Merge and Append Using Stata: How to Merge and Append Datasets

Search this Guide Search

This guide discusses basic techniques to merge and append datasets using Stata.

How to Merge and Append Datasets

Table of Contents

Merge Append References

IBM v. Digital Doors,
IPR2023-00972, 7,958,268
DigitalDoors EX2033

Ask Us

Merge

- We use merge to add variables to a dataset.

- Merging two datasets requires that both have at least one variable in common (either string or numeric), which is also called the identifier variable.

- The common variables or the identifier variables must have the same name.

- If you merge string variables, make sure the categories have the same spelling (e.g., the spelling of the name of each country should be the same).

- Inspect each dataset carefully before merging. While merging two panel datasets, for example, look for two common variables: *entity id* (e.g., country, state) and *time* (e.g., year, month).

- This guide discusses different data merging techniques available in Stata. Note down some useful jargons I use throughout this guide:

master dataset: starting dataset/dataset in memory

using dataset: dataset you want to merge

matched dataset: dataset generated after the merge

master + using = matched

1. One-to-One Merge

We use one-to-one merge when the common variable(s) uniquely identifies each observation in each of the datasets. Let us merge the following two datasets, named *mydata1* (master dataset) and *mydata2* (using dataset).

mydata1

	country	year	y	y_bin	x1	x2	x3
1	A	2011	1343	1	.2779036	-1.107956	.2825536
2	A	2012	-1900	0	.3206847	-.94872	.4925385
3	A	2013	-1123	0	.3634657	-.789484	.7025234
4	A	2014	2646	1	.246144	-.885533	-.0943909
5	B	2011	-5935	0	-.08185	1.42512	.0234281
6	B	2012	-712	0	.10616	1.649602	.2603625
7	B	2013	-1933	0	.3537852	1.593719	-.2343988
8	B	2014	3073	1	.726777	1.691758	.2562243
9	C	2011	-1292	0	1.312561	-1.293136	.204083
10	C	2012	-3416	0	1.177484	-1.344218	.2839719
11	C	2013	-3558	0	1.256408	-1.259951	.3733927
12	C	2014	1225	1	1.421545	-1.311745	-.3759656

mydata2

	country	year	x4	x5	x6
1	A	2011	10	1	9

Data Consultant

Muhammad Al Amin
He/Him/His

Email Me

Contact:
Firestone Library, A-12-F.1
609-258-6051

Data Consultant

Yufei Qin

Email Me

Contact:
Firestone Library, A.12F.2
6092582519

IBM v. Digital Doors,
IPR2023-00972, 7,958,268
DigitalDoors EX2033

2	A	2012	7	1	9
3	A	2013	7	9	4
4	A	2014	1	2	3
5	B	2011	0	5	6
6	B	2012	5	8	5
7	B	2013	9	4	5
8	B	2014	1	5	1
9	C	2011	4	5	4
10	C	2012	6	9	6
11	C	2013	6	5	3
12	C	2014	7	3	3

Notice that the above datasets have two common variables: **country** and **year**. For merging these two datasets, we need to use one-to-one method because the combination of the values of **country** and **year** uniquely identifies observations in both datasets.

Use the following Stata commands for 1:1 merge.

```
use https://dss.princeton.edu/training/mydata1.dta

merge 1:1 country
year using https://dss.princeton.edu/training/mydata2.dta
```

```
    Result                      Number of obs
    -----------------------------------------
    Not matched                             0
    Matched                                12   (_merge==3)
    -----------------------------------------
```

The resulting merged/matched dataset is as follows.

