Merge/Append using Stata

(draft)

Oscar Torres-Reyna

Data Consultant

otorres@princeton.edu

Intro

Merge – adds variables to a dataset. Type `help merge` for details.

Merging two datasets require that both have *at least* one variable in common (either string or numeric). If string make sure the categories have the same spelling (i.e. country names, etc.).

The common variables *must* have the same name.

Explore each dataset separately before merging. Make sure to use all possible common variables (for example, if merging two panel datasets you will need country and years).

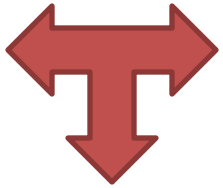
Append – adds cases/observations to a dataset. Type `help append` for details.

Appending two datasets require that both have variables with *exactly* the same name. If using categorical data make sure the categories on both datasets refer to *exactly* the same thing (i.e. 1 “Agree”, 2”Disagree”, 3 “DK” on both). Remember that Stata is case sensitive, 'Year' is not the same as 'year'. For variables that do not match, Stata will add missing values.

MERGE – EXAMPLE 1 (type `help merge` for more details)

mydata1

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	A	2002	-11	0	.36	-.79	.7
4	A	2003	2646	1	.25	-.89	-.09
5	B	2000	-5935	0	-.08	1.43	.02
6	B	2001	-712	0	.11	1.65	.26
7	B	2002	-1933	0	.35	1.59	-.23
8	B	2003	3073	1	.73	1.69	.26
9	C	2000	-1292	0	1.31	-1.29	.2
10	C	2001	-3416	0	1.18	-1.34	.28
11	C	2002	-356	0	1.26	-1.26	.37
12	C	2003	1225	1	1.42	-1.31	-.38



mydata2

	country	year	x4	x5	x6	order
1	A	2000	10	1	9	1
2	A	2001	7	1	9	2
3	A	2002	7	9	4	3
4	A	2003	1	2	3	4
5	B	2000	0	5	6	5
6	B	2001	5	8	5	6
7	B	2002	9	4	5	7
8	B	2003	1	5	1	8
9	C	2000	4	5	4	9
10	C	2001	6	9	6	10
11	C	2002	6	5	3	11
12	C	2003	7	3	3	12

```
merge 1:1 country year using mydata2

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                                12  (_merge==3)
    -----------------------------------------
```

- Make sure one dataset is loaded into Stata (in this case mydata1), then use merge.
- Make sure to map where the using data is located (in this case mydata2, for example “c:\folders\data\mydata2.dta”)*.

NOTE: For Stata 10 or older:

1) Remove the 1:1
2) Sort both datasets by all the ids and save before merging

	country	year	y	y_bin	x1	x2	x3	x4	x5	x6	order	_merge
1	A	2000	1343	1	.28	-1.11	.28	10	1	9	1	matched (3)
2	A	2001	-1900	0	.32	-.95	.49	7	1	9	2	matched (3)
3	A	2002	-11	0	.36	-.79	.7	7	9	4	3	matched (3)
4	A	2003	2646	1	.25	-.89	-.09	1	2	3	4	matched (3)
5	B	2000	-5935	0	-.08	1.43	.02	0	5	6	5	matched (3)
6	B	2001	-712	0	.11	1.65	.26	5	8	5	6	matched (3)
7	B	2002	-1933	0	.35	1.59	-.23	9	4	5	7	matched (3)
8	B	2003	3073	1	.73	1.69	.26	1	5	1	8	matched (3)
9	C	2000	-1292	0	1.31	-1.29	.2	4	5	4	9	matched (3)
10	C	2001	-3416	0	1.18	-1.34	.28	6	9	6	10	matched (3)
11	C	2002	-356	0	1.26	-1.26	.37	6	5	3	11	matched (3)
12	C	2003	1225	1	1.42	-1.31	-.38	7	3	3	12	matched (3)

*To set the working directory see here
http://dss.princeton.edu/training/StataTutorial.pdf#page=6

MERGE – EXAMPLE 2 (one dataset missing a country)

mydata1

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	A	2002	-11	0	.36	-.79	.7
4	A	2003	2646	1	.25	-.89	-.09
5	B	2000	-5935	0	-.08	1.43	.02
6	B	2001	-712	0	.11	1.65	.26
7	B	2002	-1933	0	.35	1.59	-.23
8	B	2003	3073	1	.73	1.69	.26
9	C	2000	-1292	0	1.31	-1.29	.2
10	C	2001	-3416	0	1.18	-1.34	.28
11	C	2002	-356	0	1.26	-1.26	.37
12	C	2003	1225	1	1.42	-1.31	-.38

mydata3

	country	year	x4	x5	x6	order
1	A	2000	10	1	9	1
2	A	2001	7	1	9	2
3	A	2002	7	9	4	3
4	A	2003	1	2	3	4
5	B	2000	0	5	6	5
6	B	2001	5	8	5	6
7	B	2002	9	4	5	7
8	B	2003	1	5	1	8

```
merge 1:1 country year using mydata3
```

```
    Result                           # of obs.
    -----------------------------------------
    not matched                             4
        from master                         4  (_merge==1)
        from using                          0  (_merge==2)

    matched                                 8  (_merge==3)
    -----------------------------------------
```

