Stata

Popular program in Economics, Medicine, ???

The previous versions were command line based. The current version (8) has switched to more of a graphical menu driven interface.

However, everything can be done from a command line. Somethings have to be done from the command line.

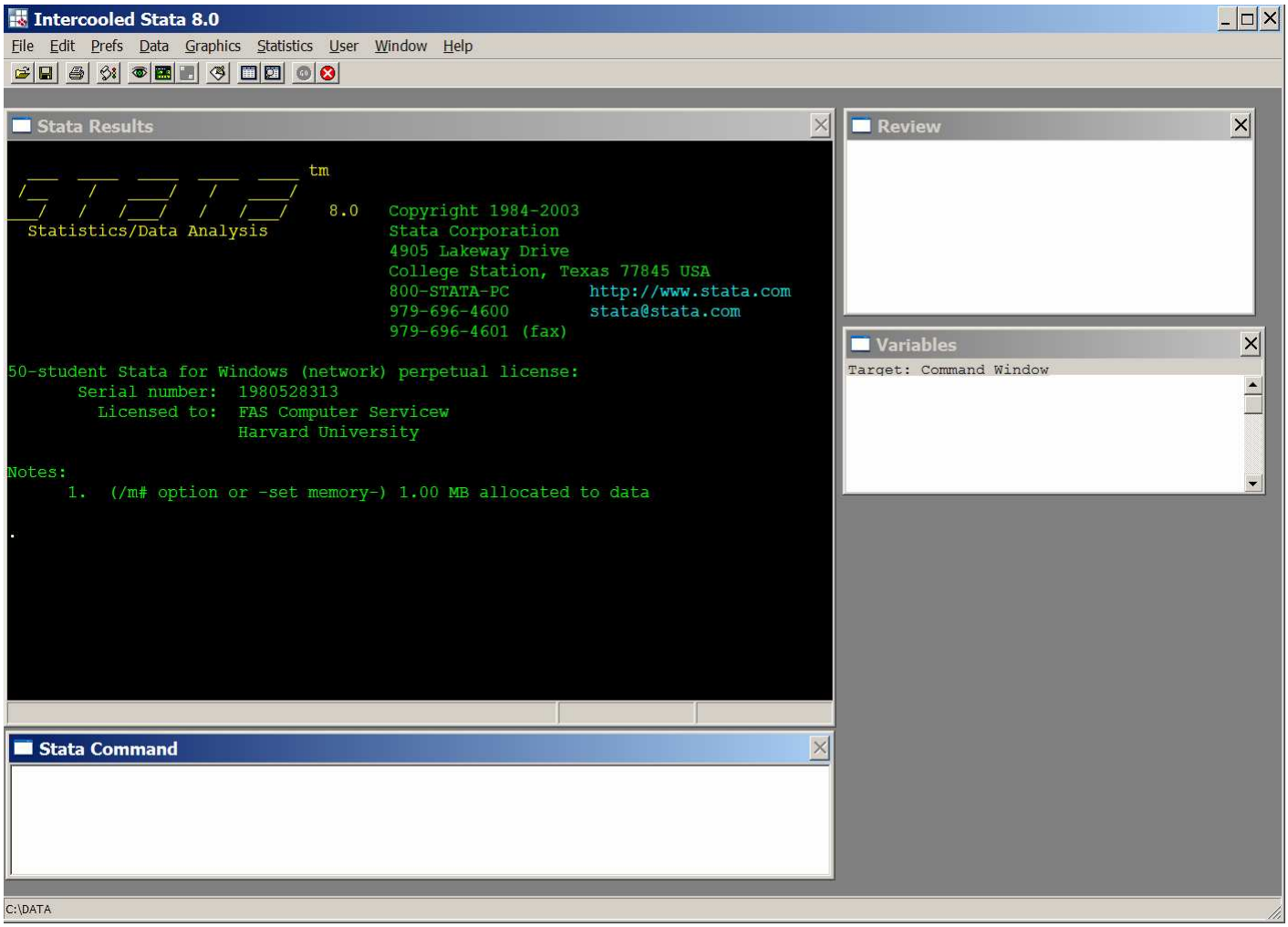
Has very good Survey Sampling routines

More of a canned program, though you can write your own routines, and even link in your own C code (with version 8.1).

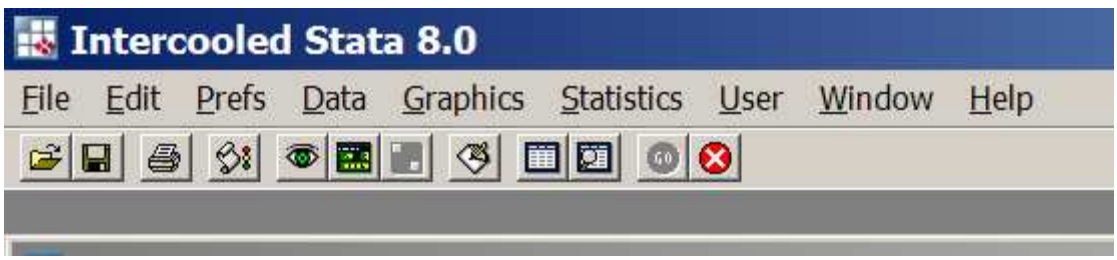
While you can do some programming with it, and a lot of people apparently do, I would not use Stata unless I was performing a "standard" analysis. Use S-Plus/R in those situations.

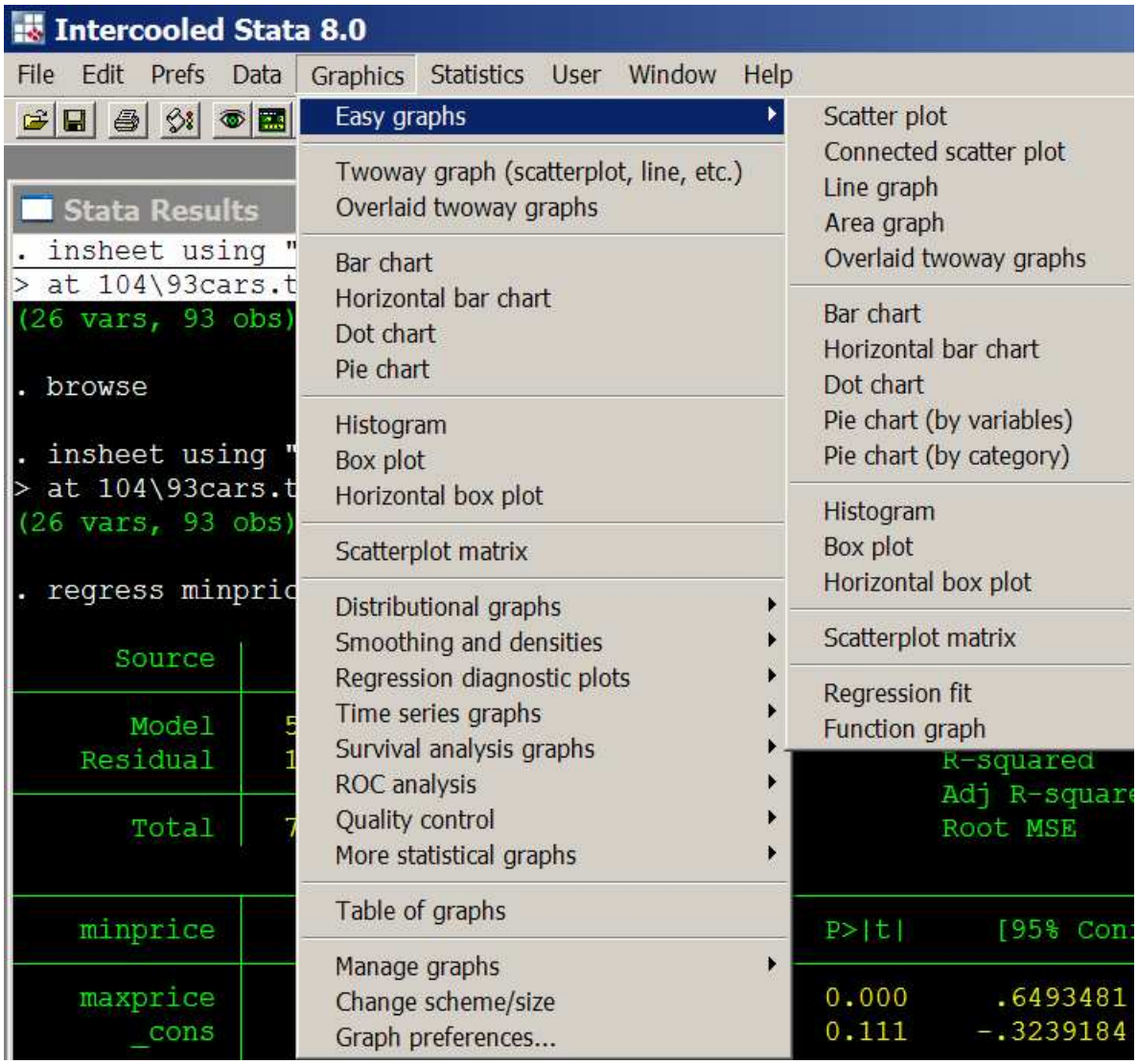
Has excellent backward compatibility with previous versions

