

Getting data in and out of Stata – and more

Svend Juul, June 2011

1

Getting data in and out of Stata

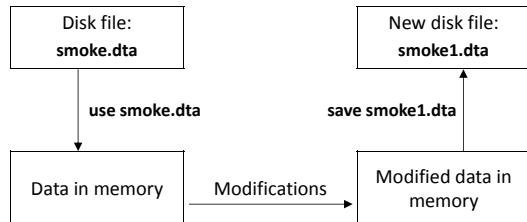
- **use** fetches a copy of a Stata dataset on disk and puts the copy in Stata memory.
If you have unsaved data in memory, Stata will refuse, unless you use the **clear** option:
`use filename.dta, clear`
- **save** copies the data in memory to a disk file.
If a disk file with that name already exists, Stata will refuse, unless you use the **replace** option:
`save filename.dta, replace`

2

Risk to destroy good data by mistake

3

STRONG recommendation:
After modifications, save dataset with a *new name*.



Always make modifications in a do-file. Give the do-file a name that tells what it does (`gen_smoke1.do`).

4

Entering very small amounts of data

The **input** command:

```

input case exposed n
0 0 100
0 1 30
1 0 23
1 1 21
end
expand n
tab2 case exposed
-> tabulation of case by exposed
      exposed
      case | 0 | 1 | Total
-----+-----+-----+
      0 | 100 | 30 | 130
      1 | 23  | 21 | 44
-----+-----+-----+
      Total | 123 | 51 | 174
  
```

5

Entering small amounts of data

Start by generating a template. Use a do-file:

```

clear
generate id="" 
generate type=.
generate sold=.
generate price=.
save aa0.dta
  
```

Open Stata's Data Editor:

Enter data.

Not good if there are more variables than can be seen in one screen.

6

How about Excel?

- Not good if there are more variables than can be seen in one screen.
- Problems transferring data from Excel to Stata's Data Editor.
- DO NOT transport data from Excel to Stata by copy-and-paste. DO NOT!

7

Using a data entry program

We recommend EpiData (freeware).
Find a short introduction at
www.folkesundhed.au.dk/uddannelse/software.

Epidata lets you:

- set up a screen resembling a questionnaire page.
- check the validity of the data entered.
- enter data twice and compare the two files.
- export the data entered to, e.g., Stata.

For your first simple task, you will be flying in 10 minutes.

8

Importing ASCII data

To read fixed-format ASCII data:

```
infix id 1 type 2 sold 3-5 price 6-11 using a.txt
```

To read ASCII data generated by a spreadsheet (tab-separated; variable names in first line):

```
insheet using a.txt , tab
```

Semicolon-separated; no variable names included:

```
insheet id type sold price using a.txt , delimit(";")
```

You may use the menus to generate the command:

File ► Import

9

Complex data structures

Stata handles simple rectangular data structures.
A database program like MS Access may be useful.
But you can come a long way with Stata.

If you have data from several sources, e.g.

- baseline clinical examination
- baseline interview data
- follow-up clinical examination
- follow-up interview data
- register based information on hospitalizations, deaths

then enter data from each source in a separate file.

Let the source be reflected in the variable names.

Make corrections and derived variables in each partial dataset before merging them.

10

Translating between packages

Use a translation program. We recommend Stat/Transfer (Aarhus GradPlan price SEK 684).

Stat/Transfer translates between, e.g., SPSS, SAS, Stata, and Excel. It is EASY to use.

But take care, anyway, especially with missing values and with date variables.

Translation from Excel: Column headings must conform Stata's rules for variable names.

11

String variables

- For most purposes, use numerical codes. Exceptions:
 - names, etc. (of course)
 - long id numbers, e.g., 10-digit CPR numbers
 - ICD-10 codes (of course)
- Diabetes has the ICD-10 codes E10.0 to E14.9. If **dx** is the variable with the ICD-10 code, you can identify the diabetics by:

```
generate diab=0
replace diab=1 if dx>="E10" & dx<"E15"
```

12

Date variables

- Date variables are numeric variables: number of days since 1 January, 1960.

```
display exp(10)
22026.466
display %td exp(10)
21apr2020
display mdy(4,21,2020)
22026
```

- If **bdate** is the variable containing the birth date, and **opdate** is the date of operation, we can calculate the age at operation by:

```
generate opage = (opdate-bdate)/365.25
```

13

Getting help, etc.

If you know the command name (e.g. **merge**):

```
help merge
```

If you search for commands on a subject (e.g. nonparametric tests):

```
search nonparametric
```

If you search outside official Stata (e.g. diagnostic tests):

```
findit sensitivity specificity
```

14