CENTER for DIGITAL SCHOLARSHIP

Intro to Stata

James Ng, PhD
james.ng@nd.edu

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries

what is Stata?

CENTER for DIGITAL SCHOLARSHIP

- statistical software package
- created in 1985 by economists

why bother when I can use Excel?

CENTER for DIGITAL SCHOLARSHIP

- documentation and reproducibility of data and results
- eases revision, collaboration
- integrates nicely with Word, Excel, LaTeX
- time and energy saver

steps in data analysis

CENTER for DIGITAL SCHOLARSHIP

- locate data
- load data into software package
- manipulate as needed *(bulk of your time)*
- analyze

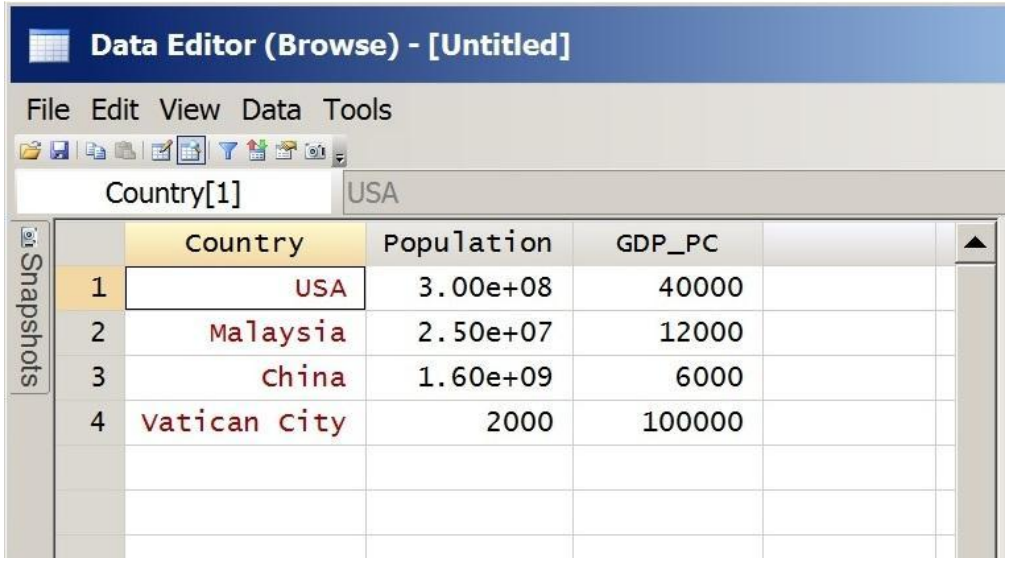
"data"

CENTER for DIGITAL SCHOLARSHIP

- a set of numbers and/or text describing specific phenomena
  - economy, weather, traffic, pollution levels
- in social sciences, always rectangular:
  - columns contain "variables"
  - rows contain "observations"

example

Country	Population	GDP per capita (in USD)
USA	300,000,000	40,000
Malaysia	25,000,000	12,000
China	1,600,000,000	6,000
Vatican City	2,000	100,000