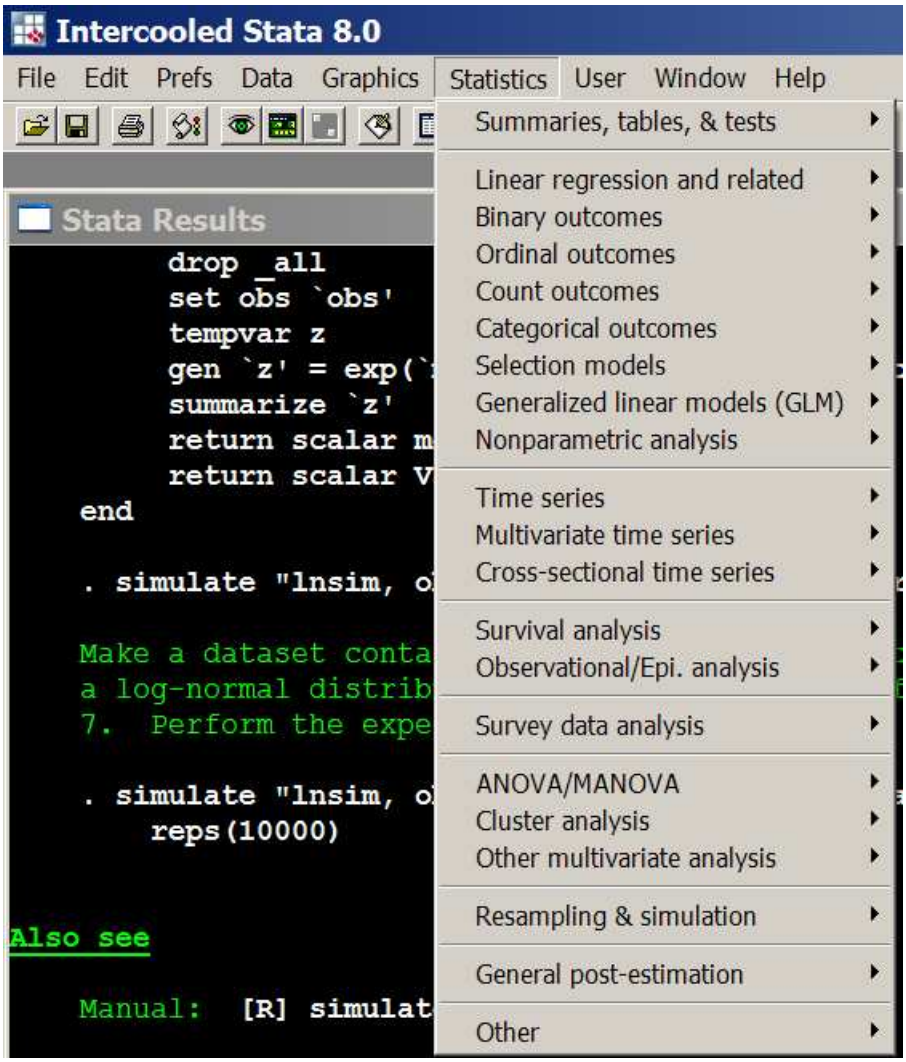
Available for Windows, Macintosh, Unix (not available at Harvard that I can find).



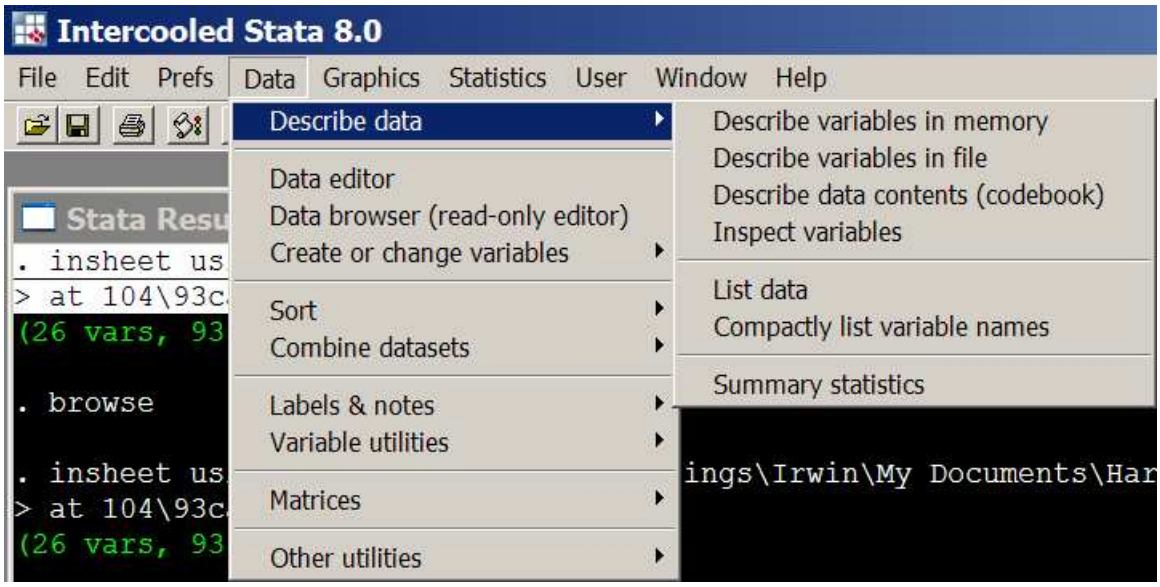
The default setup will look similar to the above. However you can change the setup, font sizes, colours, etc. These will be kept from session to session.




Intercooled Stata 8.0
File Edit Prefs Data Graphics Statistics User Window Help
Stata Results
. insheet using "
> at 104\93cars.t
(26 vars, 93 obs)
. browse
. insheet using "
> at 104\93cars.t
(26 vars, 93 obs)
. regress minpric
Source
Model
Residual
Total
minprice
maxprice
_cons
R-squared
Adj R-squar
Root MSE
P>|t|
[95% Con
0.000 .6493481
0.111 -.3239184
Easy graphs
Twoway graph (scatterplot, line, etc.)
Overlaid twoway graphs
Bar chart
Horizontal bar chart
Dot chart
Pie chart
Histogram
Box plot
Horizontal box plot
Scatterplot matrix
Distributional graphs
Smoothing and densities
Regression diagnostic plots
Time series graphs
Survival analysis graphs
ROC analysis
Quality control
More statistical graphs
Table of graphs
Manage graphs
Change scheme/size
Graph preferences...
Scatter plot
Connected scatter plot
Line graph
Area graph
Overlaid twoway graphs
Bar chart
Horizontal bar chart
Dot chart
Pie chart (by variables)
Pie chart (by category)
Histogram
Box plot
Horizontal box plot
Scatterplot matrix
Regression fit
Function graph


Intercooled Stata 8.0
File Edit Prefs Data Graphics Statistics User Window Help
Stata Results
drop _all
set obs `obs'
tempvar z
gen `z' = exp(`
summarize `z'
return scalar m
return scalar V
end
. simulate "lnsim, o
Make a dataset conta
a log-normal distrib
7. Perform the expe
. simulate "lnsim, o
reps(10000)
Also see
Manual: [R] simulat
Summaries, tables, & tests
Linear regression and related
Binary outcomes
Ordinal outcomes
Count outcomes
Categorical outcomes
Selection models
Generalized linear models (GLM)
Nonparametric analysis
Time series
Multivariate time series
Cross-sectional time series
Survival analysis
Observational/Epi. analysis
Survey data analysis
ANOVA/MANOVA
Cluster analysis
Other multivariate analysis
Resampling & simulation
General post-estimation
Other

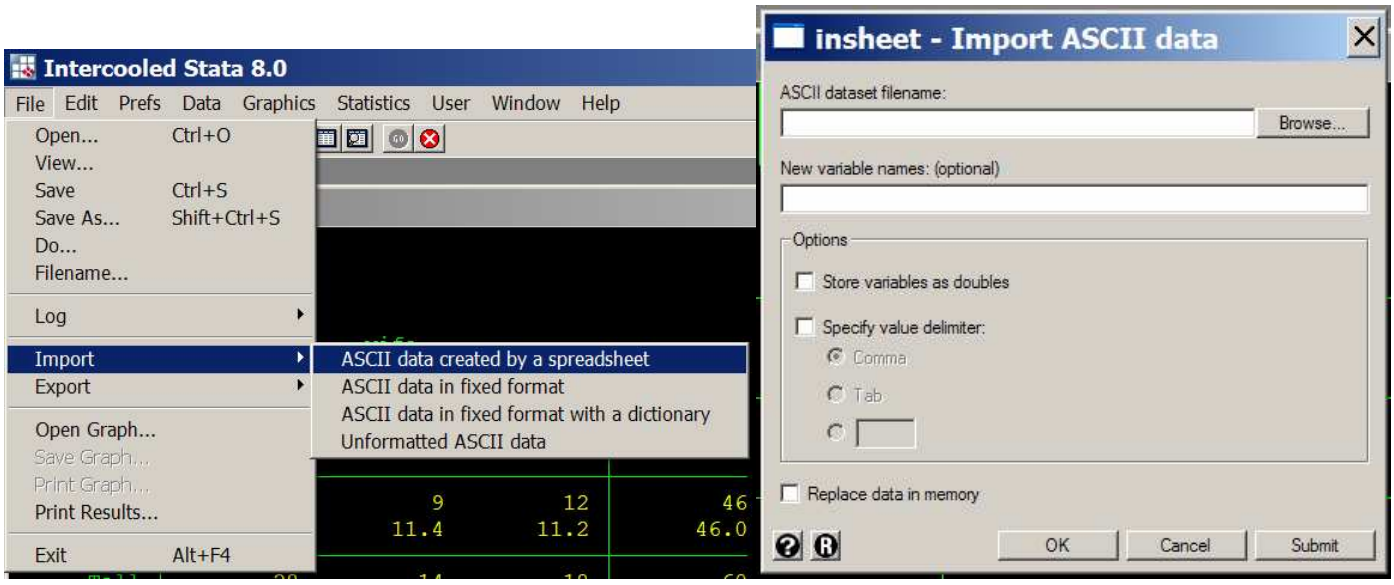
Data Input:



Stata can only deal with two file formats, its own, and text files for data.

Its own file format has changed over time, but the current versions can read older versions. However the other way doesn't work (e.g. Version 7 can read a version 8 file)

Reading in text files




```
insheet using "C:\Documents and Settings\Irwin\My
Documents\Harvard\Courses\Stat 104\93cars.txt", clear
```

When creating commands with Stata, the command you would need to type is given in the Results window and the Review window.

The Review window can be used to either repeat commands (double click on the command) or as a start to create a new command (single click).

A single click will put the command into the Commands window and it can be edited

Also the previous version of a command is kept in the dialogue boxes so you can also edit things that way as well.

The  button in a command dialogue box will clear the entries.

The underlying data structure is similar to a Spreadsheet. Rows correspond to observations and colums correspond to variables. You can also think of it like a S-Plus/R data frame.

In Stata you can only have one data file in use at a time. Its not like S-Plus/R where you can have many data frames available.

It is possible to combine datasets to get around this limitation.

All you can really read in is a table with *r* rows and *c* columns.

However the columns can be a mixture of numbers and strings. Just about any file that `read.table()` will work with, the Stata `insheet` command can handle.

```
. insheet using "filename.raw", clear
```

The default file extension for text files being read in is `.raw`, though any can be used.

