Introduction to stata

Moses Ngari

mngari@kemri-wellcome.org

Outline

Introduction to stata

- Stata window
- Log and do-files
- Reading, importing and exporting data
- Saving data

Data management

- Creating & renaming variables
- Labelling variables and values
- Merging & appending datasets

Outline2

Exploring data

- Tabulate and table function
- Summarize function
- “if” and bysort function
- Graphing

Descriptive statistics

- Means, variance and SD
- Median, IQR, min, max
- Tabulate
- Proportions and 95% CI
- Means and 95% CI

Outline3

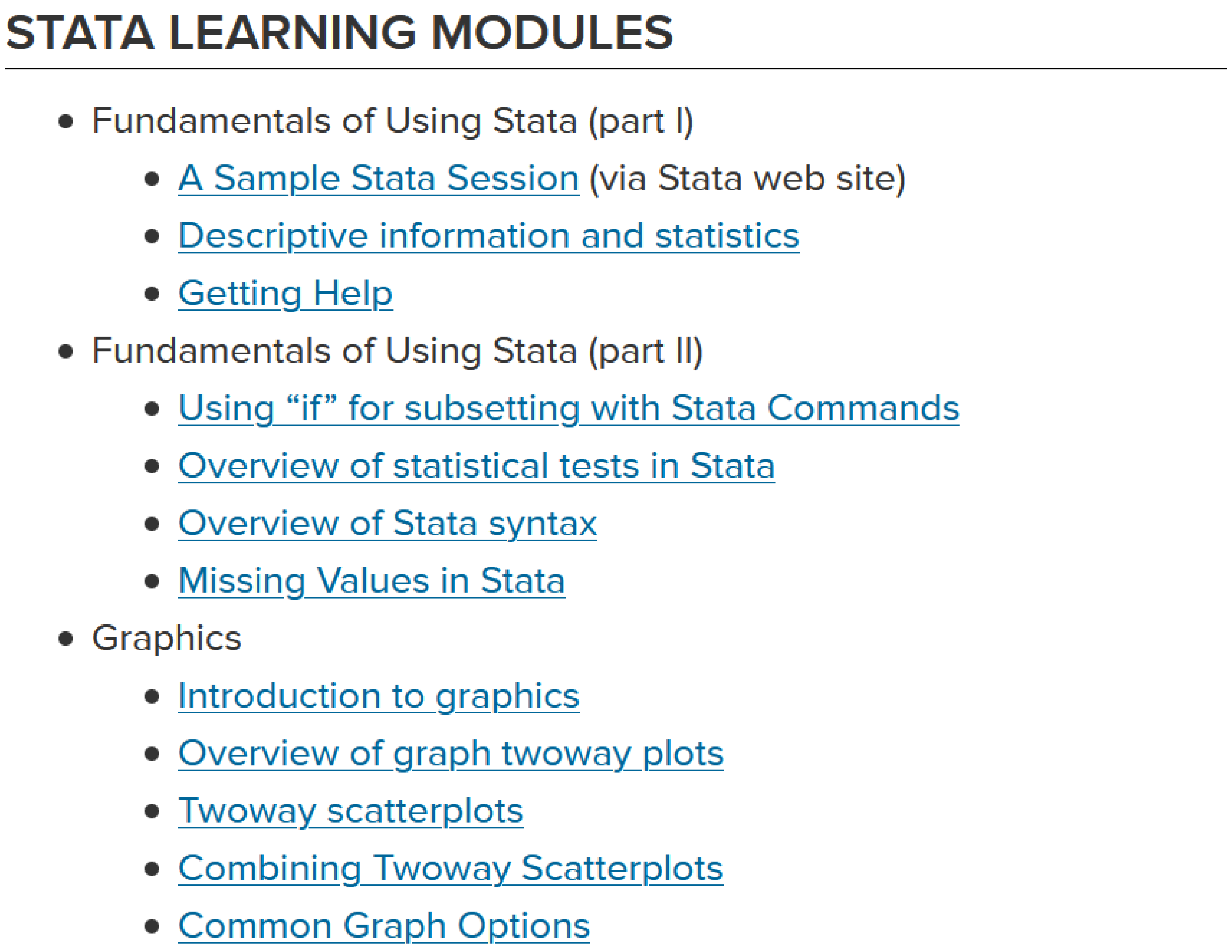
Inferential statistics

- Test of means (z-test/t-test)
- ANOVA
- Chi-square test/Fisher’s exact
- Proportions comparision

Regression modelling

- Linear regression
- Logistic regression

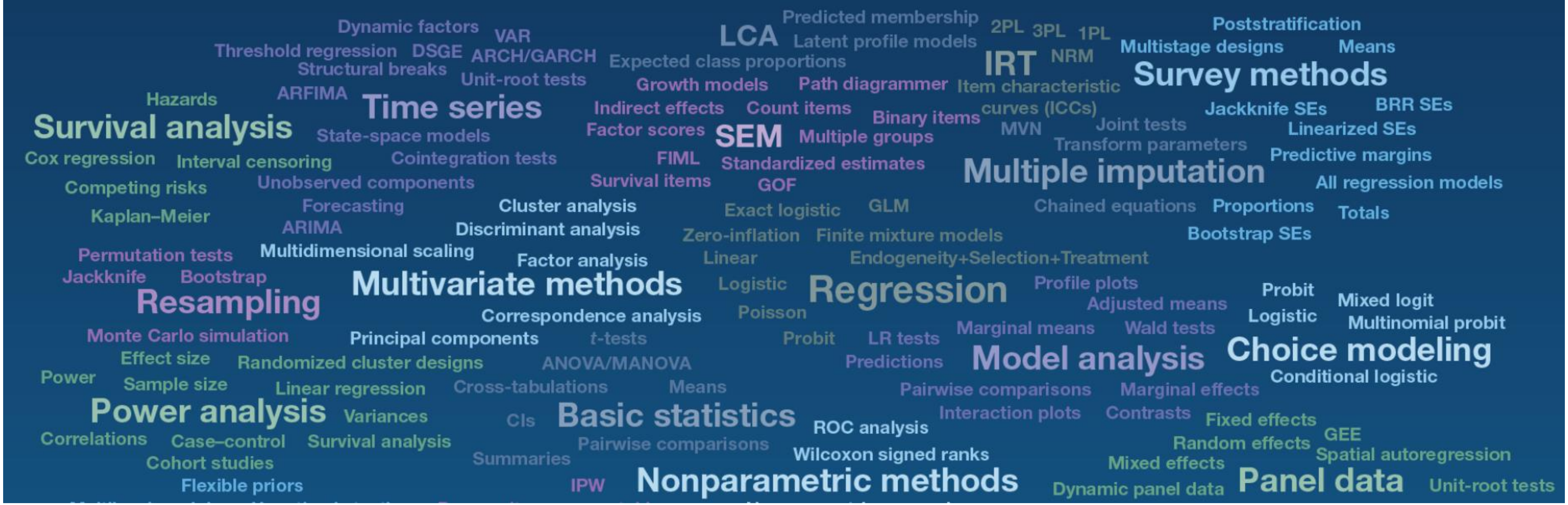
https://stats.idre.ucla.edu/stata/modules/

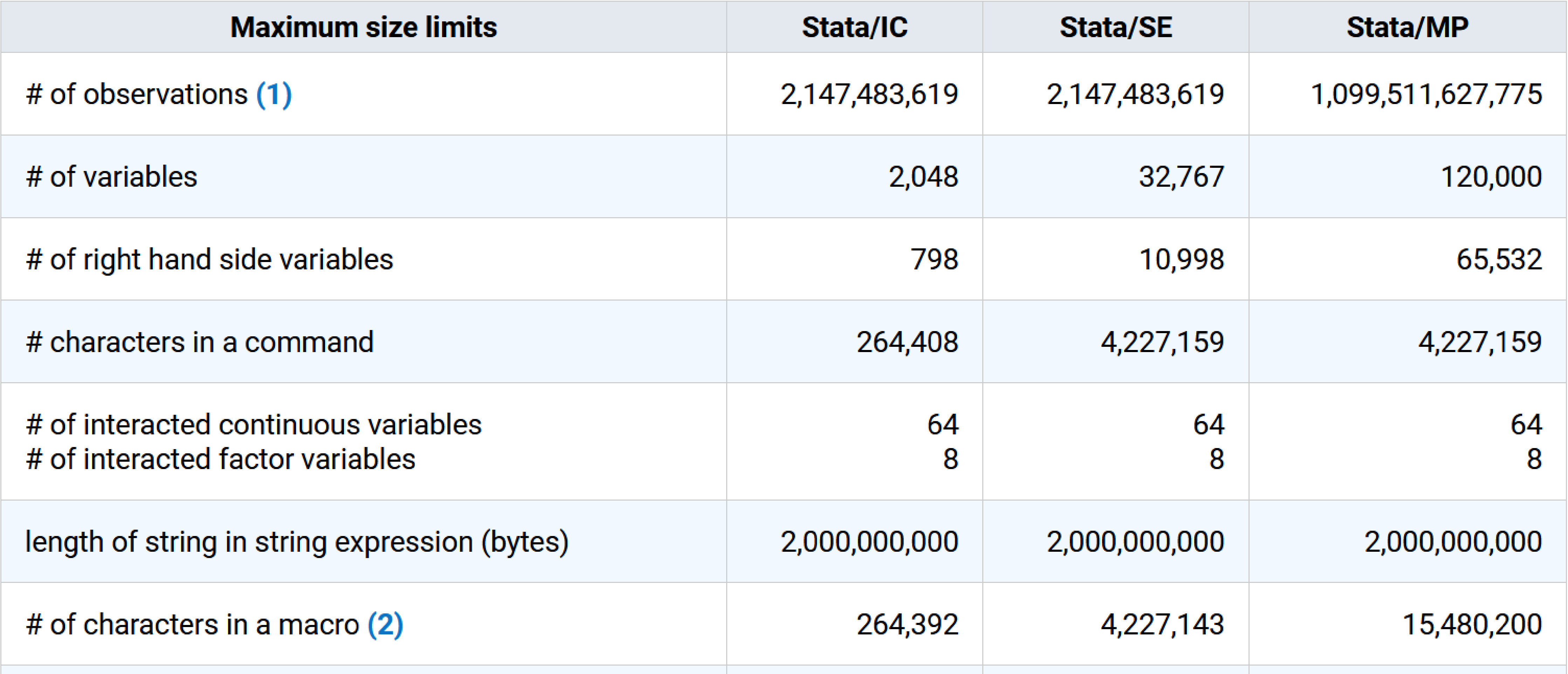


STATA LEARNING MODULES

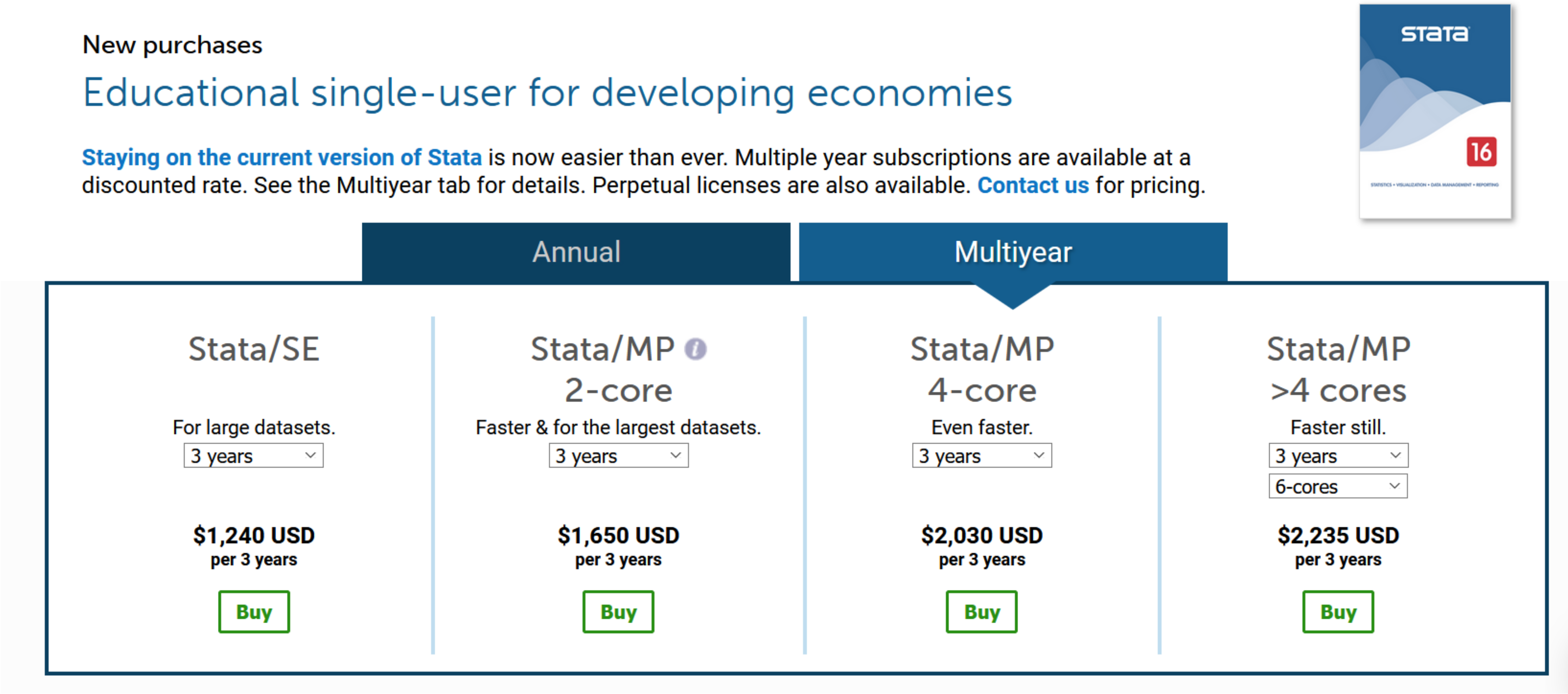
- Fundamentals of Using Stata (part I)
  - A Sample Stata Session (via Stata web site)
  - Descriptive information and statistics
  - Getting Help
- Fundamentals of Using Stata (part II)
  - Using “if” for subsetting with Stata Commands
  - Overview of statistical tests in Stata
  - Overview of Stata syntax
  - Missing Values in Stata
- Graphics
  - Introduction to graphics
  - Overview of graph twoway plots
  - Twoway scatterplots
  - Combining Twoway Scatterplots
  - Common Graph Options

Broad suite of statistical features





Maximum size limits	Stata/IC	Stata/SE	Stata/MP
# of observations (1)	2,147,483,619	2,147,483,619	1,099,511,627,775
# of variables	2,048	32,767	120,000
# of right hand side variables	798	10,998	65,532
# characters in a command	264,408	4,227,159	4,227,159
# of interacted continuous variables # of interacted factor variables	64 8	64 8	64 8
length of string in string expression (bytes)	2,000,000,000	2,000,000,000	2,000,000,000
# of characters in a macro (2)	264,392	4,227,143	15,480,200



New purchases

Educational single-user for developing economies

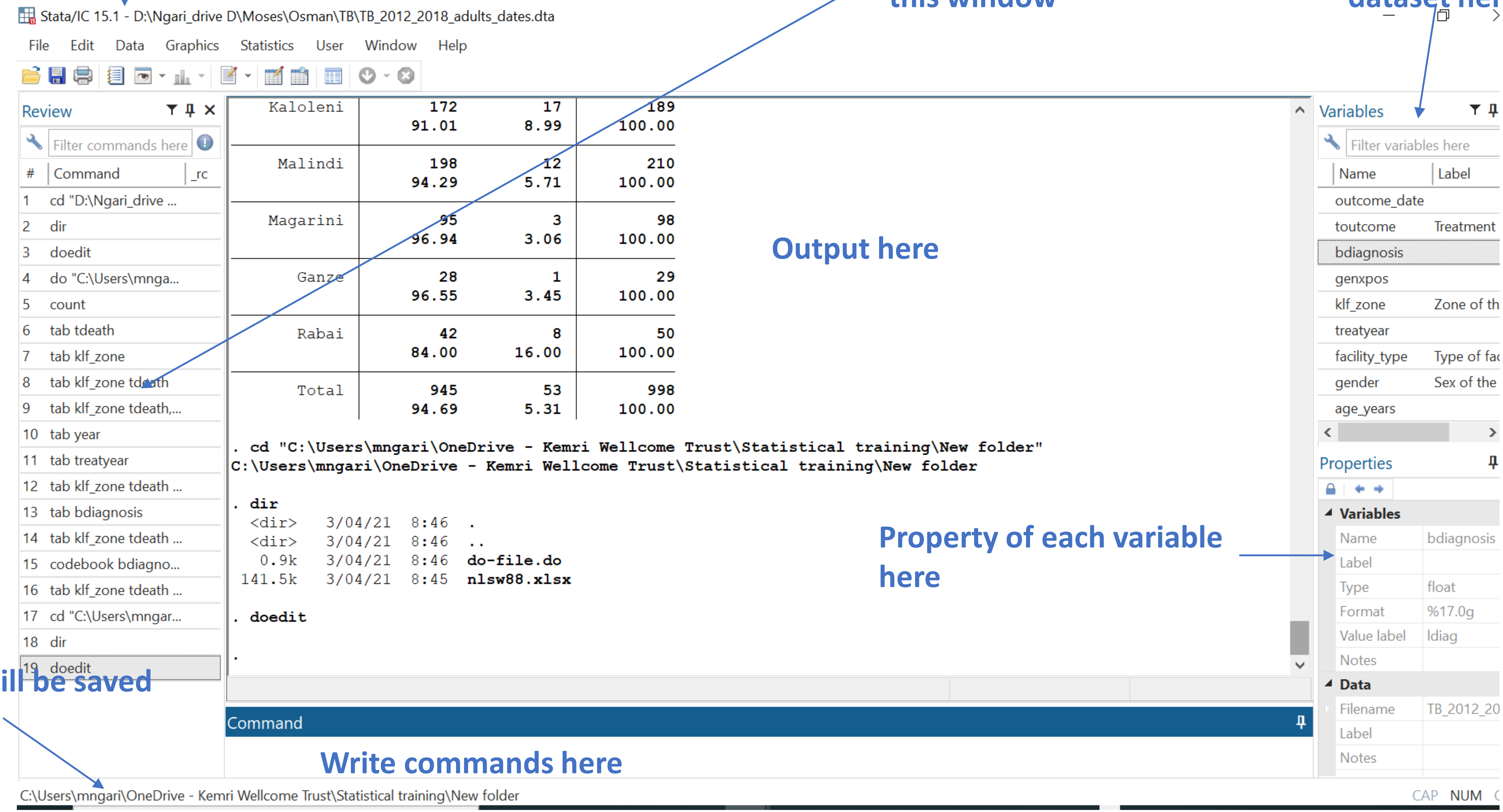
Staying on the current version of Stata is now easier than ever. Multiple year subscriptions are available at a discounted rate. See the Multiyear tab for details. Perpetual licenses are also available. **Contact us** for pricing.

Stata/SE	Stata/MP 2-core	Stata/MP 4-core	Stata/MP >4 cores
For large datasets.	Faster & for the largest datasets.	Even faster.	Faster still.
3 years	3 years	3 years	3 years 6-cores
$1,240 USD **per 3 years**	**$1,650 USD** **per 3 years**	**$2,030 USD** **per 3 years**	**$2,235 USD** **per 3 years**
Buy	Buy	Buy	Buy

Name of dataset

History of commands, this window

Variables in dataset here



#	Command	_rc
1	cd "D:\Ngari_drive ...	
2	dir	
3	doedit	
4	do "C:\Users\mnga...	
5	count	
6	tab tdeath	
7	tab klf_zone	
8	tab klf_zone tdeath	
9	tab klf_zone tdeath,...	
10	tab year	
11	tab treatyear	
12	tab klf_zone tdeath ...	
13	tab bdiagnosis	
14	tab klf_zone tdeath ...	
15	codebook bdiagno...	
16	tab klf_zone tdeath ...	
17	cd "C:\Users\mngar...	
18	dir	
19	doedit	

```
   Kaloleni |       172          17 |       189
            |     91.01        8.99 |    100.00
------------+-----------------------+----------
    Malindi |       198          12 |       210
            |     94.29        5.71 |    100.00
------------+-----------------------+----------
   Magarini |        95           3 |        98
            |     96.94        3.06 |    100.00
------------+-----------------------+----------
      Ganze |        28           1 |        29
            |     96.55        3.45 |    100.00
------------+-----------------------+----------
      Rabai |        42           8 |        50
            |     84.00       16.00 |    100.00
------------+-----------------------+----------
      Total |       945          53 |       998
            |     94.69        5.31 |    100.00

. cd "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder"
C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder

. dir
  <dir>   3/04/21  8:46  .
  <dir>   3/04/21  8:46  ..
   0.9k   3/04/21  8:46  do-file.do
 141.5k   3/04/21  8:45  nlsw88.xlsx

. doedit

.
```

Name	Label
outcome_date	
toutcome	Treatment
bdiagnosis	
genxpos	
klf_zone	Zone of th
treatyear	
facility_type	Type of fac
gender	Sex of the
age_years	

Variables	
Name	bdiagnosis
Label	
Type	float
Format	%17.0g
Value label	ldiag
Notes	
Data	
Filename	TB_2012_20
Label	
Notes	

C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder

Output here

Property of each variable here

Files will be saved here

Write commands here

To see your working directory, type

```
pwd
```

To change the working directory to avoid typing the whole path when calling or saving files,

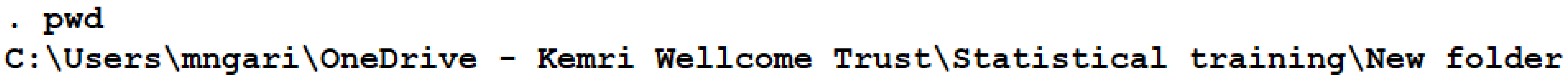
type: `cd`

Use quotes if the new directory has blank spaces, for example

```
cd “h:\stata and data” h:\stata and data. cd "h:\stata and da
```

To see your working directory, type

`pwd`

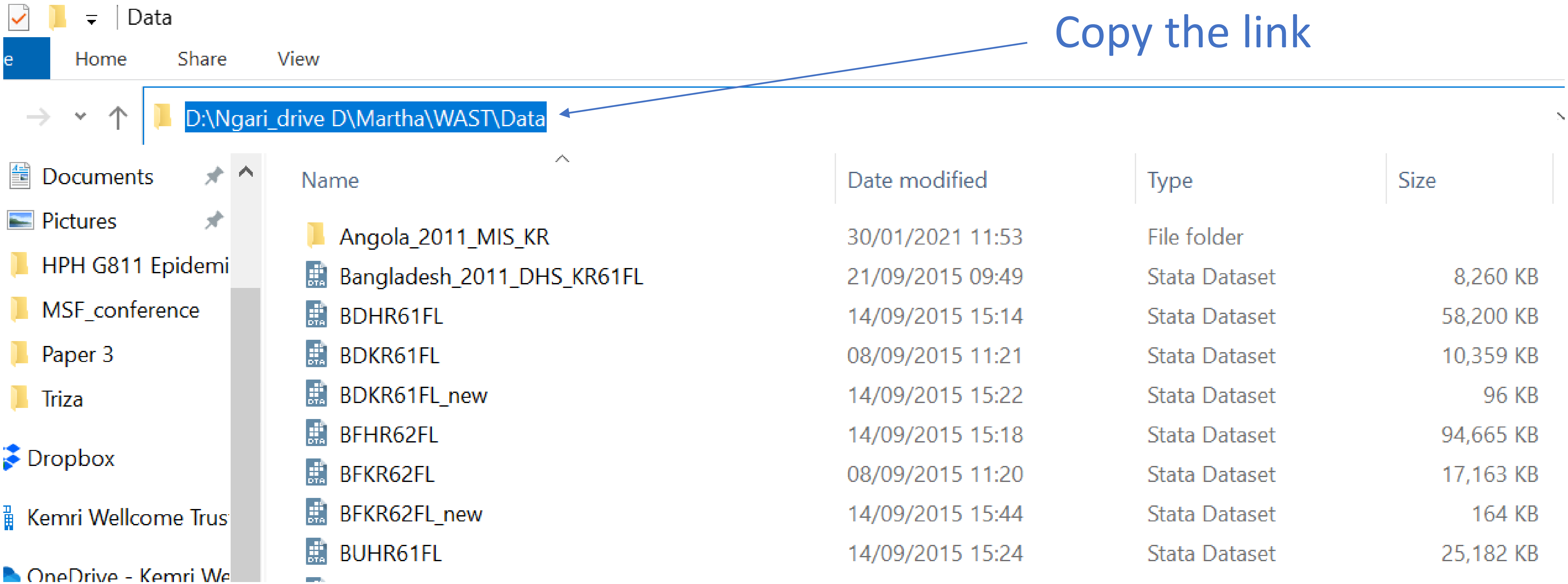


```
. pwd
C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder
```

To change the working directory to avoid typing the whole path when calling or saving files,

type: `cd`

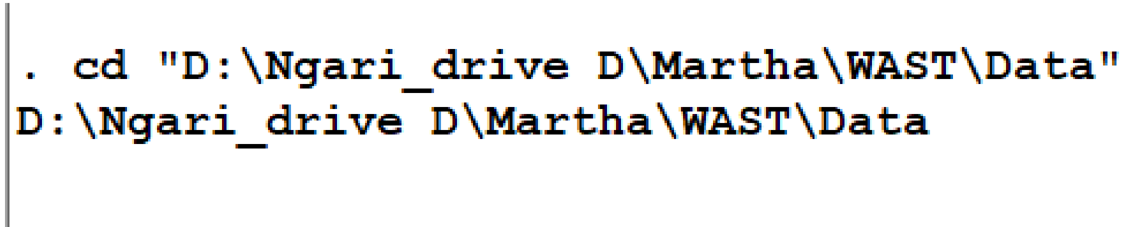
Use quotes if the new directory has blank spaces



D:\Ngari_drive D\Martha\WAST\Data

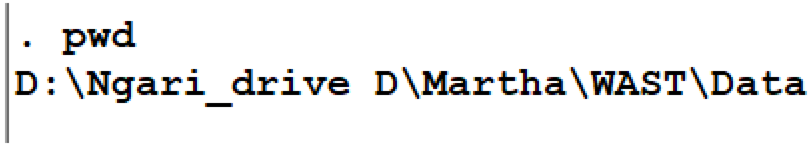
Name	Date modified	Type	Size
Angola_2011_MIS_KR	30/01/2021 11:53	File folder	
Bangladesh_2011_DHS_KR61FL	21/09/2015 09:49	Stata Dataset	8,260 KB
BDHR61FL	14/09/2015 15:14	Stata Dataset	58,200 KB
BDKR61FL	08/09/2015 11:21	Stata Dataset	10,359 KB
BDKR61FL_new	14/09/2015 15:22	Stata Dataset	96 KB
BFHR62FL	14/09/2015 15:18	Stata Dataset	94,665 KB
BFKR62FL	08/09/2015 11:20	Stata Dataset	17,163 KB
BFKR62FL_new	14/09/2015 15:44	Stata Dataset	164 KB
BUHR61FL	14/09/2015 15:24	Stata Dataset	25,182 KB

Copy the link

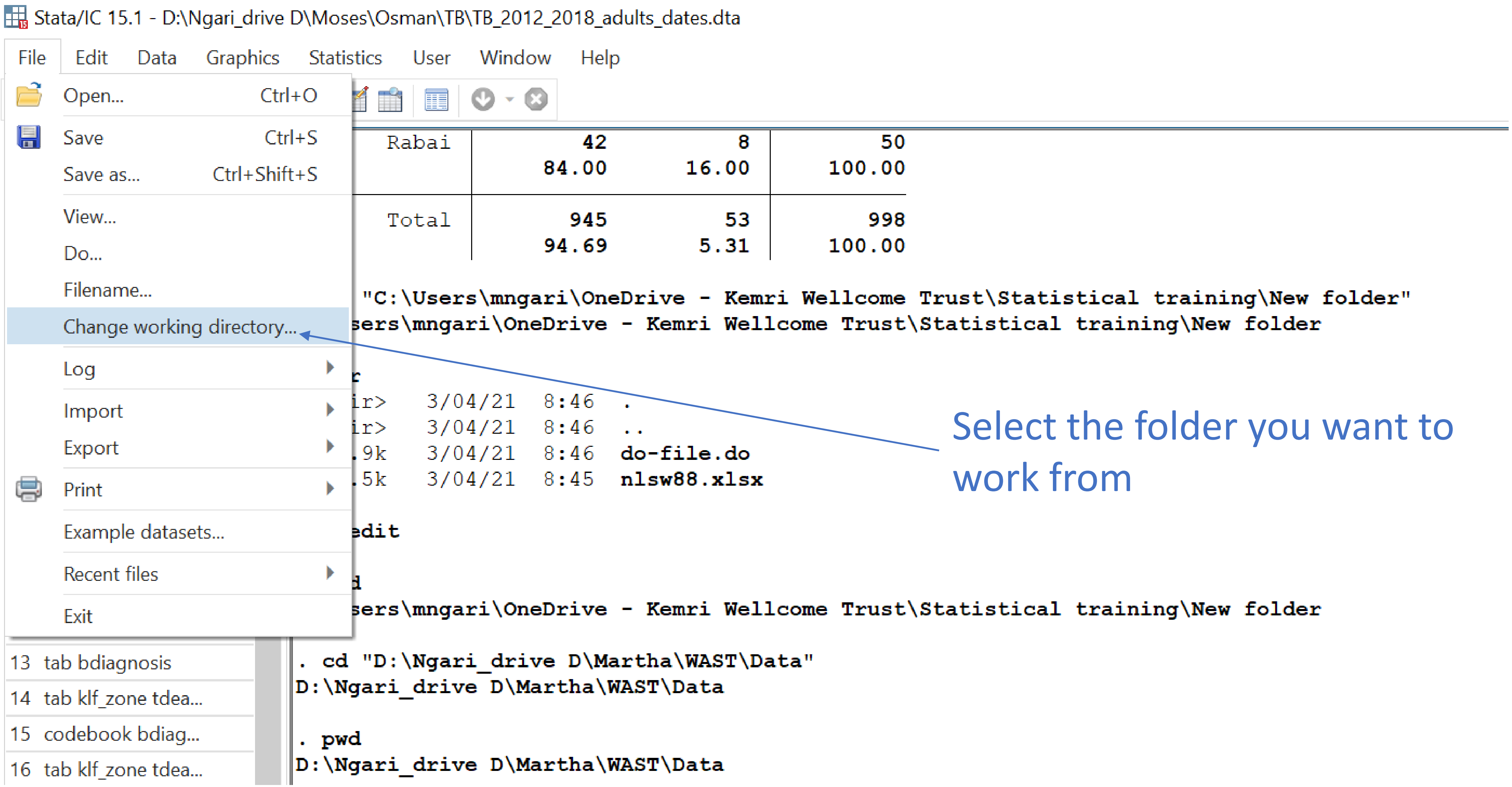


```
. cd "D:\Ngari_drive D\Martha\WAST\Data"
D:\Ngari_drive D\Martha\WAST\Data
```

Type this



```
. pwd
D:\Ngari_drive D\Martha\WAST\Data
```



```
      Rabai |        42          8 |        50
            |     84.00      16.00 |    100.00
------------+----------------------+----------
      Total |       945         53 |       998
            |     94.69       5.31 |    100.00

 "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder"
sers\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder

r
ir>   3/04/21  8:46  .
ir>   3/04/21  8:46  ..
.9k   3/04/21  8:46  do-file.do
.5k   3/04/21  8:45  nlsw88.xlsx

edit

d
sers\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder

. cd "D:\Ngari_drive D\Martha\WAST\Data"
D:\Ngari_drive D\Martha\WAST\Data

. pwd
D:\Ngari_drive D\Martha\WAST\Data
```

```
13  tab bdiagnosis
14  tab klf_zone tdea...
15  codebook bdiag...
16  tab klf_zone tdea...
```

Select the folder you want to work from

Create a ***log file***, sort of Stata's built-in tape recorder and where you can: 1) retrieve the output of your work and 2) keep a record of your work.

In the command line type:

```
log using mylog.log
```

This will create the file 'mylog.log' in your working directory. You can read it using any word processor (notepad, word, etc.).

To close a log file type:

```
log close
```

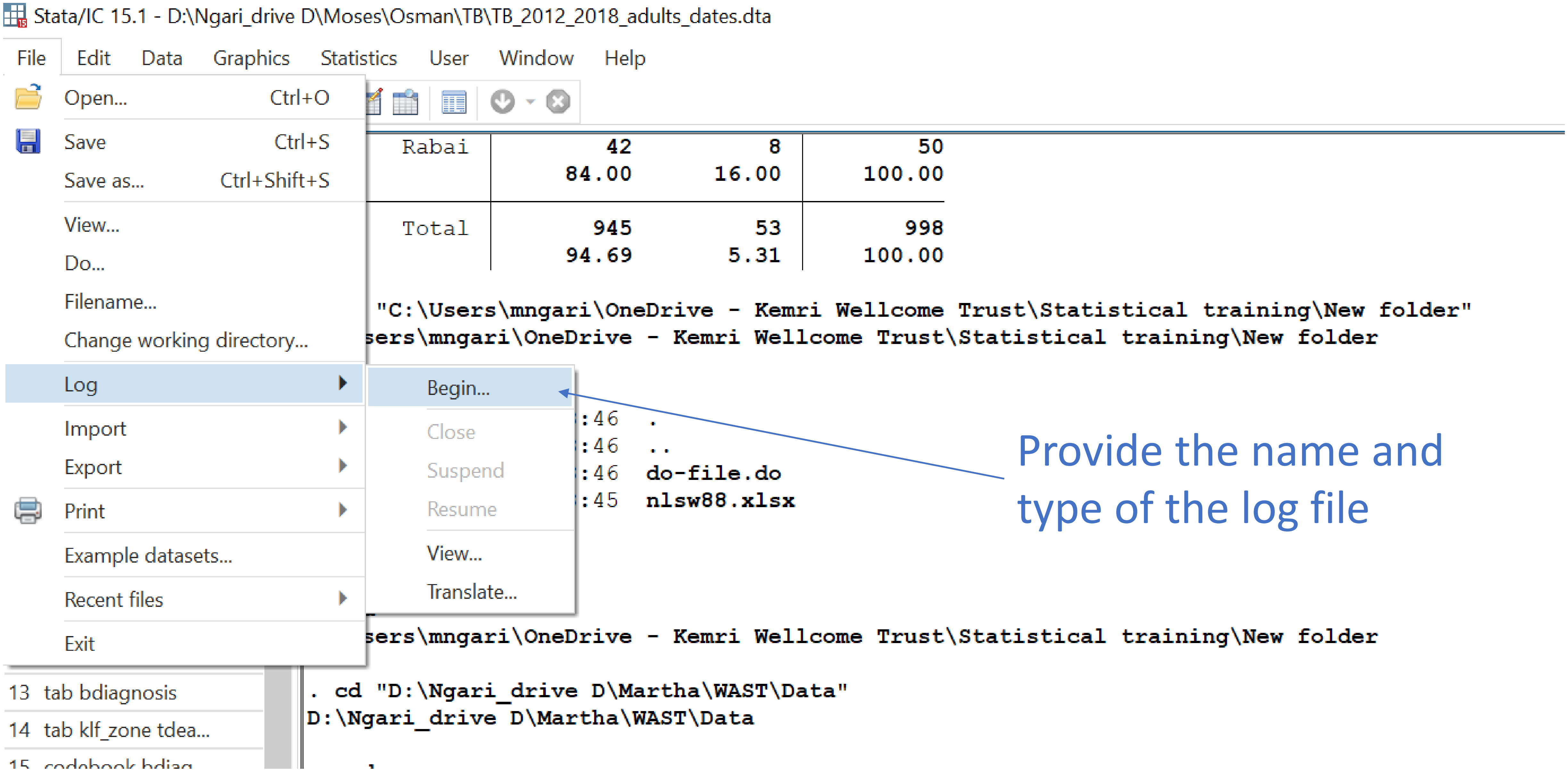
To add more output to an existing log file add the option append, type:

```
log using mylog.log, append
```

To replace a log file add the option replace, type:

```
log using mylog.log, replace
```

Note that the option replace will delete the contents of the previous version of the log.



