CENTER for DIGITAL SCHOLARSHIP

BEYOND THE NUMBERS: INTRODUCTION TO STATA

James Ng
james.ng@nd.edu
Center for Digital Scholarship
Hesburgh Libraries, University of Notre Dame

library.nd.edu/cds/

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries

what is Stata?

- statistical software package

- created in 1985 by economists

why bother when I can use Excel?

- documentation and reproducibility of data and results
- eases revision, collaboration
- reduces time/labor spent on repetitive tasks
- integrates nicely with Word, Excel, LaTeX

steps in data analysis

- locate data
- load data into software package
- manipulate as needed
- analyze

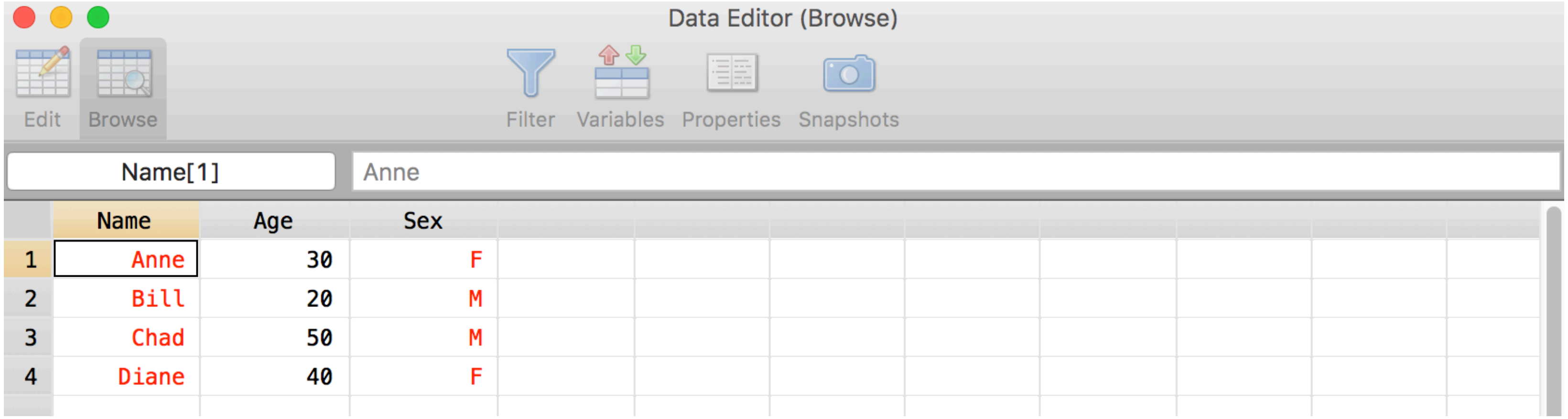
“data”

- a set of numbers and/or text describing specific phenomena
  - economy, test scores, traffic, pollution levels, etc.

- in social sciences, usually rectangular:
  - columns contain “variables”
  - rows contain “observations”