The `clear` option is there to force the current data file to be deleted from memory and to be replace by the new file. If not, you will get an error message similar to

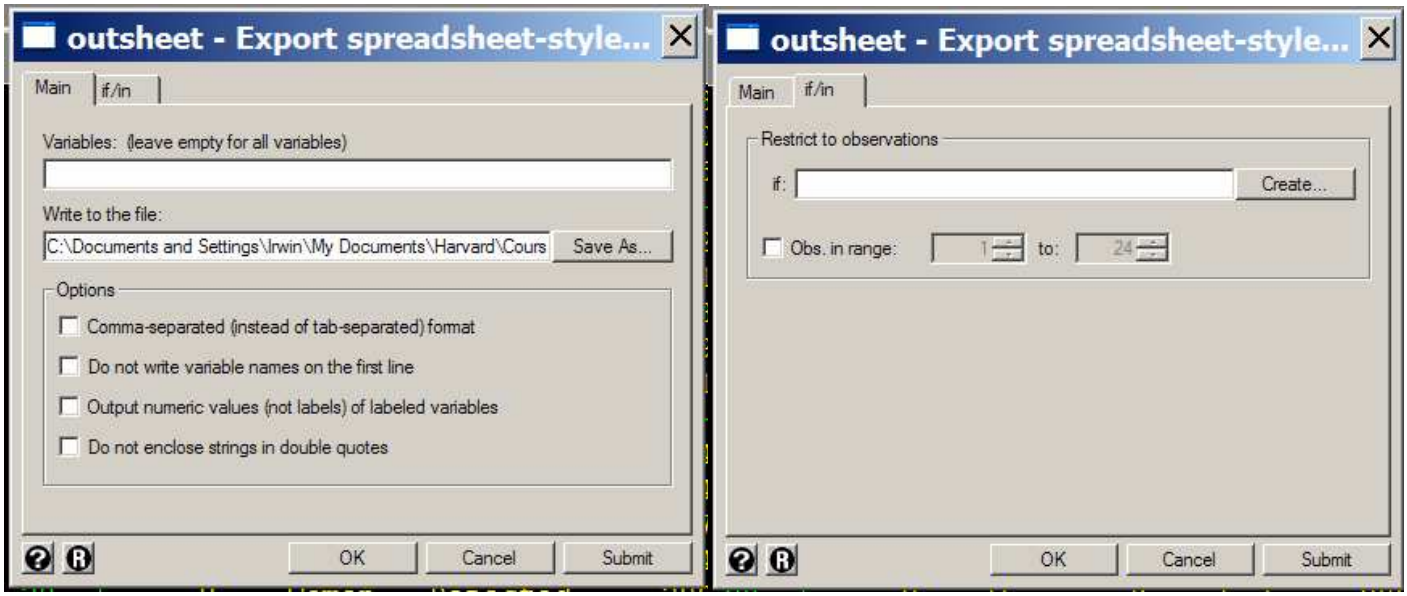
```
insheet using "Berkeley.txt"

you must start with an empty dataset

r(18);
```

While the expected column separators are either commas or tabs, any character can be used.

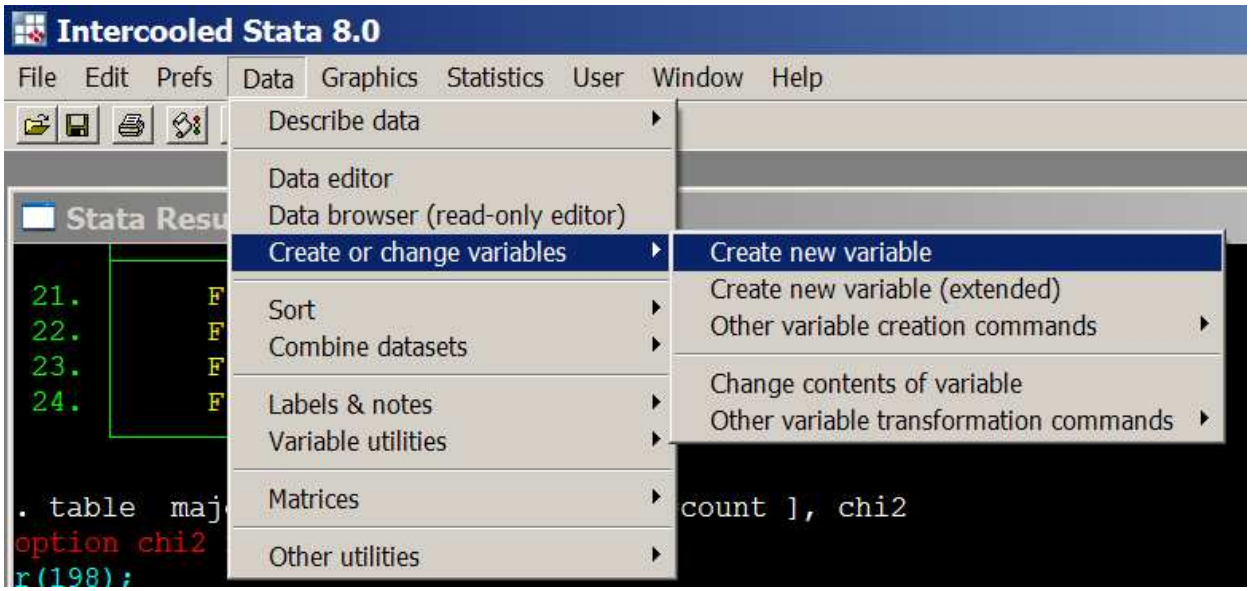
Exporting Data

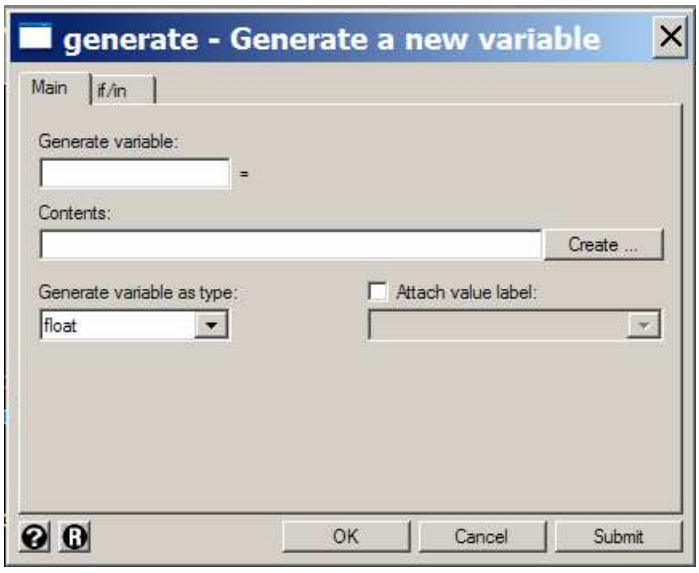


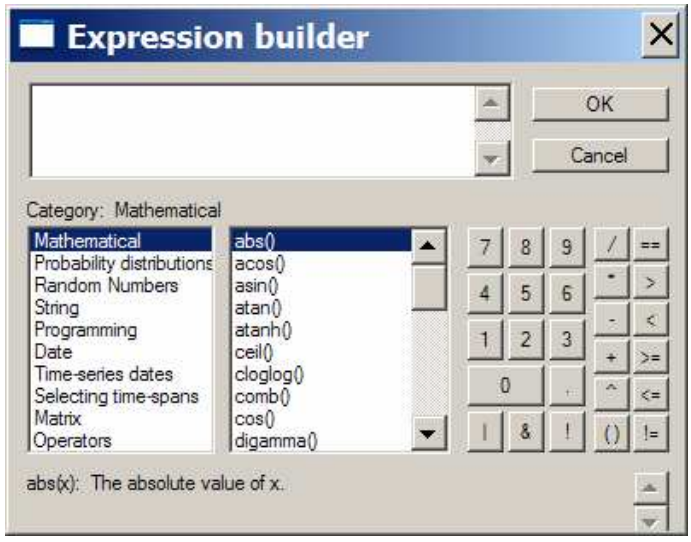

```
outsheet using " Berkeley.raw", replace

outsheet using " Berkeley-men.raw" if
gender == "Men", replace
```

Defining variables







Going through the dialogues will generate a command like

```
generate float CityFuel= 100/HighMPG

generate float CityFuel= 100/CityMPG

HighFuel already defined
```

`generate` can only be used for creating a variable for the first time. If you need to update a variable use `replace`

```
replace CityFuel= 100/CityMPG
(93 real changes made)
```

Stata has the property when creating a new variable to place results in every row, even when it may not make intuitive sense. For example

```
generate tcrit= invttail(24,0.025)
```

will place the value 2.063899 in every row of the data file. To store it only in the first row, use

```
gen tcrit= invttail(24,0.025) in 1
```

instead. Actually for this sort of calculation, you probably just want

```
disp invttail(24,0.025)
2.0638985
```

which just prints the result in the Results window.

Viewing Data

There are a number of ways that you can view your data in Stata.

