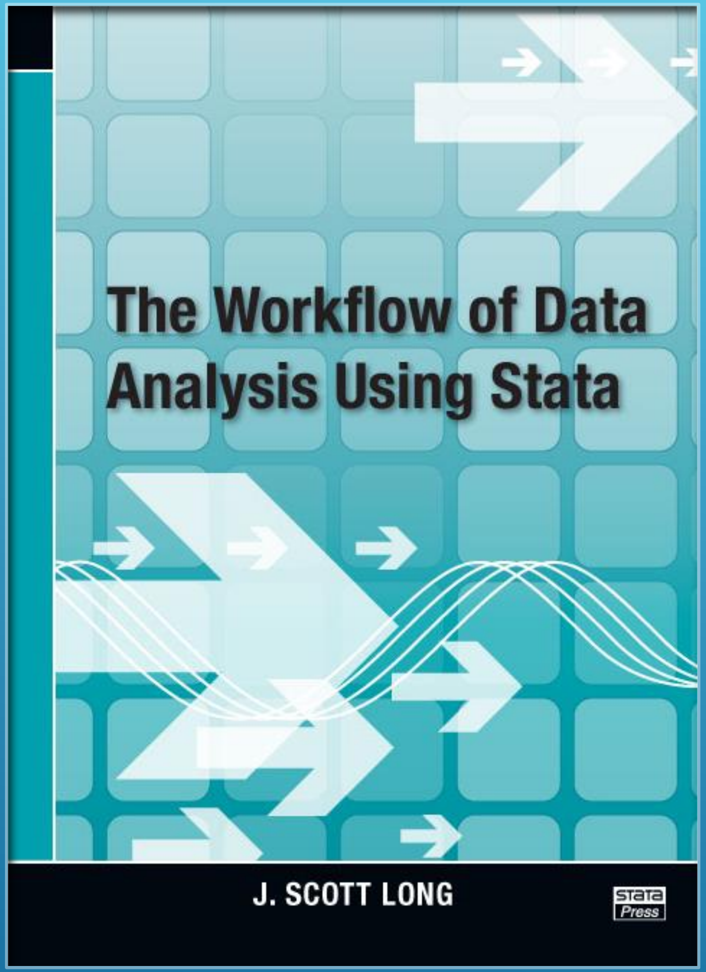
RESEARCH WORKFLOW USING STATA

How to Be an Effective Researcher

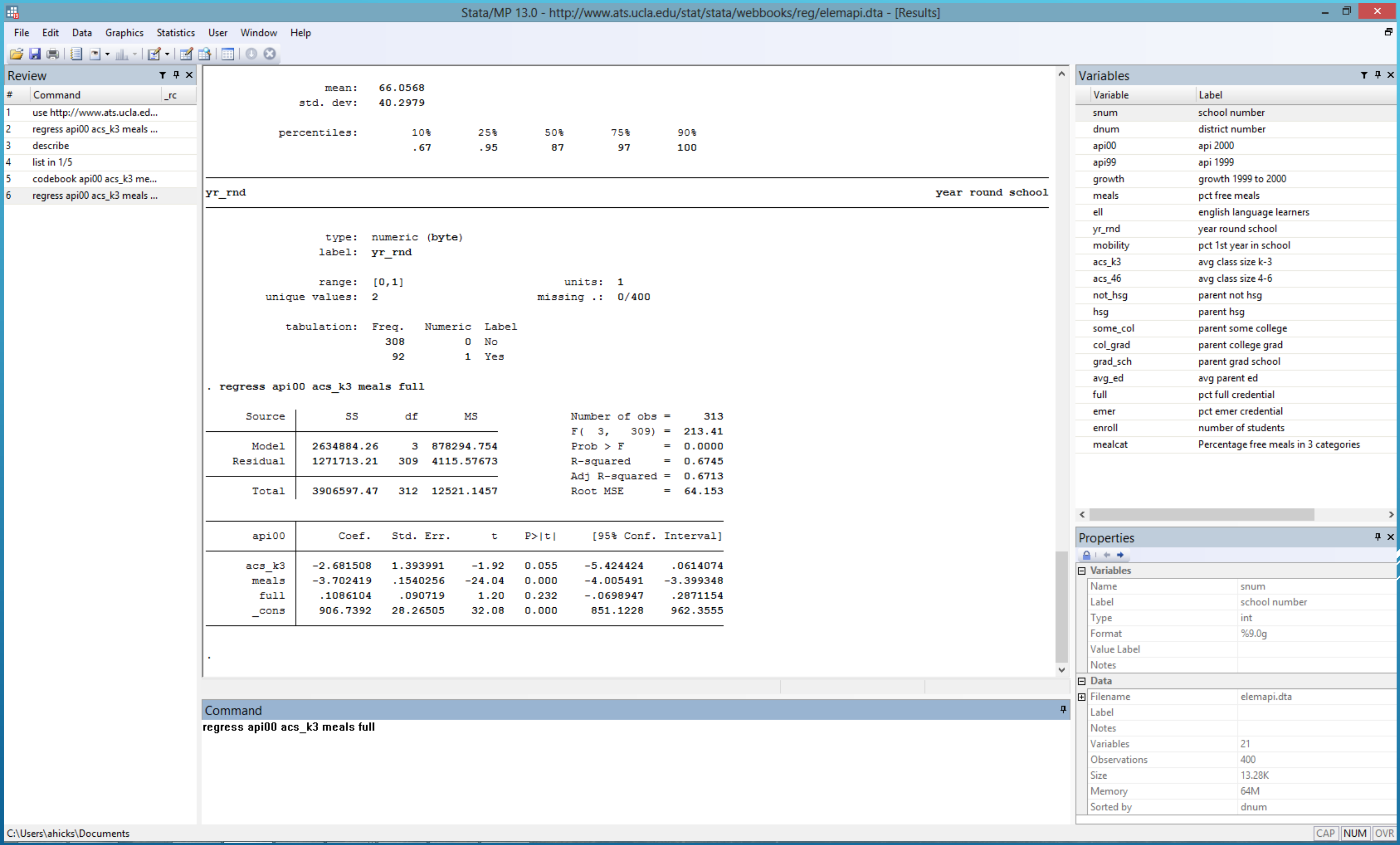
CCPR Workshop



THE WORKFLOW OF DATA ANALYSIS USING STATA

J. Scott Long

THREE WAYS TO EXECUTE COMMANDS



Stata/MP 13.0 - http://www.ats.ucla.edu/stat/stata/webbooks/reg/elemapi.dta - [Results]
File
Edit
Data
Graphics
Statistics
User
Window
Help
Review
Variables
Properties
Command
CAP
NUM
OVR