example

Name	Age	Sex
Anne	30	F
Bill	20	M
Chad	50	M
Diane	40	F



Name[1] Anne

	Name	Age	Sex
1	Anne	30	F
2	Bill	20	M
3	Chad	50	M
4	Diane	40	F

today’s agenda

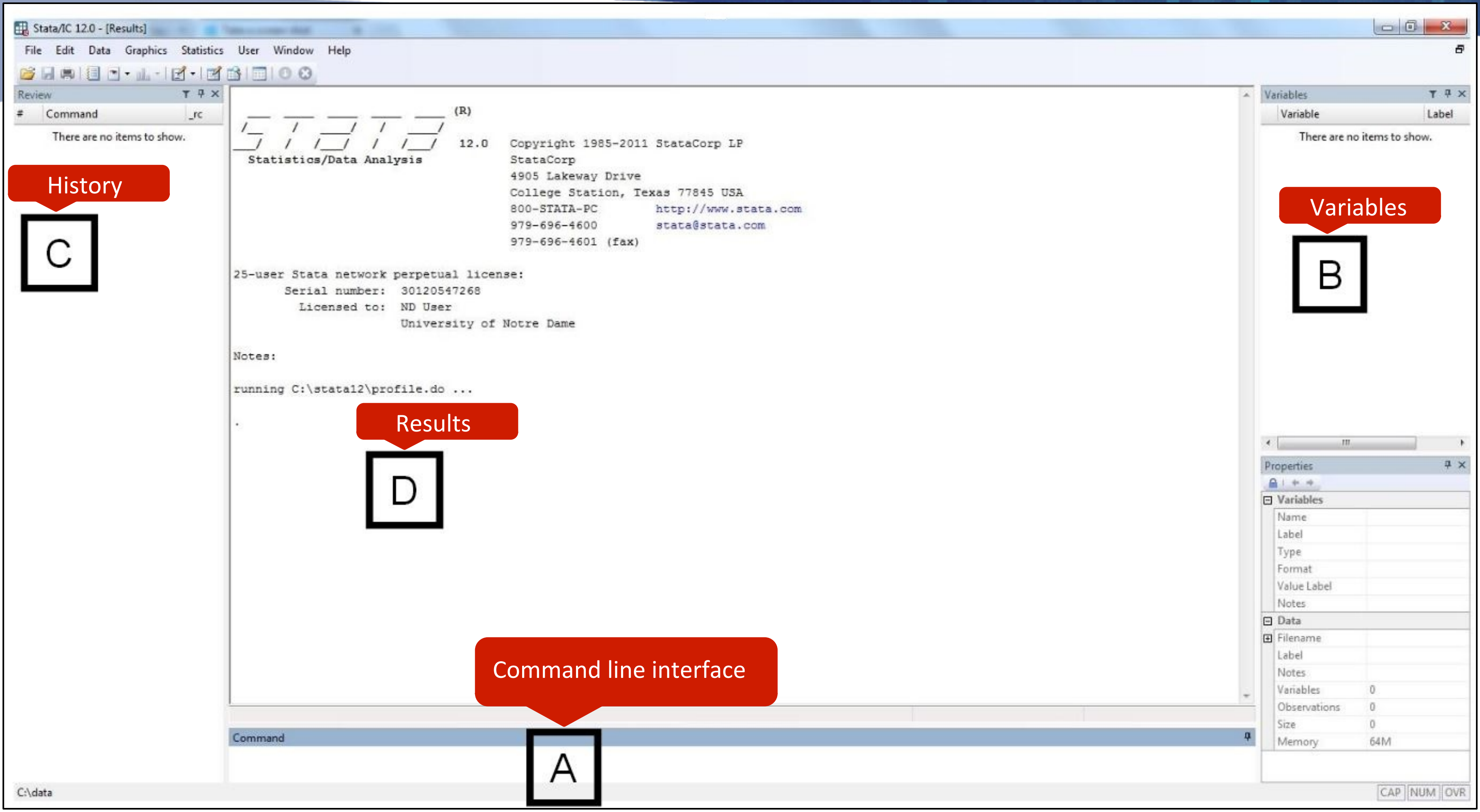
CENTER for DIGITAL SCHOLARSHIP

- how to load data
- basic manipulations, analysis

- on two widely-used, publicly-available datasets:
  - – National Health Interview Survey (NHIS)
  - – General Social Survey (GSS)

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries

Stata environment



```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   12.0   Copyright 1985-2011 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
                                      College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

25-user Stata network perpetual license:
       Serial number:  30120547268
         Licensed to:  ND User
                       University of Notre Dame

Notes:

running C:\stata12\profile.do ...

.
```

History

C

Variables

B

Results

D

Command line interface

A

ways to use Stata

- point & click
- enter commands in command line interface
- enter commands or code in a “do-file” ← do this for extended projects

good habits for every user

- set Stata’s “working directory”
  - if dataset is stored in /Volumes/~jng2/workshop/intro, or if you want any files you produce saved there, set that folder as the working directory:
- cd /Volumes/~jng2/workshop/intro
  - what is my current working directory?
- pwd

loading data into Stata (1)

