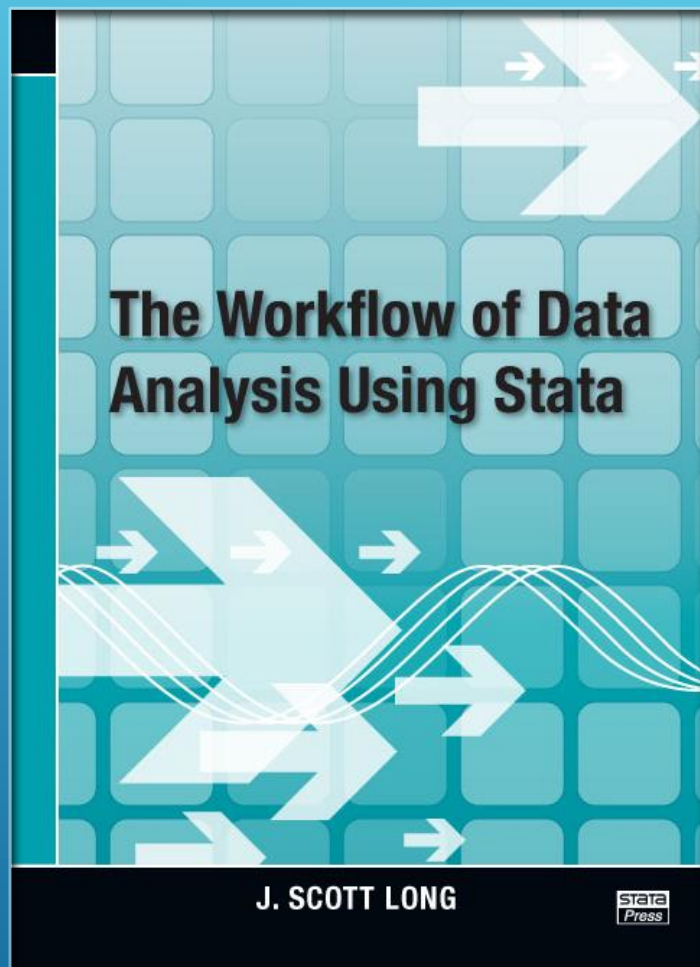


# RESEARCH WORKFLOW USING STATA

How to Be an Effective Researcher  
CCPR Workshop



# THE WORKFLOW OF DATA ANALYSIS USING STATA

J. Scott Long

# THREE WAYS TO EXECUTE COMMANDS

The screenshot displays the Stata/MP 13.0 interface with the following components:

- Review Panel:** Shows a list of commands executed, including `regress api00 acs_k3 meals full`.
- Main Output Area:** Displays the results of the regression command, including summary statistics for `yr_rnd` and the full regression output.
- Variables Panel:** Lists all variables in the dataset with their labels, such as `snum` (school number) and `acs_k3` (avg class size k-3).
- Properties Panel:** Shows details for the selected variable `snum`, including its type (int) and format (%9.0g).
- Command Panel:** Shows the command `regress api00 acs_k3 meals full` being executed.

**Summary Statistics for yr\_rnd:**

mean:	66.0568
std. dev:	40.2979
percentiles:	10% 25% 50% 75% 90%
	.67 .95 87 97 100

**Regression Results:**

Source	SS	df	MS	Number of obs =
Model	2634884.26	3	878294.754	313
Residual	1271713.21	309	4115.57673	F( 3, 309) = 213.41
Total	3906597.47	312	12521.1457	Prob > F = 0.0000

**Regression Coefficients:**

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
api00					
acs_k3	-2.681508	1.393991	-1.92	0.055	-5.424424 .0614074
meals	-3.702419	.1540256	-24.04	0.000	-4.005491 -3.399348
full	.1086104	.090719	1.20	0.232	-.0698947 .2871154
_cons	906.7392	28.26505	32.08	0.000	851.1228 962.3555

**Command:** `regress api00 acs_k3 meals full`

# THREE WAYS TO EXECUTE COMMANDS

The screenshot displays the Stata software interface with the following components:

- Statistics Menu:** A dropdown menu is open, showing 'Linear models and related' selected. The 'Linear regression' sub-menu is also open, listing options like 'Regression diagnostics', 'ANOVA/MANOVA', 'Constrained linear regression', etc.
- Review Panel:** Shows a list of commands executed, including 'regress api00 acs\_k3 meals'.
- Results Window:** Displays the output of a regression analysis. It includes a table of coefficients and a summary of statistics.
- Variables Panel:** Lists the variables in the dataset, such as 'snum', 'dnum', 'api00', 'meals', etc.
- Properties Panel:** Shows the properties of the selected variable 'snum', including its name, label, type, and format.

