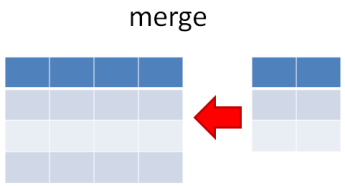
Stata Workshop 2: Data Management

Boriana Pratt

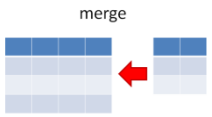
September 2016



Online Resources:

- Prof. Rodríguez: http://data.princeton.edu/stata/
- Mitchell (2010). *Data Management Using Stata.* http://www.stata.com/bookstore/data-management-using-stata/
- UCLA ATS: http://www.ats.ucla.edu/stat/stata/
- Stata Support Main: http://stata.com/support/
- Stata Resources: http://stata.com/links/

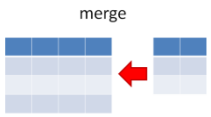
- Stata `help` command



Topics:

- Datasets
  - describe / list
  - tabulate / summarize
  - generate / replace
  - by-group / egen *
- Data to/from Excel
- Append / Merge
- Infile (reading free format, using dictionary)
- Date / Time *

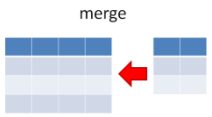
* Stata do file provided for self-study.



Setup:

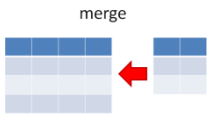
From dm.do file:

```
//check which directory I am in
pwd

//change directory to where the file (dm.do) is
cd C:\dm
```



Display

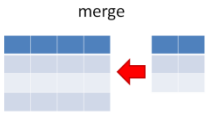
```
clear all
display 1+2

display ln(0.4/(1 - 0.4))
display logist(0.4)

//it can display strings also
display “Hello world”

//system values
display c(current_date)
```



Stata Dataset

Dataset is a matrix of observations (rows) and variables (columns)

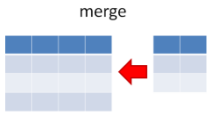
```
clear all

//describe data currently in memory ("master dataset")
describe

//create some observations
set obs 5

//create a variable x, with all values of 1
generate x = 1

//create another variable y, with values of built-in obs number _n
generate y = _n

//save the master data in a file on the hard drive
save mydata.dta, replace
```



Use a Dataset

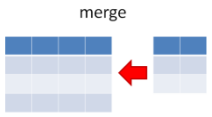
```
//load the data into memory
use mydata, clear
list
```

```
## . list
##      +--------+
##      | x    y |
##      |--------|
##   1. | 1    1 |
##   2. | 1    2 |
##   3. | 1    3 |
##   4. | 1    4 |
##   5. | 1    5 |
##      +--------+
```



Replace

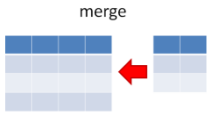
```
use mydata, clear

replace x = 2

//replace with if or in
replace x = 3 in 1/3
replace y = 9 if y == 5

//based on another variable's value
replace x = -99 if y<3

//make it missing
replace x = . if x == -99
replace y = . If y == 9

save mydata2, replace
```



See mydata2.dta

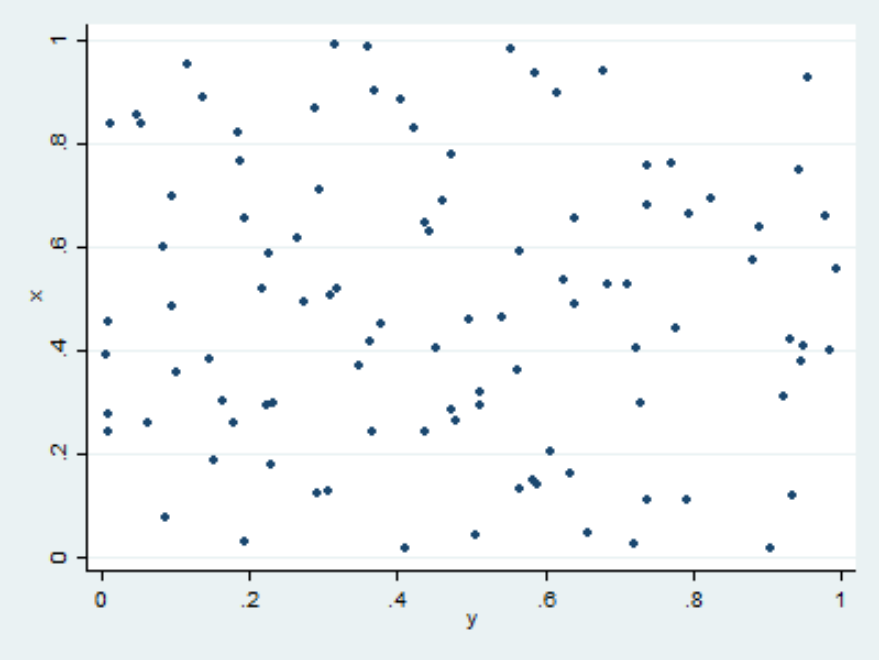
```
use mydata2, clear
list
```

```
##. list
##
##      +-------+
##      | x   y |
##      |-------|
##   1. | .   1 |
##   2. | .   2 |
##   3. | 3   3 |
##   4. | 2   4 |
##   5. | 2   . |
##      +-------+
```

