An example of "reshape long" with Stata[1]

Stata has "reshape long" and "reshape wide" commands that make it pretty easy to modify files from wide to long, and back. This example will use a subset of NLSY data.

For this example, the variables for each case, are as follows:

```
caseid race0 sex
hhkid79 hhkid80 hhkid81 (.skip variables) hhkid00  hhkid02 hhkid04 hhkid06
age79   age80   age81   (.skip variables) age00    age02   age04   age06
```

hhkid06 is 1 if there are kids in the household 2006, 0 if there are none, and missing if unknown
hhkid79 is 1 if there are kids in the household 1979, 0 if there are none, and missing if unknown

We want to transform the structure of the data so that we have the data so that we will have a separate case for every person-year, rather than for every person. The file we want will be longer, but narrower.

caseid	year	race0	sex	hhkid	age
8	1979	3	2	1	20
8	1980	3	2	1	21
8	1981	3	2	1	22
	(... skip cases)				
8	2006	3	2	0	15
9	1979	12	2	0	16
9	1980	12	2	0	18
9	1981	12	2	0	19
	(... skip cases)				
9	2006	12	2	1	43

Note that in the new format, there is a record for each person for each year. In the original data file, there are N people and so there are N cases. However, in the person-year file, there (N)*(j) cases where j= number of years. There are some fixed variables that remain the same for the person over the years (for example, race and sex). There are others that may vary (for example, age in 1979, age in 1980, etc).

On the next few pages you will see the "log" file in which our commands to Stata, and Stata's responses were recorded. We have bolded our commands, and made comments bold and green. Stata's responses are plain text.

A list of the commands used in the program follows. By reading the comments and Stata's response, you can see what they do.

```
log using "H:\Stata_long\NLSY_Part_Long.log", replace
set more off
use H:\Stata_long\testforpjg2.dta, clear
keep caseid race0 sex age* hhkid*
codebook, c
renpfix age0 age200
save H:\Stata_long\NLSY_Part.dta, replace
reshape long age hhkid, i(caseid) j(year)
replace year=year+1900 if year<1900
tab year
* Record my commands and STATA responses in the following location.
```