**Regression Results Table:**

	Std. Err.	t	P> t	[95% Conf. Interval]		
acs_k3	-2.681508	1.393991	-1.92	0.055	-5.424424	.0614074
meals	-3.702419	.1540256	-24.04	0.000	-4.005491	-3.399348
full	.1086104	.090719	1.20	0.232	-.0698947	.2871154
_cons	906.7392	28.26505	32.08	0.000	851.1228	962.3555

**Summary Statistics:**

	df	MS	Number of obs =
3	3	878294.754	313
21	309	4115.57673	F( 3, 309) = 213.41
			Prob > F = 0.0000
			R-squared = 0.6745
			Adj R-squared = 0.6713
			Root MSE = 64.153

**Command:**

```
regress api00 acs_k3 meals
```

# THREE WAYS TO EXECUTE COMMANDS

The screenshot displays the Stata/MP 13.0 interface. The main window shows the results of a regression command: `regress api00 acs_k3 meals full`. The results include summary statistics for the dependent variable `yr_rnd`, a source table, and a coefficient table. A purple circle highlights the 'Results' icon in the top toolbar. On the right, the 'Variables' window lists 21 variables, and the 'Properties' window shows details for the variable `snum`.

**Review**

#	Command
1	use http://www.ats.ucla.ed...
2	regress api00 acs_k3 meals ...
3	describe
4	list in 1/5
5	codebook api00 acs_k3 me...
6	regress api00 acs_k3 meals ...

**Summary Statistics**

```
      mean:    66.0568
      std. dev: 40.2979

      percentiles:    10%    25%    50%    75%    90%
                    .67    .95    87    97    100
```

**Variable Information**

```
yr_rnd                                     year round school

      type: numeric (byte)
      label: yr_rnd

      range: [0,1]                               units: 1
      unique values: 2                           missing .: 0/400

      tabulation: Freq.  Numeric  Label
                  308      0  No
                  92      1  Yes
```

**Regression Results**

```
. regress api00 acs_k3 meals full
```

Source	SS	df	MS	Number of obs = 313		
Model	2634884.26	3	878294.754	F( 3, 309) =	213.41	
Residual	1271713.21	309	4115.57673	Prob > F =	0.0000	
Total	3906597.47	312	12521.1457	R-squared =	0.6745	
				Adj R-squared =	0.6713	
				Root MSE =	64.153	

	api00	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
acs_k3		-2.681508	1.393991	-1.92	0.055	-5.424424 .0614074
meals		-3.702419	.1540256	-24.04	0.000	-4.005491 -3.399348
full		.1086104	.090719	1.20	0.232	-.0698947 .2871154
_cons		906.7392	28.26505	32.08	0.000	851.1228 962.3555

**Command**

```
regress api00 acs_k3 meals full
```

**Variables**

Variable	Label
snum	school number
dnum	district number
api00	api 2000
api99	api 1999
growth	growth 1999 to 2000
meals	pct free meals
ell	english language learners
yr_rnd	year round school
mobility	pct 1st year in school
acs_k3	avg class size k-3
acs_46	avg class size 4-6
not_hsg	parent not hsg
hsg	parent hsg
some_col	parent some college
col_grad	parent college grad
grad_sch	parent grad school
avg_ed	avg parent ed
full	pct full credential
emer	pct emer credential
enroll	number of students
mealcat	Percentage free meals in 3 categories