Command	File Type	File Extension
use	Stata format	.dta (always)
infix	Fixed-format ASCII	.dat, .raw, .fix, or simply nothing
infile (version 1)	Text-delimited ASCII	
infile (version 2)	Fixed-format ASCII, with a “dictionary”	
import delimited	Text-delimited ASCII	
import excel	Excel	.xls, .xlsx

loading data into Stata (2)

- Excel spreadsheets
- command: `import excel`
- GUI: File > Import > Excel spreadsheet

loading data into Stata (3)

CENTER for DIGITAL SCHOLARSHIP

- example: National Health Interview Survey
  - http://www.cdc.gov/nchs/nhis/nhis_2012_data_release.htm
- this is a fixed-format ASCII file
- Stata command: `infix`
- fixed-format data must come with a codebook
- GUI: File > Import > Text data in fixed format
- script to load data already written by data provider – really helpful!

loading data into Stata (4)

- Stata-format data
- example: General Social Survey
  - http://www3.norc.org/GSS+Website/Download/STATA+v8.0+Format/
- reading Stata-format data is trivial
- Stata command: `use`

```
use /Volumes/~jng2/workshop/intro/
GSS2012.dta, clear
```

- good practice:

```
cd /Volumes/~jng2/workshop/intro
use GSS2012, clear
```

combining datasets

- merging
  - adding variables to existing observations

id	sex
001	M
002	F

data1.dta

+

id	age
001	21
002	23

data2.dta

=

id	sex	age
001	M	21
002	F	23

```
use data1
merge 1:1 id using data2
```

- appending
  - adding observations to existing variables

id	sex
001	M
002	F

data1.dta

+

id	sex
003	F
004	M

data3.dta

=

id	sex
001	M
002	F
003	F
004	M

```
use data1
append using data3
```

more on merging (1)

- for each dataset, must know whether identifying variable/s is/are unique
- in the previous example, the identifying variable is id and clearly unique in each dataset (each value of id only occurs once)
- therefore, we performed a 1:1 (one-to-one) merge
- and here's an example of a m:1 (many-to-one) merge

more on merging (2)

- here’s an example of a 1:m (one-to-many) merge
- we want to merge the two files based on state
- state is the identifying variable

state	area
IN	36.4
MI	96.7

data1.dta

+

state	year	gdp
IN	2014	324289
IN	2013	310669
MI	2014	447221
MI	2013	431112

data2.dta

=

state	year	area	gdp
IN	2014	36.4	324289
IN	2013	36.4	310669
MI	2014	96.7	447221
MI	2013	96.7	431112

```
use data1
merge 1:m state using data2
```

more on merging (3)

- to find out whether the identifying variable/s is/are unique, use the `duplicates report` command
- if the identifying variable is unique, there will be no surplus observations reported

