cualitativas. Un ejemplo simple de ellos consistiría en mostrar las distintas medias de ideología según la edad, el sexo o ambas características de las personas entrevistadas.

La primera de ellas no es realidad una nueva instrucción, sino la misma orden *tabulate* aplicada con la opción *summarize(variable)*. Si se opta por incluir esta última, en lugar de frecuencias (además de posibles porcentajes y residuos) de las variables de la tabla, aparecen los principales estadísticos de la variable cuantitativa expresada entre paréntesis.

El uso más simple de esta opción consiste en poner una variable cualitativa tras la instrucción y una cuantitativa en la opción del sumario. Si se desean ver las diferentes atribuciones ideológicas que atribuyen al PP los encuestados de distintas edades, habría que redactar la instrucción del siguiente modo:

```
tabulate edadr, summarize(ideopp)
```

En realidad, más que una nueva instrucción, se trata de la orden analizada en el apartado anterior con una opción que permite representar en las casillas, los estadísticos de la variable expresada, en este caso, la ideología que los entrevistados atribuyen al PP.

#### ILUSTRACIÓN 8.13. Tabla de estadísticos según valores de una variable

	Summary of Atribución de ideología al PP		
Edad	Mean	Std. Dev.	Freq.
18/49	7.4254574	1.5351508	2569
50/98	7.4261548	1.5507619	1537
Total	7.4257185	1.5408244	4106

También pueden obtenerse estadísticos en función de dos variables, e incluso tres o más si se emplea para ello la preinstrucción *bysort*.

```
by sort sexo: tabulate edadr ingresos, summarize(ideopp)
```

### ILUSTRACIÓN 8.14. Tabla de estadísticos según valores de dos variables

Means, Standard Deviations and Frequencies of Atribución de ideología al PP						
Edad	Ingresos familiares mensuales entrevistado			Ns/Nc	Total	
	<150	150-300	>300			
18/49	7.38	7.3193277	7.2989691	7.4155496	7.3566642	
	1.7239994	1.5406507	1.3858272	1.439263	1.5345397	
	300	476	194	373	1343	
50/98	7.408046	7.4285714	7.1911765	7.2402235	7.3571429	
	1.6441824	1.4448856	1.458453	1.4352512	1.5338973	
	348	203	68	179	798	
Total	7.3950617	7.3519882	7.2709924	7.3586957	7.3568426	
	1.6803548	1.5124239	1.4030065	1.4390069	1.5339419	
	648	679	262	552	2141	

Means, Standard Deviations and Frequencies of Atribución de ideología al PP						
Edad	Ingresos familiares mensuales entrevistado			Ns/Nc	Total	
	<150	150-300	>300			
18/49	7.5892256	7.4656319	7.5074627	7.4680233	7.5008157	
	1.5551631	1.6098428	1.4548145	1.4404084	1.5329042	
	297	451	134	344	1226	
50/98	7.4089636	7.4129032	7.7111111	7.7032967	7.5006766	
	1.6165327	1.5699578	1.3249738	1.5050156	1.5663941	
	357	155	45	182	739	
Total	7.4908257	7.4521452	7.5586592	7.5494297	7.5007634	
	1.5902868	1.5986111	1.4224012	1.4659402	1.5451865	
	654	606	179	526	1965	

Si parece demasiada información que aparezcan tantos estadísticos, es posible omitir algunos de ellos con las siguientes opciones, cuyo nombre explica por sí solo que es lo que se deja de representar, *nomeans*, *nostandard*, *noobs*, *nofreq*<sup>11</sup>.

La segunda instrucción, *tabstat*, se puede considerar en cambio una ampliación de la instrucción *summarize*. Y esto es así en un doble aspecto. Por un lado, porque incluye la posibilidad de mostrar más estadísticos de lo

<sup>11</sup> Observaciones coincide con el número de casos, las frecuencias son los casos una vez efectuada la ponderación, si la hubiere.

que es capaz la orden más sencilla. Y, por el otro lado, porque es capaz de cruzar los estadísticos según las valores de una segunda variable cualitativa sin necesidad de utilizar la preinstrucción *by*, ni de ordenar el fichero por los valores de la mencionada variable.

Para ello, la orden mencionada ha de ir acompañada por dos opciones. La opción *statistics*(estadístico) se expresa para indicar otros estadísticos distintos de la media, que se obtiene por omisión. Dentro de esta, las posibilidades son *mean* (para la media), *median* (para la mediana), *n* (para la frecuencia de casos), *sum* (para la suma de los valores de la variable), *q* (para los cuartiles), *max* (para el valor máximo), *min* (para el valor mínimo), *range* (para el rango), *iqr* (para el rango intercuartílico), *sd* (la desviación típica), *variance* (la varianza), *cv* (el coeficiente de variación), *semean* (el error típico de la media), *skewness* (el coeficiente de simetría), *kurtosis* (el coeficiente de apuntamiento).

El siguiente ejemplo muestra para un trío de variables (la valoración ideológica de tres partidos) los siguientes estadísticos: número de casos, media, cuartiles, desviación típica, rango intercuartílico, simetría y apuntamiento. Para ello hay que escribir la siguiente instrucción

```
tabstat ideopp ideopsoe ideoiu, s(n mean q iqr sd skewness kurtosis) col(variable)
```

Y el resultado aparece en la siguiente ilustración:

**ILUSTRACIÓN 8.15. Tabla de estadísticos de diversas variables**

stats	ideopp	ideopsoe	ideoiu
N	4106	4007	3913
mean	7.425718	4.284253	2.453872
p25	6	3	2
p50	7	4	2
p75	8	5	3
iqr	2	2	1
sd	1.540824	1.445571	1.206442
skewness	-.1242612	.5195444	1.083963
kurtosis	3.065596	4.59968	5.997489

Se observa que tanto las medidas de tendencia central como las de localización van en sentido decreciente: las más altas corresponden al PP, pues representan las posiciones situadas más a la derecha. En concreto, la media de la ideología atribuida al PP (en una escala del 1 al 10) está situada en el 7,4; la del PSOE está ubicada en el 4,3, y la de IU, en el 2,5. De igual modo, en los cuartiles se produce una distancia de tres puntos entre los dos primeros partidos, y de uno o dos puntos entre los dos partidos que se sitúan a la izquierda del espectro. Por otro lado, las desviaciones de las atribuciones son ligeramente más altas en la calificación ideológica del

PP que en la de los partidos de izquierda. Finalmente, la simetría muestra resultados muy lógicos: es negativa para el PP, porque hay muchos sujetos que dan alta puntuación y pocos que la dan baja (asimetría paradójicamente en este caso a la izquierda). Todo lo contrario de lo que ocurre en la evaluación de IU, donde hay muchos que dan baja puntuación (puntuación de izquierdas) y pocos que dan puntuaciones altas (en este caso la simetría es positiva, a la derecha). Cabe destacar también cómo la evaluación de la ideología del PP sigue prácticamente una distribución normal (simetría cercana a 0 y apuntamiento próximo a 3), mientras que el PSOE y, sobre todo, IU poseen una concentración de valoraciones en torno a la media (distribución leptocúrtica).

La instrucción *tabstat* permite —además de mostrar al mismo tiempo las estadísticas de diversas variables— cruzar estos datos por los valores de una variable nominal o grupal. El mismo ejemplo anterior, con sólo tres estadísticos, se ofrece para cada uno de los dos grupos de edad y para el total, si se añade la opción *by(variable)*.

```
tabstat ideopp ideopsoe ideoiu, statistics(n mean sd sk k) col(statistics) by (edad)
```

Otra opción que se emplea de modo distinto en la anterior instrucción es *col*. En esta ocasión se ha utilizado con la modalidad *statistics* para que aparezcan los estadísticos en los encabezamientos de las columnas, en lugar de las variables, como se especificó —*col(variable)*— en el ejemplo anterior. De este modo, en la primer columna —por ejemplo— aparece el número de casos correspondientes a las tres variables, en primer lugar para los jóvenes, en segundo lugar para los mayores y finalmente para el conjunto de la muestra.

#### ILUSTRACIÓN 8.16. Tablas de estadísticos según valores de una variable

Summary for variables: ideopp ideopsoe ideoiu by categories of: edad (Edad)					
edad	N	mean	sd	skewness	kurtosis
18-49	2569	7.425457	1.535151	-.1338611	3.066827
	2520	4.413492	1.461038	.5639496	4.469687
	2484	2.526973	1.204635	.9846718	5.507637
>=50	1537	7.426155	1.550762	-.1086945	3.062845
	1487	4.065232	1.392315	.4177254	4.831858
	1429	2.326802	1.199435	1.285705	7.093494
Total	4106	7.425718	1.540824	-.1242612	3.065596
	4007	4.284253	1.445571	.5195444	4.59968
	3913	2.453872	1.206442	1.083963	5.997489

Comparando las medias puede verse que en el caso del PP la edad apenas influye y los otros dos partidos son valorados unas décimas más a la derecha por los jóvenes que por los mayores. En la dispersión de las diferencias en la valoración ideológica de los partidos se ve claramente que —al igual que en las medias, pero no de forma tan exagerada— hay mayores diferencias entre partidos que entre edades.

Finalmente, está la instrucción *table*, que permite una construcción versátil de tablas multidimensionales, con la única limitación de no ser capaz de generar porcentajes condicionales, es decir, porcentajes verticales u horizontales.

La versatilidad de esta instrucción reside en el número de dimensiones que pueden ser conjugadas en una misma tabla. En primer lugar, como en el caso de *tabulate*, *summary(variable)* o el de *tabstat variable* pueden utilizarse las casillas para representar los estadísticos de una variable cuantitativa. En segundo lugar, tras la instrucción pueden explicitarse hasta tres variables: la primera será representada en las filas, la segunda en las columnas, la tercera —y aquí reside la diferencia fundamental de esta instrucción frente al resto— anidará los valores de la segunda. Quiere ello decir que, si en la tercera dimensión se expone la edad recodificada y en la segunda los ingresos, se representarán en primer lugar todos los ingresos correspondientes a los jóvenes, y a continuación todos los correspondientes a los mayores. Y, por si más dimensiones se pudieran necesitar, la instrucción *table* permite introducir hasta cuatro variables más en la opción *by(listavar)* para anidar a la variable *dependientes* expuesta en las filas de la tabla. Además, como en tantas otras órdenes, también puede emplearse la preinstrucción *bysort*, que puede añadir cuantas dimensiones sean necesarias en la presentación de frecuencias o estadísticos.

Aunque pocas veces se necesiten tantas dimensiones, el ejemplo siguiente muestra dónde son expuestas cada una de las variables en la tabla resultante.

```
recode ccaa 13=1 .=. else=2, into(comunidad)
label define madrid 1 "Madrid" 2 "Resto"
label val comunidad madrid
bysort comunidad: table rvoto ingresos edadr if ingresos <4, by(sexo) ///
cellwidth(8)
```

De este modo, la primera y la cuarta dimensión se representan en las filas anidadas entre sí; la segunda y la tercera variables, en las columnas, también anidadas entre ellas, y la quinta y posibles subsiguientes conforman distintas tablas, como se manifiesta en la ilustración 8.17. Adviértase, asimismo, que se ha indicado un tamaño de columna de ocho posiciones para que cupieran todas las columnas en la misma línea.

**ILUSTRACIÓN 8.17. Tabla de contingencia con la orden *table***

		Edad and Ingresos familiares mensuales entrevistado			50/98		
		18/49		>300			
		<150	150-300	>300	<150	150-300	>300
<b>-&gt; comunidad = Madrid</b>							
sexu and							
Recuerdo de	voto						
Hombre	PP	9	29	24	11	18	12
	PSOE	11	18	11	12	8	4
	IU	2	13	6	5	5	1
	Otros	1	1	2			
	Blanco	1	4	2			1
	No votó	5	14	6	4	3	
	No contesta	1	1	2	8	2	1
Mujer	PP	5	30	15	25	16	9
	PSOE	10	26	9	18	6	3
	IU	6	6	4	5		
	Otros	1	1				
	Blanco	2	1	2	1		2
	No votó	10	15	8	8		1
	No contesta		2	1	12	1	
<b>-&gt; comunidad = Resto</b>							
sexu and							
Recuerdo de	voto						
Hombre	PP	68	158	39	177	86	24
	PSOE	82	77	24	101	37	6
	IU	21	31	6	9	9	3
	Nacionalistas	16	46	16	20	19	7
	Otros	6	12	16	1	2	1
	Blanco	11	11	3	4	2	1
	No votó	87	71	29	29	15	4
	No contesta	29	34	13	63	20	4
Mujer	PP	115	145	45	203	70	15
	PSOE	93	86	17	150	35	5
	IU	14	25	5	7	1	1
	Nacionalistas	16	30	10	34	13	4
	Otros	5	12	2	2	2	2
	Blanco	6	12	3	6		2
	No votó	65	96	11	78	16	2
	No contesta	53	35	6	102	19	

Además de frecuencias, también pueden mostrarse mediante la orden *table* prácticamente los mismos estadísticos que la instrucción *tabstat*. Sólo se exceptúan el rango, la varianza, el error típico de la media, la asimetría y la curtosis. Pero en este caso, en lugar de aparecer en la opción *statistics*, ha de figurar en la opción *contents*, y al lado del nombre literal del estadístico en inglés ha de aparecer la variable cuantitativa de la que se desea que se obtenga el correspondiente cálculo. En el siguiente ejemplo, se quieren las medias de la atribución ideológica del PP y del PSOE, para cada una de las edades y recuerdos de voto. Además, se especifica un formato con la opción *format* para que las medias no salgan con una larga lista de decimales y se añaden otras dos opciones (*col* y *row*) a fin de que también muestre las medias independientemente de la edad en la última columna e independientemente del voto en la última fila.

**ILUSTRACIÓN 8.18. Tabla de estadísticos con la instrucción *table***

Recuerdo de voto recodificado	Recodificación de edad		
	18-49	>=50	Total
PP	6.9 4.2	7.1 3.9	7.0 4.0
PSOE	7.8 4.0	8.0 3.8	7.9 3.9
IU	7.9 5.2	8.2 5.0	8.0 5.1
Nacionalista	7.7 4.7	7.8 4.3	7.7 4.6
Otros	7.8 5.0	7.8 4.9	7.8 5.0
Blanco	7.6 4.7	7.7 5.0	7.6 4.8
No voto	7.6 4.8	7.3 4.5	7.5 4.7
NC	7.3 4.4	7.3 4.4	7.3 4.4
Total	7.4 4.4	7.4 4.1	7.4 4.3

En resumen, la instrucción *tabulate* es la más apropiada para la obtención de frecuencias y porcentajes, aunque eventualmente sirva también para representar en una o dos dimensiones estadísticas básicas de una variable cuantitativa al añadirle la opción *summarize*. La orden *tabstat* es muy apropiada para la obtención de muy distintos tipos de estadísticos univariados (error típico, asimetría y curtosis, entre otros), pero está limitada por permitir una sola dimensión de cruce, aunque de múltiples variables cuantitativas. Finalmente, la instrucción *table* ofrece multidimensionalidad, sin necesidad de ordenar el fichero, ni ejecutar la preinstrucción *bysort*, sobresale en la posibilidades de formato y, aunque no es capaz de ofrecer porcentajes verticales ni horizontales, permite representar un elenco amplio de estadísticos de una o varias variables cuantitativas.

## 8.4. Las tablas de respuesta múltiple

Es frecuente, sobre todo en cuestionarios, tener que analizar preguntas a las que conviene o a las que simplemente es posible dar más de una respuesta. Con los análisis e instrucciones tratados hasta el momento estas preguntas tienen que ser tratadas subdividiéndolas en cada una de sus opciones de respuesta. Por ejemplo, sea la pregunta 21a del estudio postelectoral de las elecciones de 2000.

### ILUSTRACIÓN 8.19. Pregunta de respuesta múltiple (opción código binario)

P.21 ¿Ha visto Ud. por televisión algún espacio de propaganda electoral de algún partido o coalición?
- Sí ..... 1
- No ..... 2 (88)
- N.C. ..... 9
P.21a ¿Recuerda de cuál o cuáles? (RESPUESTAS ESPONTÁNEAS). (ANOTAR TODOS LOS QUE DIGA EL ENTREVISTADO).
- IU ..... 1 (89)
- PP ..... 1 (90)
- PSOE ..... 1 (91)
- EA ..... 1 (92)
- PNV ..... 1 (93)
- CiU ..... 1 (94)
- ERC ..... 1 (95)
- IC-V ..... 1 (96)
- BNG ..... 1 (97)
- PA ..... 1 (98)
- CC ..... 1 (99)
- CHA ..... 1 (100)
- Otros ..... 1 (101)
- N.C. ..... 1 (102)

Fuente: CIS. Estudio 2384.

En ella se interroga por los partidos de los que se ha visto publicidad durante la campaña electoral. En el cuestionario destaca que todos los partidos posibles están codificados con el mismo dígito, el 1, ocupando cada uno de ellos una columna. Si se atiende al fichero en Stata (panel8.dta), puede observarse que las variables se denominan p21a01-p21a14. Con lo hasta ahora aprendido, sólo podría escribirse una instrucción como<sup>12</sup>:

```
tab1 p21a01-p21a14
```

<sup>12</sup> Aprovechando que está codificada con valores 0-1, se podría utilizar la igualdad entre una proporción y la media de una variable dicotómica para emplear la instrucción:

```
tabstat p21a01-p21a04, s(mean n), col(statistics)
```

Además, si se sustituye el 1 por el 100, los resultados serían porcentajes.

Con ella, saldrían catorce tablas dicotómicas, de las que se ofrece una muestra de las tres primeras:

#### ILUSTRACIÓN 8.20. Tabulaciones de ítems de multirrespuesta

```
-> tabulation of p21a01
```

iu	Freq.	Percent	Cum.
0	1,323	43.16	43.16
1	1,742	56.84	100.00
Total	3,065	100.00	

```
-> tabulation of p21a02
```

pp	Freq.	Percent	Cum.
0	479	15.63	15.63
1	2,586	84.37	100.00
Total	3,065	100.00	

```
-> tabulation of p21a03
```

psoe	Freq.	Percent	Cum.
0	564	18.40	18.40
1	2,501	81.60	100.00
Total	3,065	100.00	

Puede comprobarse que la presentación de los resultados es larga innecesariamente. Para solventarlo, se ha de recurrir a una rutina o módulo especial creado para Stata, que ha de descargarse de Internet de las páginas de su revista<sup>13</sup>.

```
net from http://www.stata-journal.com/software/sj5-1
net install st0082
```

Para variables dicotómicas, como es el caso de las que se acaban de señalar para el ejemplo, la instrucción para obtener una tabla no puede ser más sencilla:

```
mrtab p21a01-p21a14, title("Recuerdo publicidad") nonames
```

Sin embargo, la salida, mostrada en la ilustración 8.21 requiere un comentario más extendido. En primer lugar, conviene fijarse en el final, donde

<sup>13</sup> Aunque se publique en *Stata Journal*, este procedimiento (*mrtab*) no tiene garantía de Stata Corp., pues no ha sido desarrollado por esta empresa. Se presenta y explica en Jann (2005).

se muestran los casos válidos e inválidos. En este caso hay 2.192 de estos últimos, porque esta pregunta está filtrada por la anterior. Todos aquellos que dijeron no haber visto publicidad en la televisión o no contestaron a la pregunta p21 aparecen como casos inválidos en esta tabla.

**ILUSTRACIÓN 8.21. Tabla multirrespuesta de partidos de los que se recuerda publicidad**

Recuerdo de publicidad	Frequency	Percent of responses	Percent of cases
iu	1742	19.79	56.84
pp	2586	29.38	84.37
psoe	2501	28.42	81.60
ea	87	0.99	2.84
pnv	200	2.27	6.53
ciu	436	4.95	14.23
erc	191	2.17	6.23
ic-v	104	1.18	3.39
bng	99	1.12	3.23
pa	159	1.81	5.19
cc	86	0.98	2.81
cha	50	0.57	1.63
otros partidos	307	3.49	10.02
ns/nc	253	2.87	8.25
Total	8801	100.00	287.15
Valid cases:	3065		
Missing cases:	2192		

Pero lo más importante de esta tabla son los dos porcentajes, que corresponden a dos bases distintas que hay que saber distinguir en la interpretación de la respuesta múltiple. Por un lado, está el que emplea como base para su cálculo el número de casos ( $n$ ). La última columna de la tabla de la ilustración 8.21 responde al cociente entre la frecuencia y dicho número (3.065 casos que han respondido a la pregunta). Por tanto, la lectura de dicho porcentaje ha de ser la siguiente: “Entre quienes han visto publicidad de partidos en la TV durante la última campaña electoral”, el 84% la aribuyó al PP, un 82% al PSOE, mientras que sólo un 57% lo hizo a IU.

Por otro lado, para el cálculo de porcentajes, puede utilizarse otra base distinta: el número de respuestas ( $r$ ), equivalente en este ejemplo concreto al número total de partidos vistos por el conjunto de los entrevistados. En la ilustración 8.21, este valor corresponde a la suma de las frecuencias, es decir, 8.801<sup>14</sup>. Esta cantidad, a menudo difícil de interpretar, tiene su importancia porque es el denominador de la columna de porcentajes de respues-

<sup>14</sup> Habría que prestar atención al «ns/nc», que está sumado y en realidad no es ningún partido político, y también a la categoría «otros partidos», que pueden ser uno o varios. Lo correcto en la mayor parte de los casos sería omitir de esta tabla también la no respuesta, especialmente si interesa utilizar como base de porcentajes el número de respuesta.

tas, que con ciertas precauciones podría interpretarse como el porcentaje de recuerdos de un partido sobre la totalidad de partidos recordados. Del conjunto de recuerdos, casi un 30% son de publicidad del Partido Popular, un 28% del Partido Socialista y un 20% de Izquierda Unida.

El problema que presenta la tabla anterior es que su base (los no filtrados en la pregunta anterior) es una parte no representativa de la muestra: son sólo aquellos que vieron publicidad. Por tanto, debería modificarse a fin de que la base sea el conjunto de los casos encuestados. Para ello hay que hacer dos operaciones:

- A todos los que no supieron o no contestaron a la pregunta p21 (la que hace el papel de filtro) otorgar el valor 1 en la variable múltiple que recoge el valor ns/nc (p21a14, en este caso).
- Crear una nueva variable (p21a15, por ejemplo) que recoja a aquellos sujetos que dijeron "no" en la pregunta filtro, es decir, en la pregunta p21, esto es, a todos aquellos que dijeron no haber visto publicidad de partidos políticos en la televisión.

Empleando los recursos abordados en el capítulo 5, ambas cosas pueden realizarse del siguiente modo:

```
replace p21a14=1 if p21==9
generate p21a15=0
replace p21a15=1 if p21==2
label variable p21a15 "No ha visto"
```

Tras este proceso, se puede volver a solicitar la tabla, añadiendo la nueva variable creada (p21a15), que representa a todos aquellos que no han visto publicidad de partidos políticos durante la campaña electoral.

A continuación, se repite la instrucción, pero incluyendo también la nueva variable creada:

```
mrtab p21a01-p21a14 p21a15, title("Recuerdo de publicidad")
```

La ventaja del nuevo resultado está en que el porcentaje de los casos está calculado sobre el conjunto de la muestra, en lugar de sólo sobre los que recordaron haber visto publicidad. De este modo, puede estimarse que apenas el 50% de los ciudadanos mayores de 18 años vieron publicidad en la campaña de las elecciones generales de 2000 de alguno de los dos partidos mayoritarios del sistema político español. Este porcentaje es tan bajo como consecuencia de que más del 40% de los entrevistados señalaron no haber visto ningún espacio de propaganda electoral de algún partido o coalición.

**ILUSTRACIÓN 8.22. Tabla multirrespuesta con la inclusión de la categoría “ninguno”**

Recuerdo publ.	Frequency	Percent of responses	Percent of cases
iu	1742	15.81	32.97
pp	2586	23.47	48.95
psoe	2501	22.70	47.34
ea	87	0.79	1.65
pnv	200	1.82	3.79
ciu	436	3.96	8.25
erc	191	1.73	3.62
ic-v	104	0.94	1.97
bng	99	0.90	1.87
pa	159	1.44	3.01
cc	86	0.78	1.63
cha	50	0.45	0.95
otros partidos	307	2.79	5.81
ns/nc	279	2.53	5.28
No ha visto	2192	19.89	41.49
Total	11019	100.00	208.57
Valid cases:	5283		
Missing cases:	0		

Este es el tratamiento que hace el programa *mrtab* de las variables múltiples dicotómicas. Hay, no obstante, otros tipos de variables múltiples en los cuestionarios que necesitan un tratamiento ligeramente diferente. Se trata de las variables múltiples cuyas opciones de respuesta no están codificadas dicotómicamente, sino con un valor distinto para cada categoría. La ventaja de este proceder es que se necesitan menos variables distintas por pregunta, pues basta con emplear tantas como el máximo número de respuestas distintas puede dar el entrevistado. El caso anterior no es procedente, ya que hay sujetos que pueden contestar diciendo que han visto publicidad de todas y cada una de las opciones políticas, como sucedió realmente en seis ocasiones en el estudio que se comenta. Sin embargo, en muchas ocasiones el propio redactor del cuestionario limita el número de contestaciones que una persona puede dar a una pregunta de respuesta múltiple. Es, por ejemplo, el caso de la pregunta p39 del estudio postelectoral de 2000, cuyo texto literal es el siguiente: “¿Podría decirme entre qué dos partidos u opciones dudó Ud.?”. Obviamente, podrían reservarse treinta columnas para distintos partidos u opciones codificados binariamente. Pero es mucho más cómodo emplear dos variables, cada una de ellas con la posibilidad de grabar en formato fijo dos columnas, con lo que se dispondría de 100 opciones distintas. La ilustración 8.23 muestra con claridad un ejemplo alternativo de codificación al mostrado en la ilustración 8.19.

### ILUSTRACIÓN 8.23. Pregunta de respuesta múltiple (opción multicódigo)

P.39a	¿Podría decirme entre qué dos partidos u opciones dudó Ud.?
(ENTREVISTADOR:	Espere respuesta espontánea y marque las dos opciones que señale el entrevistado, cada una en una columna).
(217-218)	(219-220)
- IU .....	01 01
- PP .....	02 02
- PSOE .....	03 03
- EA .....	04 04
- EH .....	05 05
- PNV .....	06 06
- CIU .....	07 07
- ERC .....	08 08
- BNG .....	09 09
- PA .....	10 10
- CC .....	11 11
- IC-V .....	13 13
- CHA .....	16 16
- Otros partidos .....	49 49
- Votar en blanco.....	96 96
- Abstenerse .....	97 97
- N.C. .....	99 99

Fuente: CIS. Estudio 2384.

La instrucción para elaborar tablas de distribuciones de respuesta con este tipo de codificación es la misma que la que se acaba de escribir. Lo único distinto es la opción, ya que hay que indicar que los códigos son múltiples, en lugar de binarios con la opción *poly*. En esta modalidad es conveniente especificar otra opción que restrinja los códigos que se van a recontar: se trata de *response(listavalores)*, donde la lista debe indicar los valores que se desean representar. Para obtener al unísono las frecuencias de las dos respuestas de la pregunta de la ilustración 8.23, habría que especificar los códigos comprendidos entre 1 y 99.

```
mrtab p39a01 p39a02, poly response(1/99) title("Opciones en duda") nonames
```

El resultado de esta instrucción da un total de 538 casos, que son quienes dijeron dudar en la emisión del voto en las últimas elecciones. El total de la frecuencia no tiene sentido alguno, pues mezcla partidos, abstención y no contestaciones. Por tanto, los únicos números interpretables de la tabla son las frecuencias y sus correspondientes porcentajes de los casos. La base de estos últimos son quienes dudaron. En consecuencia, podría decirse, de acuerdo con la ilustración 8.24, que el 51% de los que dudaron lo hicieron incluyendo al PSOE, el 45% al PP y el 16% a IU. Además, para casi un 30% de los que dudaron, la abstención fue una de las opciones posibles.

**ILUSTRACIÓN 8.24. Tabla de frecuencias para variable múltiple con multicódigos**

Opciones en duda	Frequency	Percent of responses	Percent of cases
iu	87	8.54	16.17
pp	243	23.85	45.17
psoe	276	27.09	51.30
ea	5	0.49	0.93
pnv	8	0.79	1.49
ciu	31	3.04	5.76
erc	11	1.08	2.04
bng	9	0.88	1.67
pa	7	0.69	1.30
cc	6	0.59	1.12
ic-v	15	1.47	2.79
cha	6	0.59	1.12
otros partidos	37	3.63	6.88
votar en blanco	49	4.81	9.11
abstenerse	159	15.60	29.55
n.c.	70	6.87	13.01
Total	1019	100.00	189.41
Valid cases:	538		
Missing cases:	4745		

En este tipo de variables es muy útil utilizar una opción de esta instrucción llamada *include*, que incorpora también todos los casos que no han dado ninguna respuesta.

```
mrtab p39a01 p39a02, poly response(1/99) title("Opciones en duda") nonames ///
include
```

El efecto de esta opción (véase la ilustración 8.25) es el de incorporar en la base del porcentaje todos los casos que no tengan un valor comprendido entre los valores dados (entre el 1 y el 99, en este ejemplo). Por tanto, en lugar de 538 casos válidos, se obtienen 5.283 (el conjunto de la muestra)<sup>15</sup>. Consecuentemente, cambia el porcentaje sobre casos y se mantiene el de respuestas. De todos modos, esta tabla presenta un pequeño defecto en el cálculo del total del porcentaje de los casos (última fila de la última columna), puesto que utiliza como numerador la suma de las frecuencias, en lugar del recuento de casos con al menos una respuesta dada. El número 1.019 es el número total de respuestas. El total del porcentaje de los casos se calcula dividiendo esta cifra (de

<sup>15</sup> No necesariamente tiene que ser el conjunto de la muestra. Si un caso tiene valor perdido en todas las variables individuales, no es contabilizado para la base. En este ejemplo, los casos filtrados por la pregunta anterior se encuentran en la base de datos con 0, por lo que cumplen la condición para ser recontados.

modo equivocado, porque deberían dividirse los que responden y no las respuestas) por el número de casos total. Por tanto, este porcentaje no tiene ninguna interpretación válida, como es de reconocer que tampoco la tiene en la ilustración 8.24, donde es más evidente por sobrepasar el 100%<sup>16</sup>.

**ILUSTRACIÓN 8.25. Tabla de frecuencias para variable múltiple con multicódigos e *include***

Opciones en duda	Frequency	Percent of responses	Percent of cases
iu	87	8.54	1.65
pp	243	23.85	4.60
psoe	276	27.09	5.22
ea	5	0.49	0.09
pnv	8	0.79	0.15
ciu	31	3.04	0.59
erc	11	1.08	0.21
bng	9	0.88	0.17
pa	7	0.69	0.13
cc	6	0.59	0.11
ic-v	15	1.47	0.28
cha	6	0.59	0.11
otros partidos	37	3.63	0.70
votar en blanco	49	4.81	0.93
abstenerse	159	15.60	3.01
n.c.	70	6.87	1.33
Total	1019	100.00	19.29
Valid cases:	5283		
Missing cases:	0		

#### 8.4.1. Cruces con variables múltiples

Además de obtener tablas de preguntas de respuestas múltiples, la instrucción *mrtab*, al igual que su correspondiente origen *tabulate*, es capaz de producir tablas de doble entrada, siempre y cuando la segunda variable, ubicada en las columnas, no sea múltiple. La estructura general de esta instrucción para tablas de doble entrada es la siguiente:

---

<sup>16</sup> Si se quisiera saber qué porcentaje expresa duda, habría que convertir a los que «no dudan» (codificados aquí como 0) en un nuevo valor (100, por ejemplo) y tabularlo como una categoría más. El porcentaje complementario de esta categoría sería el correspondiente a las personas que «dudan». En este ejemplo habría 4.745 que no dudan, esto es, un 90%. En consecuencia, los que dudan son un 10%. Más en concreto, 538 de 5.283 entrevistados.

```
mrtab listavarmul, by(variable) [column row cell noref chi2 lrchi2 mtest  
mlrchi2]
```

En esta orden hay que notar que la variable sencilla de las columnas aparece entre las opciones, antecedida por *by()*, y pueden obtenerse tipos de porcentajes similares a los de la tabla normal, así como suprimir las frecuencias. También se puede calcular un  $\chi^2$  conjunto<sup>17</sup> de Pearson u otro basado en la razón de verosimilitud, o puede optarse por múltiples pruebas por filas<sup>18</sup>, en cuyo caso deberíamos optar por *mtest* o esta seguida de *mlrchi2*, a fin de calcularlas por el segundo método.

Para probar la hipótesis de que no hay diferencias en el recuerdo de publicidad de los partidos políticos durante la campaña electoral entre hombres y mujeres, habría que solicitar una tabla de porcentajes que incluyera las pruebas múltiples y el test conjunto de la  $\chi^2$  de Pearson:

```
mrtab p21a01-p21a14 p21a15, by(sexo) noref col chi2 mtest title("Rec...") ///  
nonames
```

La tabla obtenida muestra que los hombres se han expuesto o recuerdan más los mensajes publicitarios de casi todos los partidos. Especialmente, son reseñables las diferencias en los tres partidos de mayor cobertura territorial, pues son muy próximas a diez puntos porcentuales. En consonancia, un 45% de las mujeres no recuerdan haber visto cuñas electorales, frente a un 38% de los hombres que no vieron este tipo de publicidad.

---

<sup>17</sup> Para el cálculo de este estadístico, el programa expande las filas a tantas como combinaciones empíricas existen de multirrespuestas. Un ejemplo sencillo aclarará esto último. Imaginemos sólo dos partidos optados. La tabla aparentemente tiene sólo dos filas, pero pueden expandirse a cuatro: no elige ninguno, sólo elige el primero, sólo elige el segundo o elige los dos. En general, el número de filas expandidas es  $2^i$ . Sin embargo, es preciso tener en cuenta que hay que eliminar aquellas combinaciones sin ninguna frecuencia, a fin de obtener los verdaderos grados de libertad de la tabla.

<sup>18</sup> Si se opta por múltiples pruebas, se puede elegir un ajuste de la significación por el método de Bonferroni, Holm o Sidak, poniendo sus nombres en minúsculas y entre paréntesis después de la opción *mtest*.

**ILUSTRACIÓN 8.26. Cruce de recuerdo de publicidad por sexo con pruebas múltiples**

Recuerdo de publicidad	Hombre	Mujer	Total	chi2/p*
iu	38.75	27.71	32.97	72.575/0.000
pp	53.20	45.08	48.95	34.740/0.000
psoe	52.52	42.62	47.34	51.829/0.000
ea	1.83	1.48	1.65	0.956/0.328
pnv	4.33	3.29	3.79	3.874/0.049
ciu	9.65	6.98	8.25	12.353/0.000
erc	4.25	3.04	3.62	5.525/0.019
ic-v	2.42	1.56	1.97	5.120/0.024
bng	2.30	1.48	1.87	4.809/0.028
pa	3.85	2.24	3.01	11.668/0.001
cc	1.98	1.30	1.63	3.833/0.050
cha	1.27	0.65	0.95	5.389/0.020
otros partidos	7.07	4.67	5.81	13.859/0.000
ns/nc	4.57	5.93	5.28	4.931/0.026
No ha visto	37.55	45.08	41.49	30.741/0.000
Total	225.53	193.13	208.57	
* Pearson chi2(1) / Unadjusted p-values				
Valid cases:	5283			
Missing cases:	0			
Overall Test(s) of Significance:				
Pearson chi2(164) = 238.7807 Pr = 0.000				

La única prueba en la que no se aprecian diferencias significativas (véase última columna de la ilustración 8.26) es en la de Eusko Alkartasuna. En conjunto, hay una asociación significativa entre recuerdo de publicidad y género: los hombres recuerdan haberla visto más que las mujeres.

## 8.5. Ejercicios

1. Cruza el uso de Internet en los doce últimos meses (cis2794: P.27) por sexo, edad (recojidificada en tres intervalos) y estudios. ¿Qué variable parece tener mayor influencia? (Se recomienda poner valores perdidos a los que no contestan a la pregunta sobre Internet y a los estudios).
2. Calcula los residuos ajustados de la tabla Internet por estudios. ¿En qué casillas se encuentran los residuos significativos? ¿Qué implica que unos sean positivos y otros negativos?
3. Realiza un cruce de uso de Internet por edad y estudios, controlando por sexo. ¿Es diferente la influencia de la edad y de los estudios según se sea hombre o mujer?
4. Empleando ahora el barómetro de abril de 2009 (cis2798), o cualquier otro de enero, abril, julio u octubre, haz una tabla con los estadísticos principales (*n*, media y desviación típica) de la valoración de los prin-

- cipales líderes políticos (P.17). Compara estos resultados por ideología agrupada en tres categorías (izquierda, centro y derecha).
5. En el barómetro de mayo de 2009 (cis2801) se hace una pregunta de respuesta múltiple (P.25a) sobre las personas con las que convive el entrevistado. Realiza una tabla en la que esté recogida también la categoría “nadie”. Crúzala por la edad recodificada en cuatro intervalos y extrae conclusiones sobre la relación entre ambas variables.
  6. Con el mismo estudio, obtén una tabla de distribución de frecuencias con los principales problemas del país (P.5). Crúzala por ideología y comenta la influencia de la ideología en la percepción temática de los problemas del país.



# 9

## La regresión<sup>1</sup>

Un aspecto de primordial atención en el análisis de las variables cuantitativas es el estudio de la asociación entre ellas, para averiguar si los valores de unas determinadas variables varían con la misma pauta que los de otras. Una perspectiva —ya vista en el capítulo relacionado con las comparaciones— es estudiar, por ejemplo, si la tasa de inflación es mayor o menor en Francia que en España; otra perspectiva que puede adoptarse es la de ver si varían conjuntamente, esto es, si en los momentos en que en Francia es alta, también lo es en España, mientras que cuando el ascenso de los precios se encuentra en cotas bajas en Francia, también lo hace de ese modo en España; o si, por el contrario, no existe relación alguna entre los datos de cada uno de estos países, y la evolución de la inflación en uno de estos países es independiente de la del otro. Poniendo otro ejemplo, podría indagarse en un conjunto de países si, por término medio, la esperanza de vida de los hombres es distinta —mayor o menor— que la de las mujeres, en este caso estaríamos ante una comparación. Sin embargo, si se desea averiguar si aquellos países en los que viven más tiempo los hombres son los mismos en los que también las mujeres tienen una esperanza de vida mayor, entonces se está ante el estudio de la asociación entre las variables. Naturalmente, puede darse el caso en el que la relación o asociación sea de distinto signo, pues podría ocurrir que los casos que en una variable tienen valores más altos, lo tienen más bajos en la otra variable. Por ejemplo, es claro que en los países con alta renta per cápita, la mortalidad infantil es pequeña, pues los países que tienen mejores niveles de ingresos suelen tener mejores condiciones sanitarias para sus recién nacidos, por lo que la tasa de una muerte postnatal es menor.

---

<sup>1</sup> Para los próximos dos capítulos y, en parte, también para los dos siguientes se recomiendan los manuales de econometría. Entre los más conocidos se encuentran Novales (1993), Peña (2002), Gujarati (2008), Wooldridge (2009), Green (2008), Maddala (2001)... También existen libros de econometría basados en Stata. Entre ellos, se encuentran Cameron (2005) y Baum (2006). Asimismo, enfocados desde el análisis multivariante con abundantes ejemplos, pueden citarse, entre otros muchos, Hair (2006) y Cea (2002).

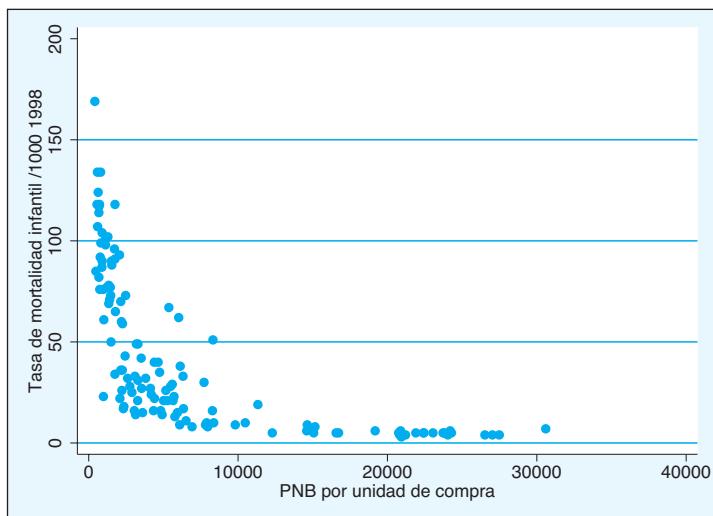
## 9.1. Nube de puntos, covarianza y correlación

El método más intuitivo para el estudio de la covariación entre dos variables cuantitativas es el gráfico de dispersión o nube de puntos, que consiste en un eje de coordenadas en el que se representa la variable  $x$ , independiente o también denominada predictor, en el eje de abscisas, y se ubica la otra variable  $y$ , dependiente o resultado (*outcome*), en el eje de ordenadas. En el espacio interior del gráfico se dibuja un punto por cada caso a una distancia horizontal proporcional al valor de la variable  $x$  y a una distancia vertical proporcional al valor de la variable  $y$ ; o, dicho de otro modo, se trata de dibujar tantos puntos como casos tenga la distribución sobre cada una de las dos dimensiones del gráfico, cuyas proyecciones equivalgan a los valores que cada caso tiene en las dos variables en cuestión. Como ya se explicó en el capítulo dedicado a los gráficos, la nube de puntos (o diagrama de dispersión) entre dos variables puede obtenerse mediante la instrucción *scatter* seguida de las dos variables que se quieren representar. Así, con la siguiente instrucción...

```
scatter tmi pnbppa, name(G1, replace)
```

... Stata representa la nube de puntos de las dos variables señaladas, en este caso la tasa de mortalidad infantil y el producto nacional bruto per cápita, consideradas como dependiente e independiente, respectivamente.

**GRÁFICO 9.1. Nube de puntos**



En este gráfico están representadas dos variables para los 125 países del mundo de los que se poseen ambos datos: en el eje horizontal está expuesto el producto nacional bruto per cápita en unidades de poder de compra y en el vertical, la tasa de mortalidad infantil. El rango de la primera se extiende desde los 414 dólares per cápita de Sierra Leona hasta los 30.600 de Estados Unidos, el de la segunda, desde los 3% de Hong Kong hasta los 169% correspondientes también a Sierra Leona. Como era lógicamente de esperar, se puede apreciar que los casos que tienen menores valores de renta poseen también altas tasas de mortalidad.

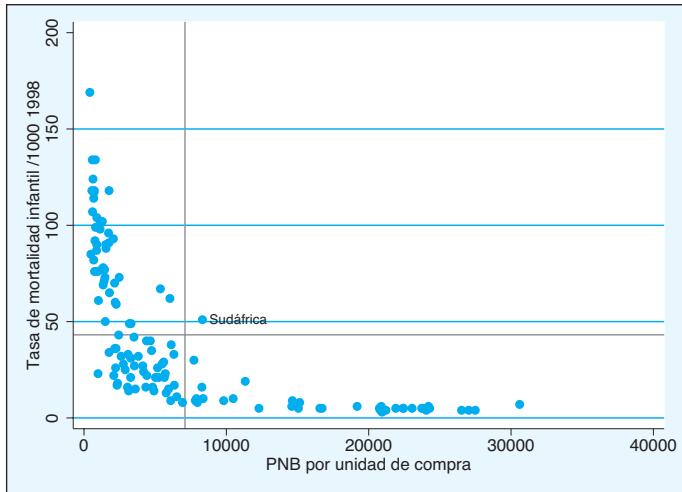
Para entender mejor la relación entre las dos variables es útil dibujar en el gráfico dos líneas de referencia que sean las medias de las variables representadas. Esto se logra calculando en primer lugar una constante (denominada *macro* en la terminología de Stata) mediante la orden *global*, para cada una de las medias, a las que se les denominará con el nombre que se desee. En este caso se ha preferido hacerlo con el mismo nombre de la variable seguido de una *x*. A continuación, la constante o macro se inserta en la instrucción *scatter* cuidando de que esté precedida del signo \$<sup>2</sup>.

```
summarize tmi
global tmix=r(mean)
summarize pnbppa
global pnbppax=r(mean)
scatter tmi pnbppa, xline($pnbppax) yline($tmix) name(G2, replace)
```

Al trazar las dos líneas de referencia que representan las medias de cada una de las variables, resulta que la mayoría de los países se encuentran o bien por encima en mortalidad infantil, pero por debajo en producto nacional bruto per cápita, o bien por debajo en mortalidad y por encima de la media en renta. Sin embargo, aunque en menor número en conjunto, también hay otro grupo de países que poseen baja mortalidad y bajo producto nacional bruto y sólo un caso que se encuentra con valores por encima en las dos variables consideradas.

---

<sup>2</sup> En los gráficos de este capítulo y siguientes se han especificado opciones que no han sido explicadas por razones de espacio en el correspondiente capítulo de gráficos. Por la denominación de estas opciones y por su contexto, el lector inteligente deducirá inmediatamente su uso. Por ejemplo, *xline* significa dibujar una línea en el eje *x*, e *ytitle* se usa para poner un título en el eje *y*.

**GRÁFICO 9.2.** Nube de puntos con las medias representadas

Un concepto clave para la comprensión de la asociación entre dos variables de tipo cuantitativo es el de covarianza. Como es fácil deducir, procede del concepto de varianza, es decir, del promedio de las distancias cuadradas de los valores con respecto a la media. En el caso de trabajar con dos variables, en lugar de una, se pueden calcular sendas distancias con respecto a la media, una para cada una de las variables:  $(x_i - \bar{x})$  e  $(y_i - \bar{y})$ . La covarianza es un promedio del producto entre estas dos distancias, y su fórmula adopta la siguiente expresión:

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})f_i}{\sum_{i=1}^n f_i} \quad (9.1)$$

La covarianza puede adoptar cualquier valor positivo o negativo. Si se divide el gráfico en cuatro sectores por las medias aritméticas de cada una de las variables, es fácil advertir que los productos de las diferencias en los casos que están en el cuadrante superior de la derecha (sólo Sudáfrica en este caso) han de ser positivos, pues tanto  $(x_i - \bar{x})$  como  $(y_i - \bar{y})$  son positivos. También los productos de las diferencias de los casos que se encuentran en el cuadrante inferior izquierdo (por ejemplo, Georgia) son superiores a 0 porque ambas diferencias en  $x$  e  $y$  son negativas. En cambio, los puntos o casos que se ubiquen en los cuadrantes superior izquierdo e inferior derecho generan productos negativos, pues una de las diferencias es positiva y la otra negativa. En este caso, como la mayor parte de los puntos se encuentran en cuadrantes con productos negativos y, sobre todo, las distancias de estos a las medias son bastante mayores, el sumatorio, y en consecuencia la covarianza, arroja un valor por debajo de 0.

Para obtener la covarianza con el programa Stata ha de utilizarse una opción del programa que hace el cálculo de la matriz de correlaciones. De este modo, para la obtención de la covarianza entre las variables *pmi* y *pnbppa*, habrá que escribir la siguiente instrucción:

```
correlate tmi pnbppa, covariance
```

El resultado no sólo otorga la covarianza entre las dos variables. También presenta en la diagonal de la matriz las varianzas de las dos variables, puesto que la covarianza de una variable consigo misma es igual a su varianza.

### ILUSTRACIÓN 9.1. Matriz de varianzas-covarianzas

(obs=125)		
	tmi	pnbppa
tmi	1493.36	
pnbppa	-194352	6.2e+07

La varianza de las tasas de mortalidad supera el millar y la del producto nacional bruto supera las decenas de millón. Son tan grandes porque están referenciadas en unidades cuadráticas. La covarianza con un valor de -194.352 sale negativa, como se dedujo de los razonamientos expuestos más arriba. Resulta evidente que la interpretación de estas cantidades depende de las unidades de medida que empleemos en cada variable. Por ello, es conveniente, para poder realizar comparaciones, transformar las variables en unidades típicas, esto es, convertirlas en otras, linealmente dependientes, en función de una fórmula que logra que se conviertan en variables con media 0 y desviación típica igual a 1.

$$z_i = \frac{x_i - \bar{x}}{s} \quad (9.2)$$

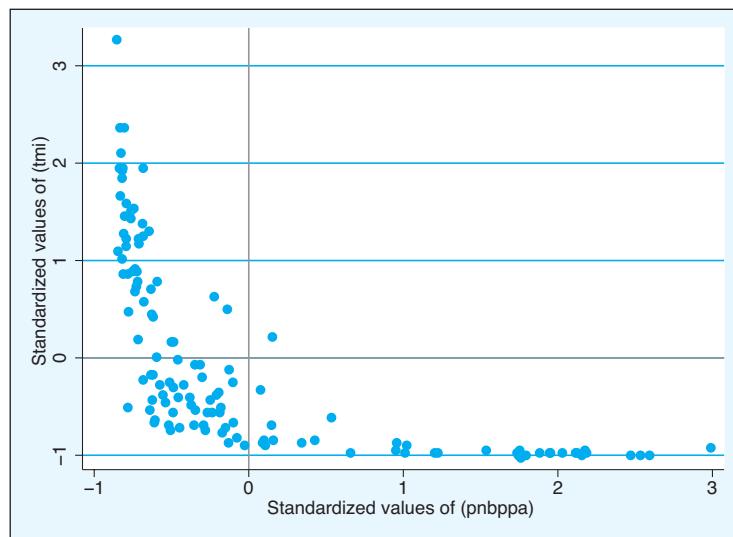
Si se realiza esta operación tanto sobre la variable *x* como sobre la variable *y*, mediante la instrucción *egen* y la función *std()*<sup>3</sup>...

<sup>3</sup> Para realizar operaciones repetitivas sobre variables, es útil la utilización del bucle *for*. Aunque ya esté descatalogado y sustituido en Stata por instrucciones más complejas y completas, no deja de ser útil y sencillo su uso. Tras la instrucción se escribe una lista de variables seguidas de dos puntos. A continuación se escribe otra orden en la que la letra X mayúscula será reemplazada por los nombres de las variables. En el ejemplo del cuerpo del texto, la X aparece dos veces en la instrucción *egen*. Incluso en una de ellas se escribe precedida de una *z*, lo que permite crear nuevas variables con el nombre de las antiguas antecedidas por dicha letra minúscula.

```
generate valido= tmi<. & pnbppa<.
for var tmi pnbppa: egen zX=std(X) if valido
scatter ztmi zpnbppa, xline(0) yline(0) xlabel(-1 0/3) ylabel(-1 0/3) name(G4, replace)
```

... al haber aplicado una transformación lineal sobre las dos variables estudiadas, el gráfico que representa su relación queda inalterado en comparación con el anterior y ofrece en consecuencia el siguiente aspecto:

**GRÁFICO 9.3. Nube de puntos de las variables estandarizadas**



Pero, aunque el gráfico sea similar, las covarianzas son distintas. Con la misma instrucción antes usada, aplicada a las variables transformadas...

```
correlate ztmi zpnbppa, covariance
```

... la matriz de covarianzas presentaría el siguiente resultado:

**ILUSTRACIÓN 9.2. Matriz de correlaciones**

		ztmi	zpnbppa
ztmi	1		
zpnbppa	-.640543	1	

Como puede apreciarse, es la misma que si se pidiera la matriz de correlaciones de las variables originales. Esto es así porque, matemáticamente, el coeficiente de correlación entre dos variables es igual a la covarianza de estas variables tipificadas o, expuesto directamente en una fórmula, el *coeficiente de correlación* se expresa como la covarianza dividida por las desviaciones típicas de ambas variables:

$$R_{xy} = \frac{S_{xy}}{S_x S_y} \quad (9.3)$$

La matriz anterior podría haberse obtenido de modo mucho más simple con la instrucción *correlate* aplicada a las variables originales:

```
correlate tmi pnbppa
```

Esta instrucción ha de ofrecer necesariamente el mismo resultado que el de las covarianzas de las variables tipificadas:

**ILUSTRACIÓN 9.3. Matriz de varianzas-covarianzas de variables estandarizadas**

(obs=125)		
	tmi	pnbppa
tmi	1	
pnbppa	-.640543	1

Por tanto, desde este punto de vista, puede entenderse el coeficiente de correlación como la covarianza de dos variables tipificadas, de modo tal que varía entre 1, en el caso de que ambas variables sean iguales, y -1, en el caso de que tengan los mismos valores tipificados, pero de distinto signo.

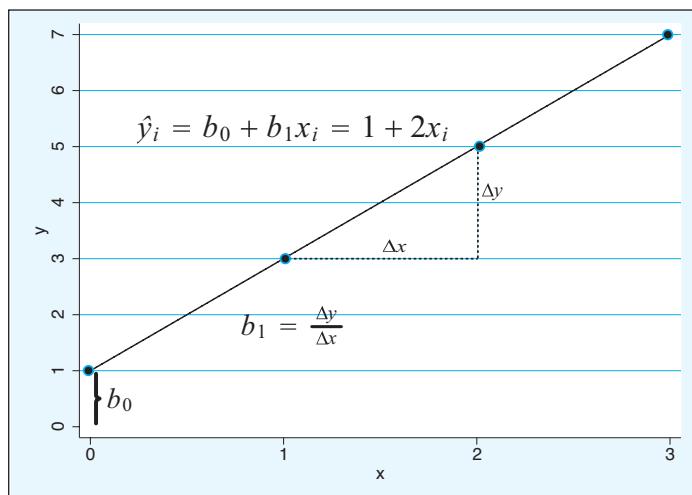
## 9.2. Regresión simple

Realizar una regresión simple consiste en buscar una línea que pase lo más cerca posible de los puntos que reflejan la distribución conjunta de dos variables. El modelo más simple de línea que puede encontrarse es una recta. Con nociones básicas de álgebra, se sabe que una línea recta puede representarse matemáticamente a través de una ecuación.

$$\hat{y}_i = b_o + b_1 x_i \quad (9.4)$$

En esta ecuación sólo resulta de momento conocido  $x_i$ , que representa los valores de la variable *predictor*, en este caso, el producto nacional bruto per cápita.  $\hat{y}_i$  son los valores teóricos (por eso, el acento circunflejo) que debería tener el resultado o variable dependiente (por eso,  $y$ ), si esta siguiera fielmente el modelo de la recta.  $b_0$  es la constante de la regresión, o punto donde la recta corta el eje de abscisas (el vertical).  $b_1$  es otra constante, que refleja la inclinación de la recta (su tangente) o, dicho de otro modo, el cambio que se produce en la variable dependiente cuando en la independiente se produce el aumento de una unidad.

**GRÁFICO 9.4. Representación matemática y geométrica de una recta**



En este gráfico se representa el modelo lineal, en el que la variable  $x$  está expuesta en el eje horizontal, la  $y$  en el vertical y en el espacio entre ellas se ha dibujado una recta, representada en la ecuación  $\hat{y}_i=1+2x_i$ , que nace en el punto 1 del eje de ordenadas ( $b_0$ ), con una pendiente igual a 2 ( $b_1$ ), es decir, que la variable  $y$  incrementa esta cantidad por cada incremento de una unidad en  $x$ , en este caso, cuando el valor de  $x$  crece un punto, la variable  $y$  sube dos.

Otro ejemplo de obtención de esta relación lineal de dos variables se ofrece a continuación utilizando el programa Stata. Se va a adoptar como variable predictora la variable tipificada del producto nacional per cápita (*zpnbpaa*) y a partir de ella se va a generar una nueva variable, la  $\hat{y}_i$ , designada en el programa como *tztmi* (la *t* para aclarar que se trata de valores teóricos y la *z* para indicar que se trata de valores tipificados) asumiendo que  $b_0$  sea igual a 0 y  $b_1$  igual al coeficiente de correlación (-0,640543):

```
generate tztmi=-.640543*zpnbpaa
```

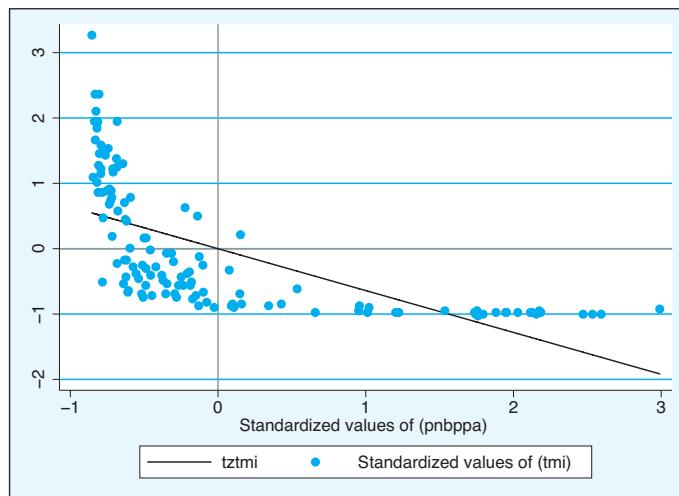
Una vez creada la nueva variable, se construye el gráfico añadiendo una serie de opciones a la orden *scatter* para obtener la línea del modelo de la regresión<sup>4</sup>:

```
scatter ztmi ztmi zpnbp if valido, connect(l i) msymbol(i O) ///
xline(0) yline(0) sort(zpnbp) name(G8, replace)
```

De este modo se obtiene el siguiente gráfico, donde, además de los puntos que representan a los países, aparece una recta que representa el modelo. Como claramente se ve, hay una discrepancia entre la realidad (los puntos) y el modelo (la recta). Cada una de las discrepancias, correspondiente a un caso, recibe el nombre de *residuo*, que se obtiene mediante la sustracción de los valores teóricos a los reales:

$$e_i = y_i - \hat{y}_i \quad (9.5)$$

**GRÁFICO 9.5.** Nube de puntos y recta estimada de variables estandarizadas



<sup>4</sup> Para que la línea solicitada en *connect* (l) se presente sin discontinuidades, es preciso que los datos estén ordenados por la variable independiente. Mediante la opción gráfica *sort(variable)* puede realizarse esta operación sólo temporalmente para producir el gráfico.

Como es tan larga la instrucción gráfica, se ha dividido en distintas líneas mediante ///. Este procedimiento es válido en Stata sólo si la orden está localizada en un programa. No funciona si se introduce desde la ventana interactiva de instrucciones. Para que lo haga, han de omitirse las tres barras.

Además de las rectas ortogonales a los ejes que representan las respectivas medias de  $x$  e  $y$ , todas las rectas que pasan por el punto  $(\bar{x}, \bar{y})$  tienen una curiosa propiedad, que es la de que la suma de las distancias de los puntos a estas rectas trazadas, es decir, el sumatorio de los residuos, es igual a 0. Esto es evidente en el caso de las rectas ortogonales correspondientes a las medias de cada una de las variables: en el caso de la recta vertical (media de  $x$ ), las distancias de los puntos a la recta se pueden expresar con la diferencia entre  $(x_i - \bar{x})$ ; y una de las propiedades de la media es que el sumatorio de estas diferencias es igual a 0. Por otro lado, la aplicación de esta propiedad a la recta horizontal (media de  $y$ ) también es evidente, pues las distancias de los puntos a ella vienen indicadas por la expresión  $(y_i - \bar{y})$ , que por extensión tiene también las mismas propiedades que  $x$ .

En la regresión basada en el criterio de *mínimos cuadrados* se pretende no sólo que la suma de residuos sea igual a 0, ya que esta propiedad la cumplen infinitas rectas, esto es, todas las que pasan por el punto  $(\bar{x}, \bar{y})$ , sino, sobre todo, encontrar la recta en la que sea mínima la suma de las distancias al cuadrado de sus puntos con respecto a los puntos empíricos, esto es, en la que la suma de los residuos elevados al cuadrado sea la menor posible.

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^n e_i^2 \quad (9.6)$$

La instrucción del programa que permite obtener los parámetros  $b_0$  y  $b_1$  correspondientes a la recta obtenida con el criterio de mínimos cuadrados ordinarios es *regress*. Su sintaxis más simple se compone de la orden seguida por la variable dependiente (resultado) y la independiente (predictora), en este orden:

```
regress vardep varindep
```

Como primer ejemplo se va a realizar la regresión de las dos variables tipificadas correspondientes a tasa de mortalidad infantil y producto nacional bruto per cápita:

```
regress ztmi zpnbpaa
```

El resultado de esta instrucción es el siguiente:

### ILUSTRACIÓN 9.4. Regresión de las variables estandarizadas

Source	SS	df	MS	Number of obs = 125		
Model	50.876649	1	50.876649	F( 1, 123) = 85.58		
Residual	73.1233526	123	.594498802	Prob > F = 0.0000		
Total	124.000002	124	1.00000001	R-squared = 0.4103		
				Adj R-squared = 0.4055		
				Root MSE = .77104		
ztmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
zpnbpaa	<b>-.6405432</b>	.0692412	-9.25	0.000	-.7776019	-.5034844
_cons	<b>3.23e-09</b>	.0689637	0.00	1.000	-.1365094	.1365094

Del conjunto del resultado se han resaltado las dos cantidades que respectivamente representan los coeficientes de la regresión. El primero, el correspondiente a la línea de la variable predictora (*zpnbpaa*), es la pendiente de la recta dibujada en el gráfico 9.5. Como puede apreciarse, es idéntica al coeficiente de correlación, por el hecho de haber empleado las variables tipificadas. E implica que un cambio en una desviación típica del producto nacional bruto de un país se traduce en una disminución de seis décimas de la desviación típica en la tasa de mortalidad. En la misma columna, pero en la línea siguiente, se encuentra el valor correspondiente a la constante o  $b_0$ . Este es prácticamente igual a 0; sólo problemas de precisión hacen que el valor no sea exactamente nulo, sino con nueve cifras decimales. El hecho de que la constante sea 0 es por estar ambas variables tipificadas, es decir, porque ambas tienen media nula y la recta de regresión basada en los mínimos cuadrados ha de pasar necesariamente por ese punto. Cuando una variable independiente tipificada adopta el valor 0, la dependiente, también tipificada, ha de tener el valor nulo.

¿Qué pasaría si se hiciera la regresión con las variables originales? Véase aplicando la instrucción con los nombres que estas dos variables tenían originalmente en el fichero, esto es, *tmi* y *pnbppa*.

```
regress tmi pnbppa
```

A partir de lo cual la regresión resultante es la siguiente:

### ILUSTRACIÓN 9.5. Regresión de las variables originales

Source	SS	df	MS	Number of obs = 125			
Model	75976.9095	1	75976.9095	F( 1, 123) = 85.58			
Residual	109199.139	123	887.797874	Prob > F = 0.0000			
Total	185176.048	124	1493.35523	R-squared = 0.4103			
				Adj R-squared = 0.4055			
				Root MSE = 29.796			
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]		
pnbppa	-.0031526	.0003408	-9.25	0.000	-.0038272	-.002478	
_cons	65.15607	3.604551	18.08	0.000	58.02108	72.29106	

En este caso, los coeficientes salen completamente distintos: el correspondiente a la variable dependiente pasa a ser de -0,003, puesto que —según la nueva recta encontrada— cada dólar que un país aumenta en su producto nacional bruto implica una reducción en 3 millonésimas<sup>5</sup> en su tasa de mortalidad infantil. Y el valor 65,1 de la constante significa que en el imposible caso de que un país tuviera una producción nula, igual a 0 dólares per cápita, la tasa de mortalidad infantil predicha por la recta sería del 65%.

¿Cómo se podrían obtener con Stata los valores predichos por este modelo? Hay dos fórmulas. Entre las instrucciones conocidas, la más conveniente es *generate*, tal como se hizo anteriormente con las dos variables tipificadas, aunque en este caso haya que emplear los dos coeficientes, puesto que la constante ( $b_0$ ) no es igual a 0:

```
generate ttmib=65.15607 + -.0031526*pnbppa
```

Pero hay una instrucción más directa<sup>6</sup> en la que no hay que transcribir ningún número, ni constante, con tal de realizarla tras la instrucción *regress* correspondiente. Se trata de la orden *predict* seguida del nombre de la variable predicha que se creará con los parámetros de la regresión y los valores disponibles de la variable independiente:

```
predict varnueva
```

<sup>5</sup> En realidad, un cambio en la variable independiente provoca una disminución en 3 milésimas de la variable dependiente, pero como esta está expresada en la matriz de datos en tantos por mil, las tres milésimas se convierten en 3 millonésimas.

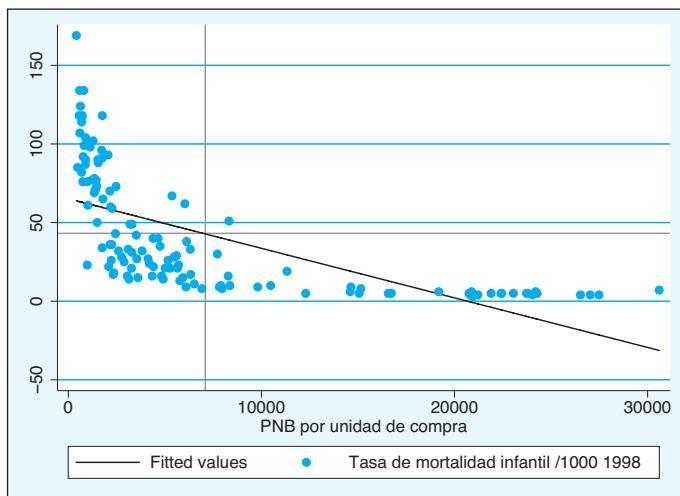
<sup>6</sup> También hay otra solución indirecta. Se trata de utilizar las estimaciones que se quedan guardadas después de cualquier análisis estadístico. Los coeficientes de las regresiones en Stata se conservan con este formato *\_b[nombre]*. De este modo, podría predecirse la *ttmi* con la siguiente expresión: *\_b[\_cons]+\_b[pnbppa]\*pnbppa*.

A partir de este momento ya se puede trabajar con esta nueva variable en cualquier instrucción, como puede ser *scatter*:

```
predict ttmi
scatter ttmi tmi pnbppa if valido, connect(l i) msymbol(i O) ///
xline($pnbppax) yline($tmix) sort(pnbppa) name(G11, replace)
```

... que genera el gráfico siguiente:

**GRÁFICO 9.6. Nube de puntos y recta estimada de variables originales**



Puede apreciarse que el gráfico 9.6 y el efectuado con los valores típicos de ambas variables (gráfico 9.5) son idénticos salvo en las escalas, puesto que cuando se tipifica una variable sus relaciones lineales con otras variables quedan inalteradas, pues esta transformación sólo produce un cambio proporcional en sus valores.

### 9.3. Bondad del ajuste de la regresión

Se acaba de obtener la recta (el modelo lineal) que mejor se ajusta a los valores empíricos de la distribución de dos variables. Pero el mejor no necesariamente quiere decir que sea bueno. Para medir la bondad del modelo se utilizan dos medidas: una absoluta y otra relativa.

La medida absoluta para evaluar una regresión es el error típico de la regresión o desviación típica de los residuales. En el fondo es un promedio

cuadrático<sup>7</sup> de los residuos de la regresión, es decir, la raíz cuadrada de la media aritmética de los cuadrados de dichos valores:

$$S_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n - k}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - k}} \quad (9.7)$$

En la regresión de la tasa de mortalidad con el producto nacional bruto per cápita, el valor obtenido (*Root MSE*) es 29,8<sup>8</sup>. Esto quiere decir que, utilizando la segunda de las variables, puede predecirse la primera con un error por término medio del 29,8%, expresado en tantos por mil, por cuanto las unidades de la variable dependiente están medidas de esta forma. O también puede considerarse como una medida de la desviación de las predicciones, de modo que podría pensarse que una gran mayoría de los valores observados de la variable dependiente estarían comprendidos en el rango  $\pm 2S_e$  en torno a los valores predichos. En el ejemplo actual el rango aproximado estaría cifrado en  $\pm 60\%$ , una cifra nada desdeñable, como se aprecia al pensar sobre ello.

