17.871 PS1 Tips

Mike Sances

MIT

March 5, 2012

Do File Basics

- A do file should always do the following
  1. Give a brief description of what the file is for
  2. Run a command to "Clear" what is currently stored in memory
  3. Set the memory limit
  4. Set the working directory
  5. Open the log file
  6. **Perform the actual commands for reading and analyzing the data**
  7. Close the log file
- Since all do files perform the same tasks, it is useful to have a template

Do File Basics

```
********************************************************************************
/*
Mike Sances
17.871 PS1
Due 2/27/12
*/
********************************************************************************
* clear anything currently in memory
clear *
* set working directory
cd "~/Dropbox/Spring 2012/17.871 - Political Science Laboratory/problem sets/ps1/"
* open log file
log using ps1.log, replace
* optional
set more off
* now do the questions
do part1
do part2
do part3
do part4
do part6
* close log file
log close
```

Do File Basics

- Do files can be linked to one another. Here is what my “part4.do” looks like:

```
clear *
set mem 100m
use CCES
* Q1 *
tab pid
replace pid =. if pid == 8
* Q2 *
collapse pid, by(state)
* Q3 *
hist pid
graph export "hist.eps", as(eps) replace
```

Part II: Getting data into STATA

- “Data comes in many forms. Here’s one way to get data into Stata. Using a text editor (such as EMACS), type the text from Exhibit 1 in the handout “How to Use the STATA infile and infix Commands” into Athena and save it in a file named scores.dat on your home directory. Write the code that will create a STATA data set from this raw data and save it as a file called “scores.dta”. Use the list command to see your data.”

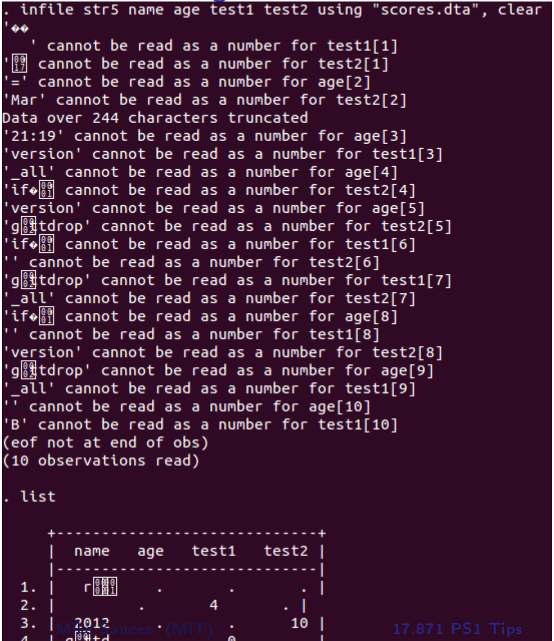
Part II: Getting data into STATA

- "Data comes in many forms."

Format	File Extension	Stata Command
Fixed Format	.dat, .raw, .txt	infix
Space-delimited	.dat, .raw, .txt	infile
Comma-separated	.csv, .txt	insheet
Tab-separated	.txt	insheet
Stata	.dta	use
Excel	.xls, .xlsx	(save as .csv)

  - Note that in general, commands are format-specific. This means that Stata does not like it when you use "infile" for a .dta format file.