The first is the list command

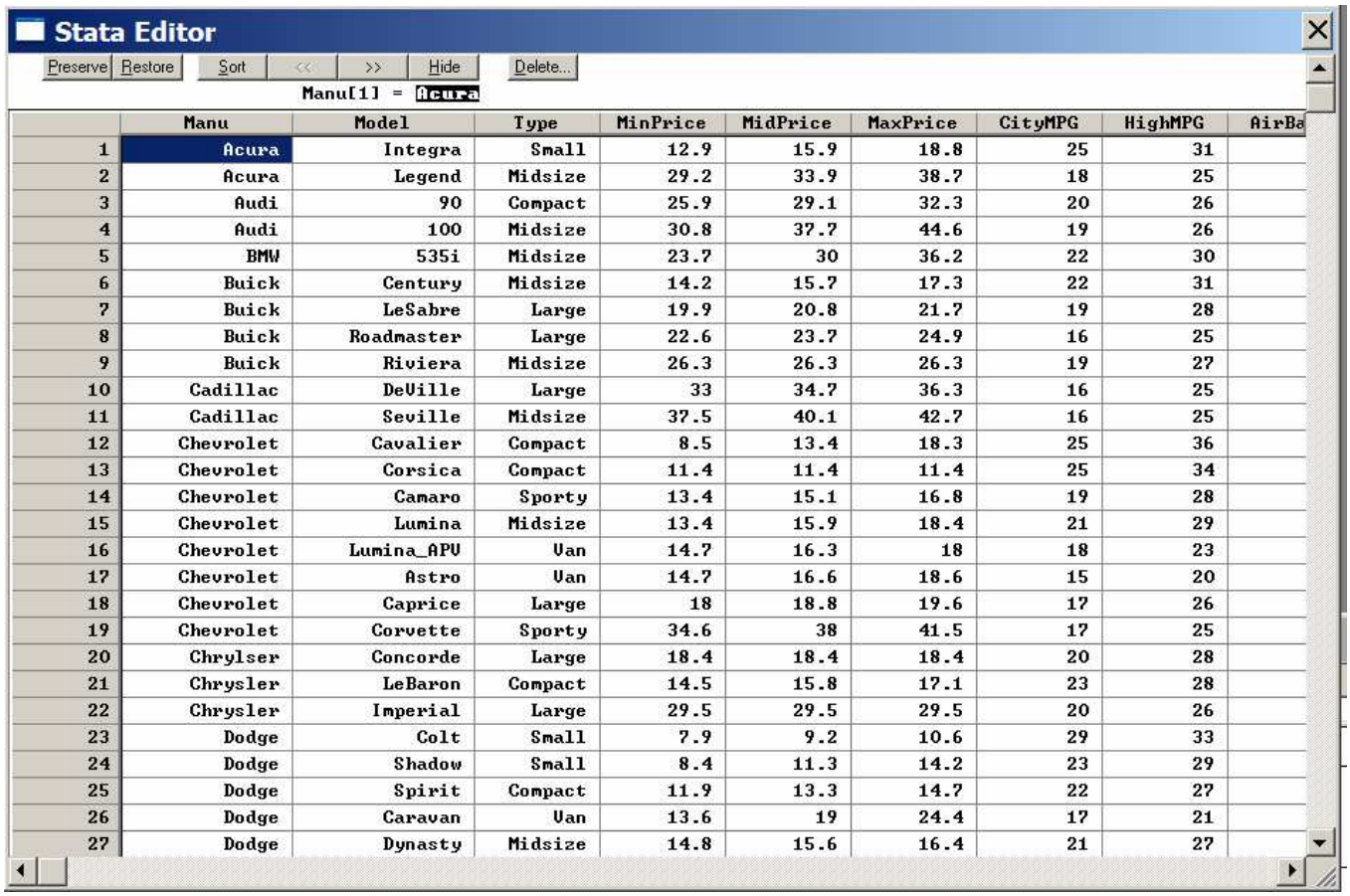
```
list MinPrice MidPrice MaxPrice in 1/10,
separator(5)
```

```
     +-----------------------------------+
     | MinPrice   MidPrice   MaxPrice |
     |-----------------------------------|
  1. |     12.9        15.9        18.8 |
  2. |     29.2        33.9        38.7 |
  3. |     25.9        29.1        32.3 |
  4. |     30.8        37.7        44.6 |
  5. |     23.7          30        36.2 |
     |-----------------------------------|
  6. |     14.2        15.7        17.3 |
  7. |     19.9        20.8        21.7 |
  8. |     22.6        23.7        24.9 |
  9. |     26.3        26.3        26.3 |
 10. |       33        34.7        36.3 |
     +-----------------------------------+
```

The other common approaches are with the Data Editor (`edit`) and the Data Browser (`browse`)

They both appear the same, but only with the Editor can you make changes.

Also when either of these is active, it is not possible to run any analyses. You must close the window to proceed.



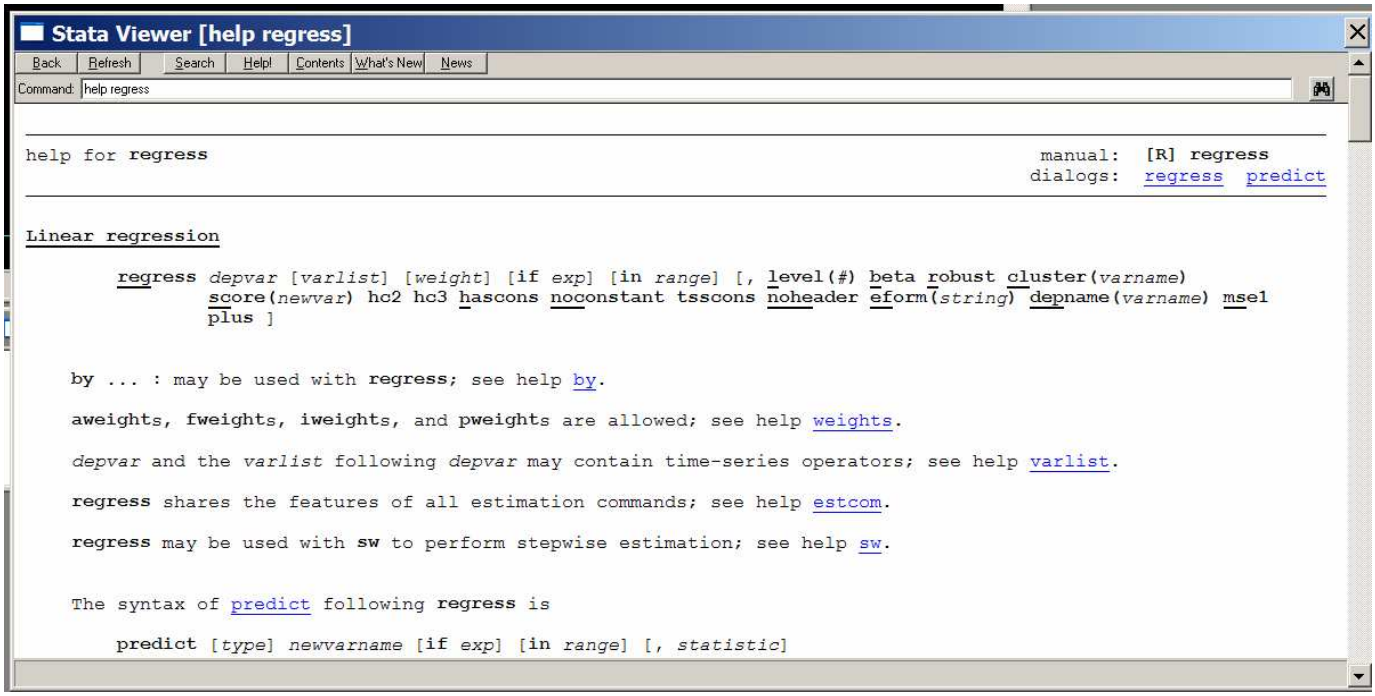
	Manu	Model	Type	MinPrice	MidPrice	MaxPrice	CityMPG	HighMPG	AirBa
1	Acura	Integra	Small	12.9	15.9	18.8	25	31	
2	Acura	Legend	Midsize	29.2	33.9	38.7	18	25	
3	Audi	90	Compact	25.9	29.1	32.3	20	26	
4	Audi	100	Midsize	30.8	37.7	44.6	19	26	
5	BMW	535i	Midsize	23.7	30	36.2	22	30	
6	Buick	Century	Midsize	14.2	15.7	17.3	22	31	
7	Buick	LeSabre	Large	19.9	20.8	21.7	19	28	
8	Buick	Roadmaster	Large	22.6	23.7	24.9	16	25	
9	Buick	Riviera	Midsize	26.3	26.3	26.3	19	27	
10	Cadillac	DeVille	Large	33	34.7	36.3	16	25	
11	Cadillac	Seville	Midsize	37.5	40.1	42.7	16	25	
12	Chevrolet	Cavalier	Compact	8.5	13.4	18.3	25	36	
13	Chevrolet	Corsica	Compact	11.4	11.4	11.4	25	34	
14	Chevrolet	Camaro	Sporty	13.4	15.1	16.8	19	28	
15	Chevrolet	Lumina	Midsize	13.4	15.9	18.4	21	29	
16	Chevrolet	Lumina_APV	Van	14.7	16.3	18	18	23	
17	Chevrolet	Astro	Van	14.7	16.6	18.6	15	20	
18	Chevrolet	Caprice	Large	18	18.8	19.6	17	26	
19	Chevrolet	Corvette	Sporty	34.6	38	41.5	17	25	
20	Chrylser	Concorde	Large	18.4	18.4	18.4	20	28	
21	Chrysler	LeBaron	Compact	14.5	15.8	17.1	23	28	
22	Chrysler	Imperial	Large	29.5	29.5	29.5	20	26	
23	Dodge	Colt	Small	7.9	9.2	10.6	29	33	
24	Dodge	Shadow	Small	8.4	11.3	14.2	23	29	
25	Dodge	Spirit	Compact	11.9	13.3	14.7	22	27	
26	Dodge	Caravan	Van	13.6	19	24.4	17	21	
27	Dodge	Dynasty	Midsize	14.8	15.6	16.4	21	27	

With the data editor, you cannot give variable names within the editor. However you can change them later with the `rename` command, e.g.

```
rename Manu Manufacturer
```