Es evidente que esta medida de ajuste estará muy determinada por las unidades que se empleen en la variable dependiente *y*, en consecuencia, si, en lugar de haber medido la tasa de mortalidad en tantos por mil, se hubiera hecho en tantos por cien, el valor de  $S_e$  habría cambiado. Por eso, y para poder comparar las regresiones efectuadas entre variables muy diferentes, es muy útil el empleo de medidas relativas de ajuste. Una de las propiedades que estas han de tener es un conocimiento preciso de sus límites, con el fin de saber el grado de ajuste que tiene la recta hallada.

La medida de ajuste relativo más empleada en la regresión es el  $R^2$  o coeficiente de determinación. En realidad, no es más que el coeficiente de correlación al cuadrado, pero se puede interpretar mejor sabiendo que es el cociente entre dos sumas cuadráticas: la correspondiente a la regresión y la correspondiente a la variable dependiente.

$$R^2 = \frac{SCReg}{SCT} \quad (9.8)$$

<sup>7</sup> En mínimos cuadrados ordinarios, se divide por *n* menos el número de parámetros de la regresión (*k*) para obtener el estimador insesgado de  $\sigma_e$ .

<sup>8</sup> Stata calcula  $S_e$  como si los datos procedieran de una muestra. Por ello, en lugar de dividir por *n*, lo hace por los grados de libertad (*n*-2), de ahí que, cuando se tenga un bajo número de casos, pueda haber una divergencia importante en el resultado entre esta fórmula y la que utiliza el programa estadístico.

La *suma cuadrática total* se define como la suma de las distancias al cuadrado de los  $n$  valores de la variable dependiente con respecto a la media aritmética:

$$SCT = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (9.9)$$

En cambio, la *suma cuadrática de la regresión* es la suma de las distancias al cuadrado de los  $n$  valores predichos por la regresión con respecto a la media aritmética:

$$SCReg = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (9.10)$$

La resta de ambas sumas cuadráticas es la ya conocida como suma de los residuos al cuadrado, o *suma cuadrática residual* (9.11), suma de la diferencia al cuadrado entre valores reales y predichos, que por el criterio de mínimos cuadrados ordinarios ha de ser la menor posible con los datos disponibles.

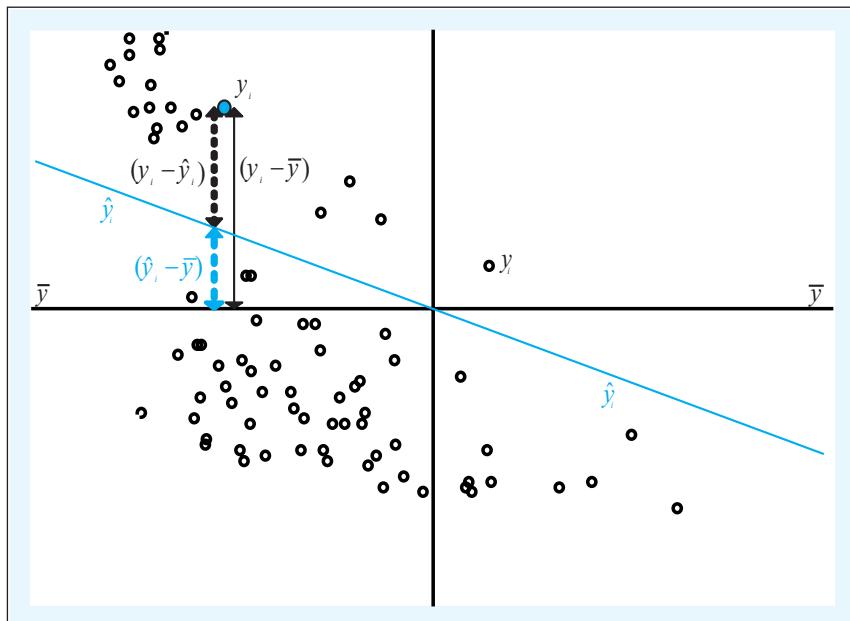
$$SCRes = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (9.11)$$

Por tanto, entre ellas se establece la siguiente igualdad:

$$SCT = SCReg + SCRes \quad (9.12)$$

Puesto que las sumas cuadráticas sólo pueden ser positivas, el valor de la *SC* de la regresión ha de tener como límite superior el valor de la *SC* total, en el supuesto de que todos los valores empíricos se encontraran sobre la línea recta del modelo. Dada esta situación, el valor de  $R^2$  sería igual a 1. En cambio, si la *SC* de la regresión fuera igual a 0, esto es, todos los valores predichos por la recta fueran igual a la media de la variable dependiente, en este caso  $R^2$  sería igual a 0. Ambos, 0 y 1, son los límites entre los que se mueve este coeficiente denominado de determinación.

**GRÁFICO 9.7.** Gráfico de la descomposición de la varianza en la regresión



Otro modo de concebir esta partición de la varianza total de la variable dependiente es de modo gráfico. En el gráfico 9.7 se han representado mediante dos líneas verticales dos distancias desde un valor empírico hasta la recta horizontal de las medias. La de la derecha aparece completa, en tanto que la de la izquierda queda dividida en dos fragmentos: el que va desde el punto empírico hasta la recta de la regresión y el que va desde esta hasta la recta horizontal de las medias.

En el ejemplo del anterior apartado (ilustración 9.5), la suma cuadrada de la variable dependiente asciende a 185.176 y se descompone en la debida a la regresión (*Model* = 75.977) y en la que no se puede explicar la regresión, esto es, la residual (*Residual* = 109.199). Por tanto, el  $R^2$  tiene el valor de 0,41, que puede ser interpretado diciendo que el 41% de la variación de la variable resultado es explicado por su regresión lineal con su predictor. Más concretamente, en este modelo, el 41% de la variación total de la mortalidad infantil se puede explicar con el producto nacional bruto per cápita.

Este  $R^2$  depende del número de variables introducidas en la regresión. Por ello, se emplea un ajuste, al que se reconoce como  $R_{adj}^2$ , tanto mayor cuanto: a) menor sea el  $R^2$  original, y b) mayor sea el número de variables en relación con el número de casos. Para su cálculo, al coeficiente de determinación original hay que restarle el producto del complemento del  $R^2$

$(1-R^2)$  y del cociente entre el número de parámetros ( $k$ ) menos 1 (esto es, el número de variables, por considerar la constante como un parámetro) y el número de casos menos el de parámetros:

$$R_{aj}^2 = R^2 - (1 - R^2) \frac{k - 1}{n - k} \quad (9.13)$$

En el ejemplo contemplado  $(1-R^2)$  es igual a 0,59 y el cociente variables/casos 1/124. El producto de ambos es tan bajo que el  $R^2$  original apenas se reduce 5 milésimas.

## 9.4. Inferencias en la regresión simple

De la ilustración 9.5 tan sólo se ha explicado la interpretación de los parámetros  $b_0$ ,  $b_1$ , las sumas cuadráticas y el coeficiente de determinación, pues el resto tiene relación con inferencias estadísticas cuya dificultad implica que se le dedique específicamente este apartado. Además, hasta el momento los estadísticos contemplados se han calculado en  $y$  para la muestra obtenida. Sin embargo, es común en estadística extrapolar los datos obtenidos en la muestra a la población de la que proceden. En regresión pasa lo mismo, los cálculos que se obtienen proceden generalmente de una muestra y para trasladarlos a la población hay que tener en cuenta las leyes de la inferencia estadística.

Ante todo, es preciso convertir la ecuación muestral de la regresión a su expresión poblacional:

$$\hat{y}_i = \beta_0 + \beta_1 x_i \quad (9.14)$$

En el capítulo relacionado con las comparaciones se introdujo el concepto de prueba estadística y cómo se procede para enunciar hipótesis estadísticas nulas y alternativas. Y de la misma forma que pueden realizarse pruebas de significación con medias, proporciones, medianas, varianzas..., también pueden efectuarse con los parámetros de la regresión, en cuyo caso se han de formular del siguiente modo:

$$\begin{cases} h_0 : \beta_1 = 0 \\ h_1 : \beta_1 \neq 0 \end{cases} \quad \begin{cases} h_0 : \beta_0 = 0 \\ h_1 : \beta_0 \neq 0 \end{cases} \quad (9.15)$$

Estadísticamente, se sabe que, siempre y cuando se cumplan una serie de supuestos que se verán detenidamente en el próximo capítulo, la distribución muestral del estadístico  $b_0$  es t-Student con  $(n-1)$  grados de libertad, media  $\beta_0$  y desviación típica:

$$\sigma_{b_1} = \frac{\sigma_e}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (9.16)$$

Por su parte, en relación con el comportamiento del estadístico  $b_0$ , también con los mismos supuestos, su distribución muestral es t-Student con  $(n-1)$  grados de libertad, media  $\beta_0$ , pero con esta otra desviación típica:

$$\sigma_{b_0} = \sigma_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (9.17)$$

Ahora bien, más importante en este contexto que las fórmulas es la interpretación y uso de estos errores típicos. Para explicarlo, es conveniente volver al resultado anterior de la regresión de la tasa de mortalidad infantil con el producto bruto nacional per cápita:

#### ILUSTRACIÓN 9.6. Regresión simple

Source	SS	df	MS	Number of obs	=	125
Model	75976.9095	1	75976.9095	F(	1,	123) = 85.58
Residual	109199.139	123	887.797874	Prob > F	=	0.0000
Total	185176.048	124	1493.35523	R-squared	=	0.4103
				Adj R-squared	=	0.4055
				Root MSE	=	29.796
<hr/>						
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	-.0031526	.0003408	-9.25	0.000	-.0038272	-.002478
_cons	65.15607	3.604551	18.08	0.000	58.02108	72.29106

El error típico se utiliza para la prueba de hipótesis. Si el coeficiente  $b_0$  o el  $b_1$  son divididos por su error típico, se obtiene su valor típico en la distribución muestral ( $t$ ), del que sabiendo que adopta la forma de distribución de la  $t$  de Student puede obtenerse su probabilidad ( $p>|t|$ ). Desde el punto de vista práctico, en tanto se cumplan los supuestos explicados más adelante, si la probabilidad es menor del 5% o del 1%, según sea el nivel de significación adoptado, se puede rechazar la hipótesis nula sobre el coeficiente.

Otro modo de plantear lo mismo es a través de los intervalos de confianza, que se obtienen con las siguientes operaciones:

$$\begin{aligned} b_0 &\pm t_c \sigma_{b_0} \\ b_1 &\pm t_c \sigma_{b_1} \end{aligned} \quad (9.18)$$

... siendo  $t_c$  el valor crítico de dos colas de la distribución  $t$  de Student con el nivel de confianza igual a  $c$  y  $n-2$  grados de libertad.

En el ejemplo de la ilustración precedente los valores de los intervalos se obtendrían con las siguientes operaciones:

$$\begin{aligned} b_0 &= 65,154 \pm 1,96 \times 3,605 \\ b_1 &= -0,003 \pm 1,96 \times 0,003 \end{aligned} \quad (9.19)$$

Como en ambos casos los dos límites del intervalo tienen el mismo signo (positivo para  $b_0$  y negativo para  $b_1$ ), la hipótesis nula puede ser rechazada con un 95% de nivel de confianza (el complementario del 0,05, como nivel de significación). Sólo cuando un límite es negativo y el otro positivo, no es posible rechazar la hipótesis nula, ya que el valor 0 se encontraría dentro del intervalo con estas últimas características enunciadas.

Una alternativa al test de Student para los coeficientes de la regresión es la prueba de Wald, que permite comprobar más de un coeficiente al mismo tiempo. La hipótesis nula se convertiría de este modo en esta fórmula:

$$\left\{ \begin{array}{l} h_0 : \beta_0 = \beta_1 = \dots = \beta_k = 0 \\ h_1 : \exists \beta_i \neq 0 \end{array} \right. \quad (9.20)$$

En la regresión simple la prueba de Wald no es muy importante, ya que sólo existen dos coeficientes (la constante y el correspondiente al predictor), cuya comparación no tiene sentido por ser de tan diferente interpretación.

Stata permite el empleo del test de Wald a través de la instrucción *test*, en la que deben especificarse las variables cuyas hipótesis quieren comprobarse, junto con el valor, en el caso de que se quieran comprobar valores distintos de 0.

Veáñese dos ejemplos de esta instrucción. En primer lugar, si se quiere hacer la hipótesis de que tanto la constante como el coeficiente son nulos. En cuyo caso:

```
test _cons pnbppa
```

... presenta el siguiente resultado:

### ILUSTRACIÓN 9.7. Prueba de hipótesis sobre los parámetros de la regresión

```
( 1) _cons = 0
( 2) pnbppa = 0

F( 2,    123) = 171.17
Prob > F = 0.0000
```

Donde es obvio que se puede rechazar la hipótesis nula de que ambos coeficientes son iguales a 0.

Esta misma instrucción permite pruebas en las que la igualdad sea inicialmente distinta de 0<sup>9</sup>. Así, si se quiere probar la hipótesis de que el coeficiente correspondiente al producto nacional bruto per cápita es igual a -0,003, se escribiría la siguiente instrucción:

```
test pnbppa = -0.003
```

En cuyo caso, el resultado sería el siguiente:

### ILUSTRACIÓN 9.8. Prueba de hipótesis específica de un parámetro de la regresión

```
( 1) pnbppa = -.003

F( 1,    123) = 0.20
Prob > F = 0.6551
```

Obvio es en esta ocasión que no puede rechazarse la hipótesis propuesta, puesto que la probabilidad del estadístico  $F$  es demasiado alta como para arriesgarse a hacerlo.

Otro estadístico de significación en la regresión es el cociente  $F$ . Este se obtiene dividiendo la media cuadrática de la regresión y la residual, obtenidas a su vez al dividir por sus correspondientes grados de libertad las sumas cuadráticas ya explicadas anteriormente. En el caso de la variación de la regresión (*Model*), sus grados de libertad son igual al número de parámetros menos 1, y en el caso de la residual, los grados de libertad se obtienen restando al número de casos el número de parámetros. Así, las fórmulas

---

<sup>9</sup> El caso más común podría ser la hipótesis de que el coeficiente de regresión es igual a la unidad, que equivaldría a decir que la variable independiente tiene un efecto directo sobre la dependiente de igual magnitud. Por ejemplo, cada año de estudios del padre o la madre implica un año de estudio en su hija o hijo.

completas de las medias cuadráticas quedarían como sigue. En el caso de la del modelo adoptaría la siguiente expresión:

$$MCReg = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{k - 1} \quad (9.21)$$

En el de la residual respondería a esta otra:

$$MCRes = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - k} \quad (9.22)$$

Y con el cociente de ambas medias cuadráticas se obtiene el estadístico F:

$$F = \frac{MCReg}{MCRes} \quad (9.23)$$

Se sabe que este nuevo estadístico tiene una distribución F de Snedecor con  $(k-1)$  y  $(n-k)$  grados de libertad, a partir de la cual puede calcularse la probabilidad de que se dé este valor o uno mayor.

La hipótesis nula con la que se trabaja en este caso es la de que el parámetro es igual a 0. Es similar a la que se formulaba anteriormente con la distribución de Student. De hecho, en la regresión simple se da la siguiente relación entre ambos estadísticos de significación:

$$F = t^2 \quad (9.24)$$

En consecuencia lógica, siempre que sea significativo el coeficiente  $b_1$ , también lo será la regresión en la que está incluido.

## 9.5. Regresión múltiple

Además de la constante y una variable independiente, en la regresión pueden introducirse otras variables con una doble finalidad: la de mejorar la predicción de la variable dependiente y la de controlar la influencia que sobre ella tienen el resto de las variables incluidas en la regresión.

Los valores teóricos o esperados del modelo responden en este caso a la siguiente ecuación:

$$\hat{y}_i = b_0 + b_1 x_{1i} + \dots + b_k x_{ki} = \mathbf{x}_i \mathbf{b} \quad (9.25)$$

...siendo  $k$  el número de variables independientes.

Para obtener una regresión múltiple con el programa Stata basta con añadir a continuación de la primera variable independiente tantas como se deseen introducir, con la limitación de que no pueden incluirse más del número de casos de que se disponga.

Por tanto, a la regresión anterior podría añadirse una nueva variable y, de este modo, la regresión ajusta el plano que pasa lo más cerca posible de los puntos que se alzan en un plano tridimensional, dos de cuyas dimensiones son las variables independientes y la tercera es la variable dependiente. En este caso, además del producto interior bruto, se introduce en la regresión el porcentaje de este que es debido al sector agrícola, con la suposición de que los países en los que tiene más peso el primer sector poseen una tasa de mortalidad infantil superior.

```
regress tmi pnbppa pibag
```

El formato de la salida es idéntico al de la regresión simple. Lo único que lo diferencia es la adición de una línea correspondiente a una variable con el valor de su coeficiente, error típico, significación e intervalos de confianza.

### ILUSTRACIÓN 9.9. Regresión múltiple

Source	SS	df	MS	Number of obs	=	125
Model	102524.348	2	51262.174	F( 2, 122)	=	75.67
Residual	82651.70	122	677.472951	Prob > F	=	0.0000
Total	185176.048	124	1493.35523	R-squared	=	0.5537
				Adj R-squared	=	0.5463
				Root MSE	=	26.028
<hr/>						
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	-.0015201	.0003958	-3.84	0.000	-.0023036	-.0007366
pibag	1.323406	.2114112	6.26	0.000	.9048969	1.741916
_cons	29.5263	6.504691	4.54	0.000	16.64962	42.40299

Es de advertir, sin embargo, que, a pesar de que el formato es el mismo, muchos datos han cambiado. Para comprender mejor la regresión múltiple,

es conveniente fijarse en estos cambios, pero previamente es preciso reparar en lo que permanece inalterado.

Puede verse también cómo lo único que no cambia, además del número de casos<sup>10</sup>, es la suma y media cuadráticas de la variación total. Eso es así por una razón muy sencilla, la variable resultado no cambia y, por tanto, la suma de las desviaciones de los valores de esta variable con respecto a su media es constante cualquiera que sea el número de variables independientes que se introduzcan en el modelo<sup>11</sup>.

En cambio, son diferentes las sumas cuadráticas de la regresión y la residual. Es obvio que cuantas más variables incorporemos a una regresión, el ajuste será tanto mayor, y sólo en el caso de introducir una variable nada relevante para la dependiente, el valor de la suma cuadrática de la regresión sería igual al anterior sin la nueva variable introducida. A la inversa, la suma de residuos al cuadrado se irá haciendo cada vez más pequeña a medida que se vayan introduciendo más variables independientes relevantes.

Los grados de libertad siguen la tendencia opuesta. Por cada variable introducida en la regresión, los grados de libertad de su suma cuadrática aumentan en una unidad, mientras que los de la residual disminuirán en un punto por parámetro calculable.

Al cambiar tanto las sumas cuadráticas de la regresión y de los residuos como sus respectivos grados de libertad, es obvio que también han de cambiar las medias cuadráticas y los estadísticos  $F$  y  $R^2$ , que de ellos se derivan por cálculo, como es obvio al examinar las fórmulas (9.23) y (9.8). Estos dos últimos son mayores a medida que el modelo incorpora más variables.

Recuérdese que  $R^2$  es el coeficiente de determinación y expresa el porcentaje de la varianza de la variable dependiente que es explicado por el conjunto de independientes, mientras que  $F$  es un estadístico de significación que es capaz de comprobar simultáneamente la hipótesis de que todos los coeficientes de la regresión sean igual a 0, es decir, prueba la certidumbre de que sea cierta la siguiente relación:

$$h_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0 \quad (9.26)$$

<sup>10</sup> En la mayor parte de las ocasiones también cambia el número de casos. En general, siempre que las nuevas variables incorporadas posean casos perdidos entre sujetos con valores válidos en las variables de modelos más simples, el número de casos será inferior en tantos enteros como casos perdidos con esas características haya.

<sup>11</sup> Podría ocurrir, sin embargo, que el número de casos en la regresión descendiera al incluir más variables independientes que contribuyeran a disminuir el tamaño muestral. En estos casos descendería necesariamente la suma cuadrática total.

En el ejemplo presente, el  $R^2$  es igual a 0,55. Comparado con el de la regresión simple, que era igual a 0,41, se ha producido un incremento de 14 puntos en la explicación de la variable dependiente al introducir la segunda independiente. Precisamente la raíz cuadrada de esta diferencia es lo que se denomina coeficiente parte de correlación. Por otro lado, se puede rechazar con tranquilidad la hipótesis de que todos los coeficientes de las variables independientes son nulos, es decir, igual a 0, puesto que el valor de la F es muy alto (75,7) y, por tanto, es muy improbable.

En cambio, los parámetros o coeficientes y sus correspondientes errores típicos cambian su valor, no siempre en la misma dirección, a medida que se introducen nuevas variables en la regresión. En algunos casos, el coeficiente se hace mayor y, en otros, se hace menor. Todo depende de la especial pauta de relación que tenga el conjunto de las variables predictoras.

En el caso del ejemplo presente, ha de recordarse que la ecuación lineal del modelo simple de regresión es:

$$\hat{y}_i = 65,2 - 0,003x_i \quad (9.27)$$

Es decir, la mortalidad infantil es inicialmente de 65,2% en un supuesto país cuyo producto interior per cápita fuera nulo, y por cada dólar que aumenta el PIB, esa cifra disminuye en 3 milésimas.

Sin embargo, al introducir la variable porcentaje del producto interior bruto atribuido a la agricultura, la ecuación cambia a la siguiente:

$$\hat{y}_i = 29,5 - 0,002x_{1i} + 1,3x_{2i} \quad (9.28)$$

En este caso, un país sin PIB ni producto agrario tendría una mortalidad infantil promedio del 29,5%, y por cada dólar de aumento en el producto interior bruto bajaría la tasa de mortalidad 2 milésimas y, en condiciones iguales de renta, por cada punto que subiera el porcentaje del producto agrícola, la tasa de mortalidad infantil subiría un 1,3%.

Dos aspectos son suficientemente importantes a la hora de interpretar estos coeficientes:

1. Lo primero es que el valor del coeficiente depende de las unidades en las que estén medidas principalmente las variables predictoras, pero también de la variable resultado, aunque esta sea menos trascendente, porque es única, en tanto que los coeficientes de cada variable independiente se refieren a unidades distintas entre sí.

Para solucionar este problema se puede recurrir a la estandarización de los coeficientes, operación que puede plantearse de dos formas, que dan lugar al mismo resultado. La primera es más compleja de rea-

lizar, pero refleja bastante mejor el planteamiento del procedimiento: se trata de convertir todas las variables de la regresión en valores típicos, esto es, media 0 y desviación típica 1. Si se realiza la regresión con las variables tipificadas, los coeficientes resultantes serían los coeficientes estandarizados, que podrían interpretarse como el cambio en unidades de desviación típica de la variable dependiente, que implica el cambio en una unidad de desviación típica de la variable independiente en cuestión, manteniendo constante el resto de las variables. La segunda es más inmediata en su cálculo y consiste en multiplicar el coeficiente original por la desviación típica de la variable dependiente y dividirlo por el de la independiente. Se les denomina coeficientes beta, aunque no deban confundirse con los parámetros  $\beta_k$  de la población que se estiman a partir de los estadísticos  $b_k$  de la muestra.

$$\text{beta}_k = b_k \frac{S_{x_k}}{S_y} \quad (9.29)$$

En Stata, en la regresión múltiple, como en la simple, pueden obtenerse estos coeficientes estandarizados, en lugar de los originales, simplemente añadiendo la opción *beta*. Además, como ya se han obtenido los datos generales del modelo de varianza, puede añadirse otra opción para que no muestre el análisis de varianza de la regresión ni el coeficiente de determinación. Se trata de la opción *nohead*.

```
regress tmi pnbppa pibag, beta nohead
```

El resultado presenta directamente los coeficientes estandarizados:

#### ILUSTRACIÓN 9.10. Coeficientes estandarizados de la regresión múltiple

	Coef.	Std. Err.	t	P> t	Beta
tmi					
pnbppa	-.0015201	.0003958	-3.84	0.000	-.3088509
pibag	1.323406	.2114112	6.26	0.000	.5033717
_cons	29.5263	6.504691	4.54	0.000	.

Estos coeficientes estandarizados, que tienen que variar entre -1 y 1, son útiles —siempre y cuando se cumplan los supuestos de regresión, especialmente los de homocedasticidad, ausencia de multicolinealidad y correcta especificación— como un medio adicional al valor *t* de juzgar la importancia de la asociación directa de cada predictor sobre el resultado. Obviamente, valores próximos a 0 reflejan una variación

conjunta pequeña de las variables, en tanto que, cuanto mayor valor absoluto (prescindiendo del signo) posea, tanto mayor cambio conjunto entre ellas será presumible, manteniendo constante las demás<sup>12</sup>.

2. El segundo aspecto que ha de ser tenido en cuenta en la interpretación de estos coeficientes es que se sobreentiende que se mantiene constante el resto de las variables introducidas en la ecuación. En concreto, con el ejemplo actual, en el caso de que un conjunto de países tuvieran el mismo producto nacional bruto per cápita, por cada punto que suba el porcentaje del que le corresponde a la agricultura, la mortalidad infantil, medida en tantos por mil, es 1,3 puntos superior. O, dicho de otro modo, entre dos naciones con el mismo producto nacional bruto per cápita y una de ellas diez puntos por encima en porcentaje agrícola, esta última tendría teóricamente una tasa de mortalidad en tantos por mil 13 puntos superior. Por esta razón sustancial, son distintos de los de la regresión simple y, no sólo eso, sino que si se introdujera una tercera variable, como se va a ver a continuación, también cambiarían porque esta también se supone que se deja constante para el cálculo del nuevo coeficiente obtenido.

```
regress tmi pnbppa pibag lintfno, nohead
```

### ILUSTRACIÓN 9.11. Regresión sin cabecera

	tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
+	pnbppa	.0020745	.000961	2.16	0.033	.0001719 .0039772
	pibag	1.142371	.2033324	5.62	0.000	.7397867 1.544955
	lintfno	-.1533458	.0376802	-4.07	0.000	-.22795 -.0787416
	_cons	34.82963	6.23195	5.59	0.000	22.4908 47.16846

Como puede apreciarse, se ha introducido una tercera variable independiente (*lintfno*, las líneas telefónicas por mil habitantes) y el coeficiente correspondiente al porcentaje agrícola del PIB ha descendido a dos décimas, hasta 1,1, y el del producto nacional bruto (*pnbppa*) se ha convertido en positivo (en igualdad del sector agrar-

<sup>12</sup> Dicho esto, hay quienes emplean esta medida para comparar los efectos *causales directos* entre las variables presentes en un modelo. No obstante, son inmensa mayoría y les asiste más la razón a quienes piensan que no se puede inferir que la variable A tenga un efecto *causal mayor* que la B, a partir de la constatación de que un cambio de una desviación en B tenga un efecto en *x* unidades de desviación típica de la variable dependiente, mientras que el efecto proporcionado por la A sea de *x+a*, es decir, también mayor. Un buen ejemplo de ello es el próximo ejemplo de regresión. Sería una barbaridad decir que el número de líneas telefónicas son la *causa mayor* de descenso en la tasa de mortalidad infantil.

rio y del número de teléfonos, el efecto del producto nacional bruto no va en la dirección esperada). Es evidente que el mejor modo de hacer comparaciones es mediante los coeficientes beta, que se obtienen al expresar junto a la regresión la opción correspondiente:

```
regress tmi pnbppa pibag lintfno, nohead beta
```

### ILUSTRACIÓN 9.12. Coeficientes estandarizados sin cabecera

	Coef.	Std. Err.	t	P> t	Beta
pnbppa	.0020745	.000961	2.16	0.033	.4221748
pibag	1.142371	.2033324	5.62	0.000	.4353203
lintfno	-.1533458	.0376802	-4.07	0.000	-.8143264
_cons	34.82963	6.23195	5.59	0.000	.

Se puede apreciar cómo —con todas las precauciones que habría que adoptar—<sup>13</sup> la variable con beta más alta es la última introducida, en cierta medida porque está reflejando la extensión del nivel tecnológico en cada país, no porque los teléfonos sean la causa del descenso de la mortalidad infantil. Con esta interpretación y todas las cautelas necesarias, se podría afirmar que, manteniendo constantes el porcentaje de producción agraria y el número de líneas telefónicas por habitante, a mayor renta per cápita corresponde parádójicamente mayor mortalidad infantil.

Siguiendo con los coeficientes, hay que terminar diciendo que, al igual que en la regresión simple, de cada coeficiente puede calcularse su error típico con fórmulas similares a las expresadas en (9.16) y (9.17). El cociente entre cada coeficiente y su error típico posee en muestras aleatorias una distribución de *t* de *Student* con *n-k-1* grados de libertad. Cuando el número de casos menos el de parámetros es superior a 30, entonces la distribución puede considerarse normal, y si el valor absoluto del mencionado cociente es superior a 1,96, podría considerarse significativo con un riesgo de error estadístico de tipo I, inferior al 5%. En este ejemplo se ve que todos los coeficientes son significativamente distintos de 0, lo cual no significa necesariamente que la relación sea fuerte. En el caso de muestras grandes, relaciones muy débiles pueden mostrar coeficientes significativos.

<sup>13</sup> Sobre todo, la precaución más importante de todas es la alta correlación entre estos tres predictores, especialmente, entre el producto nacional bruto per cápita y el número de líneas telefónicas, que es mayor de 0,95. Véase el problema de la multicolinealidad en el próximo capítulo.

## 9.6. Regresión con variables ficticias

Aunque los mecanismos matemáticos de la regresión sean propios de variables cuantitativas, también se permite la introducción de variables cualitativas, siempre y cuando se tomen precauciones. Por ejemplo, teniendo una variable en la base de datos como el continente, que incluye cinco valores arbitrariamente codificados desde al 1 (Europa) hasta el 5 (Oceania), ningún sentido tendría introducirla como variable independiente. En cambio, si tomamos uno de los valores de esta variable y se transforma en una nueva variable dicotómica con valores 1 y 0, los coeficientes de la regresión y esta misma adoptan un significado interpretable, puesto que la unidad representa la característica que represente al valor. Un ejemplo puede aclarar lo que se acaba de decir. Si se selecciona el valor “África”, codificado como 3, dentro de la variable *conti*, se transforma en 1, y el conjunto de países que no están situados en el continente africano se les otorga el valor de 0, el coeficiente propio de la nueva variable *africa* significará la diferencia media de valores en la variable dependiente, la tasa de mortalidad infantil en este caso, entre los países africanos y el resto, manteniendo constante los valores del resto de las variables incluidas en la regresión.

Para realizar la regresión con este tipo de variables, la solución más evidente es la de crear la nueva variable y, una vez que ya está creada, se introduce en la regresión:

```
generate africa=(conti==3) if conti<
regress tmi pnbppa africa
```

Tras la ejecución de estas dos instrucciones, la regresión resultante es la siguiente:

### ILUSTRACIÓN 9.13. Regresión múltiple con variable ficticia

Source	SS	df	MS	Number of obs	=	125
Model	129825.14	2	64912.5701	F( 2, 122)	=	143.08
Residual	55350.9078	122	453.695966	Prob > F	=	0.0000
Total	185176.048	124	1493.35523	R-squared	=	0.7011
				Adj R-squared	=	0.6962
				Root MSE	=	21.3
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	-.0018975	.0002695	-7.04	0.000	-.002431	-.001364
africa	50.70149	4.653904	10.89	0.000	41.48862	59.91436
_cons	41.61537	3.362865	12.37	0.000	34.95824	48.2725

El coeficiente correspondiente a la variable *africa*, que tiene el valor 50,7, indica que, en término medio y controlando por la variable del producto nacional bruto per cápita, los países africanos tienen una tasa de mortalidad infantil 50,7 puntos (en tantos por mil) por encima de los países ubicados en otros continentes.

Aunque sean algo complejas las instrucciones, es de especial interés ver el resultado gráfico de esta operación, por cuanto facilita la correcta interpretación de lo que sucede cuando se introduce una variable ficticia en una regresión. Lo primero que hay que hacer es generar dos predicciones distintas: una para los países africanos (*ptmi1*) y otra para los no africanos (*ptmi0*). Para una representación diferenciada, también conviene desdoblar la variable original en los del continente negro (*tmia*), por un lado, y en los habitantes de otros continentes (*tmir*). Estas cuatro variables pueden representarse sobre el producto nacional bruto per cápita, dos de ellas (las predicciones) en formato de línea y las otras dos (las variables reales) en formato de puntos (países africanos) o cuadrados (resto de los países).

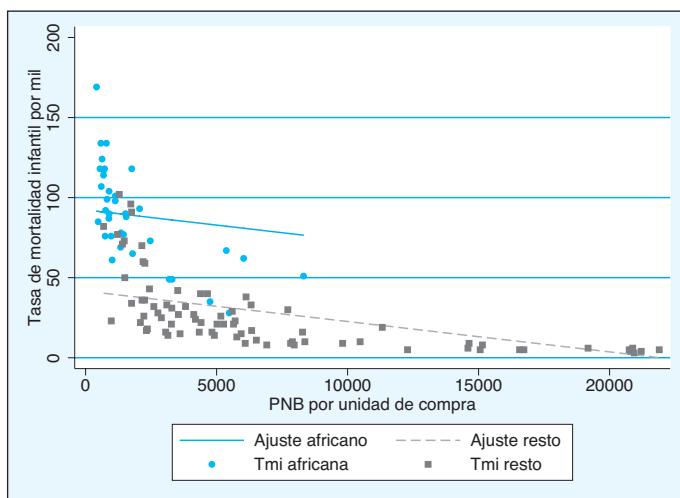
```

predict ptmi0 if !africa
label var ptmi0 "Ajuste resto"
predict ptmi1 if africa
label var ptmi1 "Ajuste africano"
generate tmia=tmi if africa
label var tmia "Tmi africana"
generate tmir=tmi if !africa
label var tmir "Tmi resto"
scatter ptmi1 ptmi0 tmia tmir pnbppa if e(sample) & ptmi0>0, ///
    connect (l l .) lpattern(solid dash) symbol(i i o s) ///
    ytitle("Tasa de mortalidad infantil por mil") ///
    sort(pnbppa) name(G21, replace)

```

En el gráfico puede verse claramente que la introducción de una variable dicotómica genera dos predicciones paralelas. En la línea continua se encuentra la de los países africanos, mientras que la discontinua se refiere al resto de los continentes. Esta última nace en el eje de ordenadas en la constante de la regresión (41,6), que es el valor esperado de la tasa de mortalidad infantil de un país no africano. En cambio, la línea africana arranca 50,7 puntos más arriba (este es el coeficiente de la variable ficticia *africa*), esto es, en torno a los 92%.

**GRÁFICO 9.8. Representación gráfica de una regresión con variable dicotómica**



Ambas rectas son paralelas y su inclinación refleja el efecto del producto nacional en la mortalidad. Se trata de rectas descendentes (coeficiente negativo: -001) en la medida en que esta variable tiene una influencia positiva en el descenso de la mortalidad infantil. Por cada mil dólares, baja prácticamente dos puntos la tasa. Este modelo asume que el efecto del producto nacional bruto es igual en África que en el resto de los continentes; lo que puede ser dudoso. Más adelante se verá cómo realizar una regresión que no asuma que ambas rectas sean paralelas.

Al haber transformado sólo uno de los cinco valores originales de la variable nominal se pierde información. No se sabe cuál es el efecto de los otros continentes. La solución está en crear tantas variables como valores -1 disponga la variable. En este caso, puesto que hay cinco continentes, se deberían crear para disponer de toda la información cuatro variables y dejar uno de los valores como referencia. Puede ser cualquiera, pero para obtener una regresión con similar información, se va a dejar como categoría base el valor “Africa” de la variable continente, codificado como el valor 3. Aunque luego se muestre una instrucción específica para ello, puede recordarse lo visto en el capítulo de transformaciones y crearse mediante una instrucción recursiva en una sola instrucción. Por ejemplo, de este modo<sup>14</sup>:

<sup>14</sup> Se presenta un nuevo y más complejo uso del *for*, que implica dos parámetros que cambian de modo paralelo: uno numérico (*num*) y otro textual, que representa textos empleados para generar nuevas variables (*any*). Ambas listas están separadas por \ y terminan con los dos puntos. La primera se llama en la instrucción con *X*, la segunda con *Y*.

for num 1 2 4 5 \ any europa asia america oceania : generate Y=(contि==X) if contि<

A continuación ya puede formularse la regresión con las cuatro nuevas variables creadas, desde *europa* hasta *oceania*:

regress tmi pnbppa europa-oceania

La regresión aparecerá con la variable independiente cuantitativa más las cuatro variables ficticias que se acaban de generar:

#### **ILUSTRACIÓN 9.14. Regresión múltiple con variables ficticias**

Source	SS	df	MS	Number of obs	=	125
Model	132516.756	5	26503.3511	F( 5, 119)	=	59.89
Residual	52659.2923	119	442.515061	Prob > F	=	0.0000
				R-squared	=	0.7156
Total	185176.048	124	1493.35523	Adj R-squared	=	0.7037
				Root MSE	=	21.036
<hr/>						
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	- .0016082	.0002996	-5.37	0.000	-.0022015	-.0010149
europa	-59.75014	6.163719	-9.69	0.000	-71.95492	-47.54536
asia	-45.83031	5.304399	-8.64	0.000	-56.33355	-35.32707
america	-52.55179	5.895357	-8.91	0.000	-64.22519	-40.8784
oceania	-46.6498	13.13383	-3.55	0.001	-72.65608	-20.64351
_cons	91.77734	3.550267	25.85	0.000	84.74745	98.80722

Como, en este caso, la categoría base es el continente africano, todos los coeficientes pertenecientes al resto de los continentes son negativos, porque en todos ellos la tasa de mortalidad infantil es menor, desde Asia, 46 puntos inferior, hasta Europa, 60 puntos por debajo de la tasa por mil africana. Además, es de notar que la influencia del producto nacional per cápita apenas ha cambiado y sigue con un coeficiente significativamente distinto de 0.

El gráfico, que para simplificarse no diferencia los valores empíricos de los distintos continentes, muestra cinco ajustes de líneas distintas, además de la inicial nube de puntos. La continua representa al continente base, África en este caso, que tiene un pronóstico de partida (la ordenada en el origen de las abscisas) de 93%. La línea más cercana es la de Asia (-45,8), seguida muy de cerca por la de Oceanía (-46,6). Más distanciado se encuentra el continente americano, y el que presenta los pronósticos de la tasas de mortalidad infantil menores es Europa.

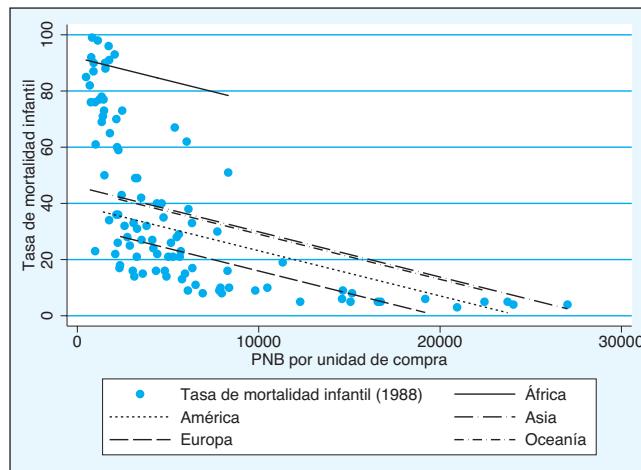
```

predict ptmi
twoway (scatter tmi pnbppa, legend(label(1 "Tasa de mortalidad infantil")) ///
(lfit ptmi pnbppa if africa, lpattern(solid) legend(label(2 "África")))) ///
(lfit ptmi pnbppa if america, legend(label(3 "América")))) ///
(lfit ptmi pnbppa if asia, legend(label(4 "Asia")))) ///
(lfit ptmi pnbppa if europa, legend(label(5 "Europa")))) ///
(lfit ptmi pnbppa if oceania, legend(label(6 "Oceanía")))) ///
if (e(sample)& tmi<100 & ptmi>0) ///
, ytitle("Tasa de mortalidad infantil") name(G9, replace)

```

Para elaborar el gráfico, en lugar de construir una predicción para cada continente con la orden *predict*, tal como se planteó en el gráfico 9.8, se realiza a través de la misma instrucción gráfica con el subtipo *lfit*. La equivalencia es posible porque los ajustes se realizan con *ptmi*. Si se hubiera hecho con *tmi*, las pendientes de las rectas no habrían sido idénticas<sup>15</sup>.

**GRÁFICO 9.9. Representación gráfica de una regresión con más de una variable dicotómica**



<sup>15</sup> Nótese, además, que la instrucción *twoway* contiene seis gráficos distintos que se representan en las mismas coordenadas cada uno con sus propias opciones. Pero, además, termina con dos líneas de códigos que afectan al conjunto del gráfico. La primera es una selección de casos que sólo evita representar los países que no están en la regresión (*e(sample)*), así como los que no tienen una tasa de mortalidad extremadamente alta (*tmi<100*), ni un pronóstico negativo de este indicador (*ptmi<0*); la segunda son las opciones que repercuten de modo general en el gráfico compuesto. Algunas de ellas, como *legend* e *ytitle*, pueden colocarse indistintamente en cada gráfico o en el conjunto.

Volviendo a la regresión, ha de advertirse que el continente africano, en el último ejemplo, está representado en la constante, y que los coeficientes del resto de los continentes significan diferencias con respecto al primero. Como puede apreciarse, todos presentan un coeficiente significativo. ¿Y si deseáramos saber si hay diferencias entre Europa y Asia? ¿O entre este último y el continente americano? Nada más sencillo, emplearíamos la instrucción *test* con el siguiente formato:

```
test europa=asia
test asia=america
```

Como puede deducirse de los resultados, la respuesta a la primera pregunta es positiva y a la segunda es negativa.

#### **ILUSTRACIÓN 9.15. Pruebas de hipótesis sobre igualdad de parámetros en la regresión**

```
( 1) europa - asia = 0
      F(  1,    119) =     5.68
      Prob > F =    0.0187

( 1) asia - america = 0
      F(  1,    119) =     1.29
      Prob > F =    0.2589
```

Otro modo más directo y mucho más cómodo de proceder a la creación de variables ficticias es mediante el uso de factores en Stata. Desde su versión undécima<sup>16</sup> cualquier variable discreta es susceptible de ser empleada en la mayor parte de las instrucciones como un conjunto de variables dicotómicas o ficticias. Cualquier variable que sólo contenga valores enteros puede ser referenciada con el prefijo *i* seguido de un punto, en cuyo caso se crean (*I*-1) variables de la variable categórica. Así, si se desea proceder de este modo con la variable *cont*, debe escribirse *i.cont*.

Un ejemplo de este uso sería el siguiente:

```
regress tmi pnbppa i.conti
```

... que daría lugar al siguiente resultado, similar al que se acaba de exponer:

---

<sup>16</sup> En versiones anteriores también era posible usar factores, aunque para ello era necesario emplear la preinstrucción *xi*. Más detalles de esta posibilidad pueden encontrarse en *help xi*. También en la versión 11 sigue existiendo esta posibilidad, como puede comprobarse en Stata (2009e: 2029-2039).

**ILUSTRACIÓN 9.16. Regresión múltiple con variables ficticias automáticas (factores)**

Source	SS	df	MS	Number of obs = 125		
Model	132516.756	5	26503.3511	F( 5, 119) = 59.89		
Residual	52659.2923	119	442.515061	Prob > F = 0.0000		
Total	185176.048	124	1493.35523	R-squared = 0.7156		
				Adj R-squared = 0.7037		
				Root MSE = 21.036		
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	-.0016082	.0002996	-5.37	0.000	-.0022015	-.0010149
contí						
2	13.91983	5.838961	2.38	0.019	2.358107	25.48156
3	59.75014	6.163719	9.69	0.000	47.54536	71.95492
4	7.198347	6.102943	1.18	0.241	-4.886089	19.28278
5	13.10034	12.66966	1.03	0.303	-11.98684	38.18753
_cons	32.0272	5.476525	5.85	0.000	21.18313	42.87126

Es de notar que la regresión no es exactamente igual que la anterior. No cambia el análisis de varianza, ni el  $R^2$ , ni la raíz del error medio cuadrático. Tampoco cambian los coeficientes, ni las significaciones de la variable cuantitativa. En cambio, aparece la variable *contí* seguida de cuatro valores, que se refieren, respectivamente, a Asia (2), África (3), América (4) y Oceanía (5). No aparece el 1 (Europa), pues, por omisión, este recurso deja como categoría base el primer valor de la variable categórica y, en este caso, corresponde a este continente. Por tanto, ahora los coeficientes no marcan la diferencia de un continente con respecto a África, tal como se hizo en el caso anterior, sino con respecto a Europa y, aunque esta regresión explica lo mismo que la anterior, entre las variables ficticias de los continentes sólo aparecen como significativas las correspondientes a África y a Asia, porque el resto de los continentes no tienen tasas de mortalidad infantil sustancialmente diferentes de las de Europa, que es la considerada en este caso base.

Hay una manera fácil de cambiar la categoría base. Es expresando entre la *i* y el punto una *b* seguida del valor que se desea tomar como tal<sup>17</sup>.

```
regress tmi pnbppa ib3.contí
```

De este modo, la categoría base de la variable *contí* sería la tercera en lugar de la primera. Por consiguiente, en el caso de que se pidiera la regre-

<sup>17</sup> Además, se puede optar por especificar el primero con *ib(first)*, el último con *ib(last)* o el enésimo valor *ib(n)* de la variable de enteros en cuestión. Asimismo, mediante el prefijo *ib(freq)*, Stata se encarga de seleccionar como base la categoría más frecuente.

sión con la misma instrucción que anteriormente, el resultado sería como el expuesto en el primer ejemplo (ilustración 9.14). Lo único que cambiaría sería el nombre de las variables índices o ficticias.

## 9.7. Regresiones con interacción

Hay dos maneras de entender las interacciones. Por un lado, si la relación entre dos variables depende de los valores de una tercera, estamos ante una clara situación de interacción. Siguiendo con el ejemplo anterior simplificando al continente africano frente al resto, habría interacción entre la tasa de mortalidad infantil, el producto nacional bruto per cápita y el continente, si la relación entre las dos primeras variables fuera distinta según si el país se encuentra en África o fuera de ella.

También se dice que hay interacción cuando dos variables tienen una influencia conjunta en una tercera. Y conjunta no significa que ambas puedan influir por su lado, sino que inciden sólo o adicionalmente en el resultado si se da una combinación específica de valores en los predictores. Se podría poner un ejemplo simple diciendo que agua y luz interactúan en el crecimiento de las plantas. Por mucha agua con que se riegue un vegetal por sí solo, o por mucha luz que se le proporcione sin que se le añada agua, este ser vivo no sobrevivirá adecuadamente. Se necesita la acción conjunta de ambos agentes.

El modo de trabajar con interacciones en una regresión es mediante la multiplicación de las variables independientes. El porqué es así se ve manifiestamente en variables dicotómicas, índices o ficticias. Tomando el ejemplo anterior y recodificando las variables *luz* y *agua* a presencia y ausencia, al multiplicar ambas, sólo da el valor unidad en el caso de que ambas sean igual a 1:

**CUADRO 9.1. Cuadro de la interacción entre dos variables**

Aqua	Luz	Interacción	Vive
No (0)	No (0)	No ( $0 \times 0=0$ )	No
No (0)	Sí (1)	No ( $0 \times 1=0$ )	No
Sí (1)	No (0)	No ( $1 \times 0=0$ )	No
Sí (1)	Sí (1)	Sí ( $1 \times 1=1$ )	Sí

Dos son las propiedades que pueden descubrirse en la cuadro 9.1. En primer lugar, que la interacción se obtiene multiplicando los valores de las variables originales y, en segundo lugar, que el valor que mejor predice el producto (si vive o no la planta) es la interacción y no las variables originales.

El que haya interacción no implica necesariamente, como era el caso anterior, que las variables no tengan influencia por separado. A esta última contribución específica de las variables se le denomina efecto principal. A continuación ponemos un ejemplo de esto con las regresiones del apartado anterior.

Como se podrá observar numérica y gráficamente, el continente y el producto nacional bruto per cápita interactúan en su asociación con la tasa de mortalidad infantil. La regresión con interacciones debería prepararse del siguiente modo:

```
generate pnbppaXafrica=pnbppa*africa
regress tmi pnbppa africa pnbppaXafrica
```

El resultado de esta regresión son cuatro coeficientes distintos: de abajo arriba, el más simple sería la constante, que estaría referida a la supuesta tasa de mortalidad infantil de un país no africano con un producto nacional bruto de cero dólares. El siguiente sería el correspondiente a África, que significa que en un país africano de nulo producto nacional bruto, tendría una tasa de mortalidad infantil 69 puntos superior a la del resto de continentes. Otro coeficiente, el de la variable *pnbppa*, indica la pendiente de la recta *pnb*-tasa de mortalidad para los países fuera de África (*africa*=0). Finalmente, la novedad del análisis es que el coeficiente *pnbppaXafrica* es la diferencia de la influencia de la renta en la mortalidad entre no africanos y africanos. Dicho de otro modo, si por cada dólar que aumenta el producto nacional bruto per cápita fuera de África, disminuye la tasa de mortalidad infantil en dos milésimas (-0,0017), en África por cada dólar adicional producido por cada habitante, baja la tasa de mortalidad infantil una centésima (-0,0017+(-0,0090)). Hay, en consecuencia, un efecto diferente del indicador económico en el indicador sanitario según se esté en África o fuera de ella. En la primera tiene mayores efectos positivos, es decir, el incremento del producto reduce más la mortalidad.

#### **ILUSTRACIÓN 9.17. Regresión con interacción de una variable cuantitativa con una ficticia**

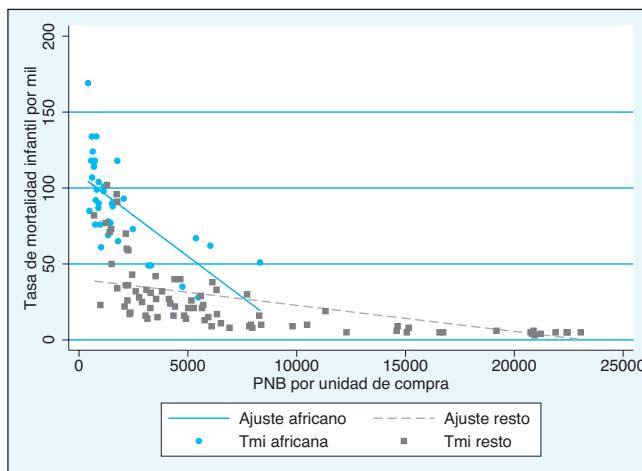
Source	SS	df	MS	Number of obs	=	125
Model	139646.774	3	46548.9246	F( 3, 121)	=	123.71
Residual	45529.2742	121	376.274994	Prob > F	=	0.0000
Total	185176.048	124	1493.35523	R-squared	=	0.7541
				Adj R-squared	=	0.7480
				Root MSE	=	19.398
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	- .0017201	.0002479	-6.94	0.000	- .0022108	- .0012294
africa	68.87338	5.53297	12.45	0.000	57.9194	79.82735
pnbppaXafr~a	- .0090408	.0017696	-5.11	0.000	- .0125441	- .0055375
_cons	39.975	3.07931	12.98	0.000	33.87869	46.0713

Una mejor comprensión de este análisis se consigue representando gráficamente el modelo. Para conseguir el gráfico 9.10, habría que hacer dos predicciones por separado: una para el continente africano y otra para el resto. Asimismo, para representar los puntos de los continentes con distinta forma, se han vuelto a utilizar las variables empíricas de la tasa de mortalidad infantil di-

ferenciada: una con los valores africanos (*tmia*) y otra con el resto de los valores (*tmir*). Las instrucciones para generar este gráfico, inmediatamente después de haber solicitado la regresión con el efecto interactivo incluido, han sido:

```
predict ptmi0X if !africa
label var ptmi0X "Ajuste resto"
predict ptmi1X if africa
label var ptmi1X "Ajuste africano"
scatter ptmi1X ptmi0X tmia tmir pnbppa if e(sample) & ptmi0X>0, ///
connect (l l ..)lpattern(solid dash) symbol(i o s) ///
ytitle("Tasa de mortalidad infantil por mil") sort(pnbppa) name(G27, replace)
```

**GRÁFICO 9.10. Representación gráfica de una regresión con interacción**



Si se tienen dos variables cuantitativas, para introducir la interacción en la ecuación de regresión, es necesario crear la nueva variable mediante su producto. Así, por ejemplo, si se quiere estudiar el efecto de la interacción del producto nacional bruto y la proporción no agrícola de este, bastaría escribir la siguiente instrucción para crear la nueva variable que represente a la interacción entre ambas variables.

```
generate pnbXpibnag=pnbppa*(100-pibag)
```

Esta instrucción genera la variable que representa la interacción señalada. Como puede apreciarse, se ha obtenido el porcentaje del PIB no agrario, restando de 100 el porcentaje correspondiente a la producción agrícola. Una vez que se obtiene la nueva variable, puede ser introducida en la regresión.

```
regress tmi pnbppa pibag pnbXpibnag
```

Y, a continuación, se genera la ecuación con los dos efectos principales (las variables originales) más el efecto interactivo de ambos.

### ILUSTRACIÓN 9.18. Regresión con interacción de dos variables cuantitativas

Source	SS	df	MS	Number of obs	=	125
Model	133472.561	3	44490.8537	F( 3, 121)	=	104.12
Residual	51703.4868	121	427.301544	Prob > F	=	0.0000
				R-squared	=	0.7208
				Adj R-squared	=	0.7139
Total	185176.048	124	1493.35523	Root MSE	=	20.671
<hr/>						
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	-0.0565354	.0064721	-8.74	0.000	-.0693487	-.0437222
pibag	1.683842	.1731587	9.72	0.000	1.341029	2.026655
pnbXpibnag	.0005579	.0000656	8.51	0.000	.0004281	.0006877
_cons	47.45577	5.578999	8.51	0.000	36.41067	58.50087

Los tres coeficientes, además del correspondiente a la constante, son significativos. La mortalidad infantil desciende con el aumento del producto nacional bruto per cápita, asciende a medida que el porcentaje del PIB sea más agrario, pero también aumenta a medida que se produce una conjunción de aumento de rentas y de producción no agraria. Sin embargo, esta ecuación posee un problema importante que se verá cuando se aborden los diagnósticos de la regresión.

También, como es obvio, a partir del primer ejemplo dado se pueden construir interacciones con variables categóricas, siempre y cuando se conviertan en variables ficticias. Incluso, puede hacerse uso de las propiedades de los factores. De este modo, si dos variables con valores enteros se separan con el signo #, Stata las considerará ficticias y construirá automáticamente las interacciones. Adicionalmente, si se separan con dos signos #, no sólo incluirá automáticamente en la regresión las interacciones, sino también los términos principales. También pueden incluirse interacciones (y términos principales) automáticamente con variables cuantitativas, siempre y cuando se precedan con c y un punto. Existen las siguientes modalidades:

vcategorica1#vcategorica2 o vcategorica##vcategorica2  
 vcategorica#c.vcuantitativa o vcategorica##c.vcuantitativa  
 c.vcuantitativa#c.vcuantitativa o c.vcuantitativa##c.vcuantitativa<sup>18</sup>

<sup>18</sup> No es necesario preceder las variables con valores enteros con la i y el punto. A partir de la versión 11, se emplean estos nuevos signos: # para indicar interacción entre variables categóricas y ## para especificar no sólo las interacciones, sino también los efectos principales.

Para comprender sus efectos, no estaría de más ver algunos ejemplos de cada una de estas modalidades. En primer lugar, se verá la interacción de dos variables categóricas:

```
regress tmi pnbppa conti##ocde
```

El continente (*conti*) tiene cinco valores, por tanto se generan cuatro variables ficticias; la variable *ocde* sólo tiene dos codificados como 0 y 1. La interacción debería incorporar, por tanto, cuatro modalidades (*I-1*(*J-1*):

### ILUSTRACIÓN 9.19. Regresión múltiple con interacciones automáticas

note: 3.conti#1.ocde identifies no observations in the sample						
Source	SS	df	MS	Number of obs = 125		
Model	135427.415	9	15047.4906	F( 9, 115) = 34.78		
Residual	49748.633	115	432.596809	Prob > F = 0.0000		
Total	185176.048	124	1493.35523	R-squared = 0.7313		
				Adj R-squared = 0.7103		
				Root MSE = 20.799		
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	-.0021084	.000425	-4.96	0.000	-.0029503	-.0012665
conti						
2	22.07487	7.015219	3.15	0.002	8.179069	35.97067
3	66.06252	6.955569	9.50	0.000	52.28488	79.84016
4	13.71847	7.503153	1.83	0.070	-1.143831	28.58077
5	37.1236	21.64243	1.72	0.089	-5.745878	79.99308
1.ocde	19.84606	9.074104	2.19	0.031	1.872002	37.82011
conti#ocde						
2 1	-21.29399	16.93196	-1.26	0.211	-54.83294	12.24497
3 1	(empty)					
4 1	-2.60739	14.92736	-0.17	0.862	-32.17561	26.96083
5 1	-37.4885	26.58474	-1.41	0.161	-90.14776	15.17076
_cons	26.64775	6.306928	4.23	0.000	14.15494	39.14056
_cons	28.78621	6.041257	4.76	0.000	16.81964	40.75277

Efectivamente, se generan cuatro coeficientes relacionados con el continente, desde 2 hasta 5, que equivalen a Asia, África, América y Oceanía, pues Europa, al estar codificada con el valor más bajo, queda como categoría base. De ellas, destacan Asia y sobre todo África, que tienen coeficientes significativos y positivos, en la medida en que en ambos continentes la tasa

---

Anteriormente, era más complejo, pues había que utilizar \* y | con la preinstrucción *xi*: Si se dispone de una versión anterior, consultese la ayuda de *xi*.

de mortalidad infantil es superior a la que se da en Europa. A continuación aparece la categoría 1 de la variable *ocde*, puesto que el valor más pequeño es el 0, que indica la no pertenencia a esta organización. El significativo coeficiente viene a señalar que, controlando por la renta per cápita, los países de la OCDE en Europa (considerado aquí el continente base)<sup>19</sup> tienen mayor mortalidad infantil que los que no pertenecen a ella. También aparecen las interacciones entre el continente y la OCDE: cuatro en la medida en que responden a la fórmula (I-1)(J-1). Entre ellas, se descarta 3 1 porque no hay país africano que pertenezca a esta organización comercial.

Si se desea introducir en la ecuación de regresión una interacción entre una variable categórica y una cuantitativa, basta con separarlas con doble almohadilla y preceder la segunda con la letra *c* seguida de punto.

```
regress tmi conti##c.pnbppa
```

La regresión incluye tanto los efectos principales como los interactivos, como puede apreciarse a continuación:

#### **ILUSTRACIÓN 9.20. Regresión múltiple con interacción entre variable cuantitativa y cualitativa**

Source	SS	df	MS	Number of obs = 125		
Model	145869.64	9	16207.7378	<i>F</i> ( 9, 115) = 47.42		
Residual	39306.4077	115	341.79485	Prob > <i>F</i> = 0.0000		
Total	185176.048	124	1493.35523	R-squared = 0.7877		
				Adj R-squared = 0.7711		
				Root MSE = 18.488		
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
conti						
2	30.91629	7.637111	4.05	0.000	15.78864	46.04393
3	90.16486	7.682314	11.74	0.000	74.94767	105.382
4	19.00068	8.447558	2.25	0.026	2.26769	35.73366
5	43.94498	21.30635	2.06	0.041	1.741206	86.14875
pnbppa	-.0006379	.0003968	-1.61	0.111	-.0014238	.000148
conti# c.pnbppa						
2	-.0016523	.0006366	-2.60	0.011	-.0029133	-.0003914
3	-.010123	.0017164	-5.90	0.000	-.0135229	-.0067232
4	-.0007495	.000697	-1.08	0.284	-.0021302	.0006312
5	-.0022423	.0013203	-1.70	0.092	-.0048576	.000373
_cons	18.68352	6.310526	2.96	0.004	6.183581	31.18346

<sup>19</sup> En general, el coeficiente de un efecto principal de una variable debe interpretarse teniendo en cuenta que sólo mide el efecto de ella, cuando la otra variable con la que está interactuando posee el valor de 0.

De todos los coeficientes relacionados con la interacción, el más significativo es el correspondiente a África (contí#c.pnbppa=3), que indica que el aumento de la renta per cápita en este continente tiene más efectos positivos (reductores) sobre la mortalidad infantil que en el europeo, como ya se vio en un ejemplo anterior.

Si la expresión de la interacción se hubiera realizado con una sola almohadilla (#), en lugar de dos (##), la regresión no habría incluido los efectos principales de la variable categórica. De este modo, si la orden se hubiera escrito de este modo...

```
regress tmi conti#c.pnbppa
```

... la regresión resultante habría sido esta otra:

### **ILUSTRACIÓN 9.21. Regresión múltiple con interacciones sin efectos principales**

Source	SS	df	MS	Number of obs = 125			
Model	81451.8192	5	16290.3638	F( 5, 119) = 18.69			
Residual	103724.229	119	871.632175	Prob > F = 0.0000			
Total	185176.048	124	1493.35523	R-squared = 0.4399			
				Adj R-squared = 0.4163			
				Root MSE = 29.523			
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]		
conti# c.pnbppa							
1	-.0029641	.0003833	-7.73	0.000	-.0037231	-.0022052	
2	-.0031422	.0006845	-4.59	0.000	-.0044975	-.0017868	
3	.0020754	.0021563	0.96	0.338	-.0021943	.006345	
4	-.0031168	.0007029	-4.43	0.000	-.0045086	-.001725	
5	-.0028194	.0010749	-2.62	0.010	-.0049478	-.0006911	
_cons	61.4754	3.926884	15.66	0.000	53.69978	69.25102	

Como puede fácilmente apreciarse no aparecen los efectos principales de ninguna de las dos variables, lo que conduce a que el coeficiente interactivo correspondiente a África sea positivo, ya que la tasa de mortalidad infantil es muy alta en este continente y al no aparecer este hecho en los efectos principales, lo incorpora en la interacción. Según este modelo, todos los continentes compartirían la constante (lo que es más que dudoso) y el efecto de la renta sobre la tasa de mortalidad sería el indicado por el coeficiente. Nótese que cada continente, incluida la base (Europa), tiene el suyo propio.

## 9.8. Otras relaciones funcionales de la regresión

Mediante mínimos cuadrados ordinarios no sólo pueden ajustarse líneas rectas (planos o hiperplanos cuando se tiene más de una variable independiente) para pronosticar los valores de la variable dependiente en función de la(s) independiente(s). También es posible ajustar curvas que en determinados casos se aproximan más a los valores empíricos que se intentan pronosticar. El procedimiento en Stata pasa por la transformación adecuada de las variables y la posterior introducción de las nuevas variables en la regresión.

Recuérdese que las relaciones funcionales no lineales más frecuentes son:

- a) Las regresiones cuadrática y cúbica.
- b) La regresión inversa.
- c) Las regresiones con variables logarítmicas.

Con el ejemplo de la regresión de la mortalidad infantil sobre el producto nacional bruto per cápita, se verá cómo se opera para obtenerlas y representarlas.

*Regresión cuadrática:* Para producirla, hay que obtener primero el cuadrado de la variable independiente para después introducirlo junto con la variable original. Por ello el primer paso consiste en generar los valores cuadráticos mediante la instrucción *generate*.

```
generate pnb_2=pnbppa^2
```

Y, una vez que se dispone de la nueva variable, esta se introduce en la ecuación de regresión junto con la original.

```
regress tmi pnbppa pnb_2
```

De esta forma, sale una ecuación con tres coeficientes, la constante, el de la variable y el de esta al cuadrado<sup>20</sup>.

---

<sup>20</sup> Otro modo de obtener el mismo resultado con adicionales ventajas es haciendo uso de las posibilidades de incorporar términos de interacción, ya que elevar al cuadrado es como multiplicar una variable por sí misma:

```
regress tmi c.pnbppa##c.pnbppa
```

De esta manera, pueden calcularse con propiedad los efectos marginales de *pnbppa* (*help margins*).

### ILUSTRACIÓN 9.22. Regresión cuadrática

Source	SS	df	MS	Number of obs = 125		
Model	115248.64	2	57624.3202	F( 2, 122) = 100.54		
Residual	69927.4075	122	573.175471	Prob > F = 0.0000		
Total	185176.048	124	1493.35523	R-squared = 0.6224		
				Adj R-squared = 0.6162		
				Root MSE = 23.941		
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	-.0118483	.0010856	-10.91	0.000	-.0139974	-.0096992
pnb_2	3.46e-07	4.18e-08	8.28	0.000	2.63e-07	4.29e-07
_cons	88.35696	4.030461	21.92	0.000	80.37826	96.33566

Toda función cuadrática se caracteriza por tener un punto de inflexión, esto es, los valores pronosticados cambian de orientación a partir de un valor determinado de  $x$ . El tener el coeficiente cuadrático positivo implica que la curva obtenida empieza descendiendo y termina ascendiendo, como es aquí el caso. Y, como el coeficiente de la variable original tiene signo contrario al cuadrático, el punto de inflexión se encuentra en un valor positivo de  $x^{21}$ .

Y, si se quisiera obtener una representación gráfica, habría que escribir las siguientes instrucciones. La primera genera el valor predicho con la nueva regresión<sup>22</sup>, la segunda ordena los casos por la variable independiente y la tercera genera propiamente el gráfico.

```

predict ttmi2 if e(sample).
label variable ttmi2 "Predicción cuadrática"
sort pnbppa
scatter ttmi2 tmi pnbppa if e(sample), connect (l) symbol(i o) name(G33, replace)

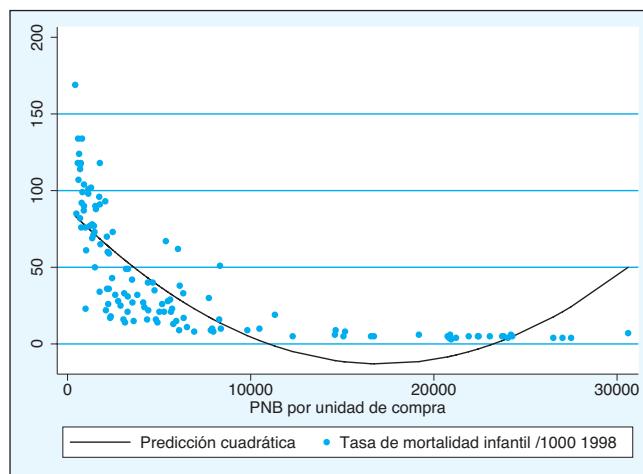
```

Como puede apreciarse a través del gráfico, el modelo cuadrático predice que la tasa de mortalidad infantil baja a medida que aumenta la renta per cápita en los países pobres, pero, a partir de determinado nivel de renta, la relación se invierte y la producción incide negativamente en este indicador.