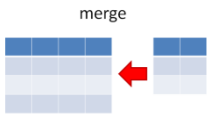


Random Data

```
clear all
set obs 100
set seed 34567
generate x = runiform()
generate y = runiform()
twoway scatter x y
graph export random.png, width(400) height(300) replace
```



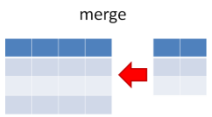
Questions thus far?



Missing Values

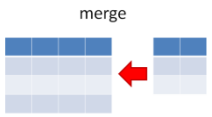
```
use mydata2, clear
list
```

```
##. list
##
##      +-------+
##      | x   y |
##      |-------|
##   1. | .   1 |
##   2. | .   2 |
##   3. | 3   3 |
##   4. | 2   4 |
##   5. | 2   . |
##      +-------+
```



Dichotomizing y

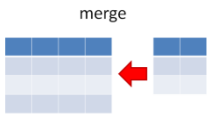
```
use mydata2, clear

// this may *not* be correct
generate high_y = 0
replace high_y = 1 if y > 2.5

// correct way
generate high_y2 = 0 if !missing(y)
replace high_y2 = 1 if y > 2.5 & !missing(y)

save mydata3, replace
```



See mydata3.dta

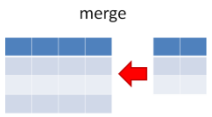
```
use mydata3, clear
list y high_y high_y2
```

```
##. list y high_y high_y2
##
##      +----------------------+
##      | y   high_y   high_y2 |
##      |----------------------|
##   1. | 1         0          0 |
##   2. | 2         0          0 |
##   3. | 3         1          1 |
##   4. | 4         1          1 |
##   5. | .         1          . |
##      +----------------------+
```



Save

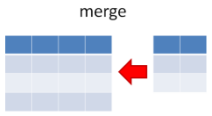
```
clear all

// let's create a dataset in memory
input id str10 name yob
1 "Amy" 1990
2 "Bob" 1991
3 "Cathy" 1989
4 “David" 1996
5 “Emma" 1998
end

rename yob year_of_birth
save birth, replace
```



Use

```
use birth, clear
assert _N == 5
list, abbreviate(15)
```

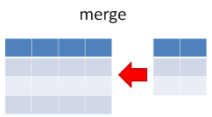
```
##. use birth.dta, clear
##. assert _N == 5
##. list, abbreviate(15)
##
##      +----------------------------+
##      | id    name   year_of_birth |
##      |----------------------------|
##   1. |  1     Amy            1990 |
##   2. |  2     Bob            1991 |
##   3. |  3   Cathy            1989 |
##   4. |  4   David            1996 |
##   5. |  5    Emma            1998 |
##      +----------------------------+
```



Labels

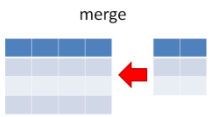
```
use birth, clear
generate int gender = 1 if name =="Amy" | name =="Cathy" | name ==“Emma"
replace gender = 2 if name =="Bob" | name ==“David"
describe gender
tabulate gender
save birth2, replace
```

```
##. generate int gender = 1 if name=="Amy" | name=="Cathy" | name==“Emma"
##(2 missing values generated)
##. replace gender = 2 if name == "Bob" | name == “David"
##(2 real changes made)
##. tabulate gender
##
##     gender |      Freq.     Percent        Cum.
##------------+-----------------------------------
##          1 |          3       60.00       60.00
##          2 |          2       40.00      100.00
##------------+-----------------------------------
##      Total |          5      100.00
```



