**joinby — Form all pairwise combinations within groups**

### Description

`joinby` joins, within groups formed by `varlist`, observations of the dataset in memory with `filename`, a Stata-format dataset. By *join* we mean to form all pairwise combinations. `filename` is required to be sorted by `varlist`. If `filename` is specified without an extension, `.dta` is assumed.

If `varlist` is not specified, `joinby` takes as `varlist` the set of variables common to the dataset in memory and in `filename`.

Observations unique to one or the other dataset are ignored unless `unmatched()` specifies differently. Whether you load one dataset and join the other or vice versa makes no difference in the number of resulting observations.

If there are common variables between the two datasets, however, the combined dataset will contain the values from the master data for those observations. This behavior can be modified with the `update` and `replace` options.

### Quick start

**Form pairwise combinations of observations from mydata1.dta in memory with those from mydata2.dta using all common variables and drop unmatched observations**

```
joinby using mydata2
```

As above, but join on v1, v2, and v3

```
joinby v1 v2 v3 using mydata2
```

As above, but include unmatched observations only from mydata2.dta and add `_merge` indicating whether the variable was in both datasets or only the using dataset

```
joinby v1 v2 v3 using mydata2, unmatched(using)
```

As above, but include unmatched observations only from mydata1.dta

```
joinby v1 v2 v3 using mydata2, unmatched(master)
```

As above, but name the variable indicating the source of the observation `newv`

```
joinby v1 v2 v3 using mydata2, unmatched(master) _merge(newv)
```

Replace missing data in mydata1.dta with values from mydata2.dta

```
joinby v1 v2 v3 using mydata2, update
```

Replace missing and conflicting data in mydata1.dta with values from mydata2.dta

```
joinby v1 v2 v3 using mydata2, update replace
```
**Syntax**

```
joinby [ `varlist' ] using `filename' [, `options']
```

<table>
<thead>
<tr>
<th><code>options</code></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>update</code></td>
<td>replace missing data in memory with values from <code>filename</code></td>
</tr>
<tr>
<td><code>replace</code></td>
<td>replace all data in memory with values from <code>filename</code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>When observations do not match:</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>unmatched(none)</code></td>
</tr>
<tr>
<td><code>unmatched(both)</code></td>
</tr>
<tr>
<td><code>unmatched(master)</code></td>
</tr>
<tr>
<td><code>unmatched(using)</code></td>
</tr>
</tbody>
</table>

| `__merge(varname)` | `varname` marks source of resulting observation; default is `__merge` |
| `nolabel` | do not copy value-label definitions from `filename` |

`varlist` may not contain `strL`s.

**Options**

`update` varies the action that `joinby` takes when an observation is matched. By default, values from the master data are retained when the same variables are found in both datasets. If `update` is specified, however, the values from the using dataset are retained where the master dataset contains missing.

`replace`, allowed with `update` only, specifies that nonmissing values in the master dataset be replaced with corresponding values from the using dataset. A nonmissing value, however, will never be replaced with a missing value.

`unmatched(none | both | master | using)` specifies whether observations unique to one of the datasets are to be kept, with the variables from the other dataset set to missing. Valid values are:

- `none`  | ignore all unmatched observations (default) |
- `both`  | include unmatched observations from the master and using data |
- `master` | include unmatched observations from the master data |
- `using` | include unmatched observations from the using data |

`__merge(varname)` specifies the name of the variable that will mark the source of the resulting observation. The default name is `__merge(__merge)`. To preserve compatibility with earlier versions of `joinby`, `__merge` is generated only if `unmatched` is specified.

`nolabel` prevents Stata from copying the value-label definitions from the dataset on disk into the dataset in memory. Even if you do not specify this option, label definitions from the disk dataset do not replace label definitions already in memory.
Remarks and examples

The following, admittedly artificial, example illustrates `joinby`.

