

Note & Syntax 6.1

Appending data: combines datasets vertically, adding observations to the end of the master dataset.

Master data: data in memory. Using data: data to be appended.

Situation 1: two datasets have the same variables

One:

```
+-----+
|   id   a   b   c |
+-----+
1. | 1001  1   2   3 |
2. | 1002  4   5   6 |
3. | 1003  7   8   9 |
+-----+
```

Two:

```
+-----+
|   id   a   b   c |
+-----+
1. | 1003  10  11  12 |
2. | 1012  13  14  15 |
+-----+
```

```
. use one,clear
. append using two
. list
```

```
+-----+
|   id   a   b   c |
+-----+
1. | 1001  1   2   3 |
2. | 1002  4   5   6 |
3. | 1003  7   8   9 |
4. | 1011  10  11  12 |
5. | 1012  13  14  15 |
+-----+
```

Note: append will only add observations to the end of the master data, it will not change any values on master data, even if they have exact same id variables and values.

For example:

```
. use one,clear
. list
```

```
+-----+
|   id   a   b   c |
+-----+
1. | 1001  1   2   3 |
2. | 1002  4   5   6 |
3. | 1003  7   8   9 |
+-----+
```

```
. save three.dta,replace
. list
```

```
+-----+
|   id   a   b   c |
+-----+
1. | 1003  7   8   9 |
2. | 1012  13  14  15 |
+-----+
```

```
. append using three
. list
```

```
+-----+
|   id   a   b   c |
+-----+
1. | 1001  1   2   3 |
+-----+
```

```

2. | 1002   4   5   6 |
3. | 1003   7   8   9 |
4. | 1003   7   8   9 |
5. | 1012  13  14  15 |
+-----+

```

Situation 2: two datasets do not have the same variables (in some states, interviewers asked additional questions)

One:

```

+-----+
|   id   a   b   c |
+-----+
1. | 1001   1   2   3 |
2. | 1002   4   5   6 |
3. | 1003   7   8   9 |
+-----+

```

Two:

```

+-----+
|   id   a   d   e |
+-----+
1. | 1011  10  11  12 |
2. | 1012  13  14  15 |
+-----+

```

```

. clear
. use one
. append using two
. list

```

```

+-----+
|   id   a   b   c   d   e |
+-----+
1. | 1001   1   2   3   .   . |
2. | 1002   4   5   6   .   . |
3. | 1003   7   8   9   .   . |
4. | 1011  10   .   .  11  12 |
5. | 1012  13   .   .  14  15 |
+-----+

```

where necessary, filling in c and d with missing.

Situation 3: Use option "keep(varlist)" to specify the variables to be kept from the using dataset.

One:

```

+-----+
|   id   a   b   c |
+-----+
1. | 1001   1   2   3 |
2. | 1002   4   5   6 |
3. | 1003   7   8   9 |
+-----+

```

Two:

```

+-----+
|   id   a   d   e |
+-----+
1. | 1011  10  11  12 |
2. | 1012  13  14  15 |
+-----+

```

Now I only want to keep var "a" and "d", and do not want var "e".

```

. clear
. use one
. append using two,keep(id a d)
. list

```

	id	a	b	c	d
1.	1001	1	2	3	.
2.	1002	4	5	6	.
3.	1003	7	8	9	.
4.	1011	10	.	.	11
5.	1012	13	.	.	14

Merging data (combines datasets horizontally, adding variables. Append: adding observations):

Example of merge: longitudinal survey.

Situation 1: two datasets have the same number of observations and in same order

One:

	a	b	c
1.	1	2	3
2.	4	5	6
3.	7	8	9

Two:

	d
1.	10
2.	11
3.	12

```
. use one,clear
. merge using two
. list
```

	a	b	c	d	_merge
1.	1	2	3	10	3
2.	4	5	6	11	3
3.	7	8	9	12	3

NOTE: _merge==1 from "master" only; _merge==2 from "using" only; _merge==3 from both

Situation 2: two datasets do not have the same number of observations, but in same order.

Three:

	d
1.	10
2.	11
3.	12
4.	13

```
. use one,clear
. merge using three
. list
```

	a	b	c	d	_merge
1.	1	2	3	10	3
2.	4	5	6	11	3
3.	7	8	9	12	3
4.	.	.	.	13	2

NOTE:for observation 4:_merge==2 means this observation comes only from USING dataset.

Situation 3: if two datasets have variables in common: The values of the master data (data in memory) will be retained ONLY

One:

	a	b	c
1.	1	2	3
2.	4	5	6
3.	7	8	9

Two:

	a	d
1.	16	17
2.	18	19
3.	20	21
4.	22	23

use one,clear
merge using two
list

	a	b	c	d	_merge
1.	1	2	3	17	3
2.	4	5	6	19	3
3.	7	8	9	21	3
4.	22	.	.	23	2

NOTE: for first three observations, values of "a" and "d" are from MASTER data only, although USING data also have these variables.

Situation 4: Match-Merge: explicitly identifying the variable(or variables) by which the records are matched.

Master:

	id	a	b
1.	1	10	11
2.	2	12	13
3.	3	14	15
4.	5	16	17

Second:

+-----+

```

      | id   d |
      |-----|
1.   |  1  18 |
2.   |  3  19 |
3.   |  4  20 |
      +-----+
use master,clear
sort id
merge id using second
. list
      +-----+
      | id   a   b   d   _merge |
      |-----|
1.   |  1  10  11  18     3 |
2.   |  2  12  13   .     1 |
3.   |  3  14  15  19     3 |
4.   |  5  16  17   .     1 |
5.   |  4   .   .  20     2 |
      +-----+

```

NOTE: Stata did not blindly match observation 1 to observation 1, observation 2 to observation 2, etc. stata smartly matched data on id=1 to id=1, id=3 to id=3, and knew that id=4 in second.dta was unique. _merge=1/2/3 also clearly told us whether the data come from master, or using, or both.

```

. tab _merge

. count if _merge!=3

```