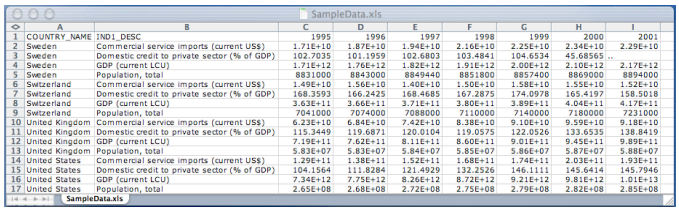
Reshaping Panel Data Using Excel and Stata

Moonhawk Kim
Department of Political Science
Stanford University

June 27, 2003

Figure 1: Downloaded Panel Data

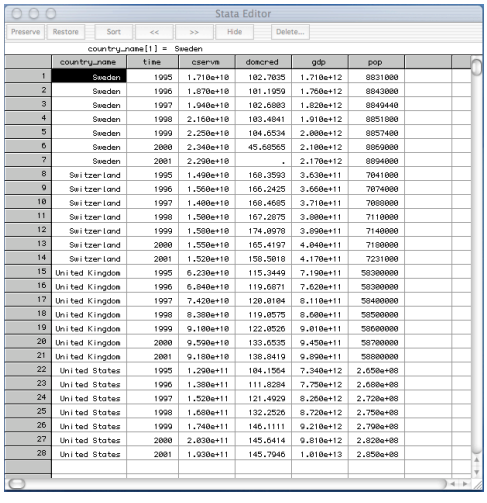


SampleData.xls

	A	B	C	D	E	F	G	H	I
1	COUNTRY_NAME	IND1_DESC	1995	1996	1997	1998	1999	2000	2001
2	Sweden	Commercial service imports (current US$)	1.71E+10	1.87E+10	1.94E+10	2.16E+10	2.25E+10	2.34E+10	2.29E+10
3	Sweden	Domestic credit to private sector (% of GDP)	102.7035	101.1959	102.6803	103.4841	104.6534	45.68565	..
4	Sweden	GDP (current LCU)	1.71E+12	1.76E+12	1.82E+12	1.91E+12	2.00E+12	2.10E+12	2.17E+12
5	Sweden	Population, total	8831000	8843000	8849440	8851800	8857400	8869000	8894000
6	Switzerland	Commercial service imports (current US$)	1.49E+10	1.56E+10	1.40E+10	1.50E+10	1.58E+10	1.55E+10	1.52E+10
7	Switzerland	Domestic credit to private sector (% of GDP)	168.3593	166.2425	168.4685	167.2875	174.0978	165.4197	158.5018
8	Switzerland	GDP (current LCU)	3.63E+11	3.66E+11	3.71E+11	3.80E+11	3.89E+11	4.04E+11	4.17E+11
9	Switzerland	Population, total	7041000	7074000	7088000	7110000	7140000	7180000	7231000
10	United Kingdom	Commercial service imports (current US$)	6.23E+10	6.84E+10	7.42E+10	8.38E+10	9.10E+10	9.59E+10	9.18E+10
11	United Kingdom	Domestic credit to private sector (% of GDP)	115.3449	119.6871	120.0104	119.0575	122.0526	133.6535	138.8419
12	United Kingdom	GDP (current LCU)	7.19E+11	7.62E+11	8.11E+11	8.60E+11	9.01E+11	9.45E+11	9.89E+11
13	United Kingdom	Population, total	5.83E+07	5.83E+07	5.84E+07	5.85E+07	5.86E+07	5.87E+07	5.88E+07
14	United States	Commercial service imports (current US$)	1.29E+11	1.38E+11	1.52E+11	1.68E+11	1.74E+11	2.03E+11	1.93E+11
15	United States	Domestic credit to private sector (% of GDP)	104.1564	111.8284	121.4929	132.2526	146.1111	145.6414	145.7946
16	United States	GDP (current LCU)	7.34E+12	7.75E+12	8.26E+12	8.72E+12	9.21E+12	9.81E+12	1.01E+13
17	United States	Population, total	2.65E+08	2.68E+08	2.72E+08	2.75E+08	2.79E+08	2.82E+08	2.85E+08

SampleData.xls

Figure 2: Reorganized Panel Data



Stata Editor

Preserve | Restore | Sort | << | >> | Hide | Delete...

country_name[1] = Sweden

	country_name	time	cservm	domcred	gdp	pop
1	Sweden	1995	1.710e+10	102.7035	1.710e+12	8831000
2	Sweden	1996	1.870e+10	101.1959	1.760e+12	8843000
3	Sweden	1997	1.940e+10	102.6803	1.820e+12	8849440
4	Sweden	1998	2.160e+10	103.4841	1.910e+12	8851800
5	Sweden	1999	2.250e+10	104.6534	2.000e+12	8857400
6	Sweden	2000	2.340e+10	45.68565	2.100e+12	8869000
7	Sweden	2001	2.290e+10	.	2.170e+12	8894000
8	Switzerland	1995	1.490e+10	168.3593	3.630e+11	7041000
9	Switzerland	1996	1.560e+10	166.2425	3.660e+11	7074000
10	Switzerland	1997	1.400e+10	168.4685	3.710e+11	7088000
11	Switzerland	1998	1.500e+10	167.2875	3.800e+11	7110000
12	Switzerland	1999	1.580e+10	174.0978	3.890e+11	7140000
13	Switzerland	2000	1.550e+10	165.4197	4.040e+11	7180000
14	Switzerland	2001	1.520e+10	158.5018	4.170e+11	7231000
15	United Kingdom	1995	6.230e+10	115.3449	7.190e+11	58300000
16	United Kingdom	1996	6.840e+10	119.6871	7.620e+11	58300000
17	United Kingdom	1997	7.420e+10	120.0104	8.110e+11	58400000
18	United Kingdom	1998	8.380e+10	119.0575	8.600e+11	58500000
19	United Kingdom	1999	9.100e+10	122.0526	9.010e+11	58600000
20	United Kingdom	2000	9.590e+10	133.6535	9.450e+11	58700000
21	United Kingdom	2001	9.180e+10	138.8419	9.890e+11	58800000
22	United States	1995	1.290e+11	104.1564	7.340e+12	2.65e+08
23	United States	1996	1.380e+11	111.8284	7.750e+12	2.68e+08
24	United States	1997	1.520e+11	121.4929	8.260e+12	2.72e+08
25	United States	1998	1.680e+11	132.2526	8.720e+12	2.75e+08
26	United States	1999	1.740e+11	146.1111	9.210e+12	2.79e+08
27	United States	2000	2.030e+11	145.6414	9.810e+12	2.82e+08
28	United States	2001	1.930e+11	145.7946	1.010e+13	2.85e+08