Part II: Getting data into STATA



```
. infile str5 name age test1 test2 using "scores.dta", clear
'��
   ' cannot be read as a number for test1[1]
'� cannot be read as a number for test2[1]
'=' cannot be read as a number for age[2]
'Mar' cannot be read as a number for test2[2]
Data over 244 characters truncated
'21:19' cannot be read as a number for age[3]
'version' cannot be read as a number for test1[3]
'_all' cannot be read as a number for age[4]
'if�� cannot be read as a number for test2[4]
'version' cannot be read as a number for age[5]
'g�tdrop' cannot be read as a number for test2[5]
'if�� cannot be read as a number for test1[6]
'' cannot be read as a number for test2[6]
'g�tdrop' cannot be read as a number for test1[7]
'_all' cannot be read as a number for test2[7]
'if�� cannot be read as a number for age[8]
'' cannot be read as a number for test1[8]
'version' cannot be read as a number for test2[8]
'g�tdrop' cannot be read as a number for age[9]
'_all' cannot be read as a number for test1[9]
'' cannot be read as a number for age[10]
'B' cannot be read as a number for test1[10]
(eof not at end of obs)
(10 observations read)

. list

     +-----------------------------+
     |  name   age   test1   test2 |
     |-----------------------------|
  1. |   r��     .       .       . |
  2. |         .       4       . |
  3. |  2012     .       .      10 |
  4. |  g�td             0         |
```

Part II: Getting data into STATA

- Scores.dat:

  ```
  Bob 18 95 18
  Carol 21 43 27
  Ted 14 67 9
  Alice 12 23 31
  ```

  - What format is this?

Part II: Getting data into STATA

- Scores.dat:

  ```
  Bob 18 95 18
  Carol 21 43 27
  Ted 14 67 9
  Alice 12 23 31
  ```

  - What format is this?
    - Space-delimited

Part II: Getting data into STATA

- Scores.dat:

```
Bob 18 95 18
Carol 21 43 27
Ted 14 67 9
Alice 12 23 31
```

  - What format is this?
    - Space-delimited
  - What command would we use?

Part II: Getting data into STATA

- Scores.dat:

```
Bob 18 95 18
Carol 21 43 27
Ted 14 67 9
Alice 12 23 31
```

  - What format is this?
    - Space-delimited
  - What command would we use?
    - infile

Part II: Getting data into STATA

- What format is this?

```
Bob 189518
Carol214327
Ted 1467 9
Alice122331
```

Part II: Getting data into STATA

- What format is this?

```
Bob 189518
Carol214327
Ted 1467 9
Alice122331
```

  - Fixed-format

Part II: Getting data into STATA

- What format is this?

  ```
  Bob 189518
  Carol214327
  Ted 1467 9
  Alice122331
  ```

  - Fixed-format
- What command would we use?

Part II: Getting data into STATA

- What format is this?

  ```
  Bob 189518
  Carol214327
  Ted 1467 9
  Alice122331
  ```

  - Fixed-format
- What command would we use?
  - infix

Part II: Getting data into STATA

- What format is this?

```
Bob,18,95,18
Carol,21,43,27
Ted,14,67,9
Alice,12,23,31
```

Part II: Getting data into STATA

- What format is this?

  ```
  Bob,18,95,18
  Carol,21,43,27
  Ted,14,67,9
  Alice,12,23,31
  ```

  - Comma-separated

Part II: Getting data into STATA

- What format is this?

  ```
  Bob,18,95,18
  Carol,21,43,27
  Ted,14,67,9
  Alice,12,23,31
  ```

  - Comma-separated
- What command would we use?

Part II: Getting data into STATA

- What format is this?

  ```
  Bob,18,95,18
  Carol,21,43,27
  Ted,14,67,9
  Alice,12,23,31
  ```

  - Comma-separated
- What command would we use?
  - insheet

Part II: Getting data into STATA

- “Write the code that will create a STATA data set from this raw data and save it as a file called “scores.dta”. Use the list command to see your data.”

  ```
  clear
  set mem 100m
  infile str5 name age test1 test2 using "scores.dat"
  ```

- Are we done?

Part II: Getting data into STATA

- "Write the code that will create a STATA data set from this raw data and save it as a file called "scores.dta". Use the list command to see your data."

```
clear
set mem 100m
infile str5 name age test1 test2 using "scores.dat"
```

- Are we done?
  - No. We still need to save in Stata format.

```
* save in Stata format:
save "scores", replace
list
```

Part III: Speed Dating

- Q11
  - b. Do any variables have missing data?