	country	year	y	y_bin	x1	x2	x3	x4	x5	x6	_merge
1	A	2011	1343	1	.2779036	-1.107956	.2825536	10	1	9	Matched (3)
2	A	2012	-1900	0	.3206847	-.94872	.4925385	7	1	9	Matched (3)
3	A	2013	-1123	0	.3634657	-.789484	.7025234	7	9	4	Matched (3)
4	A	2014	2646	1	.246144	-.885533	-.0943909	1	2	3	Matched (3)
5	B	2011	-5935	0	-.08185	1.42512	.0234281	0	5	6	Matched (3)
6	B	2012	-712	0	.10616	1.649602	.2603625	5	8	5	Matched (3)
7	B	2013	-1933	0	.3537852	1.593719	-.2343988	9	4	5	Matched (3)
8	B	2014	3073	1	.726777	1.691758	.2562243	1	5	1	Matched (3)
9	C	2011	-1292	0	1.312561	-1.293136	.204083	4	5	4	Matched (3)
10	C	2012	-3416	0	1.177484	-1.344218	.2839719	6	9	6	Matched (3)
11	C	2013	-3558	0	1.256408	-1.259951	.3733927	6	5	3	Matched (3)
12	C	2014	1225	1	1.421545	-1.311745	-.3759656	7	3	3	Matched (3)

Step-by-step procedure for your own data:

- When working on your own data, after the **use** command, write the path name of the location of your master dataset (in my case, *mydata1* is located at "C:/Users/ma3482/Documents/Merge-Append/mydata1.dta"). So, I used the following codes:

  ```
  use "C:/Users/ma3482/Documents/Merge-Append/mydata1.dta"
  ```

- Identify the common variables in your dataset. In my case, **country** and **year** are the common variables. Make sure that after the **using** command, you are writing the path name where your using dataset is located (in my case, *mydata2* is located at "C:/Users/ma3482/Documents/Merge-Append/mydata2.dta"). I used

IBM v. Digital Doors, IPR2023-00972, 7,958,268
DigitalDoors EX2033

the following codes:

```
merge 1:1 country year using "C:/Users/ma3482/Documents/Merge-Append/mydata2.dta"
```

Notes:

- 1:1 merge is less common than m:1 and 1:m merges.

- If you want to retain only selected variables from the "using" file, you need to use `keepusing` command and tell Stata which variables you want to keep. For example, if you want to merge mydata1 and mydata2, and want to merge variables x4 and x5 only from mydata2 to mydata1, use the following codes:

```
use https://dss.princeton.edu/training/mydata1.dta, clear

merge 1:1 country year
using https://dss.princeton.edu/training/mydata2.dta, keepusing
(x4 x5)
```

- In the merged/matched dataset above, notice that a new variable `_merge` has also been created because of merge. Usually, there are three possible values for this `_merge` variable. The meaning of these values are as follows:

- `_merge` = 1: A case that is present in the master dataset only.
- `_merge` = 2: A case that is present in the using dataset only.
- `_merge` = 3: A case that is present in both datasets.

- If you want to keep only observations in both datasets, type:

```
use https://dss.princeton.edu/training/mydata1.dta, clear

merge 1:1 country year
using https://dss.princeton.edu/training/mydata2.dta, keepusing
(x4 x5)keep (match)
```

Note: You need to drop the `_merge` variable if you want to conduct more merging tasks using your master dataset. For dropping the variable type:

```
drop _merge
```

2. Many-to-One Merge

We use many-to-one merge when the common variable(s) uniquely identifies the observations in the using data, but not necessarily in the master data. Let us merge the following two datasets, named *mydata1* (master dataset) and *mydata3* (using dataset).

mydata1

IBM v. Digital Doors,
IPR2023-00972, 7,958,268
DigitalDoors EX2033

	country	year	y	y_bin	x1	x2	x3
1	A	2011	1343	1	.2779036	-1.107956	.2825536
2	A	2012	-1900	0	.3206847	-.94872	.4925385
3	A	2013	-1123	0	.3634657	-.789484	.7025234
4	A	2014	2646	1	.246144	-.885533	-.0943909
5	B	2011	-5935	0	-.08185	1.42512	.0234281
6	B	2012	-712	0	.10616	1.649602	.2603625
7	B	2013	-1933	0	.3537852	1.593719	-.2343988
8	B	2014	3073	1	.726777	1.691758	.2562243
9	C	2011	-1292	0	1.312561	-1.293136	.204083
10	C	2012	-3416	0	1.177484	-1.344218	.2839719
11	C	2013	-3558	0	1.256408	-1.259951	.3733927
12	C	2014	1225	1	1.421545	-1.311745	-.3759656

mydata3

	country	x7
1	A	100
2	B	200
3	C	300

Notice that **country** is the common variable in above two datasets.