Many of us frequently find ourselves in situations of downloading panel data from having to "reshape" data from Figure 1 to Figure 2. That is, many external databases (e.g. World

Bank's World Development Indicators) download panel data in a format, in which units and data series go down the rows and time periods go across columns. This is not a helpful format for either data analysis or for importing into your own database table. Accordingly, you need to convert the format from Figure 1 to Figure 2. Before I learned this trick, I used to copy and concatenate relevant columns, which took me hours!

This memo will walk you through an example of converting the format from "wide" to "long" and then back to "wide." Each worksheet in the accompanying Excel file ("Sample-Data.xls") matches up with each of the steps below.

We are starting with the worksheet "Initial Download". This is the same as Figure 1.

1. Stata requires the variables over which we perform the `reshape` command to be numbers rather than string. Thus, we start by assigning each unit—here countries—a unique ID number:

   (a) Create column for IDs ("unit_id"). Insert "1" to the first observation of Sweden. Then type and fill-down the following formula to assign each unit the same ID number "$=$ `if(b3 = b2, a2, a2 + 1)`". (This translates into, "set the ID to the same as the ID in the observation above, if the country name is the same as the country name in the observation above. If not, increase the new ID number by one."

   (b) Insert a new column. Copy the column with values. Click on the newly-created column. Go to menu "Edit" and to "Paste Special...". Click on the "Values" option under "Paste" and click OK. This forces the ID numbers to become as if you had manually entered the numbers and gets rid of the formulas.

   (c) Delete the original "unit_id" column with formula-based cells.

2. Now we need to do the same for each data series:

   (a) Sort by data series ("ind1_desc")

   (b) Create column for IDs ("series_id"). Insert "1" to the first observation of Commercial service imports. Then type and fill-down the following formula to assign each unit the same ID number "$=$ `if(b3 = b2, a2, a2 + 1)`".

   (c) Insert a new column. Copy the column with values. Click on the newly-created column. Go to menu "Edit" and to "Paste Special...". Click on the "Values" option under "Paste" and click OK. This forces the ID numbers to become as if you had manually entered the numbers and gets rid of the formulas.

   (d) Delete the original "series_id" column with formula-based cells.

3. Stata does not accept numbers as variable names. So, we need to change the years into something else. (The string we add will also have another function later on.)

   (a) Select all the years across the first row, from 1995 to 2001. Find and replace "19" with "data19".

(b) Obviously this doesn't work for years 2000 and 2001. Change these manually to "data2000" and "data2001"

4. Lastly, Stata, World Development Indicators and Excel use different symbols for null values. WDI uses double periods; Excel uses empty cells and Stata uses single periods. There is a null value for Sweden's domestic credit to private sector in 2001. Find and replace null value indicators (e.g. from ".." to "").

5. Copy and paste data into the data editor in Stata. Close the data editor.

6. Issue the following command:

```
reshape long data, i(unit_id series_id) j(time)
```

I won't spell out the specifics of the `reshape` command. You can refer to the help file in Stata. The command should give the following output.

```
(note: j = 1995 1996 1997 1998 1999 2000 2001)

Data                                wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        16   ->     112
Number of variables                   11   ->       6
j variable (7 values)                      ->   time
xij variables:
         data1995 data1996 ... data2001   ->   data
-----------------------------------------------------------------------------
```

Issue `edit` to see what Stata did.

7. Now, delete the variable that contains the series labels ("ind1_desc") but keep the series ID number variable ("series_id"). We have to drop it, because it interferes with uniquely identifying each observation.

8. We do not want all the different data series to go down a single column. Although we still want unit-time (e.g. country-years) to go down the rows, we want the different data-series to go across columns. We accomplish that by issuing the following:

```
reshape wide data, i(unit_id time) j(series_id)

(note: j = 1 2 3 4)

Data                                long   ->   wide
-----------------------------------------------------------------------------
```

```
Number of obs.                      112   ->      28
Number of variables                   5   ->       7
j variable (4 values)         series_id   ->   (dropped)
xij variables:
                                   data   ->   data1 data2 ... data4
-----------------------------------------------------------------------------
```

Issue `edit` to see what Stata did.

9. And we clean up a little:

```
move country_name unit_id
drop unit_id
```

10. Recover series name and assign new variable names. Refer back to the Excel file to see which series ID number matches up with which series. Assign a new short variable name.