<sup>21</sup> Recuérdese que el valor del punto de inflexión en la función parabólica es  $-b/2c$ , siendo  $b$  el coeficiente de la variable original y  $c$  el de la cuadrática.

<sup>22</sup> La razón por la que está acompañada por un *if* es para que sólo prediga en caso de que los valores de la dependiente sean válidos. Si no se hace explícita esa condición, también aparecería predicción para un posible valor extremo de  $x$ , que no poseyera valor en la dependiente. Este es el caso en este ejemplo, porque de Luxemburgo, con una alta renta per cápita, no se dispone del dato de la tasa de mortalidad infantil.

**GRÁFICO 9.11. Representación gráfica de la regresión cuadrática**



*Regresión cúbica:* Para obtener una regresión cúbica hay que añadir a la cuadrática una nueva variable, la original elevada al cubo. De este modo, siempre y cuando ya se disponga de la variable al cuadrado, se pueden conseguir los coeficientes de la regresión cúbica en dos pasos<sup>23</sup>:

```
generate pnb_3=pnbppa^3
regress tmi pnbppa pnb_2 pnb_3
```

Que dará lugar a una regresión con constante y tres coeficientes, que modularán dos puntos de inflexión:

<sup>23</sup> De modo análogo a la regresión cuadrática, puede hacerse una regresión cúbica con interacciones sin necesidad de construir nuevas variables. Esto se logra mencionando la variable tres veces precedidas por *c.*, indicando su carácter cuantitativo, y separadas por doble almohadilla (##):

```
regress tmi c.pnbppa##c.pnbppa##c.pnbppa
```

### ILUSTRACIÓN 9.23. Regresión cúbica

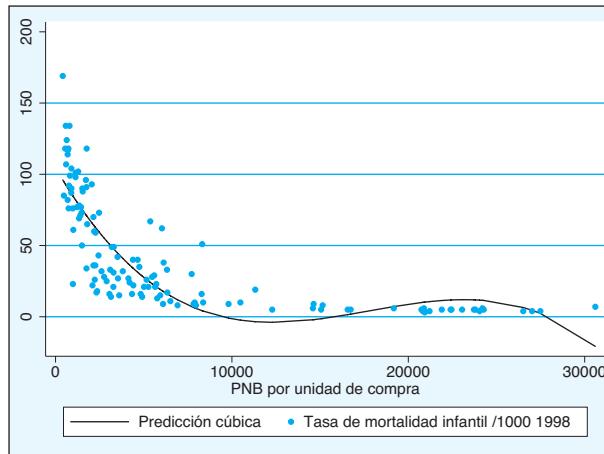
Source	SS	df	MS	Number of obs	=	125
Model	131330.579	3	43776.8596	F( 3, 121)	=	98.37
Residual	53845.4692	121	445.003877	Prob > F	=	0.0000
				R-squared	=	0.7092
				Adj R-squared	=	0.7020
Total	185176.048	124	1493.35523	Root MSE	=	21.095
<hr/>						
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	-.021411	.0018562	-11.53	0.000	-.0250858	-.0177362
pnb_2	1.33e-06	1.68e-07	7.93	0.000	9.98e-07	1.66e-06
pnb_3	-2.50e-11	4.16e-12	-6.01	0.000	-3.32e-11	-1.68e-11
_cons	104.6266	4.465046	23.43	0.000	95.7869	113.4664

Tras lo cual se pueden generar los valores teóricos de la variable dependiente a fin de obtener la representación del ajuste cúbico:

```
predict ttmi3 if e(sample)
label variable ttmi3 "Predicción cúbica"
scatter ttmi3 tmi pnbppa if e(sample), connect (l) symbol(i o) name(G35, replace)
```

El gráfico resultante de las anteriores instrucciones es el siguiente:

### GRÁFICO 9.12. Representación gráfica de la regresión cúbica



*Regresión inversa:* Para conseguir una regresión de este tipo, es suficiente convertir previamente la variable independiente en su inversa con la instrucción *generate* e introducirla como único predictor de la ecuación.

```
generate invpn=1/pnbppa
regress tmi invpn
```

El formato del resultado es idéntico al de una regresión simple.

#### ILUSTRACIÓN 9.24. Regresión inversa

Source	SS	df	MS	Number of obs	=	125
Model	139222.172	1	139222.172	F( 1, 123)	=	372.64
Residual	45953.8759	123	373.608747	Prob > F	=	0.0000
				R-squared	=	0.7518
				Adj R-squared	=	0.7498
Total	185176.048	124	1493.35523	Root MSE	=	19.329
<hr/>						
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
invpn	67857.04	3515.193	19.30	0.000	60898.93	74815.14
_cons	11.53628	2.365531	4.88	0.000	6.853862	16.21871

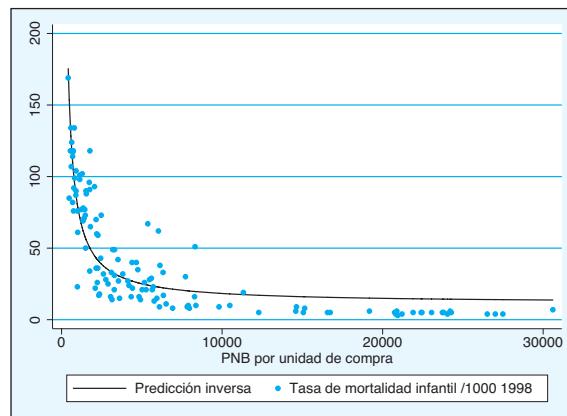
Como puede apreciarse con este simple modelo se llega a explicar el 75% de la varianza de la variable dependiente (*R-squared*).

Con los mismos principios con los que se ha hecho en las regresiones cuadrática y cúbica, en la inversa también pueden representarse los puntos empíricos y la línea del modelo.

```
predict ttminv if e(sample)
label variable ttminv "Predicción inversa"
scatter ttminv tmi pnbppa if e(sample), connect (l .)symbol(i o) name(G37, replace)
```

El resultado será un nuevo gráfico con una curva en forma de L:

#### GRÁFICO 9.13. Representación gráfica de la regresión inversa



Quedan finalmente por explicar las regresiones log-lineales que, a su vez, pueden adoptar tres modalidades:

*Regresiones log-log:* Son aquellas en las que tanto la variable dependiente como la independiente son transformadas en su correspondiente logaritmo. Su coeficiente, en lugar de indicar cuántas unidades cambia la variable dependiente por cada unidad que cambia la independiente, indica la tasa de cambio que sufre la primera por un cambio relativo en la segunda. Se puede expresar de los dos modos siguientes:

$$\ln(\hat{y}_i) = b_0 + b_1 \ln(x_i) \Leftrightarrow \hat{y}_i = \exp(b_0)x_i^{b_1} \quad (9.30)$$

La primera fórmula es tal cual se preparan los datos para que pueda realizarse la regresión como si fuera lineal. En la segunda, el valor de la variable dependiente se expresa en función de elevar el valor de la variable independiente a una determinada potencia ( $b_1$ ) y multiplicar el resultado por una constante,  $\exp(b_0)$ .

La preparación de esta regresión logarítmica implica la generación de dos nuevas variables, que sean logaritmos neperianos de las originales, y a continuación la realización de la regresión como si fuese lineal:

```
generate l_pnbppa=ln(pnbppa)
generate l_tmi=ln(tmi)
regress l_tmi l_pnbppa
```

El producto de estas tres instrucciones es también como el de la regresión simple:

#### ILUSTRACIÓN 9.25. Regresión logarítmica (log-log)

Source	SS	df	MS	Number of obs	=	125
Model	128.217683	1	128.217683	F( 1, 123)	=	653.96
Residual	24.1159321	123	.196064489	Prob > F	=	0.0000
Total	152.333615	124	1.22849689	R-squared	=	0.8417
				Adj R-squared	=	0.8404
				Root MSE	=	.44279
l_tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
l_pnbppa	-.8884191	.0347411	-25.57	0.000	-.9571869	-.8196512
_cons	10.58703	.2899	36.52	0.000	10.01319	11.16087

Posteriormente, para obtener los valores teóricos de  $y$ , habría que realizar la siguiente operación:  $y=\exp(10,8703)x^{-0,8884}=39617,6x^{-0,8884}$ . Ello implica

que el modelo predice que la duplicación del producto nacional bruto per cápita en un determinado país, por ejemplo, pasar de 5.000\$ per cápita a 10.000\$ o de 2.000\$ a 4.000\$, implicaría una caída del 54% en la tasa bruta de mortalidad infantil. Para obtener esta última cantidad hay que elevar 2 (el doble) a -0,8884 (el coeficiente  $b_1$ ). También, de modo más directo, puede interpretarse el coeficiente afirmando que una subida del 1% en el predictor, implica una bajada del 0,88% en el resultado.

Para conseguir con Stata los valores teóricos de la variable dependiente, primero se obtienen los valores de la variable logarítmica y posteriormente se convierten a su expresión original. Es necesario, pues, proceder en dos pasos<sup>24</sup>:

```
predict ttmigg if e(sample)
generate ttmgg=exp(ttmigg)
label variable ttmgg "Predicción log-log"
```

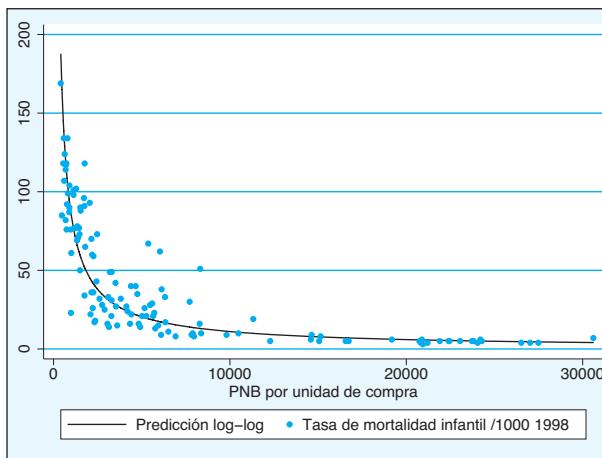
Tras disponer de los valores teóricos de la variable dependiente, se puede demandar el gráfico del modo habitual:

```
scatter ttmgg tmi pnbppa if e(sample), connect (l.) symbol (i o) name(G39, replace)
```

Como puede apreciarse, la línea del modelo se aproxima bastante a los puntos empíricos. Un  $R^2$  de 0,84 nos avala la bondad de predecir el logaritmo de la tasa de mortalidad infantil a partir del logaritmo del producto nacional bruto per cápita.

---

<sup>24</sup> Este proceder no genera la media condicional en la métrica original. Para solventarlo, Richard Goldstein ha generado un procedimiento llamado *predlog*. Está localizado en la red. Para su búsqueda e instalación, escriba *net search predlog*. Una vez incorporado, puede obtenerse ayuda de su uso y resultados, mediante *help predlog*.

**GRÁFICO 9.14.** Representación gráfica de la regresión logarítmica

*Regresión log-lin:* En este caso, sólo se transforma la variable dependiente. Por tanto, los coeficientes de la regresión indican el cambio relativo que sufre esta, cuando la independiente varía en una unidad. La expresión matemática que responde a este modelo es la siguiente:

$$\ln(\hat{y}_i) = \ln(b_0) + b_1 x_i \Leftrightarrow \hat{y}_i = b_0 e^{b_1 x_i} \quad (9.31)$$

Para poder obtener este modelo y su representación mediante las instrucciones Stata, estas deberían tener la secuencia siguiente:

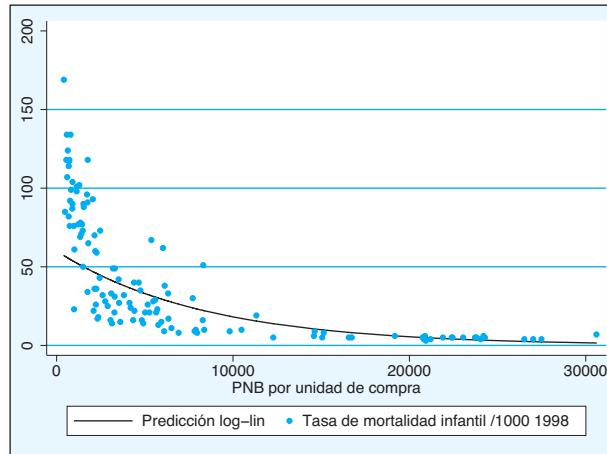
```
generate l_tmi=ln(tmi)
regress l_tmi pnbppa
predict ttmign if e(sample)
generate ttmgm=exp(ttmign)
label variable ttmgm "Predicción log-lin"
scatter ttmgm tmi pnbppa if e(sample), connect (l .) symbol (i o) name(G40, replace)
```

La ecuación resultante, empleando como resultado la tasa de mortalidad infantil (en realidad, su logaritmo) y como predictor el producto nacional bruto per cápita, ofrece un coeficiente de -0,0001196, lo que supone que para bajar un 11% esta tasa en este país, se necesita aumentar en 1.000\$ la renta per cápita de sus ciudadanos ( $1 - \exp(-0,0001196 \cdot 1000) = 0,11$ ).

**ILUSTRACIÓN 9.26. Regresión exponencial (log-lin)**

Source	SS	df	MS	Number of obs = 125		
Model	109.28253	1	109.28253	F( 1, 123) = 312.23		
Residual	43.0510845	123	.350008817	Prob > F = 0.0000		
Total	152.333615	124	1.22849689	R-squared = 0.7174		
l_tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pnbppa	-.0001196	6.77e-06	-17.67	0.000	-.000133	-.0001062
_cons	4.094556	.0715704	57.21	0.000	3.952887	4.236225

El gráfico resultante con un  $R^2$  de 0,72 —es decir, algo menos ajuste que el anterior modelo, especialmente en los valores bajos de la variable independiente— es el siguiente:

**GRÁFICO 9.15. Representación gráfica de la regresión exponencial**


*Regresión lin-log:* Finalmente, queda el modelo donde la variable dependiente no se transforma en su logaritmo, pero sí lo hace la independiente. La ecuación matriz de este modelo es:

$$\hat{y}_i = b_0 + b_1 \ln(x_i) \quad (9.32)$$

Por ello, sólo es necesaria la traducción logarítmica de la variable independiente, y la secuencia de instrucciones para representar una relación entre variables de este tipo sería la siguiente:

```
generate l_pnbppa=ln(pnbppa)
regress tmi l_pnbppa
predict ttming if e(sample)
label variable ttming "Predicción lin-log"
scatter ttming tmi pnbppa if e(sample), connect (l) symbol (i o) name(G41, replace)
```

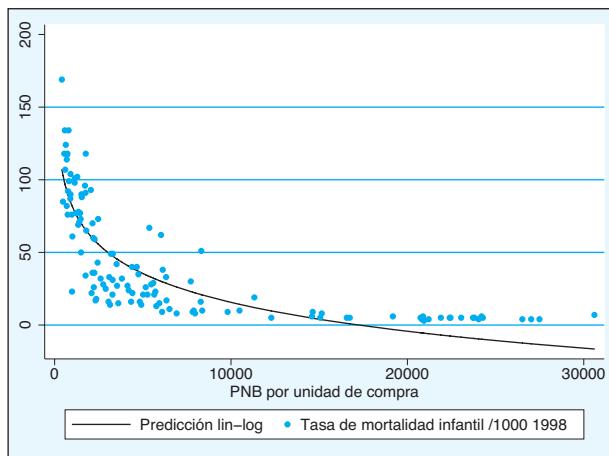
El coeficiente de la regresión significa cuánto sube o baja la variable resultado por cada punto logarítmico que suba el predictor. Una forma práctica de interpretarlo es multiplicándolo por 0,693, que es el logaritmo neperiano de 2. El valor de este producto implica el cambio en la variable dependiente cuando se duplica el predictor. De este modo, de acuerdo con la ilustración 9.26, por cada duplicación de producto nacional bruto per cápita, la tasa de mortalidad infantil se reduce en 19,9 puntos (por diez mil), cifra obtenida mediante el producto -28,698\*ln(2).

### ILUSTRACIÓN 9.27. Coeficientes de la regresión del modelo lin-log

Source	SS	df	MS	Number of obs	=	125
Model	133783.834	1	133783.834	F( 1, 123)	=	320.19
Residual	51392.2139	123	417.822877	Prob > F	=	0.0000
				R-squared	=	0.7225
				Adj R-squared	=	0.7202
Total	185176.048	124	1493.35523	Root MSE	=	20.441
<hr/>						
tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
l_pnbppa	-28.69761	1.603762	-17.89	0.000	-31.87216	-25.52306
_cons	279.9285	13.38273	20.92	0.000	253.4382	306.4188

El ajuste medido a través del  $R^2$  de este modelo es de 0,72 y, en este caso donde se relaciona la tasa de mortalidad infantil con el producto nacional bruto per cápita, el gráfico representa bastante peor que el modelo log-log el conjunto de datos empíricos, especialmente en los valores altos de la independiente:

**GRÁFICO 9.16.** Representación gráfica de la regresión lin-log



## 9.9. Ejercicios

1. Con la base de datos mundial de 2005 (mundo2005), elige como variable dependiente la esperanza de vida al nacer y como independientes la renta per cápita (en unidades de poder de compra) y el continente. Reproduce las regresiones realizadas en este capítulo y encuentra un modelo satisfactorio.
2. Con la misma base de datos, selecciona una nueva variable dependiente y selecciona las variables independientes más adecuadas para la definición de un buen modelo. Introduce también, si te parece conveniente, alguna otra variable nominal (factor), como, por ejemplo, la pertenencia o no al G20.
3. Con los datos empleados en este capítulo (mundo99), realiza una tabla que cruce el continente con la pertenencia a la OCDE poniendo en las casillas la media de la tasa de mortalidad infantil. A continuación, haz una regresión de la tasa sobre ambos factores y su interacción. Interpreta los coeficientes en función de la primera tabla elaborada. Introduce como covarianza el producto nacional per cápita y observa el cambio que experimentan los coeficientes de los factores.
4. Con el cuestionario del barómetro de abril de 2009 (cis2798), selecciona las cuatro variables que representan la probabilidad subjetiva de votar el entrevistado a los cuatro partidos con candidaturas en todo el territorio del Estado y al menos un representante en el Parlamento español (PSOE, PP, IU y UPyD). Selecciona del cuestionario las variables que teóricamente se consideren más importantes para explicar el

- comportamiento electoral. Compara, finalmente, los resultados de las cuatro ecuaciones.
5. Ahora, empleando el barómetro de marzo (cis2794), toma la primera parte de la P.28 como variable respuesta. Busca explicaciones a la variación en la disposición del tiempo libre de las personas en el resto del cuestionario. Al menos, introduce el sexo como factor y la edad como predictor cuantitativo.



# 10

## Diagnóstico de la regresión

### 10.1. Supuestos de la regresión lineal

El modelo poblacional del que se parte para el correcto funcionamiento de la estimación de los parámetros de la regresión por el procedimiento de mínimos cuadrados responde a la siguiente ecuación:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i \quad (10.1)$$

Según ella, existe en la población una relación lineal entre un conjunto  $k$  de variables independientes ( $x_i$ ) que dan lugar de modo estocástico o indeterminado a la variable dependiente ( $y_i$ ). Por ello, aparece en la ecuación la variable aleatoria  $\varepsilon_i$ , de la que se supone en el modelo que se distribuye normalmente con una media de 0.

Pero en esta ecuación del modelo no están declarados expresamente una serie de prerrequisitos para que los estimadores de los parámetros  $\beta$  obtenidos por el criterio de mínimos cuadrados tengan la propiedad de ser los mejores estimadores lineales insesgados (MELI) del teorema de Gauss-Markov.

Basándose en Gujarati y Porter (2008), aunque cambiando el orden y la terminología utilizada por estos autores, los criterios que debe satisfacer el modelo de la regresión por el método de mínimos cuadrados son los siguientes:

1. Los valores de las variables independientes han de ser fijos.
2. El número de observaciones debe ser mayor que el número de variables independientes.

$$n > k \quad (10.2)$$

3. Debe haber suficiente variabilidad en los valores de las variables independientes.

$$\text{Var}(x_i) > \ell \quad (10.3)$$

4. El término de perturbación está normalmente distribuido.

$$\varepsilon_i \sim N(0, \sigma) \quad (10.4)$$

5. Para cada conjunto de casos con una  $x_i$  dada, el valor medio de la perturbación ( $\varepsilon_i$ ) es cero.

$$\forall x_i \quad E(\varepsilon_i) = 0 \quad (10.5)$$

6. En el caso de que las  $x_i$  sean estocásticas, no existe correlación entre estas y los términos de perturbación.

$$Cov(x_i, \varepsilon_i) = 0 \quad (10.6)$$

7. Para cada conjunto de casos con una  $x_i$  dada, la varianza de  $\varepsilon_i$  es constante u homocedástica.

$$\forall x_i \quad Var(\varepsilon_i) = \sigma^2 \quad (10.7)$$

8. No hay relación exacta (no hay multicolinealidad) en los regresores.

$$Cov(z_{x_i}, z_{x_j}) < 1; \quad (i \neq j) \quad (10.8)$$

9. No existe autocorrelación entre las perturbaciones.

$$Cov(\varepsilon_i, \varepsilon_j) = 0; \quad (i \neq j) \quad (10.9)$$

10. El modelo de regresión es lineal en sus parámetros.  
 11. El modelo de la regresión está correctamente especificado.

Los tres primeros requisitos son fáciles de comprobar sin necesidad de operaciones complejas de naturaleza estadística. El primero implica que las variables independientes no son aleatorias, como puede ser el caso de que sean introducidas experimentalmente por el investigador. Sin embargo, en ciencias sociales, como es muy improbable que puedan tener esa condición los regresores, no es necesario que se cumpla en tanto en cuanto el criterio sexto esté satisfecho. El segundo es de fácil comprobación,

puesto que tanto  $n$  como  $k$  son conocidos. Por cuestiones de determinación de los parámetros, estos son imposibles de estimar siempre que  $k > n$ , pero, aun en el caso de que  $n > k$ , existen autores que recomiendan para evitar la presencia de altos errores de estimación una proporción de 5 veces superior el número de casos sobre el de parámetros (Afifi *et al.* 2003). Finalmente, el tercero puede comprobarse mediante la obtención de la desviación típica de las variables independientes. O mejor, si cabe, con el coeficiente de variación, que es el cociente entre aquella y la media aritmética de la variable. Del mismo modo que el supuesto anterior, incide principalmente en la cuantía de los errores típicos de los parámetros. Cuando la variabilidad de  $x$  es baja, automáticamente el denominador del cálculo de estos estadísticos tiende a 0 y, por tanto, el resultado del cociente se elevará hasta cantidades excesivamente altas. Obvio es que la solución a los problemas suscitados en el segundo y tercer supuesto es el incremento de la muestra<sup>1</sup>.

Los supuestos cuarto, quinto y sexto, todos ellos relacionados con el término de la perturbación pueden evaluarse con distintas instrucciones presentes en Stata. La primera y fundamental de ellas es la generación, después de realizar una regresión, de una nueva variable, que exprese los residuos de la regresión, que son los mejores indicadores muestrales del término de perturbación en la población. Hay para ellos tres modalidades que pueden seleccionarse en función de la opción que se añada al comando *predict*. Estas tres opciones son:

- a) Los residuos simples medidos en las mismas unidades que la variable dependiente (*,residuals*):

$$e_i = y_i - \hat{y}_i \quad (10.10)$$

- b) Los residuos tipificados, es decir, transformados para que tengan media de 0 y desviación típica igual a 1 (*,rstandard*):

$$z_{e_i} = \frac{y_i - \hat{y}_i}{s_e \sqrt{1 - h_i}} \quad (10.11)$$

<sup>1</sup> Un problema bastante común en la regresión múltiple es el del descenso del número de casos de la muestra original al introducir muchas variables con un alto número de casos perdidos. Incluso, aunque no lo sean, se puede dar una combinación de ausencia de información entre ellas (como en el caso de preguntas filtradas), que haga bajar sustancialmente el número de casos con los que se opera. En estas ocasiones, es conveniente prescindir de las predictores que causen un considerable descenso de la muestra, no sólo por los problemas de aumento del error típico, sino sobre todo por los de selección sesgada de individuos muestrales.

- c) Los residuos studentizados, si se divide por la desviación típica de los residuales resultante de eliminar el caso en cuestión ( $s_{e(i)}$ ) (*rstudent*):

$$t_{e_i} = \frac{y_i - \hat{y}_i}{s_{e(i)}\sqrt{1 - h_i}} \quad (10.12)$$

Estas tres variables pueden obtenerse al solicitar el comando *predict*, tras la ejecución de una regresión, con la correspondiente opción *y*, obviamente, el nombre que se le quiera dar a la nueva variable:

```
predict nueavar, residual
predict nueavar, rstandard
predict nueavar, rstudent
```

Un ejemplo de aplicación de instrucciones se puede aplicar a una de las regresiones obtenidas en el capítulo anterior:

```
regress tmi pnbppa
predict tmir, residual
predict tmirt, rstandard
predict tmirs, rstudent
```

A partir de ahí, se dispone en la base de datos abierta de tres nuevas variables con las que se puede operar como si hubieran sido introducidas al crear el fichero. De este modo, si se escribe la siguiente instrucción<sup>2</sup>...

```
summarize tmi?*
```

... se obtienen los estadísticos básicos de las tres nuevas variables:

---

<sup>2</sup> En Stata pueden emplearse los símbolos \* y ? para construir listas de variables. El primero significa reemplazo de una cadena de caracteres seguidos, mientras que el segundo sólo reemplaza un carácter al tiempo. Al escribir ?\* excluye *tmi*, puesto que ? excluye la posición en blanco, mientras que \* la incluye. Si sólo se hubiera puesto ?, sólo se incluiría *tmir*; y si se hubiera escrito *tmi??*, la lista habría sido *tmirt* y *tmirs*.

### ILUSTRACIÓN 10.1. Estadísticos de los residuos

Variable	Obs	Mean	Std. Dev.	Min	Max
tmir	125	7.63e-08	29.67555	-41.25054	105.1491
tmirt	125	.0020788	1.00298	-1.39146	3.553735
tmirs	125	.0051997	1.012505	-1.39683	3.736263

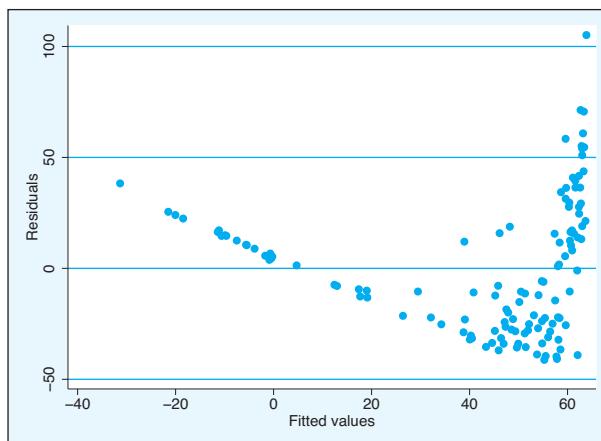
Como puede apreciarse, la media de los residuos es prácticamente igual a 0 y la desviación típica de los tipificados es igual a 1. Pero también se observa cómo los valores mínimos de los residuos tipificados y studentizados son en valores absolutos bastante menores que los máximos, lo que es indicativo de una notable asimetría en la distribución.

Para ver si esa media de 0 es constante a lo largo de los distintos valores de  $x$  (supuesto quinto), en cuyo caso también se cumpliría la no correlación entre  $\varepsilon_i$  y  $x_i$  (supuesto sexto), se puede proceder a la construcción del gráfico que cruza los residuos con los valores predichos de la variable dependiente. Este puede obtenerse de modo fácil con escribir tras una regresión el comando *rvfplot*, que representa la nube de puntos de los residuos *versus* los valores ajustados de la regresión:

```
rvfplot
```

Y, como consecuencia de esta instrucción, aparecerá el siguiente gráfico:

GRÁFICO 10.1. Nube de puntos de los residuos



En este ejemplo es obvio que el valor medio de los residuos cambia con los valores ajustados, que en este caso, como sólo se dispone de una varia-

ble independiente, coinciden linealmente con los valores de esta. Se nota cómo en los valores más bajos de la variable independiente, los residuos van descendiendo a medida que aquella aumenta, pero a partir de determinado valor (aproximadamente los 50 años de valor ajustado), la media del valor esperado de los residuos va haciéndose cada vez mayor.

Otro requisito que puede verificarse de modo fácil con Stata es la supuesta normalidad en la distribución de los residuos (supuesto cuarto). Para ello hay diversas posibilidades. La primera y más simple es a través del examen estadístico de los coeficientes de simetría y curtosis. Se dispone de un comando, que no sólo los calcula, sino que también realiza una prueba estadística sobre ellos para ver si son significativamente distintos de la hipótesis normal. La orden *sktest* permite realizar estas operaciones con tal de expresar las variables cuya normalidad se desea verificar:

```
sktest tmir tmirs tmirt
```

Esta prueba estadística es, en realidad, una comprobación de que simetría y curtosis son iguales que los que la distribución normal presenta.

### **ILUSTRACIÓN 10.2. Asimetría y curtosis de los residuos**

Variable	Skewness/Kurtosis tests for Normality			
	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
tmir	0.001	0.476	9.34	0.0094
tmirs	0.001	0.486	9.30	0.0096
tmirt	0.001	0.261	10.78	0.0046

En este caso, los residuos, tanto normales como estandarizados o studentizados, presentan una distribución asimétrica, por lo que no puede afirmarse que su distribución sea normal.

También pueden utilizarse para comprobar la distribución de la normal las pruebas de Shapiro-Wilk y Shapiro-Francia, cuyas órdenes respectivas son *swilk* y *sfrancia* seguidas de la lista de variables cuya normalidad se desea comprobar. De este modo, con las dos siguientes instrucciones:

```
swilk tmir-tmirt
sfrancia tmir-tmirt
```

... se obtienen las mismas conclusiones que con las pruebas de simetría y curtosis, pues en cada variable puede rechazarse con un nivel de significación inferior al 0,05 la hipótesis nula de que la distribución es normal.

### ILUSTRACIÓN 10.3. Pruebas de normalidad de los residuos

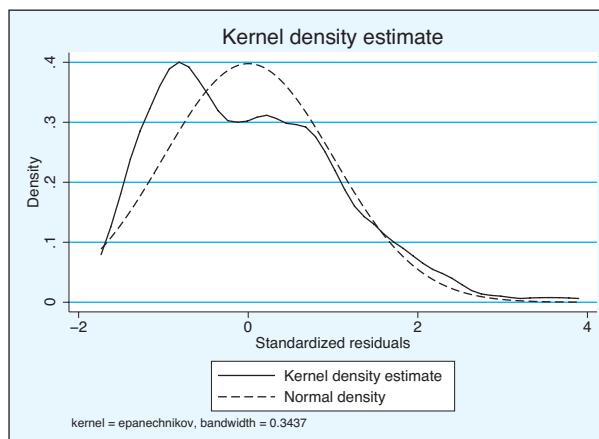
Shapiro-Wilk W test for normal data					
Variable	Obs	W	V	z	Prob>z
tmir	125	0.94111	5.866	3.972	0.00004
tmirs	125	0.94104	5.873	3.975	0.00004
tmirt	125	0.93902	6.074	4.051	0.00003
Shapiro-Francia W' test for normal data					
Variable	Obs	W'	V'	z	Prob>z
tmir	125	0.94245	6.248	3.636	0.00014
tmirs	125	0.94239	6.255	3.638	0.00014
tmirt	125	0.93985	6.531	3.719	0.00010

Finalmente, de un modo gráfico, también puede comprobarse cuán distinta es de la normal la distribución de los residuos mediante la ayuda de gráficos. A este respecto podrían utilizarse tanto un gráfico de probabilidades (*pnorm*) como de cuantiles (*qnorm*), o el de superposición de las dos distribuciones mediante la instrucción *kdensity*, seguida de la opción *normal*:

```
kdensity tmirt, normal
```

... que dará lugar al siguiente gráfico de frecuencias de una y otra distribución.

### GRÁFICO 10.2. Comprobación gráfica de la normalidad de los residuos



Otro de los diagnósticos que han de efectuarse a toda regresión es el de la homocedasticidad (supuesto séptimo). Se entiende por esta propiedad el

hecho de que las varianzas residuales sean las mismas independientemente de los valores de las variables independientes y, por extensión, de los valores predichos de la dependiente. Por ello, la formulación expresada en (10.7) puede reformularse mediante la siguiente expresión:

$$Var(\varepsilon_i | \hat{y}_i) = \sigma^2 \quad (10.13)$$

El medio gráfico idóneo para observar la presencia de heterocedasticidad es el que cruza residuos con los valores predichos de la variable dependiente, que se obtiene mediante la instrucción *rvfplot*, como se ha visto anteriormente y ya se ha ejemplificado en el gráfico 10.1.

Stata dispone, no obstante, de una prueba que da cuenta numéricamente de la existencia de la heterocedasticidad. Se trata del test de Cook-Weisberg (1983), que se obtiene especificando la instrucción *hettest* tras la ejecución de una regresión. Así pues, tras la regresión de la tasa de mortalidad infantil sobre el producto nacional bruto per cápita, al escribir la siguiente línea:

```
hettest
```

... se obtiene el siguiente resultado:

#### **ILUSTRACIÓN 10.4. Prueba de heterocedasticidad de Cook y Weisberg**

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of tmi

chi2(1)      =     12.23
Prob > chi2  =  0.0005
```

Como allí mismo se expresa, el valor de la hipótesis nula es el de varianza constante. Como en este caso el  $\chi^2$  con un grado de libertad tiene un valor superior a 12, con una probabilidad inferior al 5%, ha de rechazarse la hipótesis de homocedasticidad. Por tanto, se está ante un dato adicional que nos hace desconfiar de la estimación de mínimos cuadrados ordinarios.

La ausencia de multicolinealidad es otro de los criterios (supuesto octavo), aplicable sólo en casos de regresión múltiple. Por multicolinealidad se entiende la correlación entre las variables independientes. El criterio más utilizado para detectarla es el de la tolerancia (complementario del coeficiente de determinación múltiple de una variable independiente con el resto) o su inverso, conocido como factor de inflación de la varianza (*VIF*).

$$VIF = \frac{1}{1 - R_{x_1.x_2...x_k}^2} \quad (10.14)$$

Mediante el programa Stata se pueden obtener estos índices de multicolinealidad, al introducir la instrucción *vif* después de una instrucción. De este modo, si se introducen estas dos instrucciones:

```
regress tmi pnbppa lintfno pibag
vif
```

... se obtendrá como resultado de la segunda el siguiente listado de variables independientes:

#### **ILUSTRACIÓN 10.5. Índice de multicolinealidad en la regresión múltiple**

Variable	VIF	1/VIF
lintfno	12.37	0.080856
pnbppa	11.81	0.084651
pibag	1.85	0.539234
Mean VIF	8.68	

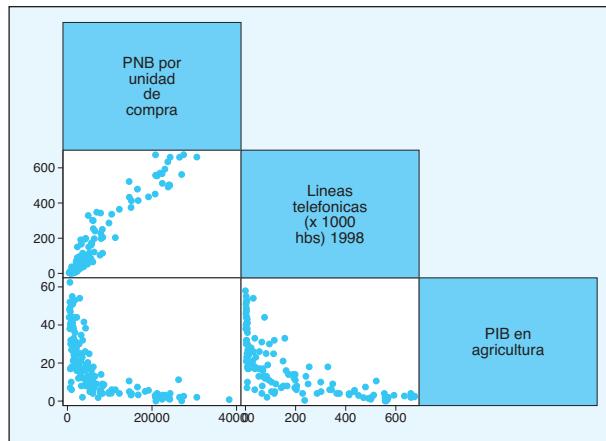
En esta tabla se detecta alta colinealidad, sobre todo entre las variables *pnbppa* y *lintfno*. Como regla sugerida, se recomienda que el factor no supere el valor de 10, lo que equivale al 0,10 de su inverso o, dicho de otro modo, cuando una variable de la ecuación tiene un coeficiente de correlación múltiple con el resto de las variables superior a 0,95, los problemas de eficiencia de los estimadores serán muy elevados. Con algo más de rigurosidad, no debería aceptarse la inclusión de variables con *VIF* superiores a 5, equivalentes a coeficientes de correlación de 0,90.

De modo gráfico, aunque imparcial por sólo recoger relaciones bivariadas, se puede recurrir a la matriz de nube de puntos entre las variables independientes para ver si entre alguna de ellas se produce alguna evidente y fuerte relación lineal.

```
graph matrix pnbppa lintfno pibag, half name(G3)
```

De este modo, se genera el siguiente resultado, en el que claramente se ve la peculiar relación lineal entre el producto nacional bruto per cápita y el número de líneas telefónicas por cada mil habitantes.

**GRÁFICO 10.3. Matriz de nubes de puntos**



Otro requisito, especialmente relevante y crítico, en las regresiones con series temporales, es el de la independencia de los residuos (supuesto noveno). El par de estadísticos más utilizados para detectarla es el de Durbin-Watson y el de Breusch-Godfrey, que deberían ser vistos con mayor profundidad en un tema relacionado con este tipo modelos que emplean datos obtenidos regularmente en distintos períodos de tiempo, que no es el caso de este capítulo.

Más importancia en este contexto tienen los requisitos décimo y undécimo, que se refieren a que en la población se dé efectivamente una relación lineal y a que la variable dependiente dependa efectivamente de los predictores que se han especificado en la ecuación. Si en la población de la que se extraen las muestras no se da una relación lineal o si se excluye alguna variable fundamental en la ecuación de regresión, los estimadores obtenidos mediante la muestra estarán sesgados, salvo en el improbable caso de que la omitida tenga correlación nula con el resto de las variables del modelo.

Indicios de modelos no lineales o de incorrectas especificaciones (supuestos décimo y undécimo) se deducen a través de bajos coeficientes de determinación, altos errores típicos de los parámetros, alta autocorrelación o distribuciones no normales de residuos. Además de ello, Stata cuenta con un test (el de Ramsey 1969, *ovtest*) que permite verificar los errores de especificación. Existen dos modalidades: en la primera se añaden los términos cuadrados, cúbicos y a la cuarta de los valores predichos, para ver si estos son significativos; en la segunda, que se obtiene mediante la opción *rhs*, lo que se añaden son las potencias de las variables independientes, siempre y cuando no sean ficticias.

Como otras instrucciones de diagnóstico, esta ha de especificarse después de la regresión correspondiente. Un ejemplo nos muestra cómo con un ligero cambio de las variables se pueden corregir estos problemas de

especificación. En primer lugar, se realizan los diagnósticos de la tasa de mortalidad infantil regresada con el producto nacional bruto per cápita:

```
regress tmi pnbppa
ovtest
```

El resultado muestra una diferencia significativa de consideración.

#### **ILUSTRACIÓN 10.6. Prueba de Ramsey sobre omisión de variables en la regresión regular**

```
Ramsey RESET test using powers of the fitted values of tmi
Ho: model has no omitted variables
      F(3, 120) =      56.11
      Prob > F =      0.0000
```

Sin embargo, al transformar las variables en sus logaritmos, es preciso realizar de nuevo la prueba.

```
for var tmi pnbppa:generate l_X=ln(X)
regress l_tmi l_pnbppa
ovtest
```

En este caso, con los datos obtenidos, puede no ser rechazada la hipótesis nula de que el modelo no ha omitido variables importantes.

#### **ILUSTRACIÓN 10.7. Prueba de Ramsey sobre omisión de variables en la regresión logarítmica**

```
Ramsey RESET test using powers of the fitted values of l_tmi
Ho: model has no omitted variables
      F(3, 120) =      1.84
      Prob > F =      0.1444
```

Los gráficos también pueden ser útiles en la detección de linealidad en la relación entre las variables relacionadas. Además del gráfico visto anteriormente que enseña las relaciones bivariadas entre las variables, pueden ser usados los gráficos de regresión parcial o gráficos de variable añadida. Se trata del cruce entre, por un lado, los residuos de la variable dependiente del resultado de su regresión con el resto de las variables independientes y, por el otro lado, los residuos de la variable independiente obtenidos tras considerarla dependiente del resto de las independientes. La pendiente de este gráfico no es otra cosa que el coeficiente parcial de la regresión.

La instrucción es fácil de ejecutar. Basta con escribir *avplots*. Si así se especifica, se generará un gráfico por cada variable independiente. Caso de que se quiera sólo el gráfico de una variable, la instrucción ha de explicitarse en singular (*avplot*), seguida de la variable independiente que se deseé representar. Y, aunque se quieran todos los gráficos, es preferible para mejor detalle obtenerlos individualmente. Otro aspecto que ha de tenerse en cuenta es que esta instrucción, como la de los gráficos anteriores, puede utilizar, siempre que le sea pertinente, las opciones propias de la instrucción *graph*. De este modo, entre otros aspectos, se podrían obtener gráficos con las etiquetas de los casos presentes en la nube de puntos. Así, en la regresión del logaritmo de la tasa de mortalidad infantil en función del logaritmo del producto nacional bruto per cápita y de las líneas telefónicas por mil habitantes, se pueden generar los gráficos de este modo:

```
regress l_tmi l_pnbppa lintfno
avplot l_pnbppa, mlabel(pais)
avplot lintfno, mlabel (pais)
```

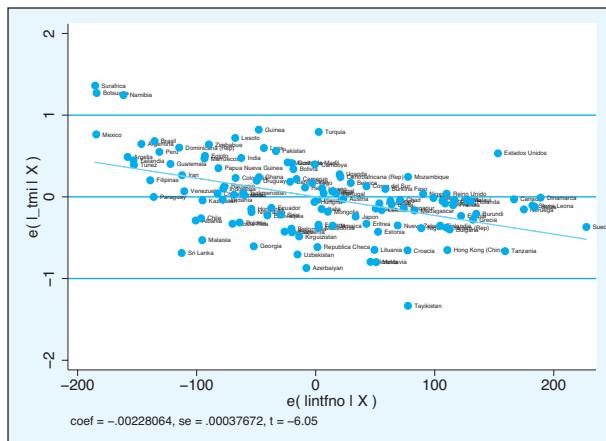
A partir de lo que se generaría una regresión (log-log) y dos gráficos, el segundo de los cuales adoptaría el aspecto del gráfico 10.4.

#### **ILUSTRACIÓN 10.8. Regresión (log-log) de la tasa de mortalidad infantil sobre producto nacional per cápita y número de líneas telefónicas**

Source	SS	df	MS	Number of obs	=	124
Model	133.661425	2	66.8307126	F( 2, 121)	=	438.32
Residual	18.4490767	121	.152471708	Prob > F	=	0.0000
Total	152.110502	123	1.23667075	R-squared	=	0.8787
				Adj R-squared	=	0.8767
				Root MSE	=	.39048
<hr/>						
l_tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
l_pnbppa	-.5244678	.0674209	-7.78	0.000	-.6579452	-.3909904
lintfno	-.0022806	.0003767	-6.05	0.000	-.0030264	-.0015348
cons	7.98921	.4993729	16.00	0.000	7.000569	8.97785

En el eje de ordenadas se representan los residuos de la variable dependiente obtenidos con su regresión sobre el logaritmo del producto nacional per cápita y en el eje de abscisas se representan los residuos de las líneas telefónicas obtenidos en su regresión sobre el logaritmo del producto nacional per cápita, esto es, el resto de las variables independientes. Es preciso notar que la inclinación de la línea representada es el coeficiente parcial de regresión múltiple, anotado también en la parte inferior del gráfico, junto con su error típico. Lo que hay que comprobar, para verificar el supuesto de linealidad, es que no haya una pauta curvilínea o plana de distribución de los casos.

**GRÁFICO 10.4.** Gráfico de residuos dependientes sobre los independientes (*avplot*)



Se han visto las herramientas de las que dispone Stata para detectar el incumplimiento de los supuestos de la regresión. Para acabar este apartado no estaría de más realizar un compendio simplificado de ellas, viendo sus efectos y el modo de detectarlas. La mayor parte de ellas inciden en la baja eficiencia de los estimadores, es decir, multiplican la posibilidad de que obtengamos una estimación alejada del valor correcto. Los incumplimientos que generan sólo problemas de eficiencia son la heterocedasticidad y la autocorrelación. La primera se detecta mediante el gráfico de residuos sobre los valores predichos (*rvfplot*) y de modo más preciso con el test de Cook-Weisberg (*hettest*). La segunda con el estadístico de Durbin-Watson, aunque en principio no debería preocupar siempre que no se tengan datos de series temporales. La multicolinealidad también genera problemas de ineficiencia e incluso puede llegar a hacer que sean incalculables los parámetros de la regresión, en el caso de que sea perfecta. Su modo de detección es a través de la tolerancia o del factor de inflación de la varianza obtenido mediante la orden *vif*. Si las perturbaciones no son normales, los estimadores, además de ineficientes, no estarán distribuidos normalmente, por lo que no serán válidas las pruebas de significación. Además, si las medias de las perturbaciones no son 0, los parámetros serán segados, especialmente peligroso, si la esperanza de las perturbaciones es además inconstante, porque afectaría no sólo a la constante, sino también a los coeficientes de las variables. Finalmente, el problema principal es que la matriz de las variables independientes sea estocástica y además correlacionen las variables regresoras con los términos de perturbación. En dicho caso, las estimaciones poseerán importantes segos y no serán ni eficientes ni consistentes. Estos problemas son detectables principalmente a través de gráficos de residuos con las variables independientes y con la prueba de Ramsey (*ovtest*).

## 10.2. Análisis de los casos en la regresión

Además de verificar que se cumplen los supuestos de la regresión, es útil examinar el comportamiento de los casos, por cuanto estos pueden sesgar el comportamiento de los estimadores de los parámetros. Hay tres tipos de medida que deben examinarse para ver si existen casos que pueden estar perturbando una regresión. En primer lugar, los ya conocidos residuos, de los que ahora se estudiará no su comportamiento conjunto, sino el particular de cada caso. En segundo lugar, están las medidas que ponderan la carga de las variables independientes, de modo que tengan puntuaciones más altas mientras valores más extremos tengan en estas. Y, finalmente, están aquellas puntuaciones que reflejan de uno u otro modo su contribución a los coeficientes, a los valores predichos o al error estimado de la regresión.

Las primeras de estas puntuaciones son los residuos. Así como anteriormente veíamos sus promedios y sus distribuciones, ahora resulta más propio el examen de los valores extremos. Para ello, se dispone en Stata tanto de herramientas numéricas como gráficas.

Entre las primeras está la instrucción *list*, que en conjunción con la instrucción *sort* y la especificación *if*, puede dar cuenta de modo ordenado sólo de los casos que tengan valores extremos.

Así, después de la regresión de la tasa logarítmica de mortalidad infantil con el logaritmo del producto nacional per cápita y las líneas telefónicas, y tras la ya efectuada generación de los distintos residuos, se pueden localizar aquellos casos con valores extraordinarios, si así se consideran aquellos cuyo valor está 1,96 desviaciones típicas por encima o por debajo de la media aritmética:

```
for any r rs rt \ any residual rstandar rstudent: predict ltmirX, Y
generate ltmira=abs(ltmir)
gsort -ltmira
list pais ltmir ltmirs ltmirt if (abs(ltmirs)>1.96 | abs(ltmirt)>1.96) & e(sample)
```

Mediante estas instrucciones, se crea una variable con los valores absolutos de los residuos para poderlos ordenar descendente por su tamaño. Y, finalmente, se seleccionan los países que han entrado en la regresión<sup>3</sup> con valores absolutos típicos o studentizados superiores al punto crítico de 1,96, correspondiente al nivel del 5% (por ello, en condiciones de normalidad, deberían aparecer aproximadamente en el listado cinco de cada cien casos).

El resultado en la regresión comentada es el siguiente:

---

<sup>3</sup> Al realizar una regresión, Stata registra una serie de resultados y estimaciones con un nombre específico. Una de las más útiles es la función *e(sample)*, que permite seleccionar los casos que han entrado en la última regresión al especificarla dentro de una cláusula *if* de una instrucción posterior.

### ILUSTRACIÓN 10.9. Listado de residuos

	país	ltmir	ltmirs	ltmirt
1.	Tayikistán	-1.156497	-2.999984	-3.105277
2.	Sri Lanka	-.9439814	-2.442333	-2.494482
3.	Suráfrica	.9388283	2.458475	2.511833
4.	Azerbaiyán	-.8882979	-2.286082	-2.327433
5.	Estados Unidos	.8813	2.323113	2.366879
6.	Namibia	.8771846	2.284423	2.325668
7.	Botsuana	.851571	2.22672	2.264379
8.	Turquía	.8011727	2.061442	2.089934

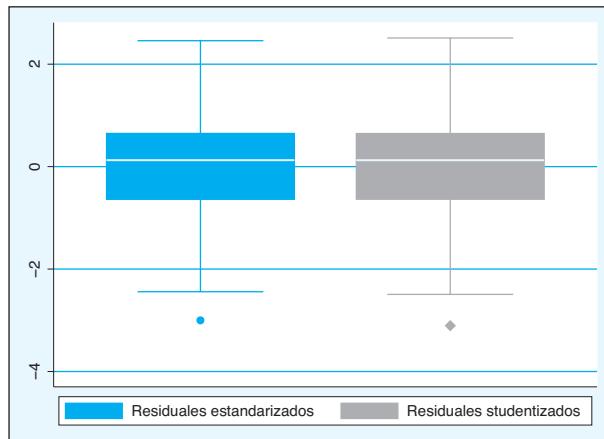
Como puede apreciarse, unos son positivos, lo que indica que son países con una tasa de mortalidad infantil superior incluso a la esperada con su producto nacional bruto per cápita y número de líneas telefónicas; otros son negativos, en el caso de que las variables independientes predigan tasas de mortalidad infantil más bajas de las reales.

Un modo gráfico y fácil de detectar los valores residuales extraordinarios es mediante los gráficos de caja, mediante los que se consideran anómalos los residuos alejados del promedio vez y media el rango intercuartílico y se denominan extremos, si se desvían del promedio tres veces dicha cantidad.

Mediante Stata puede solicitarse al mismo tiempo la representación de los residuos típicos y studentizados.

```
label var ltmirs "Residuales estandarizados"
label var ltmirt "Residuales studentizados"
graph box ltmirs ltmirt
```

### GRÁFICO 10.5. Gráfico de cajas de los residuos típicos y studentizados



Como puede fácilmente apreciarse, en el gráfico sólo considera caso desviado el de Tayikistán. Los demás están dentro de los límites marcados por las extensiones del rango intercuartílico.

Una de las medidas para indicar el peso de un caso en la regresión es la “carga” del caso, entendiendo por ello una medida de la distancia entre cada punto observado y el centro de todos ellos en el conjunto de variables **X**.

Esta medida procede de la matriz proyección (**H**), que es la que convierte los valores reales de **y** en valores predichos, de acuerdo con la siguiente expresión:

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \quad (10.15)$$

La matriz **H** se obtiene a partir de la matriz **X** de valores diferenciales de *x*, es decir,  $(x_i - \bar{x})$ , a la que se le agrega como primera columna el vector de unos, que representa la constante, de acuerdo con la siguiente expresión matricial:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (10.16)$$

La matriz **H** es una matriz de orden  $n \times n$ . Los elementos de la diagonal de **H** son las llamadas cargas (*leverage*), que toman un valor comprendido entre  $1/n$ , en la circunstancia de que un caso tenga los valores de todas las variables igual a sus respectivas medias, y 1, cuando un individuo posee valores totalmente extremos en todas las variables.

Mediante Stata estas cargas pueden calcularse mediante la opción *leverage* o *hat* de la instrucción *predict* seguida del nombre de la nueva variable con la que serán reconocidas.

```
predict carga, leverage
```

Con esta instrucción se añade una nueva variable al archivo de datos llamada *carga*, cuyos valores extremos pueden ser listados. En este caso, para listar los diez casos con mayores cargas, puede procederse del siguiente modo:

```
gsort -carga
list pais l_pnbppa lintfno carga in 1/10
```

Exponiendo en la lista la variable que identifica el caso (*pais*), las variables independientes y el índice de carga.

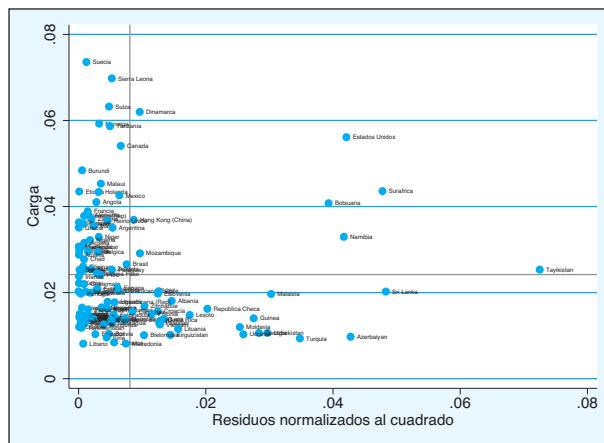
### ILUSTRACIÓN 10.10. Listado del índice de carga

	país	l_pnbppa	lintfno	carga
1.	Suecia	9.943861	674	.0735858
2.	Sierra Leona	6.025866	4	.0697807
3.	Suiza	10.22143	675	.0632281
4.	Dinamarca	10.09741	660	.0619916
5.	Noruega	10.18573	660	.0592642
6.	Tanzania	6.169611	4	.0586668
7.	Estados Unidos	10.32876	661	.0561181
8.	Canada	10.07428	634	.0540864
9.	Burundi	6.315358	3	.0484039
10.	Malawi	6.364751	3	.0453132

Como puede apreciarse han sobresalido dos tipos muy distintos de países: por un lado, Suecia, Suiza, Dinamarca, Noruega, Estados Unidos y Canadá tienen valores altos tanto en PNB como en teléfonos y, por el otro, Sierra Leona, Tanzania, Burundi y Malawi los tienen bajos en ambas variables.

Además del listado, es útil una representación gráfica de estos valores cruzados con los residuos. Esta opción se obtiene inmediatamente con la orden *lvr2plot*.

### GRÁFICO 10.6. Gráfico de cargas sobre residuos normalizados al cuadrado



En este gráfico, se ve que, con la excepción de Estados Unidos, el resto de los países que tienen alta carga (por encima de Malawi) poseen bajos residuales, por lo que no han de preocupar en la regresión, por mucho que sus cargas en las variables independientes sean considerables. Para obtenerlo, se ha introducido la siguiente instrucción:

```
lvr2plot, mlabel(pais) ytitle(Carga) xtitle("Residuos normalizados al cuadrado")
```

Una medida de la contribución que un caso tiene en un coeficiente de regresión es *dfbeta*, que representa el cambio en desviaciones típicas que sufre el coeficiente de una determinada variable al incluir un nuevo caso (Belsley 1980). Su fórmula es, pues, la resta de los dos coeficientes (con y sin la unidad añadida) dividido por el error típico de esta última, que se obtiene al dividir el error típico de la regresión sin el caso en cuestión ( $s_{e(i)}$ ) por la raíz cuadrada de la suma cuadrática de los residuos de una regresión en la que la variable dependiente es aquella ( $k$ ) de la que se calcula el coeficiente, y los predictores son el resto de las variables independientes ( $\sqrt{SCRes_k}$ ):

$$DFBETA_{ik} = \frac{b_k - b_{k(i)}}{s_{e(i)} / \sqrt{SCRes_k}} \quad (10.17)$$

En consecuencia, su valor puede ser tanto positivo (si el caso contribuye a aumentar el coeficiente) como negativo (si influye hacia la baja) y puede ser preocupante en el caso de que el valor absoluto sea superior a 1, pues modificaría el valor del coeficiente de regresión en más de un error típico. Sin embargo, Belsey (1980) sugiere que se compare con  $2/\sqrt{n}$ .

El modo de obtener estas medidas es especificando la instrucción *dfbeta* después de una regresión. En el caso de que no se mencione ninguna variable, calculará para cada caso las de todas las variables. A continuación se muestra la lista completa de instrucciones para que queden listados los casos que superen el cociente mencionado en el anterior párrafo.

```
regress l_tmi l_pnbppa lintfno  
dfbeta  
list pais _dfb* if (_dfbeta_1>2/sqrt(e(N)) | _dfbeta_2>2/sqrt(e(N))) & e(sample)
```

Es de notar que estas nuevas medidas son denominadas como *\_dfbeta\_#*<sup>4</sup>. También debe aclararse que el número de casos de una regresión queda registrado en el programa en la constante *e(N)*. Así pues, el resultado de la última instrucción muestra todos los casos en los que cualquier valor de *dfbeta* es mayor que  $2/\sqrt{n}$ :

---

<sup>4</sup> En versiones anteriores de Stata se denominaban con el nombre de la variable independiente precedida por las letras mayúsculas DF.

### ILUSTRACIÓN 10.11. Listado de las *dfbetas*

	país	_dfbeta_1	_dfbeta_2
4.	Dinamarca	-.1117166	.2098867
7.	Estados Unidos	-.1412062	.3598529
11.	Suráfrica	.4786185	-.4590031
16.	Botsuana	.4019008	-.4104398
28.	Namibia	.3554343	-.3683658
46.	Tayikistán	.3639386	-.2352885
60.	Sri Lanka	-.2223299	.2737852
65.	Malasia	-.2122332	.1815071

Como puede apreciarse, en ambas variables (*pnbppa* y *lintfno*) el caso que más perturba los coeficientes con diferencia sobre el resto es el de Suráfrica. Es recomendable realizar una regresión sin la presencia de este caso, solicitando una regresión con el resto de los países:

```
regress l_tmi l_pnbppa lintfno if país!="Suráfrica"
```

... con el siguiente resultado:

### ILUSTRACIÓN 10.12. Regresión con omisión de un caso influyente

Source	SS	df	MS	Number of obs	=	123
Model	134.109956	2	67.0549779	F( 2, 120)	=	459.08
Residual	17.5275229	120	.146062691	Prob > F	=	0.0000
Total	151.637479	122	1.24293015	R-squared	=	0.8844
				Adj R-squared	=	0.8825
				Root MSE	=	.38218

l_tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
l_pnbppa	-.5560512	.0671759	-8.28	0.000	-.6890548    -.4230475
lintfno	-.0021114	.0003748	-5.63	0.000	-.0028535    -.0013693
_cons	8.212053	.4967513	16.53	0.000	7.22852    9.195586

Si se compara esta regresión de 123 países con la de 124, se nota una variación pequeña. Sin embargo, al excluir los siete países con mayor influencia en los coeficientes, la regresión gana en explicación y los coeficientes salen bastante más diferentes que en el caso anterior.

```
regress l_tmi l_pnbppa lintfno if _dfbeta_1<2/sqrt(e(N)) & _dfbeta_2<2/sqrt(e(N))
```

... en cuyo caso se obtiene una regresión con 116 y el resultado siguiente:

### ILUSTRACIÓN 10.13. Regresión con omisión de varios casos influyentes

Source	SS	df	MS	Number of obs = 123		
Model	134.109956	2	67.0549779	F( 2, 120)	=	459.08
Residual	17.5275229	120	.146062691	Prob > F	=	0.0000
			R-squared = 0.8844			
Total	151.637479	122	1.24293015	Adj R-squared	=	0.8825
			Root MSE = .38218			
l_tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
l_pnbppa	-.5560512	.0671759	-8.28	0.000	-.6890548	-.4230475
lintfno	-.0021114	.0003748	-5.63	0.000	-.0028535	-.0013693
_cons	8.212053	.4967513	16.53	0.000	7.22852	9.195586

Existen tres medidas que son análogas a  $dfbeta$ , pero, a diferencia de esta, obtienen un único resultado para la regresión, en lugar de uno para cada variable independiente. Tratan, por tanto, de medir la influencia de un caso sobre el modelo en conjunto. Son la  $Dfits$  y las distancias de Cook ( $D_i$ ) y de Welsch ( $W_i$ ). Todas ellas son una transformación de los residuales studentizados ( $t_i$ ) o tipificados ( $z_i$ ) por el peso que tienen en cada caso los valores de la variable independiente, es decir, por su carga ( $h_i$ ).

Así, en el caso de la medida  $Dfits$ , esta se calcula mediante la siguiente expresión:

$$Dfits_i = t_i \sqrt{\frac{h_i}{1 - h_i}} \quad (10.18)$$

En tanto que para obtener la distancia de Welsch se transforma el residuo studentizado con esta otra expresión:

$$W_i = t_i \frac{\sqrt{h_i(n - 1)}}{1 - h_i} \quad (10.19)$$

Finalmente, la distancia de Cook es cuadrática y se obtiene a partir de los residuos tipificados.

$$D_i = z_i^2 \frac{h_i}{k(1 - h_i)} \quad (10.20)$$

Con Stata se obtienen de modo similar a las otras medidas aplicables a los casos, es decir, mediante la instrucción *predict*. En las respectivas opciones son *dfits*, *cooked* y *welsch*. Todas pueden obtenerse mediante una sola línea en el caso de que se utilice el bucle mediante la orden *for*, pero con anterioridad se vuelve a incluir la regresión para que queden incluidos todos los países.

```
regress l_tmi l_pnbppa lintfno
for any dfits cooksd welsch: predict l_tmi_X, X
```

Tras ser obtenidas, se puede solicitar el listado de aquellos casos con valores por encima de los recomendados:

```
list pais l_tmi_* if (abs(l_tmi_dfits)>2*sqrt((e(df_m)+1)/e(N))) | ///
l_tmi_cooksd>4*e(N) | ///
abs(l_tmi_welsch)>3*sqrt(e(df_m)+1)) & e(sample)
```

En este caso, se emplea *e(N)* para expresar el número de casos válidos en la última regresión, *e(df\_m)+1* para indicar el número de parámetros de la regresión, es decir, el número de variables independientes (grados de libertad de la suma cuadrática de la regresión) más una unidad, y se utiliza *e(sample)* para listar sólo los países con datos válidos en la última regresión estimada evitando que aparezcan en las líneas aquellos con valores omitidos.

#### ILUSTRACIÓN 10.14. Listado de otras distancias

	país	l_tmi_d~s	l_tmi_~d	l_tmi_w~h
7.	Estados Unidos	.5771238	.106956	6.588138
11.	Suráfrica	.5361364	.0917866	6.079978
16.	Botswana	.4668481	.0702527	5.286491
28.	Namibia	.4294298	.059309	4.843116
46.	Tayikistán	-.5004939	.0779316	-5.622381
60.	Sri Lanka	-.3583613	.0410365	-4.015223

En este listado aparecen las tres distancias mencionadas en los seis casos cuyas distancias están por encima (o por debajo) de los límites recomendables. Sin ser eliminados de la regresión, es evidente que se producirá una mejora del ajuste, obteniendo una *R*<sup>2</sup> ajustada por encima también de 0,91.

```
regress l_tmi l_pnbppa lintfno if ~(abs(l_tmi_dfits)>2*sqrt((e(df_m)+1)/e(N))) | ///
l_tmi_cooksd>4/e(N) | ///
abs(l_tmi_welsch)>3*sqrt(e(df_m)+1))
```

Para terminar, otra de las medidas de la influencia que un caso puede proporcionar en una regresión obtenible a través de la instrucción *predict es covratio*. Mide el cambio que supone la eliminación de un caso en la matriz de varianzas-covarianzas de los estimadores. Conceptualmente es el cociente entre los determinantes de ambas matrices, pero su valor puede calcularse también mediante la siguiente expresión:

$$Covratio = \frac{1}{1 - h_i} \left( \frac{n - k - z_i^2}{n - k} \right)^k \quad (10.21)$$

En el caso de que un determinado caso no tenga influencia alguna sobre las varianzas y covarianzas de los estimadores, el valor de este estadístico es 1. A juicio de Belsley, Kuh y Welsch (1980), el valor absoluto de esta medida menos una unidad ha de ser menor de  $3k/n$ . De otro modo tendría que examinarse cuidadosamente la observación que no tenga estas características. Para realizar esta exploración con Stata, tras la ejecución de la regresión, ha de generarse la medida, ordenar, si se desea, los casos por su valor y luego listar o representar los casos que no cumplen este criterio.

```
predict l_tmi_cov, covratio
sort l_tmi_cov
list pais l_tmi_cov if abs(l_tmi_cov-1)>=3*(e(df_m)+1)/e(N) & e(sample)
```

En este ejemplo son nueve los casos que contienen una razón de covarianzas superior a lo deseado o, dicho de otro modo, que con su eliminación hacen variar sustancialmente los errores típicos de los estimadores y las covarianzas entre ellos. Estos son:

#### **ILUSTRACIÓN 10.15. Listado de la razón de covarianzas**

	pais	l_tmi_~v
1.	Tayikistan	.8341598
2.	Sri Lanka	.8991215
3.	Azerbaiyan	.9068658
4.	Surafrica	.9191754
120.	Burundi	1.075498
121.	Suiza	1.077698
122.	Noruega	1.078582
123.	Sierra Leona	1.083683
124.	Suecia	1.102308

### 10.3. Regresiones especiales

Después de estudiar los problemas que pueden plantearse en una regresión, a continuación se dan una serie de técnicas de regresión que pueden solucionarlos, o al menos diagnosticarlos con más precisión. Es obvio que lo que se va a estudiar a continuación no es el único remedio a los problemas derivados de un no cumplimiento de los supuestos de la regresión o de la presencia de casos anómalos en el análisis. Algunos de ellos, como la transformación de las escalas de las variables o la eliminación de casos anómalos, ya han sido abordados en las páginas precedentes. A continuación, lo que se verá son otros modos de realizar la regresión que se consideran regresiones robustas en la medida en que sus estimaciones son más resistentes a la presencia de incumplimientos en los requisitos del modelo.

#### 10.3.1. *Errores típicos robustos*

La primera técnica que se presenta en esta segunda parte del segundo capítulo dedicado a las regresiones es el cálculo de errores típicos robustos.

Esta técnica está especialmente indicada para cuando los datos no cumplen la asunción de que los términos de error sean independientes de los predictores y estén homogéneamente distribuidos (homocedasticidad). Como se ha señalado anteriormente, en esas condiciones los estimadores son ineficientes, cuando no sesgados. En dichas condiciones adversas, debe asumirse que no se están realizando predicciones de la población verdadera, sino que se intenta generalizar los resultados a un conjunto de muestras realizadas en condiciones semejantes a las de aquella con la que se trabaja.

Stata ha implementado un modo fácil de calcular los errores típicos bajo este supuesto menos restrictivo de las regresiones, siguiendo los trabajos de Huber (1967) y White (1982), bajo el principio de la máxima verosimilitud. Basta con añadir a la instrucción de la regresión la opción *robust*. De este modo, sale una regresión con estimaciones idénticas de los parámetros, pero con errores típicos mayores, que conducirán a ser más exigentes a la hora de rechazar sus respectivas hipótesis nulas.