- Make sure one dataset is loaded into Stata (in this case mydata1), then use merge.
- Make sure to map where the using data is located (in this case mydata2, for example “c:\folders\data\mydata3.dta”)*.
- Unmatched data is set to missing. If you want to keep only matched data, you can type

```
keep if _merge==3
```

NOTE: For Stata 10 or older:
1) Remove the 1:1
2) Sort both datasets by all the ids and save before merging

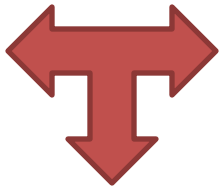
	country	year	y	y_bin	x1	x2	x3	x4	x5	x6	order	_merge
1	A	2000	1343	1	.28	-1.11	.28	10	1	9	1	matched (3)
2	A	2001	-1900	0	.32	-.95	.49	7	1	9	2	matched (3)
3	A	2002	-11	0	.36	-.79	.7	7	9	4	3	matched (3)
4	A	2003	2646	1	.25	-.89	-.09	1	2	3	4	matched (3)
5	B	2000	-5935	0	-.08	1.43	.02	0	5	6	5	matched (3)
6	B	2001	-712	0	.11	1.65	.26	5	8	5	6	matched (3)
7	B	2002	-1933	0	.35	1.59	-.23	9	4	5	7	matched (3)
8	B	2003	3073	1	.73	1.69	.26	1	5	1	8	matched (3)
9	C	2000	-1292	0	1.31	-1.29	.2	.	.	.	.	master only (1)
10	C	2001	-3416	0	1.18	-1.34	.28	.	.	.	.	master only (1)
11	C	2002	-356	0	1.26	-1.26	.37	.	.	.	.	master only (1)
12	C	2003	1225	1	1.42	-1.31	-.38	.	.	.	.	master only (1)

*To set the working directory see here
http://dss.princeton.edu/training/StataTutorial.pdf#page=6

MERGE – EXAMPLE 3 (many to one)

mydata1

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	A	2002	-11	0	.36	-.79	.7
4	A	2003	2646	1	.25	-.89	-.09
5	B	2000	-5935	0	-.08	1.43	.02
6	B	2001	-712	0	.11	1.65	.26
7	B	2002	-1933	0	.35	1.59	-.23
8	B	2003	3073	1	.73	1.69	.26
9	C	2000	-1292	0	1.31	-1.29	.2
10	C	2001	-3416	0	1.18	-1.34	.28
11	C	2002	-356	0	1.26	-1.26	.37
12	C	2003	1225	1	1.42	-1.31	-.38



mydata4

	country	x7
1	A	100
2	B	200
3	C	300

```
merge m:1 country using mydata4

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                                12  (_merge==3)
    -----------------------------------------
```

- Make sure one dataset is loaded into Stata (in this case mydata1), then use merge.
- Make sure to map where the using data is located (in this case mydata2, for example “c:\folders\data\mydata4.dta”)*.

NOTE: For Stata 10 or older:
1) Remove the m:1
2) Sort both datasets by all the ids and save before merging

	country	year	y	y_bin	x1	x2	x3	x7	_merge
1	A	2000	1343	1	.28	-1.11	.28	100	matched (3)
2	A	2001	-1900	0	.32	-.95	.49	100	matched (3)
3	A	2002	-11	0	.36	-.79	.7	100	matched (3)
4	A	2003	2646	1	.25	-.89	-.09	100	matched (3)
5	B	2000	-5935	0	-.08	1.43	.02	200	matched (3)
6	B	2001	-712	0	.11	1.65	.26	200	matched (3)
7	B	2002	-1933	0	.35	1.59	-.23	200	matched (3)
8	B	2003	3073	1	.73	1.69	.26	200	matched (3)
9	C	2000	-1292	0	1.31	-1.29	.2	300	matched (3)
10	C	2001	-3416	0	1.18	-1.34	.28	300	matched (3)
11	C	2002	-356	0	1.26	-1.26	.37	300	matched (3)
12	C	2003	1225	1	1.42	-1.31	-.38	300	matched (3)

*To set the working directory see here
http://dss.princeton.edu/training/StataTutorial.pdf#page=6

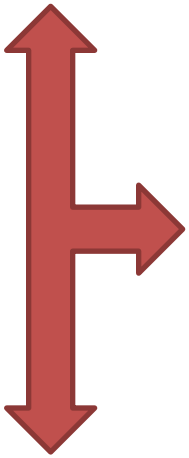
APPEND

APPEND– EXAMPLE 1

mydata7

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	B	2000	-5935	0	-.08	1.43	.02
4	B	2001	-712	0	.11	1.65	.26
5	C	2000	-1292	0	1.31	-1.29	.2
6	C	2001	-3416	0	1.18	-1.34	.28

- Make sure one dataset is loaded into Stata (in this case mydata7), then use append.
- Make sure to map where the using data is located (in this case mydata2, for example “c:\folders\data\mydata8.dta”)*.