Use the following Stata commands for m:1 merge.

```
use https://dss.princeton.edu/training/mydata1.dta

merge m:1 country
using https://dss.princeton.edu/training/mydata3.dta
```

```
    Result                      Number of obs
    -----------------------------------------
    Not matched                             0
    Matched                                12   (_merge==3)
    -----------------------------------------
```

The resulting merged/matched dataset is as follows.

	country	year	y	y_bin	x1	x2	x3	x7	_merge
1	A	2011	1343	1	.2779036	-1.107956	.2825536	100	matched (3)
2	A	2012	-1900	0	.3206847	-.94872	.4925385	100	matched (3)
3	A	2013	-1123	0	.3634657	-.789484	.7025234	100	matched (3)
4	A	2014	2646	1	.246144	-.885533	-.0943909	100	matched (3)
5	B	2011	-5935	0	-.08185	1.42512	.0234281	200	matched (3)
6	B	2012	-712	0	.10616	1.649602	.2603625	200	matched (3)
7	B	2013	-1933	0	.3537852	1.593719	-.2343988	200	matched (3)
8	B	2014	3073	1	.726777	1.691758	.2562243	200	matched (3)
9	C	2011	-1292	0	1.312561	-1.293136	.204083	300	matched (3)
10	C	2012	-3416	0	1.177484	-1.344218	.2839719	300	matched (3)
11	C	2013	-3558	0	1.256408	-1.259951	.3733927	300	matched (3)
12	C	2014	1225	1	1.421545	-1.311745	-.3759656	300	matched (3)

Step-by-step procedure for your own data:

- When working on your own data, after the **use** command, write the path name of the location of your master dataset (in my case, *mydata1* is located at "C:/Users/ma3482/Documents/Merge-Append/mydata1.dta"). So, I used the following codes:

IBM v. Digital Doors,
IPR2023-00972, 7,958,268
DigitalDoors EX2033

```
use "C:/Users/ma3482/Documents/Merge-Append/mydata1.dta"
```

- Identify the common variables in your dataset. In my case, **`country`** is the common variable. Make sure that after the **`using`** command, you are writing the path name of your using dataset's location (in my case, *mydata3* is located at "C:/Users/ma3482/Documents/Merge-Append/mydata2.dta"). I used the following codes:

```
merge m:1 country using "C:/Users/ma3482/Documents/Merge-Append/mydata3.dta"
```

3. One-to-Many Merge

One-to-many merge is similar to many-to-one merge, except that now the common variable(s) identifies unique observations in the master dataset. In this case, *mydata3* (see above) would be considered as master data and *mydata1* (see above) as using data. The Stata command for one-to-many merge is **`merge 1:m`**

4. Many-to-Many Merge

We use m:m merge when the common variable(s) does not uniquely identify the observations in either dataset. The use of m:m merge is not usually encouraged. The Stata command for many-to-many merge is **`merge m:m`**

More Notes on Merge:

- Sometimes you will find that a variable with the same name is present in both the master and using datasets. For example, the variable x4 may be there in both the master and using datasets. In this situation, by default, Stata keeps the observations for this duplicate variable from the master dataset and drops the observations from the using dataset. However, you can apply your own choice by adding **`update`** or **`update replace`** after the main command.

- If you use **`,update`** after the main command, missing values in the master dataset for the duplicate variable will be replaced by values from the using dataset. An example of this code is as follows:

```
merge 1:1 country year using mydata2, update
```

- If you use **`,update replace`** after the main command, all values in the master dataset for the duplicate variable will be replaced by values from the using dataset. An example of this code is as follows:

```
merge 1:1 country year using mydata2, update replace
```

- While merging datasets, you may sometimes get the following error message from Stata: "`variable id does not uniquely identify observations in the master data`" . You may see this message for several reasons, such as

  - using 1:m instead of m:1, or vice-versa.