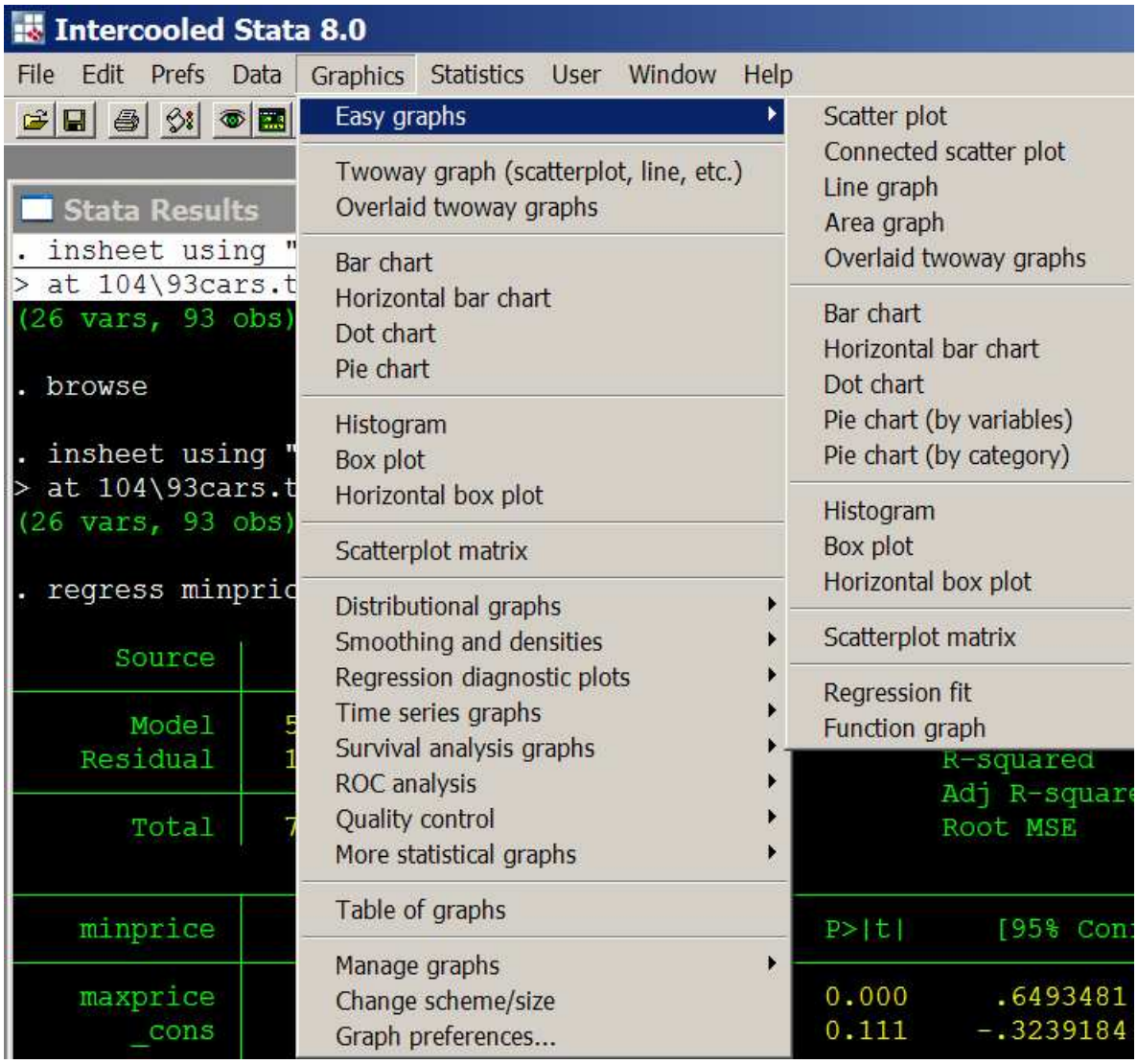
Also many command names can be abbreviated. For example you can use `ren` for `rename`, `reg` for `regress`, etc.

To see what they are, check the help pages. They can be displayed in the main results window, with a command like `help regress` or in Stata Viewer window (which you get if you click on the help button in the dialogue box)



The help facility isn't bad. One nice features is that it uses hyperlinks, allowing you to immediately go to another help page of interest.

Graphics



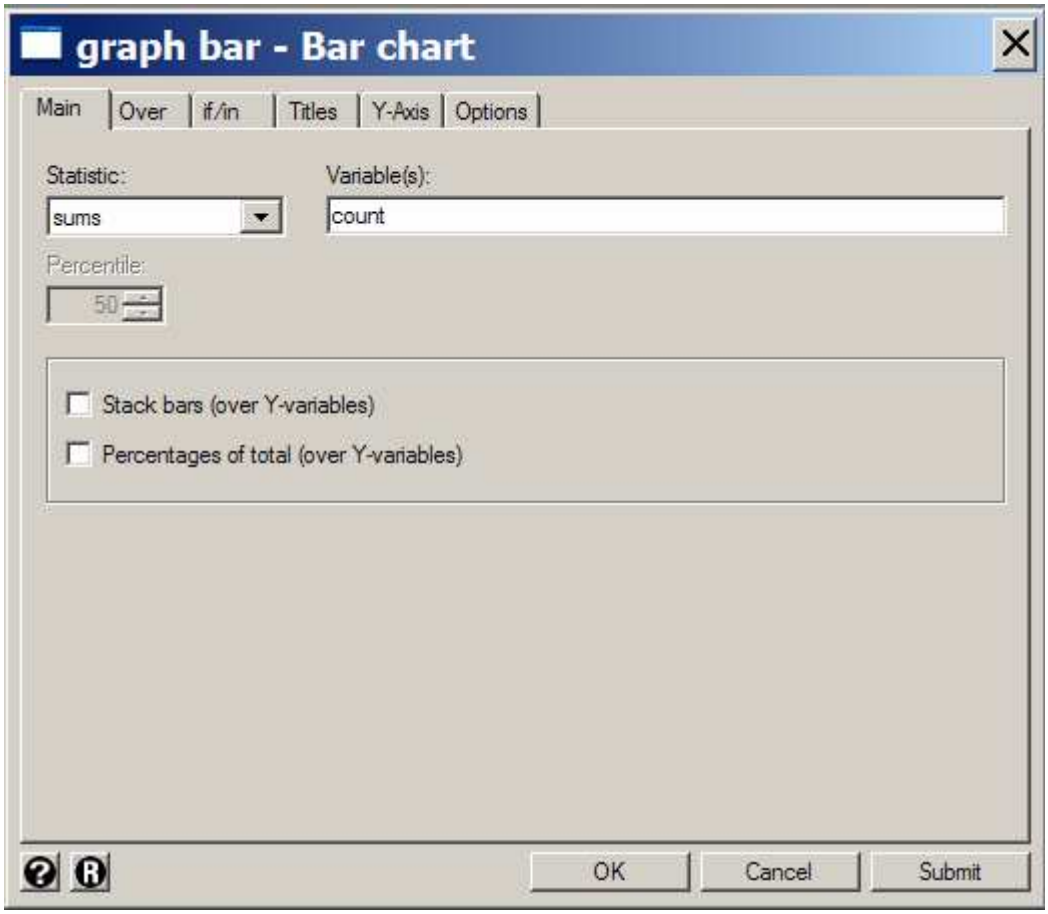
The quality of the graphs is ok. The resulting graphs are somewhat configurable (fonts, colours, etc), but there are some limitations.

For the common graphs, scatter plots, bar charts, etc, there are two approaches through the menus;

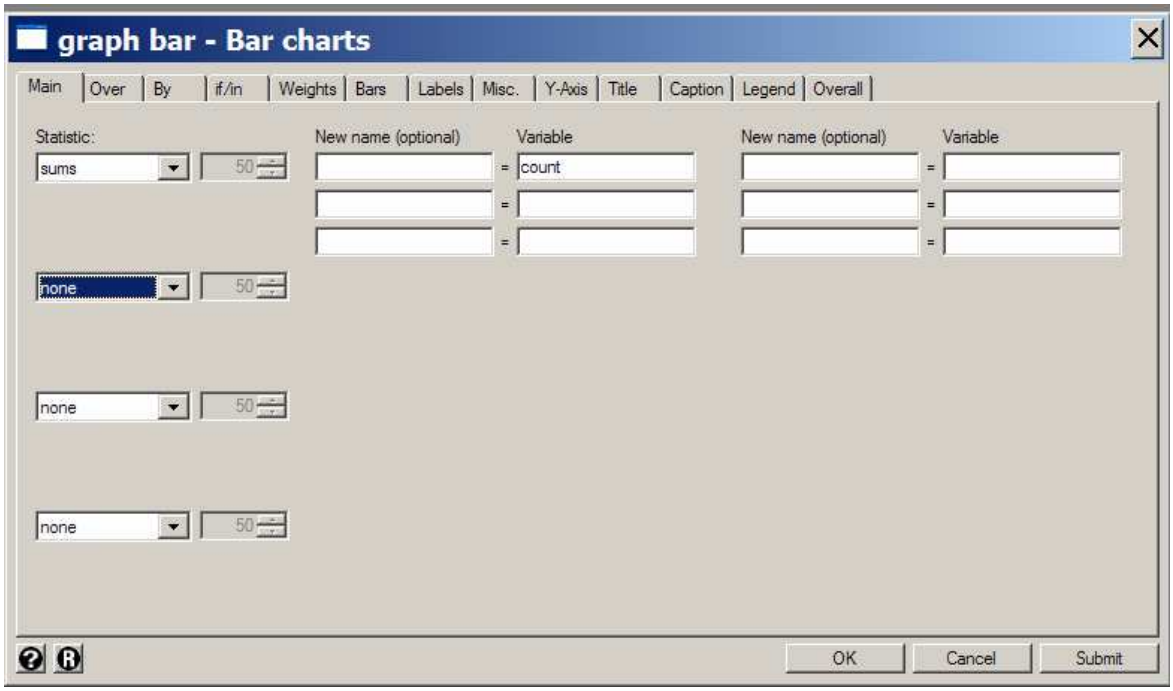
Easy graphs and main menu entry.

The Easy graph dialogue boxes are a bit friendlier, but are not as powerful (missing options).