Value Labels

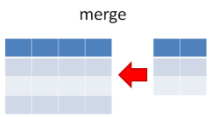
```
use birth2, clear

// 1. create the value label itself
label define gender 1 "girl" 2 "boy"
// 2. attach the value label to a variable
label values gender gender

save birth3, replace
```



Check

```
use birth3, clear

tabulate gender
```

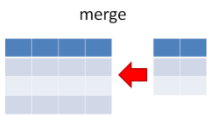
```
##. tabulate gender
##
##     gender |      Freq.     Percent        Cum.
##------------+-----------------------------------
##       girl |          3       60.00       60.00
##        boy |          2       40.00      100.00
##------------+-----------------------------------
##      Total |          5      100.00
```



Variable Label

```
use birth3, clear

//labeling variable is simple.
label variable gender "Gender of the respondent"
describe gender
```

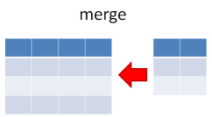
```
##. label variable gender "Gender of the respondent"
##. describe gender
##
##              storage   display    value
##variable name   type    format     label      variable label
##------------------------------------------------------------------------------
##gender          int     %8.0g      gender     Gender of the respondent
```



Look at Auto.dta

```
sysuse auto, clear
describe
```

```
##. sysuse auto, clear
##(1978 Automobile Data)
##. describe
##Contains data from C:\Program Files (x86)\Stata14\ado\base/a/auto.dta
##  obs:            74                          1978 Automobile Data
## vars:            12                          13 Apr 2014 17:45
## size:         3,182                          (_dta has notes)
##-------------------------------------------------------------------------------
##              storage   display    value
##variable name   type    format     label      variable label
##-------------------------------------------------------------------------------
##make            str18   %-18s                 Make and Model
##price           int     %8.0gc                Price
##mpg             int     %8.0g                 Mileage (mpg)
##rep78           int     %8.0g                 Repair Record 1978
##headroom        float   %6.1f                 Headroom (in.)
```

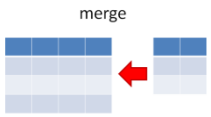


Tabulate

With Value Label

```
sysuse auto, clear
tabulate foreign
```

```
##. tabulate foreign
##
##   Car type |      Freq.     Percent        Cum.
##------------+-----------------------------------
##   Domestic |         52       70.27       70.27
##    Foreign |         22       29.73      100.00
##------------+-----------------------------------
##      Total |         74      100.00
```

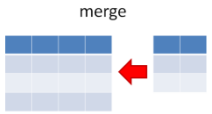


Tabulate

Without Value Label

```
sysuse auto, clear
tabulate foreign, nolabel
```

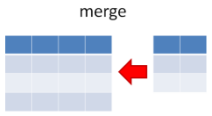
```
##. tabulate foreign, nolabel
##
##   Car type |      Freq.     Percent        Cum.
##------------+-----------------------------------
##          0 |         52       70.27       70.27
##          1 |         22       29.73      100.00
##------------+-----------------------------------
##      Total |         74      100.00
```



Summarize

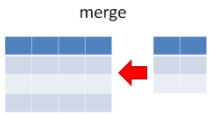
```
sysuse auto, clear
summarize price mpg
```

```
##. sysuse auto, clear
##(1978 Automobile Data)
##. summarize price mpg
##
##    Variable |        Obs        Mean    Std. Dev.       Min        Max
##-------------+---------------------------------------------------------
##       price |         74    6165.257    2949.496       3291      15906
##         mpg |         74     21.2973    5.785503         12         41
```



