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/
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”
<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anne</td>
<td>30</td>
<td>F</td>
</tr>
<tr>
<td>Bill</td>
<td>20</td>
<td>M</td>
</tr>
<tr>
<td>Chad</td>
<td>50</td>
<td>M</td>
</tr>
<tr>
<td>Diane</td>
<td>40</td>
<td>F</td>
</tr>
</tbody>
</table>
today’s agenda

• how to load data
• basic manipulations, analysis

• on two widely-used, publicly-available datasets:
  – National Health Interview Survey (NHIS)
  – General Social Survey (GSS)
Stata environment
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
<table>
<thead>
<tr>
<th>Command</th>
<th>File Type</th>
<th>File Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>use</td>
<td>Stata format</td>
<td>.dta (always)</td>
</tr>
<tr>
<td>infix</td>
<td>Fixed-format ASCII</td>
<td></td>
</tr>
<tr>
<td>infile (version 1)</td>
<td>Text-delimited ASCII</td>
<td>.dat, .raw, .fix, or simply nothing</td>
</tr>
<tr>
<td>infile (version 2)</td>
<td>Fixed-format ASCII, with a “dictionary”</td>
<td></td>
</tr>
<tr>
<td>import delimited</td>
<td>Text-delimited ASCII</td>
<td></td>
</tr>
<tr>
<td>import excel</td>
<td>Excel</td>
<td>.xls, .xlsx</td>
</tr>
</tbody>
</table>
loading data into Stata (2)

- Excel spreadsheets

- command: `import excel`

- GUI: File > Import > Excel spreadsheet
loading data into Stata (3)

- example: National Health Interview Survey

- 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
• 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

<table>
<thead>
<tr>
<th>id</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>M</td>
</tr>
<tr>
<td>002</td>
<td>F</td>
</tr>
</tbody>
</table>

  *data1.dta*

  +

<table>
<thead>
<tr>
<th>id</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>21</td>
</tr>
<tr>
<td>002</td>
<td>23</td>
</tr>
</tbody>
</table>

  *data2.dta*

  ==

<table>
<thead>
<tr>
<th>id</th>
<th>sex</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>M</td>
<td>21</td>
</tr>
<tr>
<td>002</td>
<td>F</td>
<td>23</td>
</tr>
</tbody>
</table>

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

- **Appending**
  - Adding observations to existing variables

<table>
<thead>
<tr>
<th>id</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>M</td>
</tr>
<tr>
<td>002</td>
<td>F</td>
</tr>
</tbody>
</table>

  *data1.dta*

  +

<table>
<thead>
<tr>
<th>id</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>003</td>
<td>F</td>
</tr>
<tr>
<td>004</td>
<td>M</td>
</tr>
</tbody>
</table>

  *data3.dta*

  ==

<table>
<thead>
<tr>
<th>id</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>M</td>
</tr>
<tr>
<td>002</td>
<td>F</td>
</tr>
<tr>
<td>003</td>
<td>F</td>
</tr>
<tr>
<td>004</td>
<td>M</td>
</tr>
</tbody>
</table>

  Use `data1` append using `data3`
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

<table>
<thead>
<tr>
<th>state</th>
<th>area</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN</td>
<td>36.4</td>
</tr>
<tr>
<td>MI</td>
<td>96.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>state</th>
<th>year</th>
<th>gdp</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN</td>
<td>2014</td>
<td>324289</td>
</tr>
<tr>
<td>IN</td>
<td>2013</td>
<td>310669</td>
</tr>
<tr>
<td>MI</td>
<td>2014</td>
<td>447221</td>
</tr>
<tr>
<td>MI</td>
<td>2013</td>
<td>431112</td>
</tr>
</tbody>
</table>

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

• to combine datasets:
  – joinby
  – cross

• to reshape datasets:
  – reshape
  – xpose
  – sxpose
**reshape example**

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

<table>
<thead>
<tr>
<th>patient</th>
<th>sex</th>
<th>agegrp</th>
<th>when</th>
<th>bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>30–45</td>
<td>Before</td>
<td>143</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>30–45</td>
<td>Before</td>
<td>163</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>30–45</td>
<td>After</td>
<td>153</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>30–45</td>
<td>After</td>
<td>168</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>30–45</td>
<td>Before</td>
<td>153</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>30–45</td>
<td>After</td>
<td>142</td>
</tr>
<tr>
<td>7</td>
<td>Male</td>
<td>30–45</td>
<td>Before</td>
<td>146</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
<td>30–45</td>
<td>After</td>
<td>141</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>30–45</td>
<td>Before</td>
<td>150</td>
</tr>
<tr>
<td>10</td>
<td>Male</td>
<td>30–45</td>
<td>After</td>
<td>147</td>
</tr>
</tbody>
</table>

**long form**

<table>
<thead>
<tr>
<th>patient</th>
<th>sex1</th>
<th>agegrp1</th>
<th>bp1</th>
<th>sex2</th>
<th>agegrp2</th>
<th>bp2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>30–45</td>
<td>143</td>
<td>Male</td>
<td>30–45</td>
<td>153</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>30–45</td>
<td>163</td>
<td>Male</td>
<td>30–45</td>
<td>170</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>30–45</td>
<td>153</td>
<td>Male</td>
<td>30–45</td>
<td>168</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>30–45</td>
<td>153</td>
<td>Male</td>
<td>30–45</td>
<td>142</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>30–45</td>
<td>146</td>
<td>Male</td>
<td>30–45</td>
<td>141</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>30–45</td>
<td>150</td>
<td>Male</td>
<td>30–45</td>
<td>147</td>
</tr>
<tr>
<td>7</td>
<td>Male</td>
<td>30–45</td>
<td>148</td>
<td>Male</td>
<td>30–45</td>
<td>133</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
<td>30–45</td>
<td>153</td>
<td>Male</td>
<td>30–45</td>
<td>141</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>30–45</td>
<td>153</td>
<td>Male</td>
<td>30–45</td>
<td>131</td>
</tr>
<tr>
<td>10</td>
<td>Male</td>
<td>30–45</td>
<td>158</td>
<td>Male</td>
<td>30–45</td>
<td>125</td>
</tr>
</tbody>
</table>

**wide form**
inspecting your data (1)

- read the manual / codebook / user guide

- some essential commands:
  
  sort
  order
  browse
  describe
  lookfor
  summarize
  tabulate
selecting variables

\textit{keep id happy abpoor age race sex health1 region}

\textit{– see also: drop}

- save your work data in a new file:
  \texttt{save temp\_gss2012}

- or overwrite an existing file:
  \texttt{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)

  \begin{verbatim}
  tabulate happy
  tabulate happy, nolabel
  browse happy
  browse happy, nolabel
  \end{verbatim}

- watch out for missing values!

  \begin{verbatim}
  tabulate happy, nolabel missing
  \end{verbatim}
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
creating a new variable (3)

• good practice: label all variables

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

• 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
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:

• 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:

```bash
tsc install spmap
```
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/ ➔ HIGHERLY recommend
  – Notre Dame Stata guide: http://libguides.library.nd.edu/friendly.php?action=82&s=stata
  – Google search
accessing workshop materials

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