Easy Graph approach to a bar chart

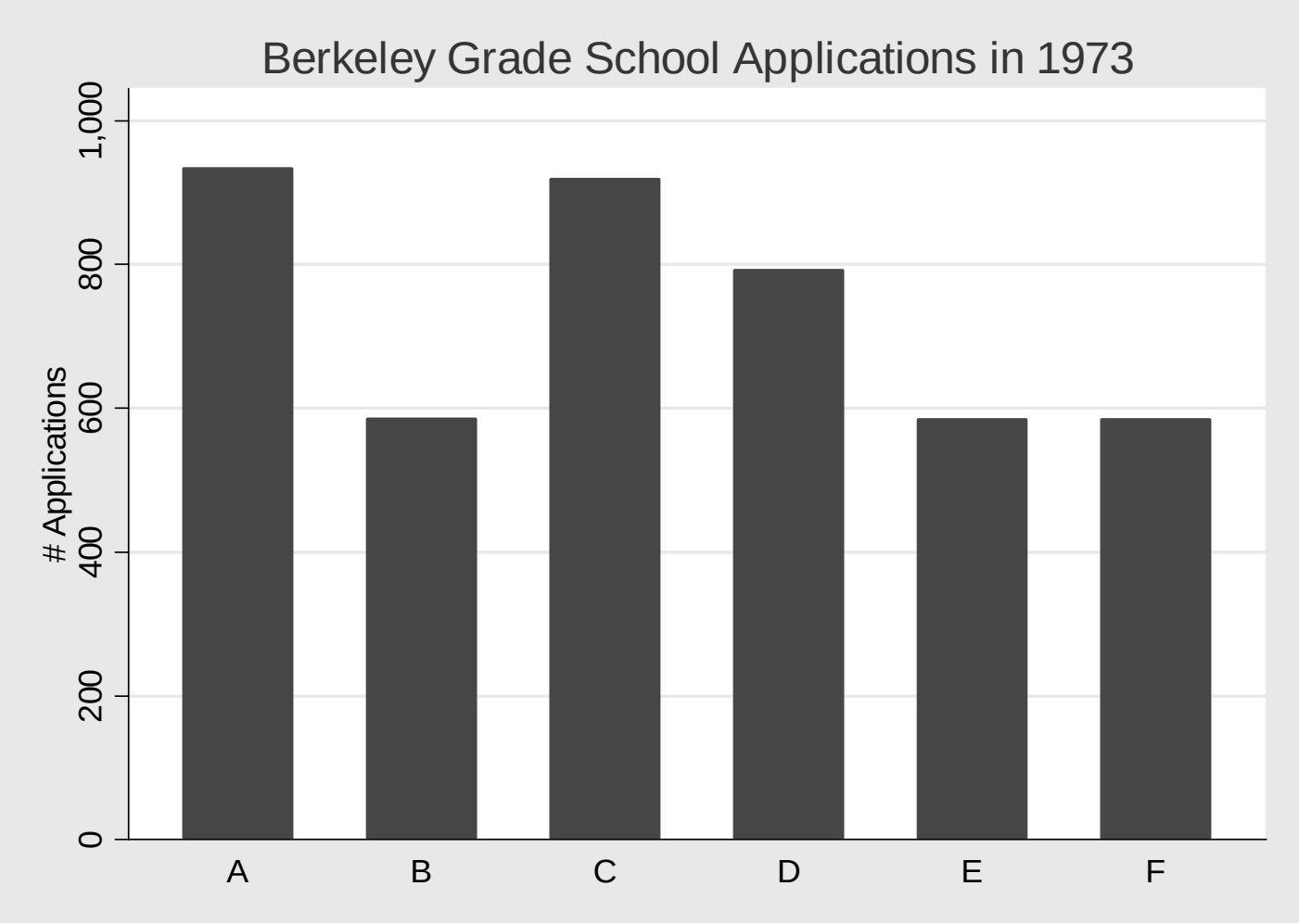


Main approach to a bar chart



Both approaches could yield the command

```
graph bar (sum) count, over(major) ytitle(# Applications) title(Berkeley Grade School Applications in 1973)
```



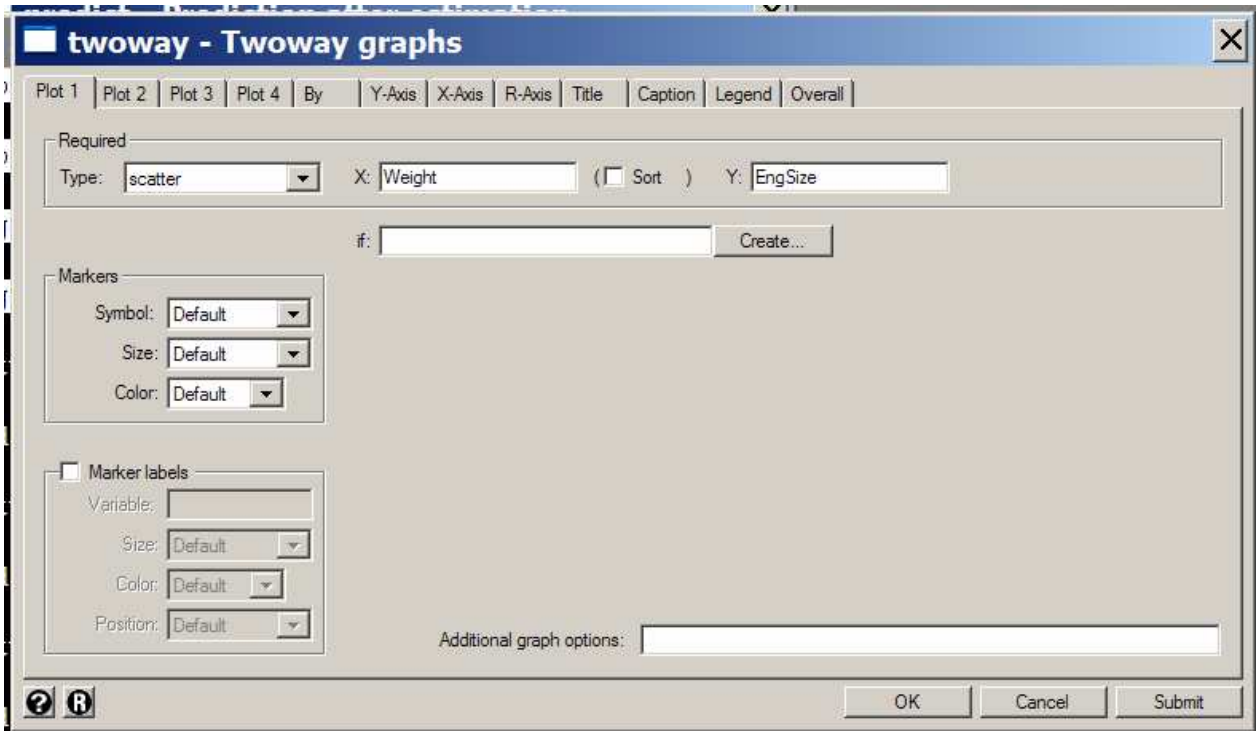
One drawback in the Stata approach to graphics is that you can't add onto figures, as with S-Plus/R and Matlab.

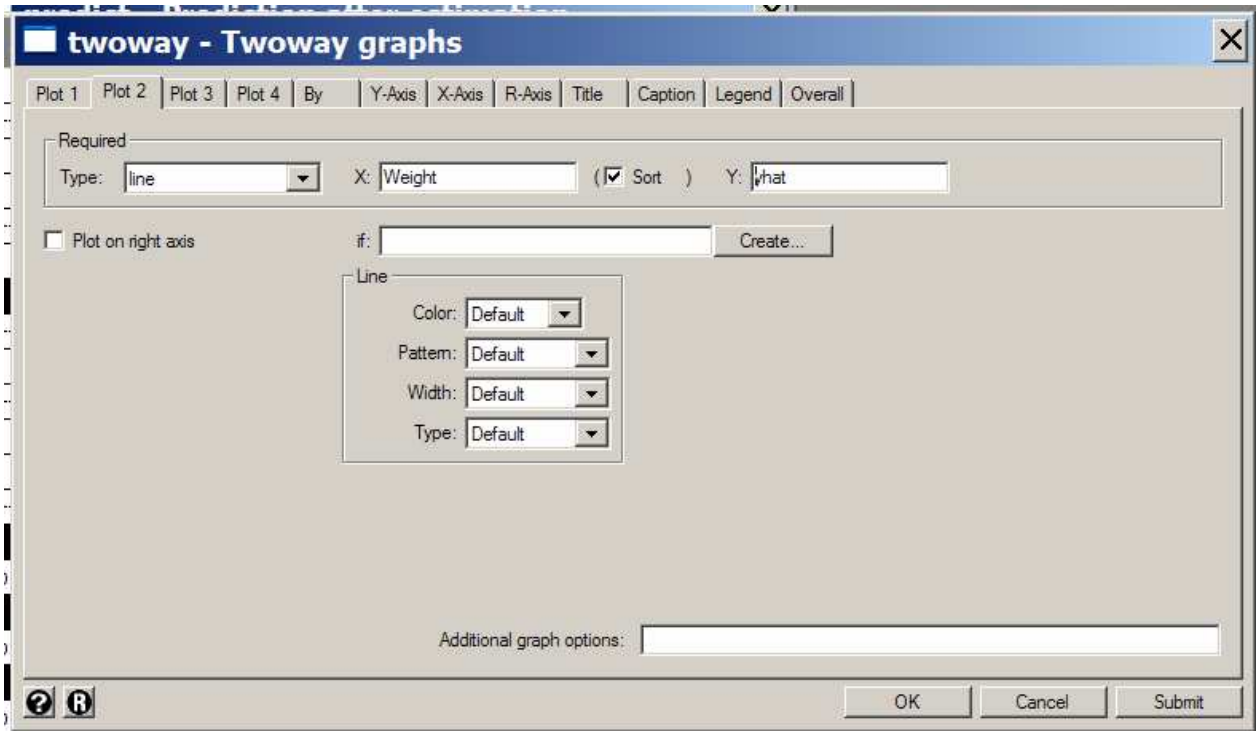
However you can create figures that overlay plots on top of each other.