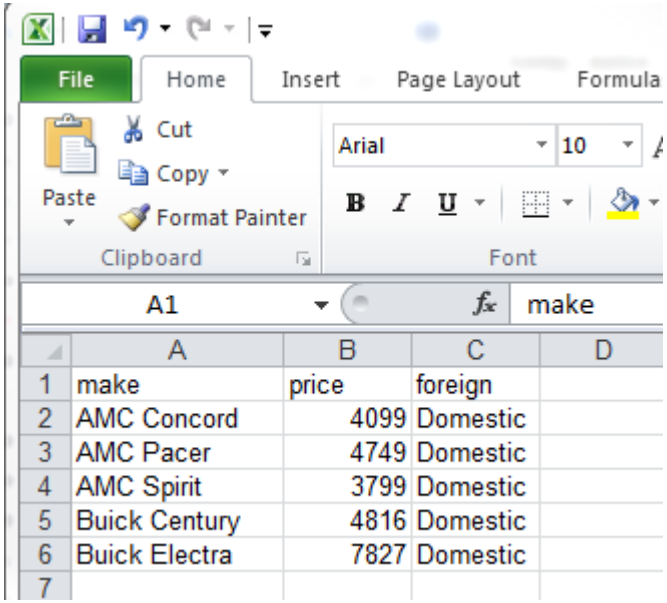
Other useful commands

```
sysuse auto, clear

describe make mpg price
inspect make mpg price
codebook make mpg price
```



Export to Excel

```
clear all
sysuse auto
keep make price foreign // drop all other variables
keep in 1/5
export excel using auto.xls, replace first(var)
!start auto.xls
// !open auto.xls // mac
```



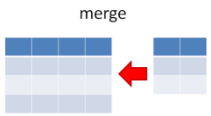
	A	B	C	D
1	make	price	foreign	
2	AMC Concord	4099	Domestic	
3	AMC Pacer	4749	Domestic	
4	AMC Spirit	3799	Domestic	
5	Buick Century	4816	Domestic	
6	Buick Electra	7827	Domestic	
7				

Import from Excel

```
clear all
import excel using auto.xls, clear firstrow
describe
```

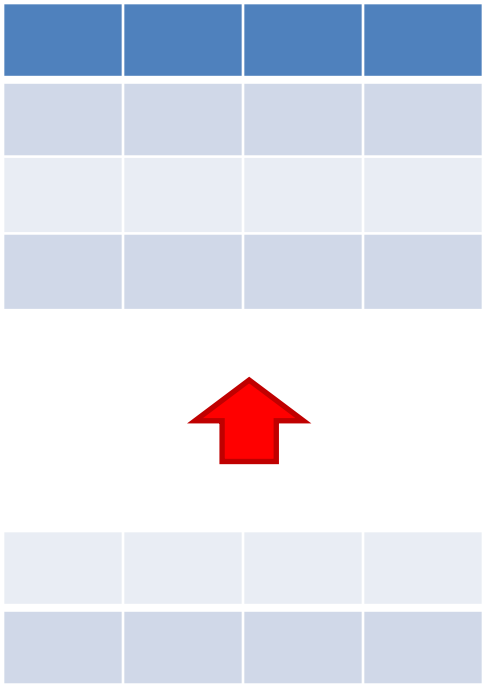
```
##. import excel using auto.xls, clear firstrow
##. describe
##Contains data
##  obs:             5
## vars:             3
## size:           115
##------------------------------------------------------------------------------
##storage   display    value
##variable name   type    format     label      variable label
##------------------------------------------------------------------------------
##make            str13   %13s                  make
##price           int     %10.0g                price
##foreign         str8    %9s                   foreign
##------------------------------------------------------------------------------
##Sorted by:
##     Note: Dataset has changed since last saved.
```

Any Questions so far?