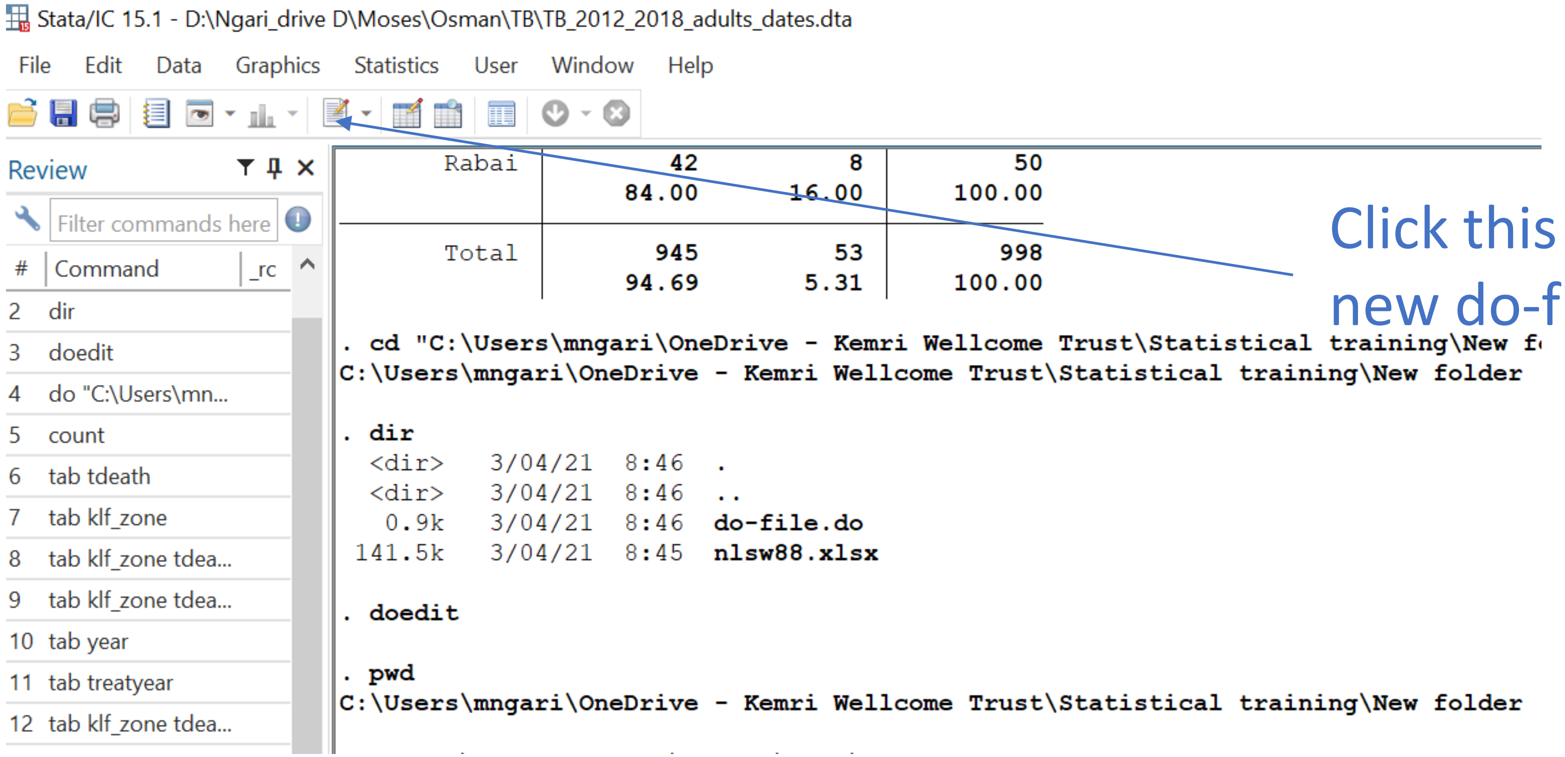
```
    Rabai   |        42          8 |        50 
            |     84.00      16.00 |    100.00 
------------+----------------------+----------
    Total   |       945         53 |       998 
            |     94.69       5.31 |    100.00 

 "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder"
sers\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder

:46  .
:46  ..
:46  do-file.do
:45  nlsw88.xlsx

sers\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder

. cd "D:\Ngari_drive D\Martha\WAST\Data"
D:\Ngari_drive D\Martha\WAST\Data
```

```
13  tab bdiagnosis
14  tab klf_zone tdea...
15  codebook bdiag
```

Provide the name and type of the log file

Create do-file

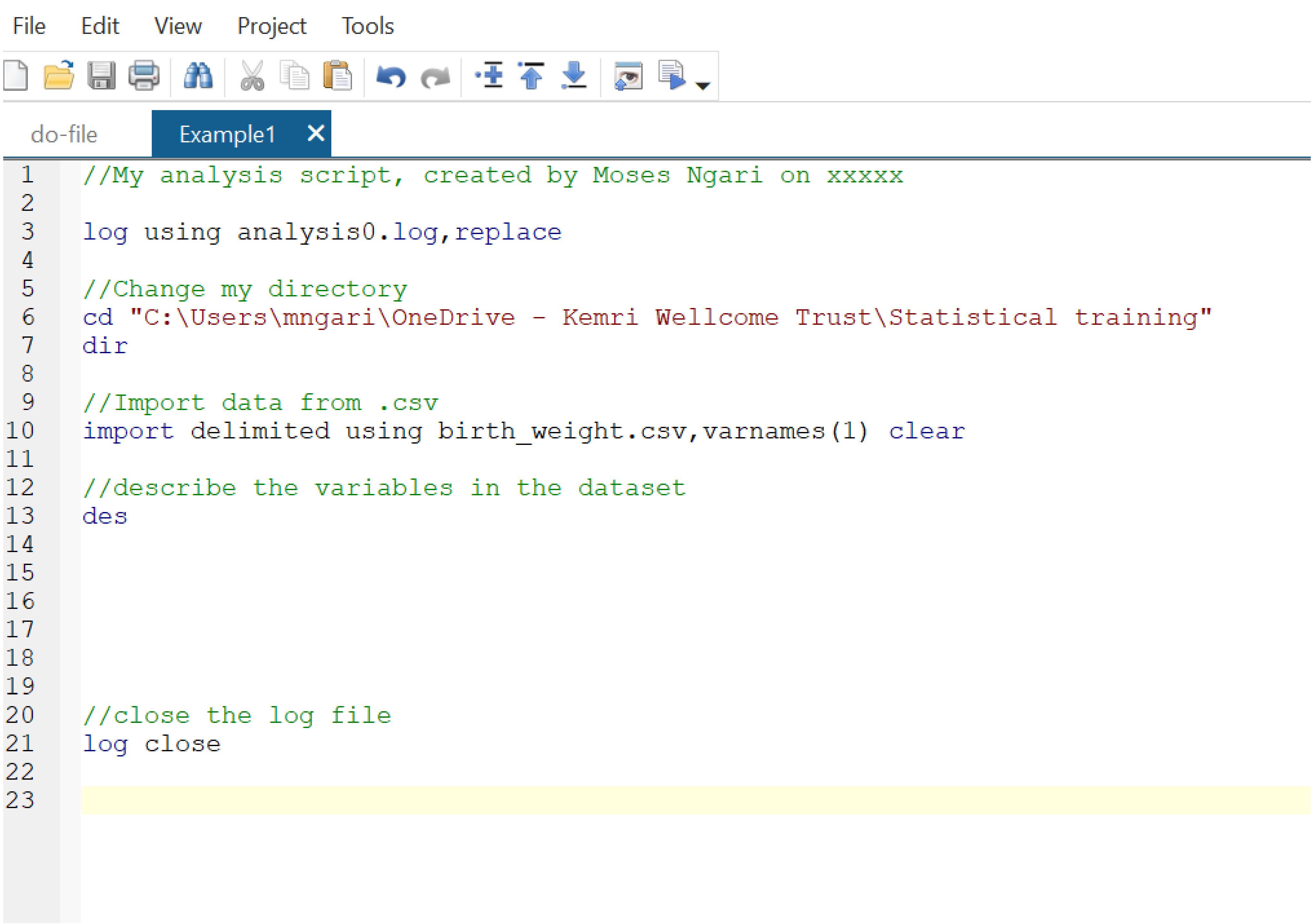
- Do-files are ASCII files that contain of Stata commands to run specific procedures. It is highly recommended to use do-files to store your commands so do you not have to type them again should you need to re-do your work.
- You can use any word processor and save the file in ASCII format, or you can use Stata's 'do-file editor' with the advantage that you can run the commands from there. Either , in the command window type:
- *doedit*


Stata/IC 15.1 - D:\Ngari_drive D\Moses\Osman\TB\TB_2012_2018_adults_dates.dta
File
Edit
Data
Graphics
Statistics
User
Window
Help
Review
Filter commands here

```
#   Command            _rc
2   dir
3   doedit
4   do "C:\Users\mn...
5   count
6   tab tdeath
7   tab klf_zone
8   tab klf_zone tdea...
9   tab klf_zone tdea...
10  tab year
11  tab treatyear
12  tab klf_zone tdea...
```

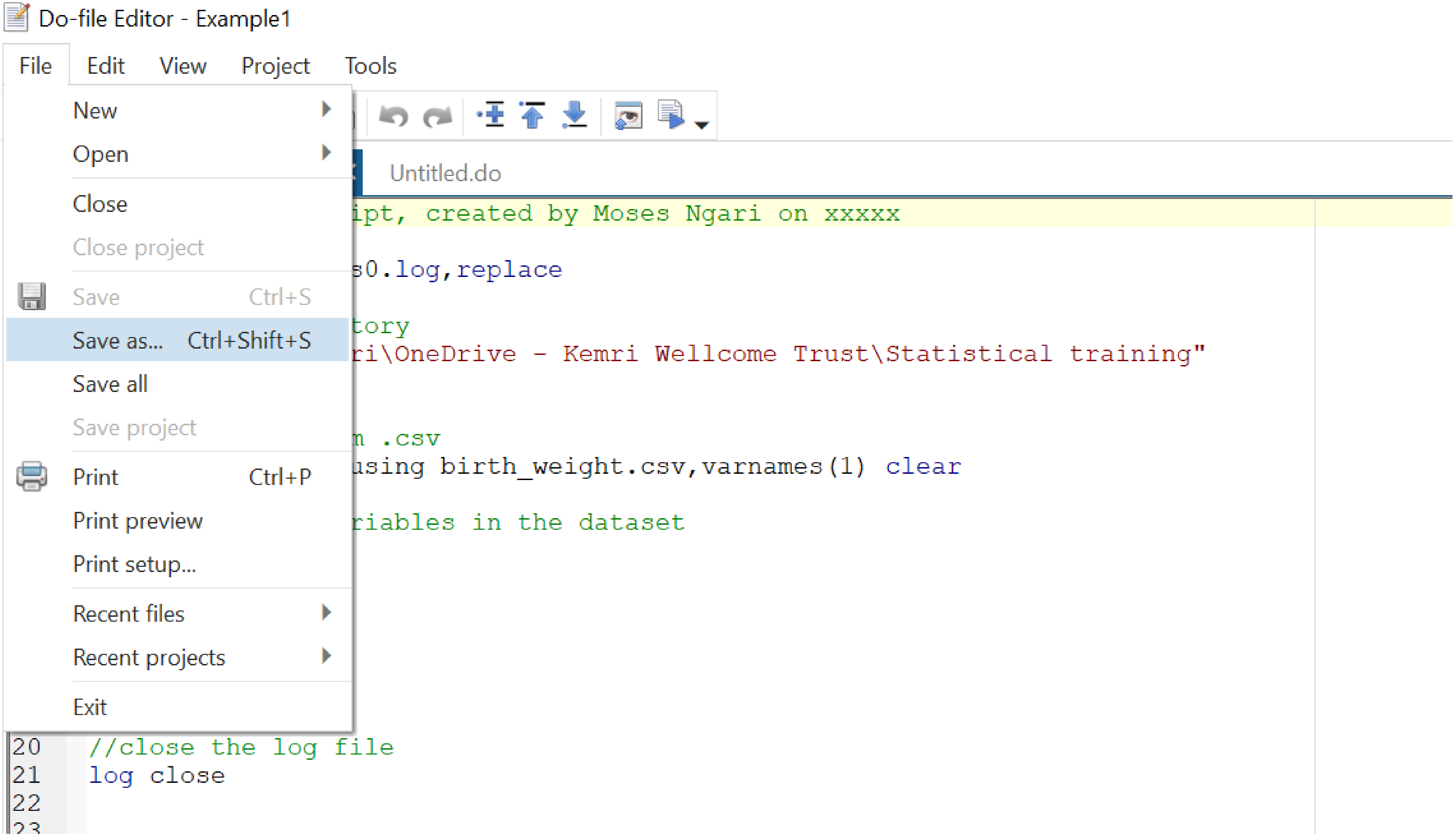
```
      Rabai |        42          8 |        50
            |     84.00      16.00 |    100.00
------------+----------------------+----------
      Total |       945         53 |       998
            |     94.69       5.31 |    100.00

. cd "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New f
C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder

. dir
  <dir>   3/04/21  8:46  .
  <dir>   3/04/21  8:46  ..
   0.9k   3/04/21  8:46  do-file.do
 141.5k   3/04/21  8:45  nlsw88.xlsx

. doedit

. pwd
C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\New folder
```

Click this icon to start new do-file



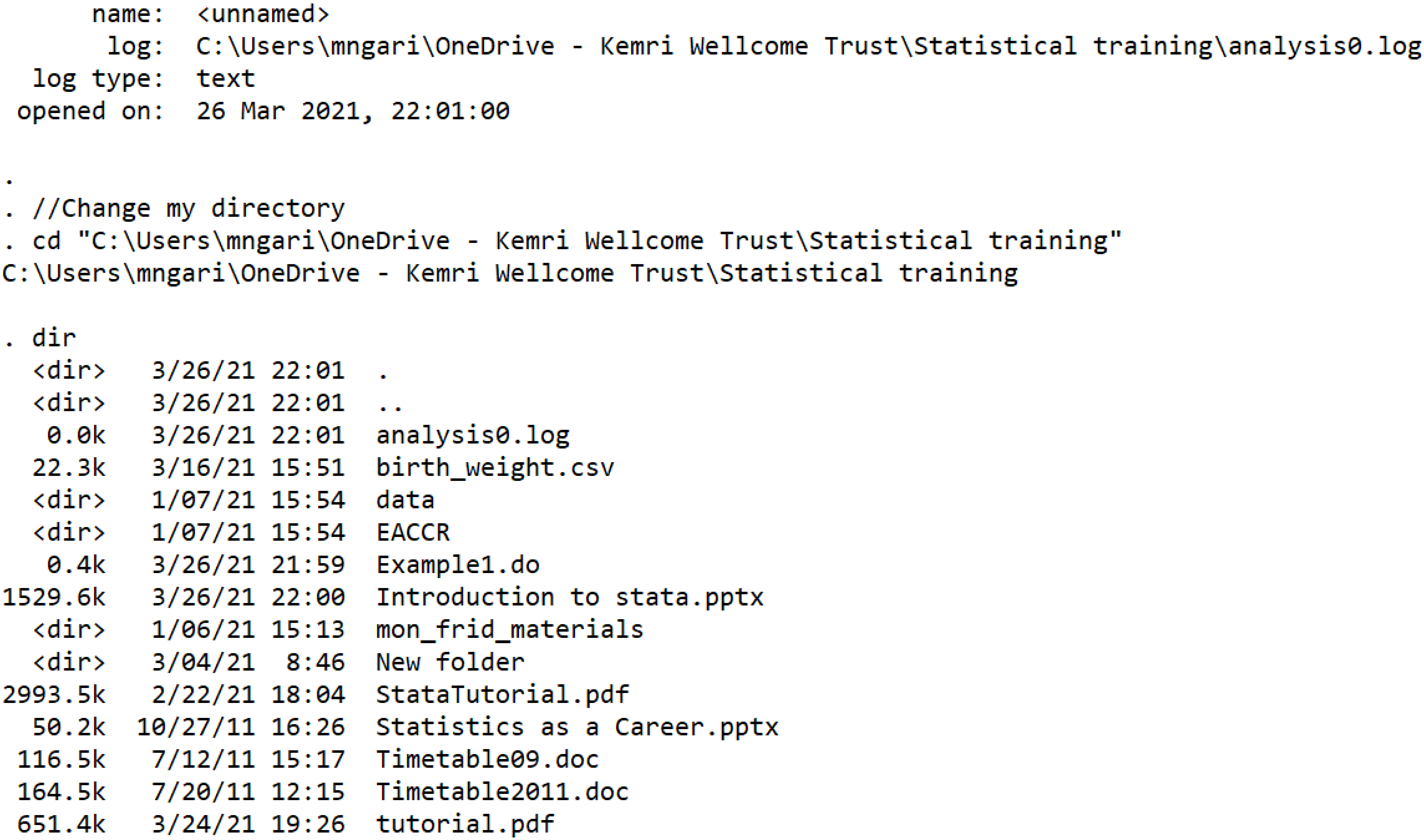
```
//My analysis script, created by Moses Ngari on xxxxx

log using analysis0.log,replace

//Change my directory
cd "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training"
dir

//Import data from .csv
import delimited using birth_weight.csv,varnames(1) clear

//describe the variables in the dataset
des






//close the log file
log close


```

Save the do-file

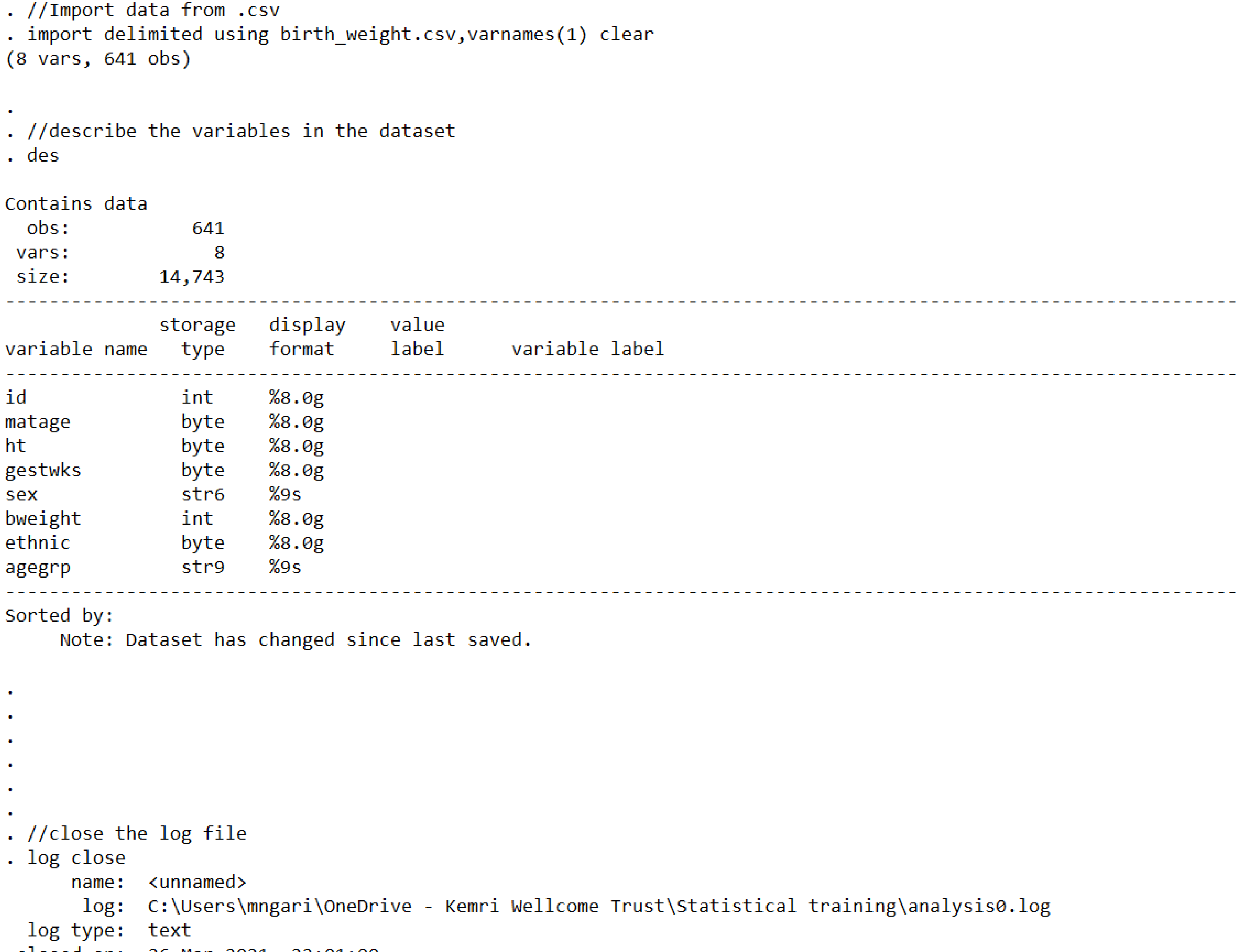


```
ipt, created by Moses Ngari on xxxxx

s0.log,replace

tory
ri\OneDrive - Kemri Wellcome Trust\Statistical training"


n .csv
using birth_weight.csv,varnames(1) clear

riables in the dataset
```

```
//close the log file
log close
```



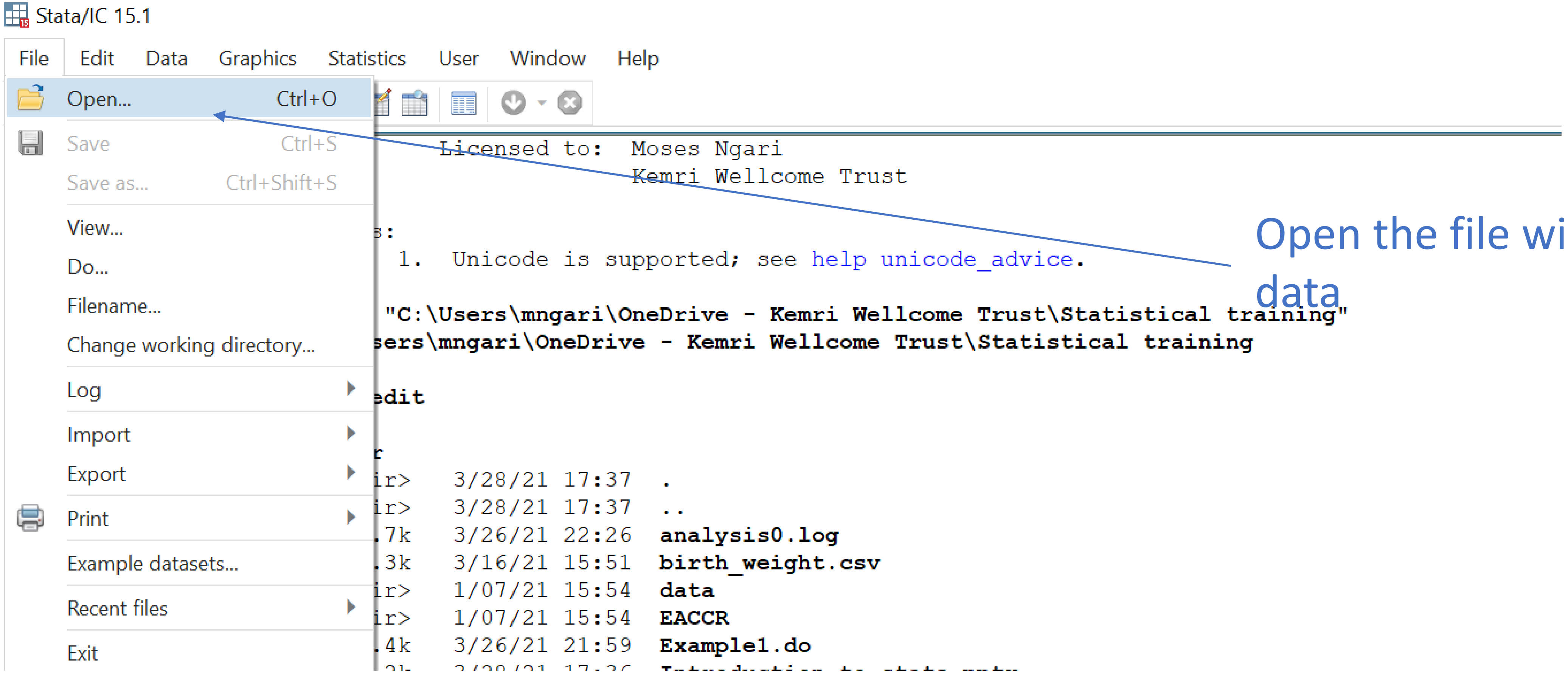
```
      name:  <unnamed>
       log:  C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\analysis0.log
  log type:  text
 opened on:  26 Mar 2021, 22:01:00

. 
. //Change my directory
. cd "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training"
C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training

. dir
  <dir>   3/26/21 22:01  .
  <dir>   3/26/21 22:01  ..
   0.0k   3/26/21 22:01  analysis0.log
  22.3k   3/16/21 15:51  birth_weight.csv
  <dir>   1/07/21 15:54  data
  <dir>   1/07/21 15:54  EACCR
   0.4k   3/26/21 21:59  Example1.do
1529.6k   3/26/21 22:00  Introduction to stata.pptx
  <dir>   1/06/21 15:13  mon_frid_materials
  <dir>   3/04/21  8:46  New folder
2993.5k   2/22/21 18:04  StataTutorial.pdf
  50.2k  10/27/11 16:26  Statistics as a Career.pptx
 116.5k   7/12/11 15:17  Timetable09.doc
 164.5k   7/20/11 12:15  Timetable2011.doc
 651.4k   3/24/21 19:26  tutorial.pdf
```



```
. //Import data from .csv
. import delimited using birth_weight.csv,varnames(1) clear
(8 vars, 641 obs)

.
. //describe the variables in the dataset
. des

Contains data
  obs:           641
 vars:             8
 size:        14,743
----------------------------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
----------------------------------------------------------------------------------------------------------------
id              int     %8.0g
matage          byte    %8.0g
ht              byte    %8.0g
gestwks         byte    %8.0g
sex             str6    %9s
bweight         int     %8.0g
ethnic          byte    %8.0g
agegrp          str9    %9s
----------------------------------------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.

.
.
.
.
.
.
. //close the log file
. log close
      name:  <unnamed>
       log:  C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\analysis0.log
  log type:  text
 closed on:  26 Mar 2021, 22:01:09
```

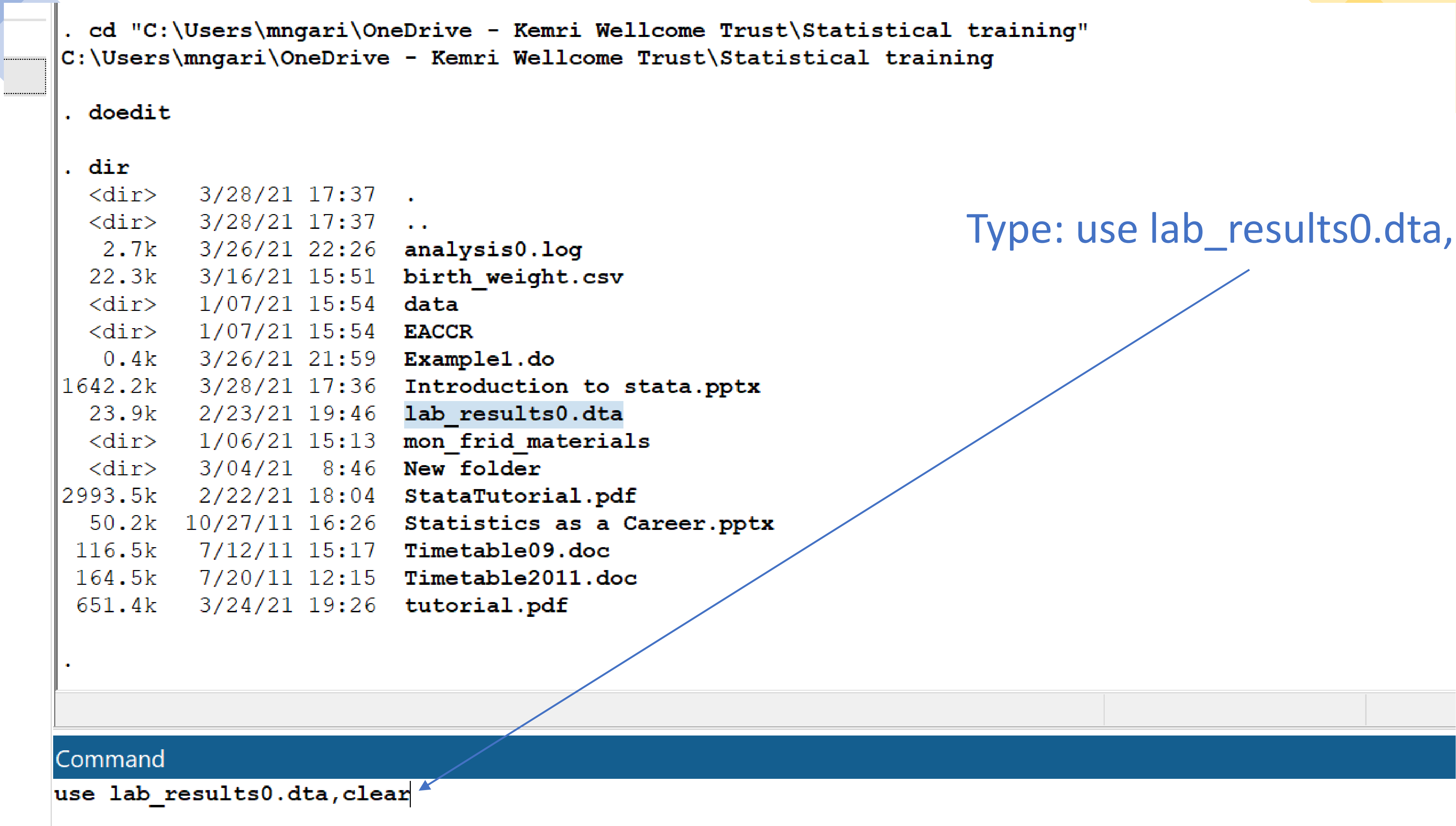
Reading data to stata

- Read data already in stata format
- To open files already in Stata with extension *.dta, run Stata and you can either:
- Type: *use data_name,clear*



```
    Licensed to:  Moses Ngari
                  Kemri Wellcome Trust

s:
 1.  Unicode is supported; see help unicode_advice.

 "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training"
sers\mngari\OneDrive - Kemri Wellcome Trust\Statistical training

edit

r
ir>   3/28/21 17:37  .
ir>   3/28/21 17:37  ..
.7k   3/26/21 22:26  analysis0.log
.3k   3/16/21 15:51  birth_weight.csv
ir>   1/07/21 15:54  data
ir>   1/07/21 15:54  EACCR
.4k   3/26/21 21:59  Example1.do
```

Open the file with stata data

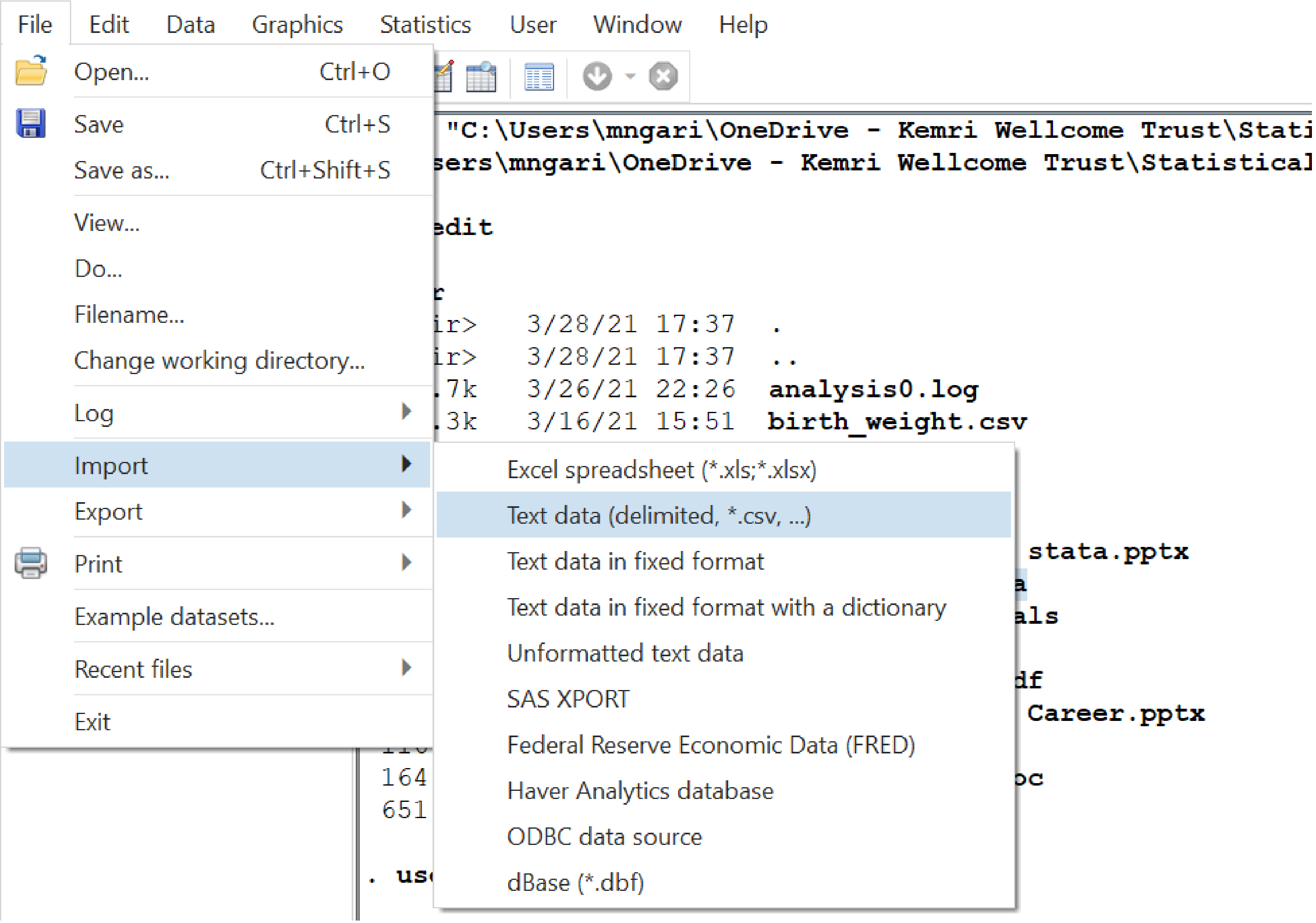


```
. cd "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training"
C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training

. doedit

. dir
  <dir>   3/28/21 17:37  .
  <dir>   3/28/21 17:37  ..
   2.7k   3/26/21 22:26  analysis0.log
  22.3k   3/16/21 15:51  birth_weight.csv
  <dir>   1/07/21 15:54  data
  <dir>   1/07/21 15:54  EACCR
   0.4k   3/26/21 21:59  Example1.do
1642.2k   3/28/21 17:36  Introduction to stata.pptx
  23.9k   2/23/21 19:46  lab_results0.dta
  <dir>   1/06/21 15:13  mon_frid_materials
  <dir>   3/04/21  8:46  New folder
2993.5k   2/22/21 18:04  StataTutorial.pdf
  50.2k  10/27/11 16:26  Statistics as a Career.pptx
 116.5k   7/12/11 15:17  Timetable09.doc
 164.5k   7/20/11 12:15  Timetable2011.doc
 651.4k   3/24/21 19:26  tutorial.pdf

.
```