```
regress l_tmi l_pnbppa lintfno, robust
```

Si se compara la ilustración 10.8 con la obtenida mediante la anterior instrucción (ilustración 10.16), se notan dos diferencias principales: la primera es que en esta no aparece la tabla de suma de cuadrados, pero más importante aún es que los errores típicos son mayores que los calculados con el método de mínimos cuadrados ordinarios. Sin embargo, las estima-

ciones de los coeficientes son exactamente iguales. En cualquier caso, con asociaciones tan manifiestas, siguen saliendo los tres coeficientes, constante incluida, significativos.

### ILUSTRACIÓN 10.16. Regresión con errores típicos robustos

Linear regression		Number of obs = 124 F( 2, 121) = 601.43 Prob > F = 0.0000 R-squared = 0.8787 Root MSE = .39048				
		Robust				
		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
l_tmi						
l_pnbppa		-.5244678	.0704985	-7.44	0.000	-.6640382 - .3848974
lintfno		-.0022806	.000407	-5.60	0.000	-.0030865 - .0014748
_cons		7.98921	.5093761	15.68	0.000	6.980765 8.997654

#### 10.3.2. Regresiones ponderadas

Además del método de mínimos cuadrados ordinarios, una regresión puede realizarse ponderando los datos con una determinada cantidad de tal forma que en la determinación de la recta influyan más unos casos que otros.

El caso más claro y radical para ponderar una regresión es cuando se utiliza una variable ficticia, puesto que esta funciona como un filtro de entrada del caso. De este modo, todas las observaciones que tengan el valor 1 en la variable de ponderación entran en la ecuación, en tanto que aquellas que tengan el valor 0 o el valor perdido no figurarán en la regresión. El modo más simple de ponderación que dispone Stata es mediante la especificación general de los pesos, es decir, escribiendo en la instrucción *regress* el modificador [*weight* = variable]. Esta operación es equivalente a la escritura de un condicional.

Por ejemplo, si se desea hacer una regresión sólo con países europeos, suponiendo que se ha creado una variable ficticia denominada *Europa* con el valor 1 para los países de este continente, la escritura de estas dos instrucciones daría resultados completamente iguales:

```
regress evn tmi if Europa==1
regress evn tmi [weight=Europa]
```

El resultado de estas dos últimas instrucciones sólo se diferencia en la advertencia sobre el tipo de ponderación que se está realizando, como se muestra a continuación:

### ILUSTRACIÓN 10.17. Regresión con ponderación analítica

(analytic weights assumed) (sum of wgt is 3.5000e+01)					
Source	SS	df	MS	Number of obs = 35 F( 1, 33) = 32.21 Prob > F = 0.0000 R-squared = 0.4939 Adj R-squared = 0.4786 Root MSE = 2.9257	
Model	275.704472	1	275.704472		
Residual	282.466957	33	8.55960475		
Total	558.171429	34	16.4168067		
<hr/>					
evn	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
tmi	-.290591	.051202	-5.68	0.000	-.3947623 - .1864197
_cons	76.97623	.7506194	102.55	0.000	75.44909 78.50338

Efectivamente, si no se especifica el tipo de ponderación que se desea, el programa que calcula la regresión ponderada asume un tratamiento analítico de los pesos. Quiere ello decir que (salvo para datos sin información [casos perdidos] o ponderaciones iguales a 0) los pesos sólo se tienen en cuenta para el cálculo de los estimadores y en el número de observaciones aparece el número de casos efectivos que hay y no la suma de las ponderaciones, como así sucedería en el caso de que las ponderaciones fueran frecuenciales (*fweight*).

Esta diferencia, así como la comprensión de lo que se hace al ponderar los casos de una regresión, puede verse claramente si se pone un ejemplo con pocos casos. Supóngase que se dispone de los tres siguientes casos: (1;5) (2;4) y (2;6), siendo, respectivamente, el primer valor el de X y el segundo el de Y. Para que quede más claro, se presenta un listado de los tres casos con todas las variables necesarias para el tratamiento del ejemplo:

### ILUSTRACIÓN 10.18. Ejemplo imaginario para ver casos de regresiones ponderadas

	caso	X	Y	peso2	peso3
1.	A	1	5	1	1
2.	B	2	4	2	1
3.	C	2	6	1	2

Una primera regresión con estos datos arrojaría una recta totalmente horizontal, que pasaría entre los puntos segundo y tercero.

```
regress Y X
```

### ILUSTRACIÓN 10.19. Regresión sin ponderar

Source	SS	df	MS	Number of obs = 3		
Model	0.00	1	0.00	F( 1, 1) = 0.00		
Residual	2.00	1	2.00	Prob > F = 1.0000		
Total	2.00	2	1.00	R-squared = 0.0000		
				Adj R-squared = -1.0000		
				Root MSE = 1.4142		
<hr/>						
Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
X	0	1.732051	0.00	1.000	-22.00779	22.00779
_cons	5	3	1.67	0.344	-33.11861	43.11861

Sin embargo, al ponderar uno de los dos últimos casos dándole un valor superior a la unidad, se obtendría una recta más orientada al punto que representa. Así, si se pondera más al segundo punto (*peso2*), el valor del coeficiente de la regresión será negativo, puesto que la recta se le aproxi- mará, y al estar colocado en Y por debajo de la media, la línea descenderá. En cambio, si se pondera más al tercero (*peso3*), el coeficiente será positivo en la medida que se le da más importancia al caso cuyo valor de *Y* está por encima de la media.

De este modo, la regresión, ponderando el doble el segundo caso, sería la siguiente:

```
regress Y X [weight=peso2], nohead
predict R2
```

### ILUSTRACIÓN 10.20. Regresión ponderada (*peso2*)

(analytic weights assumed)					
(sum of wgt is 4.0000e+00)					
<hr/>					
Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
X	-.3333333	1.885618	-0.18	0.889	-24.29238 23.62572
_cons	5.333333	3.399346	1.57	0.361	-37.85946 48.52612

Y la efectuada dándole el doble valor al tercer caso sería esta otra si- guiente:

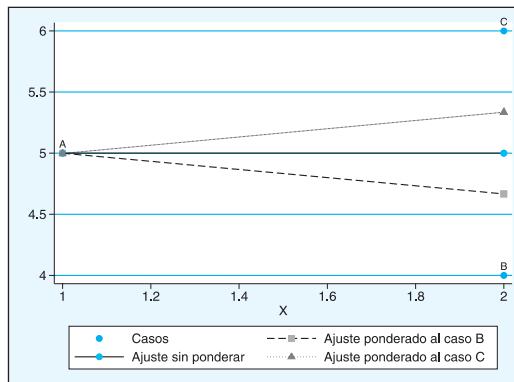
```
regress Y X [weight=peso3], nohead
predict R3
```

### ILUSTRACIÓN 10.21. Regresión ponderada (peso3)

(analytic weights assumed) (sum of wgt is 4.0000e+00)						
	Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
	X	.3333333	1.885618	0.18	0.889	-23.62572 24.29238
	_cons	4.666667	3.399346	1.37	0.401	-38.52612 47.85946

El gráfico con las tres regresiones tendría este aspecto:

GRÁFICO 10.7. Representación de las tres regresiones de distintos pesos



Para intuir el sentido geométrico de la ponderación, es conveniente ver en el gráfico el sencillo ejemplo propuesto: la recta inclinada superior tiene situado el valor predicho de la derecha (marcado con un triángulo) el doble de cerca al caso C (que se ha ponderado el doble) que al caso B. En cambio, la recta con倾inación negativa tiene su extremo derecho (con símbolo cuadrado) el doble de cerca del caso B, porque este ha sido ponderado con un valor superior en dos veces al caso C.

Como ya se sabe, existen cuatro posibilidades de ponderación en Stata: ponderación de frecuencias (*fweight*), poblacional (*pweight*), analítica (*aweight*) y específica (*iweight*). Por omisión Stata realiza la ponderación analítica, que es la más apropiada para regresiones especiales. La otra importante es la de frecuencias. La principal diferencia entre ambas es que mientras en la primera el número de casos se mantiene constante, en la segunda se recalcula el número de casos de la regresión sumando los pesos del conjunto de datos.

Para verlo con más claridad, es útil la comparación de dos regresiones iguales, con la sola diferencia del tipo de ponderación:

```
regress Y X [aweight=peso2]
```

### ILUSTRACIÓN 10.22. Regresión con ponderación analítica

(sum of wgt is 4.0000e+00)				Number of obs = 3				
Source	SS	df	MS	F( 1, 1) = 0.03				
Model	.0625	1	.0625	Prob > F = 0.8886				
Residual	2.00	1	2.00	R-squared = 0.0303				
Total	2.0625	2	1.03125	Adj R-squared = -0.9394				
				Root MSE = 1.4142				
-----								
Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]			
X	-.3333333	1.885618	-0.18	0.889	-24.29238	23.62572		
_cons	5.333333	3.399346	1.57	0.361	-37.85946	48.52612		

En esta regresión, aunque advierte que la suma de pesos es igual a cuatro (hay dos casos con una ponderación igual a la unidad y uno con el doble de peso), en el número de observaciones aparecen tres, y todos los estadísticos que dependen de ello (tanto los errores típicos y las medias cuadráticas como también las sumas cuadráticas y, en consecuencia, el  $R^2$  ajustado, en este caso con un valor irreal por la escasez de casos) se ven afectados.

En cambio, al realizar una ponderación de frecuencias, [*fweight*], el resultado es este otro:

### ILUSTRACIÓN 10.23. Regresión con ponderación de frecuencias

Source   SS df MS				Number of obs = 4				
Source	SS	df	MS	F( 1, 2) = 0.06				
Model	.083333333	1	.083333333	Prob > F = 0.8259				
Residual	2.66666667	2	1.33333333	R-squared = 0.0303				
Total	2.75	3	.916666667	Adj R-squared = -0.4545				
				Root MSE = 1.1547				
-----								
Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]			
X	-.3333333	1.333333	-0.25	0.826	-6.070204	5.403537		
_cons	5.333333	2.403701	2.22	0.157	-5.008957	15.67562		

Como puede apreciarse, sólo coinciden el  $R^2$  y los coeficientes de la regresión. Y es tan bajo el coeficiente de determinación que al ajustarlo da un resultado negativo.

### 10.3.3. Regresión de mínimos cuadrados generalizados (ponderados)

Una de las aplicaciones de la ponderación analítica de las observaciones es eliminar el error producido por la presencia de heterocedasticidad en los datos. Recuérdese que por este término se entienden varianzas desiguales en el término de perturbación según el valor de la(s) variable(s) independiente(s) y que la consecuencia estriba en que el error típico de los estimadores calculado por el método de mínimos cuadrados ordinarios es sesgado. El remedio de la heterocedasticidad consiste en ponderar los casos de la regresión por  $w_i = 1/\sigma_i^2$ , pero, como lo que realmente se ponderan son los residuos cuadráticos, se consiguen obtener los estimadores de la regresión transformando sus variables (y su constante) por  $\sqrt{w_i}$ , en consecuencia, en este caso por  $1/\sigma_i$ . Es decir, se ha de concebir una regresión con todos los términos de la ecuación divididos por  $\sigma_i$ , lo que conduce a que el término de error se transformará en constante:

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_1}{\sigma_i} + \beta_2 \frac{x_2}{\sigma_i} + \dots + \beta_k \frac{x_k}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i} \quad (10.22)$$

El resultado de esta transformación es que el nuevo término de error  $\varepsilon_i/\sigma_i$  tendrá varianza constante, puesto que al dividir cada perturbación por su desviación típica, siempre adoptará el valor constante de la unidad.

De este modo, el cálculo de los coeficientes de regresión difiere al introducir las ponderaciones, convirtiéndose en la siguiente fórmula:

$$\beta_1 = \frac{\left( \sum_{i=1}^n w_i \right) \left( \sum_{i=1}^n w_i x_i y_i \right) - \left( \sum_{i=1}^n w_i x_i \right) \left( \sum_{i=1}^n w_i y_i \right)}{\left( \sum_{i=1}^n w_i \right) \left( \sum_{i=1}^n w_i x_i^2 \right) - \left( \sum_{i=1}^n w_i x_i \right)^2} \quad (10.23)$$

Y en el cálculo de la varianza del estimador también intervienen lógicamente las ponderaciones:

$$Var(\beta_1) = \frac{\left( \sum_{i=1}^n w_i \right)}{\left( \sum_{i=1}^n w_i \right) \left( \sum_{i=1}^n w_i x_i^2 \right) - \left( \sum_{i=1}^n w_i x_i \right)^2} \quad (10.24)$$

Algebraica y lógicamente, en el caso de que todas las ponderaciones ( $w_i$ ) sean iguales, estas fórmulas son equiparables con las que se utilizan en el caso de la estimación por mínimos cuadrados ordinarios.

Stata proporciona al menos dos modos de realizar regresiones ponderadas mediante la varianza. La más común es mediante la instrucción *regress*, a la que se le especifica como peso la inversa de la varianza.

Recurriendo al ejemplo anterior de tres casos y suponiendo que la variable *peso* refleja la desviación típica de las perturbaciones y *peso2* su varianza...

#### **ILUSTRACIÓN 10.24. Matriz de ejemplo para regresión de mínimos cuadrados ponderados**

	Y	X	peso	peso2
1.	5	1	1	1
2.	4	2	1.414214	2
3.	6	2	1	1

... el modo de formular una regresión mediante mínimos cuadrados ponderados sería del siguiente modo:

```
regress Y X [aweight=1/peso2]
```

Y el resultado será el que presenta cualquier regresión ponderada:

#### **ILUSTRACIÓN 10.25. Regresión de mínimos cuadrados ponderados**

(sum of wgt is 2.5000e+00)				Number of obs = 3			
Source	SS	df	MS	F( 1, 1) = 0.05		Prob > F = 0.8600	
Model	.08	1	.08	R-squared = 0.0476	Adj R-squared = -0.9048	Root MSE = 1.2649	
Residual	1.60	1	1.60				
Total	1.68	2	.84				
<hr/>							
Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]		
X	.3333333	1.490712	0.22	0.860	-18.60796	19.27463	
_cons	4.666667	2.494438	1.87	0.313	-27.02818	36.36151	

Pero Stata contiene también un programa especial (*vwls*) que calcula el error típico bajo otras asunciones. *Variance-weighted least squares* realiza este tipo de regresiones con dos modalidades: una en la que el investigador le proporciona la varianza de cada caso y otra en la que el mismo procedimiento lo calcula, siempre y cuando los valores de las variables indepen-

dientes tengan una determinada agrupación que permita calcular la varianza de los términos de perturbación.

El primer caso es útil cuando se está ante un experimento y se piensa que las variaciones de la perturbación son solamente debidas a errores de medida. Para obtener la regresión bajo estos supuestos, se debe especificar la opción *sd(variable)*, con el nombre de la variable que recoja la desviación típica de los mencionados errores. Así, continuando con el ejemplo anterior, al escribir...

```
vwls Y X, sd(peso)
```

... se obtiene la siguiente regresión:

#### **ILUSTRACIÓN 10.26. Regresión de mínimos cuadrados ponderados por la varianza**

Variance-weighted least-squares regression					Number of obs	=	3
Goodness-of-fit chi2(1) = 1.33					Model chi2(1)	=	0.07
Prob > chi2 = 0.2482					Prob > chi2	=	0.7963
<hr/>							
Y	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]		
X	.3333333	1.290994	0.26	0.796	-2.196969	2.863636	
_cons	4.666667	2.160247	2.16	0.031	.4326606	8.900673	

Nótese que, aunque los coeficientes sean iguales, los errores típicos son diferentes, al calcularse bajo otros supuestos, en este caso el del conocimiento certero de los estimadores del error de la variable *Y*. Asimismo, en lugar del estadístico *F*, se utiliza el estadístico *Q*, con una distribución  $\chi^2$  de *n-k-1* grados de libertad. En este caso, como la probabilidad de este estadístico (el de la izquierda) es superior a 0,05, puede considerarse una regresión no significativa.

El otro modo de funcionamiento del programa *vwls* es cuando se tienen valores agrupados de  $x_i$ , en cuyo caso basta con que se especifiquen la variable dependiente y las independientes, omitiendo la opción *sd*. Si existen suficientes valores para cada  $x_i$ , entonces el programa calcula la varianza de sus respectivos términos de perturbación y pondrá la regresión con su inversa.

#### **10.4. Regresión robusta**

Una de las posibilidades de ponderar una regresión es a través de procedimientos iterativos que hagan que aquellos casos con residuos pequeños

tengan alta incidencia en el cálculo de los coeficientes, en tanto que los casos con residuos grandes tengan un peso muy pequeño, y si son incluso muy grandes, no tengan ninguna repercusión en la estimación de los parámetros. Este procedimiento es especialmente útil cuando los términos de perturbación no tienen una distribución normal y hacen que el cálculo de los coeficientes sea sesgado.

El programa Stata incluye un procedimiento para hacer este tipo de regresiones similar al que propusiera Guoing Li (1985).

En primer lugar, se realiza una regresión por mínimos cuadrados. Se calcula el valor de la  $D$  de Cook y se eliminan (o se da una ponderación igual a 0) a aquellos casos en que este estadístico arroje un valor superior a la unidad.

Con estos pesos, se realiza una nueva regresión de la que se calculan los residuales ( $e_i$ ) y estos son transformados ( $u_i$ ) del siguiente modo:

Se calcula  $M$ , que es la mediana de la diferencia en términos absolutos entre cada residuo y su mediana, esto es, al tomar las restas en términos absolutos, viene a ser una medida de la dispersión de los residuos:

$$M = \text{med} |e_i - \text{med}(e_i)| \quad (10.25)$$

Posteriormente se calcula  $u_i$ , dividiendo el residuo ( $r_i$ ) por  $M/0,6745$  con el propósito de que sean más adelante infraponderados aquellos casos cuyo residual absoluto exceda dos veces la medida  $M$ .

$$u_i = \frac{e_i}{M/0,6745} \quad (10.26)$$

A estos residuos reescalados se les aplica la función de Huber, que consiste en dar ponderación 1 a todos aquellos casos con  $|u_i|$  inferior a una determinada constante  $c$  (que para este procedimiento Stata fija en 1,345)<sup>5</sup> y una ponderación inferior a esta cantidad, si esta cantidad absoluta está por encima de la constante. La notación matemática de esta operación es la siguiente:

$$w_i \begin{cases} 1 & \text{si } |u_i| \leq c \\ \frac{c}{|u_i|} & \text{en cualquier otro caso} \end{cases} \quad (10.27)$$

---

<sup>5</sup> 1,345 dividido por el anterior 0,6745 da un valor de 2. Por ello, este procedimiento lo que hace es ponderar por debajo de 1 a todos aquellos casos cuyos residuos están alejados de la mediana dos veces su valor.

Si estas ponderaciones ( $w_i$ ) se distancian de las anteriores (en la primera ocasión, de la unidad, puesto que es el punto de partida) una cantidad ínfima, denominada convergencia y establecida por omisión en 0,01, entonces se detiene el proceso para pasar al siguiente tipo de ponderación, el de *biweight*. Pero, en el caso de que la mayor de las diferencias entre las ponderaciones no sea tan pequeña, se vuelve a realizar otra regresión con los pesos de la última, que dará lugar a nuevas ponderaciones, que de nuevo son comparadas hasta la convergencia.

Cuando se alcanza dicha convergencia, se procede a un procedimiento similar, pero en lugar de utilizar la función de Huber, se emplea la de Beaton y Tukey (1974), denominada *biweight*. Mediante esta —al revés que en la función anterior— los casos excesivamente alejados de la mediana de los residuales son ponderados con 0 y a medida que se approxima a este promedio, alcanzan el valor de 1. La expresión que calcula estos pesos es la siguiente:

$$w_i \begin{cases} \left(1 - \left(\frac{u_i}{c}\right)^2\right)^2 & \text{si } |u_i| \leq c \\ 0 & \text{en cualquier otro caso} \end{cases} \quad (10.28)$$

En esta nueva ponderación, Stata utiliza como constante por omisión el valor 4,685, que hace que todos aquellos casos cuyos residuales se alejen del promedio siete veces<sup>6</sup> la desviación mediana tengan una ponderación igual a 0. Ahora bien, en este segundo paso, de sucesivas iteraciones, el usuario puede cambiar el valor de la constante, mediante la opción *tune()*, en la que se debe expresar cuántas veces alejado de la mediana se desea que la ponderación sea nula. Este valor, fijado por omisión en siete, se recomienda que esté comprendido entre 6 y 12.

La orden que ejecuta esta regresión robusta es *rreg* seguida de variables dependiente e independientes por este orden. Así, la instrucción...

```
use mundo999
for var tmi pnbpaa: generate l_X=ln(X)
rreg l_tmi l_pnbpaa lntfno
```

... da lugar al siguiente resultado:

---

<sup>6</sup> En este caso, el siete procede de dividir 4,685 entre el 0,6745 utilizado para calcular  $u_i$ .

### ILUSTRACIÓN 10.27. Regresión robusta

Huber iteration 1:	maximum difference in weights = .58401938				
Huber iteration 2:	maximum difference in weights = .07575524				
Huber iteration 3:	maximum difference in weights = .01236351				
Biweight iteration 4:	maximum difference in weights = .19114446				
Biweight iteration 5:	maximum difference in weights = .01413694				
Biweight iteration 6:	maximum difference in weights = .00148212				
Robust regression estimates	Number of obs = 124				
	F( 2, 121) = 462.99				
	Prob > F = 0.0000				
-----	-----				
l_tmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
l_pnbppa	-.5751774	.0667243	-8.62	0.000	-.7072758    -.443079
lintfno	-.002082	.0003728	-5.58	0.000	-.0028201    -.0013439
_cons	8.382187	.4942138	16.96	0.000	7.40376    9.360613
-----	-----	-----	-----	-----	-----

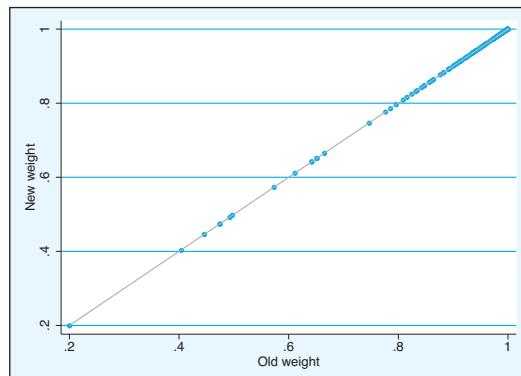
Como puede apreciarse, antes de la expresión de los estimadores aparece el historial de la obtención de la regresión robusta. En la primera iteración de Huber la máxima discrepancia en pesos era de 0,58; por eso sigue realizando otra ponderación y tras ella, el valor de la diferencia máxima cae a 0,07. En la siguiente a 0,01 y, como la próxima estaría por debajo de esta cantidad, pasa a realizar las iteraciones con el procedimiento *biweight*. En el tercero de ello, la diferencia máxima es tan pequeña que deja de buscar nuevos pesos.

El programa *rreg* tiene una opción que permite plasmar el gráfico que compara las dos últimas ponderaciones realizadas. Se trata de la opción, *graph*. Al incluirla en la instrucción del modo usual...

```
rreg l_tmi l_pnbppa lintfno, graph
```

... aparece el siguiente gráfico:

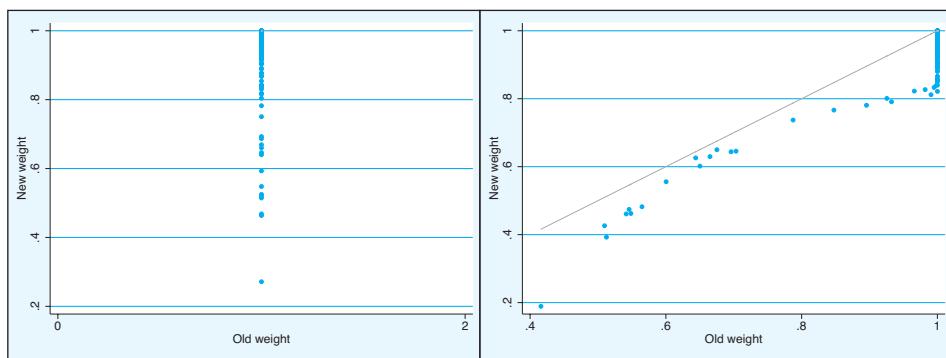
**GRÁFICO 10.8. Gráfico de comparación de las ponderaciones de la regresión robusta**



En este gráfico todos los puntos aparecen prácticamente en la recta, lo que indica una alta igualdad de los pesos en la regresión ponderada en la penúltima y en la última ponderación. Un ejemplo de ello nos lo proporcionan los pesos extremos: en el superior, no hay ninguna variación. Tanto en la escala de la “vieja” ponderación como en la de la nueva, el valor mayor es de 0,999997. En cambio, se nota una pequeña diferencia en el caso con menor ponderación, pues este en la penúltima iteración tiene un valor de 0,200245 y en la última de 0,198912. Aunque no pueda apreciarse, la recta marca los puntos en los que  $w_{i-1}=w_i$ .

Aunque el programa *rreg* no sea capaz de mostrarlos, es útil ver los gráficos de este tipo que se generarían en anteriores iteraciones distintas de la última, en la que, salvo que se ponga la tolerancia muy alta, prácticamente todos los puntos deberían coincidir con la recta.

**GRÁFICO 10.9. Secuencia de gráficos de la regresión robusta**



En el gráfico de la izquierda aparece el cruce de pesos tras la primera iteración. El punto de partida (el antiguo peso) es que todos los casos tengan la ponderación igual a 1, pero, según los criterios de Huber, aquellos que estén por dos veces alejados de la desviación mediana reciben un peso tanto menor cuanto más alejado esté de aquella. Por otro lado, a la derecha, aparece el cruce tras realizar el paso de iteraciones de Huber a Biweights. Los casos alineados en el extremo superior indican todos aquellos que por alejarse poco del valor de la mediana Huber los ponderaba con la unidad. En el caso de biweight, estos casos se ajustan no a una igualdad, sino a una fórmula y por eso no son exactamente iguales a 1 ni tan siquiera en el primer paso. El resto son tanto más parecidos a los anteriores cuanto más se ubiquen en el centro, adoptando la relación una forma curvilínea.

El programa *rreg* contiene otra opción que permite cambiar el límite a partir del cual deja de realizar iteraciones. Se trata de la opción *tolerance(#)*, que deja de realizar una nueva iteración en busca de nuevos pesos, cuando la mayor diferencia entre la anterior y la posterior es menor que el número

proporcionado, #, que obviamente ha de estar comprendido entre 0 y 1<sup>7</sup>. También puede limitarse el número de iteraciones directamente, mediante la opción *iterate(#)*, cuyo valor por omisión está fijado en 1.000. Si aún no ha convergido la solución y ya se ha alcanzado el número de iteraciones solicitadas, entonces el programa se detiene si la última ha sido una iteración *biweight*, mientras que si la última ha realizado una Huber, aplica una del otro tipo para cerrar el proceso.

Una de las opciones más útiles del programa de la regresión robusta es la que genera una nueva variable que contiene los pesos finales obtenidos tras las sucesivas iteraciones. Se trata de la opción *genwt(nombrevARIABLE)*, tras la que los resultados de la regresión son invariantes. Su utilidad consiste en poder disponer para cada caso de la ponderación que ha sido utilizada para obtener los parámetros finales. Si de desea conocer cuáles han sido los países que menos peso han tenido en la regresión, deberían añadirse dos instrucciones a la de la regresión robusta con la opción señalada.

```
rreg l_tmi l_pnbppa lintfno, genwt(w_ltmi)
sort w_ltmi
list pais w_ltmi in 1/10
```

Tras lo cual aparecen listados los diez países con pesos menores en la regresión robusta:

#### **ILUSTRACIÓN 10.28. Lista de pesos de los casos en la regresión robusta**

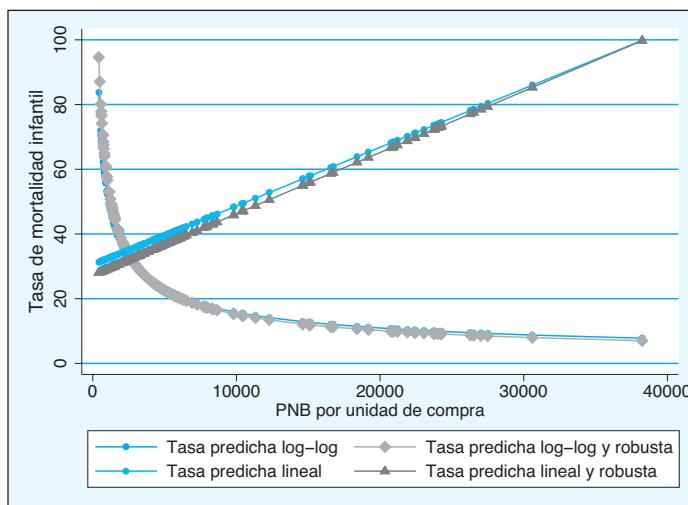
	pais	w_ltmi
1.	Tayikistan	.19891184
2.	Surafrica	.40272357
3.	Sri Lanka	.44528413
4.	Azerbaiyan	.47349819
5.	Namibia	.47353931
6.	Botsuana	.49146373
7.	Estados Unidos	.49712956
8.	Turquia	.57274883
9.	Uzbekistan	.61091855
10.	Malaisia	.6412604

Uno de ellos está manifiestamente infraponderado, contando con menos de un quinto de su valor, y otros seis más tienen una ponderación que no llega a su mitad.

<sup>7</sup> Este límite funciona para las ponderaciones *biweight*; para las Huber, se utiliza como límite este mismo número multiplicado por 5.

Es conveniente resaltar dos aspectos muy importantes de la regresión robusta. El primero de ellos es que, como en toda regresión de mínimos cuadrados ponderados, si calculamos su coeficiente de determinación, su valor será inferior al de mínimos cuadrados ordinarios. Esta es una de las razones por las que su valor no aparece en la salida. El segundo es que sólo es adecuada para solucionar problemas de los residuos, pero si no los hay se producen resultados muy similares a los de la regresión normal, pero equivocados, en el caso de que los datos poblacionales cumplan los supuestos del modelo. Un ejemplo visual de estas cuestiones lo ofrece el gráfico que a continuación se expone. Se han realizado cuatro regresiones: dos lineales y dos log-log y cada uno de estos pares se ha ajustado por mínimos cuadrados ordinarios y con regresión robusta. Como puede apreciarse, las diferencias entre las dos últimas modalidades son mínimas; en cambio, las regresiones meramente lineales son muy distintas de las otras en la medida en que están afectadas por un gran error de especificación.

**GRÁFICO 10.10. Gráfico de comparación de regresiones**



Para producir el gráfico anterior en el que la variable dependiente es la tasa de mortalidad infantil, la independiente el producto nacional bruto per cápita y la de control, el número de líneas telefónicas<sup>8</sup>, se ha utilizado la siguiente lista de instrucciones:

<sup>8</sup> Esta última variable está introducida en el gráfico en modo de control. Para ello se ha tomado como valor constante de ella su media aritmética.

```

summarize lintfno
local xtfno=r(mean)
regress tmi pnbppa lintfno
matrix coef=e(b)
generate t_tmi=coef[1,3]+coef[1,1]*pnbppa+coef[1,2]*`xtfno'
rreg tmi pnbppa lintfno
matrix rcoef=e(b)'
generate rt_tmi=rcoef[1,3]+rcoef[1,1]*pnbppa+rcoef[1,2]*`xtfno'
regress l_tmi l_pnbppa lintfno
matrix coef=e(b)
generate tl_tmix=coef[1,3]+coef[1,1]*l_pnbppa+coef[1,2]*`xtfno'
generate tl_tmiy=exp(tl_tmix)
rreg l_tmi l_pnbppa lintfno
matrix rcoef=e(b)
generate rtl_tmix=rcoef[1,3]+rcoef[1,1]*l_pnbppa+rcoef[1,2]*`xtfno'
generate rtl_tmiy=exp(rtl_tmix)
label variable t_tmi "Tasa predicha lineal"
label variable rt_tmi "Tasa predicha lineal y robusta"
label variable tl_tmiy "Tasa predicha log-log"
label variable rtl_tmiy "Tasa predicha log-log y robusta"
scatter tl_tmiy rtl_tmiy t_tmi rt_tmi pnbppa, connect (. . .) symbol(o . o .) ///
l1title("Tasa de mortalidad infantil") sort(pnbppa) name(I38, replace)

```

## 10.5. Regresión de cuantiles

Otro modo de conseguir regresiones robustas es la de realizar modelos basados en la estimación de la mediana, o cualquier otra medida de localización, en lugar de la media. Esto, en un primer momento, permite que los valores extremos de la variable dependiente tengan menos influencia en la configuración de la regresión. En efecto, en lugar de intentar predecir la media de  $y$  para cada valor de  $x$ , se trata de predecir la mediana. Por ello, la ecuación de este tipo de regresión se expresa del siguiente modo:

$$Q_p(y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon_i \quad (10.29)$$

... siendo  $p$  un número entre 0 y 1, ambos excluidos, que representa el cuantil sobre el que se quiere realizar la regresión. El caso más común, que adopta el programa por omisión, es el del valor 0,50, que representa a la mediana. De este modo, al escribir la siguiente instrucción...

```
qreg tmi l_pnbppa
```

... se ofrecen las siguientes estimaciones de los coeficientes de la regresión:

### ILUSTRACIÓN 10.29. Regresión de cuantiles

```
Iteration 1: WLS sum of weighted deviations = 2025.2391
Iteration 1: sum of abs. weighted deviations = 2007.8504
Iteration 2: sum of abs. weighted deviations = 1998.5785
Iteration 3: sum of abs. weighted deviations = 1997.8248

Median regression                               Number of obs =      125
  Raw sum of deviations      3802 (about 28)          Pseudo R2     =    0.4745
  Min sum of deviations 1997.825

-----
          tmi |     Coef.   Std. Err.      t    P>|t|    [95% Conf. Interval]
-----+
l_pnbppa |  -26.51133  2.070181   -12.81  0.000   -30.60912  -22.41353
_cons |   263.1863   17.2808    15.23  0.000    228.98   297.3926
-----+
```

Ahora bien, la mejor forma de comprender la estimación de la regresión cuantílica es con un sencillo ejemplo en el que la variable independiente sea dicotómica y se tome el cuantil 50, es decir, la mediana. Imagínese que tenemos seis casos divididos en dos grupos (a los que se les da arbitrariamente los valores de 0 y 1); en el primer grupo los valores de la variable dependiente son 1, 2 y 3; mientras que los correspondientes al segundo grupo son 7, 8 y 9. Es evidente que la mediana del primer grupo ( $x=0$ ) es igual a 2; mientras que la del segundo ( $x=1$ ) es igual a 8. En consecuencia, el valor de la constante sería igual a 2 (la mediana en el grupo con valor 0 en  $x$ ) y el valor del coeficiente es igual a 6 (la diferencia entre las dos medianas). El ejemplo contendría los siguientes datos:

### ILUSTRACIÓN 10.30. Matriz de ejemplo para la regresión de cuantiles

	X	Y
1.	0	1
2.	0	2
3.	0	3
4.	1	7
5.	1	8
6.	1	9

... y la regresión adoptaría el siguiente aspecto:

**ILUSTRACIÓN 10.31. Regresión de cuantiles sobre el ejemplo ficticio**

Iteration 1: WLS sum of weighted deviations =	4
Iteration 1: sum of abs. weighted deviations =	4
Median regression	Number of obs = 6
Raw sum of deviations 18 (about 3)	Pseudo R2 = 0.7778
Min sum of deviations 4	
<hr/>	
Y   Coef. Std. Err. t P> t  [95% Conf. Interval]	
X   6 1.379796 4.35 0.012 2.169073 9.830927	
_cons   2 .975663 2.05 0.110 -.7088748 4.708875	
<hr/>	

Los valores de los coeficientes son muy obvios. También lo es la suma de desviaciones. En este caso, no se opera con desviaciones al cuadrado, sino absolutas. Además, los residuos son ponderados en función del cuantil con el que se esté estimando la regresión, ya que es lógico que las desviaciones hacia la izquierda sean menores que las de la derecha en el cuartil primero, por ejemplo, y viceversa en el tercero. Por ello, los valores absolutos de los residuos han de ponderarse por  $2Q$ , en el caso de que sean positivos y por  $2(1-Q)$  en el de que lo sean negativos. Así, en el caso de la estimación del primer cuartil, los residuos positivos (dados por valores altos) tienen un peso de 0,5, mientras que los negativos (es, decir, los correspondientes a los valores más bajos de la variable dependiente) tendrán una ponderación de 1,5, esto es, tres veces superior.

Sin tener en cuenta la variable independiente, los seis casos arriba mostrados arrojan con respecto a la mediana (cinco, esto es, la semisuma de los casos centrales, tres y siete) una suma de desviaciones igual a 18. Al introducir la información de la variable independiente, se dispone de dos medianas (tres y ocho), y con respecto a ellas, la suma de las desviaciones sólo alcanza el valor de 4. El pseudo  $R^2$  se obtiene mediante la siguiente expresión:

$$R^2 = 1 - \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{\sum_{i=1}^n |y_i - Q_p(y_i)|} \quad (10.30)$$

... donde  $Q_p(y_i)$  es el valor de la posición cuantílica dada en el conjunto de la muestra, e  $\hat{y}_i$  el valor predicho con la ecuación cuantílica correspondiente a cada uno de los valores distintos de  $x$ . Y, en este caso, el valor asciende al 78%, que representa la mejora en la estimación de la mediana que supone el conocimiento de la(s) variable(s) independiente(s).

El cálculo de los coeficientes, como puede apreciarse en los resultados, sigue una técnica iterativa. Se comienza con una aproximación utilizando mínimos cuadrados ponderados y, a partir de ahí, se cambia la recta predictiva siempre y cuando implique una mejora en la minimización del valor absoluto de los residuos.

La opción más útil en el programa *qreg* es *quantile(#)* donde debe especificarse un valor entre 0 y 1 del cuantil de la variable dependiente del que se desea realizar la regresión. Así, si se quiere realizar la predicción del primer cuartil la instrucción ha de adoptar la siguiente expresión:

```
qreg Y X, quantile(.25)
```

... en cuyo caso el resultado es:

### ILUSTRACIÓN 10.32. Regresión del primer cuartil

Iteration 1: WLS sum of weighted deviations = 3.6000001
Iteration 1: sum of abs. weighted deviations = 4
Iteration 2: sum of abs. weighted deviations = 3.5
Iteration 3: sum of abs. weighted deviations = 3
.25 Quantile regression Number of obs = 6
Raw sum of deviations 12 (about 1)
Min sum of deviations 3 Pseudo R2 = 0.7500
-----
Y   Coef. Std. Err. t P> t  [95% Conf. Interval]
X   6 .5129548 11.70 0.000 4.575809 7.424191
_cons   1 .3627138 2.76 0.051 -.0070549 2.007055

Existen otros programas similares, como *iqreg*, que calcula regresiones de rangos intercuartílicos, y *sqreg* y *bsqreg*, que realizan regresiones cuantílicas simultáneas y calculan los errores típicos con procedimientos iterativos.

Este programa, como todos aquellos que efectúan ajustes de líneas o curvas, permite añadir variables al fichero original. En concreto son tres las posibilidades: los valores predichos, el error típico de la predicción y el residual. Una de las posibilidades es la posterior representación gráfica del ajuste, como el siguiente ejemplo, que nos permite comparar la regresión lineal con la regresión de la mediana en la relación entre la tasa de mortalidad infantil y el producto nacional per cápita:

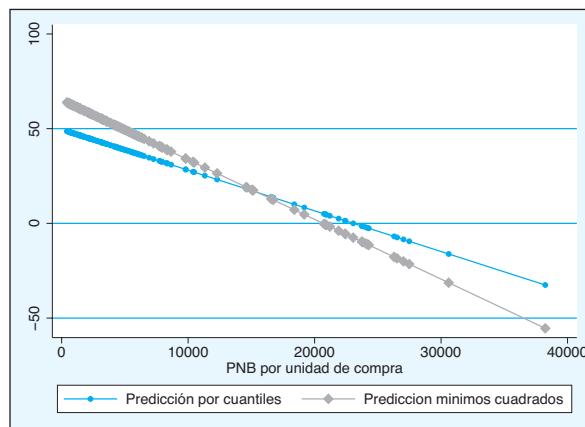
```

use mundo99, clear
reg tmi pnbppa
predict ttmi
label var ttmi "Predicción mínimos cuadrados"
qreg tmi pnbppa
predict qtmi
label var qtmi "Predicción por cuantiles"
scatter qtmi ttmi pnbppa, connect (.l) symbol (o.) sort(pnbppa) name(I43, replace)

```

Así se obtiene un gráfico donde los puntos representan la estimación de la mediana, mientras que la recta es la regresión lineal clásica. Como puede apreciarse, la primera pronostica valores más centrados en torno al promedio de la variable dependiente (está menos inclinada), al estar su predicción menos influida por los casos extremos.

**GRÁFICO 10.11. Gráfico comparativo de regresión de la mediana frente a la regresión clásica**



## 10.6. Regresión por bandas

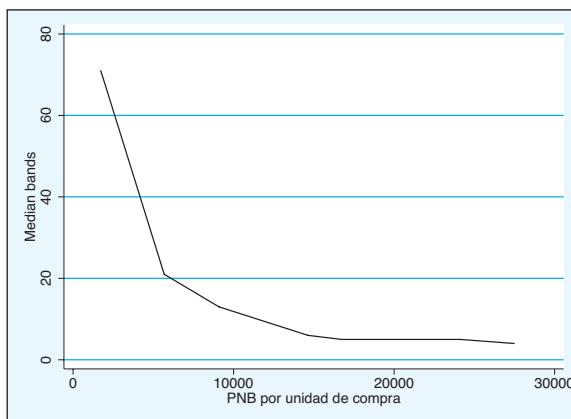
Otra posibilidad para efectuar un ajuste a los datos que permite más flexibilidad, por un lado, y también explorar la posible relación funcional entre dos variables es la regresión por bandas. En realidad, no se trata de ningún procedimiento que calcule parámetros, sino una técnica exploratoria y gráfica para representar de modo fidedigno la relación entre un par de variables. De hecho, en Stata no se realiza con ninguna instrucción similar a la de las regresiones, sino con el comando *graphics*.

Para realizarla hay que especificar el tipo *mband* en la orden *graph twoway*, como se expuso en el capítulo de gráficos. Con la opción *bands* ha de expresarse el número de zonas que quieren ajustarse.

```
graph twoway mband tmi pnbppa, bands(10)
```

Con ello, el gráfico es dividido en tantas zonas como se especifique en *bands* y en cada una de ellas se ajusta la mediana tanto de *x* como de *y*, que convenientemente unidas, forman el patrón de relación entre ambas variables.

**GRÁFICO 10.12. Gráfico de regresión por bandas**



En este caso es fácil apreciar que la sucesión de líneas muestran una relación curvilineal entre la tasa de mortalidad infantil y el producto nacional bruto per cápita.

## 10.7. Ejercicios

1. En este capítulo los ejercicios son los mismos que en el capítulo anterior. De cada uno de ellos, habría que realizar lo siguiente:
  - a) Detectar cuáles son sus principales incumplimientos de los supuestos de la regresión.
  - b) Encontrar casos anómalos y tratar de explicar su excepcionalidad, tras lo cual volver a hacer regresiones sin ellos.
  - c) Ajustar otros modelos robustos a los datos, observado las diferencias en los coeficientes con los obtenidos por el método de mínimos cuadrados ordinarios.



# 11

## La regresión logística<sup>1</sup>

El modelo de regresión lineal es una técnica de gran potencia y versatilidad. Permite predecir el comportamiento de una variable dependiente en función de una o más variables independientes y estimar con precisión la capacidad explicativa del modelo (gracias al coeficiente de determinación), entre otras muchas ventajas. Pero tiene una restricción importante para las ciencias sociales: sólo se puede utilizar con variables dependientes puramente cuantitativas (de intervalo o de razón). En sociología, la mayor parte de las variables que se estudian son cualitativas o categóricas (nominales u ordinales), por lo que la posibilidad real de uso de la regresión lineal es bastante limitada. Para este tipo de variables se pueden utilizar las técnicas de regresión logística, basadas en el modelo lineal, pero adaptadas a variables categóricas. Aunque son algo más complejas de interpretar que el modelo lineal y algo menos precisas en algunos aspectos, permiten realizar un análisis de variables categóricas equivalente al del modelo lineal. La base de todas estas técnicas logísticas es el modelo de regresión logística para variable dependiente dicotómica (logit), que es el que se incluye en este capítulo.

### 11.1. El modelo estadístico

Hay dos maneras principales de justificar el modelo estadístico de la regresión logística: la primera se basa en la relación teórica entre la variable dependiente observada (dicotómica) y una variable dependiente inobservada o latente (continua); la segunda se basa en la transformación de la variable dependiente dicotómica en una función de probabilidad no lineal (Long y Freese 2006: 132-135).

---

<sup>1</sup> Para ampliar conocimientos de este capítulo y el próximo se recomienda especialmente el libro Long y Freese (2006), cuyos argumentos, propuestas y programas aquí se reflejan. También son útiles Borroah (2001), Aldrich y Nelson (1984), Hosmer y Lemeshow (2000) y Hilbe (2009). Muy básico, en castellano, en esta misma colección, Jovell (1995). También se ha publicado otro monográfico en castellano en la colección de Cuadernos de Estadística (Silva y Barroso 2004).

### 11.1.1. *El modelo de variable latente*

Cualquier variable observada dicotómica puede concebirse como una manifestación de otra variable latente continua. Una variable dicotómica observada indica la existencia o no de un determinado atributo, o a la ocurrencia o no de un determinado suceso. Uno puede imaginarse que tras tal atributo o suceso existe una propensión o una probabilidad de ocurrencia (no observada, ni necesariamente observable), que tiene carácter continuo y que, al superar un cierto umbral, determina la existencia del atributo u ocurrencia del suceso en cuestión. La regresión logística se puede entender como una modelización de la variable latente (no observada) en función de la relación observada entre la variable dicotómica observada y la variable o variables independientes introducidas en el modelo.

En el ejemplo de asistencia a manifestaciones que se empleará a lo largo de este apartado, la variable resultado o dependiente es dicotómica, al tomar sólo dos valores: 0 si el individuo no ha asistido a una manifestación y 1 si el individuo lo ha hecho en determinado periodo. Es conceivable que esta variable dependa de otra variable subyacente continua que se puede llamar propensión a manifestarse, por ejemplo. De hecho, es excesivamente restrictivo el resumir toda la información sobre la asistencia a manifestaciones en una variable dicotómica, puesto que se puede no haber asistido nunca pese a tener una actitud muy favorable hacia ellas, o se puede haber asistido sin demasiado entusiasmo, por ejemplo. La (inobservada) propensión a manifestarse podría tomar, por ejemplo, el valor mínimo en una persona que no se ha manifestado ni probablemente se manifestará por sus convicciones o valores; valores medios en aquellas personas que se han manifestado alguna vez, pero sin mucho entusiasmo, y alcanzaría el valor máximo en aquellos que no sólo se hayan manifestado anteriormente, sino que se manifiestan y volverán a hacerlo de manera asidua. La variable dicotómica se puede entender como una manifestación de esa variable latente continua, puesto que a partir de un determinado nivel de “propensión a manifestarse” (lo que puede denominarse un nivel *umbral*) lo más probable es que el individuo se haya manifestado (y lo contrario si la propensión del individuo está por debajo de ese nivel umbral)<sup>2</sup>.

El gráfico 11.1 muestra la relación hipotética entre la variable latente “propensión a manifestarse” y una variable independiente continua cualquiera (por ejemplo, la edad). El eje vertical izquierdo representa la variable latente o teórica (de ahí el asterisco tras la *y*), y el eje horizontal la variable independiente observada (la edad). La línea que cruza el gráfico representa la relación entre

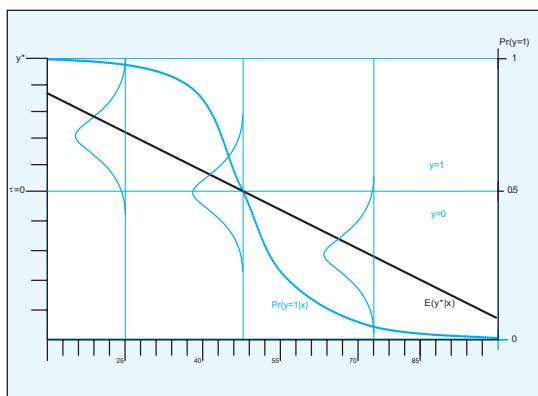
---

<sup>2</sup> Otro ejemplo, tal vez más claro, es el del acuerdo con una opinión política. Una persona puede estar de acuerdo o en desacuerdo con la frase “sin partidos no puede haber democracia” (si esta variable está codificada como variable dicotómica). Pero lo más probable es que las personas tengan opiniones más matizadas, existiendo un continuo de opiniones, desde el desacuerdo absoluto hasta el acuerdo total. De nuevo, existiría un nivel umbral a partir del cual el entrevistado respondería sí a la pregunta.

ambas variables, como una típica recta de regresión, sólo que en este caso es teórica también, puesto que muestra la relación entre la edad y la propensión a manifestarse,  $E(y^*|x)$ .

La cuestión estriba en cómo se representaría la variable observada dicotómica. Como se ha dicho antes, la probabilidad de haber asistido a una manifestación aumenta con la propensión a manifestarse, de modo que a partir de un *nivel umbral* de  $y^*$  lo más probable es que el individuo sí se haya manifestado (o sea, que si  $y^* > \text{umbral}$ ,  $y = 1$ ; si  $y^* \leq \text{umbral}$ ,  $y = 0$ ). El nivel umbral está representado gráficamente en el valor 0 de  $y^*$  (marcado con línea discontinua). La relación entre la variable latente continua y la dicotómica observada no es perfecta, sino estocástica, con un cierto nivel de error: por ello, la probabilidad de ocurrencia de  $y$  (dicotómica observada) asociada a cada nivel de  $y^*$  (continua latente) está representada por el área sombreada para tres valores ficticios de  $x$ . El valor de la variable latente (propensión a manifestarse) asociado a la edad de 25 está tan por encima del nivel de umbral que la probabilidad de haber asistido a una manifestación será prácticamente igual a 1 para los que tienen esta edad. En torno a los 40 años la línea de regresión de la variable latente se corta con el nivel umbral, por lo que a esta edad la indeterminación es máxima: la probabilidad de que un individuo haya asistido a una manifestación a esa edad es prácticamente la misma que la probabilidad de que no haya asistido. El área de probabilidad de ocurrencia del suceso está claramente por debajo del nivel de umbral para los que tienen 55 años (o más), por lo que la probabilidad de haber asistido a manifestaciones será prácticamente nula en esas edades. La probabilidad de ocurrencia del suceso para cada valor de  $x$  (de haber asistido a manifestaciones para cada edad), de acuerdo con este modelo, sería no lineal, muy semejante a la línea punteada de color gris que se muestra en el gráfico (cuyo eje está representado al lado derecho).

**GRÁFICO 11.1. Relación entre la variable latente y la variable dicotómica observada con una variable independiente**



Fuente: Reelaboración a partir de Long y Freese (2006: 133-134).

La regresión logística se puede entender, por tanto, como una modelización de la relación entre una variable dicotómica dependiente observada y una o más variables independientes, tal y como se muestra en el gráfico 11.1, asumiendo la existencia de una variable latente continua subyacente.

En términos formales, la relación entre la variable latente y las variables independientes del modelo (representada en el gráfico 11.1) es la siguiente:

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i \quad (11.1)$$

Donde la constante está representada por  $\beta_0$ , los coeficientes asociados a cada variable  $x$  por el resto de  $\beta$ , y el error aleatorio por  $\varepsilon_i$  (se trata de una ecuación estándar de regresión).

La relación entre la variable dependiente dicotómica observada y la variable latente puede formularse del siguiente modo:

$$\begin{aligned} y_i &= 1 \text{ si } y_i^* > 0 \\ y_i &= 0 \text{ si } y_i^* \leq 0 \end{aligned} \quad (11.2)$$

Por tanto, para un valor determinado de una sola  $x$ , la probabilidad de que la variable dicotómica tome un valor de 1 será la siguiente:

$$\Pr(y = 1|x) = \Pr(y^* > 0|x) \quad (11.3)$$

Sustituyendo de (11.1) y reorganizando los términos:

$$\Pr(y = 1|x) = \Pr(\varepsilon > -[\beta_0 + \beta_1 x]|x) \quad (11.4)$$

Lo que viene a demostrar que la probabilidad de ocurrencia de  $y$  depende no sólo de su relación con las variables independientes del modelo, sino también de la distribución del error de la variable latente  $\varepsilon$  (representado en la 1 por las áreas sombreadas).

La distribución del error de la variable latente no es conocida, y por tanto se ha de recurrir a alguna distribución teórica para despejar la ecuación y poder calcular el modelo de regresión no lineal para variable dependiente dicotómica. El modelo probit es el que deriva de asumir que  $\varepsilon$  se distribuye normalmente (con media 0 y varianza de 1), mientras que el modelo logit deriva de asumir que  $\varepsilon$  se distribuye de manera logística (con media 0 y varianza  $\pi^2/3$ ). La fórmula de (11.4) para el modelo probit, por tanto, es:

$$\Pr(y = 1|x) = \int_{-\infty}^{\beta_0 + \beta_1 x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt \quad (11.5)$$

Y la del modelo logit, mucho más sencilla, es:

$$\Pr(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad (11.6)$$

Como se puede inferir del gráfico 11.1 y de las ecuaciones (11.5) y (11.6), en función de la distribución que se asigne al error de la variable latente, la curva de probabilidad estimada por el modelo será ligeramente diferente. Si se asumen errores normalmente distribuidos (modelo probit), la curva tenderá a aproximarse más rápidamente a los ejes que si se consideran los errores con una distribución logística (modelo logit). Por tanto, los coeficientes variarán ligeramente. Pero, en la práctica, los resultados sustantivos serán muy similares (pese a que los coeficientes no son directamente comparables), por lo que utilizar un modelo u otro depende más de las preferencias del investigador que de ninguna otra cosa. En este libro se emplea únicamente el modelo logit, que es el que más se utiliza en las ciencias sociales (sobre todo en sociología), probablemente por su mayor facilidad de interpretación al poder expresarse en función de cocientes de razones (*odds ratio*), como se explicará más adelante.

### *11.1.2. El modelo de probabilidad no lineal*

Una justificación más sencilla, que no requiere recurrir a la existencia de variables latentes subyacentes a la variable dicotómica observada, es la que deriva de una simple transformación del modelo de regresión para variable dependiente dicotómica en un modelo de probabilidad no lineal, utilizando el concepto de cociente de razones. Esta justificación permitirá también una primera aproximación a la interpretación de los resultados del modelo de regresión logística.

Una variable dicotómica sólo puede tomar dos valores, 1 o 0. Si se utiliza el modelo de regresión lineal estándar con una variable dicotómica como variable dependiente, los valores predichos de la variable dependiente pueden ser mayores que 1 o menores que 0, en función de su relación con las variables independientes, lo que obviamente no tiene ningún sentido. El modelo de probabilidad lineal sería el siguiente:

$$\Pr(y = 1|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \varepsilon \quad (11.7)$$

¿Cómo puede modificarse el lado izquierdo de la ecuación para constreñir su rango de 0 a 1? Una manera de hacerlo es convertir las probabilida-

des en *razones*<sup>3</sup>, dividiendo la probabilidad de ocurrencia del suceso por la probabilidad de no ocurrencia.

$$\Omega(\mathbf{x}) = \frac{\Pr(y = 1|\mathbf{x})}{\Pr(y = 0|\mathbf{x})} = \frac{\Pr(y = 1|\mathbf{x})}{1 - \Pr(y = 1|\mathbf{x})} \quad (11.8)$$

Por ejemplo, la razón de haber participado en alguna manifestación para el total de los encuestados es igual a la probabilidad de haber participado dividida por la probabilidad de no haber participado, es decir,  $0,37 / 0,63 = 0,58$ . Las razones indican la relación (o proporción) entre la probabilidad de ocurrencia del suceso y la probabilidad de no ocurrencia. En este caso, la probabilidad de haber participado en alguna manifestación representa un 58% de la probabilidad de no haber participado. También podría haberse calculado la razón de no haber participado frente a la de haber participado:

$$\Omega(\mathbf{x}) = \frac{\Pr(y = 0|\mathbf{x})}{\Pr(y = 1|\mathbf{x})} = \frac{\Pr(y = 0|\mathbf{x})}{1 - \Pr(y = 1|\mathbf{x})} \quad (11.9)$$

En cuyo caso el resultado hubiese sido  $0,63 / 0,37 = 1,7$ . O sea, es 1,7 veces más probable no haber participado nunca que haber participado en alguna manifestación para un caso seleccionado al azar en la muestra.

La razón varía de 0, cuando la probabilidad de ocurrencia del suceso es 0 y la de no ocurrencia 1, a  $+\infty$ , cuando la probabilidad de ocurrencia del suceso es 1 y la de no ocurrencia es 0. Para conseguir que varíe de  $-\infty$  a  $+\infty$ , se utiliza el logaritmo neperiano de la razón. Cuando la razón es menor que 1, su logaritmo es negativo, y cuando es mayor que 1, es positivo. Al logaritmo neperiano de la razón se le denomina logit y es lo se utiliza como variable dependiente en la ecuación de la regresión logística:

$$\ln \frac{\Pr(y = 1|\mathbf{x})}{1 - \Pr(y = 1|\mathbf{x})} = \ln \Omega(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} \quad (11.10)$$

De modo que el modelo de regresión logística es equivalente al modelo de regresión lineal con la diferencia de que transforma la variable dependiente en el logaritmo de su razón, para conseguir así que varíe de

<sup>3</sup> El término inglés *odds* no tiene una traducción clara al español. En los países anglosajones, *odds* (en plural) es un término muy utilizado en el mundo de las apuestas, y se suele traducir como puntos de ventaja o simplemente *tanto contra tanto*, expresando las probabilidades a favor y en contra de una apuesta, que determinarán las ganancias relacionadas con tal apuesta. Algunos autores que han escrito sobre logit en castellano proponen el término castellanizado *ods* (Jovell 1995: 24). Sin embargo, aquí se empleará el término *razón* para denominarlo, mientras que *ratio* será traducido por *cociente*.

$-\infty$  a  $+\infty$ , y sobre ese valor estima la ecuación de la regresión. Esto es lo que hace la interpretación de la regresión logística bastante más complicada que la de la regresión lineal, puesto que los coeficientes del modelo de regresión logística no expresan de manera directa (como sí ocurre en el caso de la regresión lineal) la relación entre la variable independiente y la dependiente, sino la relación entre la variable independiente y *el logaritmo de la razón de la ocurrencia de un determinado suceso*. Por tanto, no pueden interpretarse los coeficientes directamente sobre el modelo de regresión logística estimado. Es necesario transformar la ecuación logística para que exprese los coeficientes de un modo interpretable. Hay dos formas de hacer esto: la primera es eliminando los logaritmos la ecuación logística original de tal modo que la ecuación se exprese en razones (en lugar de en sus logaritmos), y por tanto los coeficientes expresen la variación que las variables independientes producen en la razón de ocurrencia de un determinado suceso o característica. La segunda, algo más compleja de efectuar, pero más fácil de interpretar, es transformando la ecuación para que exprese directamente las probabilidades de ocurrencia del suceso estudiado.

De acuerdo con el primer procedimiento, la ecuación que expresa la variable dependiente en razones sería:

$$\Omega(\mathbf{x}) = \frac{\Pr(y=1|\mathbf{x})}{1-\Pr(y=1|\mathbf{x})} = \exp(\mathbf{x}\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1x_1 + \dots + \beta_kx_k) = \omega_0\omega_1^{x_1}\dots\omega_k^{x_k}$$

siendo  $\omega_k = \exp(\beta_k)$

(11.11)

Lo que no es más que la misma ecuación (11.10), en la que se ha despejado el logaritmo del lado izquierdo de la ecuación para que este exprese sólo la razón. Por tanto, los coeficientes  $\omega_i$  indican cómo varía la razón de la variable dependiente cuando la variable independiente varía en una unidad. Este coeficiente de la ecuación logística, expresado en razones, se denomina cociente de razones, y tiene una interpretación diferente al coeficiente de una ecuación de regresión normal. En una regresión normal, el coeficiente indica en qué medida aumenta el valor de la variable dependiente cuando aumenta en uno la independiente; en una regresión logística expresada en forma de razones, el coeficiente expresa en qué medida se multiplica la razón de la variable dependiente cuando la independiente aumenta en uno. Es decir, el cociente de razones mide el efecto en términos de tasa de cambio, no en cuántas unidades aumenta o disminuye la dependiente. Un cociente de razones superior a 1 indica que el efecto de la variable independiente en cuestión es positivo (aumenta la razón de ocurrencia del suceso estudiado), un cociente de razones inferior a 1 indica un efecto negativo (reduce la razón) y un cociente de razones de 1 indica ausencia de efecto. De este modo, si en este ejemplo (estudiando la participación en manifestaciones),

el coeficiente cociente de razones asociado a la variable género (hombre=0; mujer=1) fuera de 0,5, se diría que el hecho de ser mujer reduce a la mitad la posibilidad de haber participado en alguna manifestación, con respecto a la posibilidad de haber participado siendo hombre, esto es, disminuye a la mitad la razón de la participación. Todo esto se verá con más detalle en el apartado de análisis, con un ejemplo más concreto.

También puede despejarse la ecuación aún más para que exprese la variable dependiente en probabilidad de ocurrencia del suceso  $y=1$ :

$$\Pr(y = 1|\mathbf{x}) = \frac{\exp(\mathbf{x}\beta)}{1 + \exp(\mathbf{x}\beta)} = \frac{\omega_0\omega_1^{x_1} \dots \omega_k^{x_k}}{1 + \omega_0\omega_1^{x_1} \dots \omega_k^{x_k}} \quad (11.12)$$

Ecuación que es idéntica a la (11.6) (a la que se llegó con el modelo de variable latente), y que expresa la relación entre una variable dicotómica  $y$ , expresada como probabilidad de ocurrencia del suceso  $y=1$ , y una o más variables independientes. El resultado del modelo no se saldrá del rango 0-1, y la línea de regresión de probabilidad predicha tendrá una forma suavizada, de  $s$ , de tal modo que cuando la línea se aproxime a 0 o a 1 incrementos grandes en las variables independientes se corresponderán con incrementos cada vez menores en la probabilidad de la dependiente (como quedó mostrado anteriormente en el gráfico 11.1).

## 11.2. Estimación del modelo

Una vez contemplado el fundamento matemático del modelo de regresión logística, se examina a continuación mediante un ejemplo la estimación con Stata y la interpretación de los parámetros y coeficientes esenciales de este procedimiento estadístico. El propósito del ejemplo ya comenzado es estudiar qué tipo de persona es más frecuente o probable que haya asistido a manifestaciones alguna vez en su vida, con variable dependiente dicotómica (1: ha asistido alguna vez a una manifestación; 0: no ha asistido nunca). Como variables independientes, se emplearán distintas variables sociodemográficas tanto cuantitativas como cualitativas<sup>4</sup> (ingresos, edad, estudios, situación laboral, género y tamaño de hábitat). Lo que se persigue es construir un modelo que explique la mayor cantidad posible de variabilidad de la variable dependiente con el menor número posible de variables independientes (es decir, el modelo más parsimonioso).

---

<sup>4</sup> Las variables cualitativas que se incluyan como variables independientes en la regresión deben estar codificadas como ficticias, lo que ya se ha explicado en la sección 9.6.

La estimación del modelo de regresión logística se realiza por el método de máxima verosimilitud. Este método estima los valores de los parámetros  $b$  de la regresión que con mayor probabilidad pueden haber generado los valores de la variable dependiente de la muestra, si las asunciones del modelo son ciertas<sup>5</sup>. Se calcula una función de verosimilitud que indica cuál es la probabilidad de que para unos determinados parámetros  $b$  se hayan observado los valores muestrales. En un proceso iterativo se van probando distintos valores de los parámetros  $b$  hasta que se encuentran los coeficientes que maximizan tal función de verosimilitud<sup>6</sup> (o sea, los coeficientes que más verosímilmente corresponden a los valores muestrales): tales coeficientes serán los estimados para un determinado modelo.

Este proceso iterativo se muestra en la salida de la instrucción *logit*, cuyo formato general es el siguiente:

```
logit variable_dependiente lista_de_variables_independientes
```

En consecuencia, en el primer ejemplo mostrado la instrucción concreta es como sigue:

```
logit manif mujer edad i.estudios i.ingresos
```

Las variables que se han especificado como independientes son: *sexo* como dicotómica (mujer, 1 si es mujer, 0 si no), *edad* (continua), *estudios* (considerada como factor con tres valores: primarios, que actúa como categoría base, secundarios [2] y superiores [3]), e *ingresos* (también convertida en factor con tres categorías: menos de 150.000 pesetas al mes [categoría base], de 150.000 a 300.000 pesetas y más de 300.000). Entre todas las variables sociodemográficas se han incluido estas cuatro porque se sospecha que son las que mejor explican la frecuencia relativa de que una persona haya asistido alguna vez a una manifestación.

Lo primero que aparece en la ilustración 11.1 es el proceso iterativo de estimación del modelo a través del método de máxima verosimilitud. Stata muestra los valores sucesivos de la función de verosimilitud para los distintos parámetros que va estimando. En la iteración 0, todos los coeficientes valen 0 menos la constante, y en iteraciones sucesivas se van aproximando valores de los coeficientes que incrementan el valor de la función de ve-

<sup>5</sup> Las asunciones son las habituales de los modelos de regresión: que no falten variables importantes en el modelo, que no haya multicolinealidad entre las variables independientes, etc. Véase el segundo capítulo dedicado a la regresión.

<sup>6</sup> De hecho, no se maximiza directamente la función de verosimilitud, sino su logaritmo (*log likelihood*), lo que simplifica la computación. Este *log likelihood* es el que aparece en la salida de Stata de *logit*.

rosimilitud. Como puede apreciarse, en cada iteración el logaritmo de la verosimilitud es mayor (menos negativo en este caso), aunque dado que en cada iteración se aproxima más la función a su máximo, cada iteración añade menos valor. Cuando el mecanismo iterativo considera que ya no es necesario seguir refinando la estimación, porque ya las iteraciones no añaden prácticamente nada de verosimilitud al modelo, se detiene el proceso y muestra los coeficientes estimados. Como se puede ver, la diferencia entre el *log likelihood* de la iteración 3 y el de la iteración 2 es mínima, por lo que ya no es necesario seguir refinando la estimación. En tres iteraciones se han conseguido estimar los coeficientes que más verosímilmente pueden haber producido los valores observados de la variable dependiente.

### ILUSTRACIÓN 11.1. Regresión logística

Iteration 0:	log likelihood = -2459.6045
Iteration 1:	log likelihood = -2168.0953
Iteration 2:	log likelihood = -2164.7366
Iteration 3:	log likelihood = -2164.7337
Iteration 4:	log likelihood = -2164.7337
 Logistic regression	
	Number of obs = 3717
	LR chi2(6) = 589.74
	Prob > chi2 = 0.0000
	Pseudo R2 = 0.1199
 <hr/>	
manif	Coef. Std. Err. z P> z  [95% Conf. Interval]
<hr/>	
mujer	-.400606 .0738638 -5.42 0.000 -.5453763 -.2558357
edad	-.0145314 .0023736 -6.12 0.000 -.0191836 -.0098792
estudios	
2	.6236568 .0934946 6.67 0.000 .4404109 .8069028
3	1.42194 .1148799 12.38 0.000 1.19678 1.647101
ingresos	
2	.5415312 .0833178 6.50 0.000 .3782313 .7048311
3	.618273 .1250371 4.94 0.000 .3732047 .8633413
_cons	-.3795612 .1479147 -2.57 0.010 -.6694687 -.0896536
<hr/>	

Estos coeficientes estimados aparecen a continuación. Los resultados que proporciona Stata para el logit son similares a los de la regresión. Arriba a la derecha aparece el número de observaciones y una prueba estadística de significación del modelo basada en el  $\chi^2$ . Con un nivel de confianza del 95%, el modelo es significativo si la probabilidad que aparece es inferior a 0,05. En este caso, puede decirse que la relación entre los coeficientes del modelo y la probabilidad de haber participado alguna vez en una manifestación es significativa estadísticamente. Por último, aparece en esta columna de estadísticos el pseudo  $R^2$ . Como su propio nombre indica, es un estadístico análogo al  $R^2$ , que indica la bondad de ajuste del modelo a los datos. Aunque no tiene la inmediatez de interpretación del  $R^2$  de la regresión lineal, que directamente indica qué proporción de la varianza de la variable dependiente es explicado por el modelo, es una aproximación basada en

una comparación de la verosimilitud del modelo sólo con la constante  $\hat{L}_0$ , con la verosimilitud del modelo con todos los parámetros estimados  $\hat{L}_F$ :

$$Pseudo R^2 = 1 - \frac{\ln \hat{L}_F}{\ln \hat{L}_0} \quad (11.13)$$

Siendo  $\hat{L}_F$  la razón de verosimilitud del modelo completo (*full*) o final del que se desea estimar la bondad, y  $\hat{L}_0$  la del modelo que sólo posee la constante.

Aunque no sea tan preciso como el  $R^2$  de la regresión lineal, el *pseudo-R<sup>2</sup>* es una medida útil del ajuste del modelo a los datos, y puede servir para comparar la capacidad explicativa de modelos distintos.

Debajo de estos estadísticos aparecen los coeficientes, del mismo modo que en la regresión lineal. La primera columna muestra los coeficientes para cada variable de la regresión logística. Según estos coeficientes, la ecuación estimada sería:

$$\ln \frac{\Pr(y=1)}{\Pr(y \neq 1)} = -0,37 - 0,40mujer - 0,01edad + 0,63est2 + 1,42est3 + 0,54ingr2 + 0,61ingr3 \quad (11.14)$$

Dicho con palabras, el logaritmo de la razón de haber participado en alguna manifestación es igual a -0,37 más 0,63 si se tienen estudios secundarios, menos 0,01 por cada año de edad, etc. A la vista de esta ecuación, resulta evidente que, a diferencia de la regresión lineal, en el modelo *logit* no se pueden interpretar directamente los coeficientes. Saber que el ser mujer disminuye en 0,4 el logaritmo de la razón de haber participado alguna vez en una manifestación no sirve para mucho. El logaritmo de la razón es una medida de probabilidad ininteligible, difícilísima de interpretar tal cual. Es necesario transformar la ecuación *logit* original en una ecuación más fácilmente interpretable, que muestre la relación entre las variables independientes y la dependiente del modelo de manera más comprensible. Existen varias estrategias diferentes que permiten interpretar el modelo *logit* más fácilmente. Pero esto se verá más adelante, en el apartado de interpretación del modelo. Por el momento, ha de entenderse cómo se estima la ecuación *logit* y saberse qué son los estadísticos que se generan con esta instrucción de Stata.

El resto de las columnas de coeficientes son exactamente iguales que las que aparecen en la salida de Stata para una regresión lineal, y la interpretación también es esencialmente la misma. Aparece el error típico de cada coeficiente, su valor  $z$  y la probabilidad asociada a ese valor  $z$  (que indica si el coeficiente es estadísticamente significativo), y los intervalos de confianza de cada coeficiente. No merece la pena extenderse mucho más aquí, puesto que, como ya ha quedado dicho, en el modelo *logit* la interpretación no se hace sobre los coeficientes como en la regresión lineal, sino sobre las predicciones o los cocientes de razones, como se verá más adelante.

Mediante la orden *logit* se muestran dos pruebas de significación diferentes: por un lado, el test de chi<sup>2</sup>, que indica la significación del modelo completo (o sea, hasta qué punto la relación existente entre la variable dependiente y el conjunto de variables independientes es significativa); por otro lado, aparece el test de z para cada coeficiente, indicando hasta qué punto cada coeficiente tiene un efecto significativo en la ecuación. Adicionalmente se pueden realizar otras pruebas de hipótesis sobre los coeficientes de la ecuación: puede interesar, por ejemplo, saber si un coeficiente es igual a un determinado valor, o si dos coeficientes tienen el mismo efecto sobre la variable dependiente, o si el efecto de dos coeficientes diferentes es simultáneamente igual a 0. Hay dos instrucciones que permiten realizar pruebas de hipótesis sobre los coeficientes de un modelo logit tras la estimación: *test* (que realiza el test de Wald) y *lrtest* (que realiza un test de coeficiente de verosimilitud o *likelihood ratio*). Estadísticamente no hay muchos argumentos para preferir uno u otro: ambos son asintóticamente equivalentes (sólo difieren en muestras pequeñas, cuanto mayor es la muestra, los resultados que muestran son más cercanos, hasta llegar a ser prácticamente idénticos en muestras grandes). Por ello, sólo se explica a continuación la prueba de Wald (mediante la instrucción *test*) porque resulta más sencilla de usar.

Para realizar una prueba de hipótesis con la instrucción *test*, basta con escribir a continuación la expresión que se quiera comprobar referida a los coeficientes de un modelo logit. El test se referirá al último modelo logit estimado (hay que tener cuidado con esto, puesto que la instrucción *test* no indica a qué modelo se refiere, lo que puede inducir a error). Con unos ejemplos quedará más claro el uso de esta instrucción, que, por otro lado, es similar al correspondiente a la regresión común.

Si se desea comprobar si un coeficiente es significativo (significativamente distinto de 0), simplemente se escribe *test* seguido por el nombre de coeficiente. En el caso de la variable *edad*, la instrucción sería la siguiente:

```
test edad
```

Y el resultado de aplicarla se presenta a continuación:

#### **ILUSTRACIÓN 11.2. Prueba de hipótesis de un parámetro de la regresión logística**

```
( 1) [manif]edad = 0
      chi2( 1) =   37.48
      Prob > chi2 =  0.0000
```

Como se puede observar, en este caso Stata realiza la prueba de Wald comprobando si el coeficiente *edad* tiene un valor significativamente dis-

tinto de 0, o sea, si la edad tiene realmente efecto sobre el hecho de haber participado en alguna manifestación. El test da un  $\chi^2$  de 37,5, que para un grado de libertad (puesto que sólo se comprueba un coeficiente) tiene una probabilidad de ocurrencia menor que 0,0000. Por tanto, el coeficiente de edad es significativamente distinto de 0, puesto que la probabilidad de que el coeficiente fuera 0 en la población es mínima en el modelo.

Este uso de la instrucción *test* no es el más interesante, porque la salida normal del logit ya muestra la significación de cada uno de los coeficientes por separado (con los valores *z*). También muestra la significación de todos los coeficientes juntos (o sea, la del modelo). Pero lo que no muestra es la significación de todas las combinaciones posibles de los coeficientes: para esto sí que será imprescindible usar la orden *test*. Por ejemplo, se puede estar interesado en ver si los coeficientes *edad* y *mujer* son simultáneamente iguales a 0 (si las dos variables no tienen efecto conjunto importante sobre la participación en manifestaciones). Para ello, habría que escribir seguidos los nombres de ambas variables:

```
test edad mujer
```

En cuyo caso, el resultado se presenta de este modo:

### **ILUSTRACIÓN 11.3. Prueba de hipótesis de más de un parámetro de la regresión logística**

```
( 1) [manif]edad = 0
( 2) [manif]mujer = 0
      chi2( 2) =    65.09
      Prob > chi2 =    0.0000
```

Y, como cabría esperar, después de que fuera significativo el coeficiente aislado de la variable *edad*, se demuestra que en el modelo el valor de los coeficientes *edad* y *mujer* no son simultáneamente iguales a 0.

Además de comprobar el valor efectivo de un coeficiente, mediante la expresión variable = valor, se puede querer saber si el efecto de dos coeficientes es el mismo. Por ejemplo, puede interesar saber si los ingresos medios (2.ingresos) y altos (3.ingresos)<sup>7</sup> tienen la misma frecuencia relativa de asistencia a manifestaciones:

```
test 2.ingresos=3.ingresos
```

---

<sup>7</sup> Cualquier variable factorizable, es decir, con valores enteros, puede ser empleada como variable ficticia mediante la expresión #.nombrevar en muchas instrucciones de Stata. Excepciones, por ejemplo, a esta regla se encuentran en la órdenes *mean* y *tabulate*.

#### ILUSTRACIÓN 11.4. Prueba de hipótesis de igualdad de parámetros en la regresión logística

( 1 )	[manif]2.ingresos - [manif]3.ingresos = 0
	chi2( 1) = 0.42
	Prob > chi2 = 0.5173

En este caso, no puede rechazarse la hipótesis nula de que el efecto de tener ingresos medios sea igual que el efecto de tenerlos altos (o, como aparece reflejado, que 2.ingresos menos 3.ingresos es igual a 0): la diferencia entre los valores de ambos coeficientes no es significativamente distinta de 0 (es demasiado alta la probabilidad de que en la población el valor real de la diferencia entre el coeficiente del valor 2 en ingresos y el del valor 3 de la misma variable sea igual a 0).

### 11.3. Diagnóstico del modelo

Igual que en la regresión lineal, tras la estimación del modelo y antes de su interpretación es necesario estudiar el grado de ajuste de la regresión logística a los datos. Este paso es importante para: primero, detectar posibles problemas en el modelo, sean debidos a datos incorrectos, a una mala especificación de las variables o a cualquier otra causa, y segundo, valorar su capacidad explicativa. Este paso es, por tanto, importante, aunque a menudo se obvie y se pase directamente a la interpretación tras la estimación del modelo de regresión logística.

¿Qué significa *ajuste* del modelo? Una regresión lo que trata es de predecir el valor que toma la variable dependiente en función de la o las variables independientes. A menos que la predicción sea perfecta (lo que nunca es posible encontrar, especialmente en el ámbito de las ciencias sociales), habrá un determinado margen de error. Pues bien, a ese error es a lo que se refiere el concepto de ajuste del modelo. Cuanto mayor sea el error de predicción (menos acierto la ecuación de regresión el valor que realmente toma la variable dependiente), menor será el ajuste del modelo, y viceversa.

Véase en términos analíticos. Aunque todo modelo de regresión (incluido el logit), como muestra la ecuación (11.15), trata de expresar el valor de la probabilidad de que  $y=1$  en función de los valores que tomen las  $x$ , también hay que tener en cuenta que siempre habrá un cierto margen de error en esta predicción:

$$\ln \Omega = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon \quad (11.15)$$

Este término también puede expresarse como diferencia entre los valores observados de  $y$  y los valores esperados según la ecuación de regresión:

$$\varepsilon = \ln \Omega - \ln \Omega(\mathbf{x}) = \ln \Omega - (\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) \quad (11.16)$$

En esto se basa casi todo el análisis del ajuste del modelo. Se trata simplemente de ver en qué medida este es capaz de predecir los valores que toma en realidad la variable dependiente en función de los valores que toman las independientes. Aunque la idea básica es siempre la misma, hay muy distintos modos de estudiar el ajuste del modelo. Aquí se estudiarán dos: el estudio gráfico de residuales e influencia, y el estudio de medidas escalares de ajuste utilizando la instrucción *fitstat*.

El residuo es exactamente lo que aparece en la ecuación (11.16): es la diferencia entre el valor observado y el esperado en función de la ecuación de regresión. Puede expresarse en términos de logaritmos, como aparece en la ecuación mencionada, o en términos de probabilidades, como se utilizará a continuación para simplificación de las fórmulas. En un modelo de regresión lineal se pueden estudiar los residuos tal cual, pero en el modelo logit es necesario estandarizarlos, para su mejor interpretación. En la regresión logística, las predicciones son frecuencias o probabilidades (que van de 0 a 1) y los valores observados individuales son siempre 0 o 1. Sin embargo, siguiendo a Hosmer y Lemeshow (2000), para el cálculo de los residuos, Stata no utiliza estos valores individuales, sino la esperanza para cada determinada combinación de valores de  $x$ , es decir,  $\Pr(y=1|\mathbf{x})$ . Cuanto más cercana sea esta probabilidad a 0,5 la probabilidad predicha por el modelo, más alta será la varianza de la diferencia entre tal probabilidad y el valor observado, puesto que la fórmula de su varianza es:

$$Var(\Pr(y = 1|\mathbf{x}) - \widehat{\Pr}(y = 1|\mathbf{x})) = \frac{\widehat{\Pr}(y = 1|\mathbf{x})(1 - \widehat{\Pr}(y = 1|\mathbf{x}))}{n} \quad (11.17)$$

A partir de esta estimación de la varianza, se puede obtener el residuo de Pearson, dividiendo la diferencia entre la probabilidad real y la esperada, por su error típico, es decir, por la raíz cuadrada de (11.17).

$$r_i = \frac{(\Pr(y = 1|\mathbf{x}) - \widehat{\Pr}(y = 1|\mathbf{x}))\sqrt{n}}{\sqrt{\widehat{\Pr}(y = 1|\mathbf{x})(1 - \widehat{\Pr}(y = 1|\mathbf{x}))}} \quad (11.18)$$

Y este es el problema: la varianza de los residuales de un modelo de regresión logística no es homogénea. O sea, que el error es heteroscedástico y la estimación de residuales es incorrecta. Otro problema del residuo de Pearson es que no posee una desviación típica igual a 1. Para obtenerla, es necesario utilizar el residuo tipificado propuesto por Pregibon (1981: 720), que resuelve este problema dividiendo, como en el caso de la regresión, el anterior por  $\sqrt{1 - h_i}$ , siendo  $h_i$  la carga de las variables independientes.

$$r_i^s = \frac{r_i}{\sqrt{1 - h_i}} \quad (11.19)$$

Para obtener los residuales logísticos con Stata, de la misma manera que en la regresión, es necesario crear una nueva que almacene la diferencia para cada observación entre el valor observado de la variable dependiente y el valor predicho por el modelo. En este caso, la variable dependiente es *asistencia a manifestaciones*: los valores que toma son 1 (si ha asistido a una manifestación alguna vez) o 0 (si no ha asistido nunca). El modelo logit, por su parte, predice para cada combinación de variables independientes cuál es la probabilidad de que haya asistido a alguna manifestación en función del nivel de estudios, la edad, los ingresos y el género. Por tanto, el residual en este caso será la diferencia entre el valor observado (lógicamente entre 0 y 1) y la probabilidad predicha (también en este intervalo). Por las razones anteriormente explicadas, es preferible utilizar el residual tipificado de Pearson (11.18). Como en la regresión, con Stata, tras una estimación logística, se puede generar esta variable de residuales de Pearson estandarizados mediante la opción *rstandard* de la instrucción *predict*:

```
predict resmanif, rs
label var resmanif "Residuos estandarizados"
list manif resmanif in 1/10 if resmanif<.
```

La variable residual almacena ahora la diferencia entre la probabilidad predicha de haber asistido y el hecho de haber asistido o no (según la fórmula vista en [11.19]). Utilizando la instrucción *list* pueden examinarse los primeros casos, para obtener una idea del contenido de la variable residual:

#### ILUSTRACIÓN 11.5. Listado de residuos estandarizados

	manif	resmanif
1.	1	-.3140299
4.	1	1.006958
7.	1	1.685212

Como puede verse, la variable residual es una variable estandarizada, por lo que la mayor parte de los casos están entre -2 y +2, y obviamente cuanto mayor es su valor, es porque el caso observado está más alejado de lo predicho por el modelo.

Pero lo que interesa no es estudiar los residuales de cada observación, sino cómo se ajustan los datos observados al modelo logit generado. Para

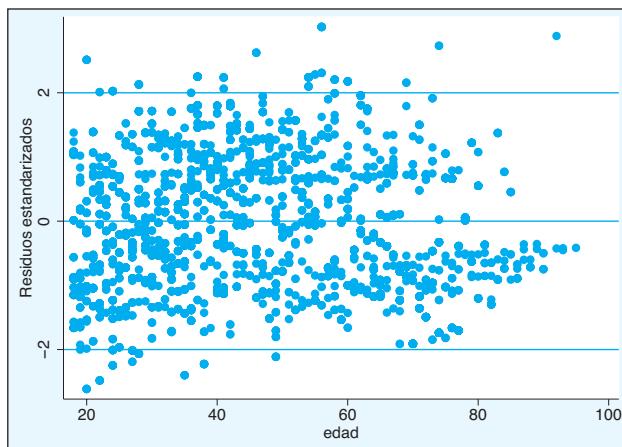
ello, es mejor representar los residuales gráficamente. Esto tiene dos ventajas. Primero, permite hacerse una idea general del ajuste del modelo a los datos, al mostrar todos los datos en un gráfico, y detectar fácilmente casos extremos para su inspección más detallada. Y, segundo, dado que se pueden representar los residuales junto con cualquier otra variable del modelo (en un gráfico de dos dimensiones de Stata), pueden detectarse sesgos o deficiencias en el ajuste. Todo ello puede verse mejor con el ejemplo utilizado.

Se van a mostrar en un gráfico los residuos en función de la variable *edad*. Es preciso fijarse en dos cosas: primero, en la existencia de casos extremos, que se salgan muy marcadamente de las predicciones del modelo; segundo, en la existencia de relación entre la edad y los residuales. En el gráfico no se debe encontrar relación alguna entre la variable *edad* y la variable residual: si hay indicios de tal relación, es que el modelo está mal especificado (falta alguna variable que está relacionada con la edad). Para hacer este gráfico, puede utilizarse la orden *scatter* de Stata (véase el capítulo de gráficos):

```
scatter resmanif edad, ylabel(-2 0 2) name(G2, replace)
```

Y aparece el siguiente gráfico:

**GRÁFICO 11.2. Gráfico de residuos estandarizados sobre una variable independiente**



Como puede apreciarse, la mayor parte de los casos está entre -2 y +2, y se distribuyen de manera bastante uniforme por arriba y por debajo del 0. Hay menos casos conforme avanza la edad, pero eso no es una deficiencia del modelo, sino el simple hecho de que en la muestra hay menos individuos de edades avanzadas. Por tanto, al menos en lo que respecta a la edad,

parece que el modelo está bien especificado. Puede hacerse lo mismo con las otras variables, incluso con variables no incluidas en la ecuación logística, para ver si se detecta alguna relación con los residuales. Si fuese así, se debería replantear el modelo y ver si es necesario introducir alguna variable o efecto nuevo.

En el gráfico de residuales por edad aparecen algunos casos extremos y extraños. Especialmente llama la atención un caso de edad muy avanzada y muy mal ajuste (residual muy grande, superior a 2). Cuando existen casos muy extremos, resulta recomendable estudiarlos individualmente con cierto detalle. Pero para ello es necesario identificarlos. ¿De qué modo? Hay una manera de detectar casos extremos, no gráfica, que resulta más conveniente cuando se cuenta con muchos casos y la representación gráfica no permite identificar con claridad las observaciones particulares. Consiste en, una vez generada la variable residual, marcar con una nueva variable (llamada aquí *extremo*) las observaciones que tengan un valor absoluto muy alto en tal variable para estudiarlas aisladamente. Se puede hacer del siguiente modo:

```
recode resmanif (. -2.7/2.7=0) (else=1), into(extremo)
sort resmanif
list manif resmanif edad sexo ingresos estudios if extremo, clean
```

De este modo<sup>8</sup>, se han marcado como extremos diez casos, aquellos con residuo estandarizado superior a 2,7. Con la instrucción *list* o simplemente estudiando la matriz de datos directamente en el revisor (*browse*) pueden analizarse más detenidamente estos casos extremos para tratar de detectar posibles problemas en el modelo logit. Una ordenación previa de los residuos estandarizados también es muy útil para mejorar la comprensión de lo que ocurre en los casos anómalos o mal predichos por la regresión logística.

#### **ILUSTRACIÓN 11.6. Listado de los residuos estandarizados extremos**

resmanif	manif	sexo	edad	estudios	ingresos
2.73281	1	Mujer	74	Primarios y menos	De 150.0000 a 300.000
2.73281	1	Mujer	74	Primarios y menos	De 150.0000 a 300.000
2.883978	1	Mujer	92	Primarios y menos	Menos de 150.000
3.025724	1	Hombre	56	Primarios y menos	Menos de 150.000
3.025724	1	Hombre	56	Primarios y menos	Menos de 150.000
3.025724	0	Hombre	56	Primarios y menos	Menos de 150.000
3.025724	1	Hombre	56	Primarios y menos	Menos de 150.000
3.025724	1	Hombre	56	Primarios y menos	Menos de 150.000
3.025724	0	Hombre	56	Primarios y menos	Menos de 150.000
3.025724	1	Hombre	56	Primarios y menos	Menos de 150.000

<sup>8</sup> Curiosamente, Stata trata los valores perdidos de una variable lógica como “verdaderos”. Por eso en la recodificación para obtener la variable lógica *extremo*, los valores perdidos, representados con un punto, son recodificados al valor 0. Fíjese en cómo delante del -2,7 hay un punto aislado, que representa cualquier valor perdido.

El estudio de los residuos permite, por tanto, comprobar el nivel de ajuste de los datos observados a los predichos por el modelo. Pero los residuales sólo indican la existencia de casos extremos, no la influencia que estos casos tienen sobre el modelo de regresión logística. Y esto último puede ser aún más importante. Supóngase que hay un caso que ha sido mal codificado y tiene un valor absurdo en la variable dependiente (con respecto a sus valores en las variables independientes), por lo que se ajusta muy mal al modelo logístico. El problema principal que presenta este caso no es un mal ajuste (después de todo el que un caso ajuste mal no tiene demasiada importancia en sí), sino el hecho de que puede haber distorsionado la propia estimación del modelo generando unos coeficientes en la ecuación logística incorrectos. Si se recuerda, más arriba la estimación del modelo logístico se realizaba por el método de máxima verosimilitud: un proceso estadístico que estima los coeficientes de la ecuación logística que más probablemente generan los valores observados de la variable dependiente. Por tanto, si hay valores erróneos en la variable dependiente, la estimación de los coeficientes estará sesgada, no será correcta, puesto que se basa en los valores observados.

Cuando se estudian los residuos, por tanto, no sólo es importante detectar los casos que se ajustan mal al modelo, sino tratar de evaluar qué influencia tienen estos casos. Evidentemente, los casos que deben estudiarse con más detalle son aquellos que no se ajusten bien a la ecuación logística y además ejerzan una influencia importante.

¿Cómo puede apreciarse la influencia de los casos individuales sobre el modelo? El concepto clave para entender el estudio de la influencia en la regresión es el siguiente: la influencia de un caso se mide a través del cambio que produce en el modelo su eliminación del proceso de estimación. O sea, si al quitar un caso en concreto el modelo cambia mucho (los coeficientes, la constante, la significación), se dice que ese caso ejerce una gran influencia; si prácticamente no hay cambios en el modelo, la influencia de tal caso será pequeña. Puesto que resultaría muy complicado realizar una estimación nueva para valorar la influencia de cada caso (quitándolo del modelo y repitiendo la estimación), lo que en la práctica se utiliza es la aproximación de Pregibon (Long y Freese, 2006: 151) llamado en Stata *dbeta*<sup>9</sup>. Dado que esta medida es equivalente a la distancia de Cook de la regresión lineal, a veces se le llama también de ese modo.

El estudio de la influencia de los casos en Stata es igual que el de los residuales: se genera una variable nueva que almacena la información de la influencia de cada caso y se estudia tal variable. La orden para generar la variable de influencia también es la misma, *predict*, sólo que en este caso con la opción *dbeta*:

<sup>9</sup> La fórmula del *dbeta* de Pregibon es:  $c_i = \frac{r_i^2 h_i}{(1-h_i)^2}$ ,  $h_i$  fue definido en (10.16).

```

predict cook, dbeta
summarize cook, detail
count if cook>.10 & cook<
list manif cook cook>.10 & cook<.

```

Posteriormente, se puede realizar una distribución de sus frecuencias y, si se encontraran valores excesivamente altos, realizar un listado de aquellos con valores extremos. Como se puede apreciar, en la ilustración 11.7, no es este el caso en tanto que el máximo valor es 0,10. De todos modos, si se examinan las personas que contienen estos valores extremos de Pregibon son 34, todas ellas mujeres de 70 años con bajos estudios y bajos ingresos.

### ILUSTRACIÓN 11.7. Estadísticas y listado de casos influyentes

Pregibon's dbeta					
Percentiles		Smallest			
1%	2.53e-06	2.00e-09			
5%	.0000521	2.00e-09			
10%	.0002681	6.39e-08	Obs	3717	
25%	.0017877	6.39e-08	Sum of Wgt.	3717	
50%	.0066305		Mean	.013848	
		Largest	Std. Dev.	.0180839	
75%	.0194973	.1024203			
90%	.0361494	.1024203	Variance	.000327	
95%	.0524679	.1024203	Skewness	2.254821	
99%	.090478	.1024203	Kurtosis	9.039001	
cook >.10 & cook<. = 34					
67.	.1024203	0	Mujer	70	Primarios y menos
68.	.1024203	0	Mujer	70	Primarios y menos
...					
					ingresos
					Menos de 150.000

Un modo pormenorizado de estudiar el ajuste del modelo a los datos es a través del análisis de residuales e influencia, tal y como se acaba de explicar. Pero otra manera de estudiar lo mismo que se usa a menudo y también resulta muy útil, sobre todo a la hora de comparar el ajuste de dos modelos diferentes, es mediante el uso de medidas resumen del nivel de ajuste. En este caso, en lugar de estudiar el ajuste de cada una de las observaciones y detectar anomalías, se trata de resumir en un solo estadístico el grado de ajuste de un modelo a los datos. Existen muchas medidas diferentes de ajuste apropiadas para modelos de regresión logística: aquí se estudiarán las más importantes que aparecen en el programa *fitstat* del conjunto de utilidades para variables dependientes categóricas, *SPost*, de Long y Freese (2006)<sup>10</sup>.

<sup>10</sup> Se trata de un conjunto de utilidades *ado* que hay que descargar por Internet, pues no viene incorporada en la versión estándar de Stata. Para instalar el paquete, basta realizar una búsqueda con la orden *net search* seguida de la especificación *Spost*, e instalar la versión que más se aproxime a la de Stata que tenga instalada el usuario.

La sintaxis de la orden *fitstat* para estudiar el ajuste es extremadamente sencilla: basta con escribir la instrucción una vez ajustado el modelo en cuestión.

```
fitstat
```

Tras lo cual aparece el resultado de la ilustración 11.8.

### ILUSTRACIÓN 11.8. Medidas de ajuste de la regresión logística

Measures of Fit for logit of manif			
Log-Lik Intercept Only:	-2459.605	Log-Lik Full Model:	-2164.734
D(3710):	4329.467	LR(6):	589.742
McFadden's R2:	0.120	Prob > LR:	0.000
ML (Cox-Snell) R2:	0.147	McFadden's Adj R2:	0.117
McKelvey & Zavoina's R2:	0.191	Cragg-Uhler(Nagelkerke) R2:	0.200
Variance of y*:	4.068	Efron's R2:	0.150
Count R2:	0.687	Variance of error:	3.290
AIC:	1.169	Adj Count R2:	0.167
BIC:	-26169.226	AIC*n:	4343.467
BIC used by Stata:	4387.012	BIC':	-540.418
		AIC used by Stata:	4343.467

Como puede apreciarse, esta orden muestra una gran cantidad de diferentes estadísticos de diagnóstico de la regresión logística. No van a explicarse aquí todos los estadísticos que aparecen, pero sí los más importantes. En la primera fila aparecen los logaritmos de la verosimilitud ( $\hat{L}_p$ ) del modelo completo (*Log-Lik Full Model*) y de aquella ( $\hat{L}_o$ ) del modelo que sólo incluye la constante (*Log-Lik Intercept Only*). Estos son los resultados principales del proceso de estimación por máxima verosimilitud. Como se explicó más arriba, la función de verosimilitud se puede entender como la probabilidad de que los datos observados en la muestra hayan sido generados por unos determinados coeficientes. Por tanto, la verosimilitud del modelo sólo con la constante es una medida de la probabilidad de que los datos observados hayan sido generados por un modelo logístico en el que todos los coeficientes valen 0 o, lo que es lo mismo, un modelo en el que las variables independientes no tienen ningún efecto importante sobre la variable dependiente. Por el contrario, la verosimilitud del modelo completo es una medida de la probabilidad de que los datos hayan sido generados por un modelo logístico en el que todos los coeficientes son importantes: o sea, en el que todas las variables independientes tienen efecto sobre la dependiente. La comparación de ambas razones de verosimilitud (la del modelo sólo con la constante y la del modelo completo) permite comprobar si realmente las variables independientes tienen efecto sobre la dependiente. Si la verosimilitud del modelo completo es significativamente mayor que la del modelo sólo con la constante, puede decirse que lo más probable es que las variables

independientes del modelo tengan realmente efecto sobre la variable dependiente<sup>11</sup>.

Para conocer la verosimilitud del modelo es necesario comparar ambas medidas de verosimilitud, realizando un test de hipótesis estadístico. Pero no es necesario que se efectúe manualmente, puesto que Stata realiza este test automáticamente. Esto es lo que aparece en la segunda fila, segunda columna de la salida de *fitstat*: el *LR* test (test de la razón de verosimilitud). La medida que aparece como *LR* (6) es una prueba de  $\chi^2$  de la significación de la diferencia entre el modelo sólo con la constante y el modelo completo. Como siempre, la hipótesis nula es que todos los coeficientes excepto la constante son iguales a 0, y la hipótesis alternativa (que se acepta si no puede aceptarse la nula) es que los coeficientes son significativamente distintos de 0. La fórmula exacta de esta prueba de hipótesis es:

$$LR = 2 \ln \hat{L}_F - 2 \ln \hat{L}_0 \quad (11.20)$$

Debajo del *LR* aparece la probabilidad  $\chi^2$  asociada al valor de la prueba y a sus grados de libertad<sup>12</sup>. La probabilidad en este caso de que en la realidad todos los coeficientes de la ecuación logística fueran iguales a 0 es inferior a 0,0001, por lo que puede rechazarse la hipótesis nula: al menos uno de los coeficientes que aparecen en el modelo logístico estimado es significativamente distinto de 0.

En la siguiente fila aparece la medida de ajuste más importante del modelo logarítmico: el *Pseudo R<sup>2</sup>* o *McFadden R<sup>2</sup>*. Como ya quedó explicado este estadístico más arriba (en el apartado de estimación del modelo), no requiere que se repita su fórmula. No obstante, es preciso comentar que junto con el *Pseudo R<sup>2</sup>* estándar aparece el ajustado, que simplemente corrige el hecho de que el primero aumenta artificialmente al añadir nuevas variables, restando al numerador del *pseudo R<sup>2</sup>* el número de parámetros (coeficientes más la constante) del modelo. Por esta razón, el ajustado es preferible al estándar.

$$PseudoR_{aj}^2 = 1 - \frac{\ln \hat{L}_F - (k + 1)}{\ln \hat{L}_0} \quad (11.21)$$

Las tres siguientes filas no se comentan, porque son sólo otros tipos de  $R^2$  menos utilizados habitualmente y unas medidas basadas en el modelo de variable latente ( $y^*$ ) que no interesan en este contexto<sup>13</sup>.

<sup>11</sup> De hecho, el  $R^2$  de McFadden (como se explica en el apartado de estimación del modelo) es precisamente el complementario de la razón de ambas verosimilitudes.

<sup>12</sup> Los grados de libertad equivalen al número de coeficientes de la ecuación ( $k$ ) o, si se prefiere, igual al número de parámetros menos el que representa a la constante  $b_0$ .

<sup>13</sup> Para explicaciones acerca de estos estadísticos, véase Long y Freese (2006, p. 110 y ss).

La fila que merece la siguiente mención es la del *Count R<sup>2</sup>*. Las dos medidas que aparecen en esta fila están basadas en la comparación de los valores observados en la muestra y los predichos por el modelo, contraste que puede obtenerse mediante la orden *estat classification*, después de la estimación del modelo.

```
estat classification
```

Por suceder a un estimador, es esta una instrucción similar a *predict*, aunque en este caso su resultado no sea crear una serie de nuevas variables, sino una tabla de contingencia, seguida de un conjunto de estadísticos sobre la correcta clasificación de la variable *manif* a partir de los predictores empleados en la ecuación logística.

### ILUSTRACIÓN 11.9. Tabla de clasificación de la regresión logística

Logistic model for manif				
Classified	True		Total	
	D	~D		
+	629	396	1025	
-	766	1926	2692	
Total	1395	2322	3717	

Classified + if predicted Pr(D) >= .5  
 True D defined as manif != 0

---

Sensitivity	Pr( +   D)	45.09%
Specificity	Pr( -   ~D)	82.95%
Positive predictive value	Pr( D   +)	61.37%
Negative predictive value	Pr(~D   -)	71.55%

---

False + rate for true ~D	Pr( +   ~D)	17.05%
False - rate for true D	Pr( -   D)	54.91%
False + rate for classified +	Pr(~D   +)	38.63%
False - rate for classified -	Pr( D   -)	28.45%

---

Correctly classified		68.74%
----------------------	--	--------

Como repetidamente se ha insistido, el modelo logit predice la probabilidad de ocurrencia de un suceso. Pues bien, en consonancia con ello, en todos aquellos casos en los que el modelo prediga más de 0,5 de probabilidad de ocurrencia, la predicción será que ocurra (+ *Classified*); y en todos los casos en los que el modelo dé una probabilidad inferior a 0,5, se pronosticará que no sucederá (- *Classified*). El *Count R<sup>2</sup>* es simplemente la proporción de predicciones correctas según este criterio: o sea, en qué porcentaje de casos la predicción derivada del modelo de regresión logística acierta. La fórmula es, por tanto:

$$R_{\text{count}}^2 = \frac{\sum_{i=j=0}^1 f_{ij}}{n} \quad (11.22)$$

Siendo  $f_{ij}$  las cuatro frecuencias de la tabla de clasificación y ofreciendo el sumatorio de las casillas con idéntico índice ( $i=j$ ), el número de casos en los que la predicción coincide con la realidad en cada una de los dos posibles tipos de resultados ( $+|D$ ) y ( $-|\sim D$ )<sup>14</sup>; el modelo acierta en el 68,7% de los casos, lo que parece un porcentaje bastante alto de acierto. Demasiado alto, de hecho. El *Count R<sup>2</sup>* puede dar una impresión excesiva de capacidad predictiva del modelo por una razón muy simple: dado que los valores que puede tomar la variable dependiente en un modelo logístico son sólo dos (0 o 1), se puede acertar en más del 50% de los casos simplemente cogiendo todos los casos de la categoría que tenga más casos. En este ejemplo, dado que se sabe que el 63% de los encuestados no han asistido a ninguna manifestación, simplemente pronosticando siempre la categoría 0 se asegura más de un 63% de aciertos<sup>15</sup>. Por tanto, el *Count R<sup>2</sup>* no sirve para comparar condiciones de partida diferentes: es necesaria una medida que tenga en cuenta cuánto se mejora la capacidad de predicción con el modelo estimado, con respecto al simple conocimiento de la categoría con más casos.

Esto es exactamente lo que hace el *Adj Count R<sup>2</sup>*. Este estadístico es una modificación del anterior, que elimina de la cuenta de aciertos los relacionados con el marginal de fila mayor:

$$R_{\text{AdjCount}}^2 = \frac{\sum_{i=j=0}^1 f_{ij} - \max_i(f_{i.})}{n - \max_i(f_{i.})} \quad (11.23)$$

Siendo  $\max_i(f_{i.})$  la frecuencia marginal más alta entre la ocurrencia o no del fenómeno que se quiere pronosticar. En este ejemplo serán los 2.322 individuos que nunca han ido a una manifestación, en lugar de los 1.395 que sí lo hicieron.

<sup>14</sup> Los subíndices  $i$  y  $j$  son peculiares en esta tabla de dimensiones 2X2.  $i$  significa la presencia (1) o ausencia (0) de la calidad de la variable dependiente (D y  $\sim D$ ).  $j$  denota el pronóstico por la regresión logística de presencia (1 o +) o ausencia (0 o -). De este modo,  $f_{11}$  es el número de casos que se han manifestado y de los que se predice que se manifestaron;  $f_{00}$  sería por el contrario el número de casos que no se han manifestado y de los que el modelo augura que no lo han hecho. Tanto  $f_{01}$  como  $f_{10}$  son errores, pues implican a los casos con predicciones distintas de la realidad. Consecuentemente, la suma de todas las  $f_{ij}$  es igual a  $n$ .

<sup>15</sup> La sensibilidad (*sensitivity* en la ilustración 11.9) es la probabilidad de clasificar certeza-ramente a alguien con la categoría positiva; mientras que la especificidad (*specificity, ibidem*) es la proporción de clasificaciones correctas para los que poseen empíricamente una categoría negativa. Se acierta más prediciendo a quienes no asisten a manifestaciones (82,9%) que a quienes lo hacen (45,1%).

Esta medida es más justa, por tanto, puesto que indica la proporción de aciertos más allá de los que derivarían simplemente de poner todas las apuestas en el mayor marginal. Como puede comprobarse fácilmente, en este caso, la proporción es mucho menor: el modelo logit estimado sólo incrementa la capacidad de acierto en un 16,7% con respecto a la que se tendría simplemente prediciendo para todos los casos el valor más común (o sea, la no participación en manifestaciones). No es un resultado demasiado alto, pero indica que el modelo tiene cierta capacidad para predecir la asistencia a manifestaciones.

Los dos últimos estadísticos que aparecen en la salida de *fitstat*, *AIC* y *BIC*, son las llamadas *medidas de información*. Son medidas de ajuste especialmente diseñadas para comparar distintos modelos, incluso con diferentes muestras. El *BIC* en concreto es una medida muy útil para comparar regresiones logísticas con distinto número de coeficientes, como podrá verse a continuación<sup>16</sup>.

El *AIC* (*Akaike Information Criteria*) se calcula utilizando la verosimilitud del modelo y el número de parámetros. La fórmula es:

$$AIC = \frac{-2 \ln \hat{L}_F + 2(k+1)}{n} \quad (11.24)$$

$\hat{L}_F$  es la verosimilitud del modelo y  $k+1$  el número de parámetros. El valor de *AIC* es interpretable sobre todo en la comparación, más que en sí mismo: el modelo con un *AIC* menor es el mejor ajustado.

El *BIC* (*Bayesian Information Criterion*) es una medida aún más útil para comparar distintos modelos logit, puesto que está mejor desarrollada teóricamente. Está basada en la verosimilitud del modelo en cuestión y en sus peculiares grados de libertad (siendo estos igual a  $n-k-1$ ):

$$BIC = -2 \ln \hat{L}_F - (n - k - 1) \ln n \quad (11.25)$$

El  $BIC'_k$  es un ajuste del *BIC* que utiliza el LR o razón de verosimilitud del modelo, el número de coeficientes ( $k$ ) y el de casos ( $n$ ):

$$BIC' = -LR + k \ln n \quad (11.26)$$

<sup>16</sup> Como puede apreciarse en la ilustración 11.8, la orden *fitstat* produce dos valores de *AIC* y *BIC*. Ello es debido a que hay ciertas discrepancias en los detalles de la fórmula de Stata y la que usan Long y Freese (2006: 112). Así, el *AIC* que calcula Stata no divide por  $n$  el resultado. La fórmula de *BIC* de Stata (2011e: 157-161) es  $BIC = -2 \ln L_F + (k+1) \ln n$ . Aun aparentemente siendo distinta, la resta de los resultados entre modelos proporciona los mismos resultados, por lo que las indicaciones del cuadro 11.1 son también aplicables. Para obtener los resultados de Stata sin instalar el paquete de *Spost*, la instrucción es *estat ic*.

Como en el caso del *AIC*, el *BIC'* es sobre todo interesante para comparar modelos distintos, más que para interpretar su valor en términos absolutos. En principio, cuanto más negativo es el *BIC'*, mejor es el ajuste. La diferencia en el *BIC'* de dos modelos distintos indica qué modelo es más correcto. Raftery (1996) propuso unas pautas de interpretación de la diferencia en el *BIC* (o *BIC'*) de dos modelos distintos. En función de la diferencia  $BIC_1 - BIC_2$ , la evidencia de que el modelo más correcto es el segundo será:

**CUADRO 11.1. Tabla de interpretación en la comparación de modelos**

Diferencia $BIC_1 - BIC_2$	Evidencia de que el segundo modelo es mejor que el primero
Entre 0 y 2	Débil
Entre 2 y 6	Razonable
Entre 6 y 10	Fuerte
Más de 10	Muy fuerte

#### 11.4. Comparación de modelos

A veces resulta difícil decidir qué variables deberán incluirse en un modelo de regresión, sea logística, lineal o de cualquier otro tipo. En principio, a la hora de elegir la inclusión o exclusión de una determinada variable en un modelo de regresión múltiple, se puede optar por razones teóricas o por razones estadísticas. La razón teórica llevaría a introducir aquellas variables que parecen relevantes en función de la teoría o las hipótesis de partida. La razón estadística llevaría a elegir las variables que muestran un mayor grado de asociación estadística con la variable dependiente que se desea explicar. Realmente, a la hora de decidir qué variables se incluyen en el modelo, han de utilizarse ambos tipos de razones. Se ha de tener en cuenta la asociación estadística entre las variables, pero el simple hecho de que exista asociación no justifica la inclusión o no de una determinada variable en el modelo de regresión. Al utilizar técnicas estadísticas multivariadas como la que contempla este capítulo, se corre el riesgo de perder completamente la sustancia teórica por desarrollar modelos matemáticos muy elaborados, con altos grados de asociación y de robustez estadística, pero ningún interés sustantivo. La estadística no es más que una herramienta de análisis, que sirve para comprobar la validez de conceptos e hipótesis desarrolladas teóricamente.

Una técnica utilizada muy a menudo a la hora de seleccionar las variables que se han de incluir en un modelo de regresión es el llamado

método de selección por pasos. Consiste básicamente en introducir o eliminar las variables independientes en etapas sucesivas, estudiando la validez del modelo en cada una de ellas, para quedarse finalmente con aquel modelo que más se ajuste a los datos. Básicamente hay dos tipos de selección por pasos: de incorporación progresiva o de eliminación progresiva. En el primer caso se trata de ir añadiendo variables independientes al modelo, comprobando la significación del modelo en cada paso y de cada variable independiente y no incorporando aquellas variables que no añaden significación al modelo o que no son significativas ellas mismas. El segundo tipo de selección por pasos de variables consiste en partir de un modelo con el mayor número posible de variables independientes según el planteamiento teórico que se realice, para eliminar progresivamente las variables que no sean significativas o cuya eliminación no afecte de manera importante a la significación del modelo. En cualquiera de los casos, hay que tener en cuenta que ambos métodos de selección por pasos sólo aportan justificaciones técnicas, estadísticas, para la inclusión o no de una o varias variables en el modelo: como se ha dicho antes, estas justificaciones estadísticas no bastan por sí solas para optar por la inclusión o no de una variable, sino que es necesario que estén vinculadas a un razonamiento sustantivo.

La instrucción *fitstat* de Stata es una potente herramienta para la selección de variables con el método por pasos, pues aporta una información muy útil para comparar dos modelos distintos y decidir cuál es mejor. Como se acaba de ver, la orden *fitstat* se utiliza después de la estimación del modelo, y muestra distintas medidas escalares de ajuste del modelo. Además de esto, *fitstat* permite guardar en memoria las medidas de ajuste de un modelo determinado, y luego compararlas con las de otro modelo para determinar cuál es el que mejor se ajusta a los datos observados. Véase a continuación un ejemplo:

```
quietly logit manif mujer edad i.estudios i.ingresos amadecasa estudiante  
quietly fitstat, save  
quietly logit manif mujer edad i.estudios i.ingresos amadecasa  
fitstat, dif
```

Con el siguiente resultado:

### ILUSTRACIÓN 11.10. Comparación de las medidas de ajuste entre modelos

	Current	Saved	Difference
Model:	logit	logit	
N:	3700	3700	0
Log-Lik Intercept Only	-2448.039	-2448.039	0.000
Log-Lik Full Model	-2146.163	-2146.123	-0.040
D	4292.326(3692)	4292.246(3691)	0.080(1)
LR	603.753(7)	603.833(8)	0.080(1)
Prob > LR	0.000	0.000	0.778
McFadden's R2	0.123	0.123	-0.000
McFadden's Adj R2	0.120	0.120	0.000
ML (Cox-Snell) R2	0.151	0.151	-0.000
Cragg-Uhler(Nagelkerke) R2	0.205	0.205	-0.000
McKelvey & Zavoina's R2	0.197	0.197	0.000
Efron's R2	0.154	0.154	-0.000
Variance of y*	4.098	4.098	0.000
Variance of error	3.290	3.290	0.000
Count R2	0.693	0.692	0.001
Adj Count R2	0.182	0.180	0.002
AIC	1.164	1.165	-0.001
AIC*n	4308.326	4310.246	-1.920
BIC	-26041.472	-26033.335	-8.137
BIC'	-546.240	-538.104	-8.137
BIC used by Stata	4358.054	4366.191	-8.137
AIC used by Stata	4308.326	4310.246	-1.920
Difference of 8.137 in BIC' provides strong support for current model.			
Note: p-value for difference in LR is only valid if models are nested.			

La palabra *quietly* antes de cualquier instrucción simplemente hace que Stata no muestre el resultado de la instrucción, aunque realice los cálculos o estime el modelo. En el ejemplo primero se estima un modelo logit con las variables *estudios*, *edad*, si el individuo es estudiante, si es ama de casa, nivel de ingresos y si el individuo es mujer. Luego, se guardan los resultados de la orden *fitstat* en memoria con la opción *save*. Se estima a continuación el otro modelo, el que se desea comparar con el primero: es el mismo modelo, pero sin la variable *estudiante*. Después se introduce la orden *fitstat* con la opción *dif*, lo que hace que Stata muestre una comparación entre las medidas de ajuste calculadas por *fitstat* para los dos modelos, el que incorpora la variable *estudiante* y el que no. La columna *current* (actual) muestra la información del último modelo estimado, y la columna *saved* (grabado), la del modelo anterior, aquel del que se guardaron los datos con la opción del mismo nombre. En este caso, el modelo *current* es el que no incorpora la variable *estudiante* y el modelo *saved* es el que sí que la incorpora. La tercera columna muestra simplemente la diferencia entre los resultados *current* y *saved*, para facilitar su comparación.

En el apartado de ajuste del modelo ya se ha explicado cómo interpretar cada una de las medidas que aparecen en la salida de *fitstat*. La comparación de las medidas de ambos modelos puede servirnos para decidir cuál es mejor. Especialmente interesante es la comparación del estadístico *BIC*,

puesto que aporta un criterio bastante fiable para elegir entre dos modelos sucesivos en una selección por pasos de variables. En la tabla 1 aparecen unos criterios para interpretar las diferencias en el *BIC* de dos modelos de regresión logística. En la propia salida de la instrucción *fitstat*, cuando es utilizada para comparar dos modelos, aparece una primera interpretación de qué indica el *BIC* con respecto a qué modelo es mejor. En este caso, la diferencia en el *BIC* sugiere que el modelo *current* (el segundo, el que no incorpora la variable *estudiante*) es más correcto que el modelo *saved* (el que sí que la incorpora). Por tanto, parece que se debería preferir el modelo sin la variable *estudiante*.

La instrucción *fitstat*, por tanto, es de gran utilidad para la selección del modelo mejor ajustado según el método por pasos (*stepwise*). Pueden irse estimando sucesivos modelos, añadiendo o eliminando variables según parezca más adecuado en función de lo que sugiera la comparación de cada par de modelos sucesivos mediante *fitstat*, además de, por supuesto, en función de lo que sea teóricamente relevante.

Además de la instrucción *fitstat*, *dif*, que proporciona el conjunto de procedimientos SPost, con las instrucciones originales de la versión 8.0 de Stata y siguientes se pueden presentar diversas comparaciones de modelos, incluyendo más de dos, que no incluye la prueba estadística de la diferencia, pero que, en contrapartida, permite realizar al mismo tiempo el contraste entre los parámetros. La instrucción en cuestión es *estimates* y para su elaboración, al igual que la que se acaba de ver, es preciso realizarla en varios pasos. Aunque su uso es ilimitado, como ejemplo se van a utilizar tres modelos jerárquicos o anidados, esto es, una serie de modelos que sólo se diferencian entre sí en que uno de ellos carece de un subconjunto de variables del otro, pero no dispone de variables de las que carezca el primero o, dicho con otras palabras, los modelos jerárquicamente inferiores tienen menos variables que los superiores y ninguna de ellas diferentes.

En el ejemplo en cuestión se considera el modelo jerárquicamente superior el compuesto por la variable *edad*, la dicotómica *mujer*, las variables ficticias correspondientes a los ingresos familiares (*i.ingresos*), las variables ficticias de estudios (*i.estudios*) y las variables ficticias de la situación laboral (*jubilado*, *parado*, *estudiante* y *amadecasa*, considerándose como base la persona ocupada)<sup>17</sup>. En el segundo modelo se eliminarán dos de estas, dejando sólo las correspondientes a *jubilado* y *amadecasa*. Y, finalmente, en el tercer modelo, se descartan estas últimas, por lo que desaparece todo indicio de situación laboral.

<sup>17</sup> La situación laboral también habría podido ser introducida como variable factor. No se ha hecho de este modo para mostrar que es posible mezclar variables indicadores con variables factores y, sobre todo, porque la salida es mucho más clara, ya que con los factores sólo se muestra el código y no la etiqueta del valor. Es más ilustrativo “estudiante” que “6”.

```

logit manif mujer edad i.estudios i.ingresos jubilado parado estudiante
      amadecasa
estimates store Modelo3, title("Modelo superior")
logit manif mujer edad i.estudios i.ingresos jubilado amadecasa if e(sample)
estimates store Modelo2, title("Modelo intermedio")
logit manif mujer edad i.estudios i.ingresos if e(sample)
estimates store Modelo1, title("Modelo inferior")
estimates table Modelo1 Modelo2 Modelo3, star stats(N ll_0 ll chi2 r2_p aic bic)

```

Como puede apreciarse, las instrucciones precedentes constan de tres pares de líneas, uno para cada modelo y una final en la que se combinan los tres modelos. Cada par consta de la instrucción *logit*, propiamente dicha<sup>18</sup>, y de otra (*estimates*), que graba (*store*) la información del modelo precedente bajo el nombre que el usuario considere más oportuno —en este caso, Modelo3, Modelo2 y Modelo1, respectivamente, y con la etiqueta que voluntariamente se escriba en la opción *title* (“Texto deseado”)—. En la última instrucción se conjugan los tres modelos con la misma instrucción *estimates* seguida de *table* y los modelos que se desean comparar. A esta última conviene añadirle la opción *star*, que es la que coloca los asteriscos a los coeficientes significativos, y la opción *stats* acompañada de los estadísticos correspondientes a cada modelo que se desea aparezcan en cada uno de ellos. Los posibles para las regresiones logísticas son *N* (número de casos), *ll\_0* (logaritmo de la verosimilitud del modelo base), *ll* (logaritmo de la verosimilitud del modelo evaluado), *chi2* (el test de la razón de verosimilitud), *r2\_p* (el pseudo R<sup>2</sup>), *aic* y *bic* (criterios de información de Akaike y bayesiano)<sup>19</sup>.

---

<sup>18</sup> Con objeto de realizar las pruebas de los modelos jerárquicos con el mismo número de caso, conviene ejecutar en primer lugar el modelo con más parámetros y, a continuación, los siguientes teniendo cuidado de incluir en la instrucción *logit* la selección *if e(sample)*, con objeto de que sólo trabaje con los casos del modelo anterior.

<sup>19</sup> La orden original de Stata *estimates table*, a diferencia de la instrucción *fitstat, dif* del módulo SPost, utiliza la siguiente fórmula  $BIC_k = -2\ln \hat{L}(M_k) + p \ln n$ , que es equivalente en las comparaciones.

**ILUSTRACIÓN 11.11. Exposición de parámetros con significación de varios modelos**

Variable	Modelo1	Modelo2	Modelo3
mujer	-.40227661***	-.25954549**	-.26264363**
edad	-.01448026***	-.00657854*	-.00597551
estudios			
2	.62288533***	.5919637***	.58675585***
3	1.4243038***	1.3544712***	1.3472467***
ingresos			
2	.54855537***	.52480584***	.52926195***
3	.63062047***	.57590225***	.58352454***
jubilado		-.47203306***	-.47685376***
amadecasa		-.60354102***	-.59545554***
parado			.02651879
estudiante			.10327419
_cons	-.38311017**	-.57134389***	-.60620518***
N	3700	3700	3700
ll_0	-2448.0392	-2448.0392	-2448.0392
ll	-2153.3079	-2140.208	-2139.9926
chi2	589.46276	615.66246	616.09327
r2_p	.12039488	.12574603	.12583403
aic	4320.6157	4298.416	4301.9852
bic	4364.1284	4354.3608	4370.3622

legend: \* p<0.05; \*\* p<0.01; \*\*\* p<0.001

En la ilustración 11.11 aparecen los tres modelos, desde el más simple a la izquierda hasta el más complejo de la derecha. En ellos aparecen los coeficientes logísticos (aunque con la opción *eform* podrían mostrarse en su lugar los cocientes de razones [*vid. infra*]), seguidos de la información de las características de cada modelo. Al seleccionar para todos ellos los mismos casos, el resultado de las dos primeras líneas de los estadísticos (*N* (*n*) y *ll\_0*, esto es,  $\ln \hat{L}_0$ ), son iguales para los tres modelos. Puede fácilmente deducirse de los datos que los tres modelos son significativos, con idénticos valores de pseudo  $r^2$ , pero por los dos criterios de información, el preferible (por tenerlos menores) es el modelo central, que es el que contiene todos los parámetros significativos del modelo superior.

En el ejemplo anterior se ha utilizado la opción *star* de la instrucción *estimates tables*, que añade estrellas a los coeficientes significativos de un modelo. Una alternativa a ella es la introducción de otros estadísticos. Existen cuatro opciones, que dan cuenta de cada uno de ellos: *b(%formato)* para los coeficientes propiamente dichos; *se(%formato)*; *t(%formato)*, y *p(%formato)* para la significación del valor de *t*. Como se deduce de su sintaxis, el usuario puede especificar el número de decimales que se van a mostrar a través del optativo (%formato). El siguiente ejemplo muestra la misma tabla de la ilustración 11.11, pero se han omitido los estadísticos del modelo, los coeficientes se muestran sólo con un decimal y aparecen los errores típicos con dos cifras decimales y las significaciones en notación científica.

estimates table Modelo1 Modelo2 Modelo3, b (%6.1f) se(%6.2f) p(%6.2e)
---

Siempre y cuando se hubieran grabado anteriormente los tres modelos citados, el resultado de la anterior instrucción aparece en la ilustración 11.12.

### ILUSTRACIÓN 11.12. Opciones de formato en la presentación de modelos

Variable	Modelo1	Modelo2	Modelo3
mujer	-0.4 0.07 5.66e-08	-0.3 0.08 1.5e-03	-0.3 0.08 1.4e-03
edad	-0.0 0.00 1.2e-09	-0.0 0.00 4.0e-02	-0.0 0.00 7.3e-02
estudios	2 0.6 0.09 3.0e-11	0.6 0.09 3.6e-10	0.6 0.09 6.1e-10
	3 1.4 0.12 0.0e+00	1.4 0.12 0.0e+00	1.3 0.12 0.0e+00
ingresos	2 0.5 0.08 5.1e-11	0.5 0.08 5.4e-10	0.5 0.09 7.4e-10
	3 0.6 0.13 5.0e-07	0.6 0.13 5.5e-06	0.6 0.13 5.7e-06
jubilado		-0.5 0.14 5.7e-04	-0.5 0.14 6.0e-04
amadecasa		-0.6 0.13 2.0e-06	-0.6 0.13 4.4e-06
parado		0.0 0.14 8.5e-01	0.0 0.14 0.1
estudiante			0.16 5.2e-01
_cons	-0.4 0.15 9.9e-03	-0.6 0.16 3.0e-04	-0.6 0.17 3.2e-04

legend: b/se/p

### 11.5. Interpretación del modelo

Se llega, por fin, a lo que realmente interesa: la interpretación de los resultados del modelo logit. Con respecto a la regresión lineal, hay un par de aspectos que hay que tener en cuenta, que hacen la interpretación de la regresión logística considerablemente más complicada.

Primero, como ya se ha comentado a lo largo de este capítulo, hay que considerar que los coeficientes del modelo logit tal cual no sirven para su interpretación. Si se recuerda la ecuación básica del logit...

$$\ln \frac{\Pr(y = 1|\mathbf{x})}{1 - \Pr(y = 1|\mathbf{x})} = \ln \Omega(\mathbf{x}) = b_0 + b_1x_1 + \dots + b_kx_k \quad (11.27)$$

... puede verse que la variable dependiente aparece en una forma no directamente interpretable, por lo que el efecto sobre ella de las variables independientes no se podrá estudiar de manera directa. Por ejemplo, en el modelo presentado, el coeficiente de la variable *edad* tiene un valor de -0,0145. ¿Qué significado sociológico tiene que por cada año de edad disminuya en 0,0145 el logaritmo de la razón de asistencia a manifestaciones? Poco se puede decir de este dato, salvo que muestra un efecto negativo. Para estudiar el modelo logit, será necesario transformar la ecuación original, como se vio en (11.8) y en (11.10), para conseguir coeficientes que puedan ser interpretados.

Hay dos transformaciones de la ecuación logit original que permiten su interpretación inmediata. La primera supone eliminar el logaritmo del lado derecho de la ecuación original. Despejando:

$$\Omega(\mathbf{x}) = \frac{\Pr(y = 1|\mathbf{x})}{1 - \Pr(y = 1|\mathbf{x})} = \exp(b_0 + b_1x_1 + \dots + b_kx_k) \quad (11.28)$$

En este caso, los valores de los coeficientes indican cómo varía la razón de ocurrencia del suceso medido por la variable dependiente en función de un cambio de magnitud 1 en el valor de las variables independientes. Esta forma de interpretar los resultados del logit será la explicada en el apartado a, pero, aunque es una forma válida de interpretar una regresión logística, sigue siendo bastante compleja, porque el cambio en la variable dependiente se expresa en términos de (cociente de razones), concepto asumible en teoría, pero poco intuitivo. Sería mucho más fácil si el cambio en la variable dependiente se expresara simplemente en términos de probabilidad de ocurrencia del suceso estudiado.

Mejor comprensión, por tanto, puede conseguirse mediante una segunda transformación de la ecuación logit original que en el lado derecho exprese las variaciones en la probabilidad de ocurrencia del suceso. Despejando:

$$\Pr(y = 1|\mathbf{x}) = \frac{\exp(b_0 + b_1x_1 + \dots + b_kx_k)}{1 + \exp(b_0 + b_1x_1 + \dots + b_kx_k)} \quad (11.29)$$

Ahora bien, esta ecuación tiene un problema, puesto que expresa una función no lineal. Esto quiere decir que el efecto de una variable independiente sobre la variable dependiente es diferente según el valor que tengan todas las demás variables independientes, además de según su propia magnitud. En la regresión lineal, el efecto de las variables del modelo es independiente y constante: el cambio en la variable independiente produce siempre el mismo cambio en la dependiente, da igual cuál sea el valor de las otras variables del modelo. La magnitud de ese cambio es la que se expresa en el valor del coeficiente asociado a cada variable independiente en la re-

gresión lineal. Pero en el caso de la regresión logística, dado que el cambio que provoca cada variable en la probabilidad de ocurrencia del suceso estudiado depende del valor de todas las demás variables, ni siquiera se puede asociar un coeficiente a cada variable independiente cuando el modelo está expresado en probabilidades como en (11.29)<sup>20</sup>. Por tanto, hay que cambiar totalmente la estrategia de análisis. La manera de estudiar una ecuación logística en forma probabilística es utilizando las probabilidades predichas por el modelo para valores específicos de las variables independientes. Esta segunda estrategia de interpretación se verá en el apartado 11.5.2.

### 11.5.1. Interpretación a través de cocientes de razones

Para que se muestre la ecuación logit en términos de cocientes de razones, hay que añadir la opción *or* (o escribir sin opción la instrucción *logistic*, en lugar de *logit*). Si ya se han pedido los coeficientes normales, no es necesario repetir el conjunto de variables, puesto que Stata recuerda la última lista de variables, en caso de que no se especifique ninguna. Véase que muestra esta opción en el ejemplo seguido:

#### ILUSTRACIÓN 11.13. Regresión logística con cociente de razones

Logistic regression						Number of obs	=	3717
						LR chi2(6)	=	589.74
						Prob > chi2	=	0.0000
						Pseudo R2	=	0.1199
Log likelihood = -2164.7337								
-----		manif		Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
		mujer		.669914	.0494824	-5.42	0.000	.5796236 .7742692
		edad		.9855737	.0023394	-6.12	0.000	.9809993 .9901695
		estudios						
		2		1.865738	.1744364	6.67	0.000	1.553345 2.240957
		3		4.145156	.476195	12.38	0.000	3.309443 5.191905
		ingresos						
		2		1.718636	.143193	6.50	0.000	1.4597 2.023505
		3		1.85572	.232034	4.94	0.000	1.452382 2.37107

Como puede verse, el formato de la salida es el mismo que el del *logit* normal. De hecho, los datos mostrados son idénticos salvo en cuatro columnas: la de cociente de razones (*odds ratio*) y la de sus correspondientes

<sup>20</sup> Este problema sólo existe cuando se utiliza la forma probabilística de la ecuación logit. Si se estudia la ecuación en términos de cociente de razones, se obtienen coeficientes de variación constantes para cada variable, por lo que el análisis se realiza de manera análoga al de la regresión lineal. De hecho, la ecuación logit en términos de cocientes de razones es una ecuación lineal. Como se verá más adelante, esto es así porque la ecuación logit no expresa probabilidades, sino cambios en el cociente de las razones.

errores típicos e intervalos de confianza. Esto es así porque realmente se trata del mismo modelo, sólo que expresado de manera diferente.

¿Cómo pueden interpretarse los cocientes de razones? Los valores de la columna correspondiente expresan cuánto varía la razón de ocurrencia del suceso en función del cambio en las variables independientes, es decir: cuando la variable independiente en cuestión aumenta en una unidad, cuánto varía la razón de asistencia a manifestaciones. Puede así decirse que el tener estudios secundarios (valor 2 de estudios) incrementa la razón de asistencia a manifestaciones 1,87 veces. Recuérdese lo que es razón: es la razón que representa la frecuencia ocurrencia de un suceso sobre la frecuencia de su no ocurrencia. El dato que muestra la columna de los cocientes de razones expresa el cambio que experimenta la mencionada razón, cuando una variable independiente varía en una unidad. Si el cociente de razones asociado a una variable es superior a 1, la razón aumenta cuando aumenta el valor de la variable (como es el caso de los dos coeficientes respectivos de estudios e ingresos): por tanto, la variable tiene un efecto positivo sobre la probabilidad de ocurrencia del suceso. Si el coeficiente mostrado es inferior a 1, la razón de ocurrencia del suceso disminuye cuando aumenta en una unidad la variable independiente en cuestión. En este caso, esto ocurre con la variable mujer: cuando el entrevistado es mujer (la variable mujer pasa de 0 a 1) la razón de asistencia a manifestaciones disminuye al tener que ser multiplicada por 0,67 (con respecto a la de los hombres).

En la interpretación de los resultados de un logit a través de los cocientes de razones, hay que tener muy en cuenta que no se está tratando directamente sobre las probabilidades de ocurrencia del suceso estudiado, sino sobre cómo varían las razones de ocurrencia del suceso en función de las variables independientes. Sólo pueden estudiarse las variaciones de las probabilidades si se tiene en cuenta el conjunto de variables independientes. Por ello, para el estudio de las probabilidades predichas para cada caso, hay que utilizar unas técnicas de análisis diferentes, que se verán en el próximo apartado. Las variaciones en las razones son de mucho interés para el estudio de un logit porque permiten cuantificar el efecto relativo de las distintas variables independientes sobre la variable dependiente. Pero no sirven para hacer predicciones.

Siguiendo con la interpretación del análisis del ejemplo en cuestión, las variables que tienen un efecto positivo sobre la probabilidad de asistencia a manifestaciones son: *estudios* (más probabilidad cuantos más estudios) e *ingresos* (ídem). Las variables *sexo* y *edad* afectan negativamente. Con respecto a la magnitud de este efecto, hay que tener en cuenta un par de cuestiones. Primero, que las variaciones positivas y negativas en los cocientes de razones son difíciles de comparar inmediatamente, porque no tienen el mismo rango de variación. Las variaciones negativas van de 0 a 1 y las positivas de 1 a infinito. Esto es así porque el cociente de razones expresa qué proporción

representa la razón después del efecto de la variable independiente frente a la razón antes de tal efecto. Por ejemplo, el valor de 1,88 en el valor 2 de *estudios* significa que cuando el individuo tiene estudios secundarios, la razón de asistencia a manifestaciones es 1,88 veces superior que cuando no tiene estudios o tiene estudios primarios. El valor de 0,67 asociado a la variable *mujer* quiere decir que cuando el individuo es mujer, la razón de asistencia a manifestaciones es 0,67 veces inferior a cuando es hombre. ¿Qué variable tiene más efectos sobre la probabilidad de asistencia a manifestaciones, *estudios secundarios* o *sexo*? Es difícil de decir porque los rangos de los efectos positivos y de los efectos negativos son distintos. Para hacerlos comparables, hay una fácil solución: se puede calcular el valor inverso de uno de los datos. El valor de 0,67 de variación asociado a la variable *mujer* es equivalente a un efecto positivo de 1,49 ( $1/0,67=1,49$ ). Por tanto, el efecto de los estudios medios es superior al efecto de la variable *mujer*.

Otra cuestión a tener en cuenta al establecer comparaciones entre distintos cocientes de razones es el rango de variación de las variables independientes. El valor de la correspondiente columna representa cuánto varía la razón cuando la variable independiente varía en una unidad. Lógicamente, la variación en una unidad de una variable dummy como *mujer*, que sólo puede tener dos valores, 0 y 1, es mucho más importante que la variación en una unidad de una variable continua como *edad*, que puede tener valores que van de 16 a 90. Puede observarse fácilmente cómo todas las variables dicotómicas aparentan un efecto sustancialmente mayor que la única variable continua incluida en el modelo, la *edad*.

Para superar este problema comparativo de los distintos coeficientes cuando tienen rangos diferentes, puede emplearse el programa del conjunto de utilidades para logit *SPost* que lista coeficientes. La instrucción es *listcoef*, que ha de emplearse tras la estimación logit estándar. Si se añade la opción *help* se obtiene una descripción de lo que significan las abreviaturas utilizadas en los encabezamientos de las columnas.

#### **ILUSTRACIÓN 11.14. Listado de coeficientes logísticos y cocientes de razones**

logit (N=3717) : Factor Change in Odds						
Odds of: 1 vs 0						
manif	b	z	P> z	e^b	e^bStdX	SDofX
mujer	-0.40061	-5.424	0.000	0.6699	0.8186	0.4997
edad	-0.01453	-6.122	0.000	0.9856	0.7650	18.4357
2.estudios	0.62366	6.671	0.000	1.8657	1.3205	0.4457
3.estudios	1.42194	12.378	0.000	4.1452	1.6754	0.3629
2.ingresos	0.54153	6.500	0.000	1.7186	1.3018	0.4870
3.ingresos	0.61827	4.945	0.000	1.8557	1.2237	0.3266

Como puede apreciarse inmediatamente, esta orden muestra más información que la instrucción original *logit, or*. Las tres primeras columnas muestran los coeficientes logit estándar (*b*), su valor *z* (*z*) y su probabilidad ( $P>|z|$ ). La cuarta columna muestra el cociente de razones, el mismo que muestra la anterior instrucción *logit, or*. La columna que interesa aquí es la quinta ( $e^bStdX$ ), que muestra el cambio en las razones para un incremento de la variable independiente de una desviación típica. Al utilizar como unidad de variación de la variable independiente su desviación típica, todos los coeficientes pueden compararse entre sí. De este modo, aparece cómo la magnitud relativa de la variable *edad* aumenta sustancialmente al medirla en desviaciones típicas: de hecho, tiene un efecto en la razón mayor (más próximo a 0) que la variable *mujer* (más próxima a 1), aunque en principio pareciera lo contrario. Calculando los valores inversos de *edad* y *mujer*, para comparar sus magnitudes con el resto de las variables, puede concluirse la interpretación de los cocientes de razones. El inverso de *edad* es 1,31 y el de *mujer* es 1,22.

Tanto el nivel de estudios como el nivel de ingresos afectan positivamente a la probabilidad de haber asistido a alguna manifestación. La edad y el género afectan, en cambio, negativamente (cuanto más edad tiene el individuo, hay menos probabilidad de que haya asistido a alguna manifestación; y, además, las mujeres tienen menos probabilidad de haber asistido que los hombres). Una vez estandarizadas todas las medidas, la variable que tiene un efecto más importante en la probabilidad de manifestarse es el nivel de estudios: cuantos más estudios, más probabilidad de haber asistido a manifestaciones. También bastante importante, aunque menos, es la variable *edad*, en el sentido ya explicado. *Ingresos* y *género* presentan también una asociación significativa, aunque menos importante.

### 11.5.2. Interpretación a través de predicciones

Una segunda manera de estudiar los resultados de una regresión logística es, como ya se ha aludido más arriba, a través de las predicciones del modelo para valores específicos de las variables independientes. Esta forma tiene la ventaja de que los resultados son más intuitivos que los derivados del estudio de los cocientes de razones, que no dejan de ser una medida relativamente compleja y difícil de interpretar. Pero esta forma tampoco es sencilla, puesto que la función logit en términos probabilísticos no es lineal, como ya se ha reflejado, lo que hace considerablemente más complejo su análisis. Es preciso notar que la regresión logística no es una técnica sencilla y requiere de una cierta práctica sobre todo en la interpretación de los parámetros.

Stata incorpora pocas herramientas para estudiar la regresión logística a través de las probabilidades predichas. Por ello, aquí se utilizarán básicamente los programas del conjunto de herramientas *SPost*.

Se ha explicado más arriba que el efecto de una variable independiente sobre la dependiente en un modelo logit en forma probabilística depende del

valor de todas las variables incluidas en el modelo, así como de su propia magnitud. Por ello, no es posible tener un coeficiente asociado a cada variable independiente que exprese el efecto de esa variable sobre la variable dependiente de manera probabilística. Hay dos maneras de estudiar este efecto teniendo en cuenta el problema aludido: una, mantener todas las variables en un valor determinado (normalmente, pero no necesariamente, la media) y hacer variar sólo una variable, estudiando cómo afecta a las predicciones del modelo; la otra forma consiste en dar valores específicos a todas las variables del modelo, según interese, y ver qué predicción arroja el modelo en esos casos.

La primera de las dos estrategias de análisis se puede desarrollar mediante la instrucción *prchange*<sup>21</sup>. Esta muestra cómo afecta la variación de una o más variables en la predicción de ocurrencia del suceso estudiado, manteniendo constantes el resto de las variables introducidas en el modelo (en la media). Por ejemplo, se puede estudiar cómo afecta el hecho de ser mujer sobre la probabilidad de haber participado en manifestaciones. La instrucción necesaria sería:

```
logit manif mujer edad estu2 estu3 ingr2 ingr3
prchange mujer, fromto
```

Y el resultado de aplicarla aparece en la próxima ilustración.

#### ILUSTRACIÓN 11.15. Efecto sobre la variable dependiente de los cambios en una independiente dicotómica

```
logit: Changes in Predicted Probabilities for manif

      from:          to:          dif:          from:          to:          dif:          from:
      x=min        x=max    min->max      x=0        x=1    0->1      x-1/2
mujer   0.4058     0.3139   -0.0919     0.4058     0.3139   -0.0919     0.4040

      to:          dif:          from:          to:          dif:
      x+1/2      -+1/2    x-1/2sd  x+1/2sd  -+sd/2 MargEfct
mujer   0.3123   -0.0917     0.3802     0.3342   -0.0459   -0.0919

      0           1
Pr(y| $\bar{x}$ )  0.6431   0.3569

      estu2      estu3       edad      ingr2      ingr3      mujer
      x=    .273339    .15604   46.6718    .386333    .121334    .518698
      sd(x)=  .445733    .362942  18.4357    .486974    .326559    .499717
```

<sup>21</sup> Las herramientas *SPost* contempladas en este apartado no funcionaban con variables factores en el momento en el que se redactó este texto. Por ello, antes de aplicarlas, se vuelve a ejecutar la regresión logística empleando variables ficticias con valores 0/1. Desde la versión 11, Stata incorpora la instrucción *margins* y, desde la 12, *marginsplot*, que puede hacer lo fundamental de *prchange* y *prvalue* y otras operaciones también con factores y con otros modelos. Más detalles y explicaciones de esta nueva orden se pueden encontrar en el manual (Stata 2011: 1027-1079 y 1099-1129).

Sin la opción *fromto* se muestra la misma información, pero menos detallada. En este caso, de hecho, la orden *prchange* muestra mucha más información de la verdaderamente necesaria, puesto que al tratarse de una variable dicotómica ficticia sólo puede tener dos valores (0 y 1), y la opción *fromto* muestra información para muchos otros valores posibles. Véase lo que puede interpretarse de la ilustración 11.15. La primera columna (*from: x=min*) muestra la probabilidad que predice el modelo para un individuo con el valor medio en todas las variables independientes salvo en la variable *mujer*, en la que adopta el valor 0. O sea, se trata de la predicción de la probabilidad (0,41) de haber asistido a alguna manifestación en el caso de que se estuviera ante un hombre con características medias en cuanto a estudios, edad e ingresos. La segunda columna (*to: x=max*) muestra la probabilidad predicha para un individuo de idénticas características pero con valor 1 en la variable *mujer*. Para la mujer media, por tanto, la probabilidad de haber asistido a alguna manifestación es de 0,31, según el modelo logístico. Por tanto, ser mujer reduce la probabilidad de haber asistido a alguna manifestación en 0,09 para una persona de características medias, lo que aparece en la tercera columna. Las siguientes columnas con información de tipo *from/to* no tienen relevancia para el caso de variables ficticias, (más tarde, se estudiará una variable continua, *edad*, en la que sí que tienen relevancia). Bajo las columnas de *from/to*, aparece la probabilidad de ocurrencia y no ocurrencia del suceso con los valores de todas las variables independientes en sus medias. Abajo del todo aparecen precisamente estos valores medios, los que se han utilizado para hacer las predicciones, y sus desviaciones típicas.

Se estudia a continuación el efecto sobre la probabilidad de asistencia a manifestaciones de la variable *edad*, en este caso, dado que la variable es continua, puede obtenerse mucha más información relevante con la instrucción *prchange*.

#### ILUSTRACIÓN 11.16. Efecto sobre la variable dependiente de los cambios en una independiente numérica

logit: Changes in Predicted Probabilities for manif							
	from:	to:	dif:	from:	to:	dif:	from:
	x=min	x=max	min->max	x=0	x=1	0->1	x-1/2
edad	0.4570	0.2156	-0.2414	0.5223	0.5187	-0.0036	0.3585
	to:	dif:	from:	to:	dif:	from:	
	x+1/2	+1/2	x-1/2sd	x+1/2sd	+sd/2	MargEfct	
edad	0.3552	-0.0033	0.3882	0.3267	-0.0614	-0.0033	
	0	1					
Pr(y x)	0.6431	0.3569					
	estu2	estu3	edad	ingr2	ingr3	mujer	
x=	.273339	.15604	46.6718	.386333	.121334	.518698	
sd(x)=	.445733	.362942	18.4357	.486974	.326559	.499717	

La probabilidad de que un individuo con características medias en todas las variables pero con el valor mínimo de la variable *edad* (18 años) haya asistido a alguna manifestación es de 0,46. La probabilidad de que un individuo con la edad máxima (95 años) haya asistido a alguna manifestación es de 0,22. Por tanto, el efecto de la edad es negativo, reduciéndose en 24 puntos porcentuales la probabilidad de asistencia a alguna manifestación cuando la edad pasa de su valor mínimo a su valor máximo. Este efecto resulta (sociológicamente) curioso, puesto que en principio, cuanto más edad, es más posible que el individuo haya tenido la oportunidad de asistir a alguna manifestación en su vida; el resultado del análisis del logit es contrario a lo que intuitivamente parece más probable, e indica probablemente un efecto generacional, según el cual las personas de más edad (con unos valores menos tendentes a la participación política, tal vez por el legado de la dictadura) tienen menos propensión a la participación política que los jóvenes. Los resultados from:  $x=0$ , from:  $x=1$  y dif:  $0>1$  no deben tenerse en cuenta en este caso, puesto que no tienen sentido. El modelo simplemente extiende el efecto negativo de la edad hasta el valor mínimo absoluto (0), pero obviamente no tiene ningún sentido el estudiar la probabilidad de asistencia a manifestaciones de una persona de cero o de un año.

Más interesantes son las siguientes columnas. La columna from:  $x-1/2$  muestra el valor de la predicción de asistencia a manifestaciones para una persona de edad media menos medio año; la columna siguiente muestra lo mismo más medio año. Por tanto, lo que se calcula aquí es la variación de la probabilidad de asistencia a manifestaciones en función de variaciones muy pequeñas de la edad: idea que es muy próxima a la derivada de una función en un punto o pendiente de la curva. Se trata de ver cuál es la tasa de cambio de la variable dependiente estudiando cómo responde a variaciones de pequeña magnitud de la variable independiente. La diferencia entre estas dos columnas proporciona la tasa de cambio estimada en torno a los valores medios de la variable independiente, lo que se puede también tomar como una estimación del efecto marginal (con terminología más propia de la economía) de la edad (en torno a su media) sobre la probabilidad de asistencia a manifestaciones (que en este caso es -0,0033).

Las tres siguientes columnas tienen una función parecida: se trata de ver cómo responde la variable dependiente a cambios en la independiente, sólo que en este caso, en vez de medio punto (año, por tratarse de la edad), se le suma y se le resta media desviación típica, lo que de algún modo estandariza la estimación de la tasa de cambio marginal (lo que podría utilizarse para comparar tasas de cambio de distintas variables, con distintos rangos). En este caso es de -0,06.

Para el caso de variables continuas, una manera diferente pero muy interesante de mostrar la misma información que la instrucción *prchange* es a través de gráficos. Puede hacerse un gráfico que muestre la relación entre la edad y la probabilidad de asistencia a manifestaciones, manteniendo

las otras variables en la media, igual que ocurría con *prchange*. Para hacer esto puede utilizarse la orden *prgen*. La sintaxis de esta es la siguiente: se teclea primero la palabra *prgen* seguida por la variable independiente que quiere analizarse (en este caso *edad*), y luego, en las opciones, se especifican los valores iniciales y finales de la variable independiente que se utilizarán en la predicción y se especifica el nombre que se le quiere poner a la variable nueva que genera la instrucción. Lo que hará entonces *prgen* es crear una nueva variable que contendrá la predicción que haga el logit de su probabilidad de asistencia a manifestaciones para cada valor de la variable *edad*. Si se añade la opción *ci* generará también los intervalos de confianza de los predictores. Todo ello se puede ver mejor con un ejemplo:

```
prgen edad, from (18) to (95) generate(edad) ci
```

La anterior instrucción muestra el resultado que se presenta a continuación, donde aparecen las medias de las variables independientes, pero, en realidad, lo que genera son tres nuevas variables con valores comprendidos entre 18 y 95 de la variable *edad*.

#### **ILUSTRACIÓN 11.17. Listado de la aplicación de la instrucción *prgen***

logit: Predicted values as edad varies from 18 to 95.
estu2        estu3        edad        ingr2        ingr3        mujer
x= .27333871 .15603982 46.671778 .38633306 .12133441 .51869787

Por omisión sólo en los 11 primeros casos del fichero<sup>22</sup>, la instrucción *prgen* crea tres nuevas variables con el nombre especificado entre paréntesis tras *generate* terminado en *x*, *p0* y *p1*. La variable *edadx* simplemente almacena valores a intervalos iguales de la mencionada variable independiente; *edadp0* contiene las probabilidades predichas de que el individuo no haya asistido a ninguna manifestación (o sea, de que manif=0) y *edadp1*, de que el individuo haya asistido (manif=1). Las variables *edadp0lb*, *edadp1lb*, *edadp0ub* y *edadp1ub* contienen los intervalos de confianza (inferior y superior) de las probabilidades para cada valor de *x*. Lo que interesa saber es cómo varía en función de la edad la probabilidad de asistencia a manifestaciones, lo que puede mostrarse con sencillez mediante un gráfico de rango que refleje los intervalos de confianza de la predicción de probabilidad de asistencia a manifestaciones por edades.

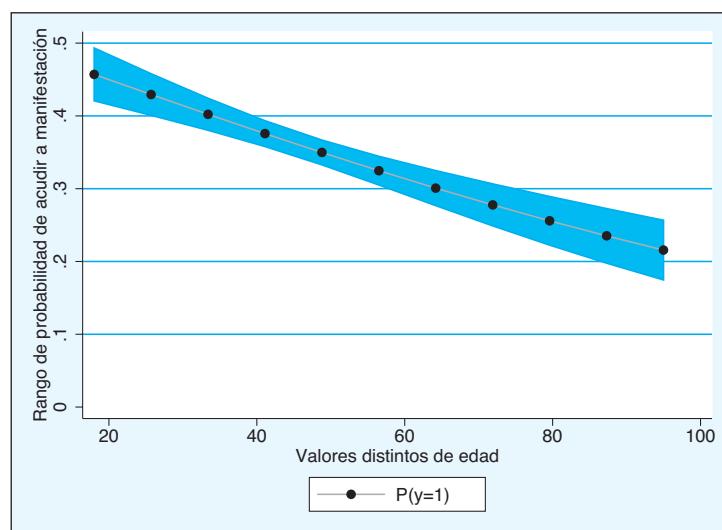
---

<sup>22</sup> Se puede modificar el número de puntos generados con la opción *ncases(#)*.

```
graph twoway rarea edadp1lb edadp1ub edadx, ///
    ylabel(0(.1).5) name(G5, replace) ///
    xtitle("Valores distintos de edad") ///
    ytitle("Rango de probabilidad de acudir a manifestación")
```

Y a continuación aparece el gráfico con los once puntos que genera la instrucción.

**GRÁFICO 11.3. Gráfico de probabilidades predichas para distintos valores de una variable independiente**



Como puede apreciarse, el gráfico 11.3 muestra cómo varía la probabilidad de asistencia a manifestaciones en función de la edad, manteniendo todas las demás variables del modelo constantes en su valor medio.

Pueden incluso complicarse algo más las cosas y plasmar en un mismo gráfico más de una variable. Por ejemplo, se puede mostrar cómo varía la probabilidad de asistencia a manifestaciones en función de la edad y del género. Ello se puede hacer aprovechando la posibilidad que brinda la orden *prgen* de especificar el valor de las otras variables del modelo, con la opción *x(variable=valor)*. Si se especifica ningún valor de ninguna variable, *prgen* mantiene todas las variables en su media salvo la que sirve para hacer la predicción. Pero con esta opción puede otorgarse un valor específico a una o más variable según se desee. Véase con un ejemplo para una mejor comprensión de este proceso.

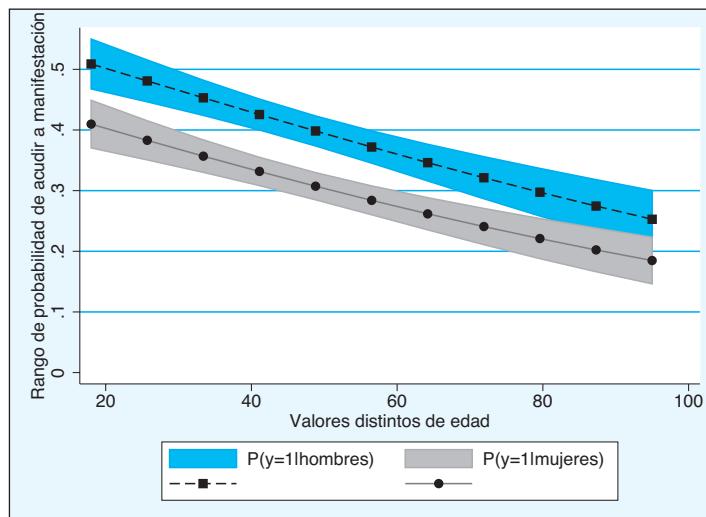
```

prgen edad, from(18) to(95) generate(edadhombre) x(mujer=0) ci
prgen edad, from(18) to(95) generate(edadmujer) x(mujer=1) ci
label variable edadhombrep1 "P(y=1|hombres)"
label variable edadmujerp1 "P(y=1|mujeres)"
graph twoway (rarea edadhombrep1lb edadhombrep1ub edadx) ///
    (rarea edadmujerp1lb edadmujerp1ub edadx), ///
    ylabel(0(.1).5) name(G6, replace) ///
    xtitle("Valores distintos de edad") ///
    ytitle("Rango de probabilidad de acudir a manifestación") ///
    legend(label(1 "P(y=1|hombres)") label(2 "P(y=1|mujeres)"))

```

Como puede comprobarse, la primera instrucción es análoga a la vista más arriba para generar las predicciones del modelo en función de la edad. La diferencia está en que se añade la opción `x(mujer=0)`, que lo que hace es mantener la variable *mujer* en el valor 0 al hacer la predicción. Por tanto, la predicción realizada sobre la probabilidad de asistencia a una manifestación en función de la edad en este caso sólo se refiere a los hombres. Tras esta instrucción para la predicción de los hombres, aparece otra similar para las mujeres. Después, sólo queda ponerles etiqueta y solicitar el gráfico compuesto.

**GRÁFICO 11.4. Gráfico de probabilidades predichas para distintos valores de una variable continua y otra variable discreta**



Se puede apreciar que este muestra la probabilidad predicha de asistencia a manifestaciones en función de la edad para hombres y mujeres. Ade-

más, puede advertirse cómo la diferencia entre hombres y mujeres es menor conforme avanza la edad, hasta el punto de que sus intervalos se cruzan.

Ya se ha examinado cómo se pueden estudiar las probabilidades manteniendo todas las variables en un valor constante menos una (o más), la(s) que interesa analizar. Se mencionó también que había otra manera de estudiar los resultados del logit en función de su probabilidad: a través de la predicción de la probabilidad de ocurrencia del suceso estudiado para un caso específico. La idea fundamental de esta segunda forma de hacer la interpretación del logit es el estudio de perfiles. Se especifica en una orden un perfil relevante para la investigación, dándole unos valores determinados de las variables independientes, y el programa proporciona una predicción utilizando el modelo estimado. La instrucción que se utiliza para realizar esta predicción es *prvalue* (también incluido en *SPost*). A continuación se muestra su uso a través del modelo empleado en este capítulo.

Si se desea conocer qué probabilidad de haber asistido a una manifestación adjudica el modelo a una persona joven de 25 años sin estudios, es preciso traducir este perfil en términos de variables (*edad*=25, por un lado; y *estu2=estu3=0*, por el otro, para referirse a no tener estudios). El resto de las variables se mantienen en sus valores medios, puesto que no interesan para este caso. La instrucción es *prvalue* y esta necesita una opción *x* para indicar los valores correspondientes al perfil y una opción *rest* para indicarle con qué estadístico se trabajará en el resto de las variables.

```
prvalue, x(estu3=0 estu2=0 edad=25) rest(mean)
```

Además de las probabilidades (y sus intervalos de confianza) de que ocurra (y no ocurra) el suceso estudiado, el resultado muestra cada uno de los valores de las variables independientes con los que se calculan dichas estimaciones:

#### **ILUSTTRACIÓN 11.18. Cálculo de probabilidades de ocurrencia de la variable dependiente para determinados valores de las variables independientes**

logit: Predictions for manif						
Confidence intervals by delta method						
				95% Conf. Interval		
Pr(y=1 x):		0.3393		[ 0.3035, 0.3751]		
Pr(y=0 x):		0.6607		[ 0.6249, 0.6965]		
	estu2	estu3	edad	ingr2	ingr3	mujer
x=	0	0	25	.38633306	.12133441	.51869787

La ilustración 11.18 es relativamente sencilla de comprender. Muestra la probabilidad de ocurrencia del suceso [ $Pr(y=1|x)$ ] y la de no ocurrencia

$[Pr(y=0|x)]$ . Por tanto, la probabilidad de que un joven de 25 años sin estudios haya asistido a alguna manifestación es de 0,34. Al lado aparece el intervalo de confianza de la predicción, con un 95% de probabilidades, y debajo los valores de las variables independientes que se utilizaron para realizar la predicción.

Como puede fácilmente apreciarse, esta forma de estudiar los resultados del logit es la más intuitiva y fácil de comprender. Permite estudiar cuál es la probabilidad de ocurrencia del suceso estudiado asociada a perfiles específicos de las variables independientes. El problema que tiene es que hacer un análisis detallado del efecto de todas las variables independientes es una operación muy tediosa, puesto que es necesario ir especificando uno por uno todos los perfiles. Como ejemplos (propuestos tan sólo para que el propio lector los interprete) se muestran a continuación un par de perfiles más (persona con 40 años y estudios altos; y persona de cincuenta y un años sin estudios). Las instrucciones serían las siguientes:

```
prvalue, x(estu3=1 estu2=0 edad=40) rest(mean)
prvalue, x(edad=51 estu3=0 estu2=0) rest(mean)
```

Y los resultados de aplicarlas son los que se muestran en la siguiente ilustración:

**ILUSTRACIÓN 11.19. Cálculo de probabilidades de ocurrencia de la variable dependiente para otro par de conjunto de valores de las variables independientes**

```
logit: Predictions for manif
Confidence intervals by delta method

Pr(y=1|x):      0.6312  [ 0.5877,      0.6748]
Pr(y=0|x):      0.3688  [ 0.3252,      0.4123]

      estu2      estu3      edad      ingr2      ingr3      mujer
x=        0         1       40   .38633306   .12133441   .51869787

logit: Predictions for manif
Confidence intervals by delta method

Pr(y=1|x):      0.2603  [ 0.2403,      0.2804]
Pr(y=0|x):      0.7397  [ 0.7196,      0.7597]

      estu2      estu3      edad      ingr2      ingr3      mujer
x=        0         0       51   .38633306   .12133441   .51869787
```

## 11.6. Ejercicios

1. Considerando como variable resultado el uso de Internet en los doce últimos meses (Estudio 2794: Pregunta 27), emplea el sexo, la edad (reco-dificada en tres intervalos) y los estudios en un modelo logístico binario. ¿Qué variable parece tener mayor influencia? Compara estos resultados con los obtenidos en el primer y en el tercer ejercicio del capítulo de tablas (8).
2. Empleando el barómetro de abril (cis2798), o cualquier otro de enero, abril, julio u octubre, toma como variable dependiente la intención de voto, eliminando a quienes no aporten una opción concreta, crea dos variables dicotómicas: votar al PP y votar al PSOE. Como variables independientes, se te sugiere que emplees la exposición a distintos *mass media* (P.13a-P.13c) la ideología, la edad y la religión.

## 12

# Regresión logística para variable ordinal y multinomial

El modelo de regresión logística binario, visto hasta ahora, es la base de toda una familia de modelos estadísticos de gran utilidad para las ciencias sociales, puesto que se pueden utilizar para variables dependientes cualitativas de distintos tipos y con distintos objetivos de investigación. En este capítulo se explicarán brevemente las dos extensiones más utilizadas del modelo de regresión logística binario, el logit ordinal y el logit nominal (*ologit* y *mlogit* en Stata). Existen otros modelos derivados de la regresión logística binaria que se pueden utilizar para variables cualitativas, pero su uso es mucho menos habitual, por lo que para su estudio se remite al lector interesado a un texto especializado como el de Long y Freese (2006) o el de Hosmer y Lemeshow (2000).

### 12.1. El modelo estadístico del logit ordinal

La regresión logística ordinal es una extensión de la regresión logística binaria. Por tanto, para explicar el modelo estadístico subyacente, se sigue la misma explicación utilizada en la sección 11.1 para explicar el modelo binario. La primera aproximación, por tanto, se basa en la idea de que tras la variable dependiente ordinal existe una variable latente continua; más tarde, se abordará otra aproximación basada en el concepto de cociente de razones y probabilidades no lineales.

Una variable ordinal es aquella en la que pueden ordenarse las categorías, pero se desconoce la distancia existente entre ellas (si se conociera, se trataría de una variable de intervalo o razón). Pese a que es relativamente común utilizar modelos de regresión lineal estándar para este tipo de variables, este tratamiento es inadecuado, pues vulnera los supuestos más básicos de la regresión lineal (principalmente por el hecho de que las distancias entre categorías son desconocidas y no constantes). El modelo de regresión logística ordinal es el modelo adecuado para este tipo de variables, tan comunes en las ciencias sociales.

Un ejemplo típico de variable ordinal es el grado de acuerdo con una pregunta de actitudes políticas. En este apartado utilizaremos un ejemplo

de este tipo sacado del estudio 2384 del CIS, que recoge el grado de acuerdo con la siguiente afirmación: “los partidos se critican mucho entre sí, pero en realidad todos son iguales”. Antes que nada conviene conocer la distribución de la variable.

tabulate p304

Las posibles respuestas son “muy de acuerdo”, “de acuerdo”, “en desacuerdo” y “muy en desacuerdo”. Cuando se realizó la encuesta (marzo de 2000), un 17% de la población se declaró muy de acuerdo con esta afirmación, un 45% de acuerdo, un 27% en desacuerdo y un 5% muy en desacuerdo, lo que refleja un considerable grado de desafección hacia los partidos políticos (las categorías de acuerdo y muy de acuerdo alcanzan en total un 62% del total de las respuestas). Casi un 6% de los encuestados no respondieron a la pregunta.

### **ILUSTRACIÓN 12.1. Distribución de frecuencias de la variable ordinal**

los partidos se	Freq.	Percent	Cum.
critican mucho	895	16.96	16.96
entre sí, pero en	2,350	44.54	61.50
realidad todos	1,443	27.35	88.86
son iguales	284	5.38	94.24
muy en desacuerdo	275	5.21	99.45
n.s.	29	0.55	100.00
Total	5,276	100.00	

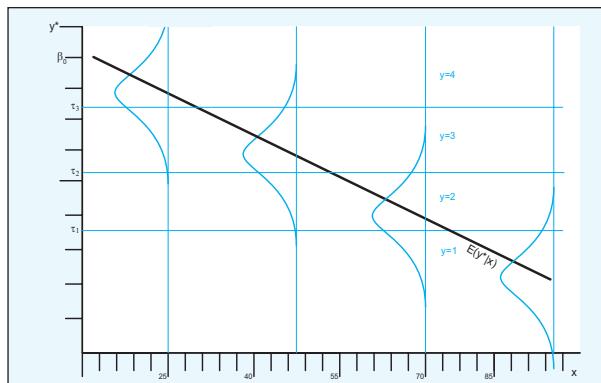
Pese a que las cuatro categorías de respuesta posible puedan recoger más o menos bien las opiniones de la gente con respecto a la pregunta p304, está claro que la opinión subyacente que se trata de captar con esta pregunta no está constituida de manera natural en estas cuatro opciones discretas y netamente diferenciadas, sino que probablemente exista un continuo mucho más diverso de opiniones desde el total acuerdo hasta el total desacuerdo. Ese continuo subyacente (que podría denominarse desafección con los partidos políticos) sería la variable latente sobre la que se construirá el modelo de regresión logística ordinal en este caso.

De manera similar a como se abordó anteriormente (gráfico 11.1) para la regresión logística binaria, el gráfico 12.1 muestra gráficamente la relación (teórica) entre la variable latente y la variable dependiente ordinal observada, con una variable independiente. De nuevo, la línea  $E(y^*|x)$  muestra la relación entre la variable latente y la variable independiente: en este

caso (puramente ilustrativo), la edad aumentaría claramente la desafección partidista.

Para la representación de la variable dependiente ordinal de desafección partidista, puede suponerse que existe una serie de *valores umbral* que permiten relacionar la variable latente continua con la variable observada ordinal. Un individuo cuyo nivel de desafección partidista sea muy bajo (en la variable latente continua) responderá con toda probabilidad “muy en desacuerdo” a la pregunta. Su nivel de desafección podría aumentar de manera moderada sin que cambiara su respuesta a esta pregunta, hasta que llegara un punto en que respondiera “en desacuerdo” en lugar de “muy en desacuerdo”. En ese momento, su nivel de desafección latente habría superado el umbral que separa la primera de la segunda categoría de respuesta en la pregunta p304. Su nivel de desafección podría seguir aumentando hasta el punto de superar el siguiente umbral y empezara a estar “de acuerdo” con la frase, o incluso podría llegar a superar el tercer y último umbral y estar “muy de acuerdo”. Estos tres niveles umbral están representados en el gráfico 12.1 por tres líneas discontinuas horizontales, etiquetadas como  $\tau_1$ ,  $\tau_2$  y  $\tau_3$ . Cuando la línea que relaciona la variable latente y la edad (en el eje de abscisas) esté por debajo del nivel  $\tau_1$ , el individuo tenderá a responder “muy en desacuerdo”, cuando esté entre  $\tau_1$  y  $\tau_2$ , tenderá a responder “en desacuerdo”, etc.

**GRÁFICO 12.1. Relación entre variable latente y variable ordinal observada con una variable independiente**



Fuente: Reelaboración a partir de Long y Freese (2006: 185).

Como en el modelo binario, la relación entre la variable latente y la variable ordinal observada es estocástica, por lo que está sometida a un cierto nivel de error representado en el gráfico 12.1 por las áreas sombreadas acopladas a cada una de las edades destacadas en la ilustración. La variable

latente de desafección política está por debajo del umbral  $t$ , para los que tienen 25 años, por lo que la predicción sería que en la mayor parte de los casos responderían “muy en desacuerdo” (aunque hay un área considerable que cae por encima de este umbral, por lo que también hay una probabilidad importante de que algunos casos respondan “en desacuerdo”). El área de probabilidad asociada a los que tienen 40 años alcanza los dos umbrales menores, por lo que la predicción sería que es muy poco probable que estén muy de acuerdo con la frase, y la probabilidad mayor sería que respondieran “en desacuerdo”, etc. Para cada uno de los niveles de umbral (o puntos de corte, como se denominarán más adelante), puede dibujarse una curva en forma de S semejante a la que se mostró para la regresión logit en el gráfico 11.1, puesto que el modelo de regresión logística ordinal es no lineal, como el de la regresión logística binaria.

Como se puede apreciar, el modelo de regresión logística ordinal es análogo al de regresión logística binario, con la importante diferencia de que en lugar de tener un único punto de corte que relaciona probabilísticamente la variable latente y la variable observada, habrá tantos puntos de corte como categorías tenga la variable ordinal (menos uno), y estos puntos de corte estarán superpuestos de manera acumulativa, como se muestra en el gráfico 12.1. De hecho, y como se verá de manera práctica más adelante, el modelo de regresión logística binario se puede entender como un modelo logístico ordinal en el que sólo hay dos categorías ordenadas.

Las ecuaciones que relacionan la variable latente y la variable ordinal observada también son análogas a las vistas más atrás para el modelo binario (ecuaciones (11.1) a (11.12)). La relación entre la variable latente y las variables independientes que se introduzcan en el modelo se puede resumir en la siguiente ecuación:

$$y_i^* = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i \quad (12.1)$$

La relación entre la variable dependiente ordinal observada y la variable latente es más compleja que en el caso de la variable binaria por la existencia de varios puntos de corte:

$$y_i = m \text{ si } \tau_{m-1} \leq y_i^* < \tau_m \quad (12.2)$$

En este caso, con cuatro categorías para la variable ordinal y, por tanto, tres puntos de corte:

$$\begin{aligned} y_i &= 1 \text{ (muy de acuerdo) si } -\infty \leq y_i^* < \tau_1 \\ y_i &= 2 \text{ (de acuerdo) si } \tau_1 \leq y_i^* < \tau_2 \\ y_i &= 3 \text{ (en desacuerdo) si } \tau_2 \leq y_i^* < \tau_3 \\ y_i &= 4 \text{ (muy en desacuerdo) si } \tau_3 \leq y_i^* < \infty \end{aligned} \quad (12.3)$$

La probabilidad de ocurrencia de cada una de las cuatro categorías para un valor de  $x$  es el área bajo las curvas sombreadas mostradas en el gráfico 12.1:

$$\Pr(y = m|\mathbf{x}) = \Pr(\tau_{m-1} \leq y_i^* < \tau_m|\mathbf{x}) \quad (12.4)$$

Sustituyendo (11.3) y despejando se llega a la fórmula de probabilidad predicha según el modelo logit ordinal:

$$\Pr(y = m|\mathbf{x}) = \Pr(\varepsilon < \tau_m - \mathbf{x}\beta) - \Pr(\varepsilon \leq \tau_{m-1} - \mathbf{x}\beta) \quad (12.5)$$

La ecuación (12.5) es equivalente a la ecuación (11.4) que se mostró para el modelo binario. Como en el modelo binario, la forma concreta que adopte el modelo depende de la distribución que se suponga para el término de error, que en el modelo logit tiene una media de 0 y varianza de  $\pi^2/3$ .

La justificación del logit ordinal mediante un modelo de probabilidad no lineal es más sencilla, y también deriva de la del modelo binario. Cada categoría de la variable dependiente ordinal se puede transformar en un cociente de razones, dividiendo la probabilidad de obtener esa categoría o una categoría menor por la probabilidad de obtener una categoría mayor. Siendo  $m$  una categoría cualquiera excepto la categoría superior de la variable dependiente:

$$\Omega_{\leq m|>m}(\mathbf{x}) = \frac{\Pr(y \leq m|\mathbf{x})}{\Pr(y > m|\mathbf{x})} \quad (12.6)$$

El modelo logit derivaría de utilizar el logaritmo neperiano de (12.6) como el lado de una ecuación de regresión para cada categoría de la variable dependiente:

$$\ln \Omega_{\leq m|>m}(\mathbf{x}) = \tau_m - \mathbf{x}\beta \quad (12.7)$$

Las fórmulas para el cálculo de cociente de razones, etc., se pueden derivar de manera análoga a como se hizo en el apartado 11.1 para el modelo binario.

## 12.2. Estimación e interpretación del modelo

Igual que en el logit binario, el modelo se estima por un procedimiento de máxima verosimilitud, por lo que no es necesario repetir cómo funciona

este procedimiento (véase apartado 11.2 más atrás). Pero, en este contexto, un ejemplo concreto de estimación de un logit ordinal permitirá entender mejor las diferencias con el modelo binario.

Se van a utilizar las mismas variables independientes que se usaron en el ejemplo anterior (el de la participación en manifestaciones) para estimar un modelo de regresión ordinal sobre la variable de desafección partidista que se han puesto como ejemplo en las páginas anteriores<sup>1</sup>. La orden de Stata para el logit ordinal es *ologit*, y se utiliza exactamente igual que la instrucción *logit*, poniendo primero la variable dependiente y detrás las variables independientes:

```
ologit dep estu2 estu3 edad ingr2 ingr3 mujer
```

Como se refleja en la ilustración 12.2, el resultado de la orden *ologit* es similar al de la instrucción *logit*, excepto en una cosa: no aparece la constante, y aparecen en su lugar tres nuevos coeficientes llamados *cut1*, *cut2* y *cut3*. Como ya habrá supuesto el lector, estos tres coeficientes corresponden a los tres valores umbral de la variable latente que se describió anteriormente. Realmente, la interpretación de estos “puntos de corte” de la regresión logística ordinal es casi idéntica a la de la constante de una regresión logística binaria. Estos tres puntos de corte representan las probabilidades acumuladas de ocurrencia de las tres categorías inferiores de la variable dependiente (la cuarta categoría no es necesario incluirla, pues la probabilidad acumulada es 1 en ese caso) cuando todas las variables independientes del modelo valen 0, aunque por supuesto expresadas en términos de logaritmos de sus cocientes de razones (*odds ratio*). El proceso de estimación del modelo empieza (la iteración cero) haciendo estos puntos de corte iguales a los porcentajes observados de respuesta de cada una de las categorías de la variable dependiente, del mismo modo que en la regresión logística binaria la estimación empieza con la constante siendo igual al porcentaje de resultados positivos en la variable dependiente. El proceso de iteración va modificando estos puntos de corte, junto con los valores de los coeficientes, buscando los valores que más verosímilmente pudieran haber producido los resultados observados (como se explica en el apartado 11.2).

---

<sup>1</sup> Se ha transformado la variable dependiente (el cuarto ítem de la tercera pregunta del estudio CIS 2384) para hacer el análisis más cómodo, eliminando los “no sabe/no contesta” (que a veces se colocan como valor intermedio, lo que no es totalmente correcto, puesto que no encajan en la métrica de la variable ordinal en cuestión), e invirtiendo los códigos otorgados en el cuestionario a las respuestas para que aquellos aumenten conforme aumenta la desafección:

```
recode p304 (1=4) (2=3) (3=2) (4=1) (8=.a) (9=.b), gen(dep)
```

### ILUSTRACIÓN 12.2. Regresión logística ordinal de la opinión sobre los partidos políticos

```

Iteration 0:  log likelihood = -4230.6463
Iteration 1:  log likelihood = -4162.4171
Iteration 2:  log likelihood = -4162.1521
Iteration 3:  log likelihood = -4162.1521

Ordered logistic regression                               Number of obs     =      3530
                                                       LR chi2(6)      =     136.99
                                                       Prob > chi2    =     0.0000
                                                       Pseudo R2       =     0.0162

Log likelihood = -4162.1521

-----+
          dep |      Coef.   Std. Err.      z   P>|z|   [95% Conf. Interval]
-----+
estu2 |  -.345338  .0846675    -4.08  0.000  -.5112833  -.1793928
estu3 |  -.7965289  .1025494    -7.77  0.000  -.9975221  -.5955358
edad |  -.006256  .0020093    -3.11  0.002  -.0101941  -.0023178
ingr2 |  -.2177021  .0731058    -2.98  0.003  -.3609868  -.0744174
ingr3 |  -.4079552  .1117292    -3.65  0.000  -.6269403  -.18897
mujer |  .1970506  .0633609    3.11  0.002  .0728655  .3212357
-----+
/cut1 |  -3.310537  .1458847
/cut2 |  -1.157784  .130701
/cut3 |  1.015081  .1305644
-----+

```

La relación entre la constante del logit binario y los puntos de corte en el logit ordinal se puede ilustrar del siguiente modo. El logit binario que se generó más atrás para la variable de asistencia a manifestaciones fue el siguiente:

### ILUSTRACIÓN 12.3. Regresión logística binaria de la asistencia a manifestaciones

```

Iteration 0:  log likelihood = -2463.521
Iteration 1:  log likelihood = -2172.6106
Iteration 2:  log likelihood = -2168.7572
Iteration 3:  log likelihood = -2168.7488

Logistic regression                               Number of obs     =      3721
                                                       LR chi2(6)      =     589.54
                                                       Prob > chi2    =     0.0000
                                                       Pseudo R2       =     0.1197

Log likelihood = -2168.7488

-----+
          manif |      Coef.   Std. Err.      z   P>|z|   [95% Conf. Interval]
-----+
estu2 |  .6304916  .0933244    6.76  0.000  .4475791  .813404
estu3 |  1.422758  .1148576   12.39  0.000  1.197642  1.647875
edad |  -.0145194  .0023696   -6.13  0.000  -.0191637  -.0098751
ingr2 |  .5353647  .0832126    6.43  0.000  .3722711  .6984584
ingr3 |  .6102257  .1249707    4.88  0.000  .3652877  .8551638
mujer |  -.4031246  .073787   -5.46  0.000  -.5477444  -.2585047
_cons |  -.3741699  .1476936   -2.53  0.011  -.663644  -.0846958
-----+

```

Si, en vez de la orden *logit*, se utiliza el *ologit* con las mismas variables...

ologit manif estu2 estu3 edad ingr2 ingr3 mujer

... el resultado es el mostrado en la ilustración 12.4:

#### **ILUSTRACIÓN 12.4. Regresión logística ordinal de la asistencia a manifestaciones**

Iteration 0: log likelihood = -2463.521	
Iteration 1: log likelihood = -2172.6106	
Iteration 2: log likelihood = -2168.7572	
Iteration 3: log likelihood = -2168.7488	
Ordered logistic regression	Number of obs = 3721
	LR chi2(6) = 589.54
	Prob > chi2 = 0.0000
	Pseudo R2 = 0.1197
Log likelihood = -2168.7488	
	-----
manif   Coef. Std. Err. z P> z  [95% Conf. Interval]	-----
estu2   .6304916 .0933244 6.76 0.000 .4475791 .813404	-----
estu3   1.422758 .1148576 12.39 0.000 1.197642 1.647875	-----
edad   -.0145194 .0023696 -6.13 0.000 -.0191637 -.0098751	-----
ingr2   .5353647 .0832126 6.43 0.000 .3722711 .6984584	-----
ingr3   .6102257 .1249707 4.88 0.000 .3652877 .8551638	-----
mujer   -.4031246 .073787 -5.46 0.000 -.5477444 -.2585047	-----
/cut1   .3741699 .1476936 .0846958 .663644	-----

Como puede apreciarse fácilmente, los resultados son absolutamente iguales, con la salvedad de que la constante se transforma en un *cutpoint* (manteniendo el mismo valor, pero cambiando el signo debido a la diferente parametrización).

La interpretación del modelo de regresión logística ordinal es similar a la de un modelo binario (y lo dicho en los apartados anteriores de este capítulo se aplica directamente), con dos diferencias importantes: primero, el modelo ordinal mostrará una serie de puntos de corte cuya interpretación es análoga (aunque obviamente no idéntica) a la de la constante en una regresión logística binaria; segundo, si el análisis se centra en la predicción de probabilidades de respuesta, se obtendrá siempre tantas probabilidades como categorías tenga la variable dependiente (en lugar de una única probabilidad de ocurrencia como se obtendría en el caso del logit binario). Si el análisis se centra en el análisis de cocientes de razones (*odds ratio*) o en el cambio marginal asociado a cada coeficiente, la interpretación del logit ordinal es prácticamente idéntica a la del logit binario.

Del mismo modo que se procedía con la regresión logística, tras la estimación de un modelo logit ordinal se puede solicitar una descripción más detallada de las variantes de los coeficientes mediante la orden *listcoef*.

```
ologit dep estu2 estu3 edad ingr2 ingr3 mujer
listcoef
```

Al mostrar los cocientes de razones (*odds ratio*) con *listcoef*, no aparecen los puntos de corte, puesto que estos cocientes son relativos, y se aplican por igual a todas las categorías. En efecto, el modelo de logit ordinal asume que el efecto de los coeficientes es el mismo (en términos relativos, recuérdese que se trata de un modelo no lineal), o que las líneas que asocian la probabilidad de ocurrencia de cada una de las categorías de la variable dependiente con las variables independientes son paralelas (en el siguiente subapartado se verá esto con más detalle).

### ILUSTRACIÓN 12.5. Coeficientes del modelo ordinal de la desafección política

ologit (N=3530): Factor Change in Odds							
	Odds of: >m vs <=m	b	z	P> z	e^b	e^bStdX	SDofX
dep							
estu2	-0.34534	-4.079	0.000	0.7080	0.8559	0.4505	
estu3	-0.79653	-7.767	0.000	0.4509	0.7464	0.3672	
edad	-0.00626	-3.114	0.002	0.9938	0.8926	18.1702	
ingr2	-0.21770	-2.978	0.003	0.8044	0.8990	0.4888	
ingr3	-0.40796	-3.651	0.000	0.6650	0.8735	0.3316	
mujer	0.19705	3.110	0.002	1.2178	1.1035	0.5000	

La interpretación de los cocientes de razones es, por tanto, idéntica a la del modelo logístico binario. La variable que tiene un mayor efecto es, claramente, el nivel de estudios: la razón de desafección política disminuye a menos de la mitad para los que tienen estudios altos con respecto a los que tienen estudios bajos, y un tercio para los que tienen estudios medios. Las mujeres tienden a una mayor desafección que los hombres, la edad tiene un efecto negativo sobre la desafección (a más edad, menos desafección política), así como los ingresos (a mayor nivel de ingresos, menor razón de desafección política). Todos los coeficientes son estadísticamente significativos, y la quinta columna (que muestra los coeficientes estandarizados, para poder compararlos) confirma que el nivel de estudios es la variable que tiene un mayor impacto (negativo) sobre la desafección política.

Como se dijo anteriormente, si se emplean las probabilidades predichas por el modelo, los resultados del logit ordinal difieren ligeramente de los del logit binario, puesto que siempre habrá que analizar varias probabilidades distintas, tantas como categorías tenga la variable dependiente. Por ejemplo, si se intenta estimar la probabilidad de desafección política para dos

perfiles distintos: uno, hombre de cuarenta años, con estudios e ingresos altos, y otro, mujer de cuarenta años, con estudios e ingresos bajos, ha de utilizarse el programa *prvalue* de *SPost* (explicado con detalle en la sección 11.5.2 del capítulo anterior).

```
prvalue, x(estu3=1 estu2=0 edad=40 ingr2=0 ingr3=1 mujer=1)
```

Como se refleja en la ilustración 12.6, la instrucción anterior genera cuatro predicciones, una para cada categoría de la variable dependiente. Por ejemplo, la probabilidad de responder “muy de acuerdo” con la frase de desafección partidista ( $y=4$ ) es de un 9% en el primer perfil, mientras que en el segundo perfil alcanza un 22% (se han escogido perfiles extremos con fines ilustrativos).

#### **ILUSTRACIÓN 12.6. Predicciones del modelo ordinal de la desafección política**

```
ologit: Predictions for dep
Confidence intervals by delta method

      95% Conf. Interval
Pr(y=1|x):      0.1138  [ 0.0898,      0.1377]
Pr(y=2|x):      0.4112  [ 0.3753,      0.4471]
Pr(y=3|x):      0.3816  [ 0.3450,      0.4183]
Pr(y=4|x):      0.0934  [ 0.0741,      0.1127]

      estu2    estu3     edad     ingr2     ingr3     mujer
x=        0       1       40         0         1         1

. prvalue, x(estu3=0 estu2=0 edad=40 ingr2=0 ingr3=0 mujer=0)

ologit: Predictions for dep
Confidence intervals by delta method

      95% Conf. Interval
Pr(y=1|x):      0.0448  [ 0.0368,      0.0528]
Pr(y=2|x):      0.2427  [ 0.2189,      0.2665]
Pr(y=3|x):      0.4924  [ 0.4747,      0.5102]
Pr(y=4|x):      0.2201  [ 0.1952,      0.2449]

      estu2    estu3     edad     ingr2     ingr3     mujer
x=        0       0       40         0         0         0
```

### **12.3. El supuesto de regresiones paralelas o razones proporcionales**

Un último apunte antes de pasar a la regresión logística multinomial. Como se ha explicado anteriormente, en el logit ordinal existe un único coeficiente para cada variable independiente. Lo que esto quiere decir es que el logit

ordinal asume que el modelo logístico que describe la relación entre las variables independientes y cada uno de los pares ordenados que se pueden formar entre categorías adyacentes de la variable dependiente es el mismo: si no fuera así, se necesitaría un modelo distinto para cada punto de corte, o para cada par posible de categorías de la variable dependiente, con distintos coeficientes para cada variable independiente (como sucede, por ejemplo, en el modelo multinomial). Este supuesto, llamado de regresiones paralelas o razones proporcionales, no siempre se cumple, y cuando no es así es necesario revisar el modelo o incluso utilizar otro para describir los datos.

Existen dos pruebas estadísticas en Stata que permiten comprobar si los datos no cumplen el supuesto de regresiones paralelas. La primera se obtiene con el programa *omodel*, que no está disponible en el paquete estándar de Stata, pero se puede descargar de Internet con la instrucción *ssc install*<sup>2</sup>:

```
ssc install omodel
```

Una vez instalada, la prueba de regresiones paralelas se ejecuta escribiendo la palabra *omodel*, seguida de *logit* (pues el modelo en cuestión es el logístico) y después la especificación del modelo del que se desea obtener información. En el ejemplo actual habría que escribir:

```
omodel logit dep estu2 estu3 edad ingr2 ingr3 mujer
```

La parte superior de su resultado simplemente repite la salida de la orden *ologit* estándar. Lo que interesa es la parte inferior, en la que se muestra una prueba de  $\chi^2$  sobre el supuesto de regresiones paralelas. El valor de  $\chi^2$  es de 26,02, que para 12 grados de libertad da una significatividad de 0,0107. El resultado de este test sugiere que los datos no se adecuan al supuesto de regresiones paralelas, puesto que el nivel de significación está por debajo del valor crítico de 0,05.

---

<sup>2</sup> *ssc* descarga paquetes y ficheros del SSC (*Statistical Software Components*) que forman parte del Archivo del *Boston College*.

**ILUSTRACIÓN 12.7. Prueba del supuesto de regresiones paralelas en la regresión logística ordinal**

```

Iteration 0:  log likelihood = -4230.6463
Iteration 1:  log likelihood = -4162.4171
Iteration 2:  log likelihood = -4162.1521
Iteration 3:  log likelihood = -4162.1521

Ordered logit estimates                                         Number of obs      =      3530
                                                               LR chi2(6)        =     136.99
                                                               Prob > chi2       =     0.0000
Log likelihood = -4162.1521                                     Pseudo R2        =     0.0162

-----+
      dep |      Coef.    Std. Err.      z   P>|z|   [95% Conf. Interval]
-----+
estu2 |   -.345338   .0846675   -4.08   0.000   -.5112833   -.1793928
estu3 |   -.7965289   .1025494   -7.77   0.000   -.9975221   -.5955358
edad |   -.006256   .0020093   -3.11   0.002   -.0101941   -.0023178
ingr2 |   -.2177021   .0731058   -2.98   0.003   -.3609868   -.0744174
ingr3 |   -.4079552   .1117292   -3.65   0.000   -.6269403   -.18897
mujer |   .1970506   .0633609   3.11   0.002   .0728655   .3212357
-----+
      _cut1 |   -3.310537   .1458847                               (Ancillary parameters)
      _cut2 |   -1.157784   .130701
      _cut3 |   1.015081   .1305644
-----+
Approximate likelihood-ratio test of proportionality of odds
across response categories:
      chi2(12) =      26.02
      Prob > chi2 =     0.0107

```

La otra prueba estadística disponible es un test de Wald diseñado específicamente para comprobar si un modelo logístico cumple el supuesto de regresiones paralelas, y forma parte del paquete de instrucciones *SPost* de Long y Freese (que se ha utilizado a menudo en este capítulo, por lo que ya debe estar instalada). Esta prueba da más información que la primera, puesto que muestra la contribución de cada coeficiente a la violación del supuesto de regresiones paralelas, lo que permite revisar el modelo y afinarlo si es necesario. La instrucción es *brant*, seguida de la opción *detail*:

```
brant, detail
```

De nuevo, lo que interesa es la parte inferior del resultado, que muestra la prueba estadístico. El valor de  $\chi^2$  de este test (el asociado a todas las variables) es muy parecido al del anterior, por lo que se confirma que el modelo es problemático. Debajo del resultado para el modelo global aparece la misma prueba para cada variable individual, que permite descubrir que sólo una variable independiente viola el supuesto de regresiones paralelas, y es el género. Como se muestra en la siguiente salida de Stata, eliminando esta variable de la ecuación, deja de incumplirse el supuesto de regresiones paralelas.

**ILUSTRACIÓN 12.8. Prueba de regresiones paralelas de Brant (A)**

```
Estimated coefficients from j-1 binary regressions

      y>1          y>2          y>3
estu2  -.5015509  -.41440488  -.1813542
estu3  -.72265037  -.87996476  -.63582219
edad   -.00963421  -.00746843  -.00308335
ingr2  -.34940201  -.28486464  -.07959308
ingr3  -.48946785  -.43298489  -.33855764
mujer   .52981587  .26463519  .01476179
_cons   3.4305133  1.2539669  -1.1783949

Brant Test of Parallel Regression Assumption

Variable |     chi2    p>chi2      df
-----+-----
All |    25.70    0.012      12
-----+-----
estu2 |    4.01    0.135      2
estu3 |    2.77    0.251      2
edad |    2.56    0.279      2
ingr2 |    4.00    0.135      2
ingr3 |    0.39    0.824      2
mujer |   11.69    0.003      2
-----+-----

A significant test statistic provides evidence that the parallel
regression assumption has been violated.
```

```
ologit dep estu2 estu3 edad ingr2 ingr3
brant, detail
```

**ILUSTRACIÓN 12.9. Prueba de regresiones paralelas de Brant (B)**

```
Brant Test of Parallel Regression Assumption

Variable |     chi2    p>chi2      df
-----+-----
All |    14.00    0.173      10
-----+-----
estu2 |    4.41    0.110      2
estu3 |    2.89    0.236      2
edad |    2.91    0.234      2
ingr2 |    4.74    0.093      2
ingr3 |    0.72    0.698      2
-----+-----

A significant test statistic provides evidence that the parallel
regression assumption has been violated.
```

**12.4. Regresión logística para variable dependiente nominal**

Por último, en este capítulo se discutirá brevemente el modelo de regresión logística multinomial. Se trata de una nueva extensión del modelo logístico.

co binario, cuya lógica es aún más sencilla que en el caso del logit ordinal (puesto que en realidad se basa en el cálculo de varios modelos binarios simultáneos), pero considerablemente más difícil de interpretar. Por ello, se hará menos énfasis en la discusión del modelo estadístico y más en la interpretación de los resultados, a través de la discusión de un ejemplo concreto.

Una variable categórica es multinomial cuando puede tomar más de dos valores pero estos valores no se pueden ordenar. Un ejemplo de variable multinomial es el recuerdo de voto o la intención de voto, por los que cada individuo puede optar por una serie de opciones distintas (partidos) que en principio no se encuentran ordenadas en función de ningún criterio. El modelo logístico binario ayuda a verificar los factores determinantes de que un individuo tome una opción frente a otra (o frente a todas las demás). Contando con esta herramienta, se podrían estimar varios modelos binarios que, uno a uno, utilizaran como variable dependiente distintas alternativas binarias de voto, tomando como referencia un partido en concreto: así, se puede elaborar un modelo explicativo de la probabilidad de votar al PSOE frente a la de votar al PP en función de los estudios, el sexo, la edad y los ingresos; luego, la probabilidad de votar a IU frente al PP con los mismos factores explicativos; más tarde, la probabilidad de voto a partidos nacionalistas frente al PP, y, por último, la probabilidad de voto a otros partidos frente al PP. En conjunto, estos cuatro modelos logísticos binarios proporcionarían una idea compleja de los factores explicativos de la opción partidista.

De modo similar a lo expresado en (11.10), cada uno de estos modelos logísticos binarios se caracterizaría por las siguientes fórmulas:

$$\begin{aligned}
 \ln \frac{\Pr(\text{PSOE}|\mathbf{x})}{\Pr(\text{PP}|\mathbf{x})} &= \mathbf{x}\beta_{\text{PSOE}/\text{PP}} \\
 \ln \frac{\Pr(\text{IU}|\mathbf{x})}{\Pr(\text{PP}|\mathbf{x})} &= \mathbf{x}\beta_{\text{IU}/\text{PP}} \\
 \ln \frac{\Pr(\text{NAC}|\mathbf{x})}{\Pr(\text{PP}|\mathbf{x})} &= \mathbf{x}\beta_{\text{NAC}/\text{PP}} \\
 \ln \frac{\Pr(\text{OTR}|\mathbf{x})}{\Pr(\text{PP}|\mathbf{x})} &= \mathbf{x}\beta_{\text{OTR}/\text{PP}}
 \end{aligned} \tag{12.8}$$

Lo que hace la instrucción *mlogit* de Stata es casi exactamente esto: estimar de manera simultánea tantos logit binarios como categorías menos una ( $j-1$ ) tenga la variable dependiente multinomial, añadiendo algunas restricciones específicas para dar cuenta del hecho de que se trata de un conjunto exhaustivo y mutuamente excluyente de elecciones (por ejemplo, los coeficientes de las comparaciones binarias deben sumar 1) y utilizando la misma

muestra para todas las comparaciones (si se solicitaran los logit binarios mostrados en (12.8) de uno en uno, cada modelo utilizaría una muestra distinta).

## 12.5. Estimación e interpretación del modelo

Utilizando la encuesta postelectoral del CIS de 2000 (estudio 2384), se recodifica la variable de recuerdo de voto en cinco categorías (PP, PSOE, IU, nacionalistas y otros), eliminando a los que no votaron y a los que no contestaron a la pregunta. Para estimar un modelo multinomial de esta variable sobre nivel de estudios, edad, ingresos y género, se emplea la instrucción *mlogit* del mismo modo que se han utilizado *logit* u *ologit*:

```
mlogit voto estu2 estu3 edad ingr2 ingr3 mujer
```

El resultado es muy similar al del logit binario, excepto en una cosa: muestra cuatro bloques de coeficientes en lugar de uno. Cada uno de los bloques de coeficientes que aparecen es un modelo logístico binario que compara la probabilidad de voto al partido mostrado al comienzo de cada bloque frente a la probabilidad de voto al PP, que es la categoría de referencia.

Por omisión, la instrucción *ologit* selecciona la primera categoría de la variable dependiente como categoría de referencia. En este caso tiene sentido (puesto que el PP fue el partido que ganó las elecciones de 2001, resulta adecuado utilizarlo como referencia para las comparaciones binarias), pero si se desea fijar otra categoría de referencia, habría que hacerlo utilizando la opción *base()* de la instrucción *mlogit*. Por ejemplo, escribiendo *base(2)*, la categoría de referencia sería el PSOE en vez del PP.

Como en los modelos binario y ordinal, la interpretación de los coeficientes del logit multinomial no es inmediata, sino que hay que recurrir a la transformación de estos coeficientes en cocientes de razones o en probabilidades. En el caso del logit multinomial, la interpretación se complica aún más al tener no un único modelo, sino tantos como categorías tenga la variable dependiente menos uno (o sea, cuatro en este caso). La interpretación de los resultados requiere el análisis simultáneo de la información contenida en todos los bloques de coeficientes.

La instrucción *listcoef* del conjunto de utilidades *Spost* (de Long y Freese)<sup>3</sup> permite una primera aproximación a la interpretación de los resultados del

---

<sup>3</sup> En el momento de escribir estas páginas, tanto la instrucción *listcoef* como *mlogplot* del módulo *SPost* no funcionaban en la versión 12 de Stata con los modelos multinomiales de la versión 12. Para poderlas ejecutar sin problemas hay que anteceder la orden *mlogit* de la instrucción *version 10*. Según los autores, se espera solucionar en una próxima revisión de estos programas *ado*.

modelo multinomial, basada en los cocientes de razones. Por defecto, esta orden, ejecutada tras un modelo multinomial, muestra el efecto de cada variable independiente sobre todas las combinaciones posibles de categorías de la variable dependiente, lo que quiere decir que en este caso la salida de *listcoef* abarcaría varias páginas.

### ILUSTRACIÓN 12.10. Regresión multinomial del voto sobre estudios, ingresos, edad y género

Iteration 0:	log likelihood = -3443.9749
Iteration 1:	log likelihood = -3355.8843
Iteration 2:	log likelihood = -3348.7016
Iteration 3:	log likelihood = -3348.5892
Iteration 4:	log likelihood = -3348.5891
Multinomial logistic regression	
Number of obs = 2745	
LR chi2(24) = 190.77	
Prob > chi2 = 0.0000	
Pseudo R2 = 0.0277	
Log likelihood = -3348.5891	
<hr/>	
voto	Coef. Std. Err. z P> z  [95% Conf. Interval]
<hr/>	
PSOE	
estu2   -.2761143 .1216149 -2.27 0.023 -.5144752 -.0377533	
estu3   -.3350645 .1536301 -2.18 0.029 -.636174 -.033955	
edad   -.0118825 .0028957 -4.10 0.000 -.017558 -.006207	
ingr2   -.4300943 .1033985 -4.16 0.000 -.6327517 -.2274369	
ingr3   -.5357216 .1666942 -3.21 0.001 -.8624362 -.209007	
mujer   .0468428 .0892647 0.52 0.600 -.1281129 .2217985	
_cons   .4292741 .1875246 2.29 0.022 .0617328 .7968155	
<hr/>	
IU	
estu2   .2634419 .2116822 1.24 0.213 -.1514476 .6783315	
estu3   .7091275 .2312174 3.07 0.002 .2559497 1.162305	
edad   -.0238329 .0054325 -4.39 0.000 -.0344804 -.0131854	
ingr2   -.0798113 .1839424 -0.43 0.664 -.4403317 .2807091	
ingr3   -.4557999 .271775 -1.68 0.094 -.9884691 .0768694	
mujer   -.4344044 .1619694 -2.68 0.007 -.7518586 -.1169502	
_cons   -.851463 .3311246 -2.57 0.010 -.1500455 -.2024708	
<hr/>	
NAC	
estu2   .1777215 .1857499 0.96 0.339 -.1863416 .5417845	
estu3   .248793 .2168983 1.15 0.251 -.1763199 .6739059	
edad   -.0052523 .0046174 -1.14 0.255 -.0143022 .0037976	
ingr2   .2194708 .1670888 1.31 0.189 -.1080172 .5469587	
ingr3   .2243626 .2354624 0.95 0.341 -.2371352 .6858604	
mujer   -.2053962 .1407716 -1.46 0.145 -.4813035 .070511	
_cons   -1.58083 .3045848 -5.19 0.000 -2.177805 -.9838552	
<hr/>	
OTR	
estu2   .0814243 .2598894 0.31 0.754 -.4279495 .590798	
estu3   .8242985 .2652403 3.11 0.002 .304437 1.34416	
edad   -.0374769 .0067912 -5.52 0.000 -.0507873 -.0241665	
ingr2   -.2593509 .2267261 -1.14 0.255 -.7027259 .186024	
ingr3   .1483741 .2768184 0.54 0.592 -.39418 .6909281	
mujer   -.2173894 .1883927 -1.15 0.249 -.5866323 .1518535	
_cons   -.7795759 .3901126 -2.00 0.046 -1.544183 -.0149693	
<hr/>	
(voto==PP is the base outcome)	

Es conveniente, por tanto, delimitar la producción de coeficientes a través de las opciones *pvalue(#)*, que hace que sólo se muestren los coeficientes que son estadísticamente significativos al nivel marcado en el paréntesis, y

*gt*, que hace que sólo se muestren las comparaciones en una dirección (si no, se muestra la misma comparación dos veces, por ejemplo PSOE frente a PP y PP frente a PSOE).

```
listcoef, pvalue(0.05) gt
```

Aun con estas opciones, la tabla (mostrada en la ilustración 12.11) sigue siendo bastante extensa, pero es mucho más manejable por haber eliminado información redundante o innecesaria. El resultado de *listcoef* para el modelo logístico multinomial es semejante a la del modelo binario, con la diferencia de que los coeficientes se refieren a las comparaciones entre las categorías mostradas en el lado izquierdo de la tabla, y los coeficientes asociados a cada variable dependiente se muestran en bloques. Por ejemplo, la tabla muestra que la probabilidad de voto al PSOE frente al PP disminuye conforme aumentan los estudios, puesto que el cociente de razones para las variables independientes dicotómicas *estu2* (estudios medios) y *estu3* (estudios superiores) tienen un valor significativo inferior a 1. La probabilidad de voto a IU frente al voto al PP o al PSOE aumenta con el nivel de estudios, así como la probabilidad del voto nacionalista o del voto a otras opciones políticas.

**ILUSTRACIÓN 12.11. Lista de coeficientes significativos  
de una regresión multinomial**

Variable: estu2 (sd=.44680195)						
Odds comparing	Alternative 1	to Alternative 2	b	z	P> z	e^b e^bStdX
PSOE -PP			-0.27611	-2.270	0.023	0.7587 0.8839
IU -PSOE			0.53956	2.460	0.014	1.7152 1.2726
NAC -PSOE			0.45384	2.321	0.020	1.5743 1.2248

Variable: estu3 (sd=.37290877)						
Odds comparing	Alternative 1	to Alternative 2	b	z	P> z	e^b e^bStdX
PSOE -PP			-0.33506	-2.181	0.029	0.7153 0.8825
IU -PSOE			1.04419	4.260	0.000	2.8411 1.4761
IU -PP			0.70913	3.067	0.002	2.0322 1.3027
NAC -PSOE			0.58386	2.504	0.012	1.7929 1.2432
OTR -PSOE			1.15936	4.187	0.000	3.1879 1.5409
OTR -PP			0.82430	3.108	0.002	2.2803 1.3599

Variable: edad (sd=17.951914)						
Odds comparing	Alternative 1	to Alternative 2	b	z	P> z	e^b e^bStdX
PSOE -PP			-0.01188	-4.103	0.000	0.9882 0.8079
IU -PSOE			-0.01195	-2.133	0.033	0.9881 0.8069
IU -PP			-0.02383	-4.387	0.000	0.9764 0.6519
NAC -IU			0.01858	2.798	0.005	1.0188 1.3959
OTR -PSOE			-0.02559	-3.696	0.000	0.9747 0.6316
OTR -NAC			-0.03222	-4.138	0.000	0.9683 0.5607
OTR -PP			-0.03748	-5.518	0.000	0.9632 0.5103

Variable: ingr2 (sd=.48991269)						
Odds comparing	Alternative 1	to Alternative 2	b	z	P> z	e^b e^bStdX
PSOE -PP			-0.43009	-4.160	0.000	0.6504 0.8100
NAC -PSOE			0.64957	3.734	0.000	1.9147 1.3747

Variable: ingr3 (sd=.3388137)						
Odds comparing	Alternative 1	to Alternative 2	b	z	P> z	e^b e^bStdX
PSOE -PP			-0.53572	-3.214	0.001	0.5852 0.8340
NAC -PSOE			0.76008	2.996	0.003	2.1385 1.2937
NAC -IU			0.68016	2.052	0.040	1.9742 1.2592
OTR -PSOE			0.68410	2.350	0.019	1.9820 1.2608

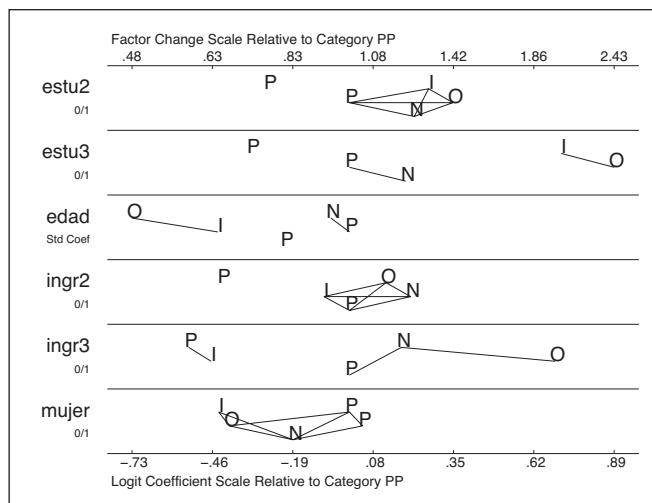
Variable: mujer (sd=.50005045)						
Odds comparing	Alternative 1	to Alternative 2	b	z	P> z	e^b e^bStdX
IU -PSOE			-0.48125	-2.872	0.004	0.6180 0.7861
IU -PP			-0.43440	-2.682	0.007	0.6477 0.8047

Pese a que la orden *listcoef* hace la interpretación del logit multinomial mucho más sencilla que la instrucción *mlogit* de Stata, sigue siendo bastante complicada por la cantidad tan enorme de coeficientes y valores que hay que tener en cuenta de manera simultánea. En el mismo conjunto de utilidades *Spost*, hay un programa específico que permite una interpretación visual mucho más sencilla de los resultados del logit multinomial. Este programa es *mlogplot*<sup>4</sup>. Tras la estimación del modelo multinomial anterior, si se introduce la siguiente línea...

```
mlogplot estu2 estu3 edad ingr2 ingr3 mujer, std(00s000) p(.1)
```

... se consigue que se representen simultáneamente todas las variables independientes del modelo (*estu2*, *estu3*, *edad*, *ingr2*, *ingr3* y *mujer*). La opción *std()* especifica qué cambio en las variables independientes quiere mostrarse en el gráfico: para variables dicotómicas se especifica el valor 0, y para *edad*, que es continua, se marca con la opción *s*, que representa el cambio en el cociente de razones asociado a un aumento de una desviación típica (hay que especificar un valor para cada variable independiente en la opción *std*). La opción *p* permite que se muestre si la diferencia entre coeficientes es significativa (con líneas, como puede apreciarse en el gráfico 12.2).

**GRÁFICO 12.2. Gráfico de distancias entre cocientes de razón de la regresión multinomial**



<sup>4</sup> La instrucción *mlogview* permite generar el mismo tipo de gráficos que *mlogplot* pero de manera interactiva, a través de un cuadro de diálogo.

En el gráfico 12.2 se muestran de manera visual los cocientes de razones asociados a cada una de las categorías de la variable dependiente (en el eje horizontal), para cada valor de las variables independientes (en el eje vertical). Cada categoría se representa a través de la primera letra de su etiqueta, con la categoría de referencia (el PP en este caso) ocupando siempre el valor de 1 (puesto que es el punto de referencia para los cocientes de razones de todas las demás categorías; de ahí se deduce que la otra P que no está en el centro siempre corresponde al PSOE). La distancia entre categorías, por tanto, refleja el impacto de cada variable independiente sobre la probabilidad de voto a cada partido político. Por ejemplo, el gráfico muestra claramente cómo la razón de voto al PSOE frente al resto de las categorías disminuye en función de los estudios. El voto a IU y otros partidos es mucho más probable para los que tienen estudios superiores que para los que tienen estudios elementales (las letras I y O están muy a la derecha). Las líneas que conectan dos partidos indican que la diferencia entre ellos *no* es significativa a un de nivel de 0,1 (el especificado en la orden): así, para los estudios superiores (segundo bloque), la diferencia entre el cociente de razones de voto al PSOE y todos los demás partidos es significativa, mientras que la diferencia entre el voto a IU y a los partidos nacionalistas o a “otros” no es significativa (puesto que están unidos por una línea).

La mayor parte de las otras técnicas para ayudar a la interpretación de los resultados del modelo logístico binario (como el uso de probabilidades, visto en el apartado 11.5 del capítulo anterior) también se pueden aplicar al modelo multinomial.

## 12.6. El supuesto de independencia de alternativas irrelevantes

En el modelo de regresión logística multinomial, las razones de ocurrencia (*odds*) de cada par de valores de la variable dependiente no deben ser afectadas por el resto de las alternativas posibles (añadir o eliminar alternativas no debe modificar los coeficientes). Se trata del supuesto de independencia de alternativas irrelevantes, que requiere que cada una de las posibles alternativas esté netamente diferenciada y sea valorada de manera independiente por el que toma la decisión. Si no se cumple este supuesto, el modelo multinomial no resulta adecuado, pues puede dar lugar a coeficientes incorrectos<sup>5</sup>.

---

<sup>5</sup> Es habitual en investigación empírica la utilización del modelo probit multinomial cuando se vulnera el supuesto de independencia de alternativas irrelevantes (en Stata, el modelo probit multinomial se puede calcular con la orden *mprobit*, pero, cuando se incumple la condición mencionada, es preferible emplear *asmprobit*). Puesto que el modelo probit asume que los errores son normales, los errores de distintas alternativas pueden estar correlacionados, y por tanto se supone que el modelo probit no se ve afectado por la vulneración de este

Es un supuesto altamente restrictivo, que es difícil que se cumpla en muchos procesos de decisión en el mundo real. Por ejemplo, ¿puede afirmarse que la opción entre dos partidos en el voto no depende del resto de los partidos existentes? La entrada de un nuevo partido de centro-izquierda en escena, por ejemplo, podría afectar a las probabilidades de voto del PSOE frente al PP. No obstante, se ha demostrado que en la mayor parte de los casos, el modelo logit multinomial es adecuado para el análisis de preferencias de voto, puesto que en la práctica el número de partidos políticos tiende a ser relativamente estable, y por tanto no se vulnera el supuesto (Dow y Endersby 2004: 112). Para tratar de entender las pautas de voto (a posteriori) para una serie de variables independientes, como en el ejemplo actual, el modelo logístico multinomial puede ser perfectamente adecuado.

La instrucción *mlogtest*, incluida en el paquete *Spost* de Long y Freese, permite la realización de un par de test estadísticos para evaluar si los datos vulneran el supuesto de independencia de alternativas irrelevantes (los test de Hausman y el de Small-Hsiao, para más detalles véanse Long y Freese 2006: p. 243 y siguientes). En esencia, estos test lo que hacen es eliminar alternativas una a una, y comprobar si los coeficientes restantes varían de manera significativa. Un resultado significativo, por tanto, obligaría a rechazar la hipótesis nula de que los datos satisfacen la restricción de independencia de alternativas irrelevantes, y por tanto debería utilizarse otro modelo para realizar el análisis (por ejemplo, volver a un modelo más simple de tipo binario).

Para realizar estas pruebas estadísticas<sup>6</sup>, se utiliza la orden *mlogtest* seguida de las opciones *hausman* y *smhsiao*.

```
mlogtest hausman smhsiao
```

---

supuesto. Pero, como explican Long y Freese (2006, p. 275), esto es incorrecto, puesto que el modelo probit multinomial también asume la independencia de alternativas irrelevantes, y los resultados son en la mayor parte de los casos idénticos a los de la instrucción *mlogit*.

<sup>6</sup> También de modo algo más complejo puede efectuarse la comprobación del supuesto de independencia de alternativas irrelevantes con la instrucción *hausman* de Stata. Para una explicación de su uso en los modelos multinomiales, véase Stata 2011: 706-714.

### ILUSTRACIÓN 12.12. Resultados de las pruebas de Hausman y Small-Hsiao

```
**** Hausman tests of IIA assumption (N=2745)

Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives.

Omitted |      chi2    df   P>chi2   evidence
-----+-----
PSOE |     0.262   21    1.000   for Ho
IU |    -0.509   19    ---    ---
NAC |     5.579   21    1.000   for Ho
OTR |    -0.059   21    ---    ---
-----+
Note: If chi2<0, the estimated model does not
meet asymptotic assumptions of the test.

**** Small-Hsiao tests of IIA assumption (N=2745)

Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives.

Omitted |  lnL(full)  lnL(omit)    chi2    df   P>chi2   evidence
-----+-----
PSOE |   -842.489  -832.172  20.634   21    0.481   for Ho
IU |   -1362.391  -1347.485 29.811   21    0.096   for Ho
NAC |   -1304.436  -1293.449 21.975   21    0.401   for Ho
OTR |   -1489.086  -1477.799 22.575   21    0.367   for Ho
-----+
```

Ambas pruebas confirman que el modelo construido no transgrede el supuesto de independencia de alternativas irrelevantes (las diferencias entre el modelo completo y los modelos a los que se ha eliminado una alternativa no son significativas).

Aunque las pruebas de la orden *mlogtest* pueden ayudar al investigador a evaluar si el modelo estimado cumple el supuesto de independencia de alternativas irrelevantes, en ocasiones pueden dar resultados contradictorios, por lo que puede ser difícil llegar a una conclusión clara. El criterio último, por tanto, es la evaluación razonada que realice el propio investigador de hasta qué punto las distintas opciones de la variable dependiente realmente están netamente diferenciadas y sus alternativas son valoradas de manera independiente.

### 12.7. Ejercicios

1. Usando el barómetro de abril de 2009 (situación de crisis económica), selecciona como variable dependiente la primera pregunta (valoración de la situación económica general). Transfórmala para que pueda ser considerada ordinal y la categoría más positiva tenga mayor puntuación. Selecciona, finalmente, como independientes al menos el estatus (con las cinco categorías que el CIS considera), la intención de voto (dos ficticias, al menos con los dos partidos principales) y otra variable que consideres relevante. Aplica un modelo logit ordinal y comenta los resultados.

2. Haz el ejercicio 2 de la página 414, pero en lugar de considerar dos variables binarias como dependientes, emplea el voto como multinomial con valores: PSOE, PP, nacionalistas, IU y otros.



# 13

## El análisis de la historia de acontecimientos con Stata

En los últimos 20 años el análisis de historia de acontecimiento (*event history analysis* en inglés) se ha aplicado de forma creciente en los estudios de sociología y de ciencia política<sup>1</sup>. En términos generales, el análisis de la historia de acontecimientos (de ahora en adelante AHA) permite investigar los factores que influyen en que suceda un acontecimiento dado. Un acontecimiento puede definirse como un cambio de tipo cualitativo de la unidad de análisis, desde el estado  $j$  al estado  $k$ , que ocurre en un momento concreto del tiempo. El ejemplo que se desarrolla en este capítulo se refiere a la transición desde la condición de parado (estado  $j$ ) a la de ocupado (estado  $k$ ).

Debido a que los límites de espacio obligan a ser muy selectivos, el objetivo de este capítulo es proporcionar una introducción simple a los fundamentos del AHA y mostrar algunas aplicaciones utilizando el programa Stata. En general han primado los aspectos aplicados sobre los detalles formales y estadísticos. En la próxima sección se proporciona un sintético compendio sobre qué es el AHA y cómo funciona. En la segunda sección se presentan las instrucciones básicas de Stata para el AHA y en la tercera las técnicas no paramétricas de análisis descriptivo. Finalmente, en la cuarta sección se describen los modelos multivariados más simples de la tasa de transición con tiempo continuo.

### 13.1. Qué es y cómo funciona el AHA

El AHA permite investigar los cambios de tipo cualitativo de la unidad de análisis que ocurren en un momento concreto del tiempo y entre un con-

---

<sup>1</sup> Tanto los temas que se tratan en este capítulo como los aspectos más sofisticados del AHA que aquí no se discuten son desarrollados de forma exhaustiva en una monografía publicada en la Colección Cuadernos Metodológicos del CIS (Bernardi 2006). El presente capítulo se basa en dicha monografía, a la que se remiten los lectores interesados para profundizar en el estudio del AHA. Para otros artículos y manuales de introducción al AHA, véase Allison (1984), Yamaguchi (1991), Strang (1994), Petersen (1995), Vermunt (1997) y Blossfeld y Rohwer (2001).

junto limitado y exhaustivo de estados. El conjunto de estados entre los cuales suceden los cambios se denomina *espacio de los estados* (Blossfeld y Rohwer 2001). Asimismo, es importante enfatizar que con las técnicas de AHA el interés reside en analizar no sólo el *tipo* de cambio, sino también *cuándo* ocurre. La propia noción de acontecimiento supone la existencia de un intervalo de tiempo anterior al cambio, desde  $j$  hasta  $k$ . El intervalo de tiempo que la unidad de análisis pasa en estado inicial ( $j$ ), antes de que suceda el acontecimiento (esto es: cuando ocurre el cambio al estado  $k$ ), se define como *episodio* o *duración*. En general, un episodio está definido por cuatro tipos de información: la fecha de inicio, la fecha de fin, el estado de origen y el estado de destino. El estado de origen  $j$  se refiere al estado que caracteriza la unidad de análisis antes de que se cumpla el acontecimiento, mientras que el estado de destino se refiere a la condición después del acontecimiento.

El tipo de proceso más simple que se puede analizar es un proceso con un *único episodio* y *dos estados*. Consideramos como ejemplo la transición desde la condición de soltero a la condición de casado por primera vez (Castro 1999). Como el primer matrimonio es un acontecimiento irrepetible, hay un único episodio para cada unidad de análisis. Además, el cambio puede suceder sólo entre dos estados: el de soltero y el de casado. Cuando los estados son más de dos, se habla de procesos *multi-estado* o con riesgos competitivos (*competing risk*). Por ejemplo, la salida del desempleo puede ocurrir tanto con una transición a la ocupación como a la inactividad. Finalmente, cuando el acontecimiento se puede repetir más de una vez para la misma unidad de análisis, el proceso se define como *multi-episódico*. Considerando de nuevo el ejemplo de la salida del desempleo, cada individuo puede caer en el desempleo más de una vez a lo largo de su carrera ocupacional y, por lo tanto, puede observarse más de un episodio de desempleo para cada individuo.

Otra distinción importante a tener en cuenta es la que se da entre datos con duraciones de tipo continuo y discreto. En el primer caso, el acontecimiento puede suceder en cualquier momento del tiempo, la duración del episodio es una variable continua y se mide con un número real positivo que en principio puede ser fraccionario. En el segundo caso, el acontecimiento sucede sólo en intervalos discretos de tiempo y la duración del episodio se mide con números enteros y positivos (1, 2, 3, etc.). Típicamente, los datos de tipo discreto se encuentran en dos situaciones. En primer lugar, cuando, aunque el acontecimiento pueda en principio suceder en cualquier momento del tiempo, las informaciones disponibles no son lo bastante precisas como para considerar las duraciones continuas. Dicho de otra manera: los datos de tipo discreto en este caso lo son por una falta de precisión en la información disponible. En segundo lugar, cuando los acontecimientos son intrínsecamente discretos, es decir, sólo suceden en momentos concretos y precisos en el tiempo. La distinción entre duración continua y discreta es

importante porque en los dos casos se precisa la aplicación de técnicas de análisis distintas. Este capítulo se centra sólo en los datos con duraciones de tipo continuo que son los más habituales en las aplicaciones de sociología y ciencias políticas<sup>2</sup>. En el cuadro 13.1 se recogen de forma sintética las definiciones de los conceptos más importantes presentados hasta aquí.

Con respecto a las técnicas tradicionales de análisis de tipo trasversal, el AHA permite tratar de forma adecuada las duraciones censuradas y especificar variables independientes que se modifican en el tiempo<sup>3</sup>. El problema de las duraciones censuradas tiene que ver con el hecho de que en la mayoría de las investigaciones las informaciones sobre las duraciones de los episodios están incompletas. El caso más típico es el de la *censura a la derecha* que ocurre cuando se conoce la fecha de inicio del episodio, pero cuando acaba el periodo de observación el acontecimiento de interés todavía no ha sucedido. Por ejemplo, se sabe que un individuo ha empezado a estar desempleado en mayo de 2005 y que sigue estando desempleado en enero de 2006, momento en el que se aplica la encuesta. El episodio de desempleo en este caso no ha terminado con un acontecimiento, y la duración correspondiente (8 meses) se considera como censurada a la derecha<sup>4</sup>. Sin entrar en más detalles, con el AHA es posible tratar las duraciones censuradas a la derecha sin sesgos en las estimaciones.

---

<sup>2</sup> Además, si los intervalos discretos del tiempo en el cual se observan los acontecimientos son pequeños, los modelos estadísticos con tiempo discreto son una aproximación a los de tiempo continuo y, en la práctica, los resultados son equivalentes. Para una discusión más profundizada del AHA para datos con duración discreta, véase Bernardi (2006).

<sup>3</sup> Hay una tercera razón de tipo estadístico para emplear el AHA que no se ilustra aquí. Se trata de la forma de las distribuciones estadísticas de las duraciones que suele violar los supuestos en la base de los modelos de regresión lineal de mínimos cuadrados (Bernardi 2006).

<sup>4</sup> Para una discusión detallada de todos los tipos de censura y los problemas a ellas asociados, véase Blossfeld y Rohwer (2001).

**CUADRO 13.1. Los conceptos básicos del AHA**

Acontecimiento	Cambio de la unidad de análisis del estado <i>j</i> al estado <i>k</i>
Episodio	Duración antes de que suceda el acontecimiento
Proceso con un solo episodio y dos estados	Proceso con un solo episodio para cada unidad de análisis y dos estados entre los que sucede el cambio (ejemplo: primer matrimonio, estados de soltero y casado)
Proceso multi-estado	Proceso con más de dos estados (ejemplo: desempleo, estados de desempleado, empleado e inactivo)
Proceso multi-episódico	Proceso en que el acontecimiento se puede repetir más de una vez para cada unidad de análisis (ejemplo: desempleo)
Tiempo continuo	El acontecimiento puede suceder en cualquier momento del tiempo
Tiempo discreto	El acontecimiento sucede sólo en momentos concretos del tiempo o las informaciones disponibles no son lo suficientemente precisas como para considerar las duraciones continuas

La otra característica fundamental del AHA es que, además de analizar específicamente la dinámica del acontecimiento investigado, permite considerar variables independientes que se modifican en el tiempo. De esta manera, el AHA consiente investigar cómo un cambio en la variable *X* en el tiempo *t* influye sobre la propensión a que suceda el acontecimiento investigado, es decir, un cambio en la variable *Y*, en un momento siguiente del tiempo *t'*. Formalmente:

$$\Delta X_t = \Delta Y_{t'} \quad (13.1)$$

Las variables *X* que se modifican en el tiempo pueden referirse tanto a características individuales de las unidades de análisis como a factores contextuales que operan a nivel macro. Por ejemplo, al estudiar la duración de los episodios de desempleo, puede investigarse cómo influyen en la transición del desempleo al empleo tanto factores macro (por ejemplo, las variaciones mensuales en la inflación o el número de puestos de trabajo creados) como factores que atañen a las características individuales de los entrevistados (por ejemplo, los cambios en su situación familiar o el fin de la percepción del subsidio de desempleo). Precisamente gracias a la posibilidad que ofrecen de evaluar el efecto del cambio de una variable independiente sobre la probabilidad de que suceda un cambio en la variable dependiente, algunos autores afirman que las técnicas del AHA

representan una nueva aproximación a la investigación empírica de las relaciones causales (Blossfeld y Rohwer 2001)<sup>5</sup>.

El concepto clave del AHA es el de la tasa de transición. Formalmente:

$$r(t)_{jk} = \lim_{t' \rightarrow t} \frac{\Pr(t \leq T < t' | T \geq t)}{t' - t} \quad (13.2)$$

... donde  $T$  es la duración antes de que suceda el acontecimiento. En otros términos,  $T$  es el tiempo que la unidad de análisis pasa en el estado de origen  $j$  hasta el momento de la transición al estado de destino  $k$ . La tasa de transición  $r(t)_{jk}$  expresa, por lo tanto, la probabilidad instantánea de que el acontecimiento ocurra en el intervalo de tiempo infinitesimal  $t'-t$ , con la condición de que el evento no haya ocurrido antes de  $t$ . La tasa de transición se interpreta como la propensión a cambiar desde el estado de origen  $j$  al estado de destino  $k$  en el momento  $t$ . La condición  $T \geq t$  significa que esta propensión al cambio es definida con respecto al conjunto de unidades de análisis todavía en riesgo de experimentar el acontecimiento en el tiempo  $t$ , es decir, el conjunto de unidades de análisis cuya duración es mayor o igual a  $t$ . Formalmente, la tasa de transición no puede ser interpretada como una probabilidad porque puede asumir valores mayores que 1. Sin embargo, si el intervalo de tiempo  $t'-t$  es pequeño, entonces:

$$\Pr(t \leq T < t' | T \geq t) \approx (t' - t)r(t)_{jk} \quad (13.3)$$

La tasa de transición se aproxima, en este caso, a la probabilidad condicional de que el acontecimiento ocurra en intervalo  $t'-t$  (Blossfeld y Rohwer 2001: 37). Es importante resaltar que la tasa de transición definida en la fórmula (13.2) contiene dos tipos de informaciones: la calidad del cambio desde  $j$  a  $k$  y la duración  $T$  antes de que el cambio ocurra. La tasa de transición está además relacionada con otros dos importantes conceptos estadísticos: la función de supervivencia y la función de densidad.

La función de la distribución  $F(t)$  describe la probabilidad de que la duración de un episodio sea menor o igual a  $t$ . Dicho con otras palabras, la probabilidad de que un acontecimiento ocurra en el intervalo de 0 a  $t$ . Formalmente:

---

<sup>5</sup> Ilustrar el funcionamiento de la “técnica de partición del episodio” (*episode splitting*) que permite definir las variables que se modifican en el tiempo requeriría mucho más espacio que el de que se dispone. Para diferentes ejemplos de definición de variables que se modifican en el tiempo con Stata, véase el capítulo 5 en Bernardi (2006).

$$F(t) = \Pr(T \leq t) \quad (13.4)$$

Además, se define la función de supervivencia  $G(t)$  como la función complementaria de  $F(t)$ :

$$G(t) = \Pr(T > t) = 1 - F(t) \quad (13.5)$$

La función de supervivencia  $G(t)$  describe la probabilidad de que la duración de un episodio sea como mínimo igual a  $t$ . Dicho de otro modo, la probabilidad de que la unidad de análisis haya sobrevivido en el estado  $j$  hasta el tiempo  $t$ . En el ejemplo de la salida del desempleo,  $G(t)$  expresaría la probabilidad de seguir estando desempleado en el tiempo  $t$ .

Finalmente, la función de densidad  $f(t)$  describe la probabilidad instantánea incondicional de que un acontecimiento ocurra en el intervalo de tiempo infinitesimal  $t'-t$ :

$$f(t) = \lim_{t' \rightarrow t} \frac{\Pr(t \leq T < t')}{t' - t} \quad (13.6)$$

Es importante destacar la diferencia entre la tasa de transición y la función de densidad. En la función de densidad la probabilidad de que ocurra el acontecimiento no está condicionada a la supervivencia hasta el tiempo  $t$ , mientras que en la tasa de transición la probabilidad se computa sólo con respecto a las unidades de análisis que se han quedado en riesgo de experimentar el acontecimiento. Existe, además, una relación entre la tasa de transición, la función de densidad y la función de supervivencia, de tal manera que (Blossfeld y Rohwer 2001: 36):

$$r(t) = \frac{f(t)}{G(t)} \quad (13.7)$$

Por último, la idea central del AHA es considerar la tasa de transición  $r(t)$  como la variable dependiente y definir un modelo de la tasa de transición de la siguiente manera:

$$r(t)_{jk} = f(\beta X, q(t)) \quad (13.8)$$

Con un modelo de la tasa de transición se estudia cómo la propensión a pasar desde el estado  $j$  al estado  $k$ , es decir, la propensión de que ocurra el

acontecimiento, varía en función de un conjunto de variables independientes  $X$  y de una función  $q(t)$  del tiempo. Los coeficientes  $\beta$  expresan la influencia de variables explicativas  $X$  sobre la tasa de transición. Volviendo al ejemplo de la salida del desempleo, con un modelo de la tasa de transición podría estudiarse cómo la propensión para encontrar un trabajo depende del género, del nivel de educación, del estado civil y de las condiciones macro del mercado de trabajo (variables  $X$ ), así como de la duración misma del episodio de desempleo (función  $q(t)$ ).

En definitiva, el modelo de la tasa de transición resume toda la lógica del AHA. En primer lugar, el análisis se centra en las dinámicas del cambio de una condición a otra a lo largo del tiempo. Así, el objetivo es explicar la propensión a que dicho cambio ocurra. Para ello se analiza cómo la propensión al cambio depende de un conjunto de variables independientes, en particular variables que se modifican en el tiempo, y de la propia duración del proceso.

### 13.2. El AHA con Stata: instrucciones para definir los datos

En Stata todas las instrucciones para analizar datos de historias de acontecimientos empiezan con el prefijo *st*, que es la abreviatura de *survival time*. Para poder utilizar las órdenes que empiezan por *st* es necesario que los datos hayan sido definidos previamente como datos de historia de acontecimientos. La instrucción para definir los datos como historia de acontecimientos en Stata es *stset*. Para ilustrar cómo funciona la instrucción *stset* se lanza a continuación un ejemplo concreto que se refiere a la duración de episodios de desempleo. En este caso, el acontecimiento investigado es la transición desde el desempleo hasta la ocupación. El fichero de datos se llama *unemployment.dta*. Los episodios de los sujetos que todavía están desempleados en el momento de la entrevista se consideran censurados a la derecha. En la ilustración 13.1 se presentan algunos episodios del fichero de datos considerado y la descripción de las variables independientes<sup>6</sup>.

---

<sup>6</sup> Para una descripción más detallada del fichero de datos, véase Bernardi (2006).

**ILUSTRACIÓN 13.1. Variables y algunos episodios del fichero de datos relativos a las duraciones de los episodios de desempleo (*unemployment.dta*)**

Variable	Descripción							
<hr/>								
id	Número de identificación							
org	Estado de origen (0=desempleado; todos los episodios tienen como estado de origen 0)							
dest	Estado de destino (0=censurado a la derecha, es decir desempleado en el momento de la entrevista, 1=empleado, 2=inactivo)							
begin	Fecha de inicio en meses del siglo							
end	Fecha de fin en meses del siglo							
dateint	Fecha de la entrevista en meses del siglo							
sex	Género (1=varón, 0=mujer)							
dbirth	Fecha de nacimiento en meses del siglo							
cohort	Cohorte de nacimiento (1=nació antes de 1940, 2=nació entre 1940 y 1959, 3=nació después de 1959)							
coh2	Cohorte 1940-1960 (1 si cohort=2, 0 si no)							
coh3	Cohorte >1960 (1 si cohort=3, 0 si no)							
<hr/>								
id	org	dest	begin	end	dateint	sex	dbirth	cohort
1000031	0	1	953	967	1168	0	708	2
1000031	0	1	1063	1099	1168	0	708	2
1000031	0	0	1167	1168	1168	0	708	2
1000043	0	0	1101	1171	1171	0	797	3
1000051	0	1	982	990	1170	1	702	2
1000061	0	0	1170	1171	1171	1	662	2
1000062	0	0	1170	1171	1171	0	669	2
1000071	0	1	982	992	1170	1	587	2
1000112	0	2	928	934	1169	0	628	2

La variable *begin* se refiere a la fecha de inicio del episodio de desempleo y la variable *end* a su fin. Todas las variables que se refieren a fechas (*begin*, *end*, *dateint* y *dbirth*) están codificadas en meses del siglo, que indican el número de meses transcurridos desde el principio del siglo XX. Así, el valor 1, en meses del siglo, significa enero de 1900, 2 febrero de 1900 y así sucesivamente. En general, la fórmula para pasar de una codificación en meses y años a una en meses del siglo es:

$$\text{meses del siglo} = (\text{año} - 1900) \times 12 + \text{mes} \quad (13.9)$$

Por ejemplo, a la fecha abril de 1980 le corresponde en meses del siglo el valor 964. Las fórmulas para la conversión inversa, de meses del siglo a mes y año, son:

$$\begin{aligned} \text{año} &= (\text{meses del siglo})/12 + 1900 \\ \text{mes} &= \text{mod}(\text{meses del siglo}, 12) \end{aligned} \quad (13.10)$$

Por ejemplo, la fecha 982 en meses del siglo corresponde a octubre del 1981<sup>7</sup>. El proceso investigado en este ejemplo es multi-episódico y multi-

<sup>7</sup> Esto porque  $982/12 = 81$  y resta 10, que corresponde al mes de octubre.

estado. Es multi-episódico porque cada entrevistado puede haber tenido más de un episodio de desempleo. Por ejemplo, el sujeto con el número de identificación igual a 1000031 tiene tres episodios de desempleo, los dos primeros terminan con una transición a la ocupación, como indica la variable *dest* con valor 1, mientras que el tercero está censurado a la derecha, como indican las variables *dest* con valor 0 y *end* igual a la fecha de la entrevista. Además, el proceso es multi-estado, ya que la salida del desempleo puede ocurrir bien con una transición a la ocupación, bien con la inactividad. Fíjémonos en el sujeto con el número de identificación igual a 1000112. Su episodio de desempleo empieza en la fecha 928 y acaba en la fecha 934 con una transición a la inactividad, como indica la variable *dest* con valor 2.

Como se dijo anteriormente, la instrucción básica para definir los datos como historia de acontecimiento en Stata es *stset*. Su formulación más simple, la siguiente:

```
stset vartemporal, failure(varinterrup)
```

... donde *vartemporal* es la variable que indica la duración antes de que ocurra el acontecimiento o antes de la censura a la derecha; *varinterrup* es la variable que indica si el episodio termina con un acontecimiento o si está censurado a la derecha. Si se omite *failure(varinterrump)*, Stata considera que todos los episodios acaban con un acontecimiento. Cuando la variable definida como *varinterrup* es mayor que 0, Stata interpreta que los episodios acaban con un acontecimiento; si es igual a 0 o inválido, Stata considera los episodios como censurados a la derecha.

Otra posibilidad es utilizar la opción *failure(varinterrup== numlist)*. Con esta opción Stata interpreta que acaban con un acontecimiento todos los episodios para los cuales la variable definida *varinterrup* es igual a uno de los valores numéricos especificados en la lista *numlist*. Todos los demás episodios son considerados como censurados a la derecha.

Existen muchas otras opciones en *stset* que permiten tratar con estructuras de datos muy complejas<sup>8</sup>. En particular, en los ejemplos de este libro se utilizarán las opciones *origin(nombrevar)* e *id(nombrevar)*:

```
stset vartemporal, failure(varinterrup) origin(nombrevar) id(nombrevar)
```

La opción *origin(nombrevar)* se utiliza para especificar una variable que identifica la fecha de inicio del episodio. Esta opción es útil cuando la fecha

---

<sup>8</sup> Para la gama completa de opciones, véase Stata (2009g).

de inicio no es igual a 0. La opción *id(nombrevar)* se utiliza para indicar que un proceso es multi-episódico<sup>9</sup>.

Para definir las duraciones de los episodios de desempleo como datos para el AHA se precisan las dos instrucciones siguientes: la primera para cargar los datos en la memoria y la segunda para definirlos:

```
use unemployment, clear
stset end, origin(begin) fail(dest==1)
```

Se utiliza la opción *origin( )* para definir que el episodio inicia en la fecha indicada por la variable *begin*. Con la opción *fail(dest==1)* se especifica que la transición de interés es la transición a la ocupación y que la transición a la inactividad tiene que ser considerada como equivalente a una censura a la derecha. En la ilustración 13.2 se presentan los resultados de la aplicación de la instrucción *stset*. De los 973 episodios de desempleo, 734 terminan con una transición a la ocupación. Además, para el sujeto con el número de identificación igual a 1000112 que sale del desempleo en la fecha 934 con una transición a la inactividad (*dest=2*), la variable *\_d* es igual a 0. Si también fuera interesante investigar la transición a la inactividad con el fin de comparar los mecanismos que, desde la condición de desempleado, determinan la salida del mercado de trabajo o la vuelta a la ocupación, sería necesario volver a definir los datos como historias de acontecimientos utilizando *stset*, pero esta vez especificando que el acontecimiento sucede cuando la variable *dest* es igual a 2<sup>10</sup>. La línea de instrucción correspondiente sería:

```
stset end, origin(begin) fail(dest==2)
```

En este caso, el número de acontecimientos es igual a 80 y las transiciones a la ocupación son consideradas como censuras a la derecha. Así, la variable *\_d* es igual a 0 para el episodio del sujeto con el número de identificación igual a 1000071.

Si no se hubiera especificado la opción *fail( )*, utilizando la siguiente instrucción:

---

<sup>9</sup> La opción *id(varname)* es además imprescindible antes de efectuar una subdivisión de los episodios (*episode splitting*) para definir variables que se modifican en el tiempo. Sobre este tema véase Bernardi (2006).

<sup>10</sup> Recuérdese que en la ilustración 13.2 se especifica que las categorías de la variable *dest* son: 0=censurado a la derecha, 1=empleado y 2=inactivo.

```
stset end, origin(begin)
```

... Stata habría considerado que todos los episodios acababan con un acontecimiento, sin distinguir entre episodios censurados a la derecha, transiciones a la ocupación y transiciones a la inactividad. Si, por otro lado, se hubiera escrito:

```
stset end, origin(begin) fail(dest)
```

### ILUSTRACIÓN 13.2. Definición de *unemployment.dta* como fichero apto para el AHA

```
*acontecimiento definido como transición a la ocupación

failure event: dest == 1
obs. time interval: (origin, end]
exit on or before: failure
t for analysis: (time-origin)
origin: time begin
-----
973 total obs.
0 exclusions
-----
973 obs. remaining, representing
734 failures in single record/single failure data
17922 total analysis time at risk, at risk from t = 0
earliest observed entry t = 0
last observed exit t = 119
list id org dest begin end _d _t _to in 1/9, nod nol noob clean
      id   org   dest   begin     end   _d   _t   _to
    1000031     0     1     953     967     1    14     0
    1000031     0     1    1063    1099     1    36     0
...
    1000071     0     1     982     992     1    10     0
1000112     0     2     928     934     0     6     0

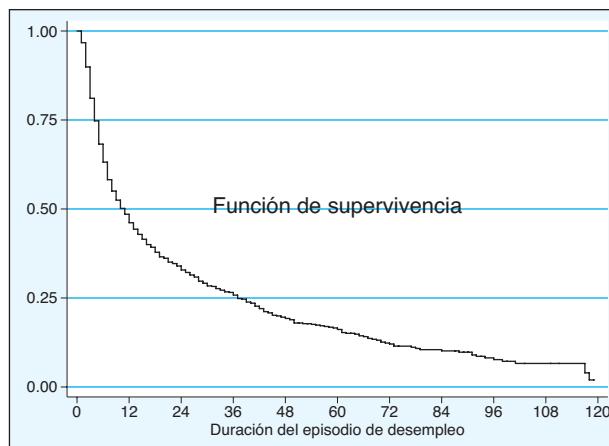
*acontecimiento definido como transición a la inactividad
failure event: dest == 2
obs. time interval: (origin, end]
exit on or before: failure
t for analysis: (time-origin)
origin: time begin
-----
973 total obs.
0 exclusions
-----
973 obs. remaining, representing
80 failures in single record/single failure data
17922 total analysis time at risk, at risk from t = 0
earliest observed entry t = 0
last observed exit t = 119
list id org dest begin end _d _t _to in 1/9, nod nol noob clean
      id   org   dest   begin     end   _d   _t   _to
    1000031     0     1     953     967     0    14     0
    1000031     0     1    1063    1099     0    36     0
...
1000071     0     1     982     992     0    10     0
    1000112     0     2     928     934     1     6     0
```

... Stata habría interpretado que todos los episodios con variable *dest* mayor de 0 y que no fueran casos perdidos terminan con un acontecimiento, sin distinguir entre las transiciones a la ocupación y las transiciones a la inactividad. Finalmente, cabe destacar que el proceso ha sido definido con un solo episodio. En otras palabras, cada episodio es analizado por su cuenta, independientemente de otros eventuales episodios de desempleo del mismo sujeto. Para definir los datos como multi-episódico se puede utilizar la opción *id* en la orden *stset*.

### 13.3. La función de supervivencia

El gráfico 13.1 presenta la función de supervivencia para la salida del desempleo y muestra la proporción de episodios de desempleo (eje vertical) que todavía no han terminado con una transición a la ocupación, en función de la duración de los episodios (eje horizontal). Puede así comprobarse que más de la mitad de los episodios de desempleo acaban con una transición a la ocupación durante el primer año, pero uno de cada cuatro tiene una duración superior a los 36 meses.

**GRÁFICO 13.1. Función de supervivencia para la salida del desempleo**



Para estimar las funciones de supervivencia se puede utilizar el estimador de Kaplan y Meier. Sin entrar en los detalles formales (para ellos véase Blossfeld y Rowher 2001), la instrucción de Stata para calcular una función de supervivencia con el estimador de Kaplan y Meier es *sts*. La orden *sts graph* produce un gráfico con la función de supervivencia, mientras que *sts list* muestra un listado con los valores de la misma. Para poder ejecutar esta orden es necesario que los datos se hayan definido previamente como

historias de acontecimientos mediante la instrucción *stset*. Si se quieren comparar las funciones de supervivencia para diferentes grupos, es decir, para diferentes valores de una variable independiente, las instrucciones que se precisan son:

```
sts graph, by(varlist)
sts list, by(varlist)
```

... donde *varlist* es una lista de variables de tipo categórico. Las instrucciones en Stata para la estimación de las funciones de supervivencia han de ser.

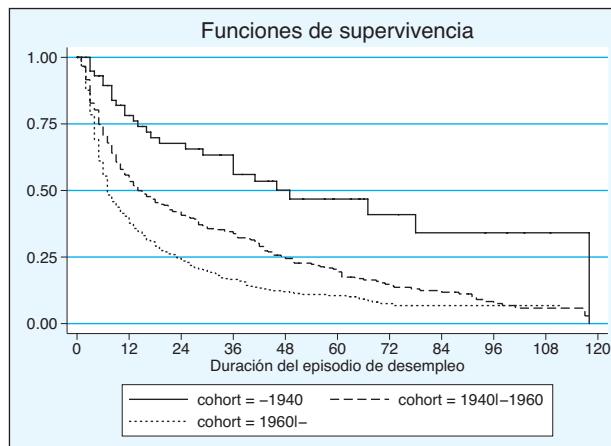
```
use unemployment.dta, clear
stset end, origin(begin) fail(dest==1)
sts graph, ylabel(, angle(horizontal)) xlabel(0 (12) 120) ///
    xtitle(Duración del episodio de desempleo) ///
    title(Función de supervivencia, position (0) ring(0))
    name(I4, replace)
sts graph, ylabel(, angle(horizontal)) xlabel(0 (12) 120) ///
    xtitle(Duración del episodio de desempleo) ///
    title(Funciones de supervivencia) ///
    by(cohort) name(I5, replace)
```

El gráfico 13.2 presenta las funciones de supervivencia para la duración de los episodios de desempleos para tres cohortes de nacimiento. La función de supervivencia para la cohorte más joven (los nacidos en 1960 o después) está por debajo de las otras funciones. Esto significa que los más jóvenes salen más rápidamente del desempleo que los demás<sup>11</sup>.

---

<sup>11</sup> La evolución con “saltos” y la brusca caída de la función de supervivencia de la cohorte más anciana al final del intervalo de observación se deben a su escaso número de casos.

**GRÁFICO 13.2. Función de supervivencia para la salida del desempleo para tres cohortes de edad. Modelos de transición con tiempo continuo**



### 13.4. Modelos de la tasa de transición con tiempo continuo

En términos generales, un modelo de la tasa de transición puede ser especificado como:

$$r(t)_{jk} = f(\beta X_t, q(t)) \quad (13.11)$$

Así, se estudia la tasa de transición en función de un vector de variables independientes  $X_t$  y de la duración  $t$  del proceso. Los coeficientes  $\beta$  expresan el efecto de  $X_t$  sobre la tasa de transición y son los factores que interesa estimar.

Entre los modelos de la tasa de transición, la especificación más común de la ecuación general (13.11) es:

$$r(t)_{jk} = \exp(\beta X_t) q(t) \quad (13.12)$$

La ecuación (13.12) describe los modelos proporcionales de la tasa de transición. La forma funcional de la relación entre las variables  $X_t$  y la tasa de transición es exponencial porque la tasa de transición no puede asumir valores negativos. Además, respecto a la formulación general de la ecuación (13.11), la función  $q(t)$  en (13.12) indica que la pauta de

dependencia temporal del proceso es igual para todas las observaciones. En términos más técnicos, esto equivale a decir que no hay efectos de interacción entre las variables  $X_t$  y el tiempo  $t$ . En otras palabras, el tiempo no condiciona el efecto de las variables  $X_t$  sobre la tasa de transición. Estos modelos son denominados “proporcionales” porque se basan en el supuesto de que los efectos de las variables  $X_t$  inducen sólo a desplazamientos proporcionales de  $q(t)$  hacia arriba o hacia abajo, sin modificar su forma.

Para estimar los coeficientes  $\beta$  es necesario formular un supuesto sobre la forma de  $q(t)$ . Con este fin se puede elegir entre varias distribuciones paramétricas. Entre las distribuciones más comunes para analizar datos de duraciones se encuentran las distribuciones exponencial, Gompertz, Weibull y log-logística y exponencial constante a intervalos. En el gráfico 13.3 se presentan ejemplos de gráficos de la tasa de transición para algunas de estas distribuciones. El modelo de la tasa de transición más simple es el exponencial y supone que  $q(t)$  es constante en el tiempo<sup>12</sup>:

$$r(t)_{jk} = a = \exp(\beta X_t) \quad (13.13)$$

El supuesto de constancia de la tasa de transición implica que el riesgo de que ocurra el acontecimiento no varía en función del tiempo. Esto significa que el proceso no “tiene memoria”: la verosimilitud de que ocurra el acontecimiento es la misma justo al principio del proceso como en momentos posteriores del tiempo.

La tasa de transición del modelo Gompertz se expresa como:

$$r(t)_{jk} = a \exp(bt) \quad (13.14)$$

Este modelo implica que la tasa de transición es monotónica creciente si el parámetro  $b$  es mayor que 0 o monotónica decreciente si  $b$  es menor que 0. Si  $b$  es igual a 0, el modelo Gompertz equivale al modelo exponencial. La solución más común para especificar los efectos de las variables independientes es introducirlo a través del parámetro  $a$ , con  $a=\exp(\beta X)$ . La misma notación se utilizará también para los otros modelos paramétricos. De este modo el parámetro  $b$  expresa la forma de dependencia temporal del modelo, y el parámetro  $a$  se utiliza para estimar los efectos de las variables independientes. En el lenguaje estadístico, el parámetro  $a$  se define como parámetro secundario (*ancillary*).

---

<sup>12</sup> La presentación de los modelos considerados se limita a sus propiedades más generales.

El modelo Gompertz de la ecuación (13.14) es de tipo proporcional. Como ya se ha mencionado poco antes, el supuesto de este modelo es que el efecto de las variables  $Xt$  no se modifica a lo largo del intervalo temporal considerado. Dicho de otra forma, que el efecto de las variables  $Xt$  se traduce en un desplazamiento (hacia arriba, si el efecto es positivo, o hacia abajo, si el efecto es negativo) de la pauta de dependencia temporal controlada por el parámetro  $b$ , pero no influye en su forma. Este modelo se ha utilizado para estudiar los acontecimientos de movilidad ocupacional, demostrando que la tasa de transición de un trabajo  $j$  a otro trabajo  $k$  disminuye en función del tiempo transcurrido en el mercado de trabajo (Sørensen y Tuma 1981).

Otro modelo empleado frecuentemente en las aplicaciones del AHA en las ciencias sociales es el modelo de Weibull. Por ejemplo, Olzak (1992) empleó este tipo de modelo en un estudio clásico sobre la acción colectiva para analizar cómo el tiempo transcurrido desde el último evento de protesta influye sobre la probabilidad de que suceda el siguiente. Asimismo, Carroll y Hannan (2000) lo utilizaron para estudiar el riesgo de quiebra de las corporaciones e industrias en función del tiempo desde su fundación. La tasa de transición para el modelo Weibull es igual a:

$$r(t)_{jk} = abt^{b-1} \quad (13.15)$$

Este modelo implica una tasa de transición monotónica creciente si el parámetro  $b$  es mayor que 1 o monotónica decreciente si el parámetro  $b$  es menor que 1. Si  $b$  es igual a 1 equivale al modelo exponencial. La solución más común es introducir el efecto de las variables independientes a través del parámetro  $a$ , con  $a=\exp(\beta X)$ . Se obtiene así un modelo Weibull de tipo proporcional.

Un ulterior modelo paramétrico que, sin embargo, no pertenece a la familia de los modelos proporcionales de la ecuación (13.12), es el modelo log-logístico. Aplicaciones de esto tipo de modelo se encuentran en los estudios de las dinámicas familiares (Diekmann 1989 y 1992) y de la demografía de las organizaciones (Carroll y Hannan 2000). La tasa de transición del modelo log-logístico es<sup>13</sup>:

$$r(t)_{jk} = \frac{ba^b t^{b-1}}{1 + (at)^b} \quad (13.16)$$

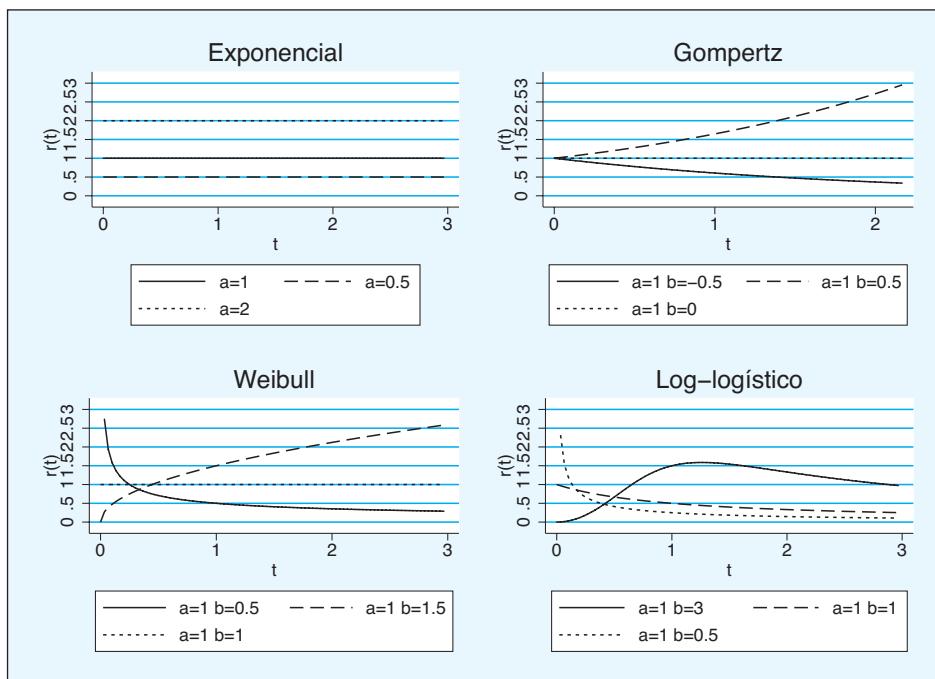
Los efectos de las variables independientes se especifican a través del parámetro  $a$ , con  $a=\exp(\beta X)$ . Este modelo es más flexible que los anteriores, ya que si  $b$  es menor o igual a 1, la tasa de transición es monotónicamente decreciente, mientras que si es mayor que 1, tiene una forma de campana (gráfico 13.3d).

---

<sup>13</sup> Existen en literatura parametrizaciones alternativas para este modelo. Aquí se sigue la parametrización utilizada en Blossfeld y Rohwer (2001).

Finalmente, el modelo exponencial constante a intervalos es una especificación del modelo exponencial simple. El intervalo temporal se divide en varios intervalos pequeños, se supone que la tasa de transición es constante en cada intervalo y que puede variar de un intervalo a otro. Formalmente,  $q(t)$  consiste en una serie de variables dicotómicas  $d_1, d_2, \dots, d_n$  con  $d_i=1$  en el intervalo temporal  $t_{i-1} \leq t < t_i$ ,  $d_2=1$  en el intervalo temporal  $t_1 \leq t < t_2$ , ...  $d_n=1$  en el intervalo temporal  $t_{n-1} \leq t < t_n$ . En este caso, el investigador tiene que elegir el número y la extensión de los intervalos en los cuales se divide el eje temporal del proceso, pero no tiene que formular ningún supuesto sobre la forma funcional de  $q(t)$ . Una vez estimado el modelo, los coeficientes relativos a las variables dicotómicas que identifican los intervalos permiten reconstruir la pauta de dependencia temporal del proceso. Precisamente, por esta flexibilidad y por no requerir supuestos a priori sobre la forma de la dependencia temporal del proceso, el modelo exponencial constante a intervalos se ha convertido en la elección más común entre los modelos de la tasa de transición para realizar un AHA.

**GRÁFICO 13.3. Ejemplo de gráficos de la tasa de transiciones**



### 13.4.1. Modelos de la tasa de transición con tiempo continuo con Stata

El modo de estimar modelos paramétricos de la tasa de transición con tiempo continuo es utilizando *streg*. Para su uso, se precisa haber definido previamente los datos como episodios con la instrucción *stset*. La sintaxis básica de la orden *streg* es la siguiente:

**streg** [varlist], **dist(distname)**

... donde *varlist* indica la lista de variables independientes *Xt* y *distname* especifica la distribución de la tasa de transición a estimar. De este modo, se puede elegir entre la distribución exponencial, Weibull, Gompertz, log-normal, log-logístico o gamma. Existe, además, un amplio abanico de opciones para *streg*. En lo que sigue se tratará de explicar aquellas que nos parecen más útiles para realizar un AHA. Tanto en la ilustración 13.3 como en la ilustración 13.4 se presentan los resultados de la estimación de un modelo exponencial y de un modelo Gompertz con las variables relativas al género (*sex*) y las cohortes de pertenencia (*coh2* y *coh3*, siendo *coh1* la categoría de referencia) para la salida del desempleo.

streg sex coh2 coh3, dist(exp) nohr

En el modelo 1 la opción *dist(exp)* especifica que la tasa de transición es de tipo exponencial, mientras que con la opción *nohr* se muestran los coeficientes de los efectos de las variables y no las ratios de las tasas de transición (en breve, se explica en qué consisten estos últimos). Si se considera, por ejemplo, la variable *sex* (igual a 1 para los hombres y a 0 para las mujeres), el coeficiente estimado (0,26) es positivo y estadísticamente significativo. Por eso, se puede concluir que la propensión a realizar la transición a la ocupación es mayor para los hombres que para las mujeres. En el caso de los modelos proporcionales de la tasa de transición, el efecto de una variable puede ser interpretado como la variación porcentual de la tasa, si todas las demás variables permanecen constantes y sólo se modifica la variable considerada.

Formalmente:

$$\Delta r = (\exp(\beta_i)^{\Delta X_i} - 1) \times 100 \quad (13.17)$$

... donde  $\Delta X_i$  corresponde a un cambio en los valores de la variable  $X_i$ ,  $\Delta r$  es la variación porcentual en la tasa de transición asociada a dicho cambio y  $\beta_i$  es el coeficiente estimado para la variable  $X_i$ . Si la variable  $X_i$  es dicotómica, como en este caso,  $\Delta X_i=1$  y...

$$\Delta r = (\exp(\beta_i) - 1) \times 100 \quad (13.18)$$

### ILUSTRACIÓN 13.3. Modelo exponencial y Gompertz para las duraciones de los episodios de desempleo (modelo 1)

* modelo 1
failure _d: dest == 1
analysis time _t: (end-origin)
origin: time begin
Iteration 0: log likelihood = -1546.1636
Iteration 1: log likelihood = -1491.1689
Iteration 2: log likelihood = -1486.9766
Iteration 3: log likelihood = -1486.9488
Iteration 4: log likelihood = -1486.9488
Exponential regression -- log relative-hazard form
No. of subjects = 973 Number of obs = 973
No. of failures = 734
Time at risk = 17922
LR chi2(3) = 118.43
Log likelihood = -1486.9488 Prob > chi2 = 0.0000
-----
_t   Coef. Std. Err. z P> z  [95% Conf. Interval]
-----+-----
sex   .2575632 .0739688 3.48 0.000 .1125871 .4025394
coh2   .8885871 .1985 4.48 0.000 .4995343 1.27764
coh3   1.445604 .1952503 7.40 0.000 1.06292 1.828287
_cons   -4.429246 .1947556 -22.74 0.000 -4.81096 -4.047532
-----

Si se comparan los hombres y las mujeres, la variación en la tasa de transición a la ocupación es igual a:  $(\exp(0,26)-1) \times 100\% = (1,3-1) \times 100\% = 30\%$ . Por lo tanto, se puede concluir que la tasa de transición al trabajo es un 30% mayor para los hombres que para las mujeres. También se puede calcular la ratio de las tasas de transición de los hombres,  $r(h)$ , y de las mujeres,  $r(m)$ :

$$\frac{r(h)}{r(m)} = \exp(0,26) = 1,30 \quad (13.19)$$

De esto modo, la tasa de transición a la ocupación de los hombres es 1,3 veces la de las mujeres. O, dicho sin números, los hombres salen del

desempleo con bastante más facilidad que las mujeres. Para obtener la ratio de las tasas de transición en lugar de los coeficientes  $\beta$ , es suficiente escribir la instrucción *streg* sin la opción *nohr*.

En el modelo 2, que se estima con la opción *dist(gomp)*, es un modelo Gompertz. Los resultados de este modelo muestran que el coeficiente del parámetro *gamma* (que corresponde al parámetro *b* de la ecuación (13.14)) es menor de 0.

```
streg sex coh2 coh3, dist(gomp) nohr
```

Este resultado indica que la tasa de transición es monotónica decreciente, es decir, la propensión a salir del desempleo disminuye en función del tiempo que se ha transcurrido en esta condición. La estimación del modelo Gompertz nos lleva, por lo tanto, a rechazar el supuesto de estabilidad en el tiempo de la tasa de transición, que, como se ha visto, está en la base del modelo exponencial. Así como se ha especificado el modelo Gompertz, sería posible estimar otros modelos paramétricos como el Weibull y el logístico, mediante las opciones *dist(weib)* y *dist(logl)*<sup>14</sup>.

---

<sup>14</sup> No se presentan aquí estas estimaciones por falta de espacio.

### ILUSTRACIÓN 13.4. Modelo exponencial y Gompertz para las duraciones de los episodios de desempleo (modelo 2)

```
* modelo 2

failure _d: dest == 1
analysis time _t: (end-origin)
origin: time begin

Fitting constant-only model:

Iteration 0: log likelihood = -1546.1636
Iteration 1: log likelihood = -1488.497
Iteration 2: log likelihood = -1484.7369
Iteration 3: log likelihood = -1484.7314
Iteration 4: log likelihood = -1484.7314

Fitting full model:

Iteration 0: log likelihood = -1484.7314
Iteration 1: log likelihood = -1447.5164
Iteration 2: log likelihood = -1445.2048
Iteration 3: log likelihood = -1445.1962
Iteration 4: log likelihood = -1445.1962

Gompertz regression -- log relative-hazard form

No. of subjects = 973 Number of obs = 973
No. of failures = 734
Time at risk = 17922
LR chi2(3) = 79.07
Log likelihood = -1445.1962 Prob > chi2 = 0.0000

-----+
_t | Coef. Std. Err. z P>|z| [95% Conf. Interval]
-----+
sex | .2224324 .074057 3.00 0.003 .0772832 .3675815
coh2 | .8207023 .198552 4.13 0.000 .4315475 1.209857
coh3 | 1.256413 .1961255 6.41 0.000 .8720137 1.640812
_cons | -3.934596 .2007817 -19.60 0.000 -4.328121 -3.541071
-----+
gamma | -.0189706 .0022907 -8.28 0.000 -.0234603 -.0144808
-----+
```

La elección entre diferentes modelos de la dependencia temporal depende fundamentalmente de la teoría que se quiere comprobar o de los conocimientos previos del investigador con respecto al proceso analizado. Imagínese, por ejemplo, que nuestra teoría de referencia prevé que la propensión a salir del desempleo disminuye de forma monótona con el paso del tiempo debido, por ejemplo, a un “efecto estigma”. En ese caso se podría ajustar la dependencia temporal con un modelo Gompertz o Weibull, ya que ambos implican una tasa de transición monótona en el tiempo. Si, por el contrario, nuestra teoría sugiere que la tasa de transición crece hasta un determinado momento del tiempo para decrecer después, el modelo más adecuado sería el log-logístico, que permite una dependencia temporal con forma de campana. En cualquier caso, resulta oportuno estimar mo-

delos con diferentes ajustes y comparar sus significaciones estadísticas. Así se puede elegir la forma funcional de dependencia temporal de la tasa de transición más adecuada para describir los datos que se están analizando (Bernardi 2006: 85). Si no se tiene una teoría precisa sobre la forma de la dependencia temporal de la tasa de transición, una solución empleada con mucha frecuencia es optar para modelos semiparamétricos que dejan la función  $q(t)$  sin especificar y estiman los efectos de las variables  $X$ . Entre los modelos semiparamétricos, los más comunes son el modelo de Cox y el modelo exponencial constante a intervalos (*piecewise constant exponential model*). A continuación se proporciona una breve introducción al modelo de Cox.

### 13.4.2. *El modelo Cox*

El modelo propuesto por Cox (1972) ofrece una estimación de los coeficientes  $\beta$  a través de un método de la verosimilitud parcial y deja la función  $q(t)$  sin especificar<sup>15</sup>. El modelo de Cox pertenece a la familia de los modelos proporcionales y, por lo tanto, se basa en el supuesto de que los efectos de las variables  $X$  inducen sólo desplazamientos proporcionales hacia arriba o abajo de  $q(t)$ , sin modificar su forma. En Stata, la instrucción para estimar un modelo de Cox es *stcox*, el cual requiere haber definido previamente los datos como historia de acontecimientos con la orden *stset*. En la ilustración 13.5 se presentan los resultados de la estimación de un modelo Cox para la duración de los episodios de desempleo. Los coeficientes estimados se interpretan de la misma manera que para los otros modelos proporcionales.

Stata utiliza por defecto el método de Breslow para controlar que no haya agrupaciones en la distribución de las duraciones, esto es, que numerosos episodios acaben en el mismo momento del tiempo, hecho que complicaría la estimación de la verosimilitud parcial<sup>16</sup>. Como se aprecia, no hay ningún coeficiente para el efecto de la duración de la legislatura sobre el riesgo de disolución de los gobiernos, es decir, la función  $q(t)$  se deja si especificar. Además, los coeficientes estimados para las variables independientes son muy parecidos a los del modelo Gompertz de la ilustración 13.3.

---

<sup>15</sup> Para una ilustración de cómo funciona el método de estimación con verosimilitud parcial, véase el Apéndice II en Bernardi (2006).

<sup>16</sup> Para más detalles sobre las complicaciones relacionadas con la existencia de agrupamientos (*ties*) en la distribución de las duraciones y sobre los varios métodos para tratarlas, véase Stata (2009g: 128-129).

### ILUSTRACIÓN 13.5. Modelo Cox para las duraciones de los episodios de desempleo

```

stcox sex coh2 coh3, nohr

failure _d: dest == 1
analysis time _t: (end-origin)
origin: time begin

Iteration 0: log likelihood = -4454.3466
Iteration 1: log likelihood = -4419.3661
Iteration 2: log likelihood = -4418.0157
Iteration 3: log likelihood = -4417.9969
Iteration 4: log likelihood = -4417.9969
Refining estimates:
Iteration 0: log likelihood = -4417.9969