#	Command	_rc
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 ...	

```
                  mean:   66.0568
              std. dev:   40.2979

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

------------------------------------------------------------------------------
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

. regress api00 acs_k3 meals full

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

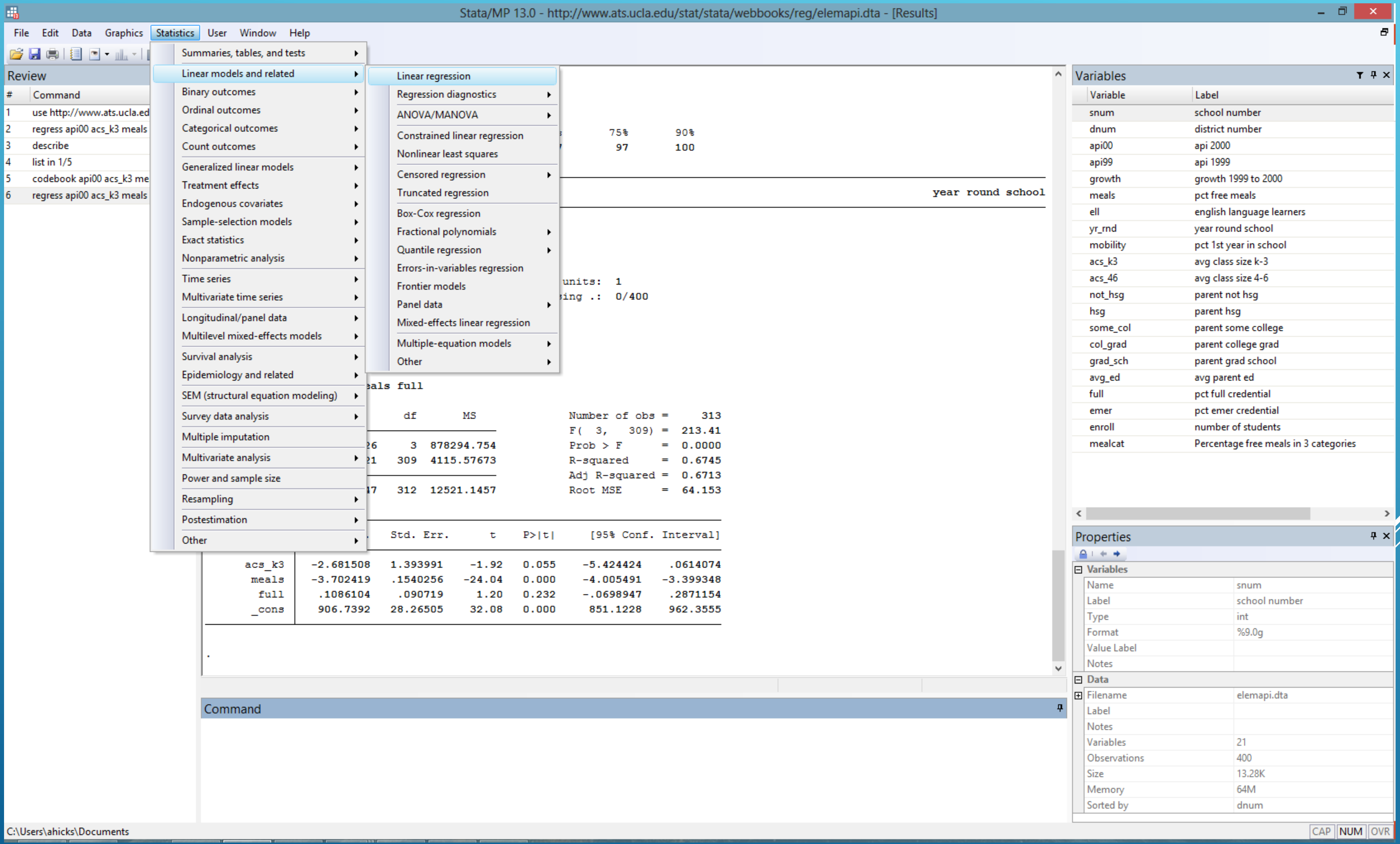
```
regress api00 acs_k3 meals full
```

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

Variables	
Name	snum
Label	school number
Type	int
Format	%9.0g
Value Label	
Notes	
Data	
Filename	elemapi.dta
Label	
Notes	
Variables	21
Observations	400
Size	13.28K
Memory	64M
Sorted by	dnum

C:\Users\ahicks\Documents

THREE WAYS TO EXECUTE COMMANDS



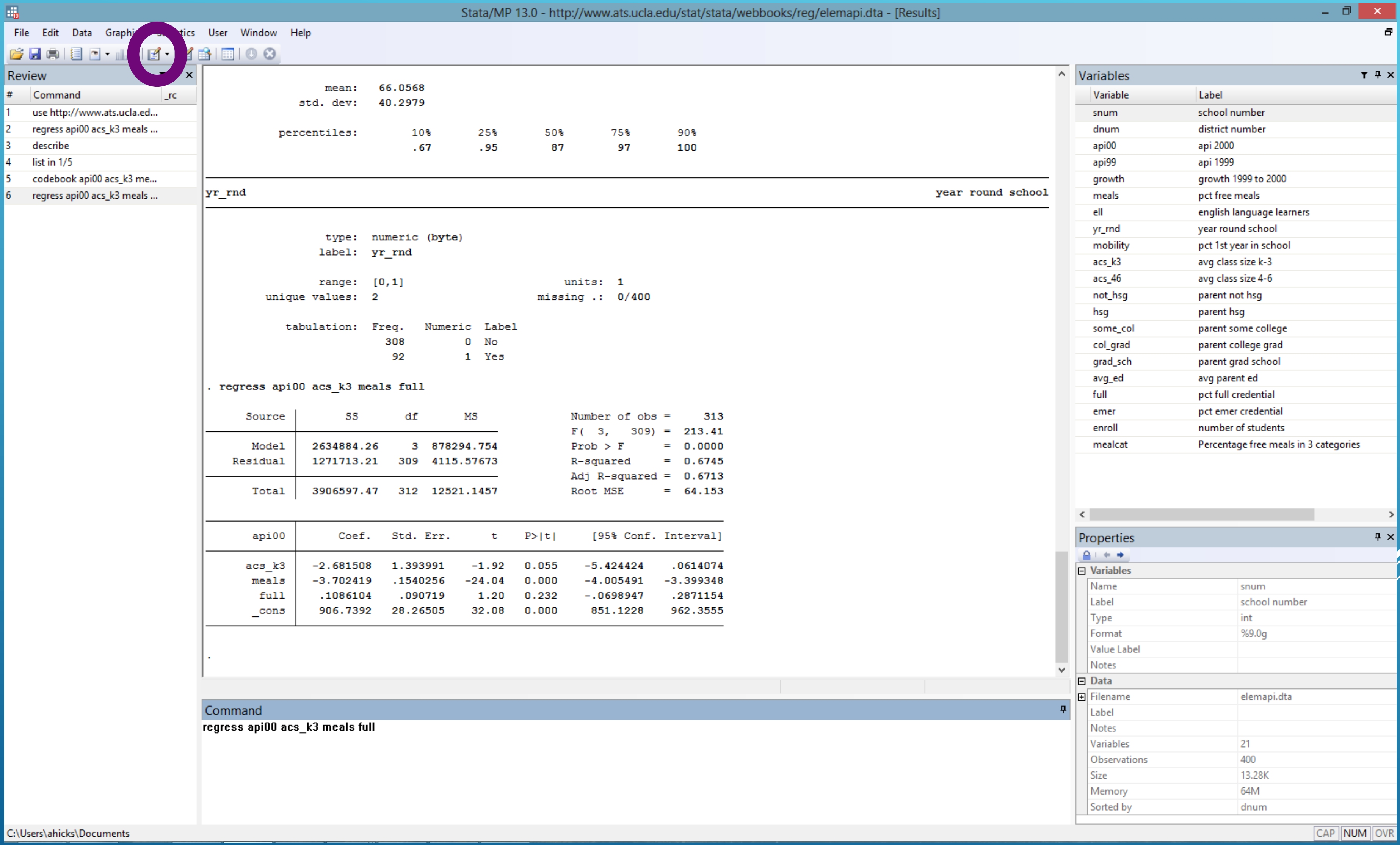
#	Command
1	use http://www.ats.ucla.ed
2	regress api00 acs_k3 meals
3	describe
4	list in 1/5
5	codebook api00 acs_k3 me
6	regress api00 acs_k3 meals

