

INTRO TO STATA

James Ng, PhD
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

steps in data analysis

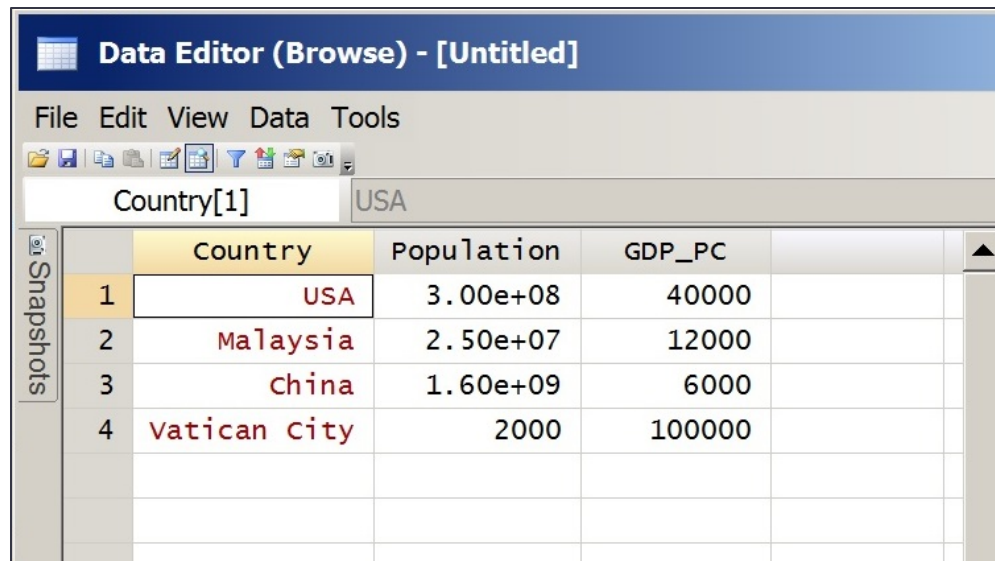
- locate data
- load data into software package
- manipulate as needed  **bulk of your time**
- analyze

“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



The screenshot shows a software window titled "Data Editor (Browse) - [Untitled]". It has a menu bar with "File", "Edit", "View", "Data", and "Tools". Below the menu is a toolbar with various icons. A search bar contains "Country[1]" and a dropdown menu shows "USA". The main area displays a table with 4 rows and 4 columns. The columns are labeled "Country", "Population", "GDP_PC", and an unlabeled column. The rows are numbered 1 to 4. The data is as follows:

	Country	Population	GDP_PC	
1	USA	3.00e+08	40000	
2	Malaysia	2.50e+07	12000	
3	china	1.60e+09	6000	
4	Vatican City	2000	100000	

today's agenda

- 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

The screenshot shows the Stata 12.0 interface with several components labeled:

- History** (C): A red callout pointing to the Command window on the left, which displays the command `_rc`.
- Variables** (B): A red callout pointing to the Variables window on the right, which is currently empty.
- Results** (D): A red callout pointing to the main Results window, which displays the Stata startup screen, including the Stata logo, version 12.0, copyright information, and license details.
- Command line interface** (A): A red callout pointing to the Command window at the bottom of the interface, which is currently empty.

The main Results window displays the following text:

```
(R)
-----
Statistics/Data Analysis

12.0 Copyright 1985-2011 StataCorp LP
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 ...
```

The Command window at the bottom shows the command `_rc`.

The Variables window on the right shows the following table:

Variable	Label
There are no items to show.	




The Properties window on the right shows the following table:

Variables	
Name	Label
Type	Format
Value Label	Notes

The Data window on the right shows the following table:

Data	
Filename	Label
Notes	Variables
Observations	0
Size	0
Memory	64M

ways to use Stata

- point & click  OK to start
- 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 (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

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

loading data into Stata (2)

- 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)

- example 2: NHIS data
 - http://www.cdc.gov/nchs/nhis/nhis_2012_data_release.htm
- fixed-format ASCII file
- Stata command: `infix`
- script to load data already written by data provider – really helpful!
 - how to use it?

combining datasets

- Merging

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

id	sex		id	age		id	sex	age
001	M	+	001	21	=	001	M	21
002	F		002	23		002	F	23
data1.dta			data2.dta					

use data1, clear
merge 1:1 id using
data2

- Appending

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

id	sex		id	sex		id	sex
001	M	+	003	F	=	001	M
002	F		004	M		002	F
data1.dta			data3.dta			003	F
						004	M

use data1, clear
append using data3

inspecting your data (1)

- 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)

- 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

```
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)

- 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 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”

- 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):
<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

- 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

- 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 Social Research: <http://csr.nd.edu>

rate this workshop

- <http://library.nd.edu/cds/WorkshopFeedbackForm.shtml>