today's agenda

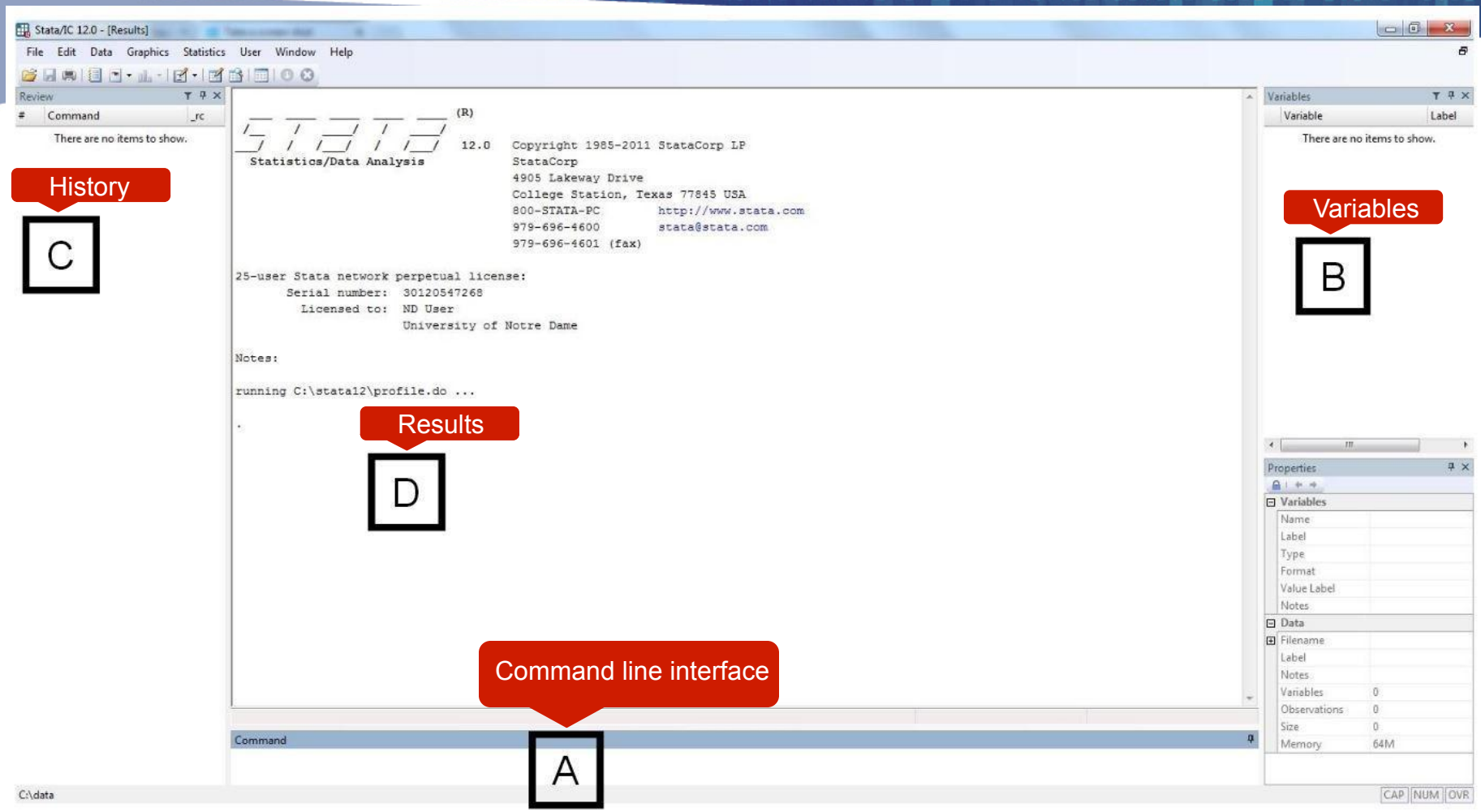
CENTER for DIGITAL SCHOLARSHIP

- how to load data and do basic manipulations and analysis

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

Stata environment

CENTER for DIGITAL SCHOLARSHIP



UNIVERSITY OF NOTRE DAME
Hesburgh Libraries

ways to use Stata

CENTER for DIGITAL SCHOLARSHIP

- point & click ← OK to start
- command line interface ← good ← today
- batch file (called a "do-file") ← best

keeping records

CENTER for DIGITAL SCHOLARSHIP

- good practice:
  keep a log file at start of each session

- Stata command:

```
log using anyfilename.log, text replace
```

loading data into Stata [1]

- there are many ways

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 [1]

CENTER for DIGITAL SCHOLARSHIP

- today:
  - ➔ load a Stata-format dataset
  - ➔ load an ASCII dataset

loading data into Stata [2]

CENTER for DIGITAL SCHOLARSHIP

- before you start, must know: file path, file name

loading data into Stata [3]

- example 1: GSS data
- reading Stata-format data is trivial
- Stata command: `use`

```
use N:\Public\GSS\GSS2012.dta, clear
```

- good practice:

```
cd N:\Public\GSS\
use GSS2012, clear
```

loading data into Stata [4]

CENTER for DIGITAL SCHOLARSHIP

- example 2: NHIS data
- fixed-format ASCII file

http://www.cdc.gov/nchs/nhis/nhis_2012_data_release.htm

- Stata command: `infix`
- script to load data already written by data provider – really helpful!