```
                      75%       90%
                       97       100

                                                         year round school

units:  1
ing .:  0/400

eals full

        df       MS              Number of obs =     313
                                 F(  3,   309) =  213.41
26       3  878294.754           Prob > F      =  0.0000
21     309  4115.57673           R-squared     =  0.6745
                                 Adj R-squared =  0.6713
17     312  12521.1457           Root MSE      =  64.153

             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

.
```

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

⊟ Variables	
Name	snum
Label	school number
Type	int
Format	%9.0g
Value Label	
Notes	
⊟ Data	
⊞ Filename	elemapi.dta
Label	
Notes	
Variables	21
Observations	400
Size	13.28K
Memory	64M
Sorted by	dnum

THREE WAYS TO EXECUTE COMMANDS


Stata/MP 13.0 - http://www.ats.ucla.edu/stat/stata/webbooks/reg/elemapi.dta - [Results]
File
Edit
Data
Graphics
Statistics
User
Window
Help
Review
Variables
Properties
Command
CAP
NUM
OVR

#	Command	_rc
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 ...	

```
                mean:   66.0568
            std. dev:   40.2979

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

-------------------------------------------------------------------------------
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

. regress api00 acs_k3 meals full

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

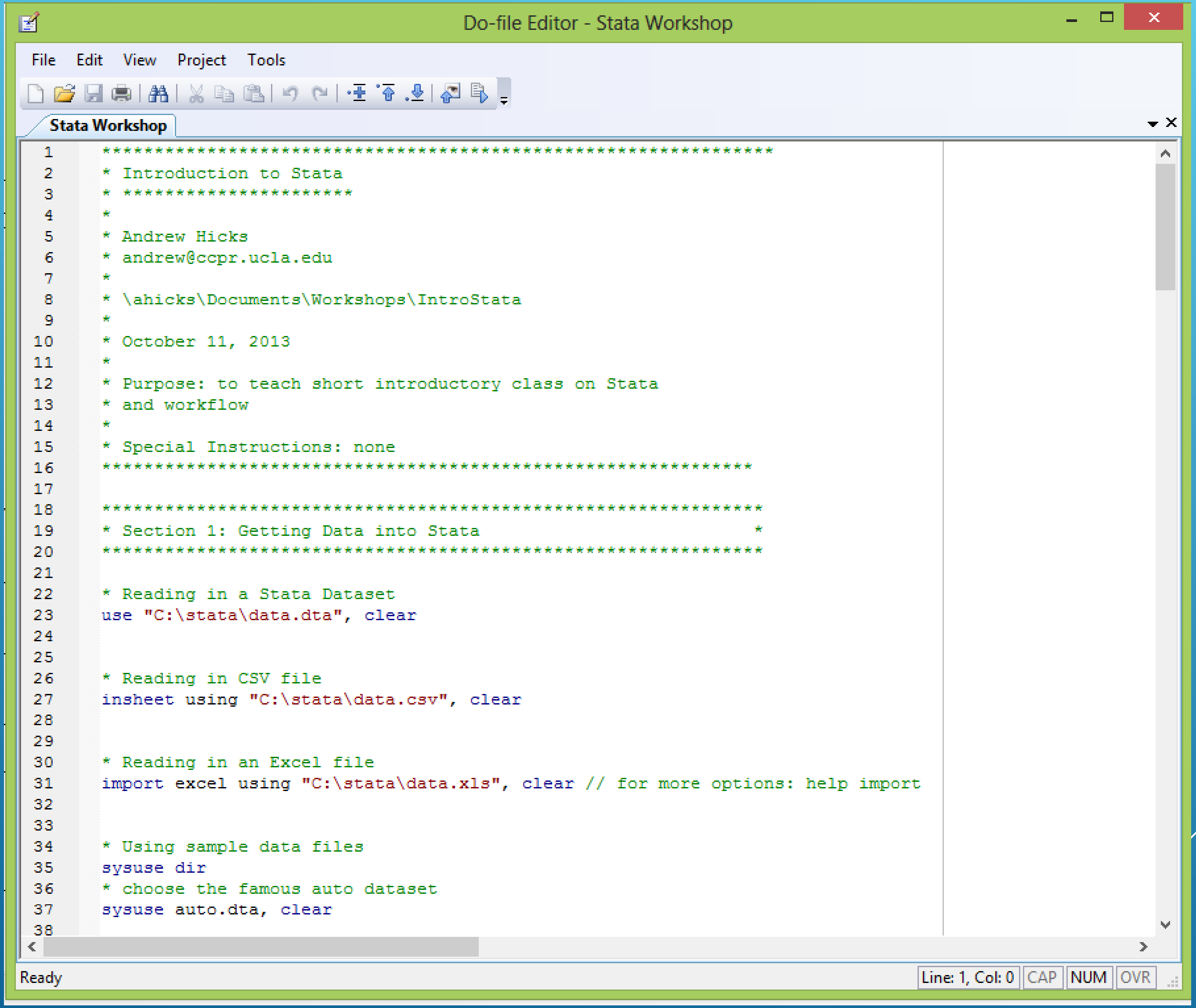
```
regress api00 acs_k3 meals full
```

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

Variables	
Name	snum
Label	school number
Type	int
Format	%9.0g
Value Label	
Notes	
Data	
Filename	elemapi.dta
Label	
Notes	
Variables	21
Observations	400
Size	13.28K
Memory	64M
Sorted by	dnum

C:\Users\ahicks\Documents

DO-FILES



```
****************************************************************
* Introduction to Stata
* **********************
*
* Andrew Hicks
* andrew@ccpr.ucla.edu
*
* \ahicks\Documents\Workshops\IntroStata
*
* October 11, 2013
*
* Purpose: to teach short introductory class on Stata
* and workflow
*
* Special Instructions: none
**************************************************************

***************************************************************
* Section 1: Getting Data into Stata                          *
***************************************************************

* Reading in a Stata Dataset
use "C:\stata\data.dta", clear


* Reading in CSV file
insheet using "C:\stata\data.csv", clear


* Reading in an Excel file
import excel using "C:\stata\data.xls", clear // for more options: help import


* Using sample data files
sysuse dir
* choose the famous auto dataset
sysuse auto.dta, clear
```

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

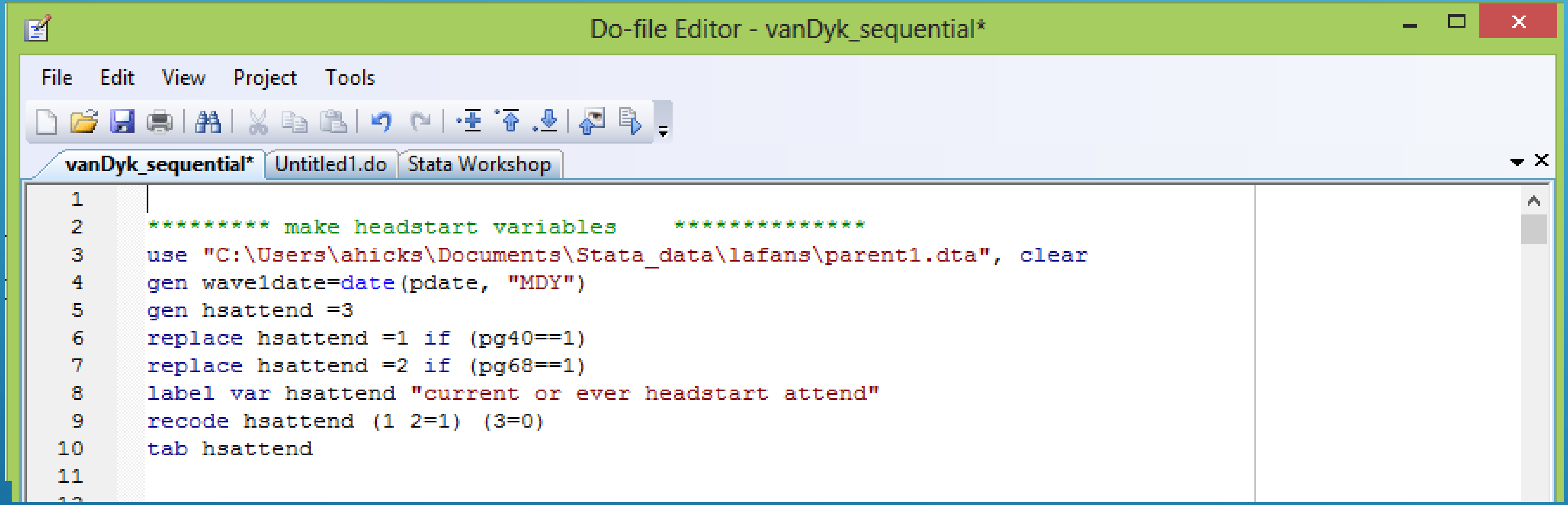
   Robust do-files produce exactly the same result when run at a later time or on another computer