IBM v. Digital Doors,
IPR2023-00972, 7,958,268
DigitalDoors EX2033

- specifying just one common variable when you need two.

- presence of duplicate IDs in your datasets.

To avoid getting this message, inspect datasets carefully before running the merge codes.

For more details about merge, type `help merge` in Stata, or consult the document, which I have used in preparing this guide, available here: https://www.stata.com/manuals/dmerge.pdf

Append

- We use append to add observations to a dataset.

- Appending two datasets requires that both datasets have variables with exactly the same name.

- When using categorical data, make sure the categories of variables on both datasets refer to exactly the same thing (e.g., both datasets should have "White" =1, "Black" = 2, “Hispanic” = 3).

Let us append the following two datasets using Stata.

mydata4

	country	year	y	y_bin	x1	x2	x3
1	A	2011	1343	1	.2779036	-1.107956	.2825536
2	A	2012	-1900	0	.3206847	-.94872	.4925385
3	B	2011	-5935	0	-.08185	1.42512	.0234281
4	B	2012	-712	0	.10616	1.649602	.2603625
5	C	2011	-1292	0	1.312561	-1.293136	.204083
6	C	2012	-3416	0	1.177484	-1.344218	.2839719

mydata5

	country	year	y	y_bin	x1	x2	x3
1	A	2013	-1123	0	.3634657	-.789484	.7025234
2	A	2014	2646	1	.246144	-.885533	-.0943909
3	B	2013	-1933	0	.3537852	1.593719	-.2343988
4	B	2014	3073	1	.726777	1.691758	.2562243
5	C	2013	-3558	0	1.256408	-1.259951	.3733927
6	C	2014	1225	1	1.421545	-1.311745	-.3759656

Notice the difference between the two datasets. *mydata4* contains information for the years 2011 and 2012 for countries A, B, C, and for the variables y, y_bin, x1, x2, and x3. On the other hand, *mydata5* contains information for the years 2013 and 2014 for the same countries and variables recorded in *mydata4*. Basically, we will update the dataset *mydata4* by adding observations for all variables for the years 2013 and 2014. Use the following Stata commands to do that.

use https://dss.princeton.edu/training/mydata4.dta
append using https://dss.princeton.edu/training/mydata5.dta

The resulting dataset is as follows.

appended data

	country	year	y	y_bin	x1	x2	x3
1	A	2011	1343	1	.2779036	-1.107956	.2825536
2	A	2012	-1900	0	.3206847	-.94872	.4925385
3	B	2011	-5935	0	-.08185	1.42512	.0234281
4	B	2012	-712	0	.10616	1.649602	.2603625

IBM v. Digital Doors, IPR2023-00972, 7,958,268 DigitalDoors EX2033

5	C	2011	-1292	0	1.312561	-1.293136	.204083
6	C	2012	-3416	0	1.177484	-1.344218	.2839719
7	A	2013	-1123	0	.3634657	-.789484	.7025234
8	A	2014	2646	1	.246144	-.885533	-.0943909
9	B	2013	-1933	0	.3537852	1.593719	-.2343988
10	B	2014	3073	1	.726777	1.691758	.2562243
11	C	2013	-3558	0	1.256408	-1.259951	.3733927
12	C	2014	1225	1	1.421545	-1.311745	-.3759656

You can sort the appended dataset, which will help you inspect your data easily. For sorting, type:

sort country year

The sorted dataset is as follows.

	country	year	y	y_bin	x1	x2	x3
1	A	2011	1343	1	.2779036	-1.107956	.2825536
2	A	2012	-1900	0	.3206847	-.94872	.4925385
3	A	2013	-1123	0	.3634657	-.789484	.7025234
4	A	2014	2646	1	.246144	-.885533	-.0943909
5	B	2011	-5935	0	-.08185	1.42512	.0234281
6	B	2012	-712	0	.10616	1.649602	.2603625
7	B	2013	-1933	0	.3537852	1.593719	-.2343988
8	B	2014	3073	1	.726777	1.691758	.2562243
9	C	2011	-1292	0	1.312561	-1.293136	.204083
10	C	2012	-3416	0	1.177484	-1.344218	.2839719
11	C	2013	-3558	0	1.256408	-1.259951	.3733927
12	C	2014	1225	1	1.421545	-1.311745	-.3759656