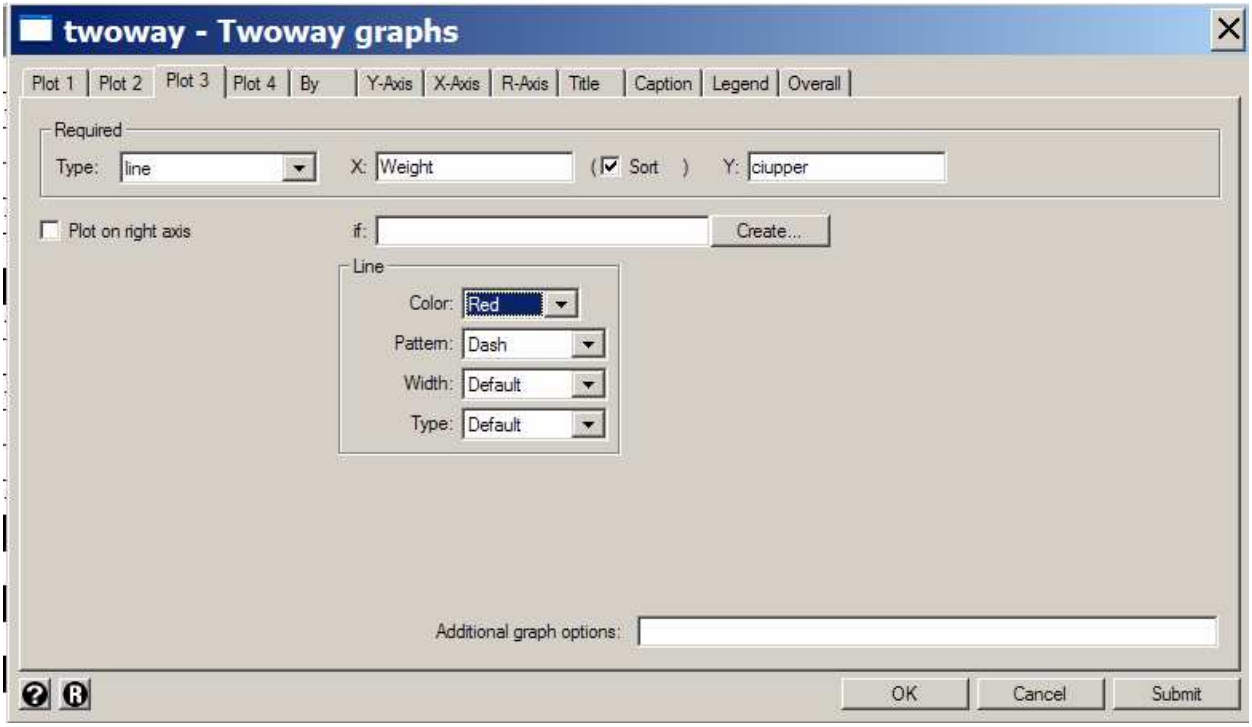
For example, one way to give a scatterplot with the regression line and 95% confidence intervals of the fit is

```
. regress EngSize Weight
  Source |       SS       df       MS              Number of obs =      93
---------+------------------------------           F(  1,    91) =  227.35
   Model |  70.7033831     1  70.7033831           Prob > F      =  0.0000
Residual |  28.2998401    91  .310987254           R-squared     =  0.7142
---------+------------------------------           Adj R-squared =  0.7110
   Total |  99.0032233    92  1.07612199           Root MSE      =  .55766
------------------------------------------------------------------------------
 EngSize |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Weight |   .0014861   .0000986    15.08   0.000     .0012903    .0016819
   _cons |  -1.898924    .308337    -6.16   0.000    -2.511398    -1.28645
------------------------------------------------------------------------------
. predict yhat, xb
. predict residual, residuals
. predict sefit, stdp
. predict sepred, stdf
. gen cilower = yhat - invttail(91,0.025)*sefit
. gen ciupper = yhat + invttail(91,0.025)*sefit
```

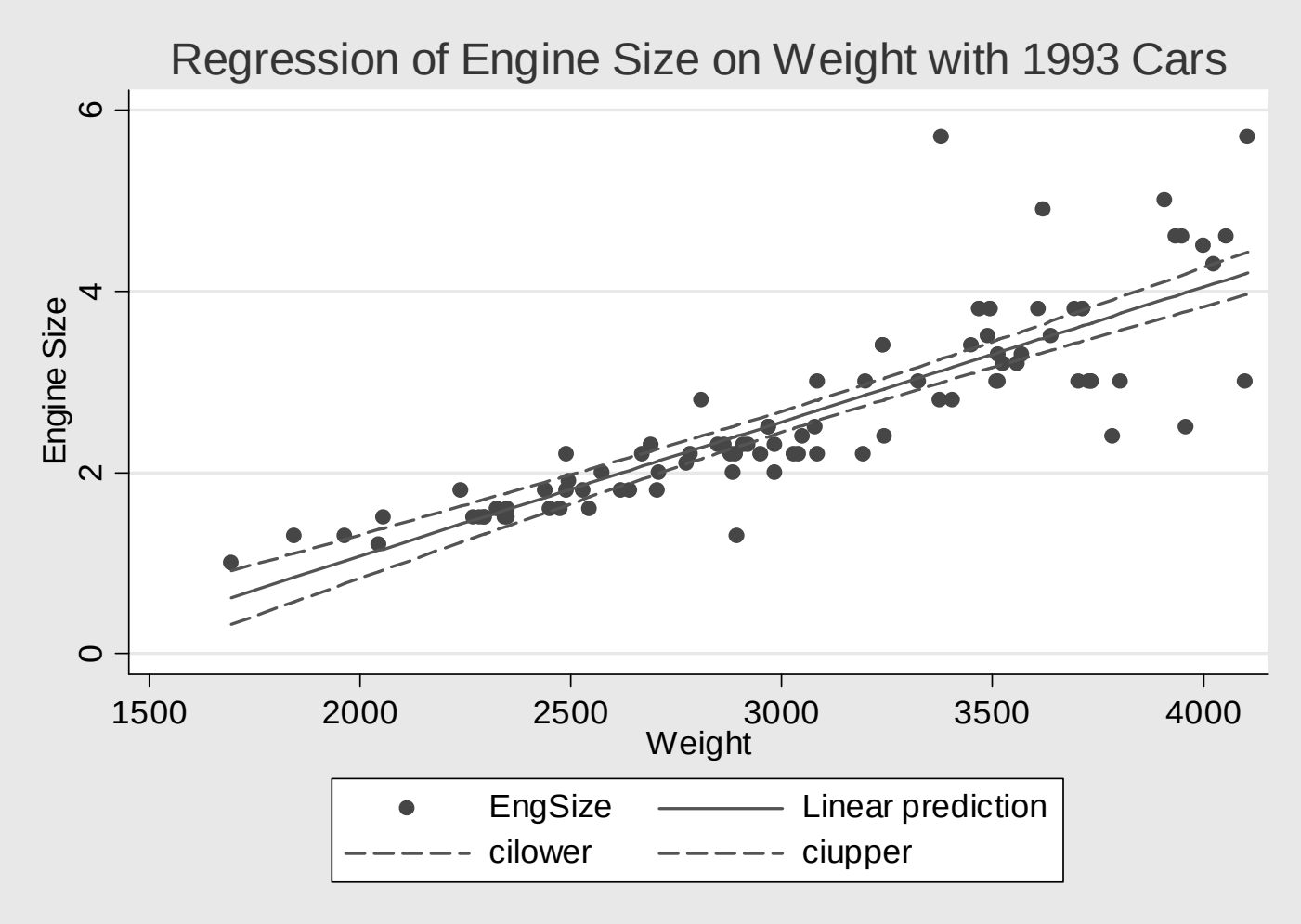
The `predict` command are available under `General Post-Estimation`


twoway - Twoway graphs
Plot 1
Plot 2
Plot 3
Plot 4
By
Y-Axis
X-Axis
R-Axis
Title
Caption
Legend
Overall
Required
Type: scatter
X: Weight
Sort
Y: EngSize
if:
Create...
Markers
Symbol: Default
Size: Default
Color: Default
Marker labels
Variable:
Size: Default
Color: Default
Position: Default
Additional graph options:
OK
Cancel
Submit


twoway - Twoway graphs
Plot 1
Plot 2
Plot 3
Plot 4
By
Y-Axis
X-Axis
R-Axis
Title
Caption
Legend
Overall
Required
Type: line
X: Weight
Sort
Y: what
Plot on right axis
if:
Create...
Line
Color: Default
Pattern: Default
Width: Default
Type: Default
Additional graph options:
OK
Cancel
Submit



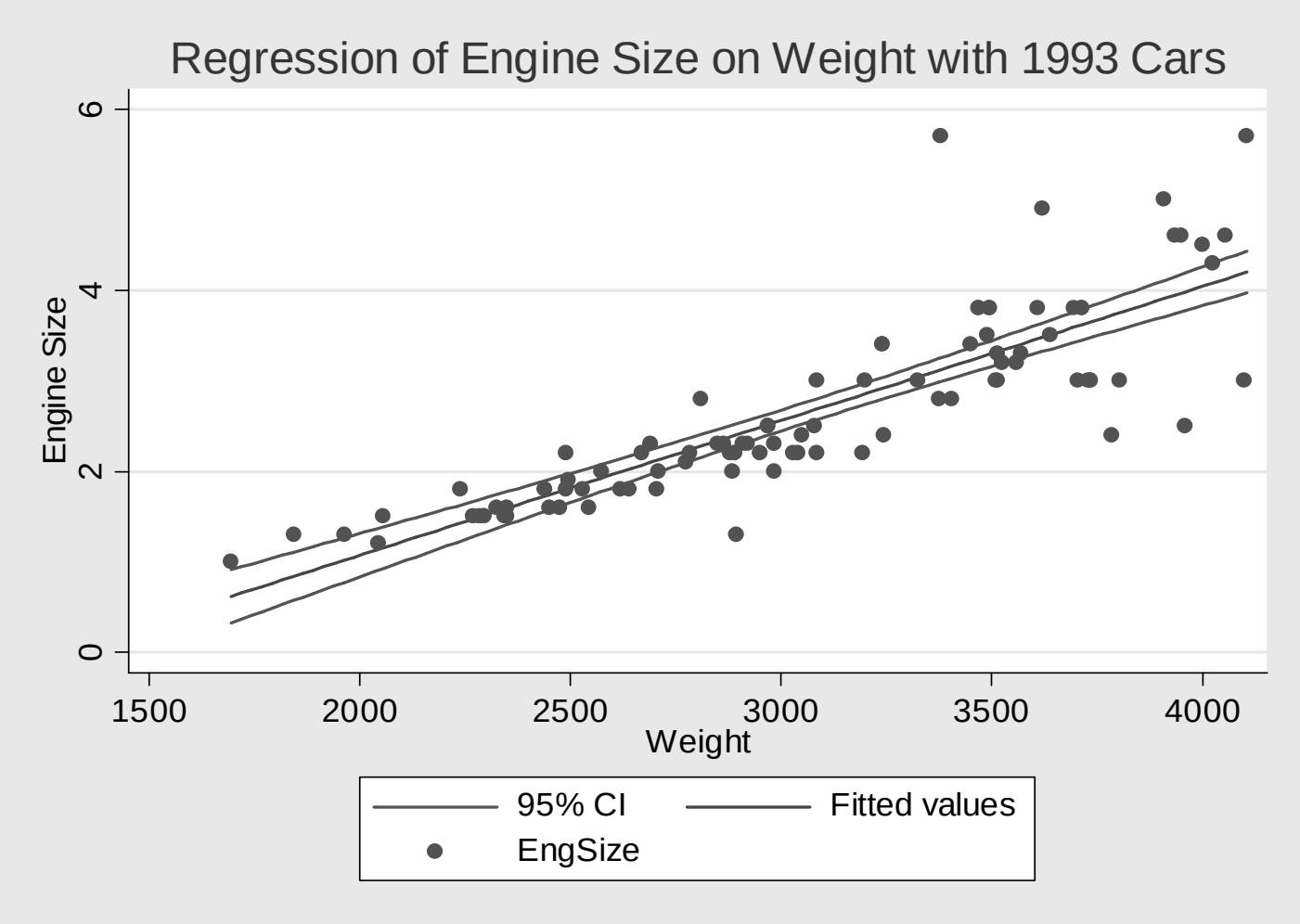

```
twoway (scatter EngSize Weight) (line yhat Weight,
sort) (line cilower Weight, sort clcolor(red)
clpat(dash)) (line ciupper Weight, sort clcolor(red)
clpat(dash)), ytitle(Engine Size) title(Regression of
Engine Size on Weight with 1993 Cars)
```



