Stata

Stata is a statistical software with two notable features:

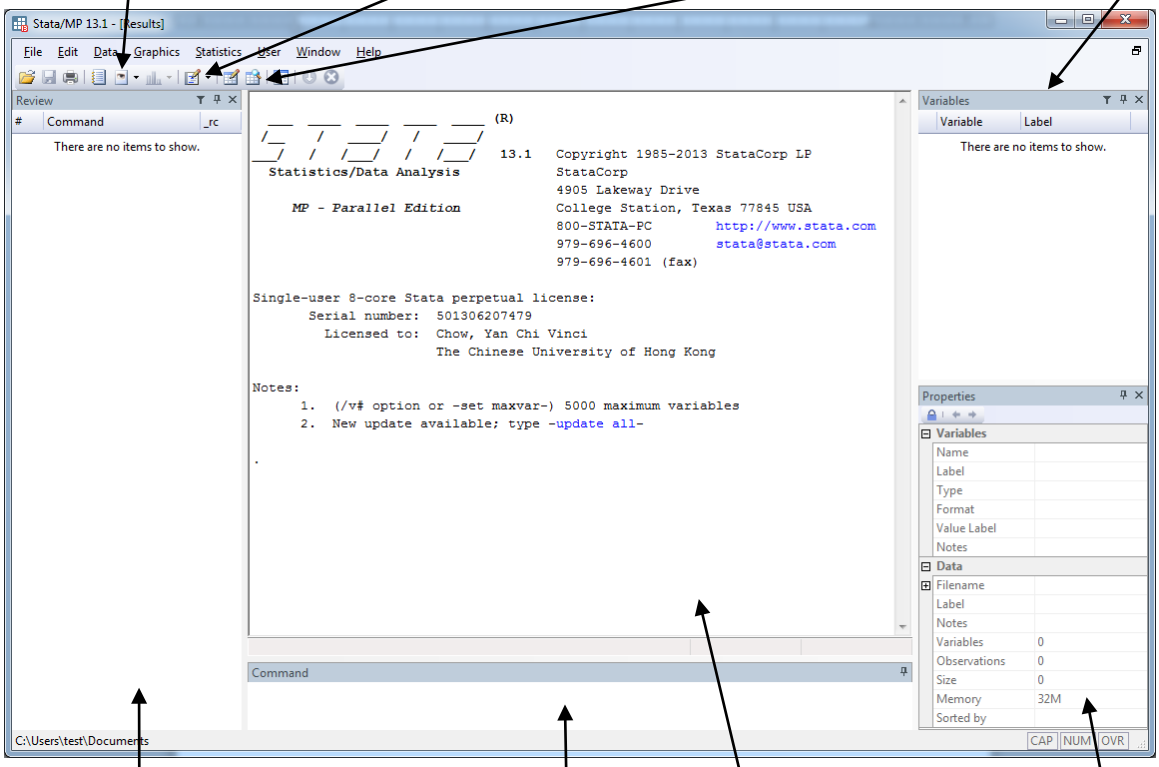
- It can be used interactively through its graphical interface and command input box, or run pre-written scripts. This makes it easier to learn than many other statistical software.
- It only operates on one dataset at a time and all data is loaded into memory. This allows Stata to operate faster than harddisk-based software such as SAS, but you can run into space problem if your dataset is very, very large.

Viewer shows you Stata's help files.

Do-File Editor is a built-in text editor to edit command files.

Data Editor & Browser shows you the data currently in Stata's memory.

Variables lists all your variables.



Review lists out commands that have previously been executed. Successful executions are in black while failed ones are in red.

Double-click on any of them to re-execute. You can also use the [Page-up] and [page-down] keys to move through the list.

Command is where you enter commands.

Results shows you Stata's display output.

Properties lists the properties of the currently-selected variable.

1. Log File: Keeping a record of the commands used and the results generated.

Description	Command	Example
Start a log: record your session into a file called a log file	**log using** *"filename"*, text	log using "D:\Economics\log.txt",text
Close log	**log close**	log close

2. Importing: We can import excel files into Stata's Data Editor and then save them as dta format.

Description	Command	Example
Import an Excel file, also known as a workbook, into Stata's Data Editor	**import excel using** *"filename"*	import excel using "D:\Economics\company_record.xlsx"
Import an Excel file and treat the first row as variable names	import excel using *"filename"*, **firstrow**	import excel using "D:\Economics\company_record.xlsx" , firstrow
Save the workbook into a dta format	**save** "*filename*"	save "D:\Economics\company_record"
Import another excel file into Stata's Data Editor Note: Data editor cannot contain two datasets, so we need to clear the previous one	Import excel using "*filename"*, firstrow **clear**	import excel using "D:\Economics\employee_survey.xlsx", firstrow clear
Save the workbook into a dta format	save *"filename"*	save "D:\Economics\ employee_survey"
If the dta file already exists, overwrite with *replace*	save *"filename", **replace***	save "D:\Economics\employee_survey", replace

3a. Use/ load a Stata dataset (dta format)

Description	Command	Example
First, clear the data in the data editor	**clear**	clear
load a dta file in the data editor	**use** "*filename"*	use "D:\Economics\company_record"
The above two steps can be combined into one command	use "*filename", **clear***	use "D:\Economics\company_record", clear

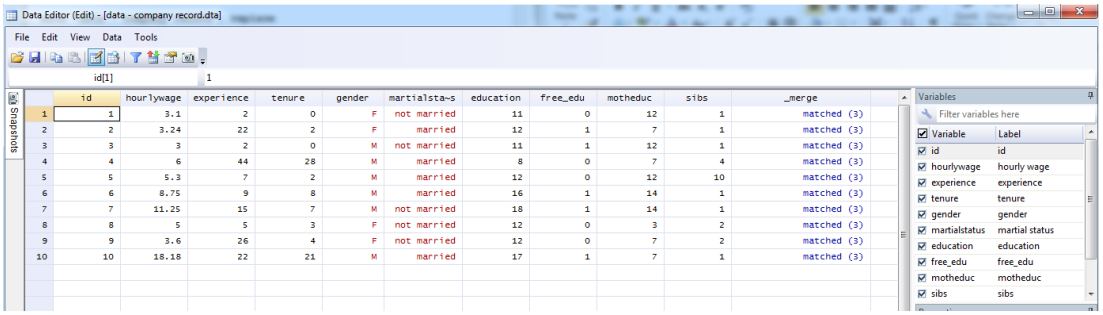
Note: A newer version of Stata can open datasets saved by an older version of Stata, but the reverse is not true.

3b. Change Working Directory

Description	Command	Example
Alternatively, we can first change the working directory before loading a stata dataset, to avoid typing again the full address.	**cd** "*directory*"	cd "D:\Economics\"
Then, load a dta file in the data editor	use "*filename*"	use "company_record"

3c. Merging Datasets

Description	Command	Example
Merge: Merging a dataset to another dataset in the memory of the Data Editor, matching on one or more key variables	**merge** 1:1 *variables* using *"filename"* merge 1:m *variables* using *"filename"* merge m:1 *variables* using *"filename"* merge m:m *variables* using *"filename"*	merge 1:1 id using "employee_survey"
Append: Adding data to bottom of the existing dataset	**append** using *"filename"*	append using "company_record_2"



	id	hourlywage	experience	tenure	gender	martialsta~s	education	free_edu	motheduc	sibs	_merge
1	1	3.1	2	0	F	not married	11	0	12	1	matched (3)
2	2	3.24	22	2	F	married	12	1	7	1	matched (3)
3	3	3	2	0	M	not married	11	1	12	1	matched (3)
4	4	6	44	28	M	married	8	0	7	4	matched (3)
5	5	5.3	7	2	M	married	12	0	12	10	matched (3)
6	6	8.75	9	8	M	married	16	1	14	1	matched (3)
7	7	11.25	15	7	M	not married	18	1	14	1	matched (3)
8	8	5	5	3	F	not married	12	0	3	2	matched (3)
9	9	3.6	26	4	F	not married	12	0	7	2	matched (3)
10	10	18.18	22	21	M	married	17	1	7	1	matched (3)

Fig. 2 Merged dataset

3d. Data frames (Stata 16 onwards)

Description	Command	Example
Create a frame	**frame create** *name*	frame create survey
Change current frame	**frame change** *name*	frame change survey frame change default
Delete frame	**frame drop** *name*	frame drop survey
Do something on a frame	**frame** *name*: … **frame** *name* { … }	frame survey: use employee_survey
Link with another frame	**frlink** m:n *variables*, frame(*name*)	frlink 1:1 id, frame(survey)
Fetch data from another frame	**frget** *varname*, from(*name*)	frget education, from(survey)

4. Manipulate Data

Description	Command	Example
Adding a new variable	**generate** *new_var* = …	gen log_edu = log(education)
Modifying a variable	**replace** *variable* = …	replace log_edu = ln(education)
Drop a variable	**drop** *variable*	drop log_edu
Drop an observation	**drop if** *variable* = …	drop if id == 11
Switch between the two common ways of storing groups of data	**reshape** …	(Read Stata's help file if you need this function)

Note: Type "function" in the viewer for a list of available functions.
Stata follows the common programming convention of using "=" for assignment(i.e. modification of data) and "==" for comparison.

5. Summarize: to obtain summary statistics

Description	Command	Example
Summarize	**sum**	sum
Summarize a variable	sum *variable*	sum hourlywage
Summarise a variable in detail	sum *variable*, **detail**	sum hourlywage, detail

6. Making a table

Description	Command	Example
Making a table of summary statistics: Make a table with certain contents	**table** *variable1 variable2*, contents(option)	table gender free_edu, contents(median hourlywage)
		table gender free_edu, contents(median hourlywage sd hourlywage)

gender	free_edu 0	1
F	3.6	3.24
	3	1
M	5.65	10
	2	4

An example of *Table* command output

7. Correlation

Description	Command	Example
Correlations (covariances) of variables	**correlate** *variable1 variable2 variable3…*	corr hourlywage experience education

```
             | hourly~e experi~e educat~n
-------------+---------------------------
  hourlywage |   1.0000
  experience |   0.1940   1.0000
   education |   0.7592  -0.2218   1.0000
```

An example of *Correlation* command output

8. T-Test

Description	Command	Example
T-test: compare the means of two variables	**ttest** *variable1* = *variable2*	ttest education = motheduc

```
. ttest education = motheduc

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
educat~n |      10        12.9    .9826269    3.107339    10.67714    15.12286
motheduc |      10         9.5    1.185561    3.749074    6.818074    12.18193
---------+--------------------------------------------------------------------
    diff |      10         3.4    1.240072    3.921451     .594763    6.205237
------------------------------------------------------------------------------
     mean(diff) = mean(education - motheduc)                      t =   2.7418
 Ho: mean(diff) = 0                               degrees of freedom =        9

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9886         Pr(|T| > |t|) = 0.0228          Pr(T > t) = 0.0114
```

An example of *ttest* command output for test of two variables

Description	Command	Example
T-test: compare the means of two groups within the same variable	ttest *variable1*, **by(*groupvar*)**	ttest hourlywage, by(gender)
Note: *groupvar* can only take on two values		

```
. ttest hourlywage, by(gender)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       F |       4       3.735    .4346167    .8692334    2.351856    5.118144
       M |       6    8.746667    2.218876    5.435114    3.042864    14.45047
---------+--------------------------------------------------------------------
combined |      10       6.742    1.528432    4.833326    3.284447    10.19955
---------+--------------------------------------------------------------------
    diff |           -5.011667    2.794796               -11.45648    1.433145
------------------------------------------------------------------------------
    diff = mean(F) - mean(M)                                      t =  -1.7932
Ho: diff = 0                                     degrees of freedom =        8

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0553         Pr(|T| > |t|) = 0.1107          Pr(T > t) = 0.9447
```

An example of *ttest* command output for test of two groups within the same variable

9a. Histogram

Description	Command	Example
Histogram of a variable	**hist** *variable*	hist education
Histogram of a variable, with n blocks (Fig 3)	hist *variable*, **bin(*n*)**	hist education, bin(5)
Histogram of a variable, with n blocks, and y axis as fraction (Fig 4)	hist *variable*, bin(*n*) **fraction**	hist education, bin(5) fraction

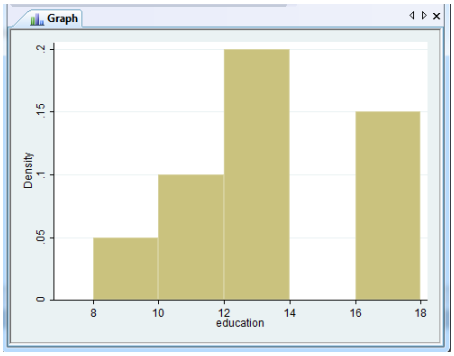


Fig 3

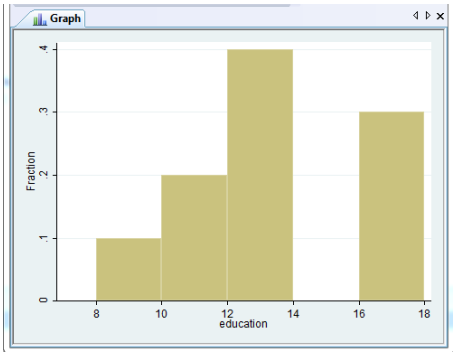


Fig 4

9b. Scatter Graph

Description	Command	Example
Plot a scatter graph	**scatter** *variable1 variable2*	scatter hourlywage education
Plot two scatter subgraphes , being placed beside each other (Fig 5)	scatter *variable1 variable2*, **by(*variable3*)**	scatter hourlywage educ, by(gender)
Plot two subgraphes, one placing on another (Fig 6)	scatter *variable1 variable2* if *variable3* == value1 **||** scatter *variable1 variable2* if *variable3* == value2	scatter hourlywage educ if gender == "M" || scatter hourlywage educ if gender == "F"

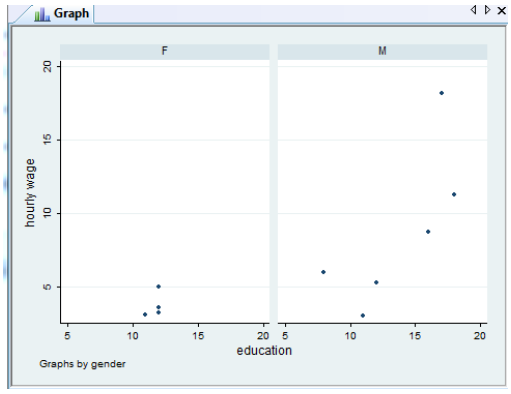


Fig 5

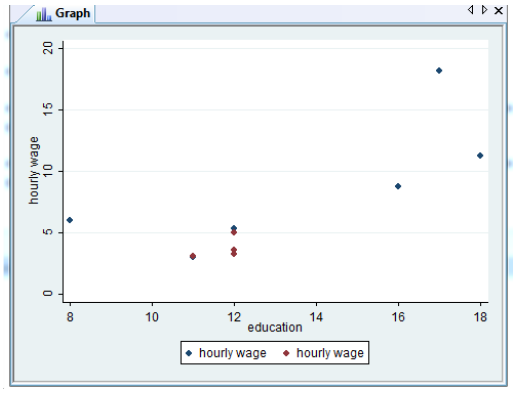


Fig 6

10a. Regression

Description	Command	Example
Ordinary Least Square	**regress** *dep_variable indep_variables*	reg hourlywage experience tenure

```
. reg hourlywage experience tenure

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  2,     7) =    3.75
       Model |  108.737034     2  54.3685168            Prob > F      =  0.0782
    Residual |  101.512326     7  14.5017609            R-squared     =  0.5172
-------------+------------------------------           Adj R-squared =  0.3792
       Total |   210.24936     9    23.36104           Root MSE      =  3.8081

------------------------------------------------------------------------------
  hourlywage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  experience |  -.2491599   .1541775    -1.62   0.150    -.6137318     .115412
      tenure |   .5712551   .2166466     2.64   0.034     .0589673    1.083543
       _cons |   6.294649   1.913857     3.29   0.013     1.769097     10.8202
------------------------------------------------------------------------------
```

An example of *regress* command

10b. Regression with dummy variables:

First, we have to generate dummy variables for qualitative variables	Step1: **generate** *the name of dummy variable* = 0 (Fig 7)	generate gender_dummy = 0
	Step2: **replace** *the name of dummy variable* = 1 if variable == “value1” replace *the name of dummy variable* = 2 if variable ==”value2” (Fig 8)	replace gender_dummy = 1 if gender=="M"
	Alternatively, use **xi**	xi i.gender
Then, we do regression with the dummy variables	reg *variable1 variable2 the name of dummy variable*	reg hourlywage experience tenure gender_dummy

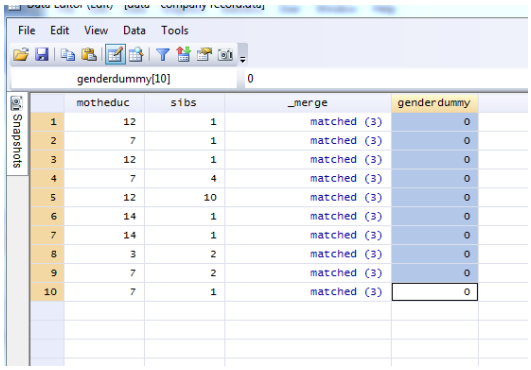


Fig 7

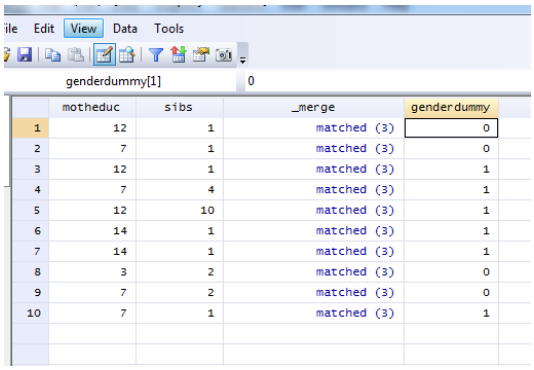
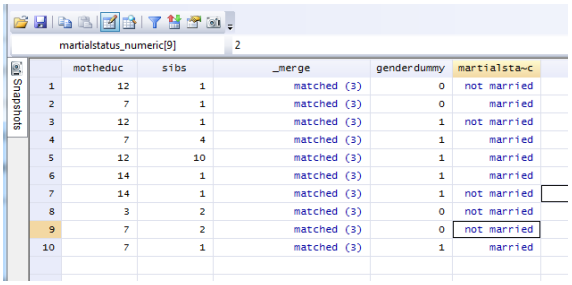


Fig 8

11. Fixed-Effect Regression

Description	Command	Example
If there are too many values for the dummy variable, we can encode the variable into numeric first	**encode** *variable*, generate(*new numeric dummy variable*)	encode martialstatus, generate(martialstatus_numeric)
Then, run fixed effect regression	**xtreg** *dep_variable indep_variables*, fe i(*new numeric dummy variable*)	xtreg hourlywage tenure, fe i(martialstatus_numeric)



	motheduc	sibs	_merge	genderdummy	martialsta~c
1	12	1	matched (3)	0	not married
2	7	1	matched (3)	0	married
3	12	1	matched (3)	1	not married
4	7	4	matched (3)	1	married
5	12	10	matched (3)	1	married
6	14	1	matched (3)	1	married
7	14	1	matched (3)	1	not married
8	3	2	matched (3)	0	not married
9	7	2	matched (3)	0	not married
10	7	1	matched (3)	1	married

Output of *encode*. The leftmost variable is in fact numeric, but is labeled.

```
. xtreg hourlywage tenure, fe i(martialstatus_numeric)

Fixed-effects (within) regression               Number of obs      =        10
Group variable: martialsta~c                    Number of groups   =         2

R-sq:  within  = 0.2532                         Obs per group: min =         5
       between = 1.0000                                        avg =       5.0
       overall = 0.3370                                        max =         5

                                                F(1,7)             =      2.37
corr(u_i, Xb)  = 0.5227                         Prob > F           =    0.1673

------------------------------------------------------------------------------
  hourlywage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   .2832131   .1838511     1.54   0.167    -.1515256    .7179517
       _cons |   4.617902   1.971669     2.34   0.052    -.0443534    9.280157
-------------+----------------------------------------------------------------
     sigma_u |  .31239775
     sigma_e |  4.4566344
         rho |   .0048896   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(1, 7) =     0.02                 Prob > F = 0.8975
```

An example of x*treg* command

12a. Correction for Heteroskedasticity

Description	Command	Example
Test if the homoscedasticity assumption holds (Run after *regress*)	**estat** hettest *indep_variables*	hettest experience tenure
Robust Standard Errors (Eicker-White Std. Err.)	regress *dep_variable indep_variables*, **robust** (Also works with xtreg)	reg hourlywage experience tenure, robust

```
. reg hourlywage experience tenure, robust

Linear regression                               Number of obs =        10
                                                F(  2,     7) =      3.25
                                                Prob > F      =    0.1002
                                                R-squared     =    0.5172
                                                Root MSE      =    3.8081

------------------------------------------------------------------------------
             |               Robust
  hourlywage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  experience |  -.2491599   .1273134    -1.96   0.091    -.5502081    .0518884
      tenure |   .5712551   .2351817     2.43   0.045     .0151388    1.127371
       _cons |   6.294649   1.705796     3.69   0.008     2.261082    10.32822
------------------------------------------------------------------------------
```

An example of robust standard errors. Note the difference in standard errors compared to 10a.

12b. Correction for Error Correlation within Group and Over Time

Clustered Standard Errors Corrects within-group error correlation	regress *dep_variable indep_variables*, **vce(cluster *clustervar*)** (Also works with xtreg)	reg hourlywage tenure, vce(cluster workplace)
Newey-West Standard Errors Corrects for equi-correlated error over time. Error beyond the number of *periods* specified are assumed to be uncorrelated	**newey** *dep_variable indep_variables*, lag(*periods*) Let Stata select optimal lag: **Ivregress** gmm *dep_var indep_vars*, wmatrix(hac nw opt)	newey hourlywage tenure, lag(2) ivregress gmm hourlywage tenure, wmat(hac nw opt)

13. Hypothesis Testing

Test linear hypothesis	**test** *varnames* test *exp1* [= *exp2* = …]	test tenure experience test tenure – experience = 0
Test non-linear hypothesis	**testnl** *exp2* [= *exp2* = …]	testnl _b[tenure]^2 = 0

14. Obtaining residuals and predicted values

Obtain predicted values after regression	**predict** *new_var*	predict predicted_hourlywage
Obtain residuals	predict *new_var*, **residuals**	predict estimated_u, r

15. Instrumental Variable Regression

Description	Command	Example
When an independent variable is correlated with the error term, OLS is biased. IV regression uses another variable uncorrelated with the error to predict the correlated one	**ivregress** estimator *dep_var exog_vars (endo_var = instrument_vars)*	ivregress 2sls hourlywage (education = free_edu)
Test for endogeneity after IV regression	**estat endogenous**	Estat endog

16a. Discrete Choice Model

Description	Command	Example
Logit: When the dependent variable takes on binary values, we can use the logit model	**logit** *dep_var indep_vars*	logit free_edu mothedu
However, the interpretation of β Estimator is different from the one we used for OLS. So we need to use odd ratios	logit *dep_var indep_vars*, **or**	logit free_edu mothedu, or

16b. Additional Discrete Choice Models

Description	Command	Example
Multinomial Logit: When the dependent variable takes on more than two discrete values	**mlogit** *dep_var indep_vars*	mlogit free_edu mothedu
Ordered Logit: When the dependent variable represents ordinal ratings (e.g. bad, good, best)	**ologit** *dep_var indep_vars*	ologit feedback budget, or
Rank-ordered Logit: When the dependent variable represents successive draws without replacement (e.g. places in a race)	**rologit** *dep_var indep_vars*, group(horse_id)	rologit position training, group(horse_id)

17. Obtaining Marginal Effects

Description	Command	Example
The marginal effect of each independent variable on the predicted value at the average value of the variable	old syntax: **mfx** new syntax: **margins**, dydx(*indep_vars*) atmeans	mfx margins, dydx(motheduc) atmeans