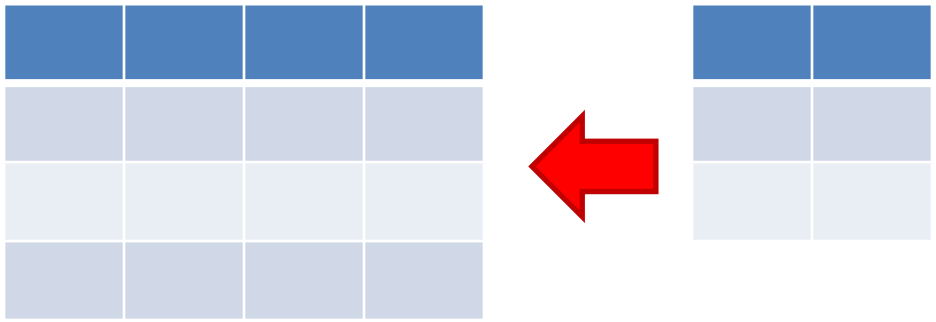


Combining Datasets

append

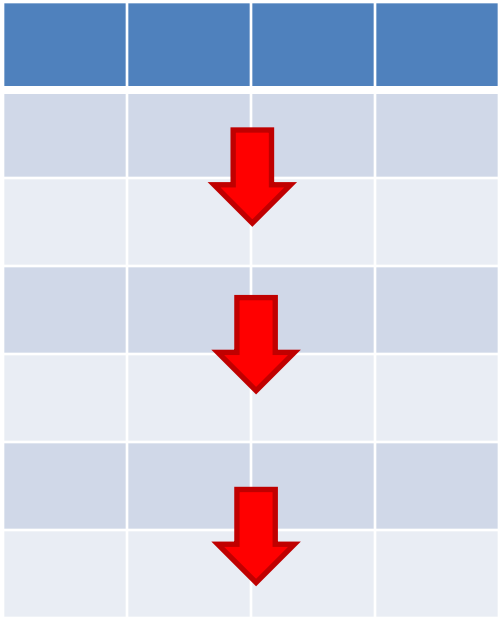


merge

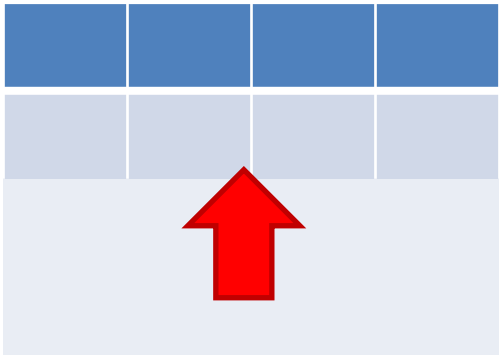


Combining Datasets

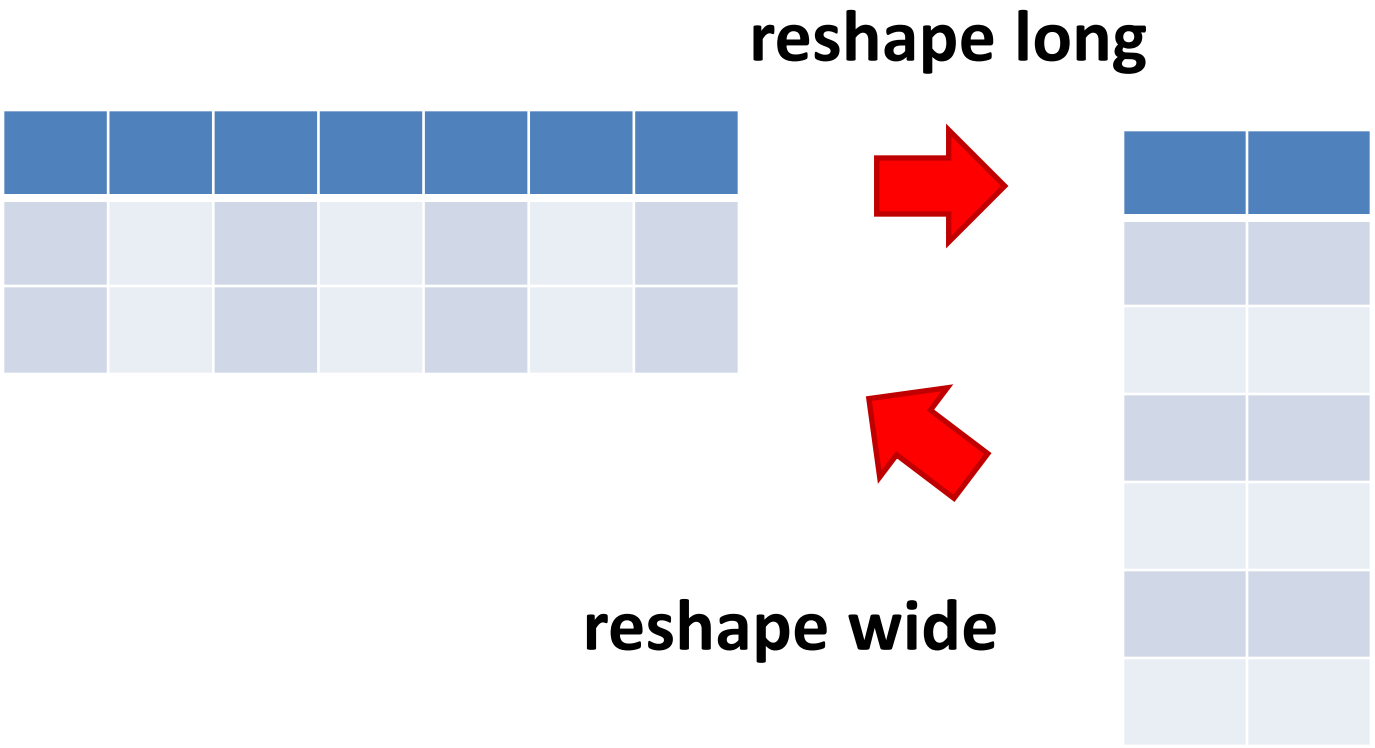
expand



collapse / contract

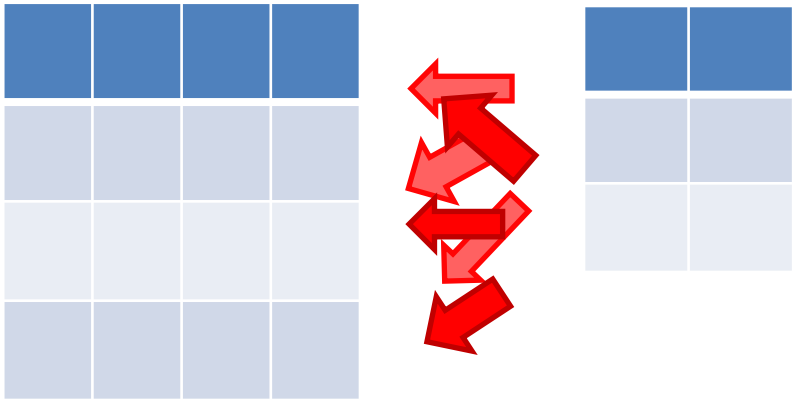


Combining Datasets



Combining Datasets

cross



joinby

key			
B			
A			
B			

key	
A	
B	

Append

Create odd.dta

```
clear all
use http://www.stata-press.com/data/r14/odd1
keep in 1/3
list
save odd.dta, replace
```

```
##. use http://www.stata-press.com/data/r14/odd1
##(First five odd numbers)
##. keep in 1/3
##(2 observations deleted)
##. list
##     +--------------+
##     | odd   number |
##     |--------------|
##  1. |   1        1 |
##  2. |   3        2 |
##  3. |   5        3 |
##     +--------------+
##. save odd.dta, replace
##(note: file odd.dta not found)
##file odd.dta saved
```

Append

Create even.dta

```
clear all
input number even odd
4 10 .
5 12 .
end
list
save even.dta, replace
```

```
##clear all
##. input number even odd
##        number       even        odd
##  1. 4 10 .
##  2. 5 12 .
##  3. end
##
##. list