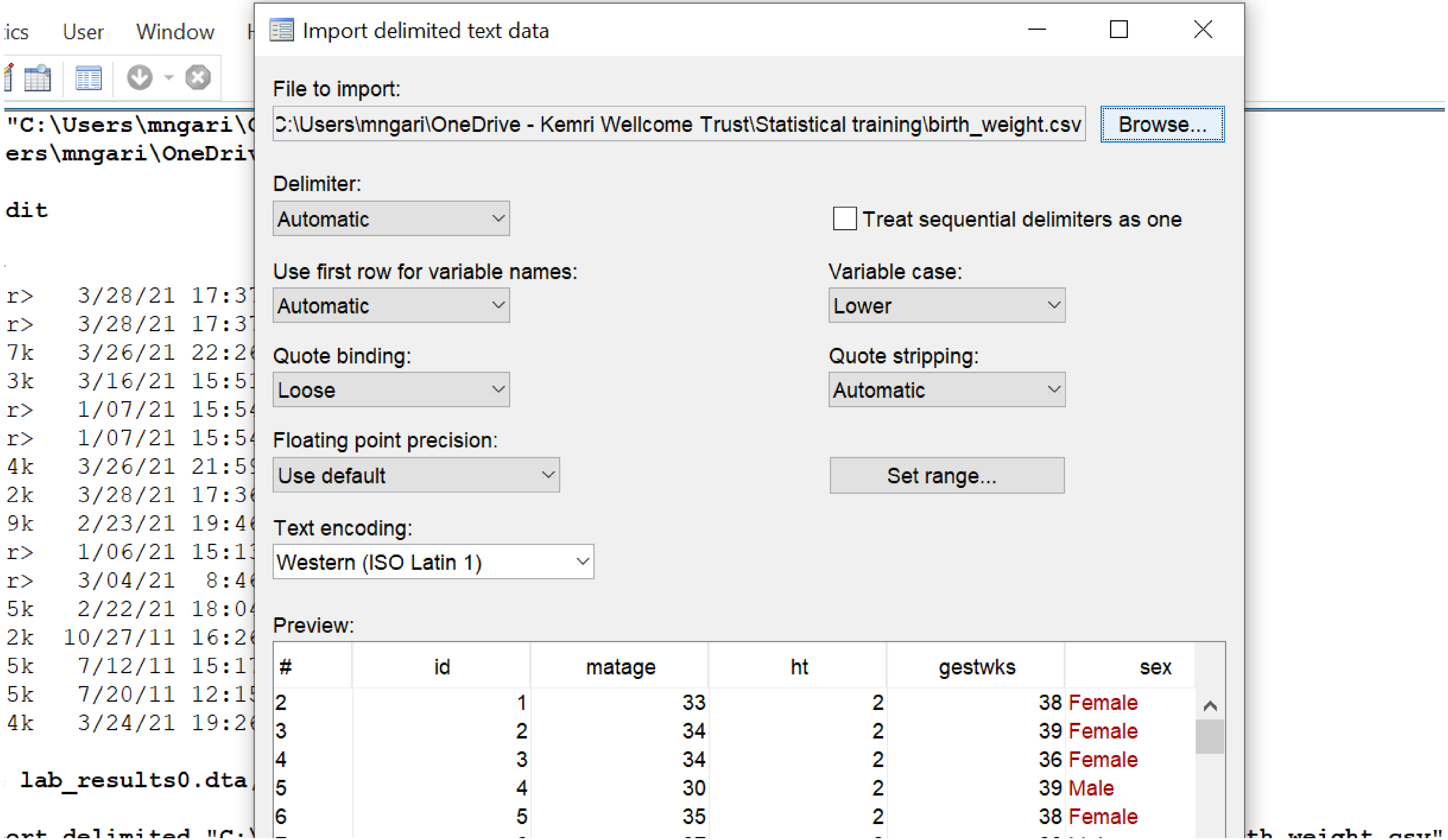
```
use lab_results0.dta,clear
```

Type: use lab_results0.dta,clear

Read data from .csv



```
 "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Stati
sers\mngari\OneDrive - Kemri Wellcome Trust\Statistical

edit

r
ir>    3/28/21 17:37  .
ir>    3/28/21 17:37  ..
.7k    3/26/21 22:26  analysis0.log
.3k    3/16/21 15:51  birth_weight.csv
                                          stata.pptx
                                         a
                                         als
                                         df
                                          Career.pptx
                                         oc
164
651

. us
```



```
"C:\Users\mngari\C
ers\mngari\OneDriv

dit

r>   3/28/21 17:37
r>   3/28/21 17:37
7k   3/26/21 22:26
3k   3/16/21 15:51
r>   1/07/21 15:54
r>   1/07/21 15:54
4k   3/26/21 21:59
2k   3/28/21 17:36
9k   2/23/21 19:46
r>   1/06/21 15:13
r>   3/04/21  8:46
5k   2/22/21 18:04
2k  10/27/11 16:26
5k   7/12/11 15:17
5k   7/20/11 12:15
4k   3/24/21 19:26

 lab_results0.dta,

ort delimited "C:\                                      th_weight.csv"
```

Import delimited text data

File to import: :\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\birth_weight.csv

Delimiter: Automatic

□ Treat sequential delimiters as one

Use first row for variable names: Automatic

Variable case: Lower

Quote binding: Loose

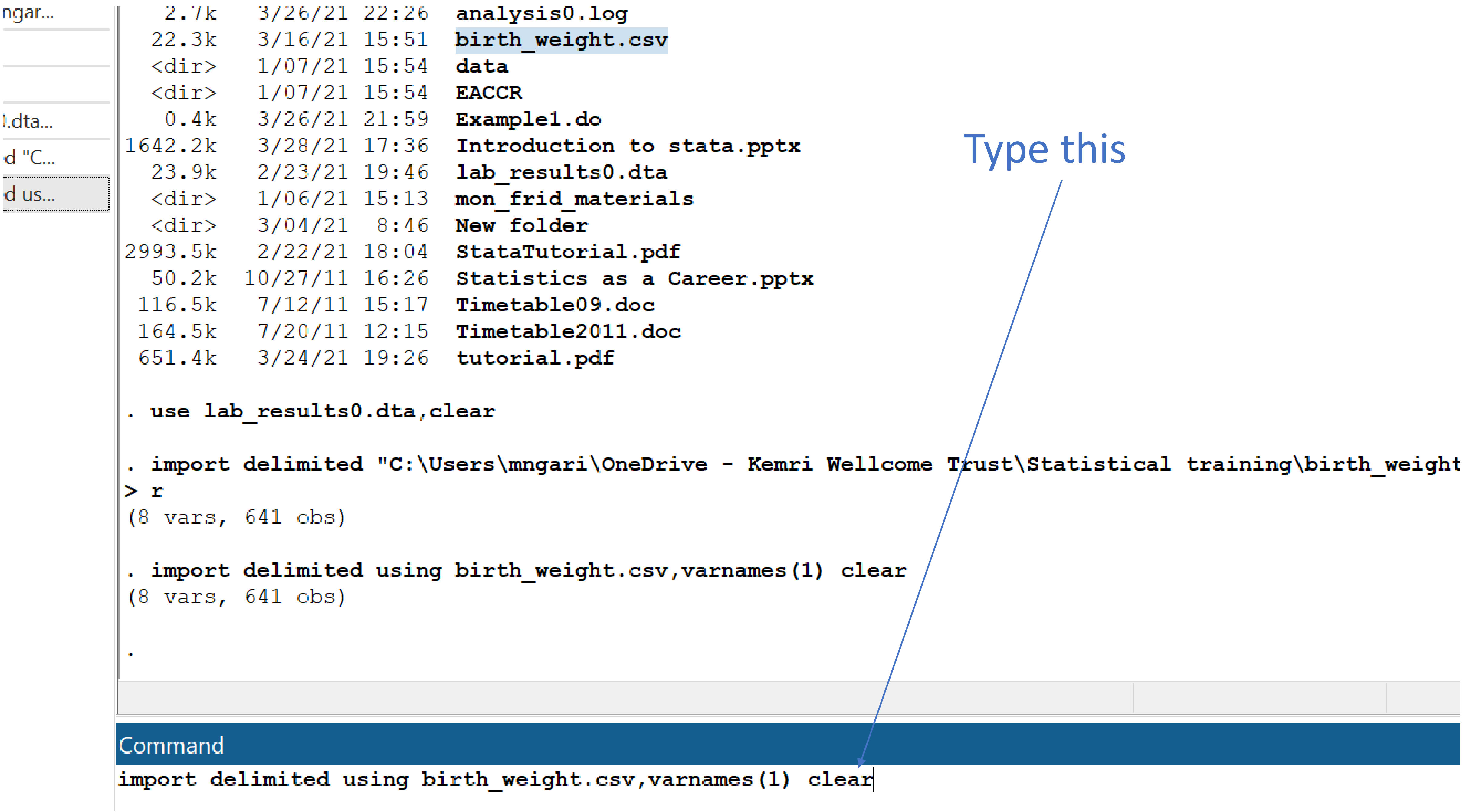
Quote stripping: Automatic

Floating point precision: Use default

Text encoding: Western (ISO Latin 1)

Preview:

#	id	matage	ht	gestwks	sex
2	1	33	2	38	Female
3	2	34	2	39	Female
4	3	34	2	36	Female
5	4	30	2	39	Male
6	5	35	2	38	Female



```
    2.7k   3/26/21 22:26  analysis0.log
   22.3k   3/16/21 15:51  birth_weight.csv
   <dir>   1/07/21 15:54  data
   <dir>   1/07/21 15:54  EACCR
    0.4k   3/26/21 21:59  Example1.do
 1642.2k   3/28/21 17:36  Introduction to stata.pptx
   23.9k   2/23/21 19:46  lab_results0.dta
   <dir>   1/06/21 15:13  mon_frid_materials
   <dir>   3/04/21  8:46  New folder
 2993.5k   2/22/21 18:04  StataTutorial.pdf
   50.2k  10/27/11 16:26  Statistics as a Career.pptx
  116.5k   7/12/11 15:17  Timetable09.doc
  164.5k   7/20/11 12:15  Timetable2011.doc
  651.4k   3/24/21 19:26  tutorial.pdf

. use lab_results0.dta,clear

. import delimited "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\birth_weight
> r
(8 vars, 641 obs)

. import delimited using birth_weight.csv,varnames(1) clear
(8 vars, 641 obs)

.
```

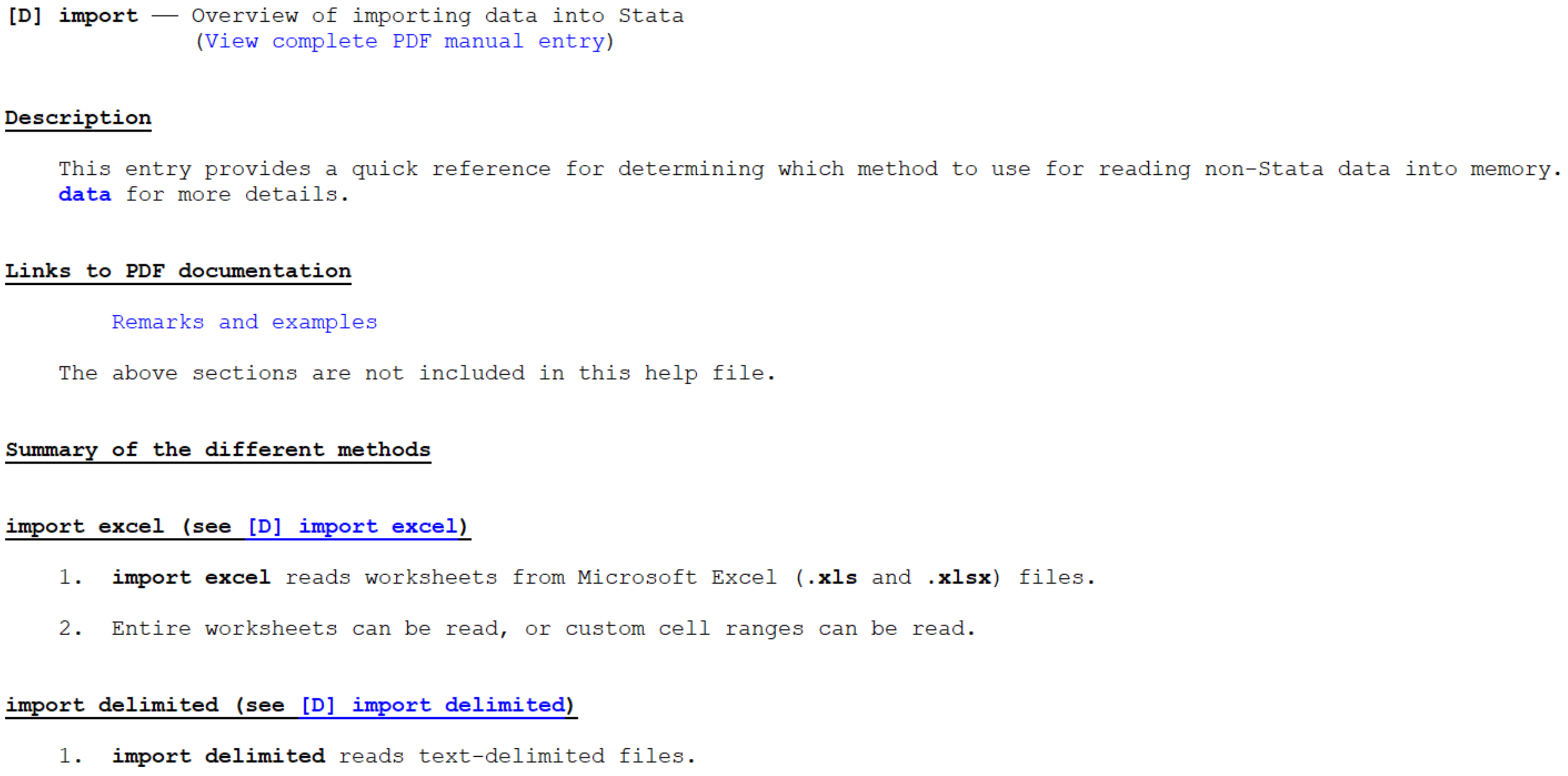
Command

```
import delimited using birth_weight.csv,varnames(1) clear
```

Type this

Import other type of data to stata

help import



[D] import — Overview of importing data into Stata
(View complete PDF manual entry)

Description

This entry provides a quick reference for determining which method to use for reading non-Stata data into memory.
data for more details.

Links to PDF documentation

Remarks and examples

The above sections are not included in this help file.

Summary of the different methods

import excel (see [D] import excel)

1. **import excel** reads worksheets from Microsoft Excel (**.xls** and **.xlsx**) files.

2. Entire worksheets can be read, or custom cell ranges can be read.

import delimited (see [D] import delimited)

1. **import delimited** reads text-delimited files.

2. An observation can be on more than one line.

3. ASCII or EBCDIC data can be read.

4. **infile** (fixed format) has the most capabilities for reading data.

import sasxport (see [D] import sasxport)

1. **import sasxport** reads SAS XPORT Transport format files.

2. **import sasxport** will also read value label information from a **formats.xpf** XPORT file, if available.

import fred (see [D] import fred)

1. **import fred** reads Federal Reserve Economic Data.

2. To use **import fred**, you must have a valid API key obtained from the St. Louis Federal Reserve.

import haver (Windows only) (see [D] import haver)

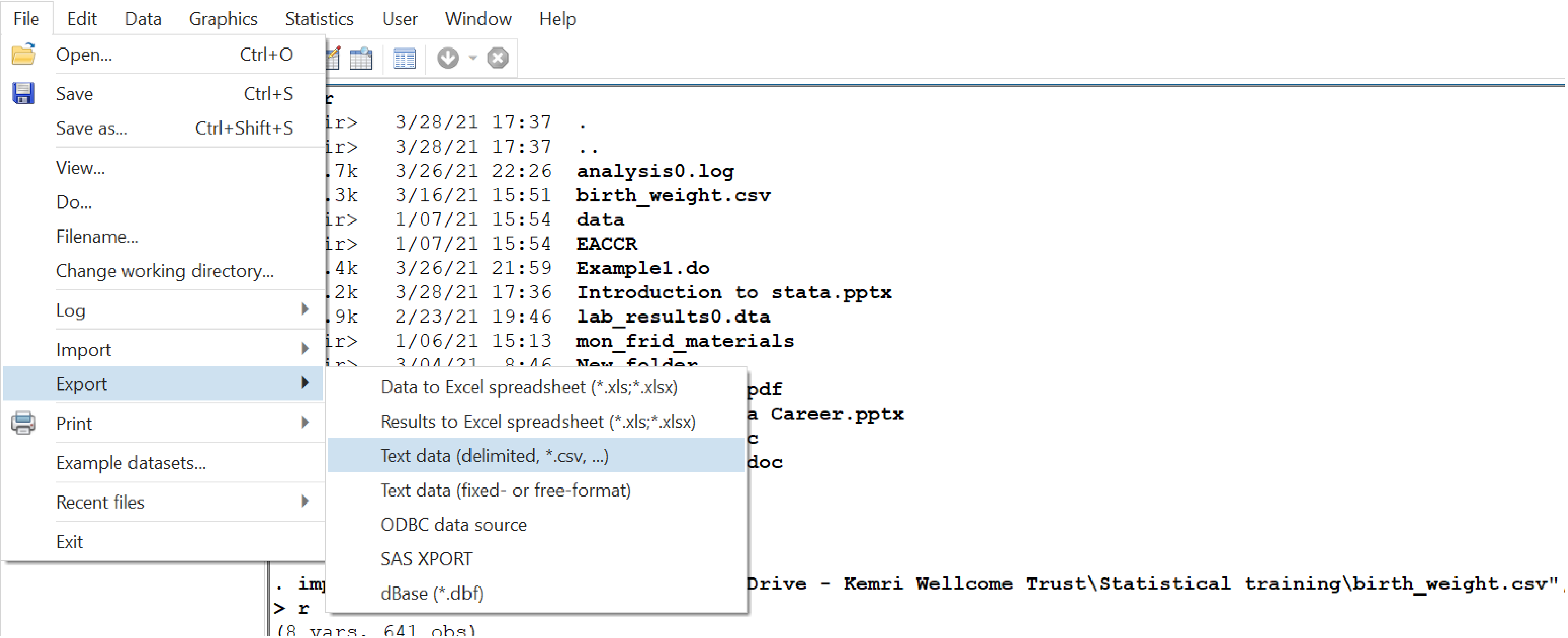
1. **import haver** reads Haver Analytics (http://www.haver.com/) database files.

import dbase (see [D] import dbase)

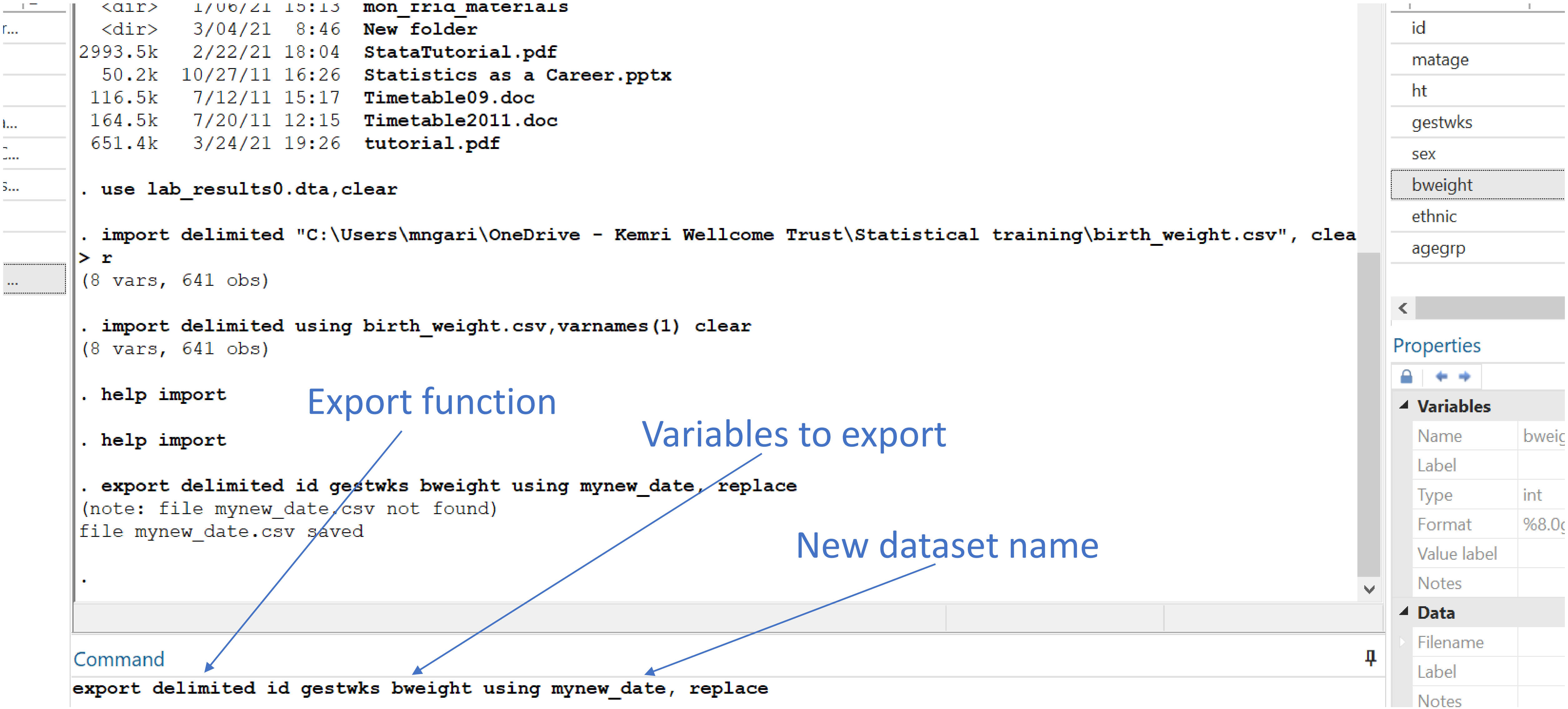
1. **import dbase** reads a version III or version IV dBase (**.dbf**) file.

spshape2dta (see [SP] spshape2dta)

Export data out of stata



```
r
ir>    3/28/21 17:37   .
ir>    3/28/21 17:37   ..
.7k    3/26/21 22:26   analysis0.log
.3k    3/16/21 15:51   birth_weight.csv
ir>    1/07/21 15:54   data
ir>    1/07/21 15:54   EACCR
.4k    3/26/21 21:59   Example1.do
.2k    3/28/21 17:36   Introduction to stata.pptx
.9k    2/23/21 19:46   lab_results0.dta
ir>    1/06/21 15:13   mon_frid_materials
ir>    3/04/21  8:46   New folder
                                           pdf
                                           a Career.pptx
                                           c
                                           doc

. im                                       Drive - Kemri Wellcome Trust\Statistical training\birth_weight.csv",
> r
(8 vars, 641 obs)
```



```
  <dir>    1/06/21 15:13  mon_frid_materials
  <dir>    3/04/21  8:46  New folder
2993.5k   2/22/21 18:04  StataTutorial.pdf
  50.2k  10/27/11 16:26  Statistics as a Career.pptx
 116.5k   7/12/11 15:17  Timetable09.doc
 164.5k   7/20/11 12:15  Timetable2011.doc
 651.4k   3/24/21 19:26  tutorial.pdf

. use lab_results0.dta,clear

. import delimited "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training\birth_weight.csv", clea
> r
(8 vars, 641 obs)

. import delimited using birth_weight.csv,varnames(1) clear
(8 vars, 641 obs)

. help import

. help import

. export delimited id gestwks bweight using mynew_date, replace
(note: file mynew_date.csv not found)
file mynew_date.csv saved

.
```

Command

```
export delimited id gestwks bweight using mynew_date, replace
```

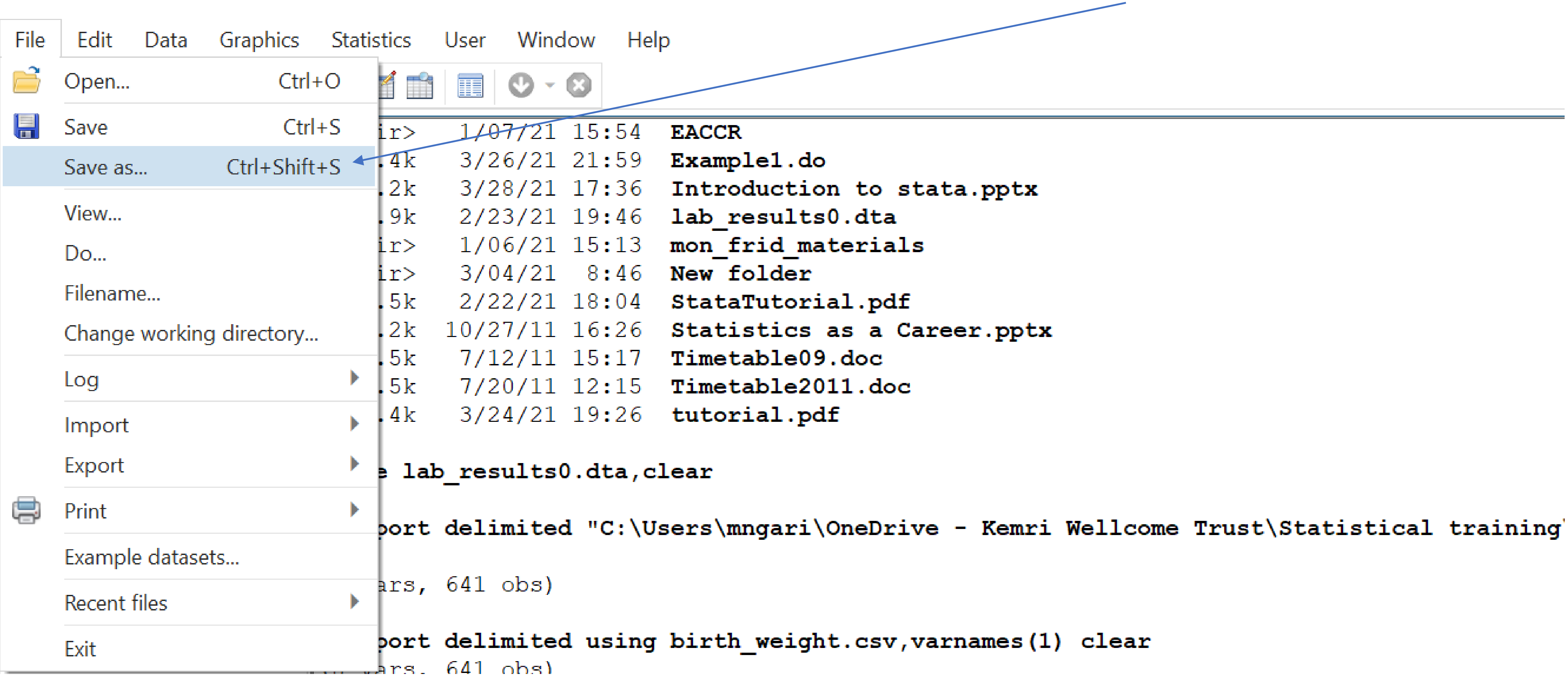
Export function

Variables to export

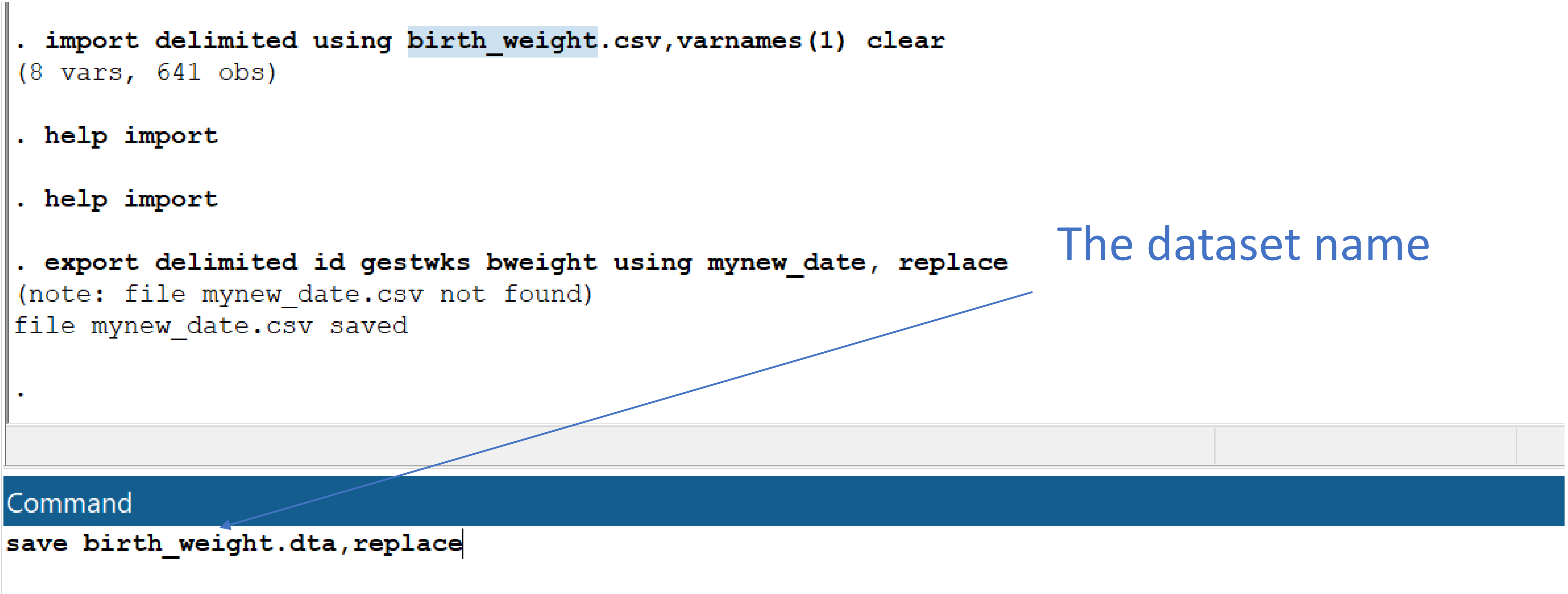
New dataset name

Saving data in stata

Provide the file to save and the dataset name



```
ir>    1/07/21 15:54  EACCR
.4k    3/26/21 21:59  Example1.do
.2k    3/28/21 17:36  Introduction to stata.pptx
.9k    2/23/21 19:46  lab_results0.dta
ir>    1/06/21 15:13  mon_frid_materials
ir>    3/04/21  8:46  New folder
.5k    2/22/21 18:04  StataTutorial.pdf
.2k   10/27/11 16:26  Statistics as a Career.pptx
.5k    7/12/11 15:17  Timetable09.doc
.5k    7/20/11 12:15  Timetable2011.doc
.4k    3/24/21 19:26  tutorial.pdf

e lab_results0.dta,clear

port delimited "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training'

ars,  641 obs)

port delimited using birth_weight.csv,varnames(1) clear
ars,  641 obs)
```



```
. import delimited using birth_weight.csv,varnames(1) clear
(8 vars, 641 obs)

. help import

. help import

. export delimited id gestwks bweight using mynew_date, replace
(note: file mynew_date.csv not found)
file mynew_date.csv saved

.
```

```
save birth_weight.dta,replace
```

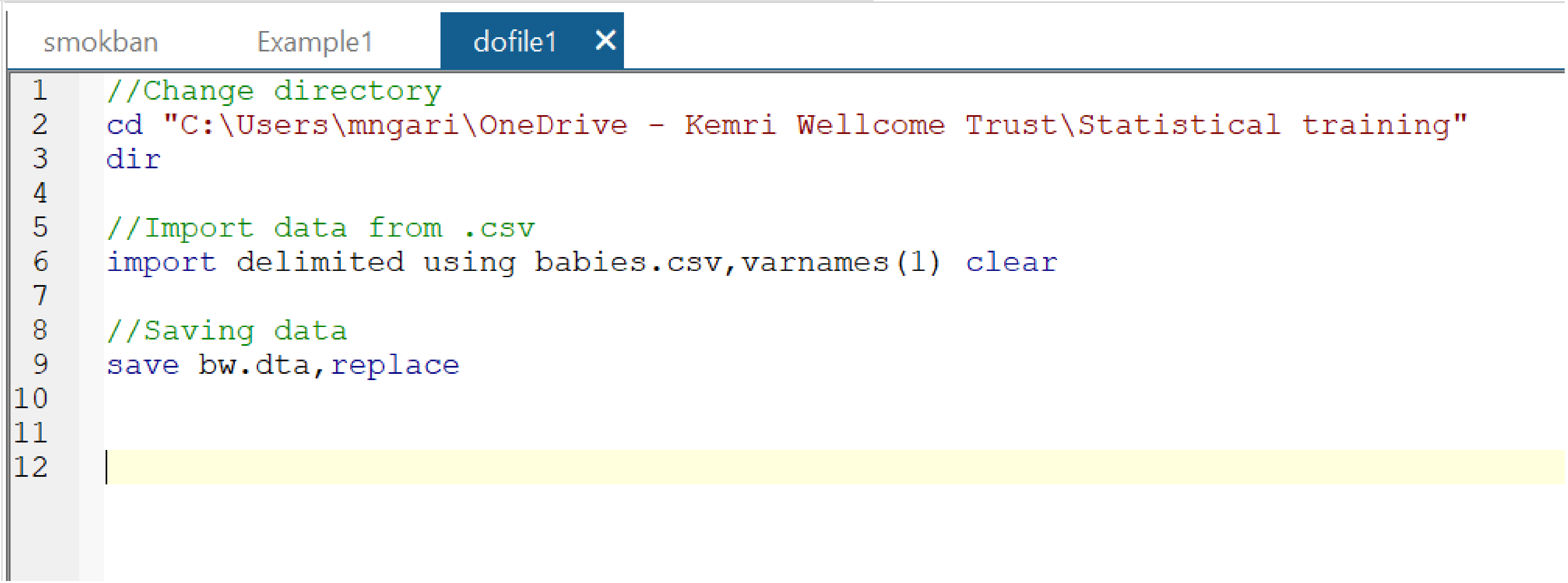
The dataset name

Replace will replace dataset in the memory with that name

Your turn now

Exercise 1

- In the folder provided there is a dataset: babies.csv
- a) Open a stata window and change directory to the folder with the dataset
- b) Open a new do-file
- c) read the babies.csv into stata
- d) Save as bw.dta
- e) save the do-file as: dofile1



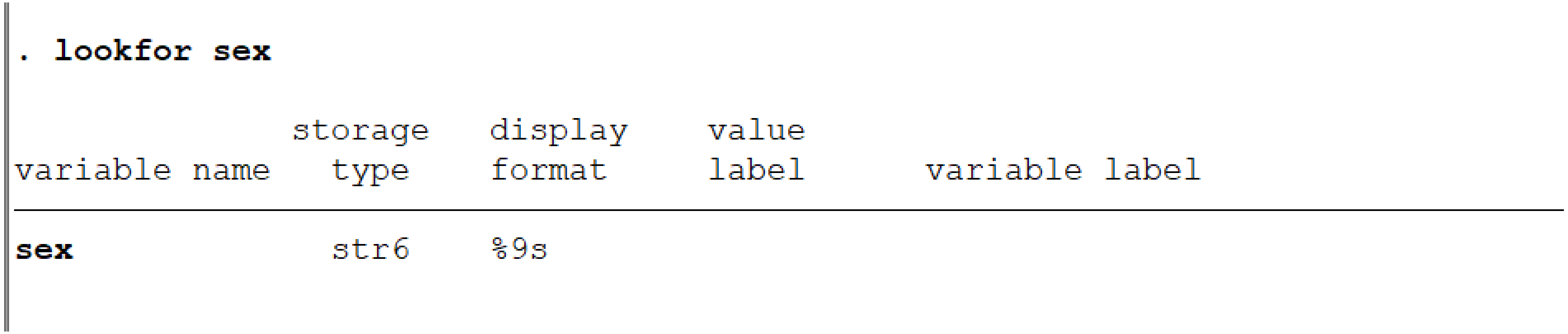
```
//Change directory
cd "C:\Users\mngari\OneDrive - Kemri Wellcome Trust\Statistical training"
dir

//Import data from .csv
import delimited using babies.csv,varnames(1) clear

//Saving data
save bw.dta,replace