Part III: Speed Dating

- How to identify missing data (efficiently)?
  1. "tab variable, m"
  2. "su variable" or "su *"
  3. use the mdesc package

Part III: Speed Dating

```
. tab date, m
(mean) date |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          7        1.27        1.27
          2 |         22        3.99        5.26
          3 |         54        9.80       15.06
          4 |        131       23.77       38.84
          5 |         99       17.97       56.81
          6 |        136       24.68       81.49
          7 |         94       17.06       98.55
          . |          8        1.45      100.00
------------+-----------------------------------
      Total |        551      100.00
```

Part III: Speed Dating

```
. su *
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         iid |       551    276.7877    159.4945          1        552
        wave |       551    11.08348    6.013947          1         21
        date |       543    4.983425     1.46852          1          7
      gender |       551    .5027223    .5004469          0          1
         dec |       551    .4277281    .2559201          0          1
-------------+--------------------------------------------------------
     attr3_1 |       542    7.092251    1.390081          2         10
     sinc3_1 |       542    8.285978    1.412038          2         10
      fun3_1 |       542    7.701107     1.54751          2         10
    intel3_1 |       542    8.385609      1.0897          3         10
      amb3_1 |       542    7.577491    1.786293          2         10
-------------+--------------------------------------------------------
       age_o |       551    26.33275    1.706116   20.44444   31.66667
      race_o |       551    2.800065    .3831047   2.166667          4
       dec_o |       551    .4251092    .2396396          0          1
      attr_o |       551    6.202797    1.185011   2.333333     8.6875
      sinc_o |       551    7.219675    .6886648   4.166667          9
-------------+--------------------------------------------------------
     intel_o |       551    7.398714    .6228981      4.875       9.15
       fun_o |       551    6.434552    1.013942      2.625   8.615385
       amb_o |       551    6.822066    .7738159        3.8   8.842105
      shar_o |       551    5.498002     .951979      1.375        7.7
      like_o |       551    6.160466    .8794204   2.333333        8.3
-------------+--------------------------------------------------------
      prob_o |       551    5.254866    .7652169          2        7.4
       met_o |       551    1.956214    .1000847      1.375   2.666667
```

Part III: Speed Dating

```
. ssc install mdesc
checking mdesc consistency and verifying not already installed...
installing into /home/michael/ado/plus/...
installation complete.
. mdesc
    Variable    |     Missing          Total     Percent Missing
----------------+-----------------------------------------------
            iid |           0            551           0.00
           wave |           0            551           0.00
           date |           8            551           1.45
         gender |           0            551           0.00
            dec |           0            551           0.00
        attr3_1 |           9            551           1.63
        sinc3_1 |           9            551           1.63
         fun3_1 |           9            551           1.63
       intel3_1 |           9            551           1.63
         amb3_1 |           9            551           1.63
          age_o |           0            551           0.00
         race_o |           0            551           0.00
          dec_o |           0            551           0.00
         attr_o |           0            551           0.00
         sinc_o |           0            551           0.00
        intel_o |           0            551           0.00
          fun_o |           0            551           0.00
          amb_o |           0            551           0.00
         shar_o |           0            551           0.00
         like_o |           0            551           0.00
         prob_o |           0            551           0.00
          met_o |           0            551           0.00
----------------+-----------------------------------------------
```

Part V: Research Design

- Q1
    - MIT faculty members were interested in determining whether ending spring-term freshman Pass/No Record had been a success. They decided to answer this question by comparing the GPA of spring-term freshmen before and after the change in Pass/No Record grading had taken effect. The average freshman GPA in the spring of 2002 is 4.0; the average freshman GPA in the spring of 2003 is 4.4. The faculty concluded that the change was a success. (Note the obvious: these are made-up data.)

Part V: Research Design

- Dependent Variable?

Part V: Research Design

- Dependent Variable?
  - GPA (note: *not* “average GPA”. why not?)

Part V: Research Design

- Dependent Variable?
  - GPA (note: *not* “average GPA”. why not?)
- Independent Variable?

Part V: Research Design

- Dependent Variable?
  - GPA (note: *not* "average GPA". why not?)
- Independent Variable?
  - Pass/No Record grading
- So,

$$
\begin{aligned}
Y_i &= X_i\beta + \epsilon_i \\
Y_i &\equiv \textit{GPA of student ``i''} \\
X_i &\equiv \begin{cases} 1 \textit{ if student ``i'' experienced pass/no record grading} \\ 0 \textit{ if not} \end{cases}
\end{aligned}
$$

Part V: Research Design

- What kind of study was this?

Part V: Research Design

- What kind of study was this?
    - Observational

Part V: Research Design

- What kind of study was this?
    - Observational
- Problems with the design?

Part V: Research Design

- What kind of study was this?
  - Observational
- Problems with the design?
  - Confounding
  - Measurement
  - Sample size

Part V: Research Design

- What kind of study was this?
  - Observational
- Problems with the design?
  - Confounding
  - Measurement
  - Sample size
- Ways to improve?

Part V: Research Design

- What kind of study was this?
  - Observational
- Problems with the design?
  - Confounding
  - Measurement
  - Sample size
- Ways to improve?
- Note: Do we care about "external validity" here?

Part V: Research Design

- Q2
  - Researchers were interested in determining whether postcards sent to registered voters encouraging them to vote actually worked. The researchers took the list of registered voters in a town (about 100,000 individuals) and randomly assigned them to one of two samples—T, a sample of voters who were sent the get-out-the-vote postcard, and C, a sample of voters who were not sent the get-out-the-vote postcard. After the election, the researchers went to the town clerk to see who voted. They discovered that 70% of the T group voted, whereas 59% of the C group voted, a highly significant difference, a highly statistically significant difference. The researchers concluded that the "causal effect" of the postcards is to increase turnout by 70%-59% = 11%.

Part V: Research Design

- Dependent variable?

Part V: Research Design

- Dependent variable?
    - Turnout (*not* % turnout)

Part V: Research Design

- Dependent variable?
  - Turnout (*not* % turnout)
- Independent variable?

Part V: Research Design

- Dependent variable?
  - Turnout (*not* % turnout)
- Independent variable?
  - Being sent the post card
- So,

$$
\begin{aligned}
Y_i &= X_i\beta + \epsilon_i \\
Y_i &\equiv \begin{cases} 1 \textit{ if registered voter ``i'' voted} \\ 0 \textit{ if not} \end{cases} \\
X_i &\equiv \begin{cases} 1 \textit{ if registered voter ``i'' was sent a postcard} \\ 0 \textit{ if not} \end{cases}
\end{aligned}
$$

Part V: Research Design

- What kind of study was this?

Part V: Research Design

- What kind of study was this?
  - Experimental

Part V: Research Design

- What kind of study was this?
  - Experimental
- Problems with the design?

Part V: Research Design

- What kind of study was this?
  - Experimental
- Problems with the design?
  - Confounding?

Part V: Research Design

- What kind of study was this?
  - Experimental
- Problems with the design?
  - Confounding?
    - No, because this is an experiment
  - External validity?

Part V: Research Design

- What kind of study was this?
    - Experimental
- Problems with the design?
    - Confounding?
        - No, because this is an experiment
    - External validity?
        - Yes.
        - Only one town.
        - Only registered voters. What if we wanted to know how effective postcards are for mobilizing unregistered voters? This study doesn't answer that question.

Part V: Research Design

- What kind of study was this?
    - Experimental
- Problems with the design?
    - Confounding?
        - No, because this is an experiment
    - External validity?
        - Yes.
        - Only one town.
        - Only registered voters. What if we wanted to know how effective postcards are for mobilizing unregistered voters? This study doesn't answer that question.
    - Ways to improve?