Cox regression -- Breslow method for ties

No. of subjects = 973 Number of obs = 973
No. of failures = 734
Time at risk = 17922
LR chi2(3) = 72.70
Log likelihood = -4417.9969 Prob > chi2 = 0.0000
-----+
_t | Coef. Std. Err. z P>|z| [95% Conf. Interval]
-----+
sex | .2272679 .0741742 3.06 0.002 .0818891 .3726467
coh2 | .814659 .1989799 4.09 0.000 .4246656 1.204653
coh3 | 1.223955 .1969942 6.21 0.000 .8378532 1.610056
-----+

```

Para concluir, el modelo de Cox ha sido tradicionalmente una elección muy popular entre los investigadores sociales, quizás por la relativa facilidad con la que se puede estimar, incluso con las primeras versiones de paquetes estadísticos estándares como SPSS y SAS. Sin embargo, su limitación principal es la de no proporcionar ninguna información sobre la pauta de dependencia temporal del proceso investigado. Por esta razón, se aprecia una clara tendencia a utilizar el modelo exponencial constante a intervalos entre los modelos semiparamétricos en los últimos años (Bernardi 2006: 107).

## 13.5. Ejercicios

1. Utiliza el fichero *marriage.dta* contenido en la página web de este libro y estima la función de supervivencia utilizando el estimador de Kaplan y Meier.
2. Usando el fichero *marriage.dta*, estima un modelo exponencial con las variables *género* y *nivel de educación*. Interpreta los coeficientes estimados para estas variables. Construye después un modelo de Cox y un modelo Gompertz.



## 14

# Análisis de datos de encuesta con Stata<sup>1</sup>

En ciencias sociales casi nunca se trabaja directamente con datos de la población objeto de estudio. Puesto que en la mayor parte de los casos se estudian poblaciones muy grandes, costosísimas de analizar directamente, se suele recurrir a la realización de encuestas administradas a una muestra probabilística del universo en estudio. El principio fundamental en el muestreo es la aleatoriedad: sobre ese principio está desarrollada toda la estadística inferencial, como se ha explicado en la segunda parte del capítulo destinado al análisis de una sola variable. Si se elige al azar un número determinado de individuos (*muestra n*) de una población de un determinado tamaño (*población N*), los resultados obtenidos con los datos de la muestra (*estadísticos muestrales*) se podrán utilizar para estimar los datos poblacionales reales (*parámetros poblacionales*), dentro de un rango proporcionado por las probabilidades de la curva normal con un determinado nivel de confianza.

Todo lo visto hasta ahora en este manual sigue estos principios. La estimación de parámetros, los errores típicos, las pruebas de hipótesis, las regresiones, etc., que hasta ahora se han explicado, parten del supuesto de que los datos analizados provienen de una muestra aleatoria simple de la población objeto de estudio. Por tanto, siempre que se analicen datos estadísticos generados tras un muestreo aleatorio simple, se pueden aplicar las técnicas tal como han sido explicadas hasta el momento. ¿Pero qué ocurre si la encuesta realizada no siguió un *muestreo aleatorio simple*? Realmente, en la mayoría de las ocasiones, en las ciencias sociales no se utiliza muestreo aleatorio simple, puesto que se requieren muestras muy grandes de poblaciones muy dispersas en el espacio geográfico. Más habitualmente, se utilizan formas de muestreo complejas, con selección no estrictamente aleatoria de los casos, a través de la construcción de estratos, conglomerados o cuotas, en los que las probabilidades de cada individuo de ser selec-

---

<sup>1</sup> En temas de errores muestrales destacan Coshran (1981) y Kish (1982). También pueden consultarse Azorín y Sánchez Crespo (1986), Pérez (2005) y, de modo aplicado a encuestas, Rodríguez Osuna (1992).

cionado varían (no son idénticas como en el muestreo aleatorio simple). Pues bien, si los datos no responden a un muestreo aleatorio simple, la aplicación *literal* de las técnicas vistas hasta ahora puede dar lugar a estimaciones y parámetros sesgados, así como a la aceptación como verdaderas de hipótesis que realmente son falsas. Por supuesto, eso no quiere decir que no se puedan utilizar las técnicas estadísticas vistas hasta ahora, pero es necesario utilizar las herramientas que Stata nos proporciona para el análisis de datos de muestras complejas. Eso es lo que se explicará en este capítulo.

### 14.1. Ajustes en el análisis de muestras complejas

Como se acaba de decir, el supuesto de todas las pruebas estadísticas analizadas en el cuerpo de este libro, que tienen como misión la generalización de los resultados de encuestas representativas, es el muestreo aleatorio simple. Pero en la mayor parte de las encuestas socioeconómicas, el muestreo que se utiliza no es (directamente) aleatorio, ni simple. El desarrollo de las técnicas de muestreo complejas ha permitido reducir los costes de la investigación mediante encuestas y a un tiempo aumentar la fiabilidad de los resultados. El inconveniente de estas técnicas complejas es que no se pueden analizar los datos directamente como si de muestras aleatorias simples se trataran, sino que deben realizarse una serie de ajustes previos.

Las ventajas del muestreo complejo se entienden mejor mediante un ejemplo. En el supuesto de que se desee realizar una muestra representativa del conjunto de los españoles mayores de 18 años. Si, sobre la base del censo, se seleccionaran de manera puramente aleatoria 2.400 individuos para hacerles una entrevista personal, el coste de la realización de estas entrevistas sería exageradamente alto, debido a la necesidad de desplazarse por toda la geografía española de los entrevistadores (2.400 desplazamientos aleatorios por toda España). Este coste se puede reducir considerablemente si en lugar de seleccionar directamente a los 2.400 individuos, se escogen aleatoriamente 240 secciones censales<sup>2</sup>, y dentro de cada sección censal se extraen diez individuos al azar. En lugar de 2.400 desplazamientos, sólo es preciso realizar 240, puesto que por definición las diez entrevistas de cada sección censal estarán muy próximas espacialmente (el desplazamiento dentro de cada sección puede realizarse a pie). Así se suele realizar habitualmente en este tipo de encuestas, y este tipo de muestreo se llama muestreo polietápico.

Además de ventajas en términos de coste, el muestreo complejo puede tener ventajas en términos de fiabilidad de los resultados. Si se pretende

---

<sup>2</sup> Las secciones censales son las zonas que pertenecen a un mismo colegio electoral.

estudiar las formas de organización del trabajo en las empresas españolas, se sabe que hay muchas diferencias en cómo organizan el trabajo las empresas grandes y las pequeñas, por lo que resulta de gran interés disponer de datos fiables para ambos tipos de empresas. Pero las empresas grandes, aunque de gran relevancia económica y social (por la gran cantidad de empleo que concentran), son pocas en términos numéricos: hay muchas más empresas pequeñas que grandes. Hasta tal punto que si se realizara una selección aleatoria simple, habría muchísimas empresas pequeñas y muy pocas empresas grandes: tan pocas que no se podrían generalizar los resultados obtenidos a su segmento. Para evitar este problema, pueden dividirse las empresas en dos grupos: grandes y pequeñas, y realizar una muestra independiente en cada grupo. Así, se dispondrá de suficientes empresas de ambos tipos y podrán generalizarse los resultados para ambos casos. Este tipo de muestreo se denomina muestreo estratificado.

En la práctica, la mayor parte de las encuestas que se realizan sobre temas sociales y económicos utilizan un muestreo complejo, utilizando muestreos polietápicos y estratificados en sucesivas etapas. Se consigue así reducir costes y ampliar la representatividad de los resultados. Pero estos procedimientos de muestreo obligan a realizar ajustes sobre los datos antes de analizarlos.

La razón por la que no se pueden utilizar los datos de muestras complejas directamente, sin ajustes, es muy sencilla: los estimadores estarían sesgados. En los ejemplos explicados, en el muestreo polietápico, es muy posible que haya relación entre los individuos pertenecientes a una misma sección censal (en un mismo barrio se comparten normalmente estatus socioeconómico y valores culturales), por lo que si se hace una estimación simple, no se tiene en cuenta que puede haber individuos cuyos valores están asociados (rompiendo uno de los supuestos básicos de la inferencia estadística); en el otro ejemplo, en el que se plantea un muestreo estratificado de las empresas españolas, si se estimara la media de uso de turnos de trabajo, saldría mucho más alta que la media real, puesto que en la selección efectuada las empresas grandes están sobrerepresentadas, y las empresas grandes utilizan más sistemas de turnos que las empresas pequeñas. Todo procedimiento complejo de muestreo requiere, pues, la aplicación de ajustes posteriores a los datos para que estos puedan ser analizados.

## 14.2. Ponderaciones, estratos y conglomerados

Los instrumentos principales para realizar los ajustes que permitan generalizar los resultados de encuestas con muestras complejas son los siguientes:

- a) *Ponderaciones*: las ponderaciones son la forma principal de ajustar los datos de encuesta a los parámetros poblacionales. En el mues-

treo aleatorio simple la probabilidad de selección de cada individuo de la muestra es la misma; en el muestreo complejo, en cambio, la probabilidad de selección de los individuos es distinta, según el estrato al que pertenezcan, según cuotas, etc. La ponderación es la inversa de la probabilidad de que un individuo haya sido seleccionado en la muestra. Al aplicarla a los individuos de la muestra, se vuelve a la proporción de la población objeto de estudio.

Por ejemplo, en el caso anterior de la muestra de empresas españolas: las empresas grandes suponen un 10% del total de las empresas españolas, pero (por las razones antes mencionadas) interesa tener suficientes empresas grandes, como un 50% de la muestra para poder generalizar. Por tanto, la probabilidad de seleccionar una empresa grande es mucho mayor que la de seleccionar una empresa pequeña: si se hubiera seleccionado por simple azar, habría una empresa grande de cada diez; pero al aplicar estratificación con afijación constante se dispone de cinco de cada diez. O sea, su probabilidad de ser seleccionadas es cinco veces mayor que la probabilidad que tendrían en un muestreo aleatorio simple. Por tanto, a las empresas grandes habrá que aplicarles una ponderación de 1/5, o sea de 0,2. Con las empresas pequeñas ocurre lo contrario. Si se hubiera hecho muestreo aleatorio, se habrían seleccionado nueve empresas; con el muestreo estratificado sólo se han extraído cinco. Su probabilidad de selección es menor (5/9) de la que hubiera sido en un muestreo aleatorio, por lo que debe aplicarse una ponderación de 9/5, o sea de 1,8 a las empresas pequeñas. Aplicando estos pesos, los resultados que se obtengan serán perfectamente representativos del conjunto de las empresas españolas.

La ponderación no sólo se utiliza cuando existe estratificación, sino siempre que haya selección no aleatoria (pero siguiendo un criterio probabilístico conocido) de los individuos. Además, también se usa para corregir a posteriori los errores de muestreo derivados de la no respuesta. Por ejemplo, si tras la realización de una encuesta se observa que la no respuesta se ha concentrado principalmente en los hombres, de modo que hay sobrerepresentación de mujeres (lo que sesga los resultados), puede aplicarse una ponderación distinta a hombres y a mujeres de modo que los hombres y las mujeres recuperen en la muestra el porcentaje que tienen en la población.

La utilización de las ponderaciones sirve principalmente para la realización de estimaciones correctas de los parámetros poblacionales. Como se acaba de ver en el ejemplo de la muestra estratificada de las empresas españolas, si no se emplean las ponderaciones en una muestra no aleatoria, las medias y proporciones de la muestra no coincidirán con las de la población.

En Stata, las ponderaciones muestrales se llaman *pweights*. Normalmente se dispone de una variable de ponderación en la base de datos (que deberá estar documentada en la metodología de la encuesta), variable que contiene para cada caso *el inverso de su probabilidad de selección en la muestra*. Para aplicar la ponderación a los casos, sólo es preciso indicar a Stata cuál es esta variable con la instrucción *svyset*, que se explica en el ejemplo del siguiente apartado.

- b) *Estratos*: como se ha visto en el caso anterior, la utilización de estratificación determina (al menos en parte) las ponderaciones de los datos muestrales. La estratificación también hace necesaria (o al menos recomendable) la utilización de información sobre los estratos en sí. Cuando se realiza estratificación de la muestra, realmente se realiza una muestra independiente en cada estrato. Por ello, si se indica a Stata cuáles son estos estratos (también con la orden *svyset*), podrá tratarlos como muestras estadísticamente independientes, lo que probablemente reducirá los errores típicos, permitiendo hacer pruebas de hipótesis más ajustadas y fiables.

Esto también tiene una explicación relativamente intuitiva. Normalmente se realizan los estratos buscando que tengan cierta homogeneidad y relación en lo que respecta a lo que ha de estudiarse. En el caso de las empresas españolas, se sabe que las empresas grandes organizan el trabajo de manera muy distinta a las pequeñas. Pero, dado que las empresas grandes son muy pocas, esta diferencia prácticamente no quedaría recogida en la muestra (habría muy pocos casos diferentes), por lo que probablemente no sería estadísticamente significativa. En cambio, al estratificar y sobrerepresentar las empresas grandes, pueden compararse sus valores con los de las empresas pequeñas de manera fiable, de modo que podrán comprobarse con mayor facilidad las hipótesis de investigación.

Si la ponderación es importante para la estimación de parámetros (esto es, para saber si la media de la muestra es igual que la media de la población, por ejemplo), la utilización de los estratos es necesaria para la estimación de los errores y para la comprobación adecuada de hipótesis. Sin especificar los estratos, los errores típicos arrojan valores mayores, por lo que se reducen las probabilidades de rechazar las hipótesis nulas cuando estas son realmente falsas.

- c) *Conglomerados*: en las encuestas sociales el uso de conglomerados o unidades primarias de muestreo es muy habitual, principalmente por el ahorro de costes que supone. Esta técnica permite evitar la dispersión propia del muestreo aleatorio simple, seleccionando los casos en agrupaciones localmente cercanas más fácilmente accesibles en menor periodo de tiempo.

El problema que presenta el muestreo por conglomerados es que las observaciones de un mismo conglomerado no son independien-

tes, vulnerando así el supuesto de independencia de la mayor parte de las técnicas estadísticas. Es presumible que personas de un mismo barrio (si el barrio es el conglomerado) se asemejen más entre sí que personas de barrios distintos. Si no se utiliza la información sobre los conglomerados, la estimación de errores típicos será menor de lo que en realidad debería ser (puesto que la agrupación en conglomerados reduce la variabilidad de manera sesgada), corriendo el riesgo de aceptar como significativas diferencias entre parámetros que no deberían serlo para un nivel de confianza determinado.

En una misma muestra se pueden realizar etapas sucesivas de selección (por ejemplo, primero ciudades, dentro de ciudades barrios y dentro de barrios individuos). Pese a que la información de conglomerados también puede utilizarse para la estimación de errores típicos, Stata no permite más que utilizar la información de una única unidad primaria de muestreo (PSU [Primary Sampling Unit]), por lo que en las circunstancias de este ejemplo sería deseable optar por la penúltima unidad seleccionada en el muestreo<sup>3</sup> (en el caso anterior, los barrios).

El cuadro 14.1 resume lo que debe tenerse en cuenta a la hora de trabajar con datos de muestras complejas.

Hay una última cuestión importante que debe tenerse en cuenta al trabajar con datos muestrales. La utilización o no de los instrumentos enumerados resulta más o menos crítica en función de los objetivos de nuestro análisis. Si lo que se pretende es estimar a través de los datos muestrales los parámetros poblacionales (por ejemplo, si se quiere estimar el porcentaje de voto a un partido en función de una encuesta preelectoral), la utilización de ponderaciones es absolutamente fundamental. Si lo que se persigue es estar muy seguros de que los resultados de la muestra son cercanos a los reales, o si el objetivo fundamental es realizar una prueba de hipótesis de significación con los datos muestrales, no sólo deben emplearse las ponderaciones, sino también los conglomerados y los estratos, siempre y cuando se disponga de la información. Ahora bien, si lo que se desea es estudiar la relación entre dos o más variables (el efecto de la clase social sobre la intención de voto, por ejemplo), la utilización de ponderaciones, conglomerados y estratos es mucho menos importante. Ciertamente, si no se utilizan, hay mayor probabilidad de equivocarse, por ejemplo, en la estimación de los parámetros de la regresión, así como de errar en la significación de estos

---

<sup>3</sup> Aunque las unidades de muestreo más allá de la señalada como primaria también pueden tener un efecto sobre los errores típicos, el error que puede derivar de su no utilización es realmente irrisorio. Si se desea, no obstante, utilizar esta información en la estimación de parámetros y errores típicos, puede emplearse un programa ya totalmente especializado como SUDAAN.

parámetros (rechazando parámetros que realmente sí son significativos o, lo que es peor [pero más difícil], aceptando parámetros realmente no significativos). Pero, sin duda, el problema es menor, puesto que de lo que se trata en este caso es de hallar relaciones entre variables, no de predecir o estimar los parámetros poblacionales. Obviamente, es mucho menos importante sobreestimar o infraestimar ligeramente (la diferencia suele ser pequeña en modelos de relaciones entre variables) una relación entre dos o más variables que equivocarse en la estimación del voto a un determinado partido. Si existe relación entre clase social y voto, aparecerá en los resultados, aunque la clase alta esté infrarrepresentada en relación con la baja —el problema estriba en que sea más difícil generalizar los resultados.

**CUADRO 14.1. Instrumentos de ajustes muestrales y consecuencias**

	Efectos sobre la estimación del parámetro poblacional	Efectos sobre los errores típicos y los test de hipótesis	Importancia de su uso para el análisis	Opción de svyset en Stata
Ponderaciones	Importante	Ninguno	Alta, evita tanto errores en las estimaciones como la aceptación de hipótesis falsas	[pweight=]
Estratos	Ninguno	Importante, los reduce	Baja, aunque puede reducir considerablemente el error de las estimaciones permitiendo resultados menos conservadores	strata()
Conglomerados	Ninguno	Importante, los amplía	Media, evita el que los errores típicos sean anormalmente bajos y de ese modo aceptemos como ciertas hipótesis dudosas	psu()

La recomendación, por tanto, es que para la estimación de parámetros poblacionales, para realizar predicciones y pruebas de hipótesis se utilicen siempre los instrumentos de análisis de encuesta. Para el estudio de la asociación entre dos o más variables, también se recomienda la utilización siempre que sea posible de la información del muestreo, para poder afinar más el análisis y hacerlo más robusto y fiable. Pero si no es posible o resulta excesivamente complicado, se puede hacer el análisis sin utilizarla, aunque teniendo mucha más cautela en lo que respecta a la validez externa de los resultados.

### 14.3. Un ejemplo práctico con Stata. Las órdenes *svy*

Hay dos maneras de trabajar con datos de muestras complejas en Stata. La primera es utilizando las instrucciones habituales de Stata, añadiendo unas opciones específicas para utilizar datos de encuesta<sup>4</sup>. La segunda manera, la más recomendable, consiste en utilizar las preinstrucciones *svy* de Stata, que son un conjunto de órdenes específicamente creadas para trabajar con datos de muestras complejas. Casi cada instrucción de análisis estadístico en Stata tiene su correlativa orden *svy*, que es exactamente la misma, pero ajustada para trabajar con datos de encuesta. Así, *regress* tiene *svy:regress*, *logit* tiene *svy:logit*, etc. En principio, funcionan exactamente igual que su equivalente sin ponderación, ya explicada, por lo que no se contempla detenidamente la explicación de cada orden *svy*, sino que más bien se explica el funcionamiento general de estas instrucciones.

Cuando se trabaja con datos de encuesta, lo primero que es necesario hacer es acudir a la documentación de los datos y estudiar en detalle cuál fue la metodología seguida en el muestreo. Si se utilizó muestreo aleatorio simple, no hace falta usar las instrucciones *svy*. Si se utilizó alguna forma de muestreo complejo, se deben buscar los tres elementos vistos más arriba (ponderación, conglomerados y estratos) y utilizar los que sean relevantes, en función del diseño muestral y de la información disponible. La variable de ponderación suele estar presente en la gran mayoría de las bases de datos de encuesta, lo que no ocurre con la información sobre unidades primarias y estratos. A veces la de conglomerados (o unidades primarias de muestreo) está disponible: en muchas encuestas, por ejemplo, una de las variables es la sección censal. La variable de estratos, en cambio, no suele aparecer en los datos proporcionados por los institutos estadísticos. En algunos casos es posible reconstruirla con la información de la documentación metodológica (por ejemplo, si hay dos estratos en función del tamaño de la empresa [más y menos de 100 trabajadores], se puede crear una variable, *estrato*, que contenga 1 si tiene menos de 100 trabajadores y 2 si tiene más), mientras que en otras ocasiones es imposible. Como ya se ha dicho, la no utilización del estrato no compromete la estimación de parámetros, sino que simplemente hace que los errores típicos sean mayores (véase sección 4.6).

Una vez conocida la información muestral, debe proporcionársela a Stata mediante la orden *svyset*. Esta instrucción sirve exclusivamente para eso, para decirle a Stata cuáles son las variables de ponderación, conglomerados y estratos.

---

<sup>4</sup> Esto es posible añadiendo a la instrucción (por ejemplo, a *regress*) la opción [*pweight* = *peso*] (antes de la coma), donde *peso* es la variable de ponderación; si la muestra es polietápica, ha de señalarse cuál es la variable que contiene información sobre las unidades primarias de muestreo con la opción *cluster(variable)*. Sin embargo, *regress* no permite tener en cuenta los estratos; para ello han de emplearse las órdenes *svy*.

### 14.3.1. Establecer la información muestral: la instrucción svyset

Véase un ejemplo. Se va a estudiar, con la *Encuesta de Calidad de Vida en el Trabajo* de 2001, la satisfacción de los asalariados con su empleo. Si se estudiaba la metodología de esta encuesta, se descubre que se realizó un muestreo trietápico con estratificación de las unidades de primera etapa. Las unidades de primera etapa fueron secciones censales, que se estratificaron en función del tamaño del municipio, estableciéndose cinco estratos con muestras independientes. Dentro de cada sección censal, se seleccionaron familias (unidades de segunda etapa) y dentro de las familias, a población ocupada de más de 16 años (unidades de tercera etapa). Tras la realización de la encuesta, se aplicaron factores de reequilibrio o ajuste para corregir las diferencias de la muestra final con la población en términos de situación profesional, edad y sexo. Por tanto, en este caso existen los tres elementos vistos más arriba.

En la documentación se advierte una variable de ponderación, llamada *pond*, así como una variable (*v294*) que incorpora información sobre el número de orden de la sección censal. De lo que no se dispone es de una variable que contenga los estratos, y tampoco puede reconstruirse, puesto que, aunque existe una variable de tamaño del municipio, las categorías no coinciden con las que se aplicaron en la estratificación de las secciones censales. Por ello, se empleará ponderación y unidades primarias, pero no los estratos (con lo que son previsibles errores típicos algo mayores de los reales)<sup>5</sup>.

Para dar a Stata esta información, se emplea la instrucción *svyset* con el peso entre corchetes y las opciones necesarias, sólo *psu()*, en este ejemplo, y para obtener una descripción del resultado se utiliza la instrucción *svydes*.

```
svyset [pweight=pond], psu(v294)
svydes
```

La pantalla de resultados de Stata muestra, en consecuencia, la disposición muestral:

---

<sup>5</sup> De modo aproximado podría utilizarse la comunidad autónoma (*v4*) y/o el hábitat (*v298*) como estratos. En el fichero con los ejemplos de este capítulo se han incluido las instrucciones con esa especificación, para que el usuario las cambie para los ejercicios finales.

### ILUSTRACIÓN 14.1. Descripción de la muestra compleja

Survey: Describing stage 1 sampling units					
pweight: pond					
VCE: linearized					
Single unit: missing					
Strata 1: <one>					
SU 1: v294					
FPC 1: <zero>					
#Obs per Unit					
Stratum	#Units	#Obs	min	mean	max
-----	-----	-----	-----	-----	-----
1	430	5998	10	13.9	15
-----	-----	-----	-----	-----	-----
1	430	5998	10	13.9	15

La utilización de la orden *svyset* requiere la siguiente información: la ponderación (*pweight*), los estratos (*strata*) y los conglomerados (*psu*). Como en *svyset* sólo se han especificado *pweight* y *psu*, Stata considera que sólo hay un estrato, que incluye a toda la muestra. Si hubiera estratos, se debería especificar *strata(estrato)* tras la opción *psu*. Tras *svyset*, se ha introducido la instrucción *svydes*, que sirve para describir la información sobre la muestra almacenada en la memoria de Stata. Tras estos elementos, aparece un resumen de los datos de los estratos y los conglomerados: el número de estratos que hay, cuántos conglomerados hay por estrato, y cuántas observaciones por estrato y por conglomerado. En este caso, sólo hay un estrato, con 430 PSU (unidades primarias) y 5.998 observaciones. Cada una de las unidades primarias tiene entre 10 y 15 observaciones, con una media de 13,9 (aproximadamente fueron entrevistados 14 individuos en cada conglomerado).

#### 14.3.2. Estimación de medias y proporciones

Una vez que se dispone de la información sobre las características de la muestra y la ponderación, se pueden realizar estimaciones con las instrucciones *svy*<sup>6</sup>. Las órdenes para realizar estimaciones poblacionales univariales son distintas de sus equivalentes sin muestreo, por lo que serán explicadas con mayor detalle; al contrario que las órdenes de análisis estadístico bivariado y multivariado, que son básicamente iguales a sus homólogas sin

<sup>6</sup> Hay que prestar atención a la versión que se emplea de Stata, ya que con anterioridad a la versión 9, las instrucciones *svy* no se formulaban como preinstrucciones, sino como instrucciones precedidas por la palabra *svy*. Esto implica dos diferencias principales: una, los dos puntos con los que hay que escribir las nuevas órdenes; la otra es que las opciones de *survey* han de aparecer antes de los dos puntos, en lugar de al final, como anteriormente.

ponderación, por lo que sólo se contemplarán algunos ejemplos. En todos los casos se compararán los resultados con y sin el uso de ponderaciones y PSU, para ilustrar su utilidad.

Véase en primer lugar cómo estimar proporciones. Una variable que sirve de indicador aproximado de la satisfacción laboral es la de búsqueda de un empleo distinto del actual (si alguien busca otro empleo, no debe de estar muy contento con el que ya tiene). En la ilustración 14.2 se muestra la estimación de la proporción de personas que buscan otro empleo sin utilizar la información de la muestra (con la orden *tabulate*, que se explica en los capítulos 4 y 8) y con la instrucción *svy: proportion*, que es su equivalente para muestras complejas:

```
tabulate v42 if asal==1
svy, subpop(asal): proportion v42
```

Estas dos instrucciones generan sendas estimaciones de las proporciones de la variable *búsqueda de otro empleo*:

#### ILUSTRACIÓN 14.2. Estimación de proporciones en muestras complejas

búsqueda de	Freq.	Percent	Cum.
otro empleo			
si	451	9.63	9.63
no	4,232	90.37	100.00
<hr/>			
Total	4,683	100.00	

Survey: Proportion estimation			
Number of strata = 1	Number of obs = 5998		
Number of PSUs = 430	Population size = 6020		
	Subpop. no. obs = 4683		
	Subpop. size = 4799.5		
	Design df = 429		
<i>_prop_1:</i> v42 = si			
<i>_prop_2:</i> v42 = no			
<hr/>			
	Linearized		
	Proportion	Std. Err.	[95% Conf. Interval]
<hr/>			
v42			
<i>_prop_1</i>   .1042389 .0055767 .0932778 .1151999			
<i>_prop_2</i>   .8957611 .0055767 .8848001 .9067222			

Como sólo son de interés los valores de los asalariados, se seleccionan aquellos casos en los que la variable *asal* es igual a 1. Aquí hay que co-

mentar una cuestión importante. Cuando se trabaja con datos de muestras complejas, y se quiere utilizar una submuestra (como, por ejemplo, en este caso los asalariados), no se pueden simplemente quitar los datos de la matriz (por ejemplo, con la instrucción *drop*) y analizar los datos seleccionados. La ponderación está hecha para que el total de las observaciones de la muestra sean representativas del total de los individuos de la población: si se trabaja con una submuestra esa ponderación no sirve. Por eso, cuando se deseen utilizar submuestras, como en este caso, habrá que utilizar la opción *subpop()* de las órdenes *svy*. Con esa opción (dentro del paréntesis, debe haber una variable que tenga un valor distinto de 0 [recomendable el 1] en aquellos casos en que se necesiten incluir en el análisis), Stata utiliza todas las observaciones, aunque las estimaciones sólo hacen referencia a los casos de la submuestra especificada. Así, pueden estudiarse submuestras sin tener resultados sesgados.

Pasando ya al análisis en sí de las salidas de ambas tablas (con y sin ponderaciones), se puede ver claramente la diferencia en las estimaciones de los parámetros poblacionales utilizando la información del muestreo. Sin ponderación, la estimación del porcentaje de asalariados que busca otro empleo es del 9,63%. Al utilizar la información del muestreo, la estimación es de 10,42% (en la salida de *svy:proportion* no aparece como porcentaje, sino como proporción sobre 1 [0,1042]), o sea, casi un punto más. Dependiendo de los objetivos del análisis, el error de la estimación puede ser más o menos importante (un punto en la estimación del porcentaje de paro puede tener una gran importancia, por ejemplo). En cualquier caso, si se dispone de la información muestral, por razones de corrección y de validez científica, siempre es aconsejable usarla en las estimaciones. Los errores típicos de las proporciones, obviamente, también son más correctos, lo que permite establecer unos intervalos de confianza creíbles.

En el caso de estimación de medias, la instrucción pasa a ser *svy:mean*, en lugar de *svy:proportion*.

```
ci v92 if valsat = =1  
svy, subpop(valsat): mean v92  
estat effects
```

### ILUSTRACIÓN 14.3. Estimación de medias en muestras complejas

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
v92	4648	7.100473	.0290105	7.043599 7.157348
<hr/>				
Survey: Mean estimation				
Number of strata	= 1	Number of obs	= 5998	
Number of PSUs	= 430	Population size	= 6020	
		Subpop. no. obs	= 4648	
		Subpop. size	= 4763.2	
		Design df	= 429	
<hr/>				
		Linearized		
		Mean	Std. Err.	[95% Conf. Interval]
v92	7.085379	.0371132	7.012433	7.158325
<hr/>				
		Linearized		
		Mean	Std. Err.	DEFF DEFT
v92	7.085379	.0371132	1.6584	1.28779

En la ilustración 14.3 se muestra en primer lugar la media de satisfacción con el trabajo utilizando la instrucción *ci*, que nos da, además de la media, los intervalos de confianza de esa media<sup>7</sup>. Existe diferencia, aunque pequeña. Sin ponderación, la media de satisfacción en el trabajo de los asalariados españoles es de 7,1; con ponderación es ligeramente inferior, de 7,085. En cambio, el error típico de la media en los datos ponderados es mayor (0,037 frente a 0,029), como corresponde a un muestreo por conglomerados en dos etapas, lo que afecta también al intervalo de confianza, que en la muestra ponderada es mayor. De nuevo, las diferencias en la estimación no son muy grandes, aunque algo mayores son los errores típicos.

Especialmente revelador es el cálculo del efecto del diseño (Kish 1982), que se obtiene con la orden *estat effects*<sup>8</sup>. Este estadístico indica la eficiencia del muestreo utilizado en los datos. Resulta de dividir la varianza obtenida con los datos muestrales entre la varianza que se habría obtenido, si la

<sup>7</sup> Como en el caso anterior, sólo se incluyen los casos de individuos asalariados. Por eso, se especifica la cláusula *if valsat == 1*. En este caso, también se excluyen aquellos casos en los que no hay respuesta a la variable *satisfacción*. La misma variable (*valsat*, igual a uno en los asalariados que responden a la pregunta de satisfacción) se utiliza para definir la subpoblación (opción *subpop*) de la instrucción *svymean*.

<sup>8</sup> Antes de la versión 10 de Stata, el efecto del diseño se proporcionaba automáticamente al solicitar la media. Desde esta versión, hay que pedirlo explícitamente mediante la orden *estat effects*.

muestra hubiera sido aleatoria simple. Cuanto más pequeño sea el valor de  $Deff$ , más eficiente es el muestreo aplicado, y viceversa. Valores menores a 1 indican una reducción en la varianza (y, por tanto, una muestra mejor a la que se habría obtenido si se hubiera realizado muestreo aleatorio simple); valores superiores a 1 indican un aumento en la varianza y, por tanto, un muestreo menos eficiente que el aleatorio simple. En este caso el valor es 1,6, indicando una muestra menos eficiente que la obtenida con muestreo aleatorio simple. Esta menor eficiencia se explica por dos cosas: primero, porque aunque sea menos eficiente es más barata (como decíamos, los conglomerados reducen los costes aunque aumentan los errores típicos); y, segundo, porque al no disponer de la información necesaria sobre los estratos para introducirla en las estimaciones con órdenes *svy*, la varianza estimada es anormalmente alta (debería ser menor). Finalmente,  $Deft$  es la raíz cuadrada del efecto del muestreo.

#### *14.3.3. Comparación de medias y tablas de contingencia*

Si se desea realizar estimaciones realmente ajustadas y fiables, es absolutamente necesario utilizar las ponderaciones y el resto de la información muestral. Para analizar relaciones entre variables, en principio, no es tan crítico el uso de la información muestral, aunque también es muy aconsejable, pues reduce la posibilidad de cometer errores.

Para mostrar un ejemplo de diferencias, se presenta a continuación la relación entre búsqueda de otro empleo y satisfacción en el trabajo, realizando una comparación de medias:

```
bysort v42: ci v92 if valsat == 1
svy, subpop(valsat): mean v92, over(v42)
```

Se supone que se conoce la instrucción *bysort*, que sirve para ejecutar instrucción en dos o más categorías de una variable determinada. En este caso, se desea obtener mediante la orden *ci* la media y el intervalo de confianza de la satisfacción en el trabajo en función de si el trabajador está buscando otro empleo. Los trabajadores que están buscando otro empleo tienen una satisfacción laboral bastante menor (casi dos puntos) que los que tienen intención de permanecer en él, al menos a corto plazo. Las diferencias entre las estimaciones con y sin ponderación son pequeñas, de nuevo, aunque algo mayores en el caso de los trabajadores que están buscando otro empleo (de 5,528 a 5,562). Los errores típicos son algo mayores también en ambos casos, lo que agranda los intervalos de confianza de la estimación de la media. Hay que llamar la atención sobre la importancia de este hecho. Aunque aquí no sucede, podría ocurrir que el error típico se

agrandase tanto con la ponderación que los intervalos de confianza al 95% se cruzasen, de modo que no se tuviera seguridad para decir que la diferencia observada entre las medias de ambos grupos existiera realmente en la población y no se debiera a errores de muestreo. Por ello, es aconsejable utilizar la información sobre ponderaciones y conglomerados en la estimación de medias y proporciones.

#### ILUSTRACIÓN 14.4. Estimación por intervalos de las medias en dos grupos

```
-> v42 = si
      Variable |       Obs        Mean      Std. Err.      [95% Conf. Interval]
-----+-----+
      v92 |     449     5.52784     .1076234      5.31633     5.739349

-----+-----+
-> v42 = no
      Variable |       Obs        Mean      Std. Err.      [95% Conf. Interval]
-----+-----+
      v92 |     4199    7.268635     .0287981      7.212176    7.325095

Survey: Mean estimation
Number of strata =          1      Number of obs      =      5998
Number of PSUs   =      430      Population size   =      6020
                                         Subpop. no. obs =      4648
                                         Subpop. size    =    4763.2
                                         Design df       =      429

_subpop_1: v42 =  si
_subpop_2: v42 =  no

-----+-----+
      Over |      Linearized
      Over |       Mean      Std. Err.      [95% Conf. Interval]
-----+-----+
      v92 |           |
      _subpop_1 |  5.562022     .1067634      5.352177    5.771866
      _subpop_2 |  7.2633      .0365404      7.19148     7.335121
```

El ejemplo siguiente es de tablas de contingencia. Como en los casos anteriores se solicita en primer lugar el análisis sin ponderar y luego el ponderado para que se adviertan las diferencias.

```
tabulate v42 v146 if asal ==1, col
svy, subpop(asal): tabulate v42 v146, obs col per se ci
```

Con la instrucción *svy:tabulate* puede obtenerse un cruce de variables idéntico al realizado con la instrucción *tabulate*, ya explicada en el capítulo de tablas. El formato es parecido, aunque para conseguir que muestre las

observaciones como la instrucción *tabulate* y las proporciones en porcentajes, deben añadirse las opciones *obs* y *per*. Lo que muestran ambas tablas es un cruce de la variable que indica si el trabajador está buscando otro empleo, por horario nocturno (parece razonable que el horario nocturno lleve a los trabajadores a no estar tan contentos con su empleo y, por tanto, a buscar otro). La relación, tanto en la tabla con ponderación como en la tabla sin ella, no es significativa (según las pruebas del chi2 y de la F). Al aplicar ponderación y PSU, los porcentajes de los que quieren cambiar su empleo se elevan, sobre todo en el caso de los que trabajan siempre de noche. Los estadísticos de asociación de las variables no aparecen en la orden *svy:tabulate*; si se requieren, pueden usarse los de los datos sin ponderar, pues la información sobre la relación entre variables no se ve tan afectada por el uso o no de ponderación o demás información muestral. En cualquier caso, se recomienda que siempre que se pueda se utilicen las ponderaciones y la información muestral para aumentar la fiabilidad de las estimaciones.

#### ILUSTRACIÓN 14.5. Tablas de contingencia en muestras complejas

Number of strata	=	1	Number of obs	=	5998
Number of PSUs	=	430	Population size	=	6020
			Subpop. no. of obs	=	4683
			Subpop. size	=	4799.505
			Design df	=	429
<hr/>					
búsqueda					
de otro			horario nocturno		
empleo		siempre	a veces	nunca t	Total
-----+-----					
si	15.23	11.06	10.04	10.42	
	(3.024)	(1.296)	(.6139)	(.5577)	
[10.19, 22.17]	[8.756, 13.87]	[8.892, 11.31]	[9.378, 11.57]		
	26	75	350	451	
no	84.77	88.94	89.96	89.58	
	(3.024)	(1.296)	(.6139)	(.5577)	
[77.83, 89.81]	[86.13, 91.24]	[88.69, 91.11]	[88.43, 90.62]		
	164	725	3343	4232	
Total	100	100	100	100	
190	800	3693	4683		
<hr/>					
Key: column percentages					
(linearized standard errors of column percentages)					
[95% confidence intervals for column percentages]					
number of observations					
 Pearson:					
Uncorrected	chi2(2)	=	7.4185		
Design-based	F(2.00, 857.09)	=	2.1268	P = 0.1199	

Además de mostrar los porcentajes y el test de independencia de la tabla, *svy:tabulate* puede calcular otros estadísticos que no se obtienen con la

instrucción normal *tabulate*. Las opciones *se* y *ci* hacen que, junto con las proporciones de cada casilla, Stata muestre el error típico de la proporción, así como el intervalo de confianza al 95% de esa proporción (ilustración 14.5). Así, se detecta claramente por qué la relación entre la búsqueda de otro empleo y el tener un horario nocturno no es significativa: los intervalos de confianza de las distintas casillas se cruzan. O sea, según los datos, puede afirmarse (con un 95% de seguridad) que entre el 10,2% y el 22,2% de los que trabajan siempre por la noche buscan otro empleo; que entre el 8,8% y el 13,9% de los que trabajan a veces de noche buscan otro empleo, y que entre el 8,9% y el 11,3% de los que nunca trabajan de noche buscan otro empleo. O sea, que es perfectamente posible que los tres grupos busquen empleo con un mismo porcentaje. Por ejemplo, podría ser (de acuerdo con los datos obtenidos) que en los tres casos hubiese un 11% de trabajadores buscando otro empleo. Por tanto, puede concluirse que la relación que se observa en la tabla no es significativa.

#### 14.3.4. *Otras instrucciones svy*

Como se ha comprobado, las órdenes más habituales de Stata tienen su correspondiente instrucción para muestras complejas. En general, todas se usan del mismo modo que la orden normal, con algunas ligeras diferencias derivadas del hecho de que están ajustadas para su uso con datos de muestras complejas (la mayor parte de las opciones que aceptan las órdenes *svy:mean*, *svy:proportion* y *svy:tabulate* vistas hasta ahora también se pueden utilizar con las otras instrucciones *svy*). Pero esas ligeras diferencias son más técnicas que otra cosa: en términos prácticos, para hacer análisis multivariable de datos de muestras complejas, basta con seguir las pautas que se han aconsejado: primero estudiar las características de la muestra, luego darle la información a Stata con la orden *svyset* y hacer el análisis utilizando la instrucción *svy* que sea pertinente.

En concreto, de las técnicas que se han estudiado en este manual, cabe destacar la regresión y el logit. En ambos casos existe la instrucción *svy* específica (*svy:regress* y *svy:logit*), que permite utilizar la ponderación y otra información muestral para hacer análisis y estimaciones complejas y multivariadas. La forma de hacerlo es exactamente igual que la que se ha explicado anteriormente: especificación de las características de la muestra con *svyset* y empleo de la instrucción precedida por *svy:*. Si se necesita realizar predicciones para interpretar los resultados o hacer diagnósticos de ajuste del modelo, se realizan también exactamente del mismo modo, utilizando la instrucción *predict* tras la estimación del modelo lineal o logarítmico. Para estas estimaciones posteriores no hay un *svypredict* ni nada parecido, puesto que la información sobre el muestreo ya ha afectado previamente a la estimación del modelo y a la de los parámetros, y es a

partir de estos con lo que se generan las predicciones, residuos o medidas similares.

#### 14.4. Ejercicios

1. Este capítulo fue originalmente escrito en la versión 8 y ha sido posteriormente modificado para la versión 10 de Stata. Las instrucciones vienen en dos ficheros: *capitulo14* y *capitulo14b*. Con la precaución de indicar la versión anterior al inicio, Stata es capaz de ejecutar ficheros con instrucciones obsoletas en las nuevas versiones. Comprueba esto y analiza las diferencias que se producen en los resultados.
2. Utiliza el fichero de programa de la versión 10 (*capitulo14.do*) cambiando la especificación del modelo de muestreo. Usa, en primer lugar, la variable *hábitat* (v268) como estrato, además de las unidades primarias. Después, reemplaza el estrato por la comunidad autónoma. Finalmente, realízalo por el cruce de comunidad autónoma por hábitat.
3. Con los datos (*daecvt01*) con los que se ha realizado este capítulo, realiza una regresión lineal con datos sin ponderar y ponderados de la satisfacción por el trabajo (v92) sobre el número de horas trabajadas (v150), el horario nocturno (v146), el trabajo en sábado (v309) y el trabajo en domingo (v310).
4. Del mismo modo que en el ejercicio anterior, haz ponderada y no ponderadamente una regresión logística de buscar empleo (v42) sobre las mismas variables.

# 15

## Bibliografía comentada

- Acock, A. C. (2006): *A Gentle Introduction to Stata*, College Station (TX): Stata Press.  
Introducción a Stata muy accesible de la versión 10.0, en su segunda edición, basada principalmente en un uso de este programa mediante menús. Orientado hacia la psicología y las ciencias, sociales, persigue el aprendizaje de buenos hábitos estadísticos e informáticos entre los usuarios de Stata. Incluye análisis factorial.
- Afifi, A. A. et al. (2003): *Computer-Aided Multivariate Analysis* (2<sup>a</sup> ed.), Nueva York: Chapman & Hall.  
Libro de análisis multivariado muy básico, donde se explican mediante ejemplos los principales conceptos estadísticos. Trabaja con BMDP, SYSTAT y en su segunda versión con S-PLUS.
- Agresti, A. (2002): *Categorical Data Analysis*, Hoboken (NJ): John Wiley & Sons.  
Tratado de tipo medio sobre el análisis de datos nominales. Además de logit y mlogit contiene modelos lineales-logarítmicos.
- Aldrich, J. H. y F. D. Nelson (1984): *Linear Probability, Logit, and Probit Models*, Londres: Sage.  
Libro básico de la colección verde de Sage. Ideal para introducirse en regresión logística.
- Allison, P. (1984): *Event History Analysis. Regression for Longitudinal Event Data*, Londres: Sage.
- Andersen, E. B. (1997): *Introduction to the Statistical Analysis of Categorical Data*, Berlín-Nueva York: Springer.  
Aunque se llame introducción es un manual intermedio para el análisis de tablas de contingencia. Presta especial atención a los modelos log-lineales y de éstos aborda los modelos logit.
- Azorín, F. y J. L. Sánchez-Crespo (1986): *Métodos y aplicaciones del muestreo*, Madrid: Alianza.  
Un clásico español del muestreo.
- Baum, C. F. (2006): *An Introduction to Modern Econometrics Using Stata*, College Station (TX): Stata Press.  
Libro de naturaleza muy práctica para estudiantes de econometría que quieran usar Stata. Incluye series temporales, datos de panel y variables instrumentales, además de los análisis básicos. Está orientado a generar sencillos ficheros .do para las tareas repetitivas.
- (2009): *An Introduction to Stata Programming*, College Station (TX): Stata Press.

- Libro especializado en la programación con Stata. Sólo para aquellos que quieran hacer programas propios.
- Beaton, A. E. y W. Tukey (1974): «The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data», *Technometrics*, 16: 146-185.  
Artículo de referencia para las regresiones robustas.
- Belsley, D. A. et al. (1980): *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Nueva York: John Wiley & Sons.
- Bernardi, F. (2006): *El Análisis de la historia de acontecimientos*, Madrid: Centro de Investigaciones Sociológicas.  
Monográfico sobre análisis de la historia de acontecimientos con numerosos ejemplos realizados con Stata publicado en esta misma colección por uno de los autores de este libro.
- Blalock, H. M. (1966): *Estadística Social*, México: FCE.  
Clásico de la Estadística con abundantes explicaciones de la materia para personas con poca base matemática.
- Blossfeld, H. P. y G. Rohwer (2001): *Techniques of Event History Modeling. New Approaches to Causal Analysis* (2<sup>a</sup>ed.), Mahwah (NJ): Erlbaum.
- Borroah, V. K. (2001): *Logit and Probit. Ordered and Multinomial Models*, Londres: Sage.  
Libro específico en la conocida serie verde de Sage para logits y probits ordinales y multinomiales. Aun siendo de 2001, ya usa Stata como programa principal.
- Cahuzac, E. y C. Bontemps (2008): *Stata Par la Pratique: Statistiques, Graphiques et Éléments de Programmation*, College Station (TX): Stata Press.  
Manual francés para iniciación del programa. Se centra especialmente en las órdenes de Stata, sin olvidar de explicar con claridad los conceptos y la interpretación de los resultados. Contiene también una introducción a la programación y recurre también a instrucciones adicionales del programa que pueden obtenerse de internet. Dedica un capítulo a la exportación de resultados a otros programas como procesadores de textos, webs e incluso LaTeX.
- Cameron, A. C. y P. K. Trivedi (2005): *Microeconometrics Using Stata*, College Station (TX): Stata Press.  
Manual de Stata especialmente dirigido a la econometría. Incluye temas avanzados como simulación, mínimos cuadrados generalizados, variables instrumentales, datos de panel, regresiones no lineales. Todo ello con elementos básicos de programación matricial.
- Carroll, G. y M. Hannan (2000): *The Demography of Corporations and Industries*, Princeton (NJ): Princeton University Press.
- Castro, T. (1999): «Pautas recientes en la formación de pareja», *Revista Internacional de Sociología*, 23: 332-373.
- Cea, M. Á. (2002): *Análisis multivariable. Teoría y práctica en la investigación social*, Madrid: Síntesis.  
Contiene capítulos con muchos y útiles ejemplos prácticos de regresión y logística, pero sin ninguna referencia a Stata.
- Cleves, M. et al. (2008): *An Introduction to Survival Analysis Using Stata*, College Station (TX): Stata Press.  
Libro de análisis de supervivencia mediante el programa Stata para quienes necesitan aplicar este tipo de análisis a sus datos. Principalmente orientado a científicos de la salud, pero también útil a economistas, sociólogos y polítólogos.

- Cochran, W. G. (1981): *Técnicas de Muestreo*, México: CECSA.  
Uno de los manuales clásicos de muestreo.
- Cook, R. D. Y S. Weisberg (1983): «Diagnostic for Heterocedasticity in Regression», *Biometrika*, 70(1): 1-10.
- Cox, D. (1972): «Regressions Models and Life-Tables», *Journal of the Royal Statistical Society*, 34: 187-220.
- Cox, N. J. (1999): *Tab\_Chi: Stata Modules for Tabulation and Chi-Square Tasks*, Boston College, Department of Economics.  
Rutinas de Cox para el cálculo de residuos ajustados en las tablas de contingencia.
- (2004): «Speaking Stata: Graphing Categorical and Compositional Data», *The Stata Journal*, 4(2): 190-215.  
Rutinas de Cox para la obtención de gráficos categóricos. Contiene las explicaciones de programas imprescindibles para la representación de variables cualitativas. Se recomiendan especialmente *catplot* y *tabplot*.
- Cuadras, C. M. et al. (1996): *Fundamentos de estadística. Aplicación a las ciencias humanas*, Barcelona: Voluminosa y rigurosa introducción a la Estadística para quienes tengan buena base matemática.
- Diekmann, A. (1989): «Diffusion and Survival Models for the Process of Entry into Marriage», *Journal of Mathematical Sociology*, 14: 31-44.
- (1992): «The Log-Logistic Distribution as a Model for Social Diffusion Processes», *Journal of Scientific & Industrial Research*, 51: 285-290.
- Dow, J. K. y J. W. Endersby (2004): «Multinomial Probit and Multinomial Logit: A Comparison of Choice Models for Voting Research», *Electoral Studies*, 23: 107-122.
- Everitt, B. S. (1977): *The Analysis of Contingency Tables*. Londres: Chapman and Hall.  
Básica introducción a las tablas de contingencia.
- Escobar, M. (1999): *Análisis gráfico/exploratorio*, Madrid: La Muralla/Hespérides.  
Una introducción a la Estadística bajo la aproximación del análisis exploratorio.
- García Ferrando, M. (1999): *Socioestadística: introducción a la estadística en sociología*, Madrid: Alianza.  
Introducción a la Estadística con multitud de ejemplos sociológicos.
- González, J. J. (1995): «Clases y alineamiento electoral al final del ciclo político», en J. Carabaña (ed.), *Desigualdad y Clases Sociales*, Madrid: Fundación Argentaria.
- Greene, W. H. (2008): *Econometric Analysis* (6<sup>a</sup> ed.), Englewood Cliffs (NJ): Prentice Hall.  
Introduce a los estudiantes en la econometría aplicada, incluyendo técnicas básicas de análisis de regresión. Comienza con una serie de capítulos instrumentales sobre álgebra matricial, probabilidades y estadística.
- Gujarati, D. N. y D. C. Porter (2008): *Basic Econometrics*, Nueva York: McGraw-Hill Education.  
Un clásico de la econometría. Por sus ejemplos, ejercicios y, sobre todo, explicaciones claras es un buen libro para introducirse en el estudio de la regresión y sus problemas.
- Hair, J. F. et al. (2006): *Multivariate Data Analysis*. Londres: Prentice-Hall International.  
Libro con muchas técnicas de análisis multivariadas acompañadas de artículos que la emplean. Prescinde de fórmulas matemáticas y se centra en la comprensión de los conceptos y en la interpretación de las tablas y los gráficos estadísticos.

- Hamilton, L. C. (2009): *Statistics with Stata. Updated for Version 10*, Belmont (CA): Thomson.
- Este es uno de los mejores manuales de Stata. Incluye las instrucciones y las interpretaciones estadísticas. Prácticamente con cada versión de Stata sale una edición distinta de este libro.
- Hilbe, J. M. (2009): *Logistic Regression Models*, Nueva York: Chapman & Hall /CRC.
- Libro monográfico sobre la regresión logística y sus extensiones, incluyendo logística con datos de panel. Muestra ejemplos y usos con Stata y R.
- Hosmer, D. W. y S. Lemeshow (2000): *Applied Logistic Regression*, Nueva York: John Wiley & Sons.
- Un buen manual de referencia de las regresiones logísticas, ordinales y multinomiales con muchas referencias a Stata.
- Huber, P. J. (1967): «The Behaviour of Maximum Likelihood Estimates under Non-Standard Conditions, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley (CA): University of California Press
- Jann, B. (2005): «Tabulation of Multiple Responses», *The Stata Journal*, 5(1): 92-122.
- Artículo donde se documenta el uso de la instrucción mrtab para la elaboración de tablas de respuesta múltiple con Stata.
- Jovell, A. J. (1995): *Análisis de Regresión Logística*, Madrid: Centro de Investigaciones Sociológicas.
- Manual de esta colección sobre regresión logística. Elemental.
- Kish, L. (1982): *Muestreo de encuestas*, México: Trillas.
- Otro de los clásicos del muestreo.
- Kohler, U. y F. Kreuter (2009): *Data Analysis Using Stata*, College Station (TX): Stata Press.
- Manual de análisis elemental con Stata que emplea numerosos ejemplos del panel socio-económico alemán y del DIW (Instituto Alemán de Investigaciones Económicas). Su primera versión fue escrita en ese idioma. Su traducción al inglés es una señal de su carácter práctico. Su núcleo son las tablas, los gráficos y las regresiones lineales y logísticas.
- Lawal, B. (2003): *Categorical Data Analysis with Sas and Spss Applications*, Mahwah (NJ): Lawrence Erlbaum Associates.
- Aunque no incluye Stata, se trata de un buen libro para avanzar en el análisis de datos nominales: se extiende desde las tablas de frecuencias hasta clasificación de casos dudosos, incluyendo también modelos log-lineales y regresiones logísticas.
- Li, G. (1985): «Robust Regression», en D. C. Hoaglin, F. Mosteller y J. W. Tukey (eds.), *Exploring Data Tables, Trends, and Shapes*, Nueva York: Wiley.
- Long, J. S. (2009): *The Workflow of Data Analysis Using Stata*, College Station (TX): Stata Press.
- Manual de cómo planificar el trabajo con Stata con el fin de realizar análisis buenos y eficientes. Presta también especial atención a cómo escribir programas útiles. Lleno de ejemplos en Ciencias Sociales.
- y J. Freese (2006): *Regression Models for Categorical Dependent Variables Using Stata*. College Station (TX): Stata Press.
- Un libro muy pedagógico para aprender regresiones logísticas y sus derivadas si necesidad de tener conocimientos matemáticos. Contiene un primer capítulo sobre instrucciones de Stata. También es la base de los programas Spost, que son explicados en este libro en los capítulos correspondientes.

- Maddala, G. S. (2001): *Introduction to Econometrics* (3<sup>a</sup> ed.), Chichester: John Wiley & Sons.  
Otro clásico de la econometría.
- Mitchell, M. N. (2008): *A Visual Guide to Stata Graphics*, College Station (TX): Stata Press.  
Más que por sus explicaciones, este libro sobresale por ser un amplio catálogo de gráficos acompañados por los códigos con los que puede obtenerse con Stata.  
Muy útil para quien quiera pasar de los gráficos de Excel o SPSS a Stata.
- Neter, J. et al. (1993): *Applied Statistics*, Boston (MA): Allyn and Bacon.  
Libro completo de Estadística básica, entre lo más recomendados en la década de los 90 para cursos intermedios de la materia.
- Novales, A. (1989): *Econometría*, Madrid: McGraw Hill.  
Manual introductorio a la econometría, principalmente destinado a cursos de grado en Economía.
- Olzak, S. (1992): *The Dynamics of Ethnic Competition and Conflict*, Stanford (CA): Stanford University Press.
- Paramio, L. (2000): «Clase y voto: intereses, identidades y preferencias», *Revista Española de Investigaciones Sociológicas*, 90: 79-93.
- Peña, D. (1989a): *Estadística, modelos y métodos. (Vol 1): Fundamentos*, Madrid: Alianza Univ.  
— (1989b): *Estadística, Modelos y Métodos. (Vol. 2): Modelos Lineales y Series Temporales*. Madrid, Alianza Universidad.
- (2002): *Regresión y diseño de experimentos*, Madrid: Alianza Editorial.  
Versión revisada del segundo volumen de la obra Estadística: Modelos y Métodos. En primer lugar se abordan los modelos de diseño experimental. En segundo lugar, se presentan los modelos de regresión, que ocupan la mayor parte del libro, y donde se estudia la relación entre una variable respuesta y un conjunto de variables explicativas que, en general, no son controladas por el investigador.
- (2008): *Fundamentos de estadística*, Madrid: Alianza Editorial.  
Versión revisada del primer volumen la obra Estadística: Modelos y Métodos. Se estructura siguiendo las etapas de construcción de un modelo estadístico.
- y J. Romo (2003): *Introducción a la estadística para las ciencias sociales*, Madrid: McGraw-Hill.  
Libro riguroso y muy básico, ideal para principiantes.
- Pérez López, C. (2005): *Muestreo estadístico. Conceptos y problemas resueltos*, Madrid: Pearson Educación.  
Manual de muestreo con gran cantidad de problemas muestrales resueltos con Excel y SPSS.
- Petersen, T. (1995): «Analysis of Event History», G. Arminger et al. (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Nueva York: Plenum Press.
- Pregibon, D. (1981): «Logistic Regression Diagnostics», *The Annals of Statistics*, 9(4): 705-724.  
Artículo en el que se desarrollan medidas de diagnóstico para detectar comportamientos extraños en los modelos logísticos.
- Rabe-Hesketh, S. y B. Everitt (2007): *A Handbook of Statistical Analyses Using Stata* (4<sup>a</sup> ed.). Boca Raton (FL): Chapman & Hall.  
Es un libro formado por capítulos independientes de investigaciones médicas y epidemiológicas, empleando modelos y análisis complejos. Cada capítulo está acompañado de buenos ejercicios.

- y A. Skrondal (2008): *Multilevel and Longitudinal Modeling Using Stata*, College Station (TX): Stata Press.  
Libro centrado en los análisis multinivel que permiten la combinación de efectos fijos y aleatorios. Para quien tenga un buen nivel de estadística.
- Raftery, A. E. (1996): «Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models», *Biometrika*, 83(2): 251-266.
- Ramsey, J. B. (1969): «Test for Specification Error in Classical Linear Least-Squares Regression Analysis», *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 31: 350-371.  
Artículo donde se presenta el test de Ramsey
- Rodríguez Osuna, J. (1991): *Métodos de Muestreo*, Madrid: Centro de Investigaciones Sociológicas.  
Texto básico sobre técnicas de muestreo que incluye tanto cuestiones teóricas como prácticas de diseños muestrales.
- Ruiz-Maya, L. (dir.) (1990): *Metodología estadística para el análisis de datos cualitativos*, Madrid: Centro de Investigaciones Sociológicas.  
Libro dedicado al análisis de encuestas. Desde la tabla de contingencia a los modelos log-lineales.
- et al. (1995): *Análisis Estadístico de Encuestas: Datos Cuantitativos*, Madrid: AC.  
Libro dedicado al análisis estadístico de datos nominales. Desde la tabla de contingencia a los modelos log-lineales.
- Sánchez Carrión, J. J. (1989): *Ánalisis de tablas de contingencia: el uso de los porcentajes en las ciencias sociales*, Madrid: Centro de Investigaciones Sociológicas..  
Técnicas elementales para el estudio de tablas de contingencia. Centrado en diferencias de porcentajes. Incluye un capítulo sobre el estudio del cambio.
- Sørensen, A. y N. Tuma (1981): «Labor Market Structures and Job Mobility», *Research in Social Stratification and Mobility*, 1: 67-94.
- Spiegel, M. R. (1970): *Estadística*, México: McGraw Hill.  
Libro de clásico de Estadística elemental, lleno de ejercicios resueltos. Muy útil como autoguía para resolver problemas de Estadística.
- y L. J. Stephens (2008): *Schaum's Outline of Theory and Problems of Statistics* (4<sup>a</sup> ed.): Nueva York, McGraw-Hill.  
Última edición del libro de clásico de Estadística elemental, con explicaciones básicas y repleto de ejercicios resueltos.
- Stata Corporation (2011a): *Stata Quick Reference and Index. Release 12*, College Station (TX): Stata Press.  
Es el índice de todos los demás volúmenes. Contiene también una interesante clasificación de los comandos. En el manual digitalizado en formato pdf se encuentra al principio (*Contents*) y al final (*Index y Subject index*).
- (2011b): *Getting Started with Stata. Release 1*, College Station (TX): Stata Press.  
Una buena guía para quienes comienzan a usar el programa con los aspectos y estadísticos esenciales. Hay tres versiones, una para cada sistema operativo (Mac, Unix, Windows).
- (2011c): *Stata User's Guide. Release 12*, College Station (TX): Stata Press.  
Volumen de la documentación de Stata elemental. Después de la referencia básica es lo que debe aprenderse para dominar los elementos comunes de Stata, como las funciones, los formatos, la sintaxis y las cuestiones básicas de programación.
- (2011d): *Stata Data-Management Reference Manual. Release 12*, College Station (TX): Stata Press.

- Este libro contiene todas las instrucciones relacionadas con el manejo de ficheros: creación, modificación, lectura, escritura, fusión, recorte y transformaciones de formato.
- (2011e): *Stata Base Reference Manual. Release 12*, College Station (TX): Stata Press.  
Este libro con tres volúmenes contiene la mayor parte de instrucciones de análisis básicos de Stata: estadísticos, tablas y regresiones salvo las temporales, las de panel, las de supervivencia e imputación de valores.
  - (2011f): *Stata Graphics Reference Manual. Release 12*, College Station (TX): Stata Press.  
Todo un volumen dedicado a las instrucciones generales de gráficos, pues las específicas se documentan con su respectiva técnica. Salvo las primeras páginas, sólo útil para quienes no empleen el editor de gráficos.
  - (2011g): *Stata Survival Analysis and Epidemiological Tables Reference Manual. Release 12*, College Station (TX): Stata Press.  
Volumen de la documentación de Stata especializado en las órdenes *st* del análisis histórico de acontecimientos y las tablas de supervivencia.
  - (2011h): *Stata Survey Data Reference Manual. Release 12*, College Station (TX): Stata Press.  
Volumen dedicado a las instrucciones *svy*, para la ponderación.
  - (2011i): *Stata Programming Reference Manual. Release 12*, College Station (TX): Stata Press.  
Volumen dedicado a la programación.
  - (2011j): *Stata Structural Equation Modeling Reference Manual. Release 12*, College Station (TX): Stata Press.  
Volumen dedicado a los modelos de ecuaciones estructurales.
- Strang, D. (1994): «Introduction to Event History Analysis», en T. Janoski *et al.* (eds.), *The Comparative Political Economy of the Welfare State*, Cambridge: Cambridge University Press.
- Vermunt, J. (1997): *Log-Linear Models for Event Histories*, Londres: Sage.
- White, H. (1982): «Maximun Likelihood Estimation of Misspecified Models», *Econometrica*, 50(1): 1-25.
- Wooldridge, J. M. (2009): *Introductory Econometrics: A Modern Approach* (4<sup>a</sup> ed.), Australia: Thomson South Western.  
Muy interesante libro de econometría con un enfoque muy moderno de nivel intermedio y lleno de ejemplos útiles. Estos se encuentran desarrollados en Stata en <http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge.html>.
- World Bank, The (2008): *World Development Indicators 2008* (en CD-ROM), Washington D. C.: The World Bank.
- Yamaguchi, K. (1991): *Event History Analysis*, Londres: Sage.



# 16

## Índice de instrucciones

### 1. Generales

*append*, 79

*aweight*, 102

*browse*, 30

*by*, 117

*bysort*, 126, 222, 225

*cd*, 64

*cmdlog close*, 42

*cmdlog off*, 42

*cmdlog on*, 42

*cmdlog using*, 42

*codebook*, 86

*compress*, 49, 65

*db*, 37

*describe*, 45

*dir*, 26

*display*, 53, 142

*do*, 32

*doedit*, 31

*drop*, 55, 120

*edit*, 29, 59

*egen*, 139

*estimates*, 403

*for*, 145, 281, 306

*format*, 50

*fweight*, 100

*generate*, 129

*global*, 143, 279

*gsort*, 119

*help*, 28

*if*, 120

*in*, 51, 120

*infile con formato libre*, 67

*infile de ancho fijo*, 69

*infix*, 68

*insheet*, 65

*iweight*, 103

*joinby*, 80

*keep*, 55

*label data*, 46

*label define*, 46

*label drop*, 47

*label list*, 47

*label save*, 47

*label values*, 46

*label variable*, 46

*labelbook*, 47

*list*, 50

*log close*, 42

*log off*, 42

*log using*, 41

*mark*, 235

*markout*, 235

*merge*, 79

*net from*, 247, 266

*net install*, 235, 247, 266

*outfile*, 73

*pweight*, 102

*quietly*, 402

*recode*, 134

*rename*, 61

*replace*, 130

*reshape*, 127, 230

*return*, 142  
*run*, 32  
*sample*, 231  
*save*, 63  
*saveold*, 74  
*set dp*, 50  
*set memory*, 44  
*set seed*, 132  
*sort*, 117  
*ssc install*, 157, 431  
*stset*, 451  
*svydes*, 477  
*svyset*, 477  
*sysuse*, 27  
*sysuse dir*, 19  
*use*, 44  
*varmanage*, 54  
*view*, 43

## 2. Estadísticas

*anova*, 229  
*asmprobit*, 440  
*brant*, 432  
*ci*, 113  
*cii*, 112  
*correlate*, 281  
*dfbeta*, 348  
*estat classification*, 397  
*estat effects*, 481  
*estat ic*, 399  
*fitstat*, 394  
*friedman*, 235  
*hausman*, 441  
*hettest*, 338  
*kwallis*, 221  
*listcoef*, 410  
*logit*, 383  
*margins*, 412  
*mlogit*, 435  
*mlogtest*, 441  
*mprobit*, 440  
*mrtab*, 266

*ologit*, 426  
*omodel*, 431  
*oneway*, 224  
*ovtest*, 340  
*prchange*, 412  
*predict*, 288, 334, 390  
*prgen*, 415  
*prtest*, 199  
*prvalue*, 418  
*qreg*, 369  
*ranksum*, 219  
*regress*, 286, 334  
*robvar*, 224  
*rreg*, 363  
*sdtest*, 216  
*sfrancia*, 336  
*signrank*, 211  
*signtest*, 203  
*sktest*, 336  
*streg*, 462  
*sts list*, 456  
*summarize*, 96, 120, 142  
*svy: proportion*, 479  
*svy:logit*, 485  
*svy:mean*, 480  
*svy:regress*, 485  
*svy:tabulate*, 483  
*swilk*, 222, 336  
*tab1*, 88, 238  
*tabchi*, 247  
*table*, 262  
*tabstat*, 232, 259  
*tabulate*, 88, 239  
*test*, 295  
*ttest*, 201  
*vif*, 339  
*vwls*, 360

## 3. Gráficas

*avplots*, 342  
*catplot*, 157  
*cluster dendrogram*, 150

- dotplot*, 150  
*graph*, 150  
*graph bar*, 156  
*graph box*, 167, 345  
*graph combine*, 151  
*graph copy*, 151  
*graph describe*, 151  
*graph dir*, 151  
*graph display*, 151  
*graph drop*, 151  
*graph export*, 152  
*graph hbar*, 162  
*graph matrix*, 174, 339  
*graph pie*, 153  
*graph query*, 186  
*graph rename*, 151  
*graph save*, 152  
*graph twoway*, 169  
*graph twoway area*, 176  
*graph twoway bar*, 173  
*graph twoway connected*, 175  
*graph twoway dot*, 173  
*graph twoway dropline*, 173  
*graph twoway fpfit*, 178  
*graph twoway fpfitci*, 181  
*graph twoway function*, 182  
*graph twoway kdensity*, 166  
*graph twoway lfit*, 177  
*graph twoway lfitci*, 181  
*graph twoway line*, 175  
*graph twoway lowess*, 178  
*graph twoway mband*, 178, 373  
*graph twoway mspline*, 178  
*graph twoway qfit*, 178  
*graph twoway qfitci*, 181  
*graph twoway rarea*, 416  
*graph twoway rbar*, 180  
*graph twoway rcap*, 180  
*graph twoway rcapsim*, 180  
*graph twoway rconnected*, 180  
*graph twoway rline*, 180  
*graph twoway rspike*, 180  
*graph twoway scatter*, 170  
*graph twoway spike*, 173  
*greigen*, 150  
*histogram*, 28, 163  
*kdensity*, 337  
*lvr2plot*, 347  
*marginsplot*, 412  
*mlogplot*, 435  
*mlogview*, 439  
*query graphics*, 186  
*rvfplot*, 150, 335  
*scatter*, 278, 368, 391  
*set scheme*, 187  
*stem*, 150  
*sts graph*, 456



































## Números publicados

1. **Métodos de muestreo**  
Jacinto Rodríguez Osuna
2. **Metodología de la evaluación de programas**  
Francisco Alvira Martín
3. **Métodos de análisis causal**  
Juan Díez Medrano
4. **Análisis de regresión múltiple**  
Mauro F. Guillén
5. **El método biográfico: el uso de las historias de vida en ciencias sociales**  
Juan José Pujadas Muñoz
6. **Métodos de muestreo. Casos prácticos**  
Jacinto Rodríguez Osuna
7. **Gráficos**  
Antonio Alaminos
8. **Programación de la investigación social**  
Ignasi Pons
9. **Encuestas telefónicas y por correo**  
J. Lluís C. Bosch y Diego Torrente
10. **Investigación participativa**  
Luis R. Gabarrón y Libertad Hernández Landa
11. **Encuestas de salud**  
María D. Navarro Rubio
12. **Modelos probabilísticos de elección**  
Silvia de la Vega Gómez
13. **Fuentes de información demográfica en España**  
David-Sven Reher y Ángeles Valero Lobo
14. **Análisis de datos con SPSS/PC+**  
José Luis Álvaro Estramiana y Alicia Garrido Luque

- 15. Análisis de regresión logística**  
Albert J. Jovell
- 16. Análisis y estructural y de redes**  
Josep A. Rodríguez
- 17. Auto/biografías**  
Jesús M. de Miguel
- 18. Redes sociales y cuestionarios**  
Félix Requena Santos
- 19. Escalas de prestigio profesional**  
Julio Carabaña Morales y Carmuca Gómez Bueno
- 20. Observación participante**  
Óscar Guasch
- 21. Metodología del análisis comparativo**  
Jordi Caïs
- 22. Metodología cualitativa en España**  
Bernabé Sarabia y Juan Zarco
- 23. Evaluación de la investigación**  
Joan Bellavista, Elena Guardiola, Aida Méndez y María Bordons
- 24. Bancos de datos**  
Magdalena Cordero Valdavia
- 25. Análisis dinámico**  
Emilio J. Castilla
- 26. Cuestionarios**  
María José Azofra
- 27. Análisis de datos electorales**  
Pablo Oñate y Francisco A. Ocaña
- 28. Metodología de la Ciencia Política**  
Eva Anduiza Perea, Ismael Crespo y Mónica Méndez Lago
- 29. Elección racional**  
Pau Marí-Klose

- 30. Estudio de casos**  
Xavier Coller
- 31. Diarios de campo**  
Juan M. García Jorba
- 32. Entrevistas cualitativas**  
Miguel S. Valles
- 33. Introducción a las matemáticas para las ciencias sociales**  
Francisca Blanco Moreno
- 34. Teoría de juegos**  
Ignacio Sánchez-Cuenca
- 35. La encuesta: una perspectiva general metodológica**  
Francisco Alvira Martín
- 36. Manual de trabajo de campo en la encuesta**  
Vidal Díaz de Rada
- 37. «Grounded Theory»: La constitución de la teoría a través del análisis interpretacional**  
Antonio Trinidad Requena, Virginia Carrero Planes y Rosa M.<sup>a</sup> Soriano Miras
- 38. Análisis de la Historia de Acontecimientos**  
Fabrizio Bernardi
- 39. El análisis de segmentación: técnicas y aplicaciones de los árboles de clasificación**  
Modesto Escobar Mercado
- 40. Evolución de la Teoría Fundamentada como técnica de análisis cualitativo**  
Jaime Andréu Abela, Antonio García-Nieto y Ana M<sup>a</sup> Pérez Corbacho
- 41. Dinámica del grupo de discusión**  
Jesús Gutiérrez Brito
- 42. Encuesta deliberativa**  
María Cuesta, Joan Font, Ernesto Ganuza, Braulio Gómez y Sara Pasadas

- 43. Análisis sociológico del sistema de discursos**  
Fernando Conde Gutiérrez del Álamo
- 44. La investigación sobre el uso del tiempo**  
M<sup>a</sup> Ángeles Durán Heras, Jesús Rogero García
- 45. Análisis de datos con Stata**  
Modesto Escobar Mercado, Enrique Fernández Macías,  
Fabricio Bernardi
- 46. Análisis de datos incompletos en Ciencias Sociales**  
Gonzalo Rivero Rodríguez

**Modesto Escobar Mercado**, es doctor en Sociología por la Universidad Complutense de Madrid y catedrático de Sociología en el Departamento de Sociología y Comunicación de la Facultad de Ciencias Sociales de la Universidad de Salamanca, del que fue su primer director. Ha publicado libros como *El análisis gráfico/exploratorio* (1999) y *El análisis de segmentación: técnicas y aplicaciones de los árboles de clasificación* (2007), aparecido también en esta colección. Es autor además, entre otros trabajos, de «Redes semánticas en textos periodísticos: una propuesta metodológica para su descubrimiento» (*Empiria*, 2009), «La presentación del self en el ciberespacio. Un análisis de las autodefiniciones personales en blogs y redes sociales» (*RPS*, 2011) y «La calidad democrática: una propuesta para su medición por expertos» (*REIS*, 2011). Su área principal de trabajo son las técnicas de investigación social.

**Enrique Fernández Macías**, es profesor en el Departamento de Sociología y Comunicación de la Universidad de Salamanca, e investigador en la Fundación Europea para la Mejora de las Condiciones de Vida y Trabajo, de Dublín. Es doctor en Sociología por la Universidad de Salamanca, y sus áreas de investigación son la sociología del trabajo y la economía laboral, principalmente en el ámbito europeo. Algunas de sus publicaciones recientes son «Job Polarization in Europe?» (*Work and Occupations*, 2012); «E Pluribus Unum? A Critical Survey of Job Quality Indicators» (*Socio-Economic Review*, 2011) con Muñoz de Bustillo, Esteve y Antón; *Transformations of the Employment Structure in the EU and the US, 1995–2007* (Palgrave-Macmillan, 2012) con Storrie y Hurley, y *Measuring More than Money: the Social Economics of Job Quality* (Edward Elgar, 2011) con Muñoz de Bustillo, Esteve y Antón.

**Fabrizio Bernardi**, es profesor de Sociología en el Instituto Universitario Europeo, Florencia. Doctor en Sociología por la Universidad de Trento, ha sido profesor de Estructura Social Contemporánea en la UNED y en la Universidad de Bielefeld y de Métodos de Investigación en la Universidad de Bolonia. Sus publicaciones más recientes incluyen: «Unequal Transitions: Selection Bias and the Compensatory Effect of Social Background in Educational Careers» (*Research in Social Stratification and Mobility*, 2012); «Female Education and Marriage Dissolution: Is it a Selection Effect?» (*European Sociological Review*, 2011) con Martínez-Pastor, y «The Recent Fast Upsurge of Immigrants in Spain and their Employment Patterns and Occupational Attainment» (*International Migration*, 2011) con Garrido y Miyar. Sus áreas de investigación principales son la desigualdad social y las dinámicas familiares y laborales.

ISBN 978-84-7476-483-3



9 788474 764833



GOBIERNO  
DE ESPAÑA

MINISTERIO  
DE LA PRESIDENCIA

**CIS**

Centro de Investigaciones Sociológicas