other commands for manipulating data

CENTER for DIGITAL SCHOLARSHIP

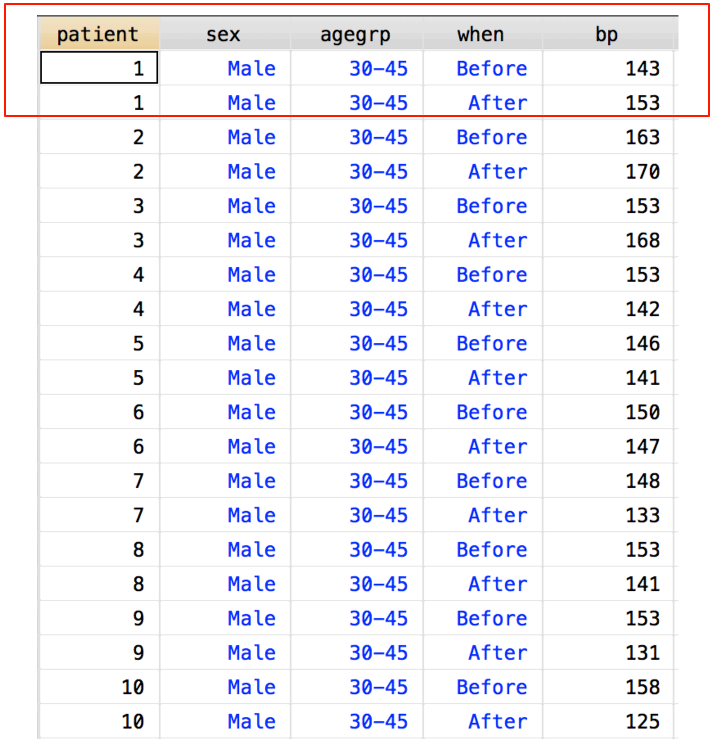
- to combine datasets:
  - `joinby`
  - `cross`

- to reshape datasets:
  - `reshape`
  - `xpose`
  - `sxpose`

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries

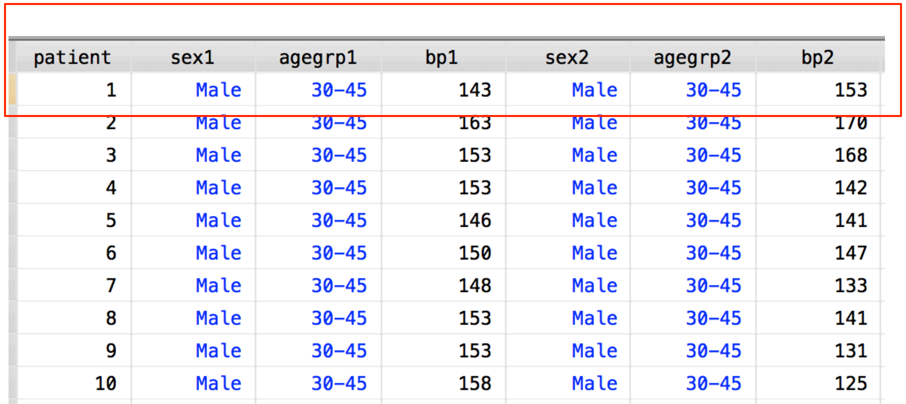
reshape example

```
sysuse bplong, clear
reshape wide sex agegrp bp, i(patient) j(when)
```



patient	sex	agegrp	when	bp
1	Male	30-45	Before	143
1	Male	30-45	After	153
2	Male	30-45	Before	163
2	Male	30-45	After	170
3	Male	30-45	Before	153
3	Male	30-45	After	168
4	Male	30-45	Before	153
4	Male	30-45	After	142
5	Male	30-45	Before	146
5	Male	30-45	After	141
6	Male	30-45	Before	150
6	Male	30-45	After	147
7	Male	30-45	Before	148
7	Male	30-45	After	133
8	Male	30-45	Before	153
8	Male	30-45	After	141
9	Male	30-45	Before	153
9	Male	30-45	After	131
10	Male	30-45	Before	158
10	Male	30-45	After	125

long form



patient	sex1	agegrp1	bp1	sex2	agegrp2	bp2
1	Male	30-45	143	Male	30-45	153
2	Male	30-45	163	Male	30-45	170
3	Male	30-45	153	Male	30-45	168
4	Male	30-45	153	Male	30-45	142
5	Male	30-45	146	Male	30-45	141
6	Male	30-45	150	Male	30-45	147
7	Male	30-45	148	Male	30-45	133
8	Male	30-45	153	Male	30-45	141
9	Male	30-45	153	Male	30-45	131
10	Male	30-45	158	Male	30-45	125

wide form

inspecting your data (1)