```

Data management

- Import birth_weight.csv dataset to stata

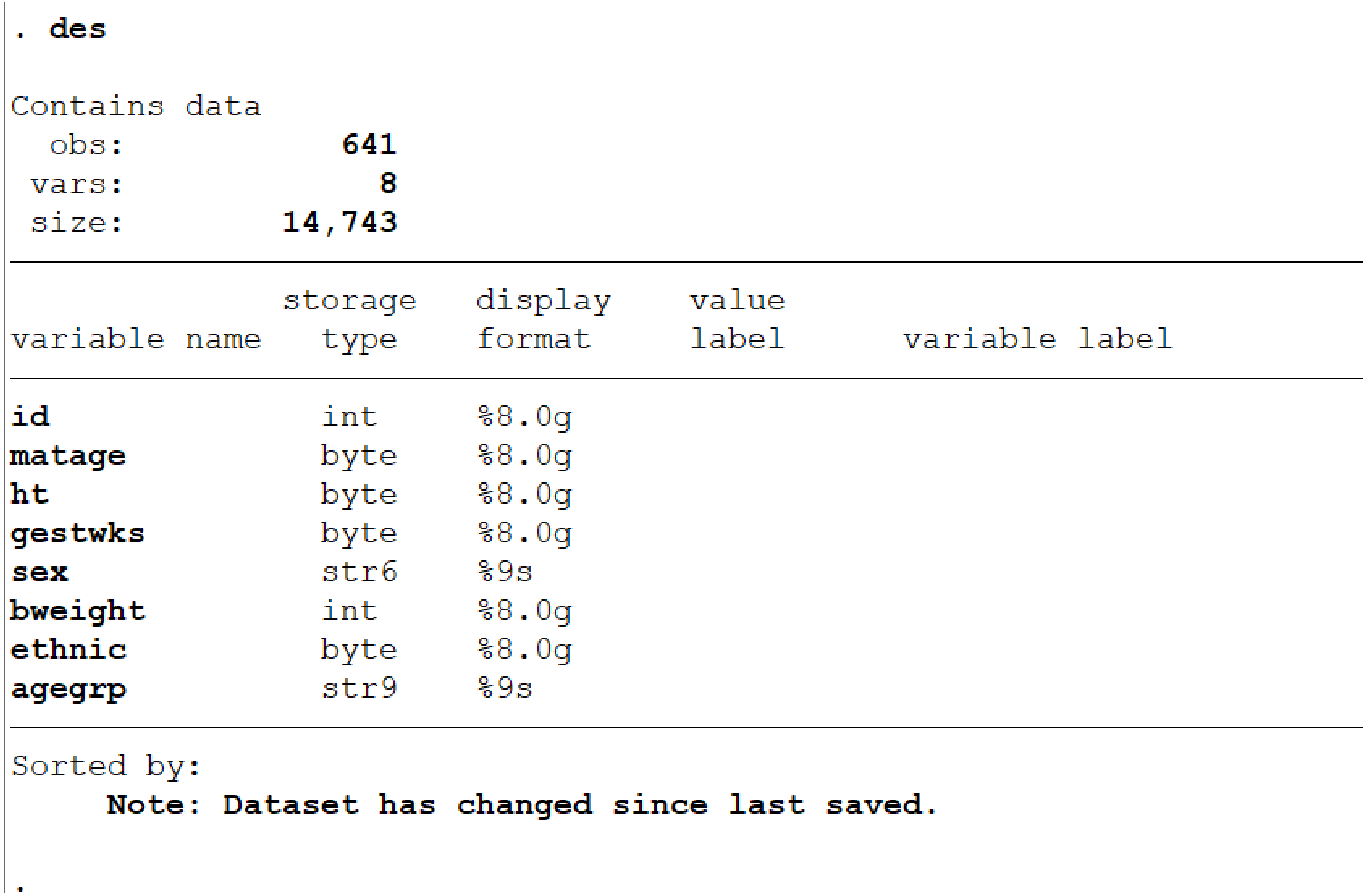
import delimited using birth_weight.csv,varnames(1) clear

You can use the command *lookfor* to find variables in a dataset



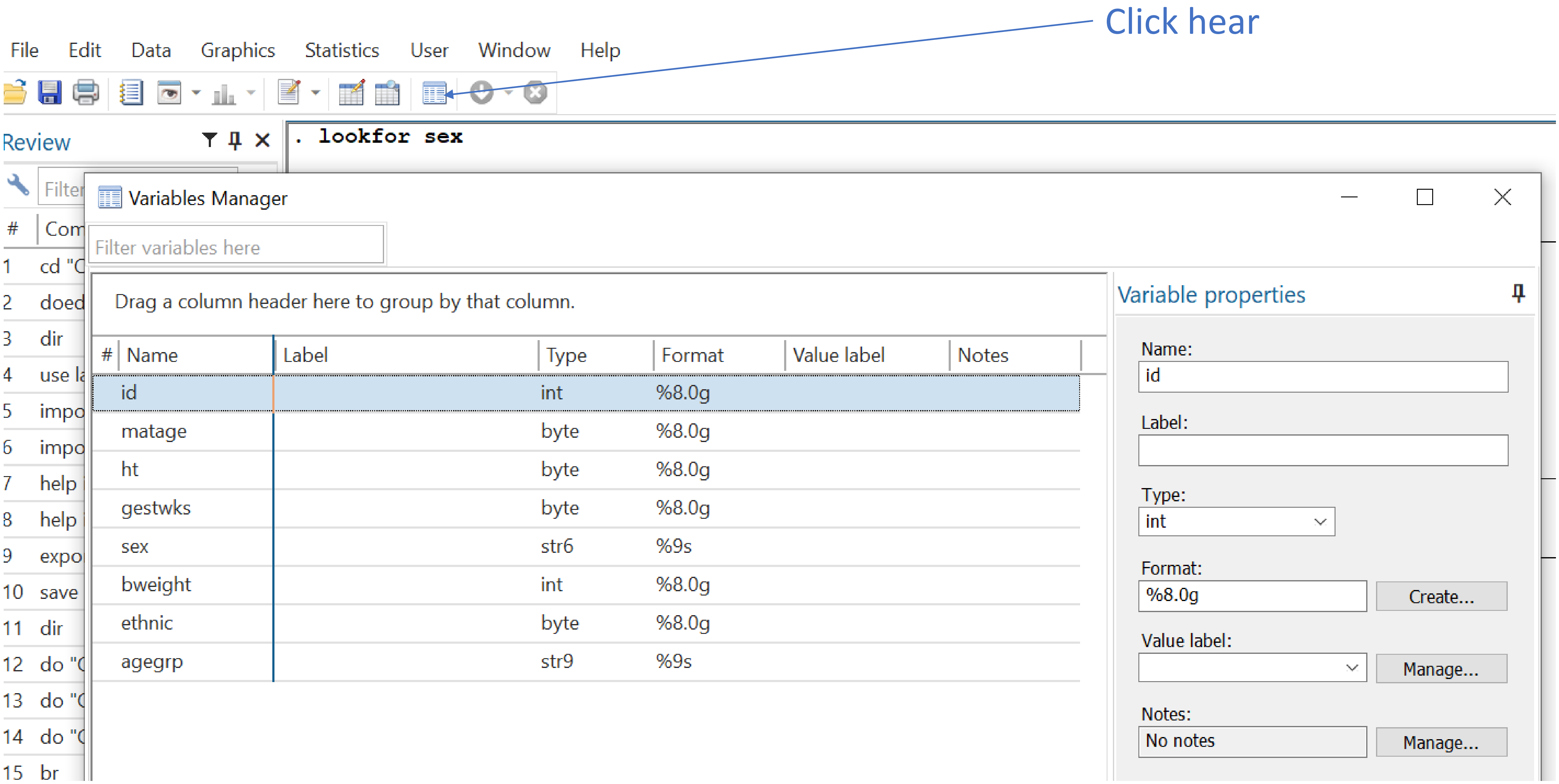
```
. lookfor sex

                storage   display    value
variable name     type    format     label      variable label
------------------------------------------------------------------------------
sex               str6    %9s
```

Data variables description



```
. des

Contains data
  obs:           641
 vars:             8
 size:        14,743
----------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
----------------------------------------------------------------------
id              int     %8.0g
matage          byte    %8.0g
ht              byte    %8.0g
gestwks         byte    %8.0g
sex             str6    %9s
bweight         int     %8.0g
ethnic          byte    %8.0g
agegrp          str9    %9s
----------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.

.
```

Click hear



```
. lookfor sex
```

#	Name	Label	Type	Format	Value label	Notes
	id		int	%8.0g		
	matage		byte	%8.0g		
	ht		byte	%8.0g		
	gestwks		byte	%8.0g		
	sex		str6	%9s		
	bweight		int	%8.0g		
	ethnic		byte	%8.0g		
	agegrp		str9	%9s		

Name: id

Label:

Type: int

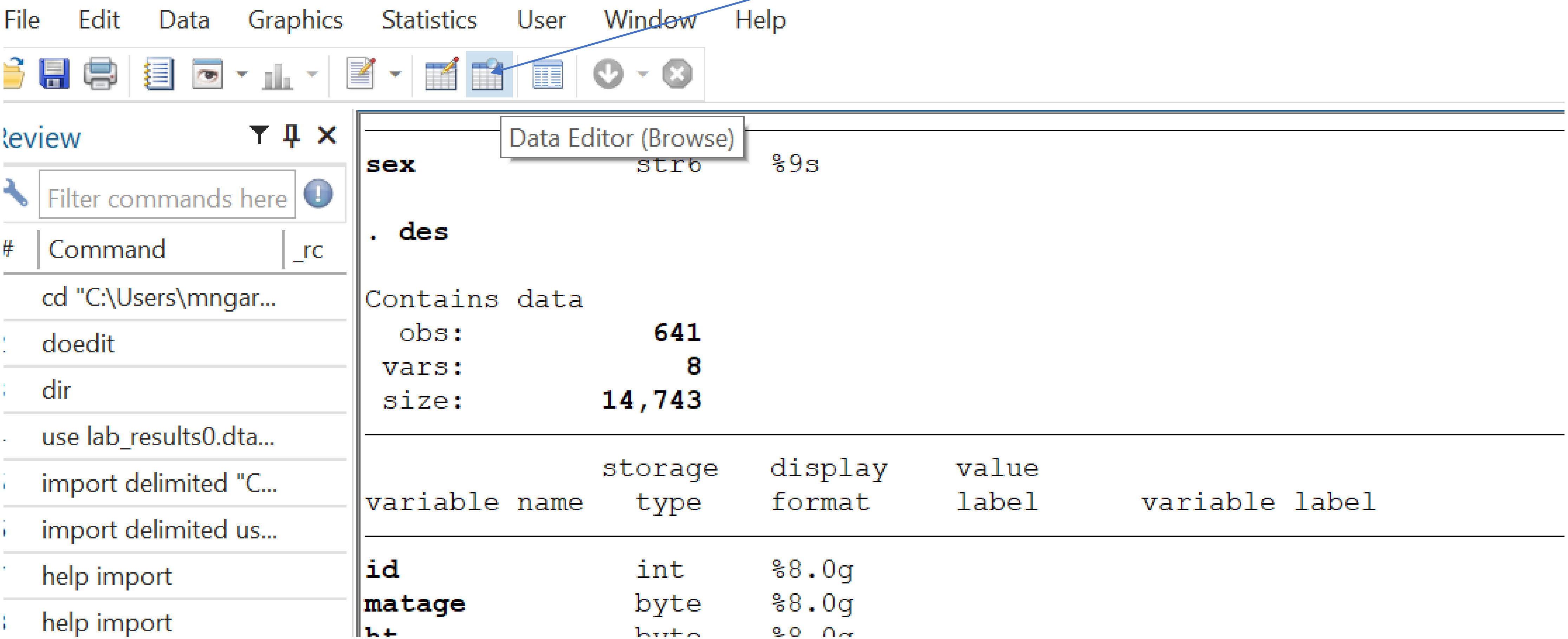
Format: %8.0g

Value label:

Notes: No notes

View the variables as spreadsheet

- Type *browse*

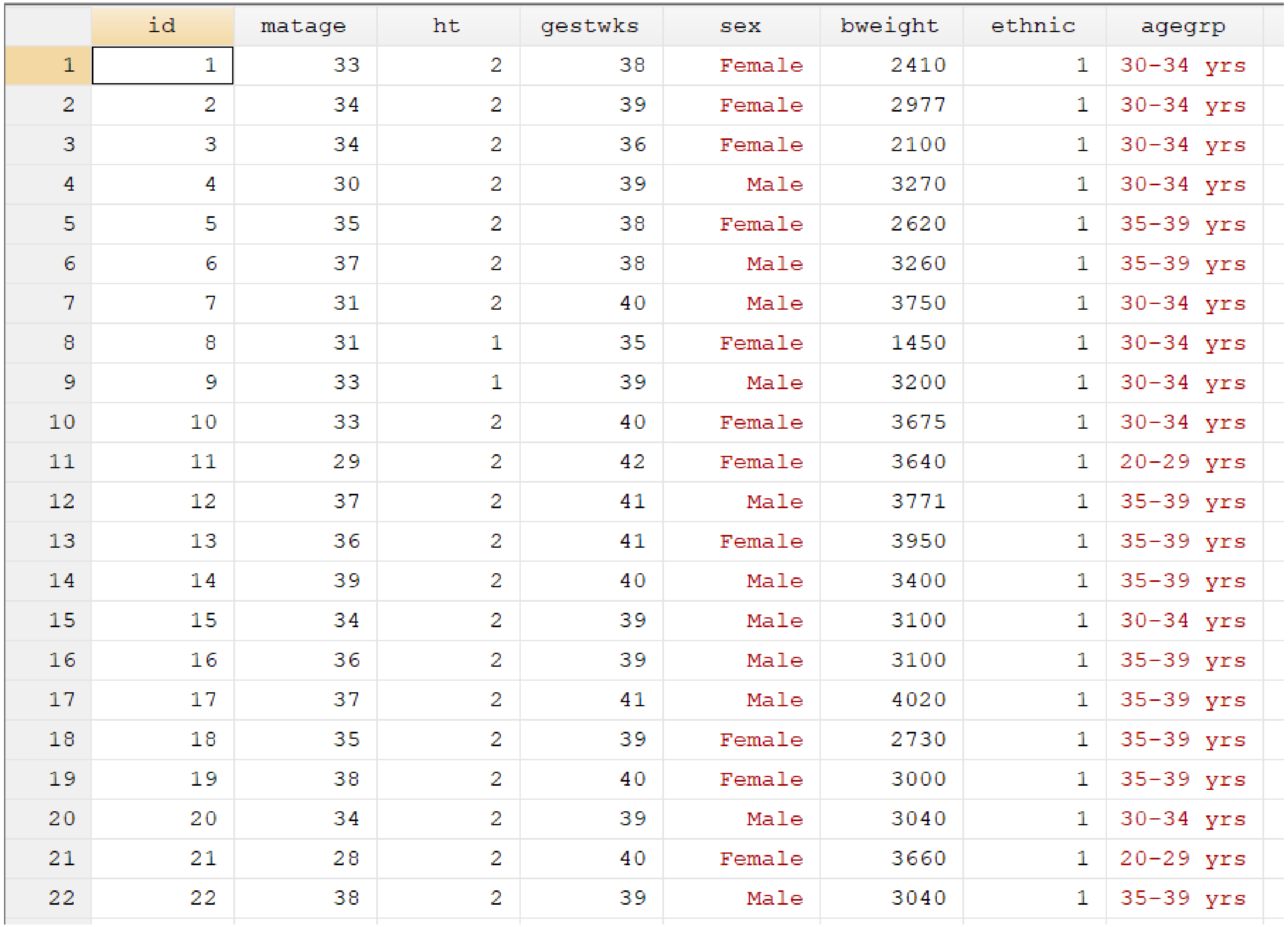


Click here

```
cd "C:\Users\mngar...
doedit
dir
use lab_results0.dta...
import delimited "C...
import delimited us...
help import
help import
```

```
sex                      str6      %9s

. des

Contains data
  obs:                 641
 vars:                   8
 size:              14,743

              storage   display    value
variable name   type    format     label      variable label

id              int     %8.0g
matage          byte    %8.0g
ht              byte    %8.0g
```

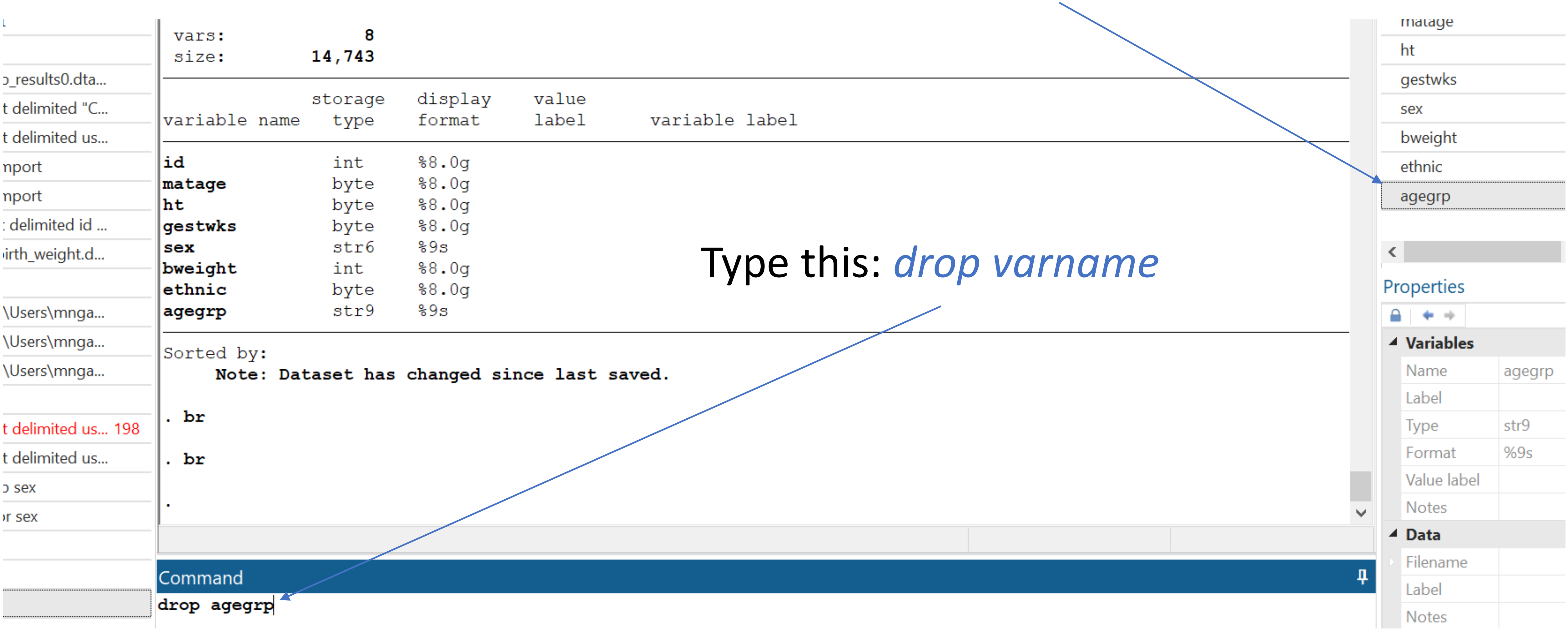
Stata has a color-coded system for each type. Black is for numbers, red is for text or string and blue is for labeled variables.



	id	matage	ht	gestwks	sex	bweight	ethnic	agegrp
1	1	33	2	38	Female	2410	1	30-34 yrs
2	2	34	2	39	Female	2977	1	30-34 yrs
3	3	34	2	36	Female	2100	1	30-34 yrs
4	4	30	2	39	Male	3270	1	30-34 yrs
5	5	35	2	38	Female	2620	1	35-39 yrs
6	6	37	2	38	Male	3260	1	35-39 yrs
7	7	31	2	40	Male	3750	1	30-34 yrs
8	8	31	1	35	Female	1450	1	30-34 yrs
9	9	33	1	39	Male	3200	1	30-34 yrs
10	10	33	2	40	Female	3675	1	30-34 yrs
11	11	29	2	42	Female	3640	1	20-29 yrs
12	12	37	2	41	Male	3771	1	35-39 yrs
13	13	36	2	41	Female	3950	1	35-39 yrs
14	14	39	2	40	Male	3400	1	35-39 yrs
15	15	34	2	39	Male	3100	1	30-34 yrs
16	16	36	2	39	Male	3100	1	35-39 yrs
17	17	37	2	41	Male	4020	1	35-39 yrs
18	18	35	2	39	Female	2730	1	35-39 yrs
19	19	38	2	40	Female	3000	1	35-39 yrs
20	20	34	2	39	Male	3040	1	30-34 yrs
21	21	28	2	40	Female	3660	1	20-29 yrs
22	22	38	2	39	Male	3040	1	35-39 yrs

Remove a variable

This will remove the agegrp variable from your data permanently



```
  vars:              8
  size:         14,743
------------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------------------------
id              int     %8.0g
matage          byte    %8.0g
ht              byte    %8.0g
gestwks         byte    %8.0g
sex             str6    %9s
bweight         int     %8.0g
ethnic          byte    %8.0g
agegrp          str9    %9s
------------------------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.

. br

. br

.
```

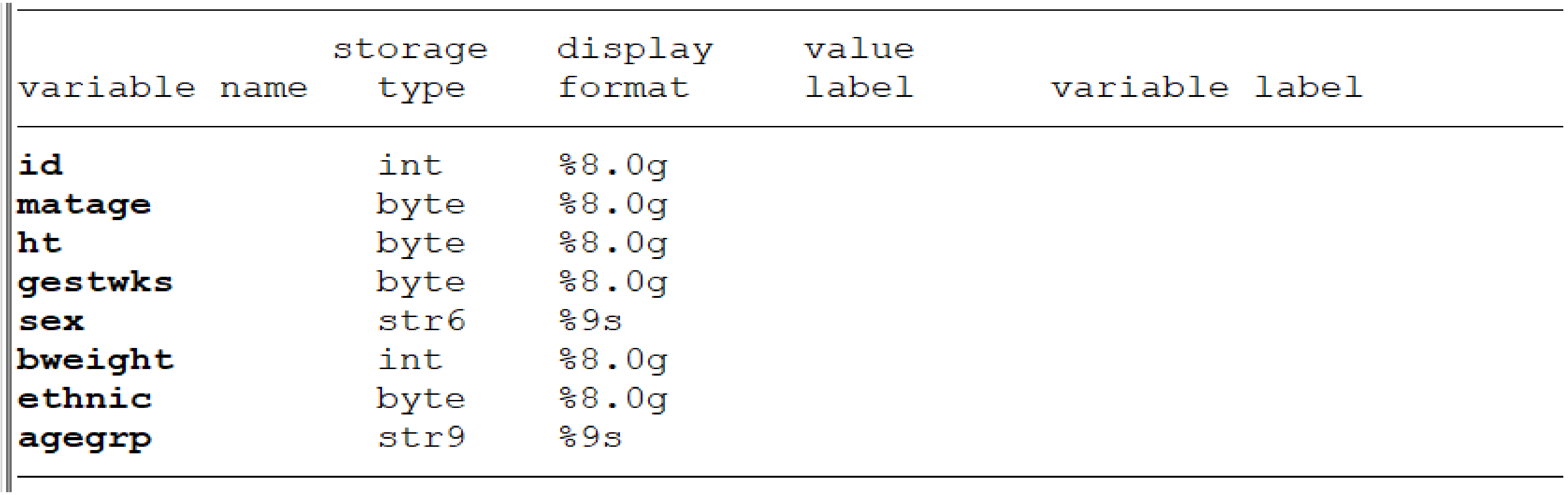
Command

```
drop agegrp
```

Type this: *drop varname*

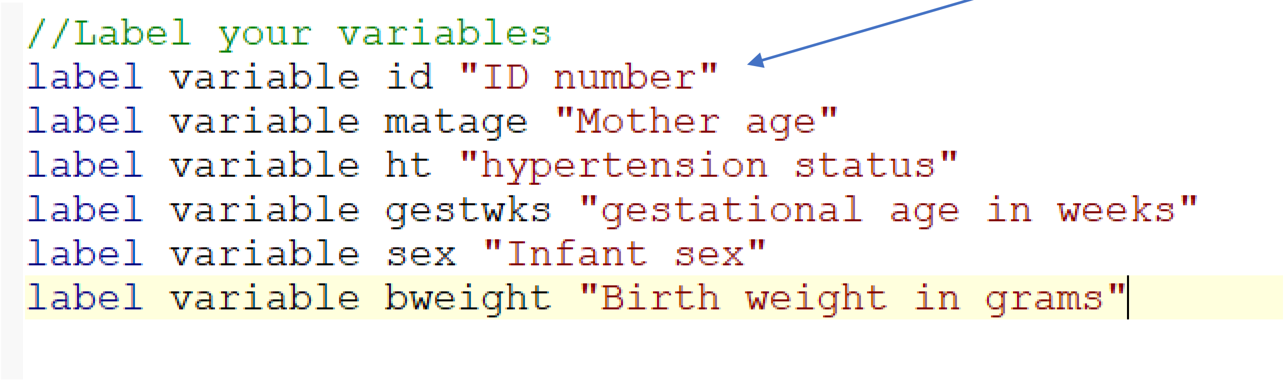
Adding/changing variable labels, type:

Before



```
              storage   display    value
variable name   type    format     label      variable label

id              int     %8.0g
matage          byte    %8.0g
ht              byte    %8.0g
gestwks         byte    %8.0g
sex             str6    %9s
bweight         int     %8.0g
ethnic          byte    %8.0g
agegrp          str9    %9s
```

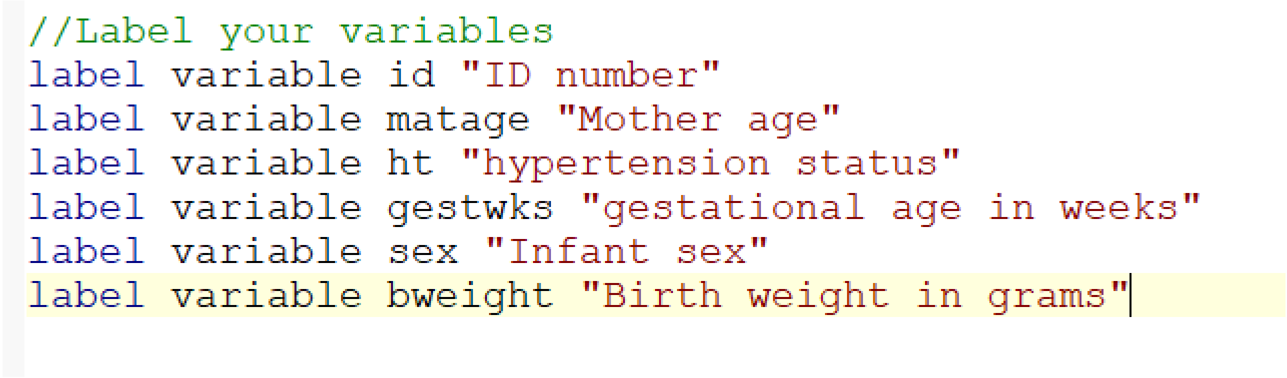
Use the function: label variable varname varlabel



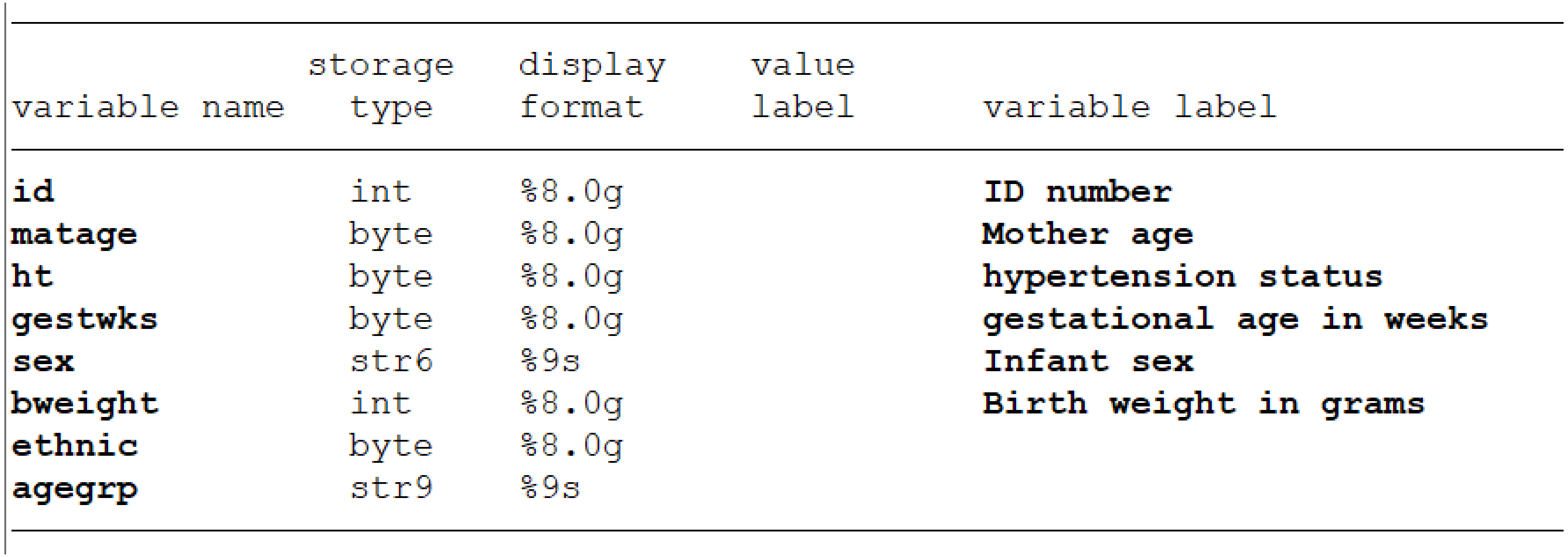
```
//Label your variables
label variable id "ID number"
label variable matage "Mother age"
label variable ht "hypertension status"
label variable gestwks "gestational age in weeks"
label variable sex "Infant sex"
label variable bweight "Birth weight in grams"
```

Adding/changing variable labels, type:

After



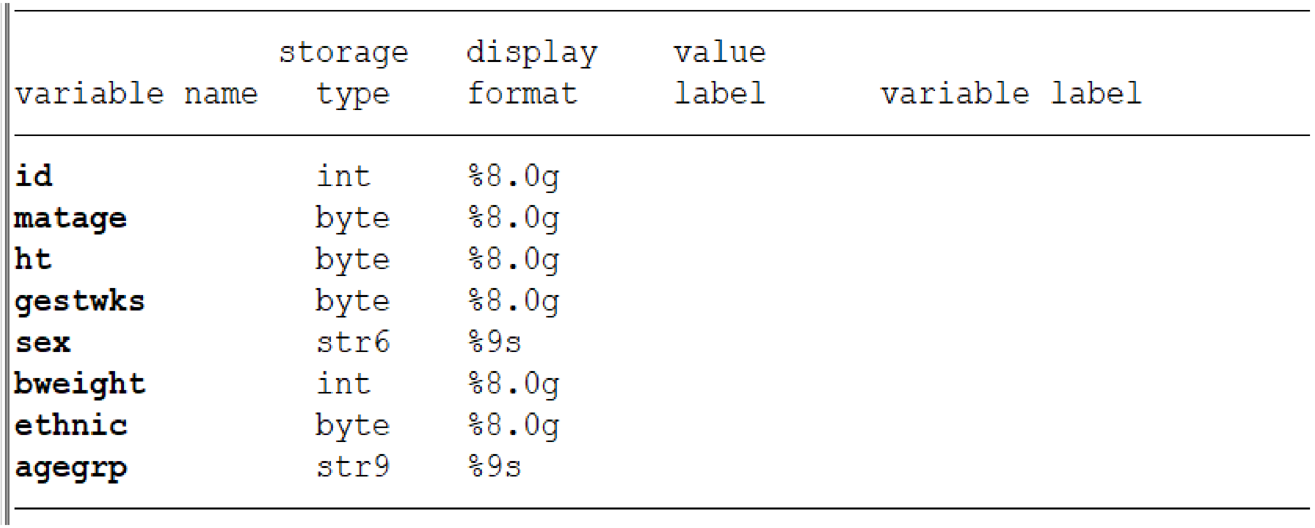
```
//Label your variables
label variable id "ID number"
label variable matage "Mother age"
label variable ht "hypertension status"
label variable gestwks "gestational age in weeks"
label variable sex "Infant sex"
label variable bweight "Birth weight in grams"
```



```
                      storage   display    value
variable name           type    format     label      variable label

id                      int     %8.0g                 ID number
matage                  byte    %8.0g                 Mother age
ht                      byte    %8.0g                 hypertension status
gestwks                 byte    %8.0g                 gestational age in weeks
sex                     str6    %9s                   Infant sex
bweight                 int     %8.0g                 Birth weight in grams
ethnic                  byte    %8.0g
agegrp                  str9    %9s
```

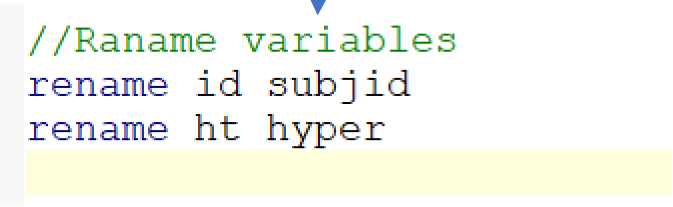
Renaming variables, type:

Before



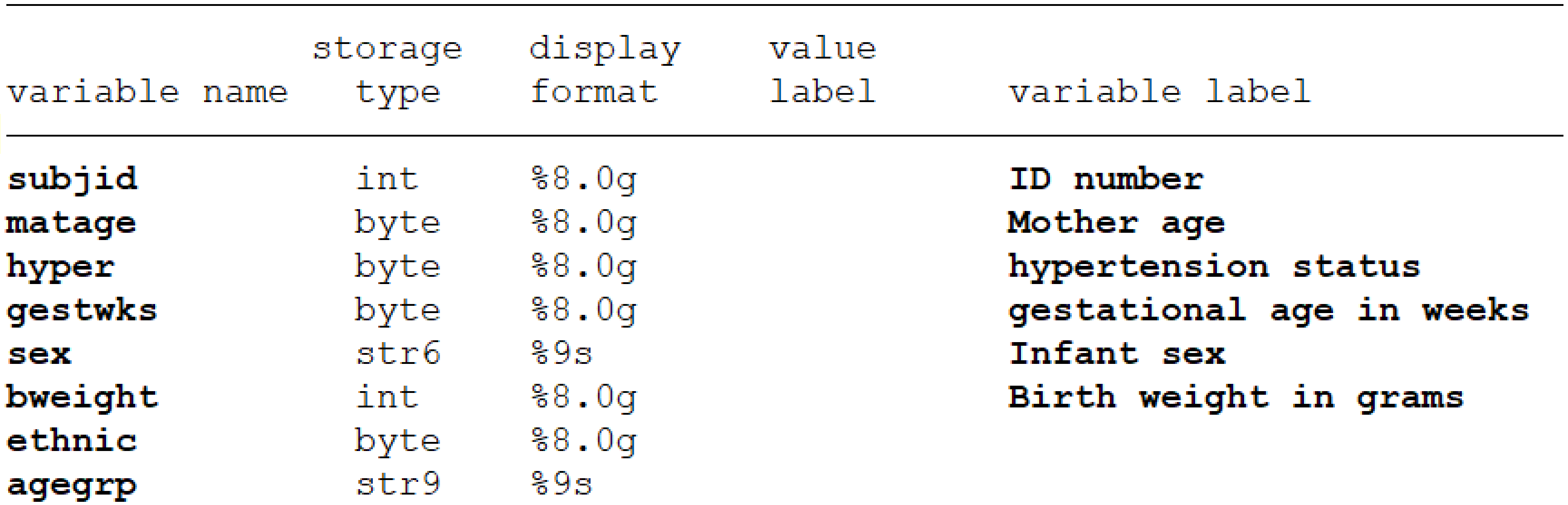
```
                  storage   display    value
variable name       type    format     label      variable label

id                  int     %8.0g
matage              byte    %8.0g
ht                  byte    %8.0g
gestwks             byte    %8.0g
sex                 str6    %9s
bweight             int     %8.0g
ethnic              byte    %8.0g
agegrp              str9    %9s
```

To rename: rename varname new_varname



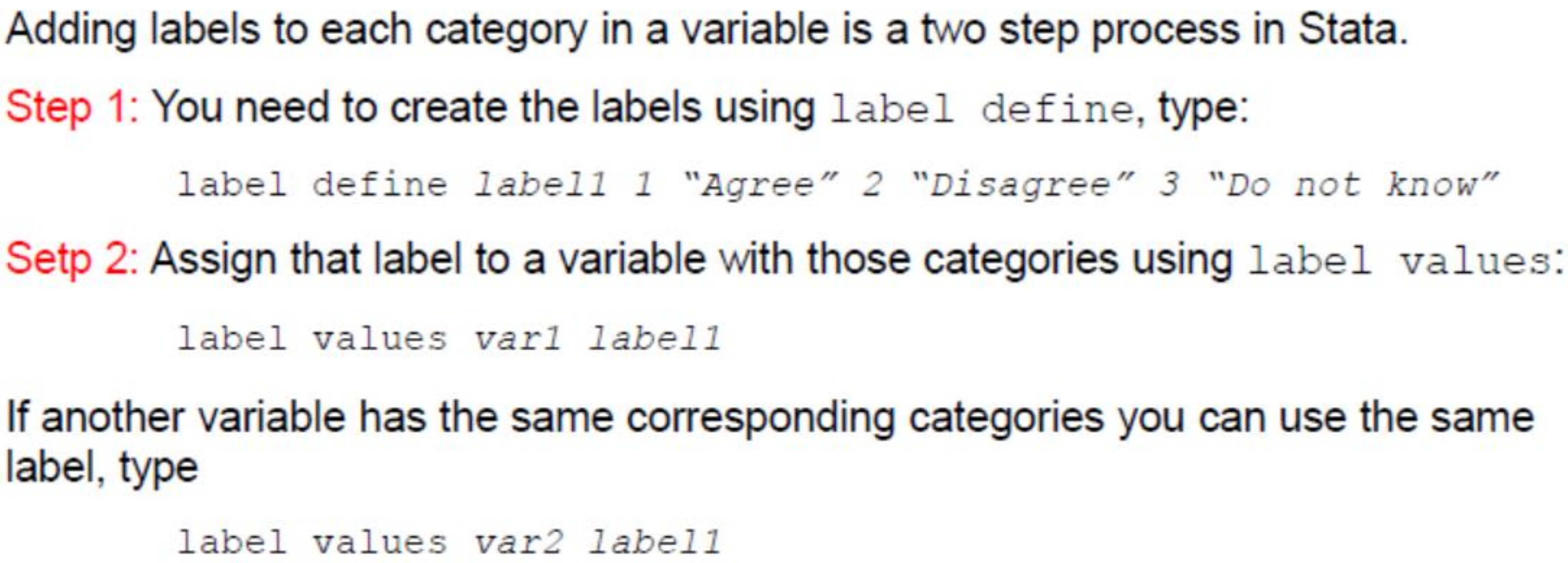
```
//Raname variables
rename id subjid
rename ht hyper
```

After



```
                  storage   display    value
variable name       type    format     label      variable label

subjid              int     %8.0g                 ID number
matage              byte    %8.0g                 Mother age
hyper               byte    %8.0g                 hypertension status
gestwks             byte    %8.0g                 gestational age in weeks
sex                 str6    %9s                   Infant sex
bweight             int     %8.0g                 Birth weight in grams
ethnic              byte    %8.0g
agegrp              str9    %9s
```

Assigning value labels



Adding labels to each category in a variable is a two step process in Stata.