##     +----------------------+
##     | number   even   odd |
##     |----------------------|
##  1. |      4     10     . |
##  2. |      5     12     . |
##     +----------------------+
```

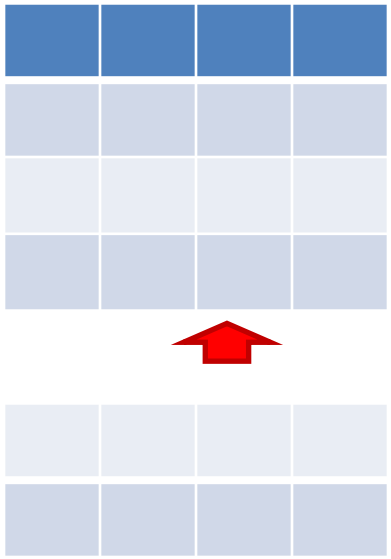
Append

Put *odd* and *even* together

```
use odd.dta, clear
append using even.dta, generate(Froma)
list
```

```
##. use odd.dta, clear
##(First five odd numbers)
##. append using even.dta, generate(Froma)
##. list
##
##     +-----------------------------+
##     | odd   number   Froma   even |
##     |-----------------------------|
##  1. |   1        1       0      . |
##  2. |   3        2       0      . |
##  3. |   5        3       0      . |
##  4. |   .        4       1     10 |
##  5. |   .        5       1     12 |
##     +-----------------------------+
```

Append



- Syntax: `append using` *`filename`* `[, options]`
- Appends a dataset stored on disk (the *using file*) to dataset in memory (master dataset)
- New dataset will have more observations (and possibly more variables)
- Variables are matched by *name* (not by order)
- Non-matched variables on the using side will be included

