Intro to Stata

James Ng
james.ng@nd.edu

Center for Digital Scholarship
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
- integrates nicely with Word, Excel, LaTeX
- time and energy saver for advanced user

steps in data analysis

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

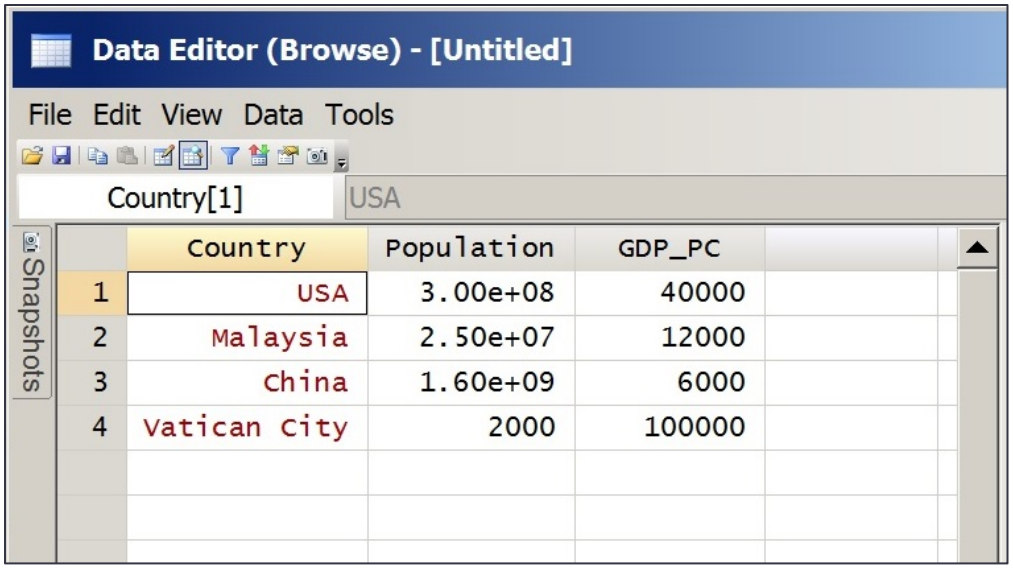
“data”