Step 1: You need to create the labels using `label define`, type:

```
        label define label1 1 “Agree” 2 “Disagree” 3 “Do not know”
```

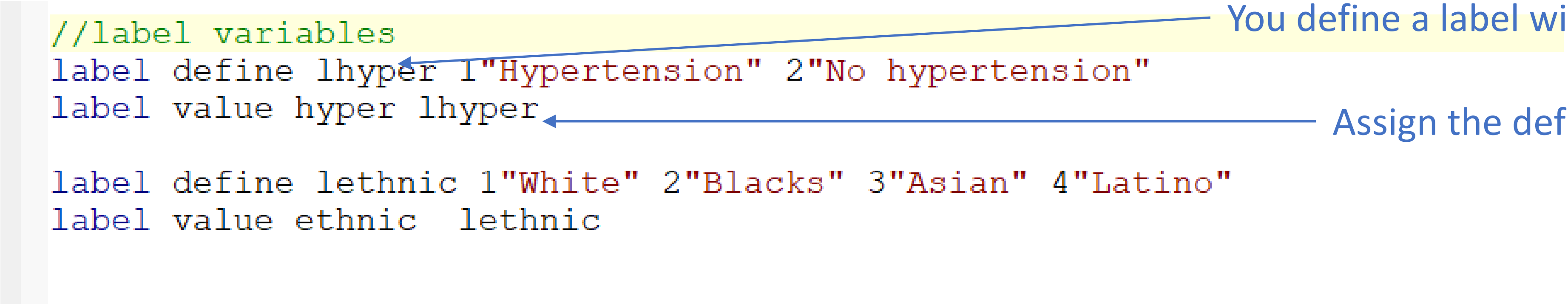
Setp 2: Assign that label to a variable with those categories using `label values`:

```
        label values var1 label1
```

If another variable has the same corresponding categories you can use the same label, type

```
        label values var2 label1
```

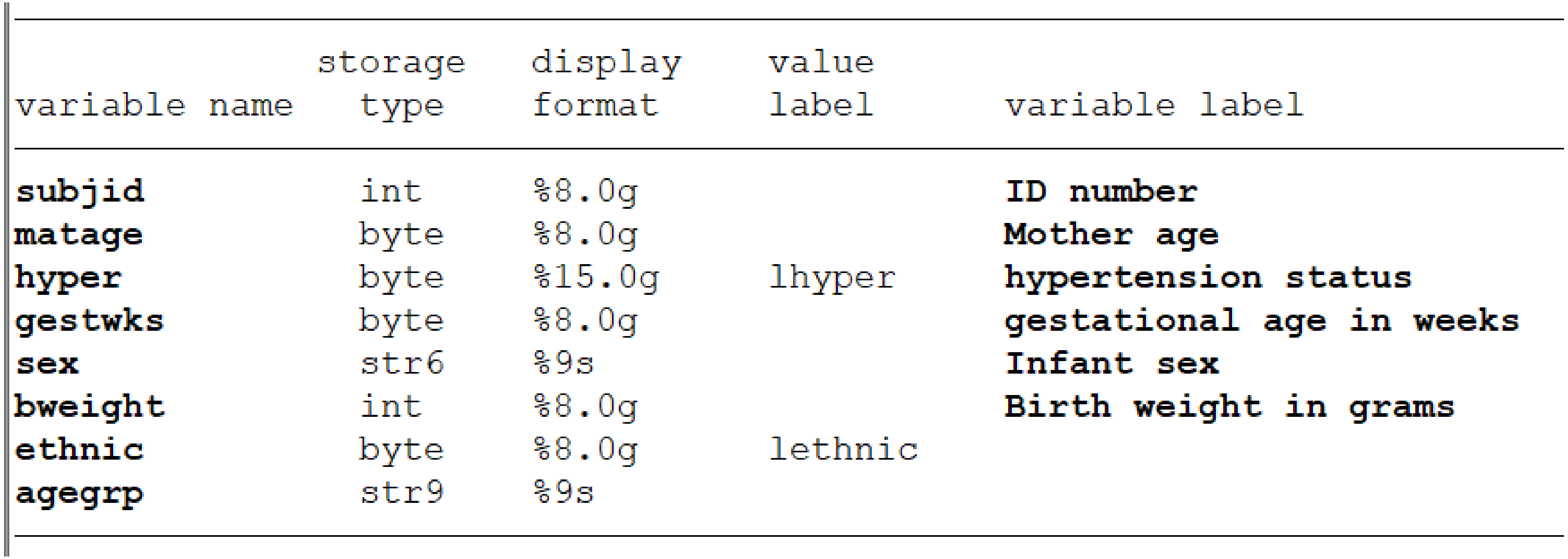
Assigning value labels



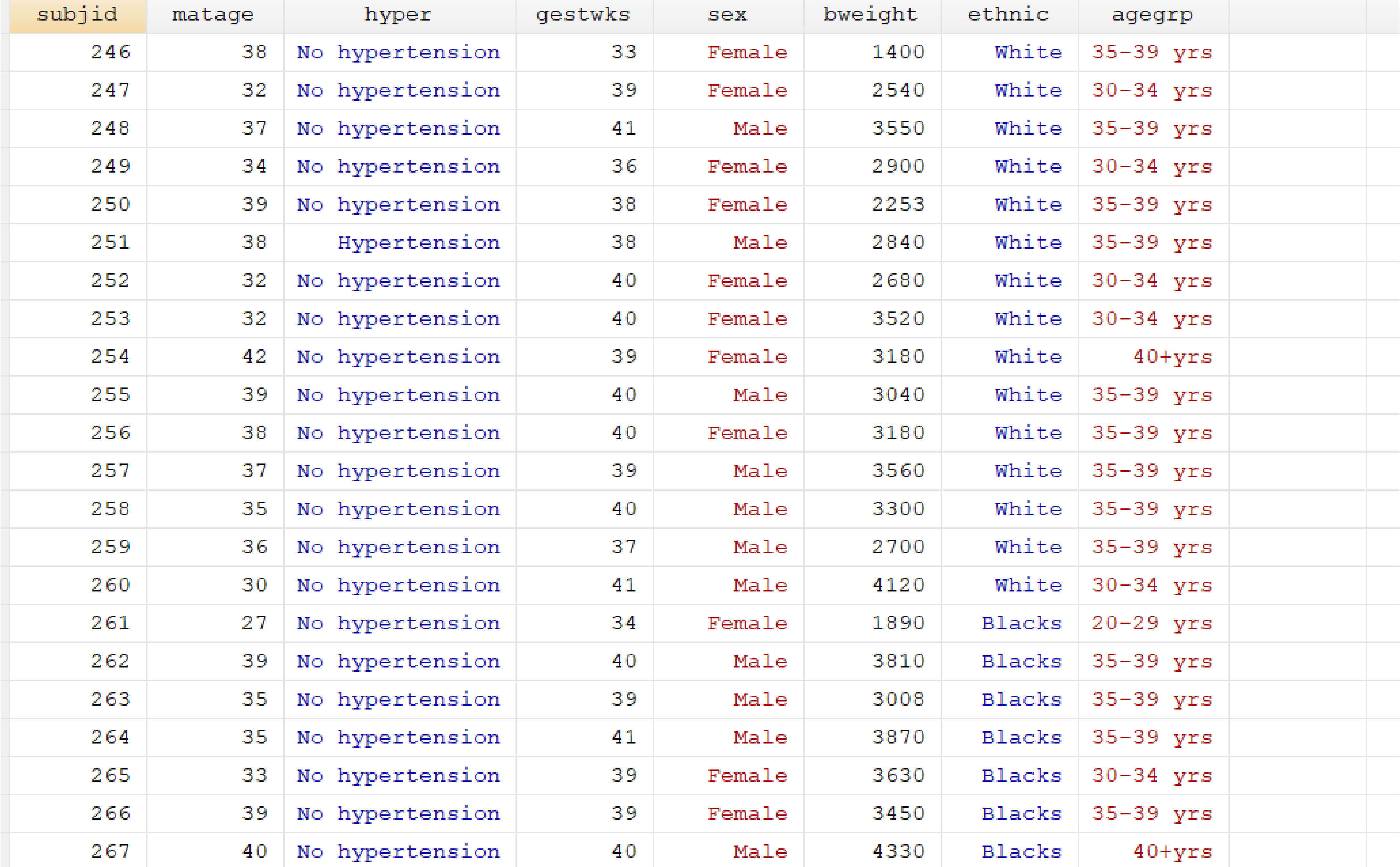
```
//label variables
label define lhyper 1"Hypertension" 2"No hypertension"
label value hyper lhyper

label define lethnic 1"White" 2"Blacks" 3"Asian" 4"Latino"
label value ethnic  lethnic
```

You define a label with 1=hyper and 2=not hyper

Assign the define label to variable hyper

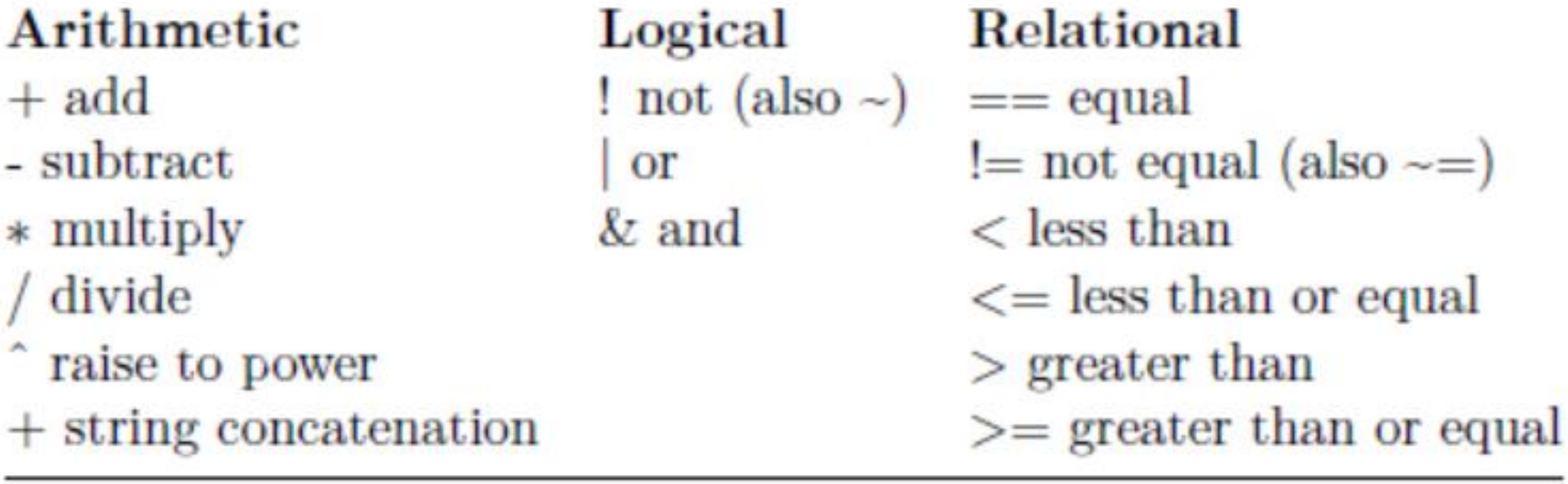


```
              storage   display    value
variable name   type    format     label      variable label

subjid          int     %8.0g                 ID number
matage          byte    %8.0g                 Mother age
hyper           byte    %15.0g     lhyper     hypertension status
gestwks         byte    %8.0g                 gestational age in weeks
sex             str6    %9s                   Infant sex
bweight         int     %8.0g                 Birth weight in grams
ethnic          byte    %8.0g      lethnic
agegrp          str9    %9s
```



subjid	matage	hyper	gestwks	sex	bweight	ethnic	agegrp
246	38	No hypertension	33	Female	1400	White	35-39 yrs
247	32	No hypertension	39	Female	2540	White	30-34 yrs
248	37	No hypertension	41	Male	3550	White	35-39 yrs
249	34	No hypertension	36	Female	2900	White	30-34 yrs
250	39	No hypertension	38	Female	2253	White	35-39 yrs
251	38	Hypertension	38	Male	2840	White	35-39 yrs
252	32	No hypertension	40	Female	2680	White	30-34 yrs
253	32	No hypertension	40	Female	3520	White	30-34 yrs
254	42	No hypertension	39	Female	3180	White	40+yrs
255	39	No hypertension	40	Male	3040	White	35-39 yrs
256	38	No hypertension	40	Female	3180	White	35-39 yrs
257	37	No hypertension	39	Male	3560	White	35-39 yrs
258	35	No hypertension	40	Male	3300	White	35-39 yrs
259	36	No hypertension	37	Male	2700	White	35-39 yrs
260	30	No hypertension	41	Male	4120	White	30-34 yrs
261	27	No hypertension	34	Female	1890	Blacks	20-29 yrs
262	39	No hypertension	40	Male	3810	Blacks	35-39 yrs
263	35	No hypertension	39	Male	3008	Blacks	35-39 yrs
264	35	No hypertension	41	Male	3870	Blacks	35-39 yrs
265	33	No hypertension	39	Female	3630	Blacks	30-34 yrs
266	39	No hypertension	39	Female	3450	Blacks	35-39 yrs
267	40	No hypertension	40	Male	4330	Blacks	40+yrs

Operators and Expressions



Arithmetic	Logical	Relational
+ add	! not (also ~)	== equal
- subtract	| or	!= not equal (also ~=)
∗ multiply	& and	< less than
/ divide		<= less than or equal
ˆ raise to power		> greater than
+ string concatenation		>= greater than or equal

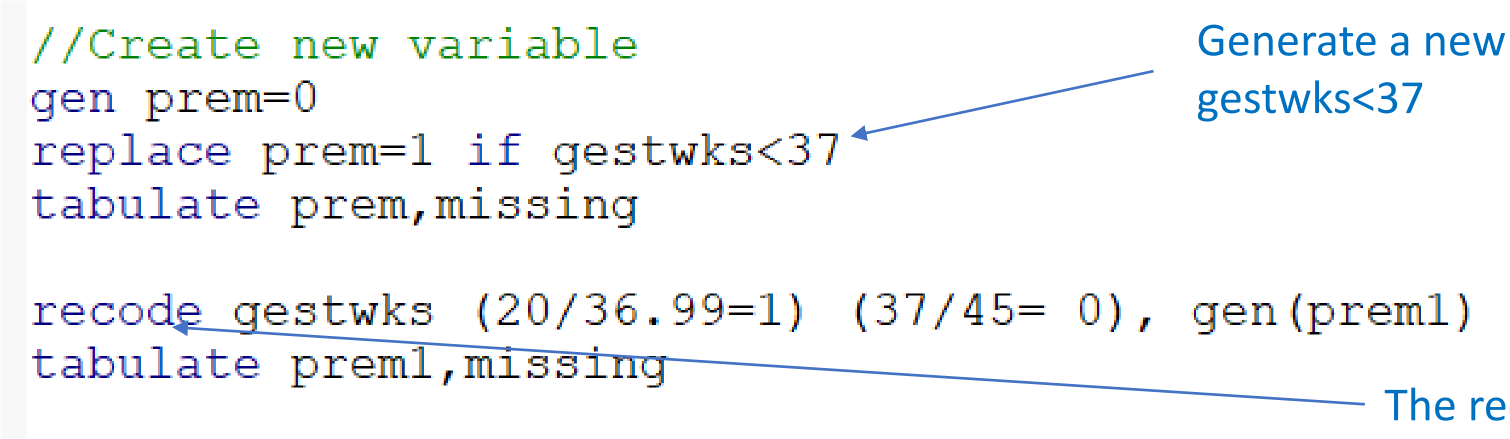
Creating new variables

To generate a new variable use the command `generate` (`gen` for short), type

`generate [newvar] = [expression]`

The variable `gestwks` is the gestation age in weeks ranging from 25 to 42 weeks. We need to create a need variable called `prem` for born premature (0=37weeks and above; 1=less than 37 weeks).

Two ways:



```
//Create new variable
gen prem=0
replace prem=1 if gestwks<37
tabulate prem,missing

recode gestwks (20/36.99=1) (37/45= 0), gen(prem1)
tabulate prem1,missing
```

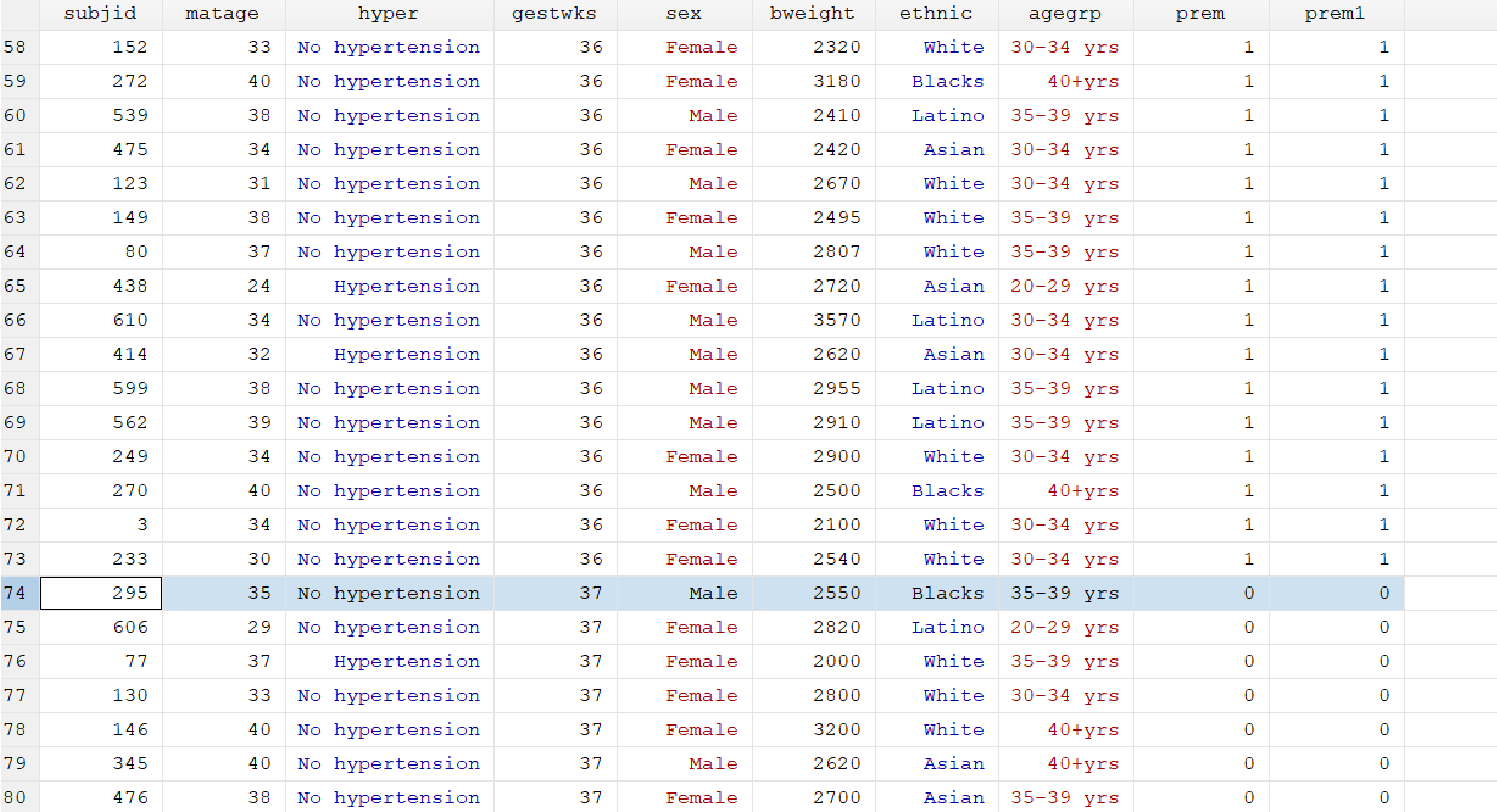
Generate a new variable prem with zero and 1 for gestwks<37

The recode function works as well

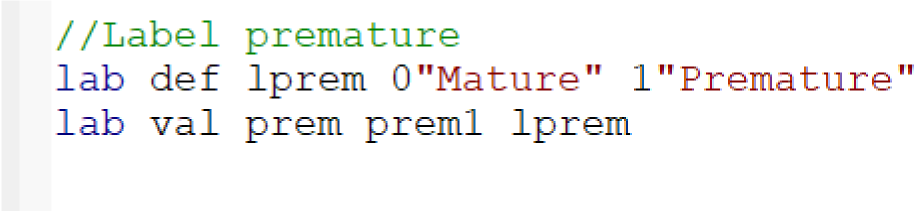
View the data now

br

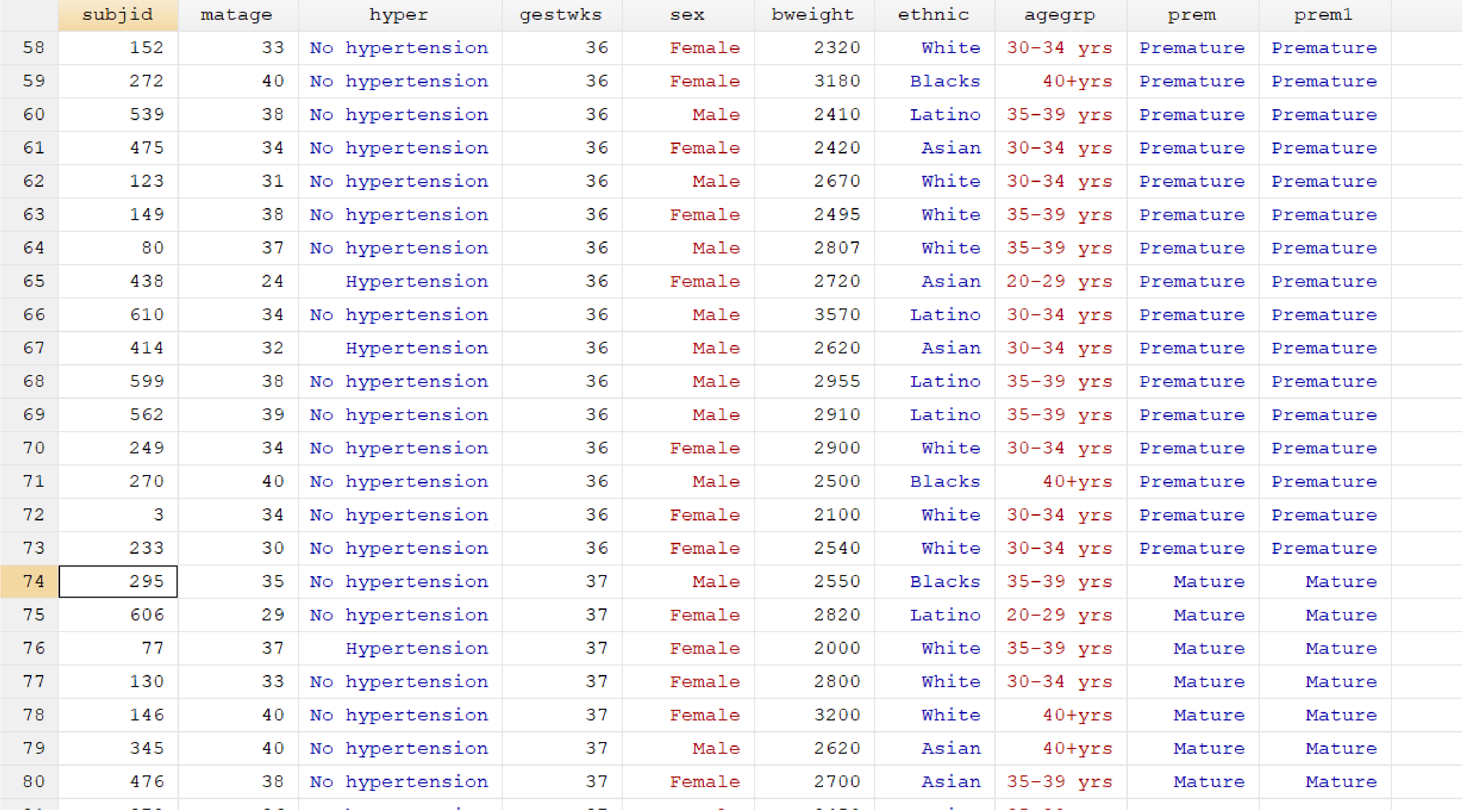
sort gestwks



	subjid	matage	hyper	gestwks	sex	bweight	ethnic	agegrp	prem	prem1
58	152	33	No hypertension	36	Female	2320	White	30-34 yrs	1	1
59	272	40	No hypertension	36	Female	3180	Blacks	40+yrs	1	1
60	539	38	No hypertension	36	Male	2410	Latino	35-39 yrs	1	1
61	475	34	No hypertension	36	Female	2420	Asian	30-34 yrs	1	1
62	123	31	No hypertension	36	Male	2670	White	30-34 yrs	1	1
63	149	38	No hypertension	36	Female	2495	White	35-39 yrs	1	1
64	80	37	No hypertension	36	Male	2807	White	35-39 yrs	1	1
65	438	24	Hypertension	36	Female	2720	Asian	20-29 yrs	1	1
66	610	34	No hypertension	36	Male	3570	Latino	30-34 yrs	1	1
67	414	32	Hypertension	36	Male	2620	Asian	30-34 yrs	1	1
68	599	38	No hypertension	36	Male	2955	Latino	35-39 yrs	1	1
69	562	39	No hypertension	36	Male	2910	Latino	35-39 yrs	1	1
70	249	34	No hypertension	36	Female	2900	White	30-34 yrs	1	1
71	270	40	No hypertension	36	Male	2500	Blacks	40+yrs	1	1
72	3	34	No hypertension	36	Female	2100	White	30-34 yrs	1	1
73	233	30	No hypertension	36	Female	2540	White	30-34 yrs	1	1
74	295	35	No hypertension	37	Male	2550	Blacks	35-39 yrs	0	0
75	606	29	No hypertension	37	Female	2820	Latino	20-29 yrs	0	0
76	77	37	Hypertension	37	Female	2000	White	35-39 yrs	0	0
77	130	33	No hypertension	37	Female	2800	White	30-34 yrs	0	0
78	146	40	No hypertension	37	Female	3200	White	40+yrs	0	0
79	345	40	No hypertension	37	Male	2620	Asian	40+yrs	0	0
80	476	38	No hypertension	37	Female	2700	Asian	35-39 yrs	0	0



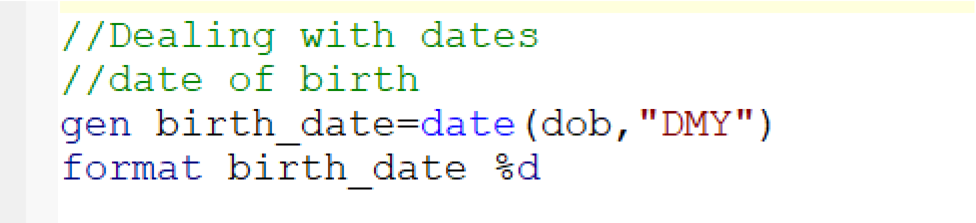
```
//Label premature
lab def lprem 0"Mature" 1"Premature"
lab val prem prem1 lprem
```



	subjid	matage	hyper	gestwks	sex	bweight	ethnic	agegrp	prem	prem1
58	152	33	No hypertension	36	Female	2320	White	30-34 yrs	Premature	Premature
59	272	40	No hypertension	36	Female	3180	Blacks	40+yrs	Premature	Premature
60	539	38	No hypertension	36	Male	2410	Latino	35-39 yrs	Premature	Premature
61	475	34	No hypertension	36	Female	2420	Asian	30-34 yrs	Premature	Premature
62	123	31	No hypertension	36	Male	2670	White	30-34 yrs	Premature	Premature
63	149	38	No hypertension	36	Female	2495	White	35-39 yrs	Premature	Premature
64	80	37	No hypertension	36	Male	2807	White	35-39 yrs	Premature	Premature
65	438	24	Hypertension	36	Female	2720	Asian	20-29 yrs	Premature	Premature
66	610	34	No hypertension	36	Male	3570	Latino	30-34 yrs	Premature	Premature
67	414	32	Hypertension	36	Male	2620	Asian	30-34 yrs	Premature	Premature
68	599	38	No hypertension	36	Male	2955	Latino	35-39 yrs	Premature	Premature
69	562	39	No hypertension	36	Male	2910	Latino	35-39 yrs	Premature	Premature
70	249	34	No hypertension	36	Female	2900	White	30-34 yrs	Premature	Premature
71	270	40	No hypertension	36	Male	2500	Blacks	40+yrs	Premature	Premature
72	3	34	No hypertension	36	Female	2100	White	30-34 yrs	Premature	Premature
73	233	30	No hypertension	36	Female	2540	White	30-34 yrs	Premature	Premature
74	295	35	No hypertension	37	Male	2550	Blacks	35-39 yrs	Mature	Mature
75	606	29	No hypertension	37	Female	2820	Latino	20-29 yrs	Mature	Mature
76	77	37	Hypertension	37	Female	2000	White	35-39 yrs	Mature	Mature
77	130	33	No hypertension	37	Female	2800	White	30-34 yrs	Mature	Mature
78	146	40	No hypertension	37	Female	3200	White	40+yrs	Mature	Mature
79	345	40	No hypertension	37	Male	2620	Asian	40+yrs	Mature	Mature
80	476	38	No hypertension	37	Female	2700	Asian	35-39 yrs	Mature	Mature

Dates

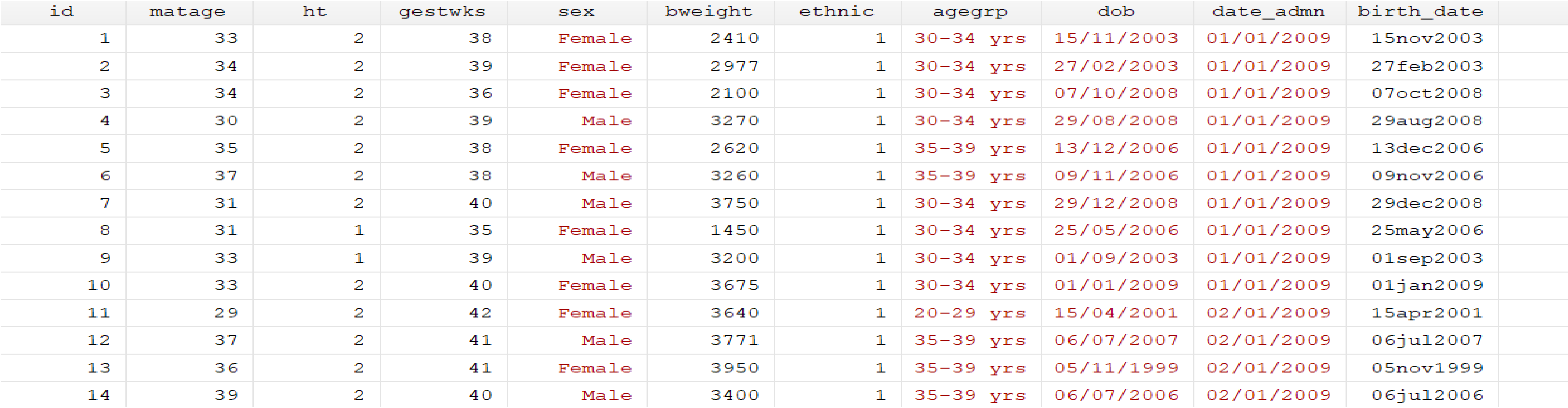
- Dates in stata appear as numbers since 01jan1960. To create a data variable, you first define a “date” variable and provide the correct format



```
//Dealing with dates
//date of birth
gen birth_date=date(dob,"DMY")
format birth_date %d
```

