The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata and Mata are registered trademarks of StataCorp LP.

The Stata Journal (2007)
7, Number 2, pp. 268–271

Stata tip 45: Getting those data into shape

Christopher F. Baum
Department of Economics
Boston College
Chestnut Hill, MA 02467
baum@bc.edu

Nicholas J. Cox
Department of Geography
Durham University
Durham City, UK
n.j.cox@durham.ac.uk

Are your data in shape? That is, are they in the structure that you need to conduct the analysis you have in mind? Data sources often provide the data in a structure that is suitable for presentation but clumsy for statistical analysis. One of the key data management tools that Stata provides is `reshape`; see [D] **reshape**. If you need to modify the structure of your data, you should be familiar with `reshape` and its two functions: `reshape wide` and `reshape long`. In this tip, we discuss how two applications of `reshape` may be the solution to some knotty data management problems.

As a first example, consider this question posted on Statalist by an individual who has a dataset in the wide form:

country	tradeflow	Yr1990	Yr1991
Armenia	imports	105	120
Armenia	exports	90	100
Bolivia	imports	200	230
Bolivia	exports	80	115
Colombia	imports	100	105
Colombia	exports	70	71

He would like to reshape the data into long form:

country	year	imports	exports
Armenia	1990	105	90
Armenia	1991	120	100
Bolivia	1990	200	80
Bolivia	1991	230	115
Colombia	1990	100	70
Colombia	1991	105	71

We must exchange the roles of years and tradeflows in the original data to arrive at the desired structure, suitable for analysis as `xt` data. This exchange can be handled by two successive applications of `reshape`:

```
. reshape long Yr, i(country tradeflow)
(note: j = 1990 1991)
Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        6   ->      12
Number of variables                   4   ->       4
j variable (2 values)                     ->   _j
xij variables:
                           Yr1990 Yr1991   ->   Yr
-----------------------------------------------------------------------------
```

This transformation swings the data into long form with each observation identified by `country`, `tradeflow`, and the new variable `_j`, taking on the values of year. We now perform `reshape wide` to make imports and exports into separate variables:

```
. rename _j year
. reshape wide Yr, i(country year) j(tradeflow) string
(note: j = exports imports)
Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       12   ->       6
Number of variables                   4   ->       4
j variable (2 values)         tradeflow   ->   (dropped)
xij variables:
                                      Yr   ->   Yrexports Yrimports
-----------------------------------------------------------------------------
```

If we transform the data to wide form once again, the `i()` option contains `country` and `year`, as those are the desired identifiers on each observation of the target dataset. We specify that `tradeflow` is the `j()` variable for `reshape`, indicating that it is a `string` variable. The data now have the desired structure. Although we have illustrated this double-reshape transformation with only a few countries, years, and variables, the technique generalizes to any number of each.

As a second example of successive applications of `reshape`, consider the World Bank's World Development Indicators (WDI) dataset.[1] Their extract program generates a comma-separated value (CSV) database extract, readable by Excel or Stata, but the structure of those data hinders analysis as panel data. For a recent year, the header line of the CSV file is

```
"Series code","Country Code","Country Name","1960","1961","1962","1963",
"1964","1965","1966","1967","1968","1969","1970","1971","1972","1973",
"1974","1975","1976","1977","1978","1979","1980","1981","1982","1983",
"1984","1985","1986","1987","1988","1989","1990","1991","1992","1993",
"1994","1995","1996","1997","1998","1999","2000","2001","2002","2003","2004"
```

1. See http://econ.worldbank.org.

That is, each row of the CSV file contains a *variable* and *country* combination, with the columns representing the elements of the time series.[2]

Our target dataset structure is that appropriate for panel-data modeling, with the variables as columns and rows labeled by country and year. Two applications of `reshape` will again be needed to reach the target format. We first `insheet` (see [D] **insheet**) the data and transform the triliteral country code into a numeric code with the country codes as labels:

```
. insheet using wdiex.raw, comma names
. encode countrycode, generate(cc)
. drop countrycode
```

We then must address that the time-series variables are named `var4-var48`, as the header line provided invalid Stata variable names (numeric values) for those columns. We use `rename` (see [D] **rename**) to change `v4` to `d1960`, `v5` to `d1961`, and so on:

```
forv i=4/48 {
        rename v`i' d`=1956+`i''
}
```

We now are ready to carry out the first `reshape`. We want to identify the rows of the reshaped dataset by both country code (`cc`) and `seriescode`, the variable name. The `reshape long` will transform a fragment of the WDI dataset containing two series and four countries:

```
. reshape long d, i(cc seriescode) j(year)
(note: j = 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972
> 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987
> 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002
> 2003 2004)
Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        7   ->     315
Number of variables                  48   ->       5
j variable (45 values)                    ->   year
xij variables:
                  d1960 d1961 ... d2004   ->   d
-----------------------------------------------------------------------------
```

2. A variation occasionally encountered will resemble this structure, but with periods in reverse chronological order. The solution here can be used to deal with that problem as well.

```
. list in 1/15
```

	cc	seriesc~e	year	countryname	d
1.	AFG	adjnetsav	1960	Afghanistan	.
2.	AFG	adjnetsav	1961	Afghanistan	.
3.	AFG	adjnetsav	1962	Afghanistan	.
4.	AFG	adjnetsav	1963	Afghanistan	.
5.	AFG	adjnetsav	1964	Afghanistan	.
6.	AFG	adjnetsav	1965	Afghanistan	.
7.	AFG	adjnetsav	1966	Afghanistan	.
8.	AFG	adjnetsav	1967	Afghanistan	.
9.	AFG	adjnetsav	1968	Afghanistan	.
10.	AFG	adjnetsav	1969	Afghanistan	.
11.	AFG	adjnetsav	1970	Afghanistan	-2.97129
12.	AFG	adjnetsav	1971	Afghanistan	-5.54518
13.	AFG	adjnetsav	1972	Afghanistan	-2.40726
14.	AFG	adjnetsav	1973	Afghanistan	-.188281
15.	AFG	adjnetsav	1974	Afghanistan	1.39753

The rows of the data are now labeled by year, but one problem remains: all variables for a given country are stacked vertically. To unstack the variables and put them in shape for `xtreg` (see [XT] **xtreg**), we must carry out a second `reshape` that spreads the variables across the columns, specifying `cc` and `year` as the i variables and `seriescode` as the j variable. Since that variable has string content, we use the `string` option.

```
. reshape wide d, i(cc year) j(seriescode) string
(note: j = adjnetsav adjsavC02)
Data                                long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       315   ->     180
Number of variables                    5   ->       5
j variable (2 values)        seriescode   ->   (dropped)
xij variables:
                                       d   ->   dadjnetsav dadjsavC02
-----------------------------------------------------------------------------
. order cc countryname
. tsset cc year
       panel variable:  cc (strongly balanced)
        time variable:  year, 1960 to 2004
```

After this transformation, the data are now in shape for `xt` modeling, tabulation, or graphics.

As illustrated here, the `reshape` command can transform even the most inconvenient data structure into the structure needed for your research. It may take more than one application of `reshape` to get there from here, but it can do the job.