

Stata Tips para la Encuesta Panel CASEN

Stata Tip 2: Pon tus datos en forma. Estrategias para declarar datos en formato de panel en Stata

Existen distintos formatos de datos en Stata. Si el objetivo de la investigación que se pretende realizar con la EPCasen contempla abordar temas como heterogeneidad no observada, correlación serial, o efectos dinámicos, evidentemente se tienen que ocupar análisis específicos para datos de panel. En el caso de Stata, los comandos para tales análisis comienzan con las letras `xt` (Stata 2007a). Una condición para ocupar dichas opciones es que los datos estén en formato “long”. Sin embargo, las bases de datos de EPCasen disponibles en Internet se encuentran en formato “wide”. En este Tip se muestran estrategias a través de las cuales las bases de datos de EPCasen pueden ser transformadas desde formato wide a long y vice versa.

Con el fin de ejemplificar la diferencia entre ambas estructuras de datos se va a trabajar con la base `panelcasen_m04.dta`, la que está en formato wide. Sin embargo, antes de ilustrar las estructuras de los dos formatos ya mencionados, se mostrará cómo trabajar con el working directory en Stata, lo cual es muy útil, especialmente cuando se tienen distintas bases de datos, como, por ejemplo, en el caso de la EPCasen, o cuando se quiere automatizar los análisis. Se puede determinar el actual working directory con el cual Stata está trabajando a través del siguiente comando

```
. pwd  
C:\data
```

Stata está trabajando con el directorio `C:\data` en mi computadora. Sin embargo, la base de datos mencionada arriba se encuentra en el directorio `c:\data\Tips`. Y mi objetivo es trabajar directamente con esta dirección. En este sentido, para cambiarme a éste último directorio, usaré el comando `cd`. Específicamente, para moverse al propuesto directorio y obtener una descripción de él, escribe

```
. cd c:\data\Tips  
. dir  
  
<dir> 11/09/08 17:02 .  
<dir> 11/09/08 17:02 ..  
40.8M 11/26/07 12:30 panelcasen_m04.dta  
0.7k 10/03/08 16:24 Tip2.do
```

`dir` por sí solo entrega información de todos los archivos en el directorio, sin embargo, también existe la opción de obtener información sólo de las bases de datos o de los Do files. En tal caso, se tiene que escribir `dir *.dta` para las bases o `*.do` para los Do files.

Una vez que hemos establecido el directorio con el que se quiere trabajar, la base de datos se puede obtener simplemente escribiendo `use` y el nombre de los datos con los que estamos trabajando. Del mismo modo, los Do files se obtienen con el comando `doedit` más el nombre del respectivo archivo. Se escriben los siguientes comandos

```
. set mem 50000  
. use panelcasen_m04  
. doedit Tip2
```

Con los datos ya en Stata, mi objetivo es ilustrar la diferencia entre formatos wide y long. La base de datos `panelcasen_m04` es una típica base en formato wide. Para ver de qué trata esto, es escrito

```
. keep idpersona pesos_long_96_01_06 ypccorh_96 ypccorh_01 ypccorh_06

. list idpersona ypccorh_96 ypccorh_01 ypccorh_06 ///
pesos_long_96_01_06 if idpersona==7 | idpersona ==12
```

```
+-----+
| idpers~a  ypcco~96  ypcco~01  ypcco~06  pesos~06 |
+-----+
3. |          7      51333      59272      52338  175.0152 |
10. |         12      47500      71833     143750  60.41674 |
+-----+
```

con el comando `keep` se modifica la base de datos de modo que se dispone solo de las variables `idpersona` `pesos_long_96_01_06` `ypccorh_96` `ypccorh_01` `ypccorh_06`¹. A su vez, con el comando `list` se puede ver el modo en que luce la base de datos para dichas variables en los casos de las personas con los números de identificación 7 y 12². Como queda en evidencia, las variables `idpersona` y `pesos_long_96_01_06` son constantes en el tiempo, mientras que el resto, que son las variables sobre ingreso per cápita corregido por imputación, presenta varianza en la dimensión temporal. Una descripción general de los datos se puede obtener mediante el comando `describe` o `des` en corto

```
. des

Contains data from panelcasen_m04.dta
  obs:          26,882
  vars:           5                26 Nov 2007 12:30
  size:        645,168 (98.7% of memory free)
(output omitido)
```