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

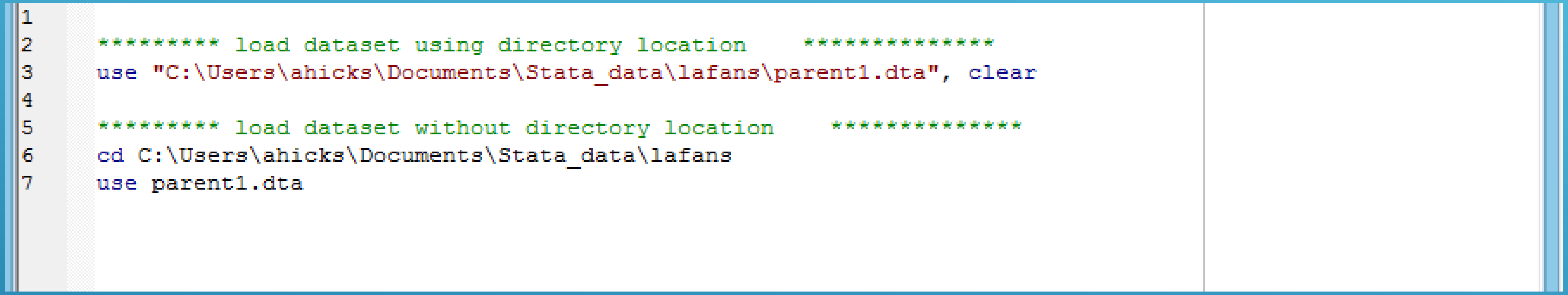
Robust do-files are self-contained



```

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

```

MAKING DO-FILES ROBUST

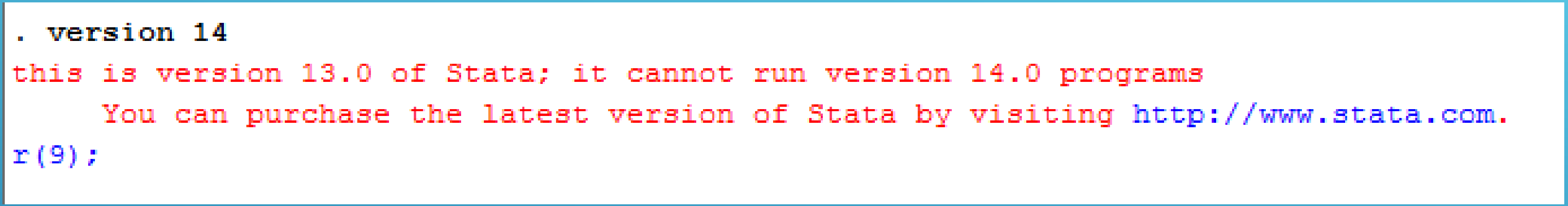
Exclude directory information



```

********* load dataset using directory location    ***************
use "C:\Users\ahicks\Documents\Stata_data\lafans\parent1.dta", clear

********* load dataset without directory location    ***************
cd C:\Users\ahicks\Documents\Stata_data\lafans
use parent1.dta
```