Date as string variable

Date variable

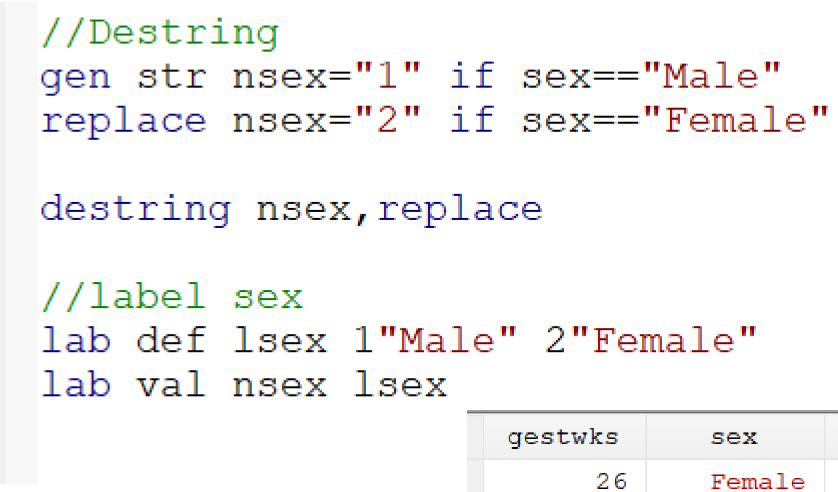


id	matage	ht	gestwks	sex	bweight	ethnic	agegrp	dob	date_admn	birth_date
1	33	2	38	Female	2410	1	30-34 yrs	15/11/2003	01/01/2009	15nov2003
2	34	2	39	Female	2977	1	30-34 yrs	27/02/2003	01/01/2009	27feb2003
3	34	2	36	Female	2100	1	30-34 yrs	07/10/2008	01/01/2009	07oct2008
4	30	2	39	Male	3270	1	30-34 yrs	29/08/2008	01/01/2009	29aug2008
5	35	2	38	Female	2620	1	35-39 yrs	13/12/2006	01/01/2009	13dec2006
6	37	2	38	Male	3260	1	35-39 yrs	09/11/2006	01/01/2009	09nov2006
7	31	2	40	Male	3750	1	30-34 yrs	29/12/2008	01/01/2009	29dec2008
8	31	1	35	Female	1450	1	30-34 yrs	25/05/2006	01/01/2009	25may2006
9	33	1	39	Male	3200	1	30-34 yrs	01/09/2003	01/01/2009	01sep2003
10	33	2	40	Female	3675	1	30-34 yrs	01/01/2009	01/01/2009	01jan2009
11	29	2	42	Female	3640	1	20-29 yrs	15/04/2001	02/01/2009	15apr2001
12	37	2	41	Male	3771	1	35-39 yrs	06/07/2007	02/01/2009	06jul2007
13	36	2	41	Female	3950	1	35-39 yrs	05/11/1999	02/01/2009	05nov1999
14	39	2	40	Male	3400	1	35-39 yrs	06/07/2006	02/01/2009	06jul2006

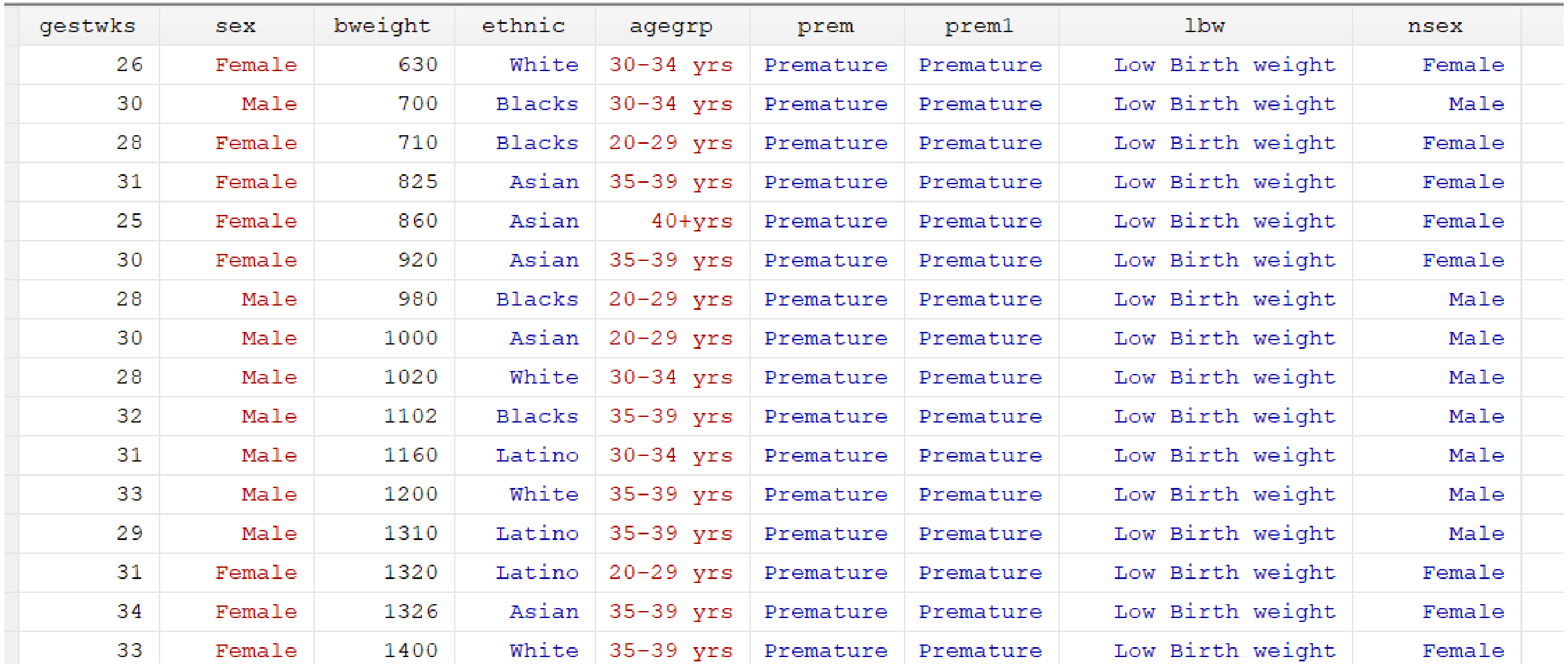
Exercise 2

- Read the birth_weight.csv to stata
- Use the *bweight* variable to create a new variable
- lbw for low birth weight code as
- 0 for birthweight >=2500
- 1 for birthweight <2500
- Label the new variable as “Low birth weight”
- Attach the variables: 0 “Normal birth weight” 1”Low birth weight”
- Create a date variable for date_admn

Change variables type



```
//Destring
gen str nsex="1" if sex=="Male"
replace nsex="2" if sex=="Female"

destring nsex,replace

//label sex
lab def lsex 1"Male" 2"Female"
lab val nsex lsex
```

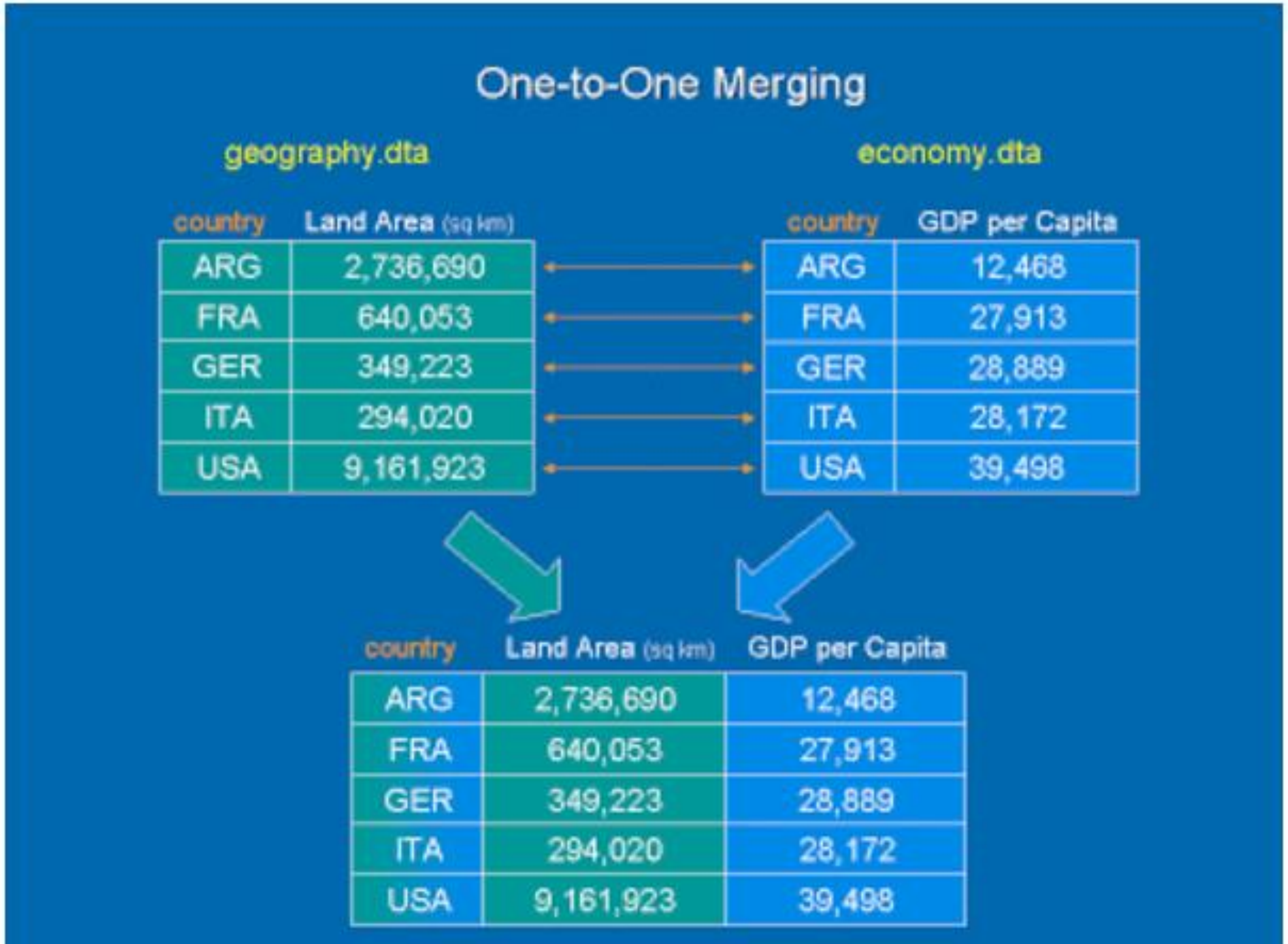


gestwks	sex	bweight	ethnic	agegrp	prem	prem1	lbw	nsex
26	Female	630	White	30-34 yrs	Premature	Premature	Low Birth weight	Female
30	Male	700	Blacks	30-34 yrs	Premature	Premature	Low Birth weight	Male
28	Female	710	Blacks	20-29 yrs	Premature	Premature	Low Birth weight	Female
31	Female	825	Asian	35-39 yrs	Premature	Premature	Low Birth weight	Female
25	Female	860	Asian	40+yrs	Premature	Premature	Low Birth weight	Female
30	Female	920	Asian	35-39 yrs	Premature	Premature	Low Birth weight	Female
28	Male	980	Blacks	20-29 yrs	Premature	Premature	Low Birth weight	Male
30	Male	1000	Asian	20-29 yrs	Premature	Premature	Low Birth weight	Male
28	Male	1020	White	30-34 yrs	Premature	Premature	Low Birth weight	Male
32	Male	1102	Blacks	35-39 yrs	Premature	Premature	Low Birth weight	Male
31	Male	1160	Latino	30-34 yrs	Premature	Premature	Low Birth weight	Male
33	Male	1200	White	35-39 yrs	Premature	Premature	Low Birth weight	Male
29	Male	1310	Latino	35-39 yrs	Premature	Premature	Low Birth weight	Male
31	Female	1320	Latino	20-29 yrs	Premature	Premature	Low Birth weight	Female
34	Female	1326	Asian	35-39 yrs	Premature	Premature	Low Birth weight	Female
33	Female	1400	White	35-39 yrs	Premature	Premature	Low Birth weight	Female

Merge datasets

- Merge – adds variables to a dataset. Type help merge for details. Merging two datasets require that both have at least one variable in common (either string or numeric). If string make sure the categories have the same spelling (i.e. country names, etc.). The common variables must have the same name. Explore each dataset separately before merging. Make sure to use all possible common variables (for example, if merging two panel datasets you will need country and years).

One to one merging



One-to-One Merging

geography.dta

country	Land Area (sq km)
ARG	2,736,690
FRA	640,053
GER	349,223
ITA	294,020
USA	9,161,923

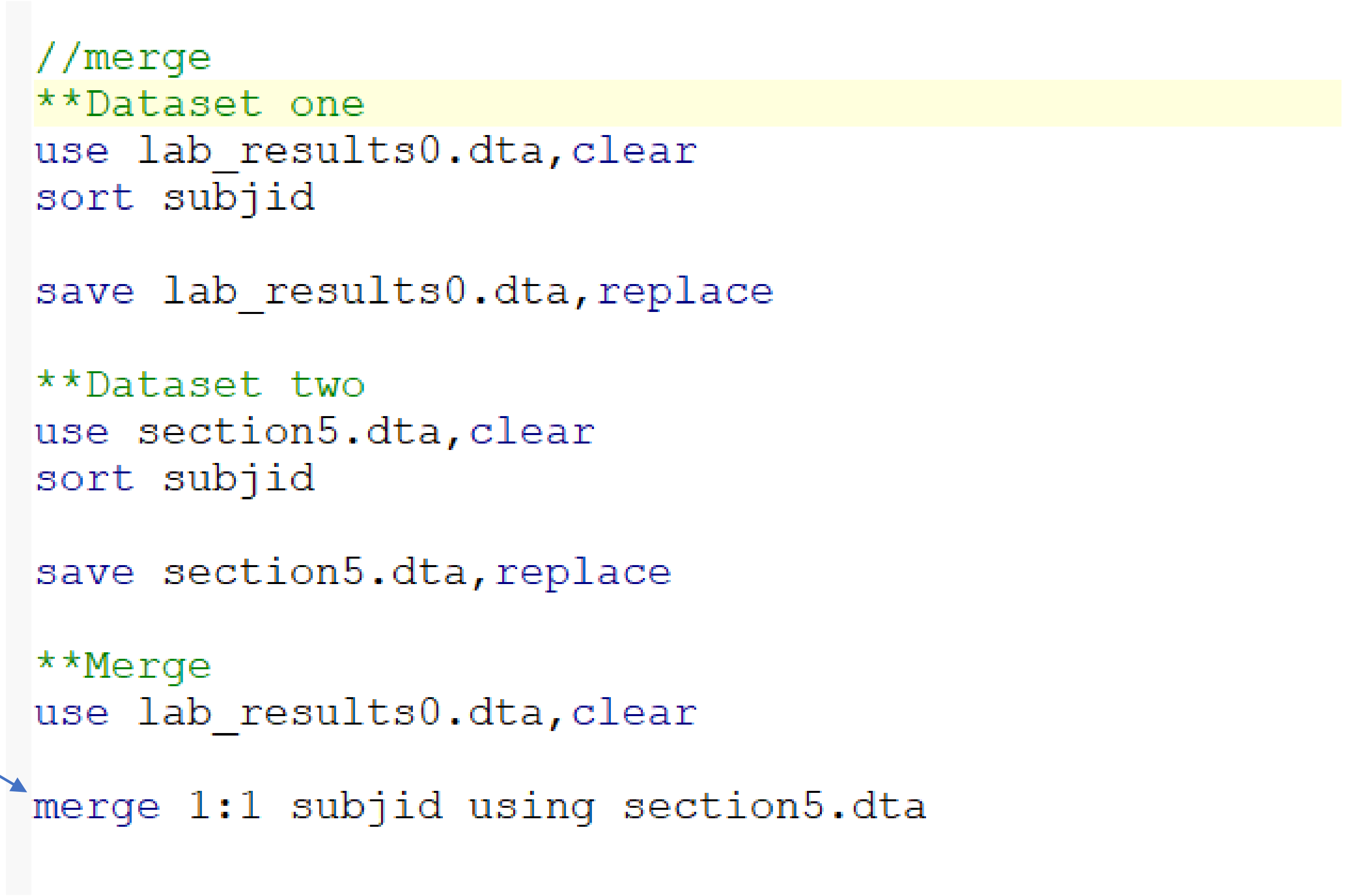
economy.dta

country	GDP per Capita
ARG	12,468
FRA	27,913
GER	28,889
ITA	28,172
USA	39,498

country	Land Area (sq km)	GDP per Capita
ARG	2,736,690	12,468
FRA	640,053	27,913
GER	349,223	28,889
ITA	294,020	28,172
USA	9,161,923	39,498

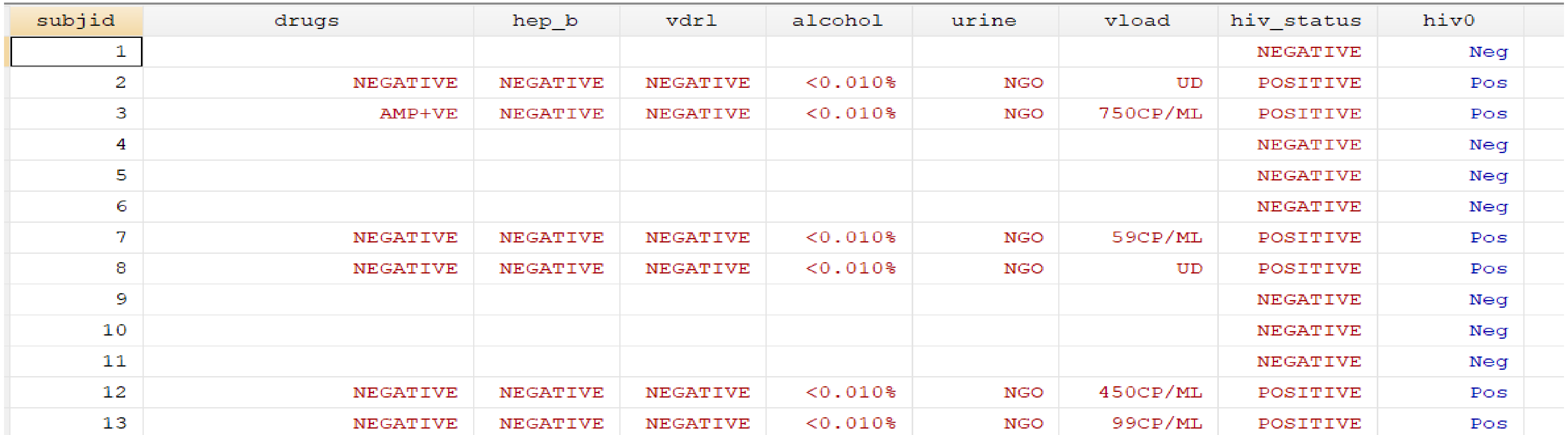
Use the *lab_results0.dta* data,

Sort the *subjid* variable



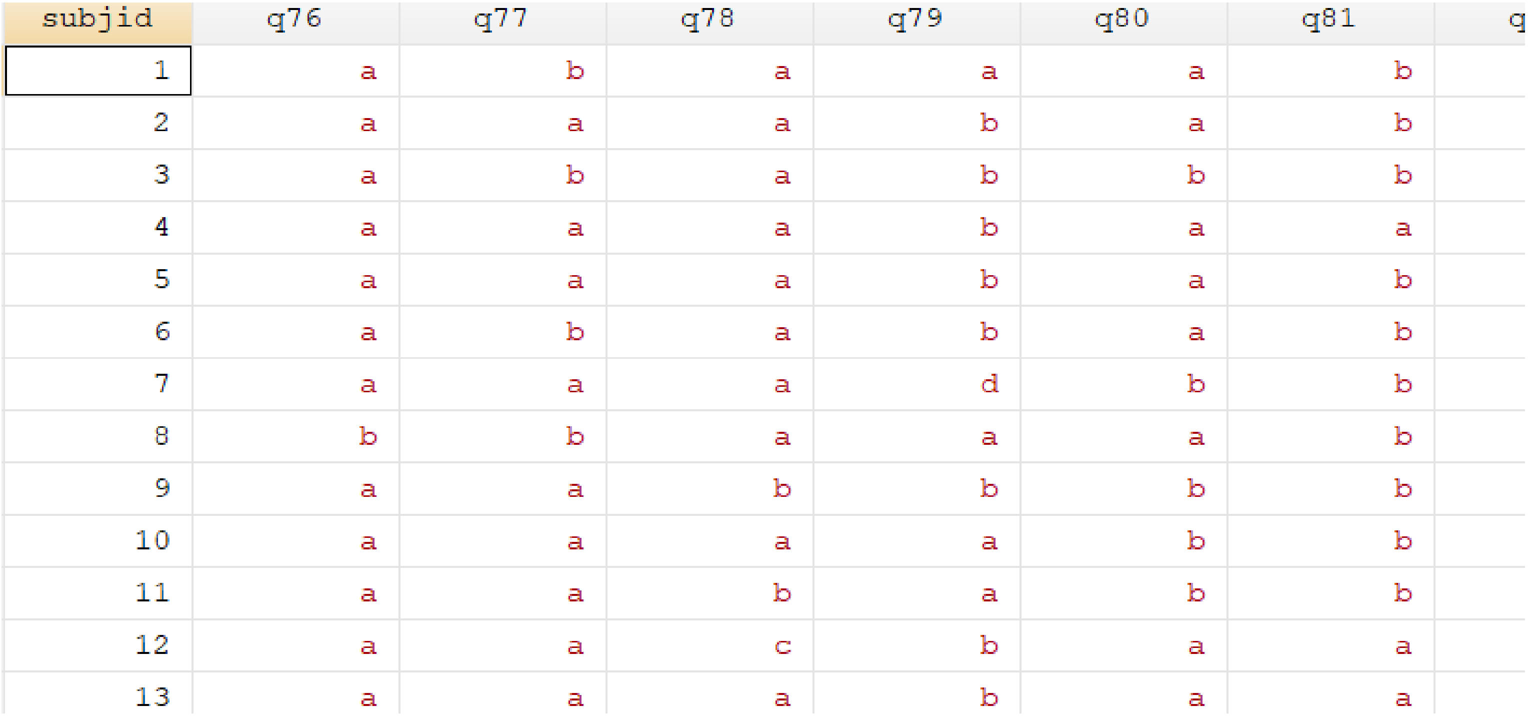
```
//merge
**Dataset one
use lab_results0.dta,clear
sort subjid

save lab_results0.dta,replace

**Dataset two
use section5.dta,clear
sort subjid

save section5.dta,replace

**Merge
use lab_results0.dta,clear

merge 1:1 subjid using section5.dta
```

1:1 means each data has one record
Section5.dta the data to merge with



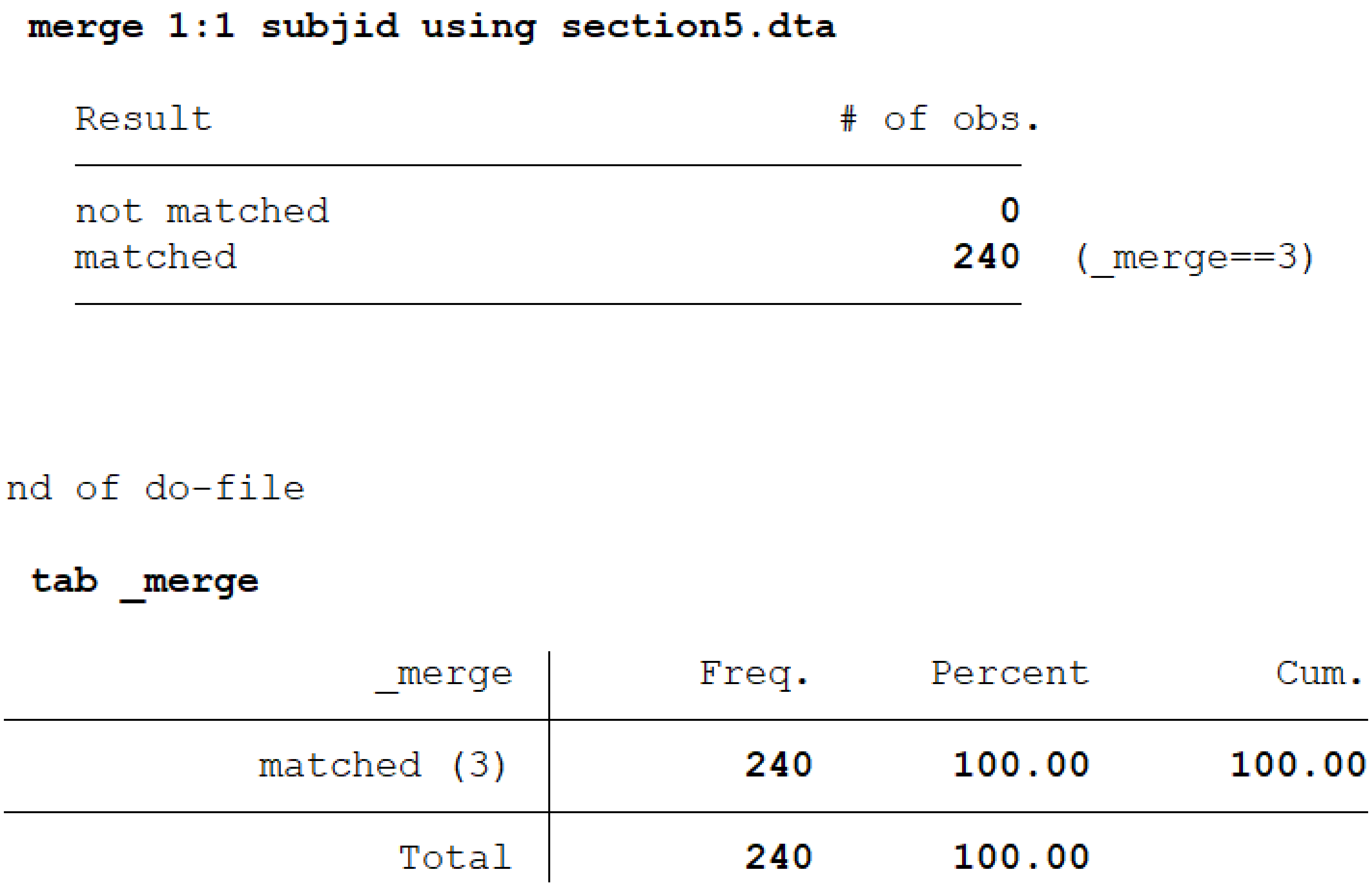
subjid	drugs	hep_b	vdrl	alcohol	urine	vload	hiv_status	hiv0
1							NEGATIVE	Neg
2	NEGATIVE	NEGATIVE	NEGATIVE	<0.010%	NGO	UD	POSITIVE	Pos
3	AMP+VE	NEGATIVE	NEGATIVE	<0.010%	NGO	750CP/ML	POSITIVE	Pos
4							NEGATIVE	Neg
5							NEGATIVE	Neg
6							NEGATIVE	Neg
7	NEGATIVE	NEGATIVE	NEGATIVE	<0.010%	NGO	59CP/ML	POSITIVE	Pos
8	NEGATIVE	NEGATIVE	NEGATIVE	<0.010%	NGO	UD	POSITIVE	Pos
9							NEGATIVE	Neg
10							NEGATIVE	Neg
11							NEGATIVE	Neg
12	NEGATIVE	NEGATIVE	NEGATIVE	<0.010%	NGO	450CP/ML	POSITIVE	Pos
13	NEGATIVE	NEGATIVE	NEGATIVE	<0.010%	NGO	99CP/ML	POSITIVE	Pos

lab_results0.dta



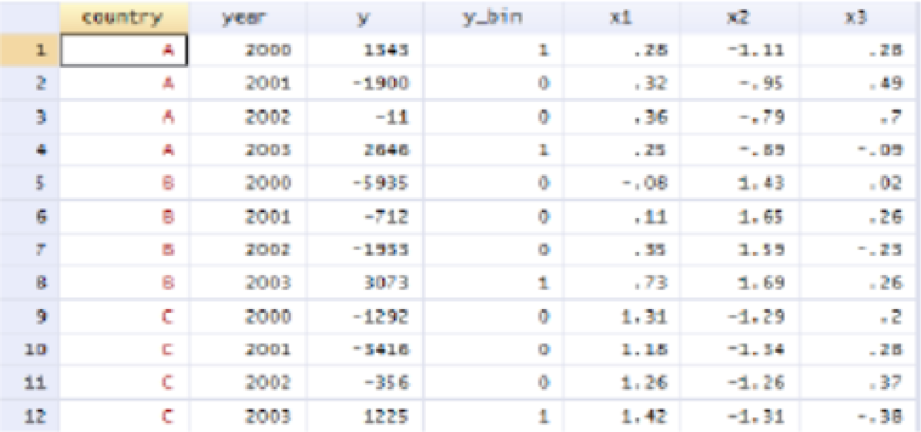
subjid	q76	q77	q78	q79	q80	q81	q
1	a	b	a	a	a	b	
2	a	a	a	b	a	b	
3	a	b	a	b	b	b	
4	a	a	a	b	a	a	
5	a	a	a	b	a	b	
6	a	b	a	b	a	b	
7	a	a	a	d	b	b	
8	b	b	a	a	a	b	
9	a	a	b	b	b	b	
10	a	a	a	a	b	b	
11	a	a	b	a	b	b	
12	a	a	c	b	a	a	
13	a	a	a	b	a	a	

section5.dta



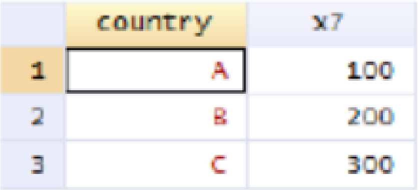
```
 merge 1:1 subjid using section5.dta

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                               240  (_merge==3)
    -----------------------------------------


nd of do-file

 tab _merge

                 _merge |      Freq.     Percent        Cum.
------------------------+-----------------------------------
            matched (3) |        240      100.00      100.00
------------------------+-----------------------------------
                  Total |        240      100.00
```

mydata1



	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	A	2002	-11	0	.36	-.79	.7
4	A	2003	2646	1	.25	-.89	-.09
5	B	2000	-5935	0	-.08	1.43	.02
6	B	2001	-712	0	.11	1.65	.26
7	B	2002	-1933	0	.35	1.59	-.23
8	B	2003	3073	1	.73	1.69	.26
9	C	2000	-1292	0	1.31	-1.29	.2
10	C	2001	-3416	0	1.18	-1.34	.28
11	C	2002	-356	0	1.26	-1.26	.37
12	C	2003	1225	1	1.42	-1.31	-.38

mydata4



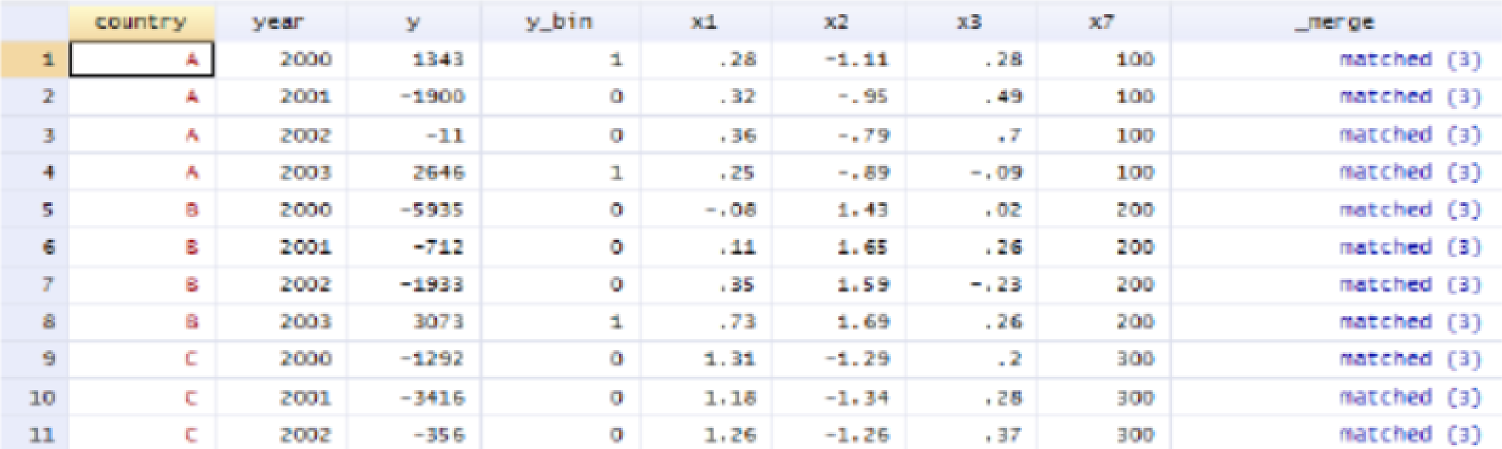
	country	x7
1	A	100
2	B	200
3	C	300

```
merge m:1 country using mydata4

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                                12  (_merge==3)
    -----------------------------------------
```

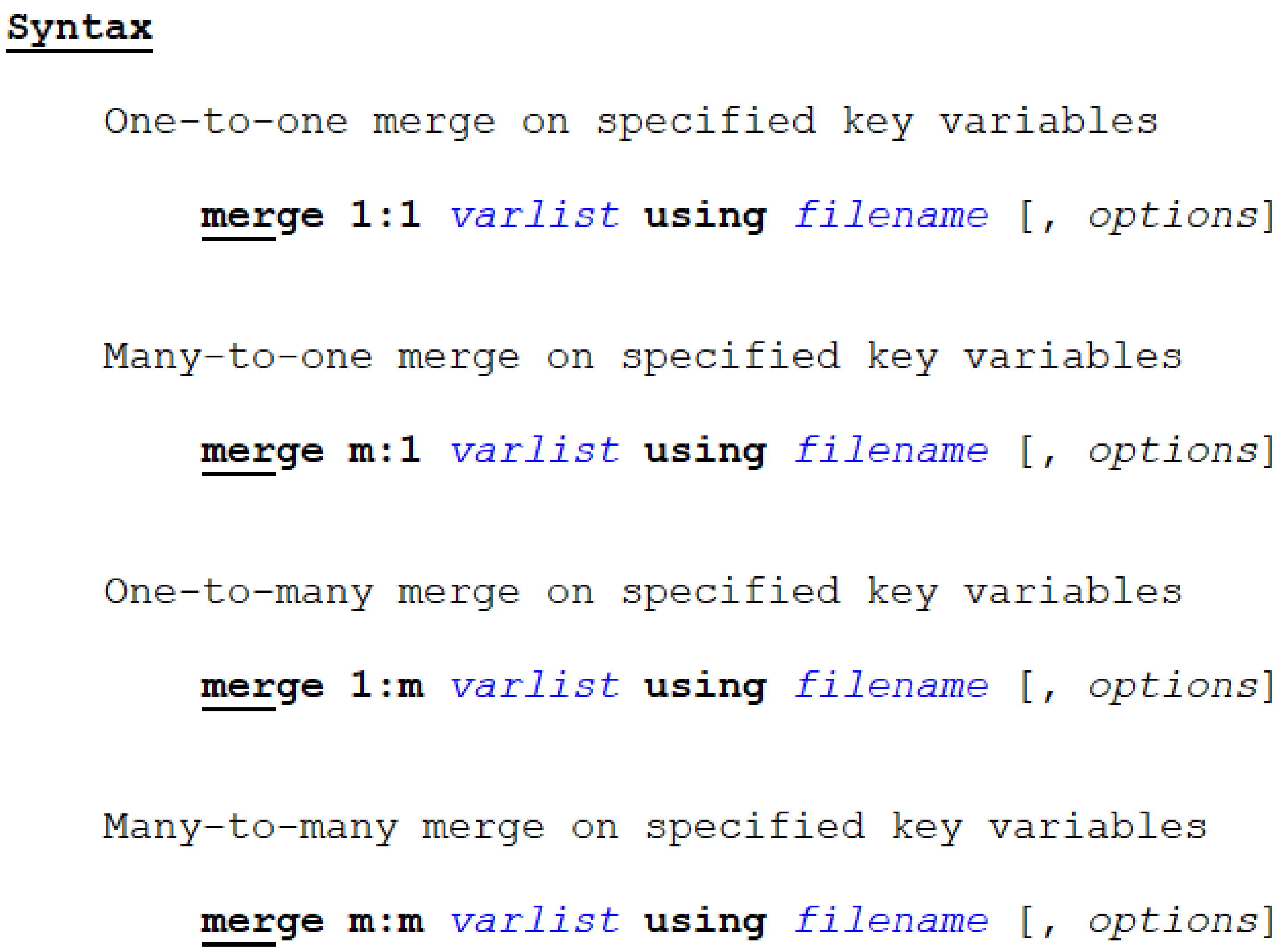
- Make sure one dataset is loaded into Stata (in this case mydata1), then use merge.
- Make sure to map where the using data is located (in this case mydata2, for example “c:\folders\data\mydata4.dta”)*.