Step-by-step procedure for your own data:

- When working on your own data, after the **`use`** command, write the path name of the location of your master dataset (in my case, *mydata4* is located at "C:/Users/ma3482/Documents/Merge-Append/mydata4.dta"). So, I used the following codes:

  `use "C:/Users/ma3482/Documents/Merge-Append/mydata4.dta"`

- Make sure that after the **`using`** command, you are writing the path name of your using dataset's location (in my case, *mydata5* is located at "C:/Users/ma3482/Documents/Merge-Append/mydata5.dta"). I used the following codes:

  `append using "C:/Users/ma3482/Documents/Merge-Append/mydata5.dta"`

- For sorting, type:

  sort country year

Example of appended data when one dataset is missing one variable

Example of appended data when one dataset is missing one variable

Let us append *mydata4* and *mydata6.*

mydata4

	country	year	y	y_bin	x1	x2	x3
1	A	2011	1343	1	.2779036	-1.107956	.2825536
2	A	2012	-1900	0	.3206847	-.94872	.4925385
3	B	2011	-5935	0	-.08185	1.42512	.0234281
4	B	2012	-712	0	.10616	1.649602	.2603625
5	C	2011	-1292	0	1.312561	-1.293136	.204083
6	C	2012	-3416	0	1.177484	-1.344218	.2839719

mydata6

	country	year	y	y_bin	x1	x2
1	A	2013	-1123	0	.3634657	-.789484
2	A	2014	2646	1	.246144	-.885533
3	B	2013	-1933	0	.3537852	1.593719
4	B	2014	3073	1	.726777	1.691758
5	C	2013	-3558	0	1.256408	-1.259951
6	C	2014	1225	1	1.421545	-1.311745

Notice that *mydata6* does not contain variable x3.

For appending *mydata4* and *mydata6,* use the following Stata codes.

use https://dss.princeton.edu/training/mydata4.dta
append using https://dss.princeton.edu/training/mydata6.dta

appended dataset

	country	year	y	y_bin	x1	x2	x3
1	A	2011	1343	1	.2779036	-1.107956	.2825536
2	A	2012	-1900	0	.3206847	-.94872	.4925385
3	B	2011	-5935	0	-.08185	1.42512	.0234281
4	B	2012	-712	0	.10616	1.649602	.2603625
5	C	2011	-1292	0	1.312561	-1.293136	.204083
6	C	2012	-3416	0	1.177484	-1.344218	.2839719
7	A	2013	-1123	0	.3634657	-.789484	.
8	A	2014	2646	1	.246144	-.885533	.
9	B	2013	-1933	0	.3537852	1.593719	.
10	B	2014	3073	1	.726777	1.691758	.
11	C	2013	-3558	0	1.256408	-1.259951	.
12	C	2014	1225	1	1.421545	-1.311745	.

Notice the missing observations in the above appended dataset.

References

Baum, C. F. (2011). Introduction to Stata. Faculty Micro Resource Center, Boston College. Available ate: http://fmwww.bc.edu/GStat/docs/StataIntro.pdf, *accessed August, 23* (2022).

DSS Online Training Section https://dss.princeton.edu/training/

Princeton DSS Libguides https://libguides.princeton.edu/dss

STATA FAQ Website http://stata.com/support/faqs/

Stata Manual for Merge https://www.stata.com/manuals/dmerge.pdf

UCLA Resources https://stats.oarc.ucla.edu/stata/

Comments or Questions?

If you have questions or comments about this guide or method, please email data@Princeton.edu.

Last Updated: May 26, 2023 4:01 PM | **URL:** https://libguides.princeton.edu/stata-merge-append | Print Page Login to LibApps

Subjects: Data & Statistical Services **Tags:** append, merge, stata