[1]Prepared by Patty Glynn and Stephanie Ewert, University of Washington, 3/04/09

```
  log using "H:\Stata_long\NLSY_Part_Long.log", replace
-------------------------------------------------------------------------------
      log:  H:\Stata_long\NLSY_Part_Long.log
  log type:  text
 opened on:  25 Feb 2009, 14:00:55


. * NLSY_Part_Long.do  February 25, 2009
. * create person-year files
. * Ask stata to not pause for screen breaks, making the log more readable.
. set more off
. use H:\Stata_long\testforpjg2.dta, clear
. * Stata has a cool shortcut for listing variables.
. * You can give the “stump” of the variable name (such as “age”)
. * plus a wild card ‘*’ and all variables that begin with the stump
. * will be included.
. keep caseid race0 sex age* hhkid*
. Ask for a concise codebook.
. codebook, c

Variable   Obs Unique      Mean  Min  Max  Label
-------------------------------------------------------------------------------
caseid      50     50      25.5    1   50
race0       50      9      8.28    3   28
sex         50      2      1.54    1    2
age79       50      9     18.06   14   22
age80       44      9  18.97727   15   23
age81       47      9  20.06383   16   24
age82       48      9  21.04167   17   25
age83       46      9  21.97826   18   26
age84       45      9  22.95556   19   27
age85       45      9  23.93333   20   28
age86       42      9  24.97619   21   29
age87       42      8  25.97619   23   30
age88       44      8     27.25   24   31
age89       45      8  28.28889   25   32
age90       43      8  29.18605   26   33
age91       44      8     30.25   27   34
age92       47      8  31.25532   28   35
age93       47      8  32.14894   29   36
age94       44      8  33.31818   30   37
age96       40      9    35.075   31   39
age98       38      9  37.02632   33   41
age00       36      9  39.22222   35   43
age02       35      8      41.2   38   45
age04       35      9  43.31429   39   47
age06       37      9  45.24324   41   49
hhkid79     50      2       .02    0    1
hhkid80     44      2  .0227273    0    1
hhkid81     47      2  .0851064    0    1
hhkid82     48      2  .1041667    0    1
hhkid83     46      2  .2173913    0    1
hhkid84     45      2  .2444444    0    1
hhkid85     45      2  .2444444    0    1
hhkid86     42      2  .2380952    0    1
hhkid87     42      2  .2857143    0    1
hhkid88     44      2  .3409091    0    1
hhkid89     45      2        .4    0    1
hhkid90     43      2  .4883721    0    1
hhkid91     44      2  .5454545    0    1
hhkid92     47      2  .5319149    0    1
hhkid93     47      2  .5744681    0    1
hhkid94     44      2  .6136364    0    1
hhkid96     40      2        .6    0    1
hhkid98     38      2  .7105263    0    1
hhkid00     36      2  .6388889    0    1
```

```
hhkid02     35      2  .6571429    0    1
hhkid04     35      2  .6571429    0    1
hhkid06     37      2  .5675676    0    1
-------------------------------------------------------------------------------

. ** We need to change the variable names that begin with 0 so that they
. ** begin with 20.  Otherwise, problems are encountered in the “reshape”
. ** command later.  After these “renpfix” commands, hhkid00 will have the
. ** name of hhkid2000, hhkid02 will be hhkid2002, and so on.
. ** (This also uses the cool “stump” feature of Stata).
.
. renpfix age0 age200
.
. renpfix hhkid0 hhkid200
.
. codebook, c

Variable   Obs Unique      Mean  Min  Max  Label
-------------------------------------------------------------------------------
caseid      50     50      25.5    1   50
race0       50      9      8.28    3   28
sex         50      2      1.54    1    2
age79       50      9     18.06   14   22
age80       44      9  18.97727   15   23
age81       47      9  20.06383   16   24
age82       48      9  21.04167   17   25
age83       46      9  21.97826   18   26
age84       45      9  22.95556   19   27
age85       45      9  23.93333   20   28
age86       42      9  24.97619   21   29
age87       42      8  25.97619   23   30
age88       44      8     27.25   24   31
age89       45      8  28.28889   25   32
age90       43      8  29.18605   26   33
age91       44      8     30.25   27   34
age92       47      8  31.25532   28   35
age93       47      8  32.14894   29   36
age94       44      8  33.31818   30   37
age96       40      9    35.075   31   39
age98       38      9  37.02632   33   41
age2000     36      9  39.22222   35   43
age2002     35      8      41.2   38   45
age2004     35      9  43.31429   39   47
age2006     37      9  45.24324   41   49
hhkid79     50      2       .02    0    1
hhkid80     44      2  .0227273    0    1
hhkid81     47      2  .0851064    0    1
hhkid82     48      2  .1041667    0    1
hhkid83     46      2  .2173913    0    1
hhkid84     45      2  .2444444    0    1
hhkid85     45      2  .2444444    0    1
hhkid86     42      2  .2380952    0    1
hhkid87     42      2  .2857143    0    1
hhkid88     44      2  .3409091    0    1
hhkid89     45      2        .4    0    1
hhkid90     43      2  .4883721    0    1
hhkid91     44      2  .5454545    0    1