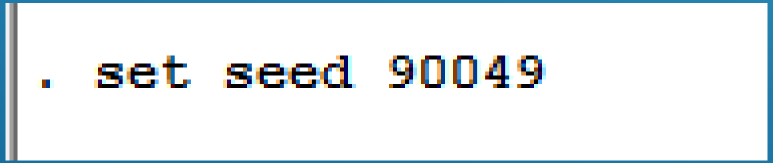
MAKING DO-FILES ROBUST

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

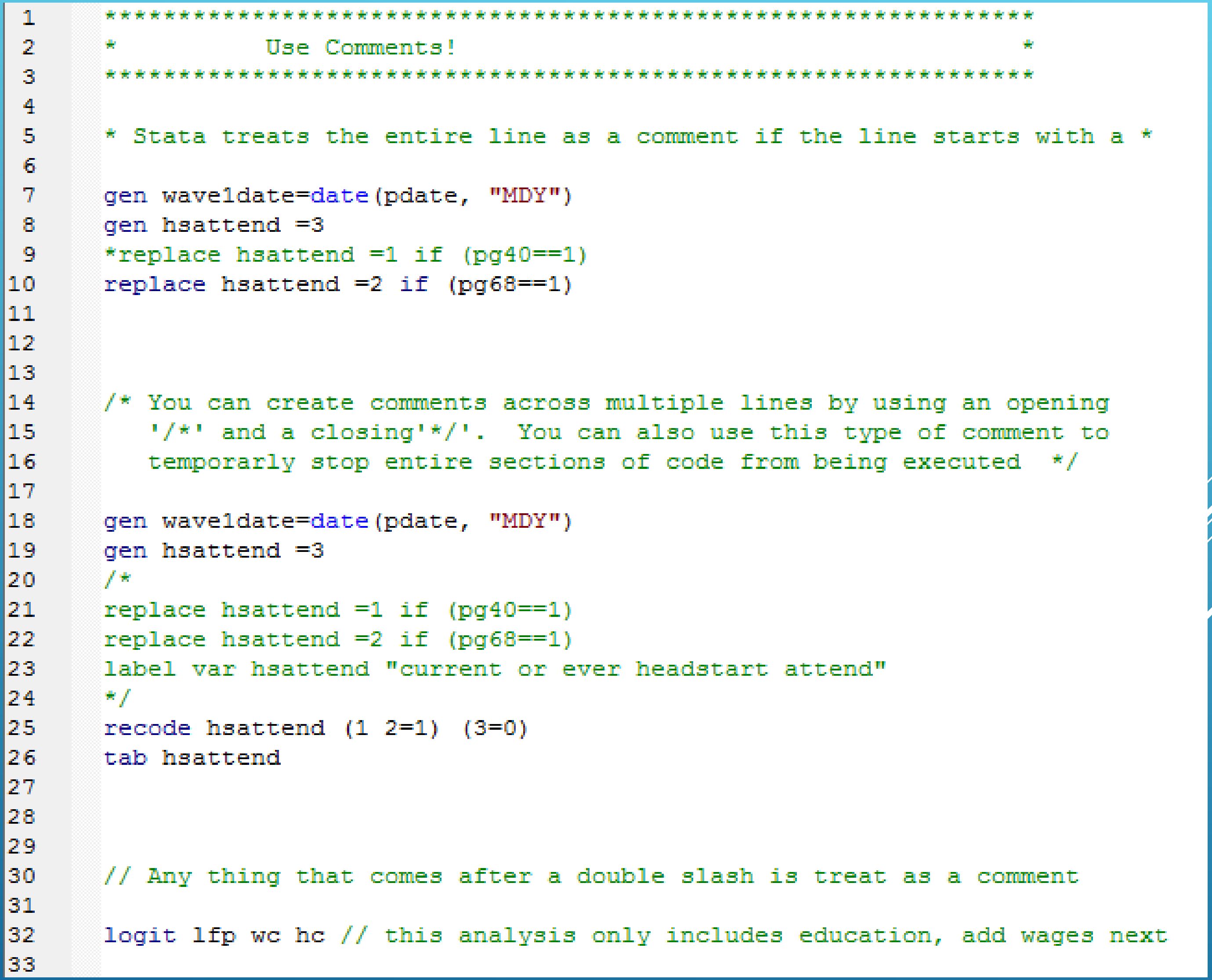


```
. set seed 90049
```

MAKING DO-FILES LEGIBLE

Legible do-files are internally documented and formatted

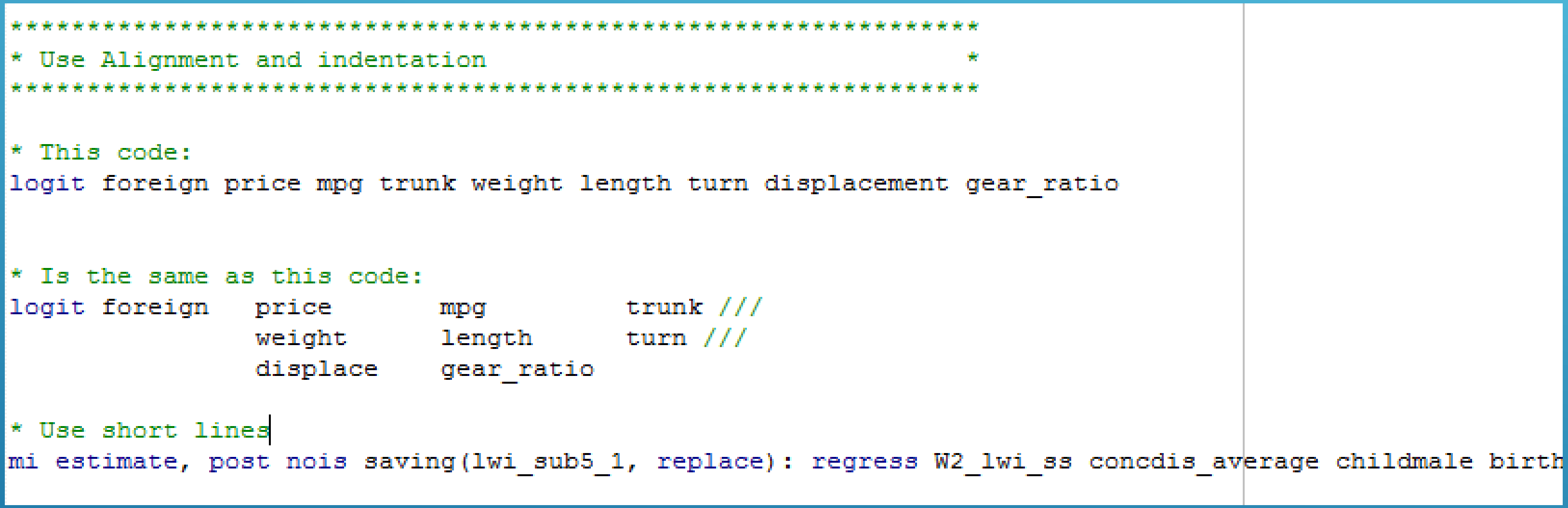
Use comments!!



```
****************************************************************
*          Use Comments!                                       *
****************************************************************

* Stata treats the entire line as a comment if the line starts with a *

gen wave1date=date(pdate, "MDY")
gen hsattend =3
*replace hsattend =1 if (pg40==1)
replace hsattend =2 if (pg68==1)



/* You can create comments across multiple lines by using an opening
   '/*' and a closing'*/'.  You can also use this type of comment to
   temporarly stop entire sections of code from being executed  */

gen wave1date=date(pdate, "MDY")
gen hsattend =3
/*
replace hsattend =1 if (pg40==1)
replace hsattend =2 if (pg68==1)
label var hsattend "current or ever headstart attend"
*/
recode hsattend (1 2=1) (3=0)
tab hsattend



// Any thing that comes after a double slash is treat as a comment

logit lfp wc hc // this analysis only includes education, add wages next

```

MAKING DO-FILES LEGIBLE

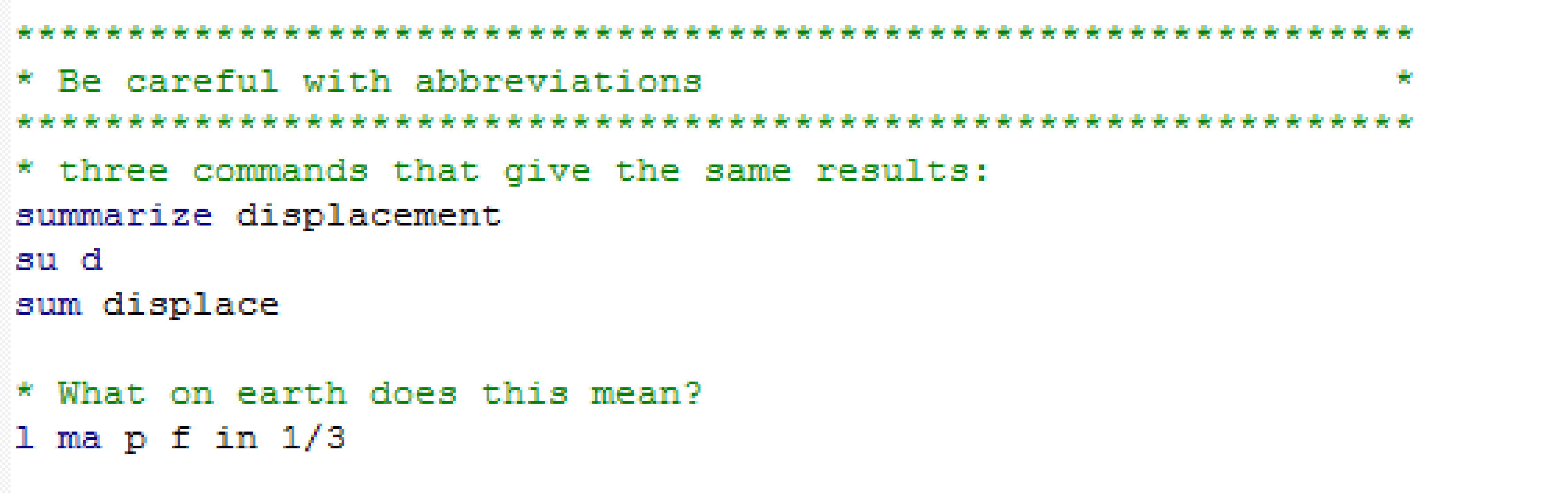
Use alignment and indentation



```
****************************************************************
* 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 concdis_average childmale birth
```

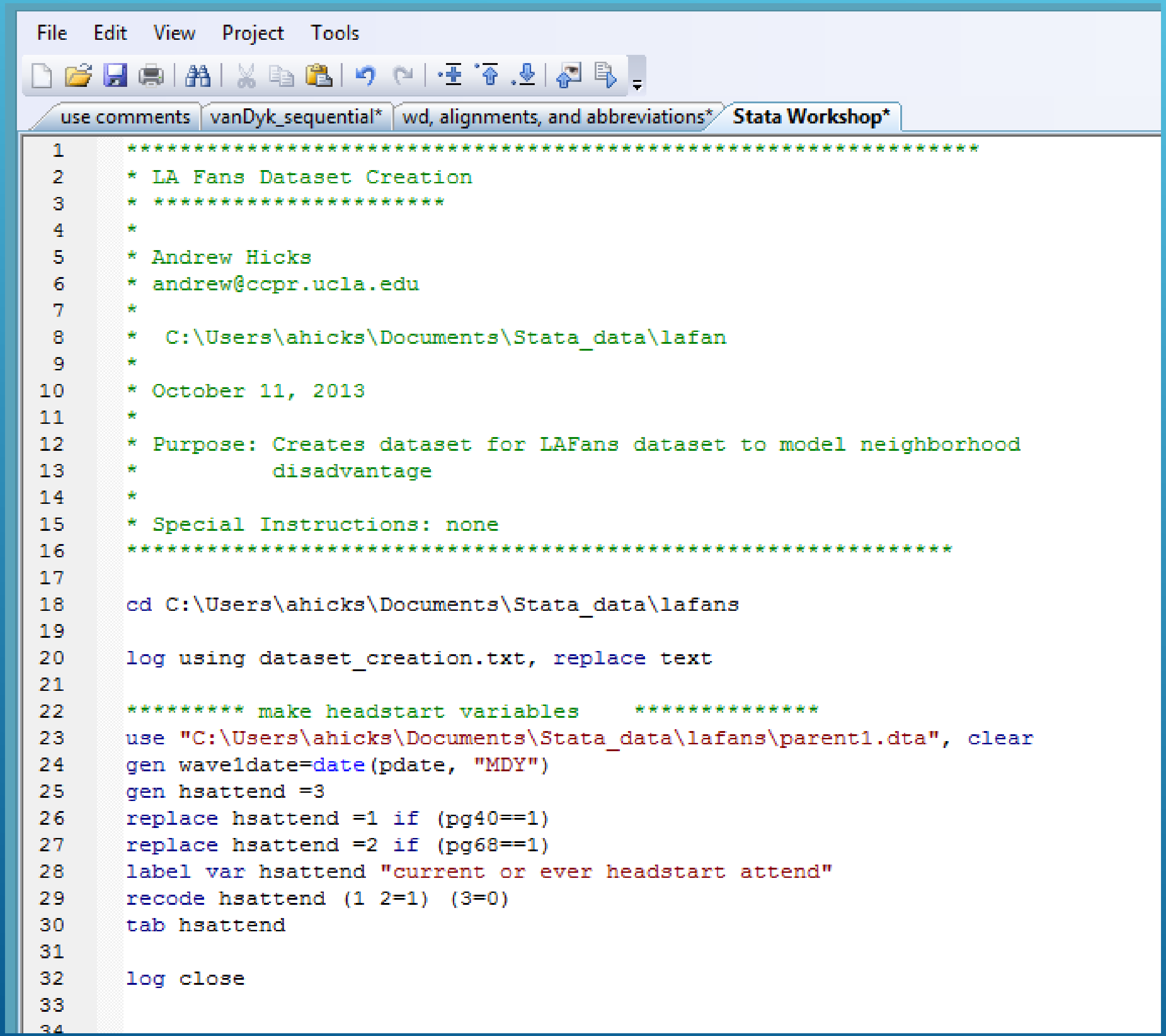
MAKING DO-FILES LEGIBLE

Limit your abbreviations

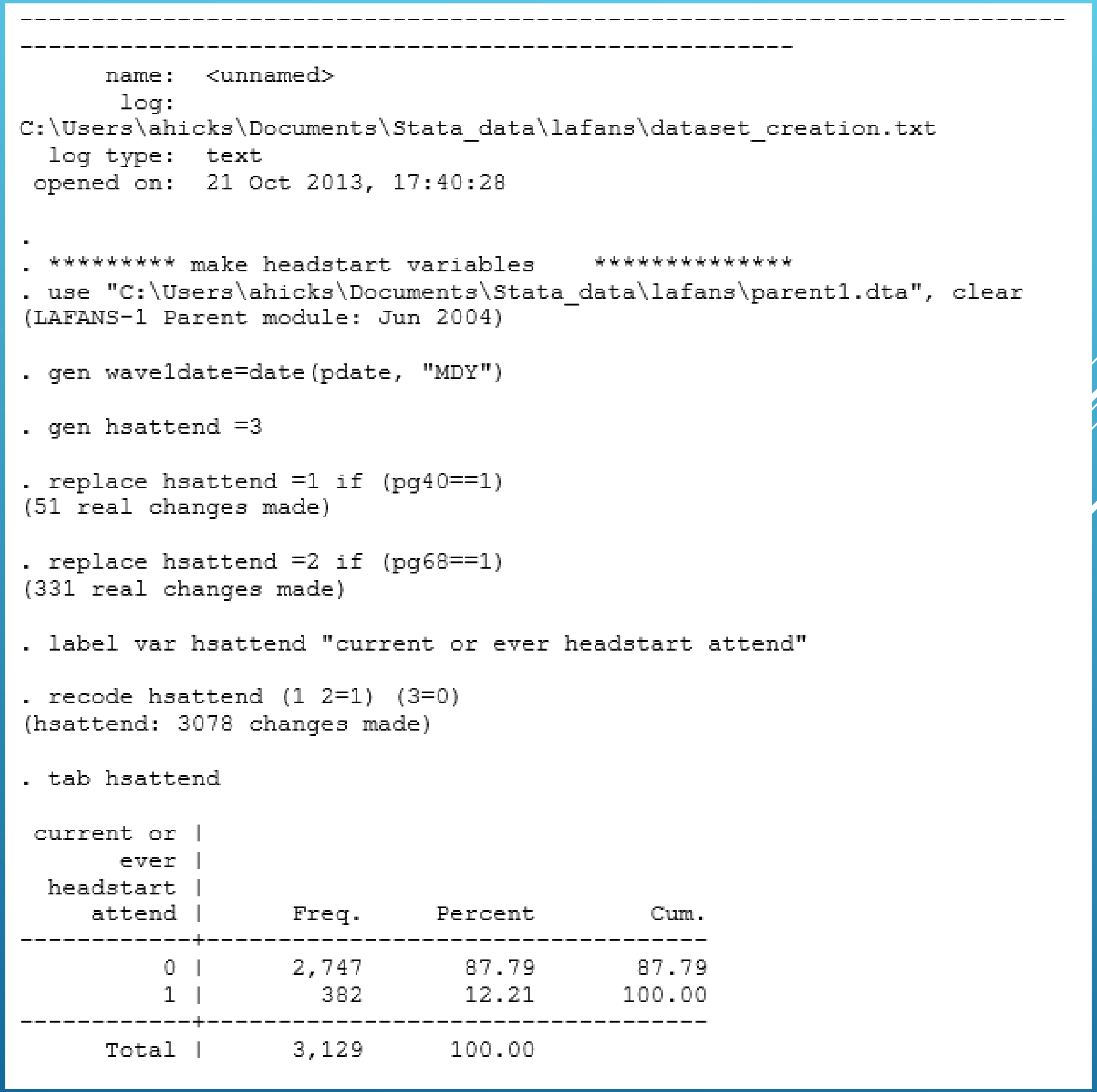


```
****************************************************************
* Be careful with abbreviations                                *
****************************************************************
* three commands that give the same results:
summarize displacement
su d
sum displace

* What on earth does this mean?
l ma p f in 1/3
```

SAVING YOUR SESSION TO A LOG



```
****************************************************************
* LA Fans Dataset Creation
* **********************
*
* Andrew Hicks
* andrew@ccpr.ucla.edu
*
*  C:\Users\ahicks\Documents\Stata_data\lafan
*
* October 11, 2013
*
* Purpose: Creates dataset for LAFans dataset to model neighborhood
*         disadvantage
*
* Special Instructions: none
**************************************************************

cd C:\Users\ahicks\Documents\Stata_data\lafans

log using dataset_creation.txt, replace text

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

log close
```



```
---------------------------------------------------------------------------
-------------------------------------------------------
      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 wave1date=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
```