append using mydata8

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	B	2000	-5935	0	-.08	1.43	.02
4	B	2001	-712	0	.11	1.65	.26
5	C	2000	-1292	0	1.31	-1.29	.2
6	C	2001	-3416	0	1.18	-1.34	.28
7	A	2002	-11	0	.36	-.79	.7
8	A	2003	2646	1	.25	-.89	-.09
9	B	2002	-1933	0	.35	1.59	-.23
10	B	2003	3073	1	.73	1.69	.26
11	C	2002	-356	0	1.26	-1.26	.37
12	C	2003	1225	1	1.42	-1.31	-.38

mydata8

	country	year	y	y_bin	x1	x2	x3
1	A	2002	-11	0	.36	-.79	.7
2	A	2003	2646	1	.25	-.89	-.09
3	B	2002	-1933	0	.35	1.59	-.23
4	B	2003	3073	1	.73	1.69	.26
5	C	2002	-356	0	1.26	-1.26	.37
6	C	2003	1225	1	1.42	-1.31	-.38

*To set the working directory see here
http://dss.princeton.edu/training/StataTutorial.pdf#page=6

APPEND – EXAMPLE 1 (cont.) – sorting by country/year

```
sort country year
```

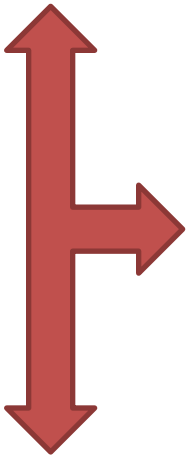
	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	A	2002	-11	0	.36	-.79	.7
4	A	2003	2646	1	.25	-.89	-.09
5	B	2000	-5935	0	-.08	1.43	.02
6	B	2001	-712	0	.11	1.65	.26
7	B	2002	-1933	0	.35	1.59	-.23
8	B	2003	3073	1	.73	1.69	.26
9	C	2000	-1292	0	1.31	-1.29	.2
10	C	2001	-3416	0	1.18	-1.34	.28
11	C	2002	-356	0	1.26	-1.26	.37
12	C	2003	1225	1	1.42	-1.31	-.38

APPEND– EXAMPLE 2 (one dataset missing one variable)

mydata7

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	B	2000	-5935	0	-.08	1.43	.02
4	B	2001	-712	0	.11	1.65	.26
5	C	2000	-1292	0	1.31	-1.29	.2
6	C	2001	-3416	0	1.18	-1.34	.28

- Make sure one dataset is loaded into Stata (in this case mydata7), then use append.
- Make sure to map where the using data is located (in this case mydata2, for example “c:\folders\data\mydata9.dta”)*.
- Notice the missing data.



append using mydata9

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	B	2000	-5935	0	-.08	1.43	.02
4	B	2001	-712	0	.11	1.65	.26
5	C	2000	-1292	0	1.31	-1.29	.2
6	C	2001	-3416	0	1.18	-1.34	.28
7	A	2002	-11	0	.36	-.79	.
8	A	2003	2646	1	.25	-.89	.
9	B	2002	-1933	0	.35	1.59	.
10	B	2003	3073	1	.73	1.69	.
11	C	2002	-356	0	1.26	-1.26	.
12	C	2003	1225	1	1.42	-1.31	.

mydata9

	country	year	y	y_bin	x1	x2
1	A	2002	-11	0	.36	-.79
2	A	2003	2646	1	.25	-.89
3	B	2002	-1933	0	.35	1.59
4	B	2003	3073	1	.73	1.69
5	C	2002	-356	0	1.26	-1.26
6	C	2003	1225	1	1.42	-1.31

*To set the working directory see here
http://dss.princeton.edu/training/StataTutorial.pdf#page=6

Useful links / Recommended books

- DSS Online Training Section http://dss.princeton.edu/training/
- UCLA Resources to learn and use STATA http://www.ats.ucla.edu/stat/stata/
- DSS help-sheets for STATA http://dss/online_help/stats_packages/stata/stata.htm
- *Introduction to Stata* (PDF), Christopher F. Baum, Boston College, USA. "A 67-page description of Stata, its key features and benefits, and other useful information." http://fmwww.bc.edu/GStat/docs/StataIntro.pdf
- STATA FAQ website http://stata.com/support/faqs/
- Princeton DSS Libguides http://libguides.princeton.edu/dss

Books

- *Introduction to econometrics* / James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.
- *Data analysis using regression and multilevel/hierarchical models* / Andrew Gelman, Jennifer Hill. Cambridge ; New York : Cambridge University Press, 2007.
- *Econometric analysis* / William H. Greene. 6th ed., Upper Saddle River, N.J. : Prentice Hall, 2008.
- *Designing Social Inquiry: Scientific Inference in Qualitative Research* / *Gary King, Robert O. Keohane, Sidney Verba, Princeton University Press*, 1994.
- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* / Gary King, Cambridge University Press, 1989
- *Statistical Analysis: an interdisciplinary introduction to univariate & multivariate methods* / Sam Kachigan, New York : Radius Press, c1986
- *Statistics with Stata (updated for version 9)* / Lawrence Hamilton, Thomson Books/Cole, 2006