

Building your data set

Merging it together

- ▶ Sometimes, we want to add variables to one of our data sets.
- ▶ This process of combining data sets is called **merging**.
- ▶ In Stata's terminology, the data set in memory is called the **master** data and the data set to be merged the **using** data.
- ▶ For merging, you have to link the different data sets by an **identifier** (e.g., person ids, households ids, years)
 - ▶ The identifier should uniquely identify observations in at least one of the data sets.
 - ▶ If the identifier uniquely identifies observations in both data sets, Stata calls this a 1:1 merge.
 - ▶ If the identifier uniquely identifies observations in only one data set, Stata calls this an m:1 (or 1:m) merge.
 - ▶ To make an identifier unique, you may have to combine different variables (e.g., person ids and years uniquely identify person-year observations in an annual panel).

1:1 merges

- Use 1:1 merges if your identifier uniquely identifies your observations in your master and in your using data:

id	syear	inc
101	2015	2322
101	2016	2367
101	2017	0
202	2016	3500
202	2017	3700

Master data

id	syear	lsf
101	2015	7
101	2016	7
101	2017	5
202	2016	8
202	2017	9

Using data

id	syear	inc	lsf
101	2015	2322	7
101	2016	2367	7
101	2017	0	5
202	2016	3500	8
202	2017	3700	9

Merged data

- In panel data, this works when merging annual person data in long format (person-year-observations).

m:1 merges

- Use m:1 merges if the observations of your master file are multiples of the observations of your using data.

id	year	lsf
101	2015	7
101	2016	7
101	2017	5
202	2016	8
202	2017	9

Master data

id	sex
101	male
202	female

Using data

id	year	lsf	sex
101	2015	7	male
101	2016	7	male
101	2017	5	male
202	2016	8	female
202	2017	9	female

Merged data

- In panel data, such merges are useful when merging person-year data with time-invariant personal characteristics and household data.
- If you have members of the same household in your data, you have many person-year observations that should get identical annual household characteristics.

Example

- ▶ Let's **merge** some anonymized practice data from the SOEP:
- ▶ The file **pracdata_pl.dta** will be our **master** data.
- ▶ We want to subsequently **merge** the files **pracdata_inc.dta** and **pracdata_sex.dta** as **using** data.
- ▶ Command syntax:

```
merge 1:1 [identifier(s)] ///  
using "using data"
```



```
merge m:1 [identifier(s)] ///  
using "using data"
```
- ▶ If necessary, use a list of identifiers!

```
1 clear all
2 set more off
3 capture log close
4
5 [*****]
6 Session 6: Building your data set
7 [*****]
8
9 //Setting up your project folder
10 global wd "YOUR PATH"
11 global data "${wd}\Data"
12 global do "${wd}\Do"
13 global output "${wd}\Output"
14 global log "${wd}\Log"
15
16 //Change working directory to project folder
17 cd "${wd}"
18
19 [*****]
20 Merging the data
21 [*****]
22
23 //Open master data
24 use "${data}\pracdata_pl.dta", clear
25
26 merge 1:1 pid syear using "${data}\pracdata_inc.dta", nogen
27
28 merge m:1 pid using "${data}\pracdata_sex.dta", nogen
```

Successful merge

- ▶ The result of the `merge` will be displayed in the **Output Window**

```
merge 1:1 pid syyear using "${data}\pracdata_inc.dta"
```

Result	Number of obs
Not matched	0
Matched	19.585 (_merge==3)

- ▶ Observations that Stata identifies in your **master** and in your **using** data are **Matched**.
- ▶ If observations are not matched, Stata will tell you where these observations were found.
- ▶ Stata generates a new variable `_merge` that stores this information.
- ! Stata does not automatically overwrite variables. You will have to **drop** `_merge` before the next merge or use the option `nogenerate`.

Merging problems

- ▶ If observations are only in your **master**, but not in your **using** data, Stata cannot match anything during the **merge**.
- ▶ Stata will insert a **missing** and set `_merge` for these observations to 1 (master only) or 2 (using only).
- ▶ Most merges fail because of a wrong identifier. This happens when you tell Stata to match the master and using data on variables that are not unique within each data set (for 1:1 merge).

```
. merge 1:1 pid using "${data}\pracdata_inc.dta"  
variable pid does not uniquely identify observations in the master data  
r(459);
```

Merging options

- ▶ `keepusing(varlist)` lets you keep only specific variables from the **using** data.
- ▶ `keep(results)` lets you keep observations that were in your master data only (*results*=1), in your using data only (*results*=2) or matched (*results*=3).
- ▶ `generate(newvar)` lets you choose a different name for the variable `_merge`. This way, you will not have to **drop** `_merge` before starting the next merge.

```
merge 1:1 pid syear using "using data", ///  
      nogen keepusing(labinc_y) keep(master match)
```


- ▶ Thus far, we built our data set horizontally by adding variables.
- ▶ `append` allows you to add more observations (rows) to your data.
- ▶ Survey data is often published in waves and each wave gets a different data file. In these cases, you will need to append all files to have a full panel!
- ▶ `append` is simple, since you do not need to identify observations.
- ▶ If appended waves contain new variables, Stata assigns **missings** for older waves.
- ▶ Appending may become work-intensive if you have to harmonize variables.

Important final note

- ▶ Working with survey data may demand a combination of `merge` and `append`.

Step 1: `append` all data files that belong together across waves/years.

Step 2: `merge` all data files that now contain all waves/years.

- ▶ The SOEP provides most files in long-format, so you only need to merge.
- ▶ For BHPS/US we will provide a do-file to help you appending the data.