Merge

```
use age.dta, clear
merge 1:1 id using weight, report
drop _merge
save ageWeight, replace
```

id	age
1	22
2	56
5	17

id	wt
1	130
2	180
4	110

Id	age	wt	_merge
1	22	130	3
2	56	180	3
5	17	.	1
4	.	110	2

How Stata Merges

- Manual says (*Stata Data Management Manual v14* [D], p.456):

  The formal definition for merge behavior is the following: Start with the first observation of the master. Find the corresponding observation in the using data, if there is one. Record the matched or unmatched result. Proceed to the next observation in the master dataset. When you finish working through the master dataset, work through unused observations from the using data. By default, unmatched observations are kept in the merged data, whether they come from the master dataset or the using dataset.

- See also Bill Gould's two-part blog on "Merging data" at:
  - Part 1: Merges gone bad @ http://tinyurl.com/jvtloka
  - Part 2: Multiple-key merges @ http://tinyurl.com/krhs7xn

One-to-one Match Merge Pointers

master

id	age
1	22
2	56
5	17

using

id	wt
1	130
2	180
4	110

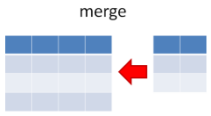
merge 1:1 id using "*using file name*"

id	age	wt	_merge
1	22	130	3
2	56	180	3
5	17	.	1
4	.	110	2

- Syntax: `merge 1:1 id using "using file name"`
- Joins corresponding observations from master and using datasets matching on the key variable(s)
- Mater data are *invoilable*, i.e. if there exists a variable in master, the values are not replaced
- By default merger creates a new variable, `_merge`, which indicates:
  - 1 (master) – this obs is from master only
  - 2 (using) – this obs is from using only
  - 3 (match) – this obs is from both master and using datasets

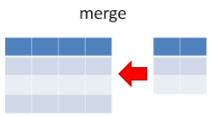
Break

Any questions?



Inputting raw data

- Stata stores data in a proprietary format - the `.dta` file
- Once data are stored in a `.dta` file, we can quickly load the data into memory by the `use` command
- If data are given in other formats, we have to input / read / import them into stata first
- One common such format is known as a raw data file, which stata assumes to have an extension of `.raw`



`infile` Example

```
infile str14 country setting effort change using "test.raw", clear
list in 1/3
```

```
##. infile str14 country setting effort change using "test.raw", clear
##(20 observations read)
##. list in 1/3
##     +-------------------------------------+
##     | country   setting   effort   change |
##     |-------------------------------------|
##  1. | Bolivia        46        0        1 |
##  2. |  Brazil        74        0       10 |
##  3. |   Chile        89       16       29 |
##     +-------------------------------------+
```

```
Bolivia          46    0     1
Brazil           74    0    10
Chile            89   16    29
Colombia         77   16    25
CostaRica        84   21    29
Cuba             89   15    40
DominicanRep     68   14    21
Ecuador          70    6     0
ElSalvador       60   13    13
Guatemala        55    9     4
Haiti            35    3     0
Honduras         51    7     7
Jamaica          87   23    21
Mexico           83    4     9
Nicaragua        68    0     7
Panama           84   19    22
Paraguay         74    3     6
Peru             73    0     2
TrinidadTobago   84   15    29
Venezuela        91    7    11
```

Free format raw data

- values are separated by space, comma or tab
- string value is quoted if imbeds space or comma
- if one observation per line then consider using `insheet` command

Fixed Column Format

- `test.raw` can also be read as fixed-column format since the values of each variable appear in fixed columns, for example :
  - country names are always in columns 4-17
  - settings values are always in columns 23 and 24
- This information can be stored in a separate *dictionary file*:
- `test.dct`

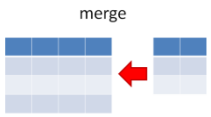
```
dictionary using test.raw {
   _column(4)  str14 country  %14s "country name"
   _column(23) int   settings %2.0f "settings"
   _column(31) int   effort   %2.0f "effort"
   _column(40) int   change   %2.0f "change"
}
```

- Using the dictionary file, the data can be read into Stata like this:

```
infile using test.dct, clear
```

Import/Export Pointers

- Stat/Transfer can import/export data from and to various formats
- But don’t blindly trust any software that translates data from one system / application to another
- Be careful and double-check everything
- Ask for help



Thank you