- read the manual / codebook / user guide

- some essential commands:

```
sort
order
browse
describe
lookfor
summarize
tabulate
```

selecting variables

```
keep id happy abpoor age race sex health1 region
```

  - see also: `drop`

- save your work data in a new file:

```
save temp_gss2012
```

- or overwrite an existing file:

```
save temp_gss2012, replace
```

- be careful to not unintentionally overwrite dataset if it isn’t your intention to overwrite it

creating a new variable (1)

- create a variable to indicate unhappiness based on existing happy variable
- don't be misled by "value labels" (text labels for numeric values)

```
tabulate happy
tabulate happy, nolabel
browse happy
browse happy, nolabel
```

- watch out for missing values!

```
tabulate happy, nolabel missing
```

creating a new variable (2)

CENTER for DIGITAL SCHOLARSHIP

- create a variable to indicate unhappiness based on existing happy variable
- here’s how

```
gen unhappy = .
replace unhappy = 1 if happy == 3
replace unhappy =0 if happy == 1 | happy == 2
```

- cross-check:
  - `tab unhappy happy, nolabel missing`

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries

creating a new variable (3)

- good practice: label all variables

```
label var unhappy "Is respondent unhappy? 1-yes 0-
no"
```

creating a new variable (3)

- create a variable indicating whether a person feels poor

(note, the following is a shorthand way of creating a dummy variable)

```
gen poor = abpoor==1
replace poor = . if missing(abpoor)

label var poor "Does respondent feel poor? 1-yes 0-
no"
```

basic analysis (1)

- descriptive statistics

```
summarize
su age
tabulate race
tab race, nolabel
tab poor
tab unhappy if race==1
tab unhappy poor
tab unhappy poor, row column
```

basic analysis (2)

- distribution of a variable

```
histogram age, normal
```

- comparison of means

```
ttest unhappy, by(poor)
```

basic analysis (3)

CENTER for DIGITAL SCHOLARSHIP

- what is the relationship between poverty and unhappiness?

```
correlate unhappy poor
regress unhappy poor
```

- what is this relationship controlling for some other factors?

```
recode sex (2=0), gen(male)
xi: reg unhappy poor male age i.health1
```

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries

basic analysis (4)

- how did average happiness change over time?

- use data compiled across years

```
use combined1972_2012, clear

browse

collapse (mean) ave_unhappiness=unhappy, by(year)

label var ave_unhappiness "fraction of respondents who felt unhappy"
```

- we can now graph it:

```
scatter ave_unhappiness year,  xlabel(1972 1982 1991 2002 2012, grid)
```

maps

CENTER for DIGITAL SCHOLARSHIP

- color code Census Divisions according to average level of unhappiness
- Command: `spmap`
- not part of default installation; download and install from Stata server in one easy step:

```
ssc install spmap
```

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries

some other useful commands

- `sysuse`
  - access example datasets; useful learning tool
- `ssc install`
  - install user-written commands
  - e.g. the `estout` package generates nice, publication-quality tables of summary statistics and regression results
  - to install, type `ssc install estout`

using a “do-file”

- send commands to Stata through a batch file (.do)
  - “do-file”

- Stata reads each line as an executable statement
  - To add comments:
    - start a line with an asterisk * or two slashes //

```
*this is a comment and will be ignored by Stata
```

    - enclose successive lines with /* and */

```
/*these two lines are comments
and will be ignored by Stata*/
```

if you get stuck

- Stata has an extensive internal help system
  - need help with how to load data?

    ```
    help loading data
    ```

  - need help with `regress` command?

    ```
    help regress
    ```

- online resources
  - UCLA: http://www.ats.ucla.edu/stat/stata/ ←HIGHLY recommend
  - Notre DameStata guide: http://libguides.library.nd.edu/friendly.php?action=82&s=stata
  - Google search

accessing workshop materials

CENTER for DIGITAL SCHOLARSHIP

- This PowerPoint, Stata datasets and do-files are on Box:
  - https://notredame.box.com/s/vs4aq0x64ovdk4zsoat6
  - http://library.nd.edu/cds/workshops/resources/index.shtml

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries