

Lecture 2: .do files, log files and workflow

BIOSTAT 212, Summer 2016



Kristen Aiemjoy
Department of Epidemiology and Biostatistics
University of California San Francisco

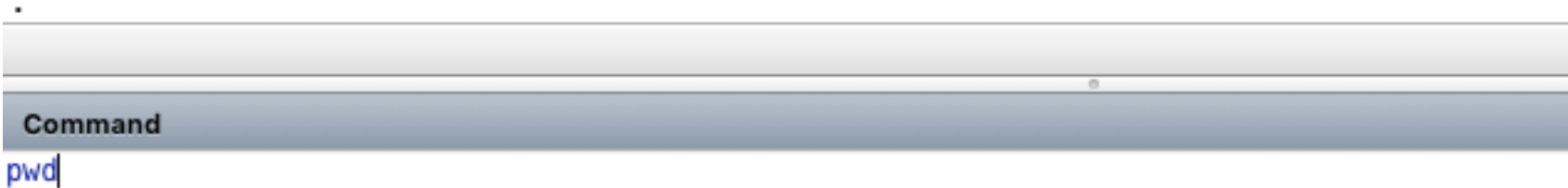
Coronary calcium dataset

- ◇ Coronary calcium – marker for ***atherosclerosis***
- ◇ Cohort study
- ◇ M/F
- ◇ Enrolled in 1985
- ◇ Age 18-30 (at enrollment)
- ◇ Data from year 15 visit

Working Directories

- ◇ Important! Where all your files will be saved/ retrieved from
- ◇ Create sub folders for different projects
- ◇ To see current working directory:
 - Command line
 - `pwd` displays the path of the current working directory

```
.  
$ pwd  
/Users/kristenaiemjoy/Box Sync
```

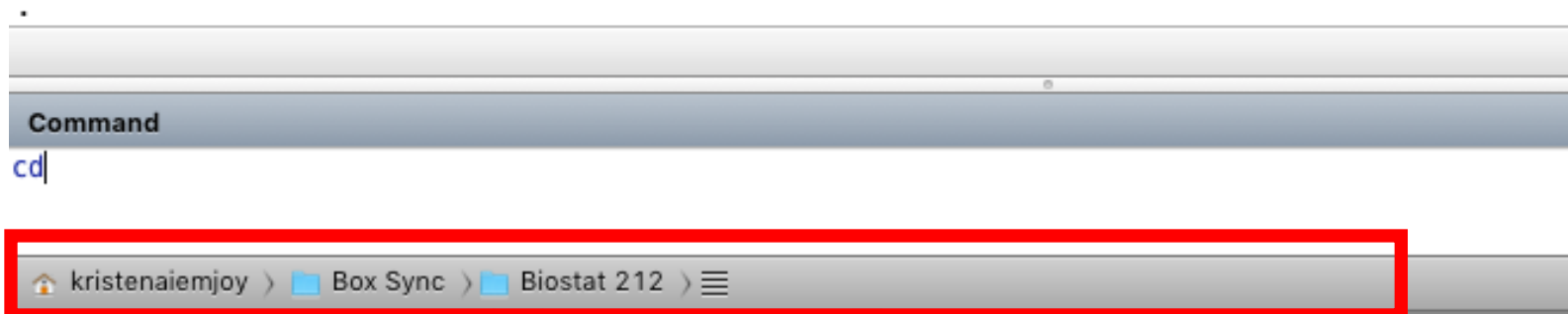


Change working Directories

Three different ways:

1. Command line
 - `cd` change current working directory
 - `dir` view contents of directory
2. Drop-down menu
 - File < Change Working Directory
3. Window shortcut (bottom left)

```
. cd "/Users/kristenaiemjoy/Box Sync/Biostat 212"  
/Users/kristenaiemjoy/Box Sync/Biostat 212
```



Working directories: PC vs. Mac

PC Users

"C:\"

"C:\data\"

"C:\data\biostat212"

cd "C:\data\biostat212\"

Mac Users equivalent

Home directory (also specified by
"/users/yourusername")

"~/Documents" – your Documents folder in your
home directory

Folder called "biostat212" in your Documents
folder

cd "~/Documents/biostat212"

- ◇ *The “~” indicates the use of the Home Directory so you don’t have to specify “/users/yourusername”.*
- ◇ *Mac uses / windows uses *

Opening a data set

Three different ways:

1. Command line

- use `"yourStataDataFile.dta"`, `clear`
 - *Note: this will open only from your current working directory*
 - *Type `dir` to see what files are in your directory*

2. File < Open

3. Import a web-based data set

- webuse `"data_set_name"`

Workflow in STATA

◊ What happens if...

- A question arises about your results?
- You decide to do something differently?
 - Add a new variable to your model
 - Categorize a variable differently
- You get a newer version of the data?
- You lose something?
 - Overwrite your data file, computer crash, etc

ALL OF THESE THINGS WILL HAPPEN TO YOU!

Workflow in STATA

◇ Cardinal Principles

1. Keep your source data pristine and secure
2. Document everything you do to it
3. Document every analysis
4. Make sure you can repeat everything you do easily and quickly and accurately

.do files

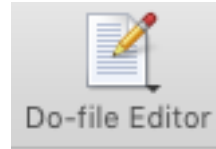
- ◇ A 'recipe' book for your analysis
- ◇ Add comments remind you what is occurring at each stage of analysis
- ◇ Both comments and commands
- ◇ Allows you to repeat analyses quickly
- ◇ Cut and paste
- ◇ Find and replace

Open a .do file

Three different ways:

1. File<new do file

2. Do file button



3. CTRL/CMND + 9

.do files: running commands

Four different ways:

1. Do button in the do file editor
 2. Command line: `do "yourdofile.do"`
 3. "Do..." in the File menu
 4. Control-D button on the keyboard
 1. CMND+shift+d (MAC)
- ◇ .do specific blocks of commands by highlighting them

.do files: making comments

- ◇ “Comments” are a way to document your analysis – here are the options

* Anything after “*” is comment (entire line)

/* Anything until you reach the reciprocal symbol
is comment even on multiple lines*/

//anything after “//” is a comment (same line)

.do files: best practices

??

```
table phototherapy, ///
  c(min birth_wt p5 birth_wt p50 birth_wt p95 birth_wt max birth_wt) ///
  format(%4.2f)
twoway ///
  (kdensity birth_wt if phototherapy==1, bw(0.2) lpattern(solid)) ///
  (kdensity birth_wt if phototherapy==0, bw(0.2) lpattern(longdash)), ///
  ytitle("Density") xtitle("Birth weight (kg)") ///
  legend(order(1 "Phototherapy" 2 "Untreated")) ///
  name(bw_overlap, replace)
foreach x in bwcat male qual_TSB gest_age age_days {
  tab `x' phototherapy, col nokey
}
logistic over_thresh i.phototherapy ib4.qual_TSB ib40.gest_age ///
  ib1.age_days male birth_wt, vce(cluster hospital) cformat(%5.2f) baselevel
scalar b0 = _b[1.phototherapy]
estat gof, group(10)
logistic over_thresh i.phototherapy ib4.qual_TSB ib40.gest_age ///
  ib1.age_days male bw_csp*, vce(cluster hospital)
testparm bw_csp*
testparm bw_csp2-bw_csp4
label list bwcat // reminder of what the categories are
logistic over_thresh i.photo ib4.qual_TSB ib40.gest_age ib1.age_days male ///
  ib3.bwcat, vce(cluster hospital)
testparm i.bwcat
```

```
*/Insider tip: use
*****
To separate sections/*
```

Labels help organize your analysis

```
3 set more off
4 clear all
5
6 /* Modify the cd statement to give the path to where you have put the phototherapy
7 data and this do-file. Note: "~/" is a MAC shortcut for "/Users/username/". */
8
9 cd "~/work/mb/atcr/215/labs"
10
11 /* Prefacing a command with "capture" or "cap" prevents Stata from stopping
12 when it detects an error, as it would in the next line if no log were open. */
13
14 capture log close
15
16 /* If you prefer the log as a text rather than a formatted smcl file, which only
17 looks right in a Stata viewer, remove the two forward slashes before the
18 option "text" in the next line. (Stata ignores anything after the two slashes.) */
19
20 log using "lab1", replace // text
21
22 /* You may need to modify the next line to read in the data. If you put the
23 dataset in your current working directory, the next line should work. */
24
25 cap use "phototherapy.dta", clear
26
27 /* Otherwise, you may need to specify the whole path, as this line does for
28 TN's computer. */
29
30 cap use "/Users/thomasnewman/Box Documents/phototherapy.dta"
31
32 ***** PART 1: checking for covariate overlap *****
```

.do file template

```
1  cd "C:\data\biostat212\"    // state the working directory where your files are
2
3  capture log close           //close any log files that may be open
4  log using "Lab2_yourname.smcl", replace
5
6  set more off                // tells stata not to pause or display the -more- message
7
8  version 13                  // assumes you are using Stata 13
9
10
11 /*  Lab 2
12
13 */
14
15 use "Coronary Calcium.dta", clear
16
17 ***** Analysis
18 tab dm
19 tab smoke
20 tab chol
21 *****
22
23 log close
24
25 save "Coronary Calcium_NEW.dta", replace
26
```

Log files

- ◇ A record of everything of all your analyses Time/date
 - Commands + output
 - Including commands run from the .do file and command line
- ◇ Documents analysis on data *as it was at that time* (.do files don't do this)
- ◇ Two file types
 - **.log**: plain text, easy to open share
 - **.smcl**: native Stata format, difficult to make changes (advantage)
- ◇ **Do not edit**

Log files

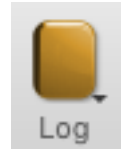
```
log using yourlogname.log  
    , append  
    , replace  
log close  
log off  
log on
```

```
open  
add to end  
replace  
close  
pause  
resume
```

Open a log file

Four different ways:

1. File <log <begin/append



2. Log button (begin/append)

3. Command line

- `log using yourlogname.smcl, replace`

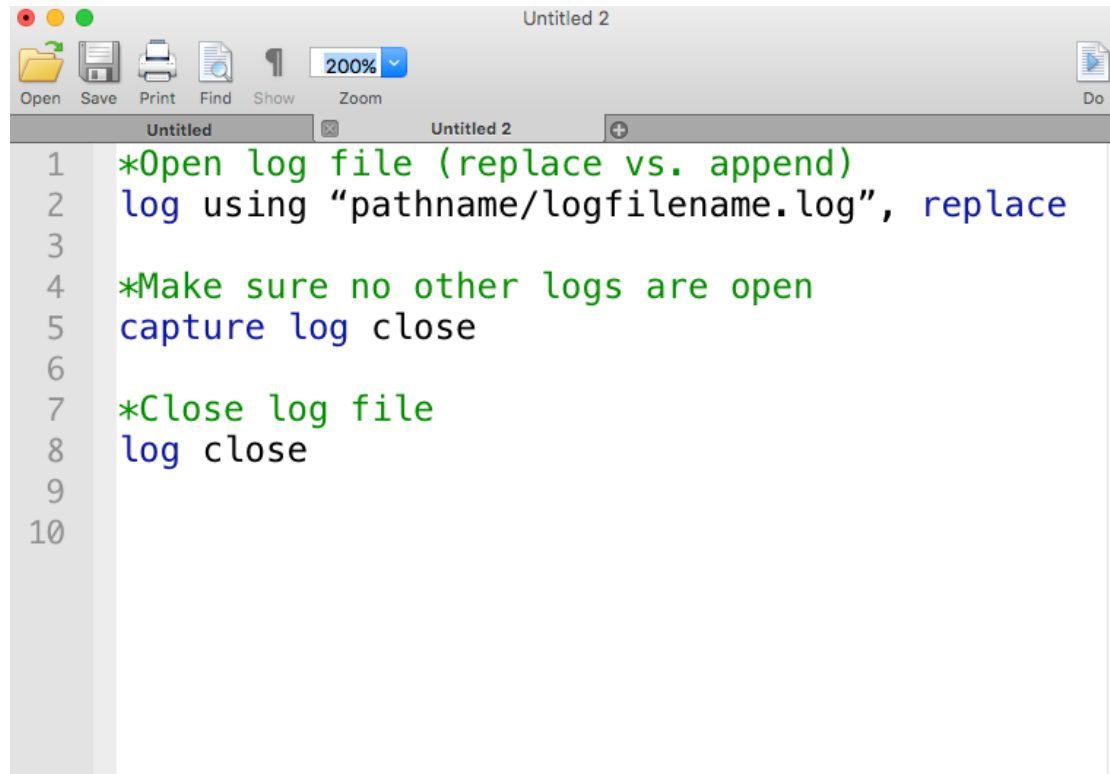
4. From .do file ****recommended**

- `log using yourlogname.smcl, replace`

.do files and log files together

- ◇ Open/close the log file within the .do file
 - Everything documented every time
 - Easily repeatable

.do files and log files together



```
1  *Open log file (replace vs. append)
2  log using "pathname/logfilename.log", replace
3
4  *Make sure no other logs are open
5  capture log close
6
7  *Close log file
8  log close
9
10
```

FYI command log files

- ◇ List of all commands you enter
 - From both .do file and command line
 - cmdlog using
 - cmdlog close
 - cmdlog off
 - cmdlog on

****Preview: exploring data**

tabulate `varname`

– One-way table of frequencies

tabulate `varname1 varname1`

– Two-way table of frequencies

tabulate `varname1 varname1, if`

– Use logic operators

example `tab chol smoke if age<50`

tabulate `varname1 varname1, missing`

-include missing data

In class exercise