**Properties**

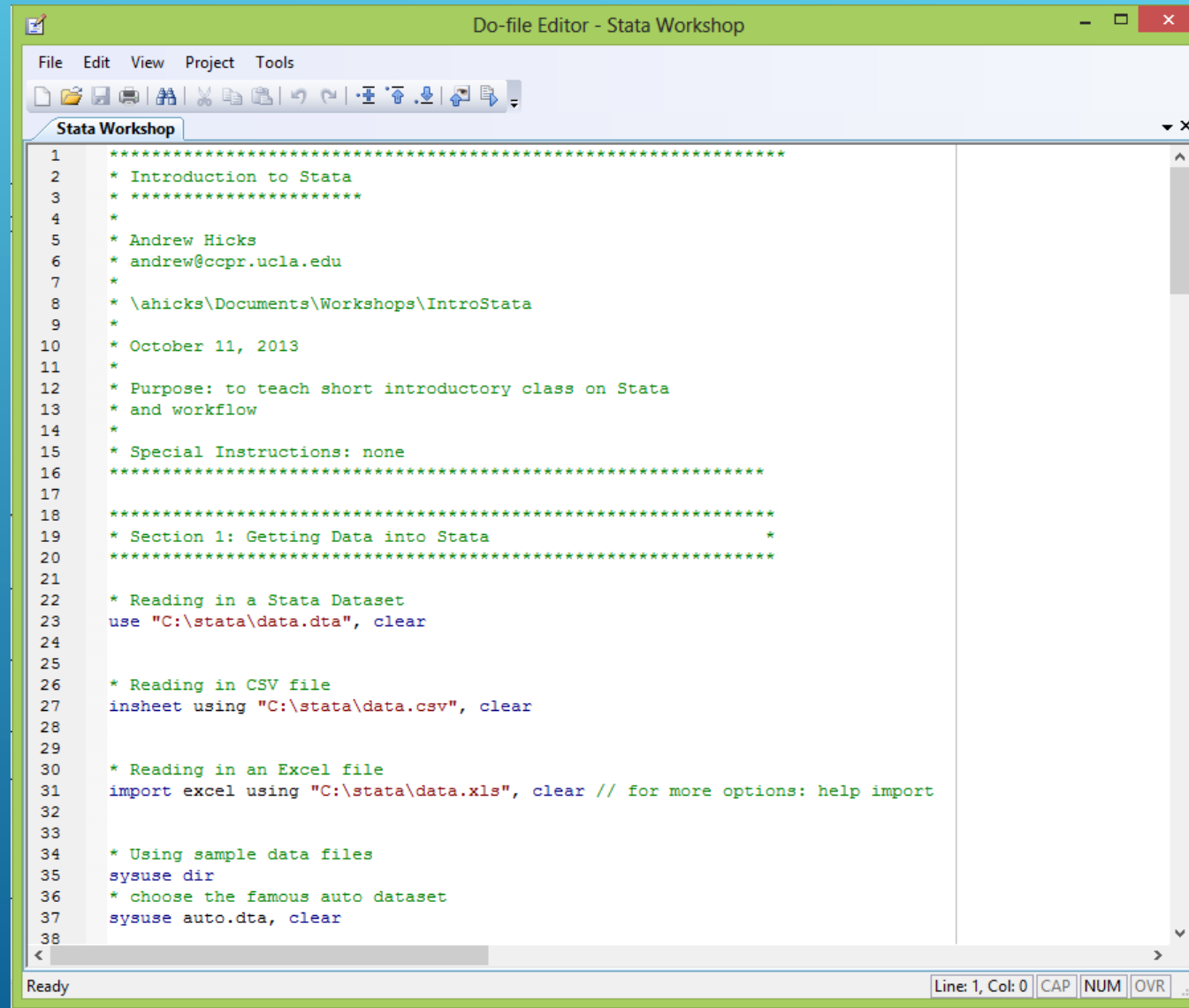
Variables	
Name	snum
Label	school number
Type	int
Format	%9.0g
Value Label	
Notes	

**Data**

Filename	
Filename	elemapi.dta
Label	
Notes	
Variables	21
Observations	400
Size	13.28K
Memory	64M
Sorted by	dnum

C:\Users\ahicks\Documents CAP NUM OVR

# DO-FILES



The screenshot shows the 'Do-file Editor - Stata Workshop' window. The menu bar includes 'File', 'Edit', 'View', 'Project', and 'Tools'. The toolbar contains icons for file operations like opening, saving, and printing. The main text area contains a Stata do-file script with the following content:

```
1 *****
2 * Introduction to Stata
3 * *****
4 *
5 * Andrew Hicks
6 * andrew@ccpr.ucla.edu
7 *
8 * \ahicks\Documents\Workshops\IntroStata
9 *
10 * October 11, 2013
11 *
12 * Purpose: to teach short introductory class on Stata
13 * and workflow
14 *
15 * Special Instructions: none
16 *****
17
18 *****
19 * Section 1: Getting Data into Stata *
20 *****
21
22 * Reading in a Stata Dataset
23 use "C:\stata\data.dta", clear
24
25
26 * Reading in CSV file
27 insheet using "C:\stata\data.csv", clear
28
29
30 * Reading in an Excel file
31 import excel using "C:\stata\data.xls", clear // for more options: help import
32
33
34 * Using sample data files
35 sysuse dir
36 * choose the famous auto dataset
37 sysuse auto.dta, clear
38
```

The status bar at the bottom indicates 'Ready' and 'Line: 1, Col: 0'.