La base de datos en formato wide tiene cinco variables y se compone de 26.882 individuos. Ahora bien, para transformar los datos desde formato wide a long se requiere usar el comando `reshape`. Adicionalmente, mostraré el comando `soepren`, que es un complemento del primero. Específicamente, `soepren` no es un archivo oficial de Stata. Técnicamente, este tipo de archivo es clasificado como user-written Stata command porque ha sido creado por un usuario privado de Stata. El comando `soepren` ha sido creado por Ulrich Kohler especialmente para re-nombrar las variables de panel alemán GSOEP (Kohler and Kreuter 2008), sin embargo, éste puede ser usado con otros paneles como el EPCasen. El comando está disponible en el SSC archivo online. Para instalarlo en tu computador, se escribe

```
. ssc install soepren
```

Una descripción del comando lo puedes obtener a través de `help soepren`. Para efectos de este Tip, escribe

```
. soepren ypccorh_96 ypccorh_01 ypccorh_06, newstub(ypccorh) ///
waves(1996 2001 2006)
```

¹ El ejemplo aquí presentado se limita a estas variables. Sin embargo, los comandos mostrados en este Tip son extensibles, bajo la misma lógica aquí mostrada, a un grupo mayor de variables.

² Las líneas `///` que están presentes en el comando estrictamente no son parte de éste. Estas significan que el comando continúa en la fila siguiente. Por lo general, dicha estrategia es de gran ayuda cuando el texto escrito en el comando es extenso.

Como se puede ver, las variables de ingreso ahora son etiquetadas como `ypccorh1996` `ypccorh2001` `ypccorh2006`. Dichas etiquetas contienen una estructura específica. La primera parte del nombre de la variable, es decir, “`ypccorh`”, en el caso de la base de datos con la que se está trabajando y que es especificada a través de la opción `newstub()`, refiere al contenido de la variable, mientras que la segunda parte refiere a la año en el cual la variable ha sido observada y es determinada a través de la opción `waves()`. Usar este tipo de convención hace más fácil formatear los datos desde `wide` a `long`. Obviamente, puedes hacer manualmente el cambio de etiqueta con el comando `rename`. Sin embargo, re-nombrar todas las variables de la base de datos en formato `wide` puede ser muy ineficiente en términos de tiempo, de ahí que es recomendable utilizar el comando `soepren`. Otra opción, que puede ser complementaria a la que representa este comando, es construir loops con los comandos `foreach` y `forvalues`. Sin embargo, por razones de espacio, éstas no serán descritas en este Tip.

Como ya se dijo, el comando en Stata para cambiar datos entre formatos `wide` y `long` es `reshape`. `reshape long` cambia una base de datos desde `wide` a `long`, y `reshape wide` hace lo mismo pero en la otra dirección. Stata necesita conocer tres piezas de información para “`reshape`” los datos: 1) La variable que identifica a las unidades de análisis en los datos (en el caso de EPCasen es `idpersona`); 2) las características que están bajo observación; y 3) los momentos en que dichas características fueron observadas.

La primera parte es fácil de obtener. Como ya se dijo, ésta corresponde a la variable `idpersona`, la cual identifica a cada persona en EPCasen. Las otras dos piezas de información están codificadas en la etiqueta de las variables. Como ya se dijo, la primera parte del nombre de la variable contiene la característica bajo observación, y la segunda parte contiene el momento de la observación. Por lo tanto, necesitamos decirle a Stata dónde la primera parte del nombre de la variable finaliza y dónde comienza la segunda parte. Esto se logra a través de un listado del nombre de la variable que corresponde a la característica bajo observación. En el caso del ejemplo de este Tip, tenemos lo siguiente:

```
. reshape long ypccorh, i(idpersona) j(wave)
(note: j = 1996 2001 2006)

Data                                wide  ->  long
-----
Number of obs.                       26882 ->  80646
Number of variables                    5    ->    4
j variable (3 values)                  ->  wave
xij variables:
    ypccorh1996 ypccorh2001 ypccorh2006 ->  ypccorh
-----
```