NOTE: For Stata 10 or older:

1) Remove the m:1
2) Sort both datasets by all the ids and save before merging



	country	year	y	y_bin	x1	x2	x3	x7	_merge
1	A	2000	1343	1	.28	-1.11	.28	100	matched (3)
2	A	2001	-1900	0	.32	-.95	.49	100	matched (3)
3	A	2002	-11	0	.36	-.79	.7	100	matched (3)
4	A	2003	2646	1	.25	-.89	-.09	100	matched (3)
5	B	2000	-5935	0	-.08	1.43	.02	200	matched (3)
6	B	2001	-712	0	.11	1.65	.26	200	matched (3)
7	B	2002	-1933	0	.35	1.59	-.23	200	matched (3)
8	B	2003	3073	1	.73	1.69	.26	200	matched (3)
9	C	2000	-1292	0	1.31	-1.29	.2	300	matched (3)
10	C	2001	-3416	0	1.18	-1.34	.28	300	matched (3)
11	C	2002	-356	0	1.26	-1.26	.37	300	matched (3)



Syntax

One-to-one merge on specified key variables

```
merge 1:1 varlist using filename [, options]
```

Many-to-one merge on specified key variables

```
merge m:1 varlist using filename [, options]
```

One-to-many merge on specified key variables

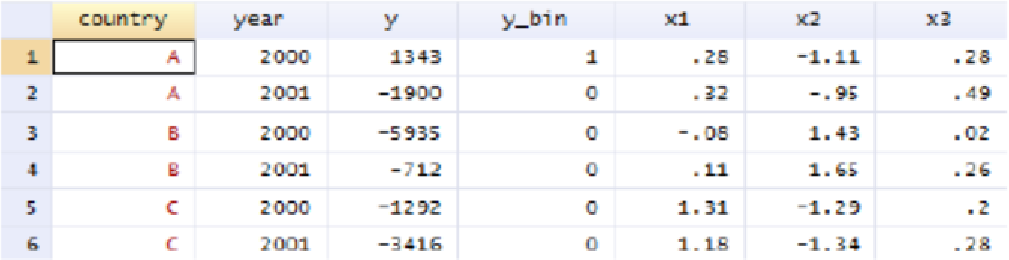
```
merge 1:m varlist using filename [, options]
```

Many-to-many merge on specified key variables

```
merge m:m varlist using filename [, options]
```

APPEND

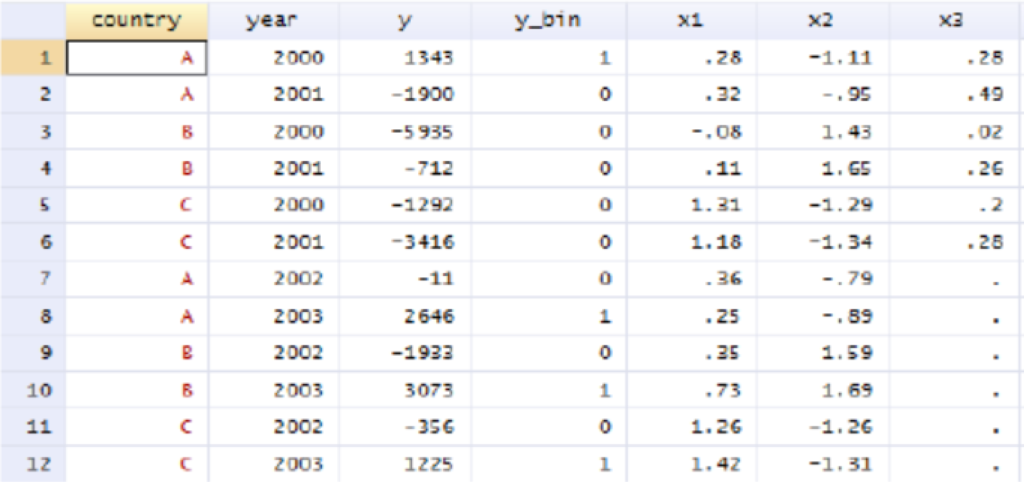
mydata7



	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	B	2000	-5935	0	-.08	1.43	.02
4	B	2001	-712	0	.11	1.65	.26
5	C	2000	-1292	0	1.31	-1.29	.2
6	C	2001	-3416	0	1.18	-1.34	.28

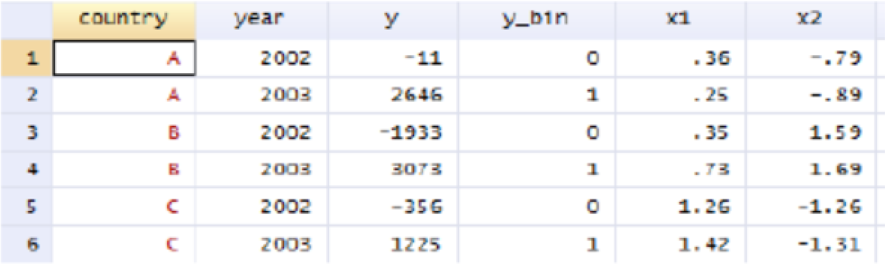
- Make sure one dataset is loaded into Stata (in this case mydata7), then use `append`.
- Make sure to map where the using data is located (in this case mydata2, for example "c:\folders\data\mydata9.dta")*.
- Notice the missing data.

append using mydata9



	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	.28	-1.11	.28
2	A	2001	-1900	0	.32	-.95	.49
3	B	2000	-5935	0	-.08	1.43	.02
4	B	2001	-712	0	.11	1.65	.26
5	C	2000	-1292	0	1.31	-1.29	.2
6	C	2001	-3416	0	1.18	-1.34	.28
7	A	2002	-11	0	.36	-.79	.
8	A	2003	2646	1	.25	-.89	.
9	B	2002	-1933	0	.35	1.59	.
10	B	2003	3073	1	.73	1.69	.
11	C	2002	-356	0	1.26	-1.26	.
12	C	2003	1225	1	1.42	-1.31	.

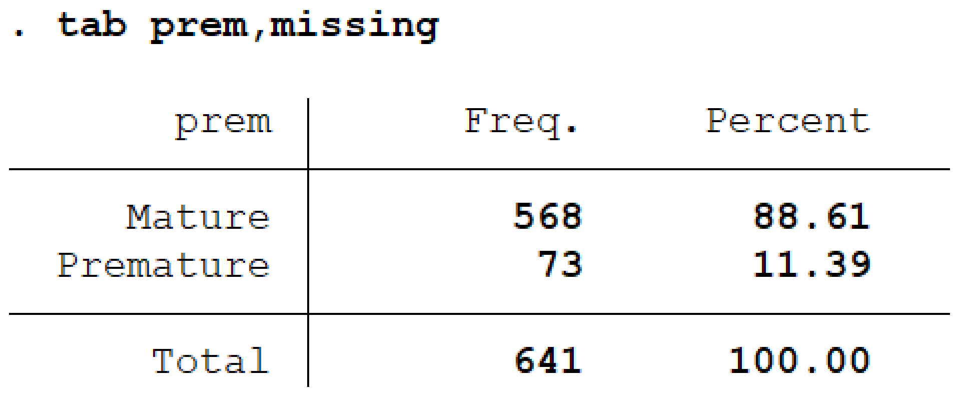
mydata9



	country	year	y	y_bin	x1	x2
1	A	2002	-11	0	.36	-.79
2	A	2003	2646	1	.25	-.89
3	B	2002	-1933	0	.35	1.59
4	B	2003	3073	1	.73	1.69
5	C	2002	-356	0	1.26	-1.26
6	C	2003	1225	1	1.42	-1.31

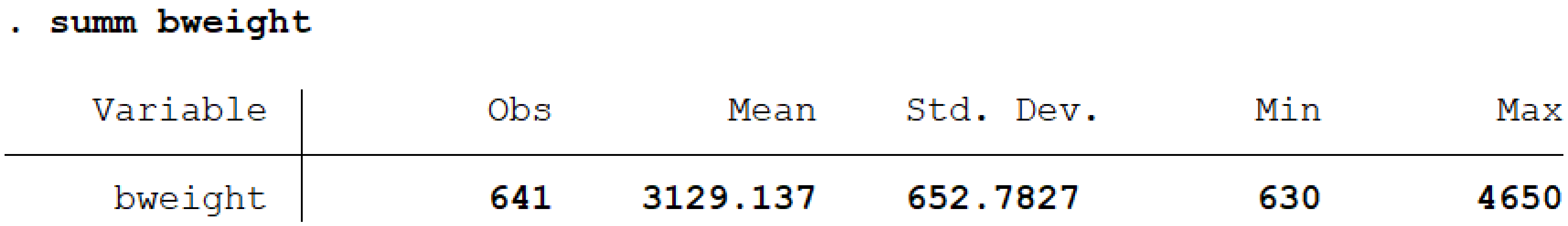
Summary in stata

- Use tabulate to get summary of categorical variables



```
. tab prem,missing

       prem |      Freq.     Percent
------------+-----------------------
     Mature |        568       88.61
  Premature |         73       11.39
------------+-----------------------
      Total |        641      100.00
```

- Use summarize to get summary of continuous variables



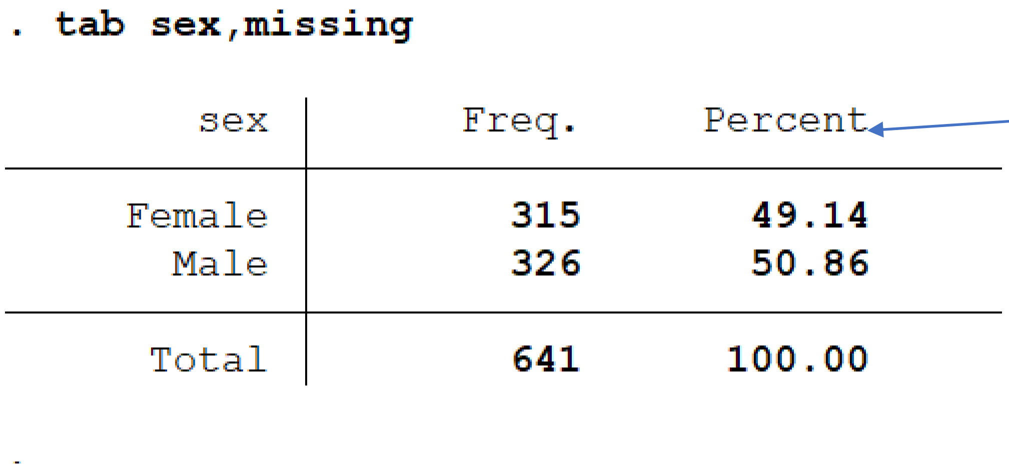
```
. summ bweight

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     bweight |        641    3129.137    652.7827        630       4650
```

Exploring data

Frequency refers to the number of times a value is repeated. Frequencies are used to analyze categorical data. The tables below are *frequency tables*, values are in ascending order. In Stata use the command **`tab`** ***`varname`*****.**

`Read the birth_weight.csv`

`What proportion of children were females?`



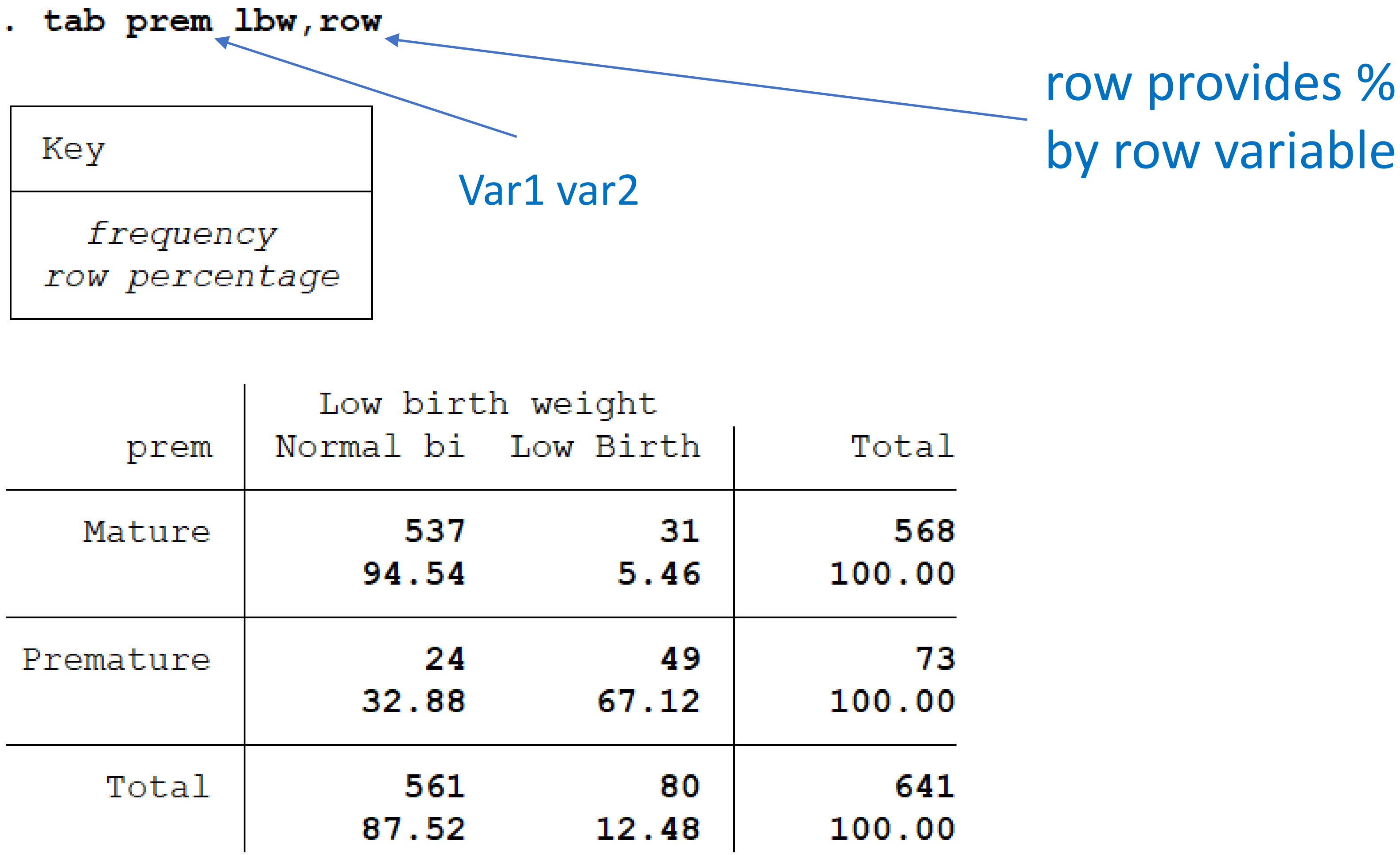
```
. tab sex,missing

        sex |      Freq.     Percent
------------+-----------------------
     Female |        315       49.14
       Male |        326       50.86
------------+-----------------------
      Total |        641      100.00

.
```

Freq. provides the count

Percent provides relative frequency

Contingency tables

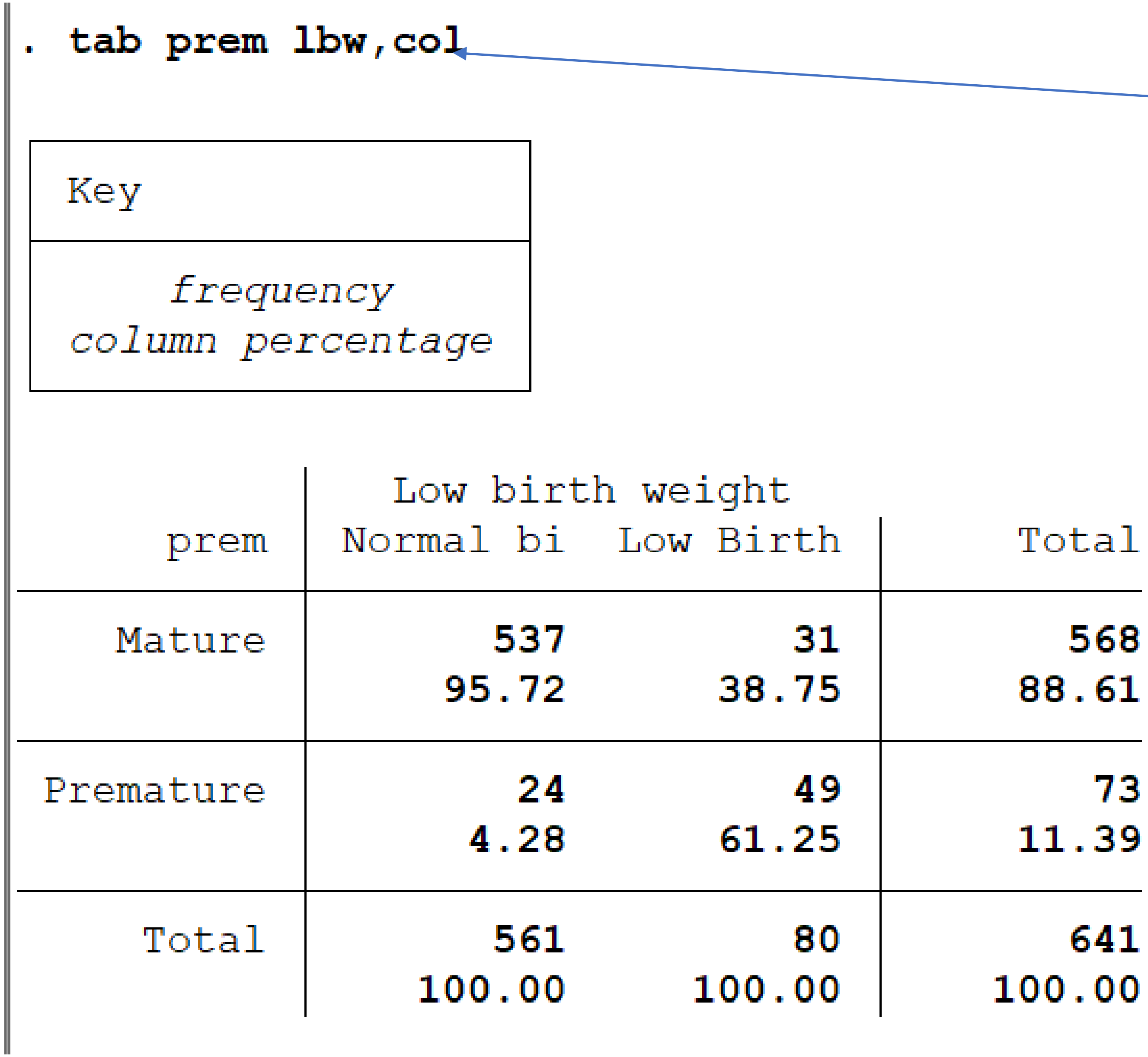
Contingency tables or crosstabs help you to analyze the relationship between two or more categorical variables.



```
. tab prem lbw,row

+--------------------+
| Key                |
|--------------------|
|     frequency      |
|   row percentage   |
+--------------------+

           |   Low birth weight
      prem | Normal bi  Low Birth |     Total
-----------+----------------------+----------
    Mature |       537         31 |       568
           |     94.54       5.46 |    100.00
-----------+----------------------+----------
 Premature |        24         49 |        73
           |     32.88      67.12 |    100.00
-----------+----------------------+----------
     Total |       561         80 |       641
           |     87.52      12.48 |    100.00
```

Var1 var2

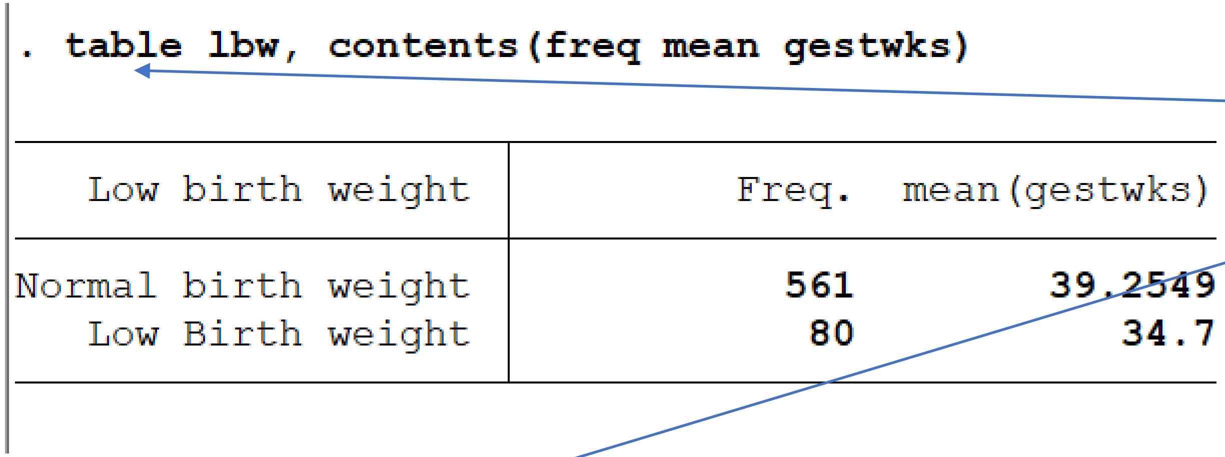
row provides %
by row variable



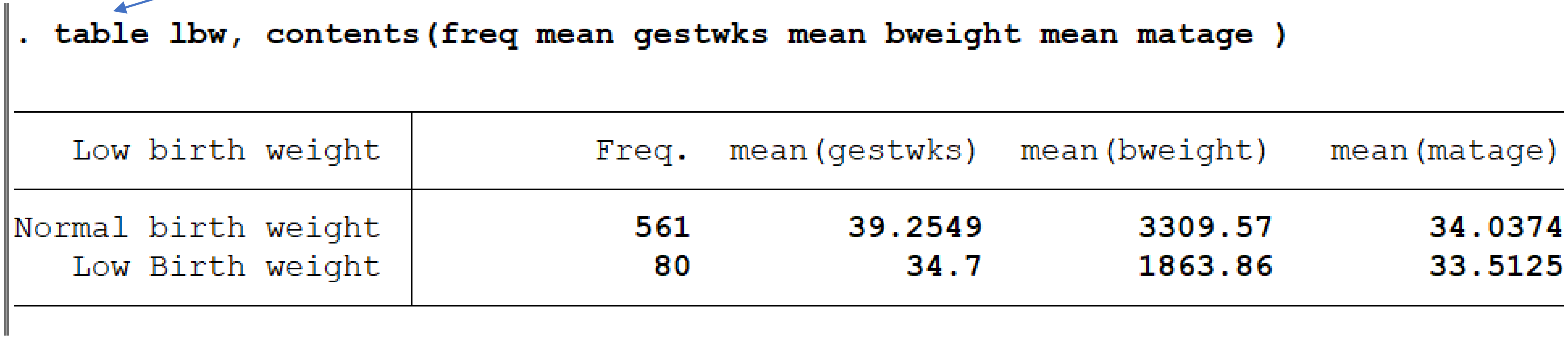
```
. tab prem lbw,col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |   Low birth weight
      prem | Normal bi  Low Birth |     Total
-----------+----------------------+----------
    Mature |       537         31 |       568 
           |     95.72      38.75 |     88.61 
-----------+----------------------+----------
 Premature |        24         49 |        73 
           |      4.28      61.25 |     11.39 
-----------+----------------------+----------
     Total |       561         80 |       641 
           |    100.00     100.00 |    100.00 
```

col provides the % by the column variable

Exploring data: frequencies and descriptive statistics



```
. table lbw, contents(freq mean gestwks)

--------------------------------------------------
   Low birth weight |       Freq.   mean(gestwks)
--------------------+-----------------------------
Normal birth weight |         561         39.2549
   Low Birth weight |          80            34.7
--------------------------------------------------
```

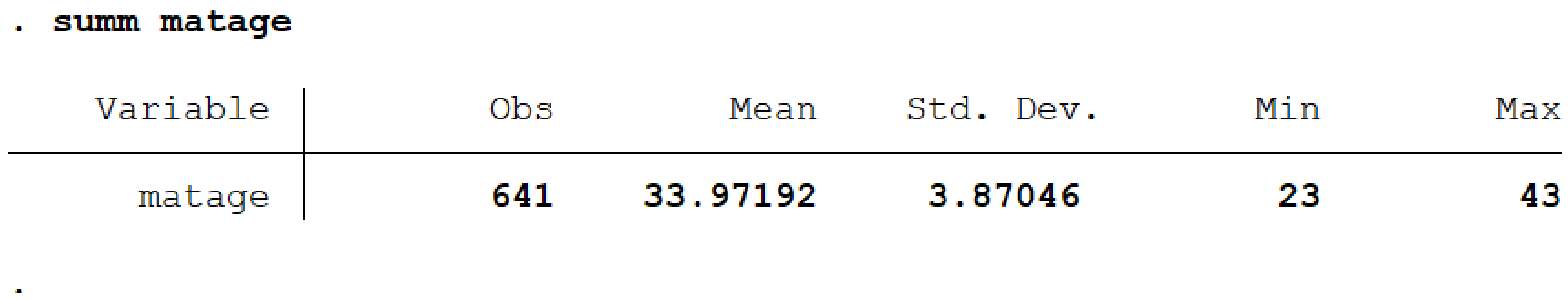


```
. table lbw, contents(freq mean gestwks mean bweight mean matage )

------------------------------------------------------------------------------
   Low birth weight |       Freq.   mean(gestwks)   mean(bweight)   mean(matage)
--------------------+---------------------------------------------------------
Normal birth weight |         561         39.2549         3309.57        34.0374
   Low Birth weight |          80            34.7         1863.86        33.5125
------------------------------------------------------------------------------
```

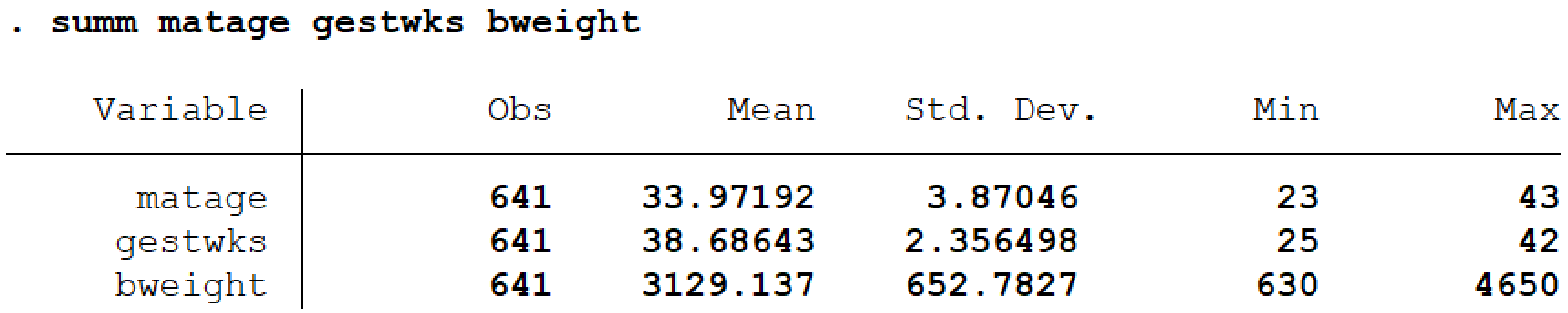
Command `table` produces frequencies and descriptive statistics per category

summarize

Type `summarize` to get some basic descriptive statistics.



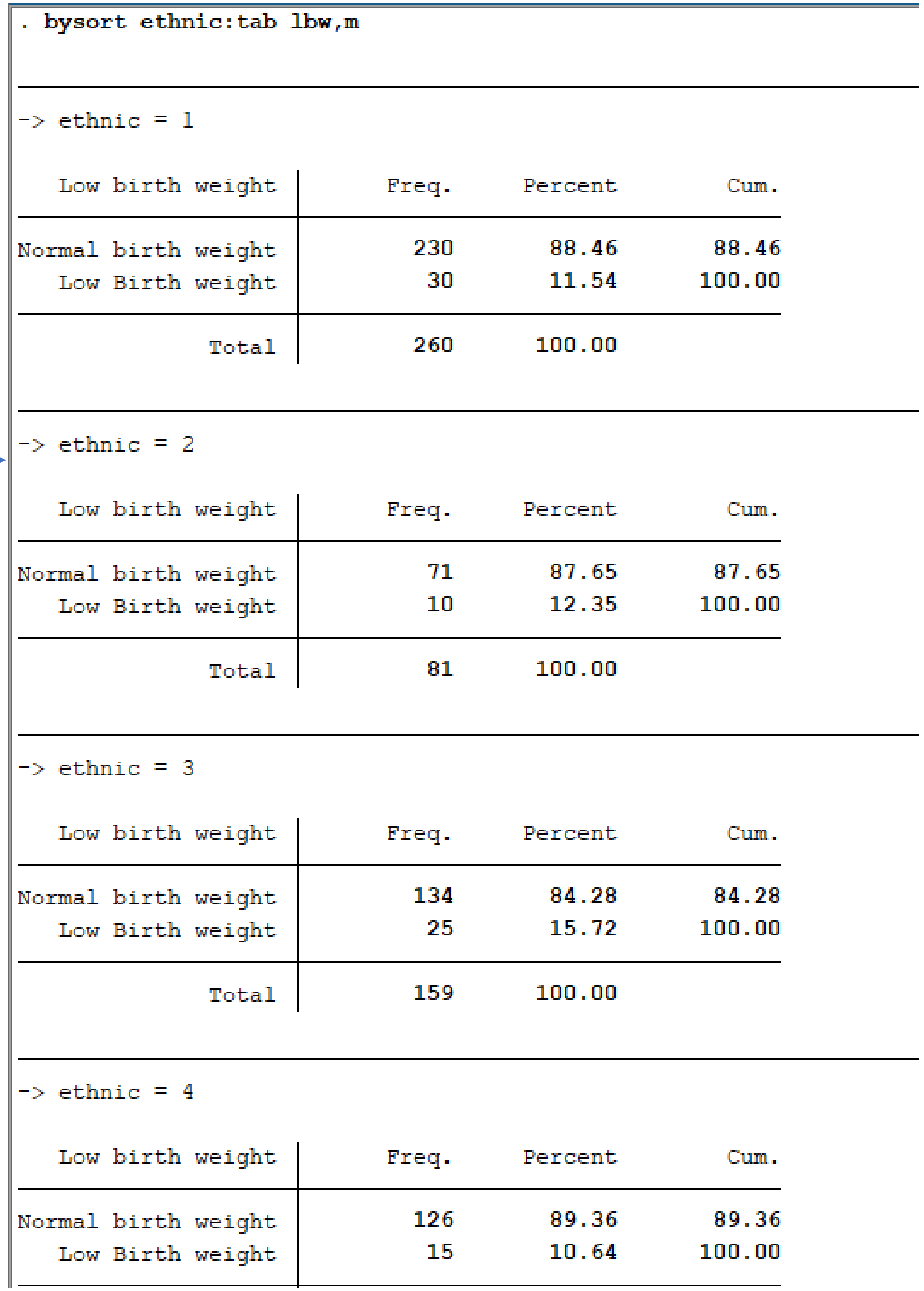
```
. summm matage

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      matage |        641    33.97192     3.87046         23         43

.
```



```
. summm matage gestwks bweight

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      matage |        641    33.97192     3.87046         23         43
     gestwks |        641    38.68643    2.356498         25         42
     bweight |        641    3129.137    652.7827        630       4650
```

bysort

bysort allows to run simple loop by a variable with many categories

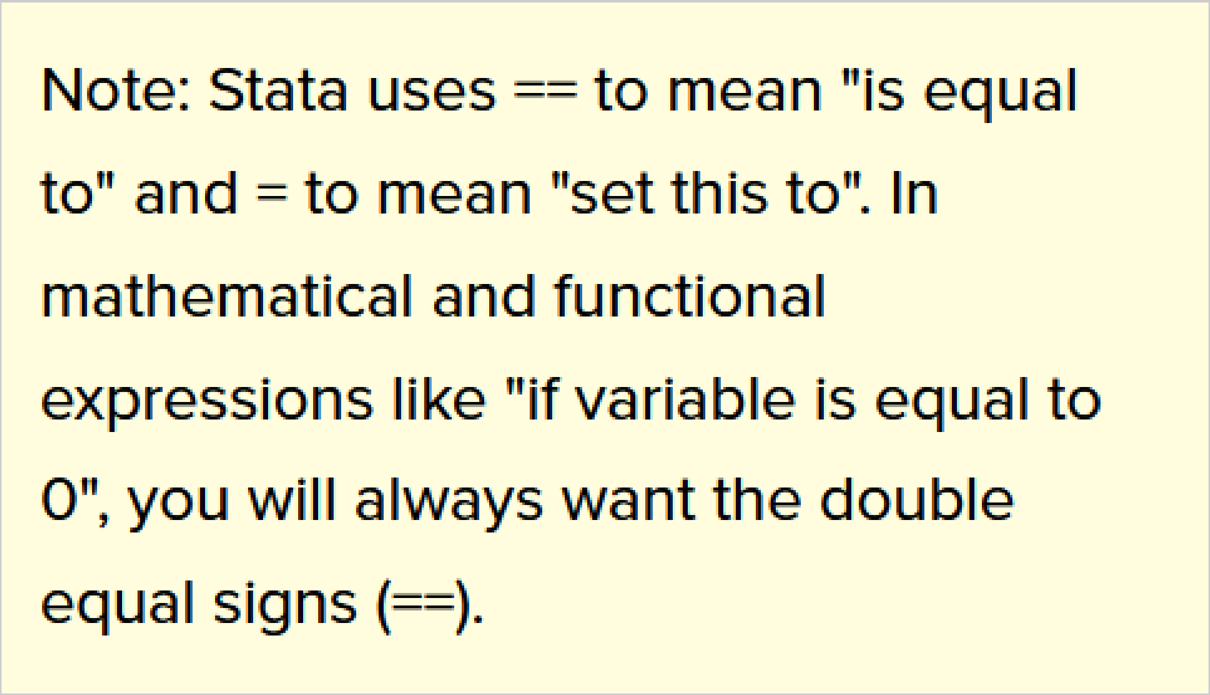


```
. bysort ethnic:tab lbw,m

-----------------------------------------------------------------------

-> ethnic = 1

   Low birth weight |      Freq.     Percent        Cum.
--------------------+-----------------------------------
Normal birth weight |        230       88.46       88.46
   Low Birth weight |         30       11.54      100.00
--------------------+-----------------------------------
              Total |        260      100.00

-----------------------------------------------------------------------

-> ethnic = 2

   Low birth weight |      Freq.     Percent        Cum.
--------------------+-----------------------------------
Normal birth weight |         71       87.65       87.65
   Low Birth weight |         10       12.35      100.00
--------------------+-----------------------------------
              Total |         81      100.00

-----------------------------------------------------------------------

-> ethnic = 3

   Low birth weight |      Freq.     Percent        Cum.
--------------------+-----------------------------------
Normal birth weight |        134       84.28       84.28
   Low Birth weight |         25       15.72      100.00
--------------------+-----------------------------------
              Total |        159      100.00

-----------------------------------------------------------------------

-> ethnic = 4

   Low birth weight |      Freq.     Percent        Cum.
--------------------+-----------------------------------
Normal birth weight |        126       89.36       89.36
   Low Birth weight |         15       10.64      100.00
--------------------+-----------------------------------
```

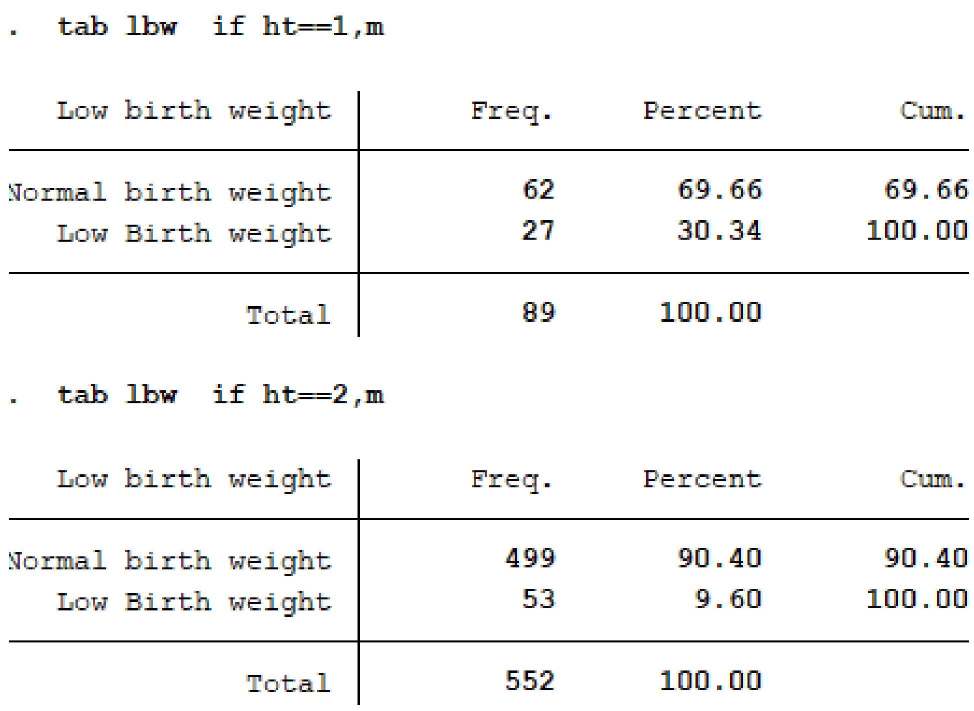
Conditional statement in stata

- **The "if" Suffix**

The “if” command suffix is used to restrict on which data a command is run.