hhkid92     47      2  .5319149    0    1
hhkid93     47      2  .5744681    0    1
hhkid94     44      2  .6136364    0    1
hhkid96     40      2        .6    0    1
hhkid98     38      2  .7105263    0    1
hhkid2000   36      2  .6388889    0    1
hhkid2002   35      2  .6571429    0    1
hhkid2004   35      2  .6571429    0    1
hhkid2006   37      2  .5675676    0    1
-------------------------------------------------------------------------------
```

```
.
. save H:\Stata_long\NLSY_Part.dta, replace
file H:\Stata_long\NLSY_Part.dta saved
.
. use H:\Stata_long\NLSY_Part.dta, clear

. ** This is where we issue the “reshape long” command.
. ** With this command, each person year will become a case,
. ** rather than each person (as the file is organized now).
. ** We are asking that all of the variables that begin with “age”
. ** and “hhkid” be reshaped. Any variables not listed after
. ** “reshape long” will be included as a variable for each person year.
. ** For example, race0 will be included as a variable for the observation
. ** for person 1 in 1979, the observation for person 1 in 1980, etc.


. reshape long age hhkid, i(caseid) j(year)
(note: j = 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 96 98 2000 2002 2004
>  2006)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       50   ->    1100
Number of variables                  47   ->       6
j variable (22 values)                    ->   year
xij variables:
                age79 age80 ... age2006   ->   age
          hhkid79 hhkid80 ... hhkid2006   ->   hhkid
-----------------------------------------------------------------------------

. tab year

       year |      Freq.     Percent        Cum.
------------+-----------------------------------
         79 |         50        4.55        4.55
         80 |         50        4.55        9.09
         81 |         50        4.55       13.64
         82 |         50        4.55       18.18
         83 |         50        4.55       22.73
         84 |         50        4.55       27.27
         85 |         50        4.55       31.82
         86 |         50        4.55       36.36
         87 |         50        4.55       40.91
         88 |         50        4.55       45.45
         89 |         50        4.55       50.00
         90 |         50        4.55       54.55
         91 |         50        4.55       59.09
         92 |         50        4.55       63.64
         93 |         50        4.55       68.18
         94 |         50        4.55       72.73
         96 |         50        4.55       77.27
         98 |         50        4.55       81.82
       2000 |         50        4.55       86.36
       2002 |         50        4.55       90.91
       2004 |         50        4.55       95.45
       2006 |         50        4.55      100.00
------------+-----------------------------------
      Total |      1,100      100.00

. *** You see that the values of year range from 79 to 2006.
. *** I want the values to be 1979 to 2006.  The following command
. *** will fix this.
.
. replace year=year+1900 if year<1900
(900 real changes made)
```

```
.
. save H:\Stata_long\NLSY_Part_Long.dta, replace
file H:\Stata_long\NLSY_Part_Long.dta saved

. use H:\Stata_long\NLSY_Part_Long.dta, replace
. tab year

       year |      Freq.     Percent        Cum.
------------+-----------------------------------
       1979 |         50        4.55        4.55
       1980 |         50        4.55        9.09
       1981 |         50        4.55       13.64
       1982 |         50        4.55       18.18
       1983 |         50        4.55       22.73
       1984 |         50        4.55       27.27
       1985 |         50        4.55       31.82
       1986 |         50        4.55       36.36
       1987 |         50        4.55       40.91
       1988 |         50        4.55       45.45
       1989 |         50        4.55       50.00
       1990 |         50        4.55       54.55
       1991 |         50        4.55       59.09
       1992 |         50        4.55       63.64
       1993 |         50        4.55       68.18
       1994 |         50        4.55       72.73
       1996 |         50        4.55       77.27
       1998 |         50        4.55       81.82
       2000 |         50        4.55       86.36
       2002 |         50        4.55       90.91
       2004 |         50        4.55       95.45
       2006 |         50        4.55      100.00
------------+-----------------------------------
      Total |      1,100      100.00
.
. codebook, c

Variable    Obs Unique      Mean   Min   Max  Label
------------------------------------------------------------------------------
caseid     1100     50      25.5     1    50
year       1100     22  1990.455  1979  2006
race0      1100      9      8.28     3    28
sex        1100      2      1.54     1     2
age         944     36  28.93008    14    49
hhkid       944      2  .3845339     0     1
------------------------------------------------------------------------------
. log close
       log:  H:\Stata_long\NLSY_Part_Long.log
  log type:  text
 closed on:  25 Feb 2009, 14:00:57
```