# ADVANTAGES OF USING DO-FILES

1. You have a record of the commands you ran.

You can rerun them in the future to replicate your results  
You can quickly modify your code

2. You can use the features of the text editor

i.e. copy/paste, find and replace, select all

# TWO RULES OF USING DO-FILES

## 1. Do-files must be robust

Robust do-files produce exactly the same result when run at a later time or on another computer

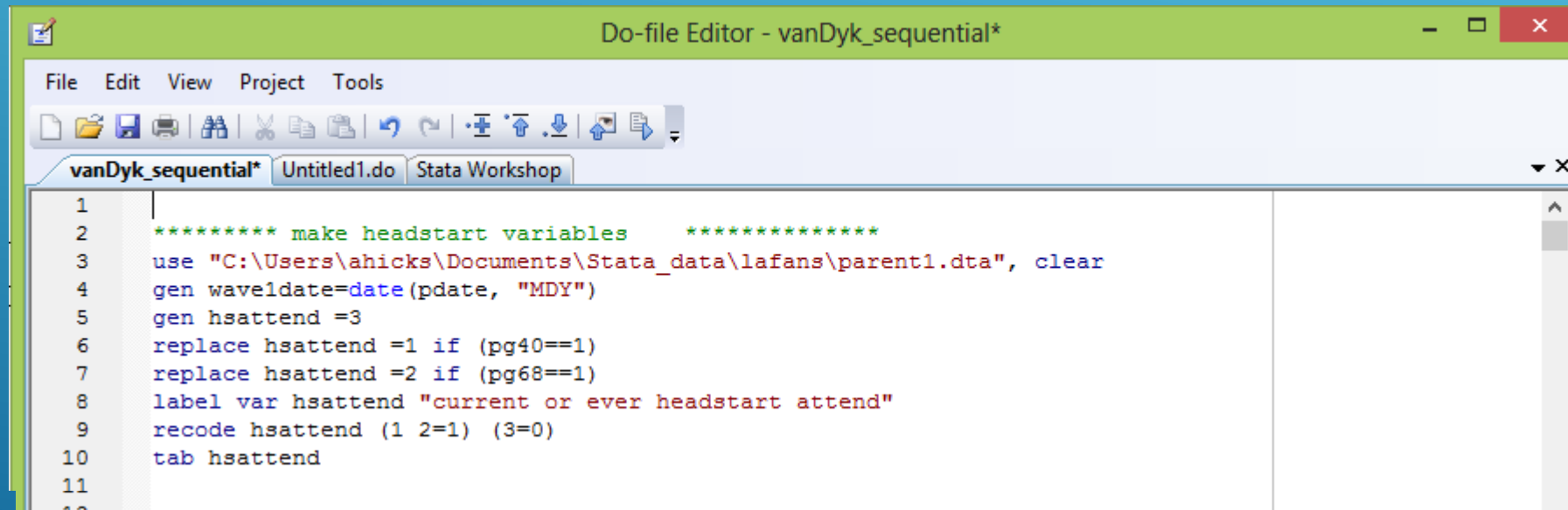
## 2. Do-files must be legible

Legible do-files are documented and formatted so that it is easy to understand what is being done



# MAKING DO-FILES ROBUST

Robust do-files are self-contained



The screenshot shows a window titled "Do-file Editor - vanDyk\_sequential\*" with a menu bar (File, Edit, View, Project, Tools) and a toolbar. The editor contains the following Stata code:

```
1  
2 ***** make headstart variables *****  
3 use "C:\Users\ahicks\Documents\Stata_data\lafans\parent1.dta", clear  
4 gen wave1date=date(pdate, "MDY")  
5 gen hsattend =3  
6 replace hsattend =1 if (pg40==1)  
7 replace hsattend =2 if (pg68==1)  
8 label var hsattend "current or ever headstart attend"  
9 recode hsattend (1 2=1) (3=0)  
10 tab hsattend  
11  
12
```

# MAKING DO-FILES ROBUST

Exclude directory information

```
1  
2 ***** load dataset using directory location *****  
3 use "C:\Users\ahicks\Documents\Stata_data\lafans\parent1.dta", clear  
4  
5 ***** load dataset without directory location *****  
6 cd C:\Users\ahicks\Documents\Stata_data\lafans  
7 use parent1.dta
```