-how to use it?

combining datasets

CENTER for DIGITAL SCHOLARSHIP

Merging

- adding variables to existing observations
- similar to SQL join, SPSS match files

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, clear
merge 1:1 id using data2
```

Appending

- adding observations to existing variables
- similar to SPSS add files

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, clear
append using data3
```

inspecting your data [1]

CENTER for DIGITAL SCHOLARSHIP

- read the manual / codebook / user guide

- some essential commands:

```
sort
order
browse
describe
lookfor
sum
tab
```

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 existing file:

```
save temp_gss2012, replace
```

- be careful not to unintentionally overwrite dataset

creating a new variable [1]

- create an variable to indicate unhappiness based on an existing variable

- don't be misled by "value labels"

```
tab happy, nolabel
```

- watch out for missing values!

```
tab happy, nolabel missing
```

creating a new variable [2]

- 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`

creating a new variable [3]

CENTER for DIGITAL SCHOLARSHIP

- good practice: label all variables

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

creating a new variable [3]

CENTER for DIGITAL SCHOLARSHIP

- create a variable indicating whether a person feels poor

```
gen poor = abpoor==1
replace poor = . if missing(abpoor)
label var poor "Does respondent feel poor? 1-yes 0-no"
```

creating a new variable [4]

- you can also label a variable's values
- let's label values of unhappy
- 2-step process:
    - define labels for variable's values:

```
label define labels_for_unhappy 0 "happy" 1 "unhappy"
```

    - assign value labels to variable:

```
label values unhappy labels_for_unhappy
```

basic analysis [1]

CENTER for DIGITAL SCHOLARSHIP

- descriptive statistics

```
sum
sum age
tab race
tab race, nolabel
tab poor
tab unhappy if race==1
tab unhappy poor
tab unhappy poor, row column
```

basic analysis [2]

CENTER for DIGITAL SCHOLARSHIP

- distribution of a variable

```
histogram age, normal
```

- comparison of means

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

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries

basic analysis [3]

CENTER for DIGITAL SCHOLARSHIP

- what is the relationship between poverty and unhappiness?

```
corr unhappy poor
reg 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, cl(region)
```

- does it vary by gender?

```
xi: reg unhappy i.poor*male age i.health1, cl(region)
xi: reg unhappy poor age i.health1 if male==1, cl(region)
xi: reg unhappy poor age i.health1 if male==0, cl(region)
```

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 [pw=wtssall], 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

- map Census regions according to level of unhappiness

- Command: `spmap`

- not part of basic installation; download and install from Stata server in one easy step:

```
ssc install spmap
```

using a "do-file"

CENTER for DIGITAL SCHOLARSHIP

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

- all commands in this session can be found in a do-file available on Box: https://notredame.box.com/s/vs4aq0x64ovdk4zsoat6

- Stata reads each line as an executable statement
  - ignores lines beginning with an asterisk, * ← documentation, good practice!

if you get stuck

CENTER for DIGITAL SCHOLARSHIP

- Stata has an extensive internal help system

- need help with how to load data?

```
help loading data
```

- need help with `regress` command?

```
help regress
```

- WWW is your friend
  - http://www.ats.ucla.edu/stat/stata/
  - Google

ending your session

CENTER for DIGITAL SCHOLARSHIP

```
log close
```

```
Exit
```

or simply close Stata with your mouse

accessing workshop materials

CENTER for DIGITAL SCHOLARSHIP

- This PowerPoint is on CDS website:
  - http://library.nd.edu/cds/workshops.shtml#DataAnalysis
- Stata datasets and do-files are on Box:
  - https://notredame.box.com/s/vs4aq0x64ovdk4zsoat6

other resources on campus

CENTER for DIGITAL SCHOLARSHIP

- Center for Social Research: http://csr.nd.edu

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries

rate this workshop

CENTER for DIGITAL SCHOLARSHIP

http://library.nd.edu/cds/WorkshopFeedbackForm.shtml

UNIVERSITY OF NOTRE DAME
Hesburgh Libraries