Primero, note la opción `i()`. Esta es usada para especificar la variable que identifica a los individuos en la base de datos. Segundo, observe que después de `reshape long` no se ha hecho un listado de las variables de ingreso, sino se ha especificado el nombre “`ypccorh`”, que refiere a la característica bajo observación, es decir, ingreso. La restante parte del nombre de la variable es interpretado automáticamente por Stata como información acerca del punto en el tiempo de la observación. Esta información es conservada en una nueva variable, la cual he llamado `wave`, nombre que es especificado en la opción `j()`. Si no se especifica esta opción, Stata usa el nombre de `_j` para la nueva variable.

Ahora podemos ver cómo lucen los mismos datos en formato long

```
. list idpersona wave ypccorh pesos_long_96_01_06 ///  
if idpersona==7 | idpersona ==12
```

```
+-----+  
| idpers~a   wave   ypccorh   pesos~06 |  
+-----+  
19. |         7   1996     51333   175.0152 |  
20. |         7   2001     59272   175.0152 |  
21. |         7   2006     52338   175.0152 |  
34. |        12   1996     47500   60.41674 |  
35. |        12   2001     71833   60.41674 |  
+-----+  
36. |        12   2006    143750   60.41674 |  
+-----+
```

La nueva base de datos tiene 4 variables en vez de 5. Por su puesto, aún hay 26.882 individuos en la base de datos, pero dado que tenemos observaciones hechas en varios momentos en el tiempo, tenemos $n*t$ observaciones, es decir 80.646. Note que la variable `pesos_long_96_01_06` no aparece en el comando `reshape` pero, de todos modos, es incluida por Stata en la base en formato long. La razón de tal exclusión es que, tal como se puede ver, esta variable es constante en el tiempo y pasa a formar parte de la nueva base de datos automáticamente.

Finalmente, es importante mencionar que después de utilizar el comando `reshape` se puede fácilmente retornar al formato wide y vice versa

```
. reshape wide  
. reshape long
```

Cabe hacer notar que `reshape` es un comando de una gran flexibilidad en lo que respecta a la modificación de la estructura de los datos. Ilustraciones de distintas facetas de dicho comando pueden encontrarse en Baum and Cox (2007) y Stata (2007b).

Luis Maldonado

lmaldona@smail.uni-koeln.de

Lehrstuhl für Empirische Social- und Wirtschaftsforschung
Universidad de Colonia
Alemania

Referencias

Baum, C.F. & Cox, N.J., 2007. Stata tip 45: Getting those data into shape. *The Stata Journal*, 7 (2), 268-271.

Kohler, U. & Kreuter, F., 2008. *Data analysis using stata* Texas.

Stata, 2007a. *Longitudinal/panel-data reference manual*.

Stata, 2007b. *Data management reference manual* Texas.



Anexo

***Stata 10.1

```
pwd  
cd c:\data\Tips  
dir
```

```
set mem 50000  
use panelcasen_m04, clear  
doedit Tip2  
keep idpersona pesos_long_96_01_06 ypccorh_96 ypccorh_01 ypccorh_06
```

```
list idpersona ypccorh_96 ypccorh_01 ypccorh_06 ///  
pesos_long_96_01_06 if idpersona==7 | idpersona ==12  
des
```

```
ssc install soepren  
help soepren
```

```
soepren ypccorh_96 ypccorh_01 ypccorh_06, newstub(ypccorh) waves(1996 2001 2006)
```

```
reshape long ypccorh, i(idpersona) j(wave)  
list idpersona wave ypccorh pesos_long_96_01_06 if idpersona==7 | idpersona ==12
```

```
reshape wide  
reshape long
```