

17.871

Primavera de 2002

Cómo utilizar los comandos **merge** y **reshape** de *STATA*

La mayoría de los proyectos del curso 17.871 y, de hecho, la mayor parte de las investigaciones interesantes, requieren la combinación de conjuntos de datos. Esta entrega repasa el comando más valioso para gestionar múltiples conjuntos el comando **merge**. Además, muchas veces queremos combinar varias observaciones de una unidad de análisis (países, estados, personas) para crear un *panel*. El comando **reshape** sirve para moverse entre las diferentes organizaciones de los datos.

El comando **merge**

Asumamos que estamos investigando el efecto que tiene la utilización de distintos sistemas para votar sobre la tendencia de los votantes a emitir votos “en blanco” o “nulos” (esto es, no registrar ningún voto o registrar múltiples votos). Es posible que un solo conjunto de datos registre el número de votos emitidos y el número de votos en blanco en cada ciudad. Observemos este ejemplo:

Ejemplo 1. *ballots.dta*

town	blank92	ballot92	blank96	ballot96
Barnstable	1163	22747	232	22467
Bourne	125	7885	61	8032
Brewster	42	5480	62	5498
Chatham	35	4558	103	4475
Dennis	294	8732	86	8493
Eastham	20	3013	26	3151
Falmouth	381	16377	169	16224
Harwich	184	6741	89	6737
Mashpee	22	4619	46	4962
Orleans	109	4251	113	4229
Provincetown	15	2488	6	2268
Sandwich	35	9151	62	9757
Truro	3	1161	7	1156
Wellfleet	14	1815	24	1768
Yarmouth	441	12862	104	12710

La variable *town* identifica una ciudad, *blank92* y *blank96* registran el número de votos en blanco en esa ciudad recabados en 1992 y en 1996, y *ballot92* y *ballot96* registran el número total de votos emitidos en la ciudad. Supongamos que hemos guardado este conjunto de datos en el archivo *ballots.dta*

Otro conjunto de datos podría registrar el método para votar utilizado en la población, así:

Ejemplo 2. machines.dta

town	method92	method96
Barnstable	Paper	AccuVote
Bourne	AccuVote	AccuVote
Brewster	Optech	Optech
Chatham	Paper	Optech
Dennis	Optech	Optech
Eastham	Paper	Optech
Falmouth	Paper	AccuVote
Harwich	Optech	Optech
Mashpee	Paper	Optech
Orleans	Optech	Optech
Provincetown	Paper	Paper
Sandwich	Optech	Optech
Truro	Paper	Paper
Wellfleet	Paper	Paper
Yarmouth	Optech	Optech

Como antes, *town* es la ciudad. Las variables *method92* y *method96* registran la clase de mecanismo para votar que se utilizó en 1992 y 1996. Los datos se guardan en el archivo machines.dta.

Tenga en cuenta la siguiente información importante: ballots.dta y machines.dta tienen una variable común, *town*, que identifica de forma exclusiva los casos. (En este ejemplo, el nombre de la ciudad identifica los casos. Para casos más complejos, es preferible utilizar una variable numérica como el identificador conjunto).

Para combinar ambos conjuntos de datos, siga los siguientes pasos:

- (1) Ordene ambos conjuntos por la variable de identificación común y guárdelos ordenados.
- (2) Utilice uno de los conjuntos de datos.
- (3) Introduzca el comando **merge**, con la siguiente sintaxis:
merge commonvariable using remotefilename

Este conjunto de comandos aumentará el conjunto de datos que había utilizado (en el paso 2), añadiendo las variables del archivo remoto. Además, se creará una nueva variable denominada *_merge*. La variable *_merge* equivaldrá a 3 si el caso estaba en ambos conjuntos de datos, a 1 si el caso estaba en el conjunto “maestro” (esto es, el utilizado en el paso 2) pero no en el conjunto “en uso” y a 2 si el caso estaba en el conjunto “en uso” pero no originalmente en el conjunto de datos “maestro”.

Por tanto, los siguientes comandos:

```
use ballots
sort town
save ballots,replace
use machines
sort town
```

**save machines,replace
use ballots
merge town using machines**

producirían el siguiente conjunto de datos:

Ejemplo 3. Conjunto de datos combinado

town	blank92	ballot92	blank96	ballot96	method92	method96	_merge
Barnstable	1163	22747	232	22467	Paper	AccuVote	3
Bourne	125	7885	61	8032	AccuVote	AccuVote	3
Brewster	42	5480	62	5498	Optech	Optech	3
Chatham	35	4558	103	4475	Paper	Optech	3
Dennis	294	8732	86	8493	Optech	Optech	3
Eastham	20	3013	26	3151	Paper	Optech	3
Falmouth	381	16377	169	16224	Paper	AccuVote	3
Harwich	184	6741	89	6737	Optech	Optech	3
Mashpee	22	4619	46	4962	Paper	Optech	3
Orleans	109	4251	113	4229	Optech	Optech	3
Provincetown	15	2488	6	2268	Paper	Paper	3
Sandwich	35	9151	62	9757	Optech	Optech	3
Truro	3	1161	7	1156	Paper	Paper	3
Wellfleet	14	1815	24	1768	Paper	Paper	3
Yarmouth	441	12862	104	12710	Optech	Optech	3

La variable *_merge* confirma que hemos introducido un conjunto equilibrado de casos de ambos conjuntos de datos. (En otras palabras, confirma que no hay ciudades con resultados electorales para las que no tengamos los datos del sistema para votar y viceversa.) Sin embargo, se trata de una variable molesta una vez completada la combinación, así que es recomendable **eliminarla** una vez que esté convencido de que todo es correcto.

El comando **reshape**

El conjunto de datos del Ejemplo 3 podría utilizarse para analizar el porcentaje de votos en blanco en 1992 y en 1996, por separado, en función de la tecnología de voto utilizada. Y, podría utilizarse para estudiar el *cambio* en el número de votos en blanco desde 1992 a 1996, en función del *cambio* en la tecnología de voto. Sin embargo, tal vez nos interese tratar el conjunto de datos como observaciones por año de 30 ciudades en lugar de 15 observaciones separadas. Para ello, tendríamos que “apilar” las diferentes observaciones de las ciudades, y obtener en primer lugar las variables de 1992 correspondientes a los votos en blanco, los emitidos y la tecnología de voto (identificadas por año), seguidas de las de 1996 (también identificadas en función del año). Para ello utilizamos el comando **reshape**.

Observe primero la convención nominal que hemos adoptado en los conjuntos anteriores: cada nombre de variable consiste en una raíz (*blank* o *ballot*) seguida por el año en cuestión (92 o 96).

Entonces, el siguiente comando:

reshape long blank ballot method, i(town) j(year)

“reestructurará” el conjunto de datos del Ejemplo 3 del siguiente modo:

Ejemplo 4. Conjunto de datos reestructurado y combinado

town	blank	ballot	method	year
Barnstable	1163	22747	Paper	92
Bourne	125	7885	AccuVote	92
Brewster	42	5480	Optech	92
Chatham	35	4558	Paper	92
Dennis	294	8732	Optech	92
Eastham	20	3013	Paper	92
Falmouth	381	16377	Paper	92
Harwich	184	6741	Optech	92
Mashpee	22	4619	Paper	92
Orleans	109	4251	Optech	92
Provincetown	15	2488	Paper	92
Sandwich	35	9151	Optech	92
Truro	3	1161	Paper	92
Wellfleet	14	1815	Paper	92
Yarmouth	441	12862	Optech	92
Barnstable	232	22467	AccuVote	96
Bourne	61	8032	AccuVote	96
Brewster	62	5498	Optech	96
Chatham	103	4475	Optech	96
Dennis	86	8493	Optech	96
Eastham	26	3151	Optech	96
Falmouth	169	16224	AccuVote	96
Harwich	89	6737	Optech	96
Mashpee	46	4962	Optech	96
Orleans	113	4229	Optech	96
Provincetown	6	2268	Paper	96
Sandwich	62	9757	Optech	96
Truro	7	1156	Paper	96
Wellfleet	24	1768	Paper	96
Yarmouth	104	12710	Optech	96

Ahora tenemos un nuevo conjunto de datos con el doble de observaciones que el anterior: cada ciudad tiene dos observaciones, una para 1992 y la otra para 1996. Para cada observación anual correspondiente a una ciudad tenemos el número de votos en blanco, el total de votos emitidos y la tecnología de voto utilizada.

El comando

reshape wide blank ballot method, i(town) j(year)

invertirá el proceso, restaurando el conjunto de datos al Ejemplo 3 (sin *_merge*).