Example 1

We have two datasets: `child.dta` and `parent.dta`. Both contain a `family_id` variable, which identifies the people who belong to the same family.

```
. use http://www.stata-press.com/data/r15/child
   (Data on Children)
. describe
Contains data from http://www.stata-press.com/data/r15/child.dta
    obs:      5  Data on Children
    vars:     4  11 Dec 2016 21:08
    size:    30

    variable name   storage  display format    value  variable label
                   variable type    format label
    family_id       int      %8.0g label     Family ID number
    child_id        byte     %8.0g label     Child ID number
    x1              byte     %8.0g label
    x2              int      %8.0g label

Sorted by: family_id
. list

+---------+--------+---+---+---+---------+---------+---------+
| family- | child_ |   |   |   |         |         |
|    d    |   id   |   |   |   |         |         |
|---------+--------+---+---+---+---------+---------+---------|
|    1    |  1025  |  3 | 11| 320|         |         |
|    2    |  1025  |  1 | 12| 300|         |         |
|    3    |  1025  |  4 | 10| 275|         |         |
|    4    |  1026  |  2 | 13| 280|         |         |
|    5    |  1027  |  5 | 15| 210|         |         |
```

. use http://www.stata-press.com/data/r15/parent
   (Data on Parents)
. describe
Contains data from http://www.stata-press.com/data/r15/parent.dta
    obs:      6  Data on Parents
    vars:     4  11 Dec 2016 03:06
    size:    84

    variable name   storage  display format    value  variable label
                   variable type    format label
    family_id       int      %8.0g label     Family ID number
    parent_id       float    %9.0g label     Parent ID number
    x1              float    %9.0g label
    x2              float    %9.0g label

Sorted by:
We want to join the information for the parents and their children. The data on parents are in memory, and the data on children are posted at http://www.stata-press.com. child.dta has been sorted by family_id, but parent.dta has not, so first we sort the parent data on family_id:

```
. sort family_id
. joinby family_id using http://www.stata-press.com/data/r15/child
```

```
Contains data
obs: 8
vars: 6
size: 136
```

```
+------------+----------------+---------+---------+------------+---------+---------+---------+---------+---------+---------+---------+
| family_id | parent_id | x1 | x3 | child_id | x2 | family_id | parent_id | x1 | x3 | child_id | x2 |
|-----------|-----------|----|----|----------|----|-----------|-----------|----|----|----------|----|-----------|-----------|----|----|----------|----|-----------|-----------|
| 1025      | 12        | 27 | 721| 1        | 300| 1025      | 12        | 27 | 721| 4        | 275| 1025      | 12        | 27 | 721| 3        | 320| 1025      | 11        | 20 | 643| 3        | 320| 1025      | 11        | 20 | 643| 1        | 300| 1025      | 11        | 20 | 643| 4        | 275| 1025      | 11        | 20 | 643| 1        | 300| 1025      | 11        | 20 | 643| 4        | 275|
| 1026      | 14        | 26 | 668| 2        | 280| 1026      | 14        | 26 | 668| 2        | 280| 1026      | 14        | 26 | 668| 2        | 280| 1026      | 14        | 26 | 668| 2        | 280| 1026      | 14        | 26 | 668| 2        | 280|
```

1. The x1 variable is in both datasets. Values for this variable in the joined dataset are the values from parent.dta—the dataset in memory when we issued the joinby command. If we had child.dta in memory and parent.dta on disk when we requested joinby, the values for x1

Note: Dataset has changed since last saved.

```
. list, sepby(family_id) abbrev(12)
```

1. family_id of 1027, which appears only in child.dta, and family_id of 1030, which appears only in parent.dta, are not in the combined dataset. Observations for which the matching variables are not in both datasets are omitted.

2. The x1 variable is in both datasets. Values for this variable in the joined dataset are the values from parent.dta—the dataset in memory when we issued the joinby command. If we had child.dta in memory and parent.dta on disk when we requested joinby, the values for x1
would have been those from child.dta. Values from the dataset in memory take precedence over the dataset on disk.

Acknowledgment

joinby was written by Jeroen Weesie of the Department of Sociology at Utrecht University, The Netherlands.

Reference

Baum, C. F. 2016. *An Introduction to Stata Programming*. 2nd ed. College Station, TX: Stata Press.

Also see

[D] append — Append datasets
[D] cross — Form every pairwise combination of two datasets
[D] fillin — Rectangularize dataset
[D] merge — Merge datasets
[D] save — Save Stata dataset
[U] 22 Combining datasets