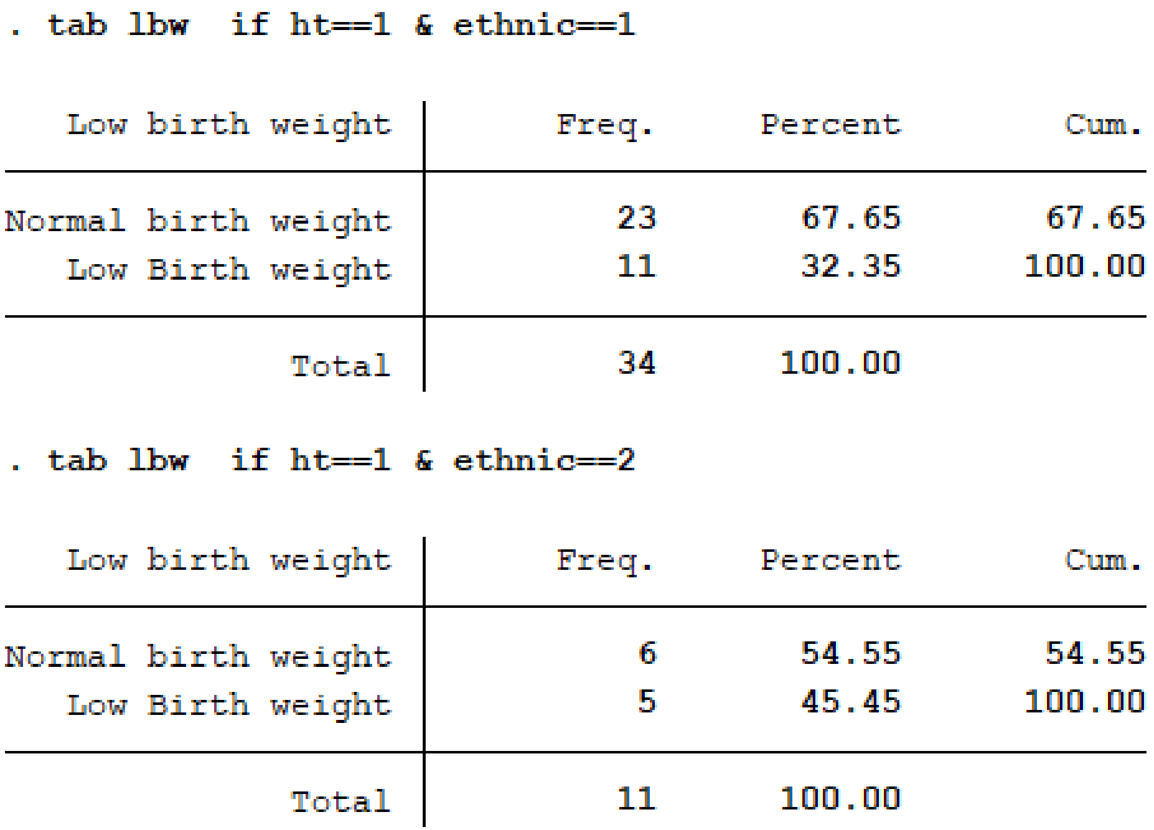


Note: Stata uses == to mean "is equal to" and = to mean "set this to". In mathematical and functional expressions like "if variable is equal to 0", you will always want the double equal signs (==).



```
. tab lbw  if ht==1,m

   Low birth weight |      Freq.     Percent        Cum.
--------------------+-----------------------------------
Normal birth weight |         62       69.66       69.66
   Low Birth weight |         27       30.34      100.00
--------------------+-----------------------------------
              Total |         89      100.00

. tab lbw  if ht==2,m

   Low birth weight |      Freq.     Percent        Cum.
--------------------+-----------------------------------
Normal birth weight |        499       90.40       90.40
   Low Birth weight |         53        9.60      100.00
--------------------+-----------------------------------
              Total |        552      100.00
```

You can combine two conditions: if and &



```
. tab lbw  if ht==1 & ethnic==1

   Low birth weight |      Freq.     Percent        Cum.
--------------------+-----------------------------------
Normal birth weight |         23       67.65       67.65
   Low Birth weight |         11       32.35      100.00
--------------------+-----------------------------------
              Total |         34      100.00

. tab lbw  if ht==1 & ethnic==2

   Low birth weight |      Freq.     Percent        Cum.
--------------------+-----------------------------------
Normal birth weight |          6       54.55       54.55
   Low Birth weight |          5       45.45      100.00
--------------------+-----------------------------------
              Total |         11      100.00
```

Exercise 3

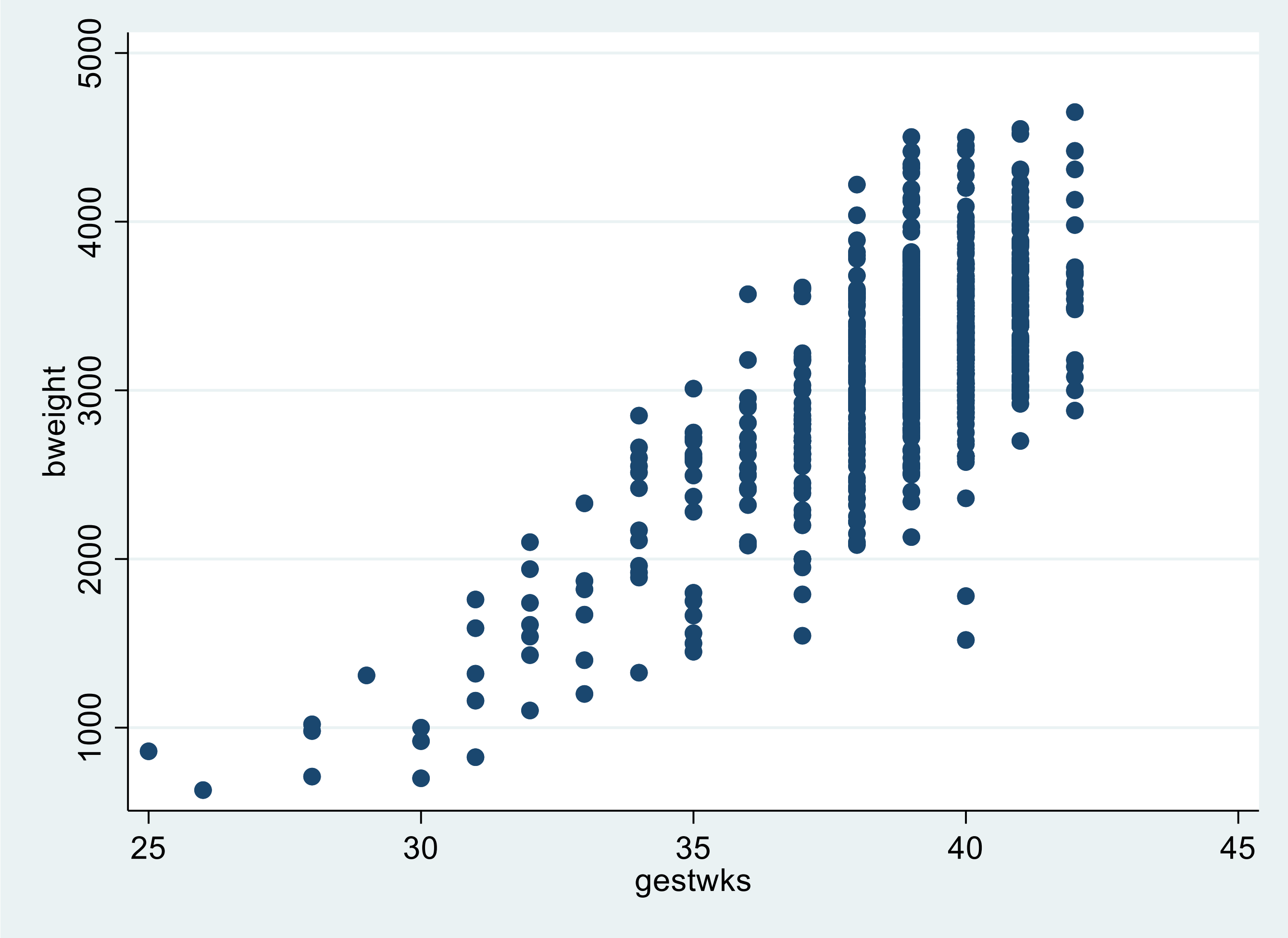
- Read the birth_ weight.csv dataset.

- What proportion of children were low birthweight?
- What proportion of boys were LBW?
- What was the mean age of the mothers?
- What was the mean age of mothers in the four ethnic groups?

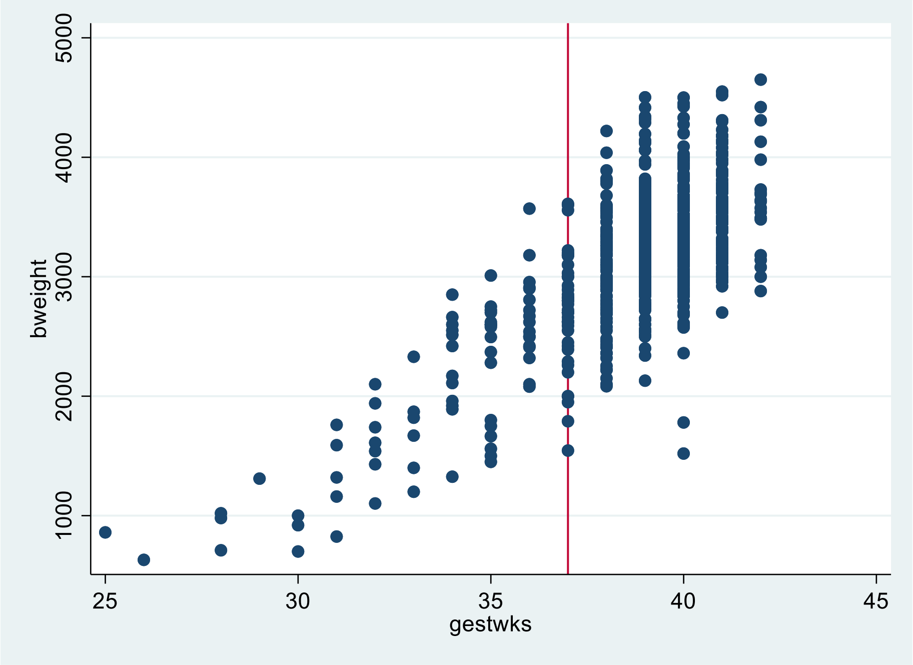
Graphs

Scatterplots are good to explore possible relationships or patterns between variables and to identify outliers.

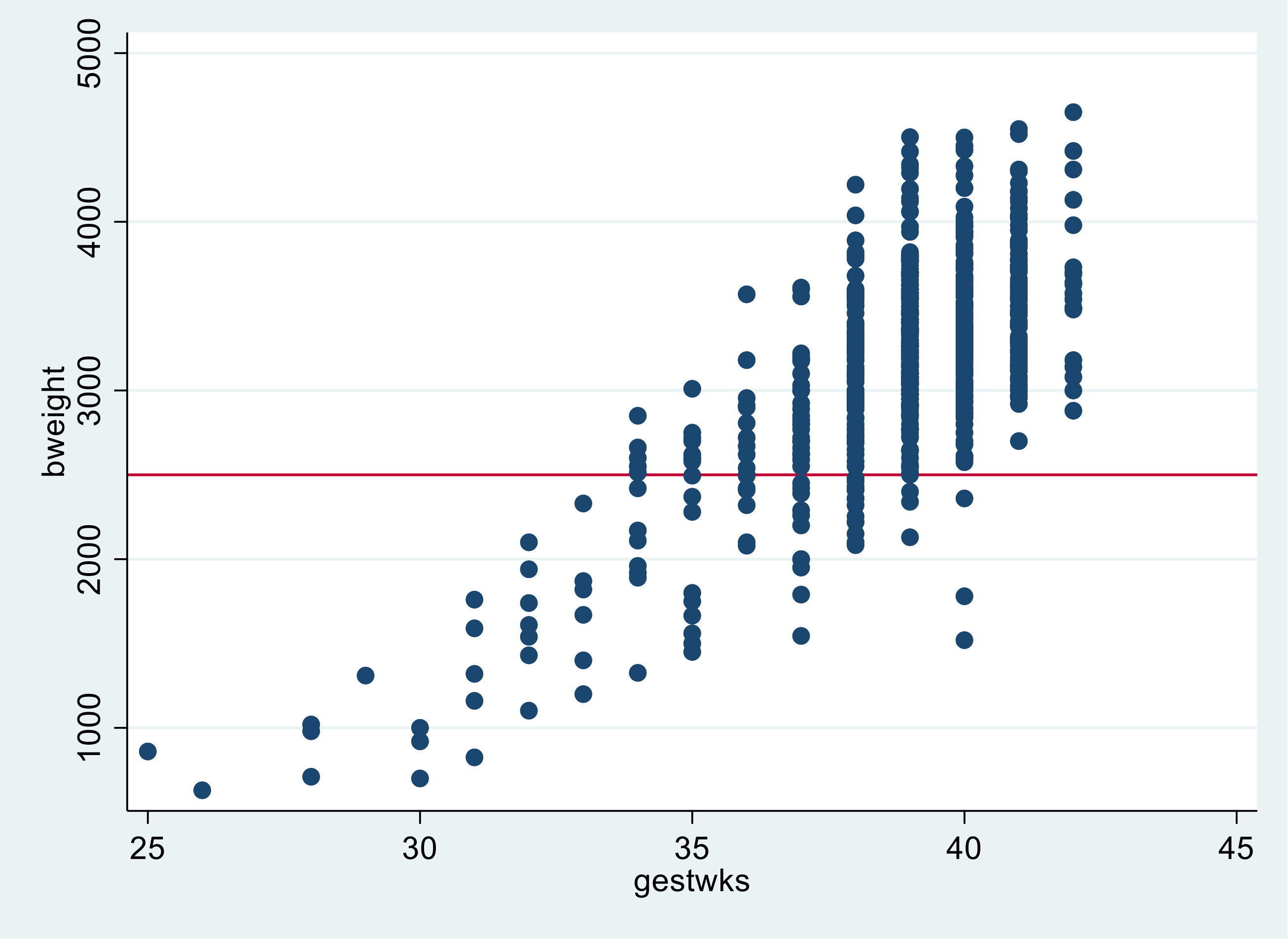
```
. twoway scatter bweight gestwks
```



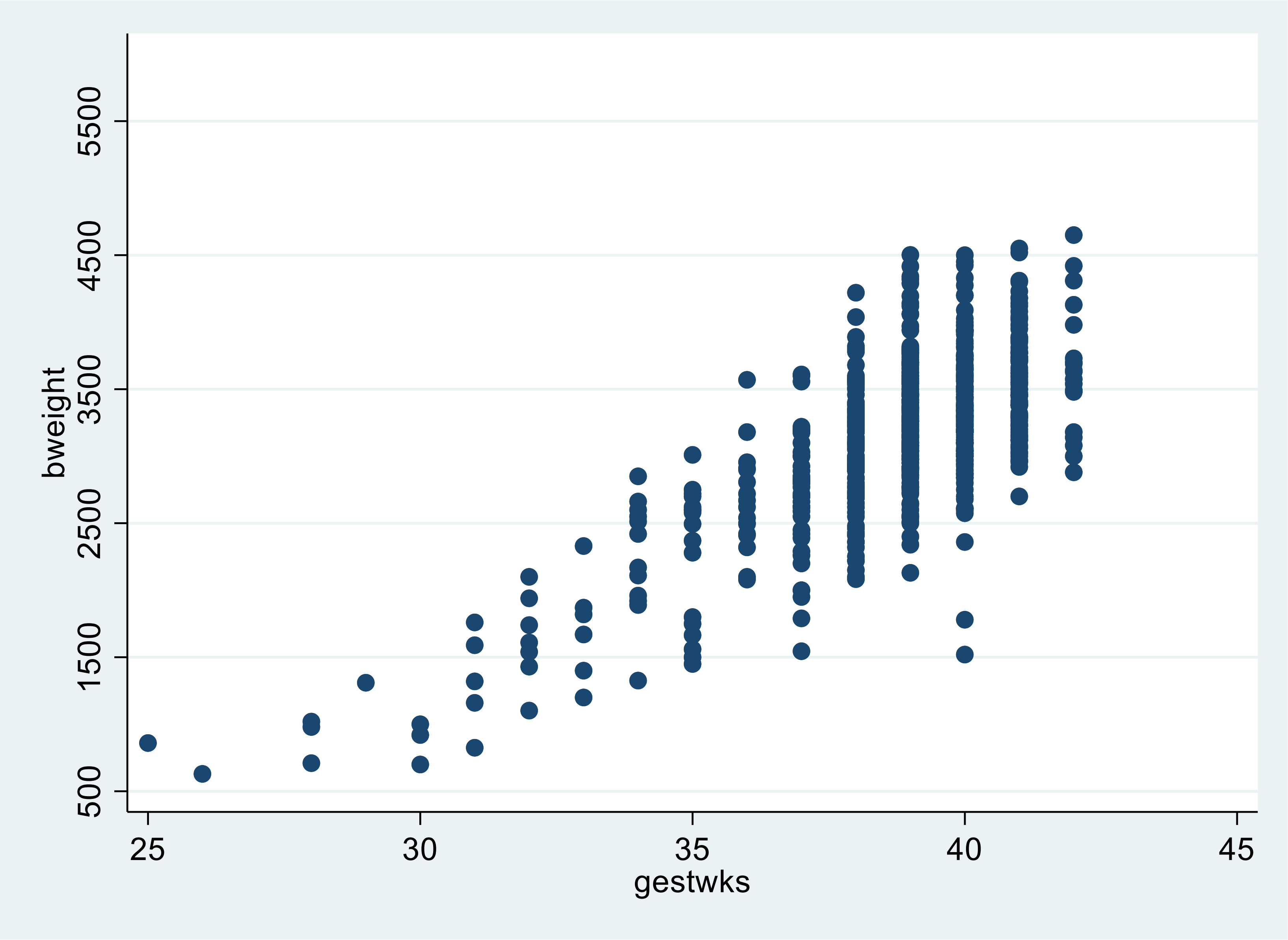
```
twoway scatter bweight gestwks,xline(37)
```



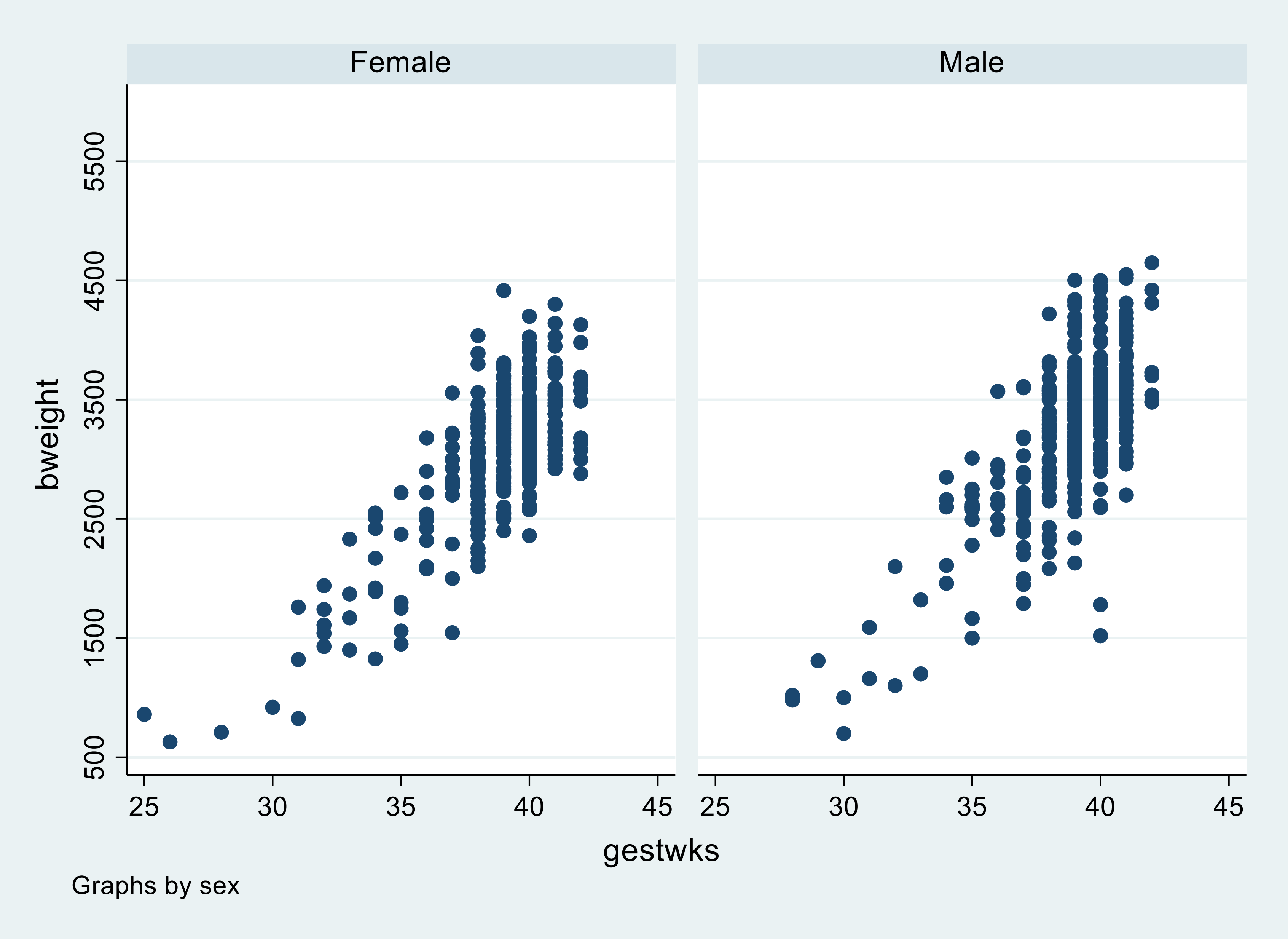
```
twoway scatter bweight gestwks,yline(2500)
```



```
twoway scatter bweight gestwks,xlab(25(5)45) ylab(500(1000)6000)
```

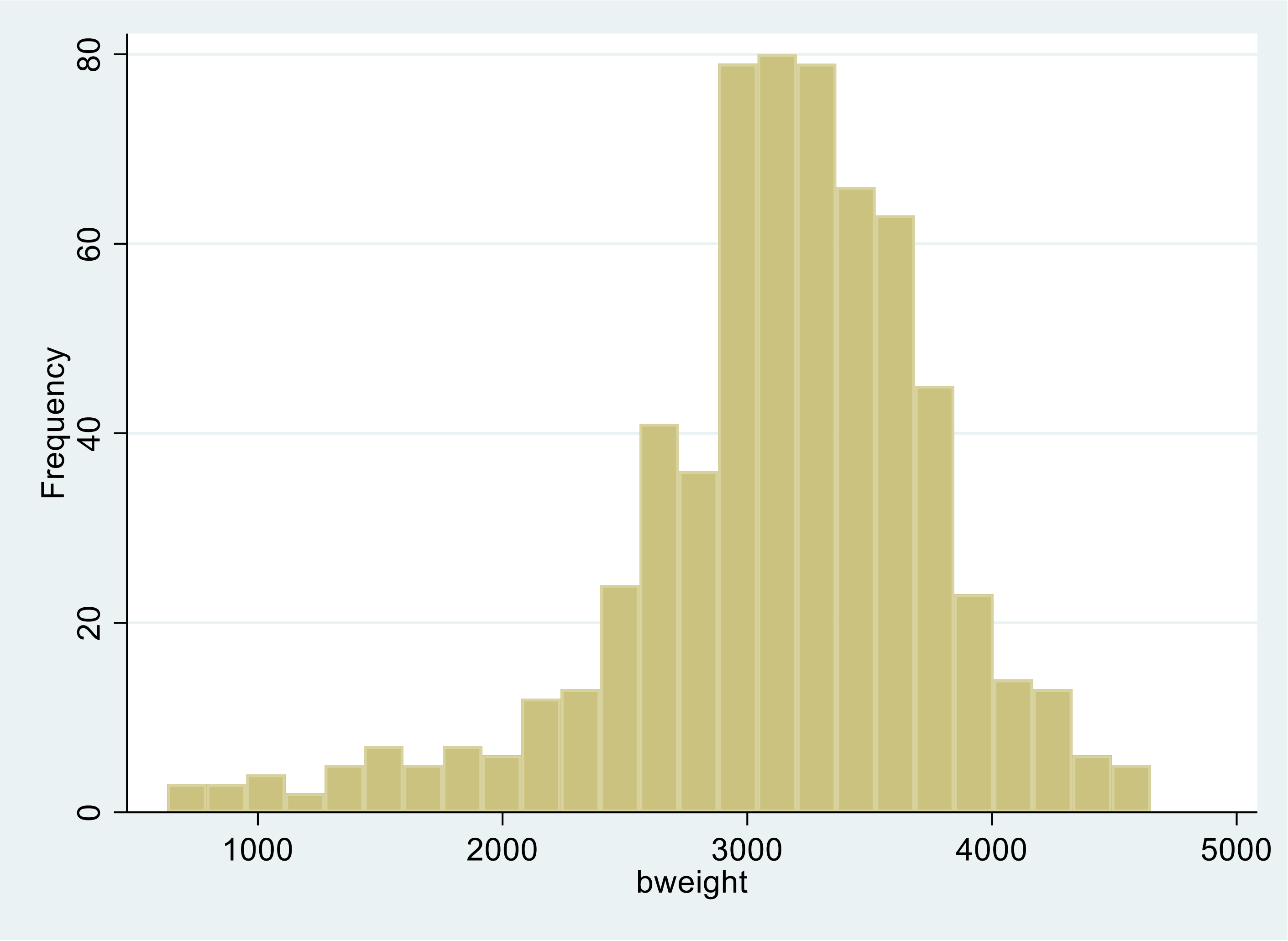


```
twoway scatter bweight gestwks,xlab(25(5)45) ylab(500(1000)6000) by(sex)
```

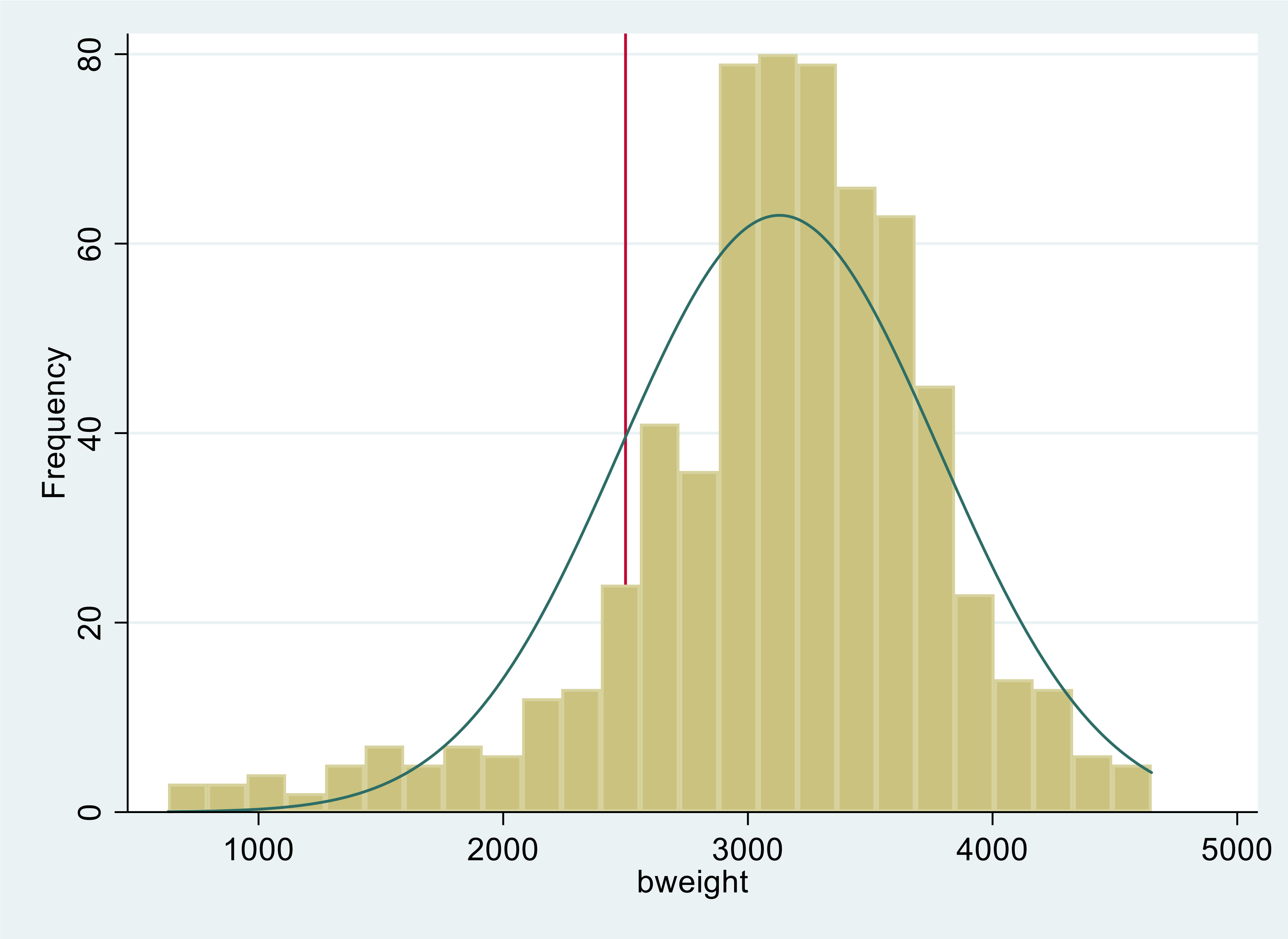


Histograms are another good way to visually explore data, especially to check for a normal distribution. Type `help histogram` for details.

```
. histogram bweight,freq
```



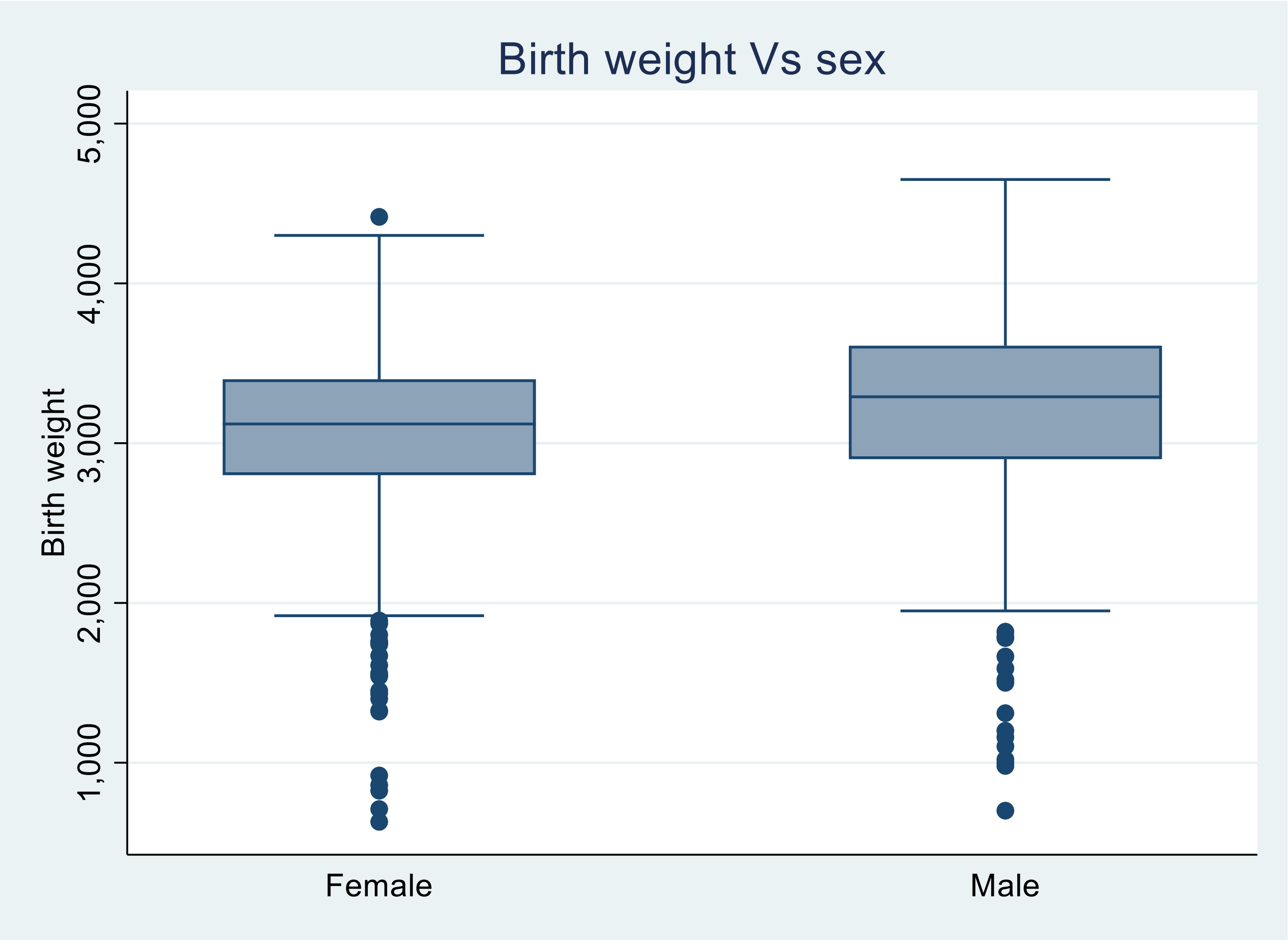
```
histogram bweight,normal freq xline(2500)
```



Box plots

- Plots medians

```
graph box bweight, over(sex) title (Birth weight Vs sex) ytitle(Birth weight)
```



Exercise 4

- Read the birth_weight.csv dataset.
- Plot scatter plot of bweight Vs matage. Add a y-axis line for bweight=2500
- Plot a box plot of gestwks Vs sex
- Plot histogram of gestwks, add a normal distribution curve