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

- 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

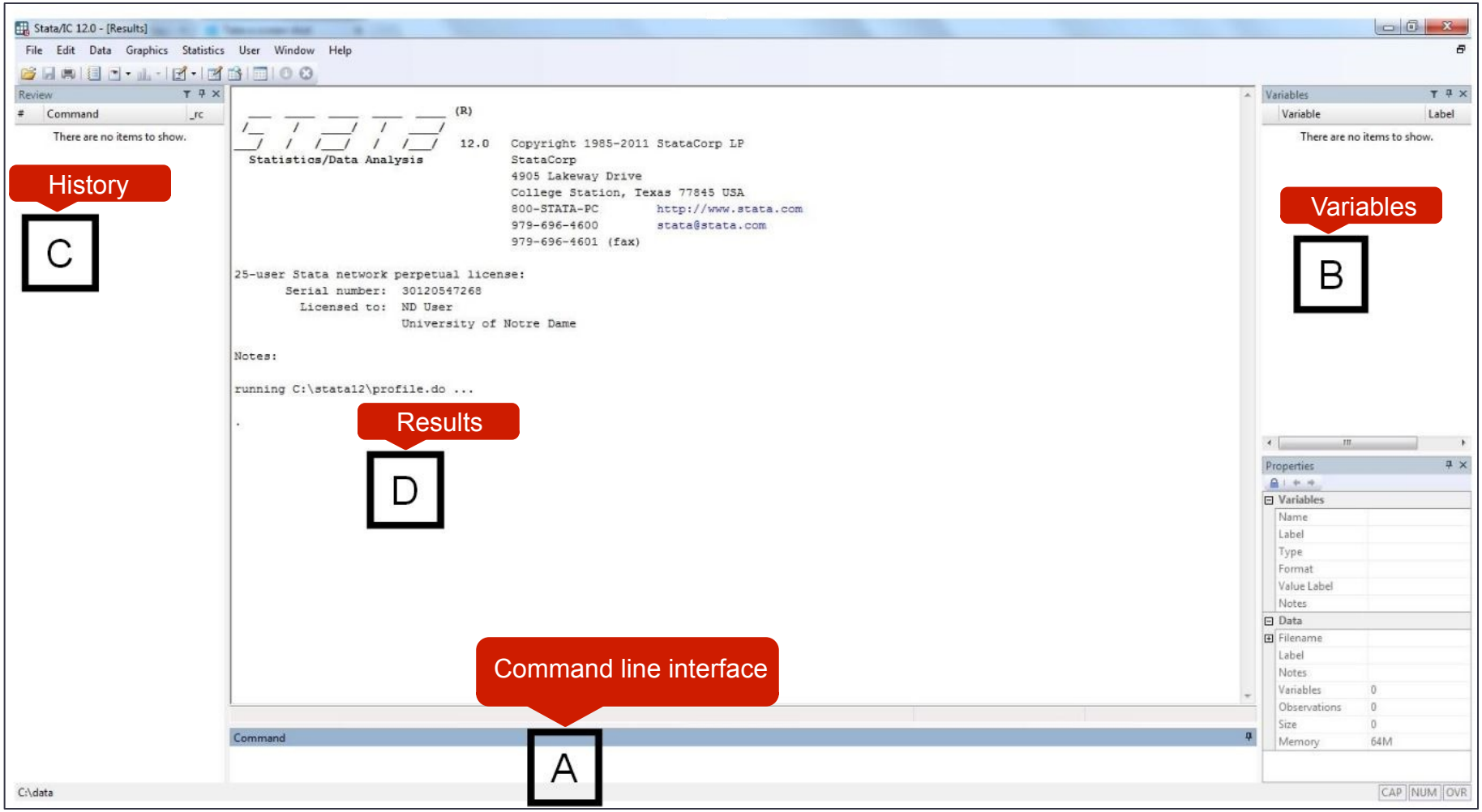


today's agenda

- demonstrate basic manipulation and analysis in Stata

- on happiness data (General Social Survey)
  - http://www3.norc.org/gss+website/

Stata environment



ways to use Stata

- point & click ⟵ **avoid**
- command line interface ⟵ **good**
- batch file (called a “do-file”) ⟵ **best**

keeping records

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

- Stata command:

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

loading data into Stata

- there are many ways
- today: load a Stata-format dataset
- must know: file path, file name

- Stata command:

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

- good practice:

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

inspecting your data (1)

- commands to use:

```
browse
describe
lookfor
sum
tab
```

inspecting your data (2)

```
lookfor happy

tab happy
```

- watch out for missing values!

```
tab happy, missing
tab happy, nolabel missing

tab abpoor
tab abpoor, nolabel missing
```

selecting variables

```
keep happy abpoor age race id
```

- careful: never overwrite original dataset
- save your work data in a new file:

```
save temp_gss2012
```

creating a new variable (1)

- create a variable indicating whether a person feels unhappy

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

- equivalently:

```
gen unhappy = happy == 3
replace unhappy = . if happy == .d | happy == .n
```

creating a new variable (2)

- good practice: label your variables

```
label var unhappy "Is respondent unhappy?"
```

creating a new variable (3)

- create a variable indicating whether a person feels poor

```
gen poor = .
replace poor = 1 if abpoor == 1
replace poor =0 if abpoor == 2

label var poor “Does respondent feel poor?”
```

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)

- 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)

- distribution of a variable

```
histogram age, normal
```

- comparison of means

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

basic analysis (3)

- what is the association between poverty and unhappiness?

```
regress unhappy poor
```

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 finally graph it:

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

fancier stuff: maps

- map Census regions according to level of unhappiness
- Command: `spmap`
- not part of basic installation; download and install from Stata server

```
ssc install spmap
```

using a “do-file”

- send commands to Stata through a batch file with the extension .do
  - “do-file”

- all commands in this session can be found in a do-file (available on Box)

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

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
```

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

ending your session

```
log close
```

```
exit
```

or simply close Stata with your mouse

accessing workshop materials

- PowerPoint slides, Stata datasets and do-files from this session are available on Box:
  - https://notredame.box.com/s/vs4aq0x64ovdk4zsoat6

other resources on campus

- Center for Social Research workshop series
  - First workshop: October 17
  - http://csr.nd.edu/events/