# MAKING DO-FILES ROBUST

Use version control

```
. version 14  
this is version 13.0 of Stata; it cannot run version 14.0 programs  
    You can purchase the latest version of Stata by visiting http://www.stata.com.  
r(9);
```

Include seeds for random numbers

```
. set seed 90049
```

# MAKING DO-FILES LEGIBLE

Legible do-files are internally documented and formatted

Use comments!!

```
1 *****
2 *           Use Comments!           *
3 *****
4
5 * Stata treats the entire line as a comment if the line starts with a *
6
7 gen waveldate=date(pdate, "MDY")
8 gen hsattend =3
9 *replace hsattend =1 if (pg40==1)
10 replace hsattend =2 if (pg68==1)
11
12
13
14 /* You can create comments across multiple lines by using an opening
15  '/*' and a closing'*/'. You can also use this type of comment to
16  temporarily stop entire sections of code from being executed */
17
18 gen waveldate=date(pdate, "MDY")
19 gen hsattend =3
20 /*
21  replace hsattend =1 if (pg40==1)
22  replace hsattend =2 if (pg68==1)
23  label var hsattend "current or ever headstart attend"
24  */
25 recode hsattend (1 2=1) (3=0)
26 tab hsattend
27
28
29
30 // Any thing that comes after a double slash is treat as a comment
31
32 logit lfp wc hc // this analysis only includes education, add wages next
33
```

# MAKING DO-FILES LEGIBLE

Use alignment and indentation

```
*****
* Use Alignment and indentation                                     *
*****

* This code:
logit foreign price mpg trunk weight length turn displacement gear_ratio

* Is the same as this code:
logit foreign    price      mpg      trunk ///
               weight     length    turn  ///
               displace    gear_ratio

* Use short lines
mi estimate, post nois saving(lwi_sub5_1, replace): regress W2_lwi_ss condis_average childmale birth
```

# MAKING DO-FILES LEGIBLE

Limit your abbreviations

```
*****  
* Be careful with abbreviations *  
*****  
* three commands that give the same results:  
summarize displacement  
su d  
sum displace  
  
* What on earth does this mean?  
l ma p f in 1/3
```

# SAVING YOUR SESSION TO A LOG

```
File Edit View Project Tools
use comments vanDyk_sequential* wd, alignments, and abbreviations* Stata Workshop*
1 *****
2 * LA Fans Dataset Creation
3 * *****
4 *
5 * Andrew Hicks
6 * andrew@ccpr.ucla.edu
7 *
8 * C:\Users\ahicks\Documents\Stata_data\lafan
9 *
10 * October 11, 2013
11 *
12 * Purpose: Creates dataset for LAFans dataset to model neighborhood
13 *      disadvantage
14 *
15 * Special Instructions: none
16 *****
17
18 cd C:\Users\ahicks\Documents\Stata_data\lafans
19
20 log using dataset_creation.txt, replace text
21
22 ***** make headstart variables *****
23 use "C:\Users\ahicks\Documents\Stata_data\lafans\parent1.dta", clear
24 gen waveldate=date(pdate, "MDY")
25 gen hsattend =3
26 replace hsattend =1 if (pg40==1)
27 replace hsattend =2 if (pg68==1)
28 label var hsattend "current or ever headstart attend"
29 recode hsattend (1 2=1) (3=0)
30 tab hsattend
31
32 log close
33
34
```

```
-----
name: <unnamed>
log:
C:\Users\ahicks\Documents\Stata_data\lafans\dataset_creation.txt
log type: text
opened on: 21 Oct 2013, 17:40:28

.
. ***** make headstart variables *****
. use "C:\Users\ahicks\Documents\Stata_data\lafans\parent1.dta", clear
(LAFANS-1 Parent module: Jun 2004)

. gen waveldate=date(pdate, "MDY")

. gen hsattend =3

. replace hsattend =1 if (pg40==1)
(51 real changes made)

. replace hsattend =2 if (pg68==1)
(331 real changes made)

. label var hsattend "current or ever headstart attend"

. recode hsattend (1 2=1) (3=0)
(hsattend: 3078 changes made)

. tab hsattend

current or |
ever |
headstart |
attend |      Freq.      Percent      Cum.
-----+-----
          0 |      2,747      87.79      87.79
          1 |       382      12.21     100.00
-----+-----
Total |      3,129     100.00
```