Actually there is an easier approach that doesn't involve running the regression ahead of time

```
. twoway (lfitci EngSize Weight,
ciplot(rline) blcolor(red)) (scatter
EngSize Weight, sort), ytitle(Engine Size)
title(Regression of Engine Size on Weight
with 1993
```

```
Cars)
```



Graphs in Stata can be saved in the following formats:

Stata graph (.gph), Windows metafile (.wmf), Enhanced metafile (.emf), Portable Network Graphics (.png), Postscript (.ps), and Encapsulated Postscript (.eps)

Multiway tables

Relationships between categorical variables

Example: 1973 graduate applications at Berkeley for the 6 largest programs

```
. table  major status gender [fweight=count],row col
```

	Gender and Status					
	Men			Women		
Major	Admitted	Rejected	Total	Admitted	Rejected	Total
A	511	314	825	89	19	108
B	353	207	560	17	8	25
C	120	205	325	202	391	593
D	138	279	417	131	244	375
E	54	137	191	94	299	393
F	22	351	373	24	317	341
Total	1,198	1,493	2,691	557	1,278	1,835

When dealing with data like this, it can be entered a couple of ways. For the above table, it was entered as

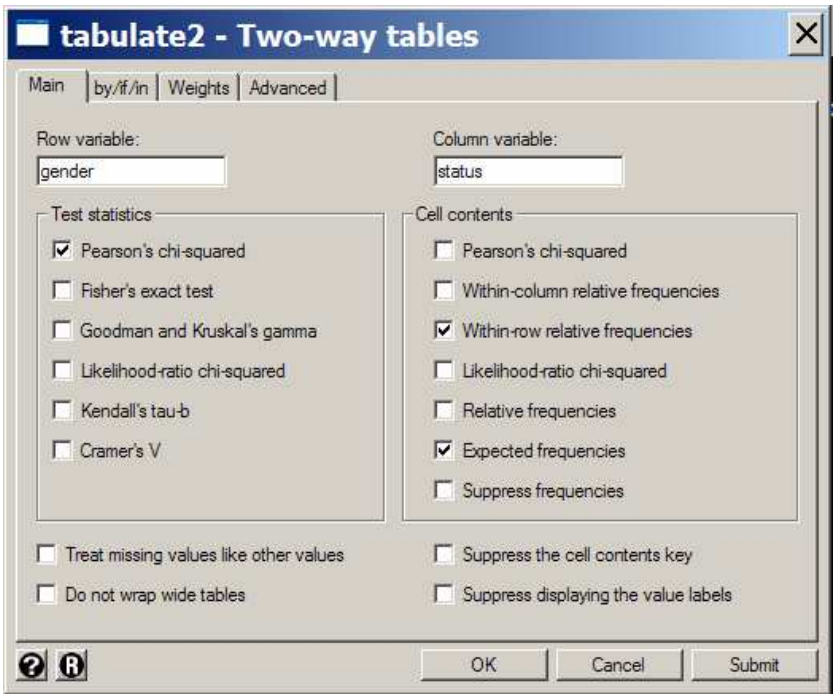
```
list
     +-----------------------------------+
     | major   gender     status   count |
     |-----------------------------------|
  1. |     A      Men   Admitted     511 |
  2. |     A      Men   Rejected     314 |
  3. |     A    Women   Admitted      89 |
  4. |     A    Women   Rejected      19 |
  5. |     B      Men   Admitted     353 |
  6. |     B      Men   Rejected     207 |
  7. |     B    Women   Admitted      17 |

 19. |     E    Women   Admitted      94 |
 20. |     E    Women   Rejected     299 |
 21. |     F      Men   Admitted      22 |
 22. |     F      Men   Rejected     351 |
 23. |     F    Women   Admitted      24 |
 24. |     F    Women   Rejected     317 |
     +-----------------------------------+
```

Instead, there could be a line for each person. For example you would need 551 lines with A Men Admitted, 314 lines with A Men Rejected, etc, for a total of 4526 entries.

The `fweight` (frequency weight) entry is saying that each line should be treated as existing `count` times.

Note that `fweight` can be used with many commands. Also there are other types of weights (sampling, analytic, importance)




```
tabulate gender status [fweight=count ], chi2
expected row
```

```
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|   row percentage   |
+--------------------+
```

```
           |        Status
    Gender |  Admitted   Rejected |     Total
-----------+----------------------+----------
       Men |     1,198      1,493 |     2,691
           |   1,043.5    1,647.5 |   2,691.0
           |     44.52      55.48 |    100.00
-----------+----------------------+----------
     Women |       557      1,278 |     1,835
           |     711.5    1,123.5 |   1,835.0
           |     30.35      69.65 |    100.00
-----------+----------------------+----------
     Total |     1,755      2,771 |     4,526
           |   1,755.0    2,771.0 |   4,526.0
           |     38.78      61.22 |    100.00

          Pearson chi2(1) =  92.2053   Pr = 0.000
```

So it appears that women had a much harder time to get into Berkeley in 1973. However, lets look at each Major separately

```
by major, sort: tabulate gender status
[fweight=count ], chi2 expect exact
```

```
-> major = A

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
+--------------------+

           |        Status
    Gender |  Admitted   Rejected |     Total
-----------+----------------------+----------
       Men |       511        314 |       825
           |     530.5      294.5 |     825.0
-----------+----------------------+----------
     Women |        89         19 |       108
           |      69.5       38.5 |     108.0
-----------+----------------------+----------
     Total |       600        333 |       933
           |     600.0      333.0 |     933.0

          Pearson chi2(1) =  17.4307   Pr = 0.000
           Fisher's exact =                 0.000
```

```
-> major = B

           |        Status
    Gender |  Admitted   Rejected |     Total
-----------+----------------------+----------
       Men |       353        207 |       560
           |     354.2      205.8 |     560.0
-----------+----------------------+----------
     Women |        17          8 |        25
           |      15.8        9.2 |      25.0
-----------+----------------------+----------
     Total |       370        215 |       585
           |     370.0      215.0 |     585.0

          Pearson chi2(1) =   0.2537   Pr = 0.614
           Fisher's exact =                 0.677
```

```
-> major = C

           |        Status
    Gender |  Admitted   Rejected |     Total
-----------+----------------------+----------
       Men |       120        205 |       325
           |     114.0      211.0 |     325.0
-----------+----------------------+----------
     Women |       202        391 |       593
           |     208.0      385.0 |     593.0
-----------+----------------------+----------
     Total |       322        596 |       918
           |     322.0      596.0 |     918.0

          Pearson chi2(1) =   0.7535   Pr = 0.385
           Fisher's exact =                 0.387
```

```
-> major = D
           |        Status
    Gender |  Admitted   Rejected |     Total
-----------+----------------------+----------
       Men |       138        279 |       417
           |     141.6      275.4 |     417.0
-----------+----------------------+----------
     Women |       131        244 |       375
           |     127.4      247.6 |     375.0
-----------+----------------------+----------
     Total |       269        523 |       792
           |     269.0      523.0 |     792.0
          Pearson chi2(1) =   0.2980   Pr = 0.585
           Fisher's exact =                 0.599
```

```
-> major = E

           |        Status
    Gender |  Admitted   Rejected |     Total
-----------+----------------------+----------
       Men |        54        137 |       191
           |      48.4      142.6 |     191.0
-----------+----------------------+----------
     Women |        94        299 |       393
           |      99.6      293.4 |     393.0
-----------+----------------------+----------
     Total |       148        436 |       584
           |     148.0      436.0 |     584.0

          Pearson chi2(1) =   1.2877   Pr = 0.256
           Fisher's exact =                 0.266
```

```
-> major = F

           |        Status
    Gender |  Admitted   Rejected |     Total
-----------+----------------------+----------
       Men |         22        351 |       373
           |       24.0      349.0 |     373.0
-----------+----------------------+----------
     Women |         24        317 |       341
           |       22.0      319.0 |     341.0
-----------+----------------------+----------
     Total |         46        668 |       714
           |       46.0      668.0 |     714.0

          Pearson chi2(1) =   0.3841   Pr = 0.535
           Fisher's exact =                 0.546
```

Summary of tests

Major	Chi2	P-value
A	17.4307	0.000
B	0.2537	0.614
C	0.7535	0.385
D	0.2980	0.585
E	1.2877	0.256
F	0.3841	0.535

So in every major, expect A, the data is consistent with the acceptance probabilities being the same for men and women. In major A, women appear to be accepted at a higher rate than men 82.4% vs 61.9

The data set gives an example of Simpson's paradox, where aggregating over third factor can switch the direction of association between to factors.

For a further description of Simpson's Paradox, see Moore and McCabe or a short article I wrote for the OSU Biostat Center's newsletter a few years ago (link on the overheads page)

Pearson Chi-square test

Can be used to check for independence between two categorical variables (or homogeneity of proportions between one factor, for each level of a second factor)

Compares observed counts within each cell with the expected counts assuming independence (or homogeneity)

These are gotten by

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}$$

The common test statistic for examining whether there is association between two variables in a $r \times c$ table is

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed count - expected count})^2}{\text{expected count}}$$

If the null hypothesis isn't true, there should be some cells in the table where the observed and expected counts are very different. This would lead to a large value for X^2.

If the number of observations is large, the sampling distribution of X^2 is approximately Chi-square (χ^2) with $(r - 1)(c - 1)$ degrees of freedom.

What to do if the sample sizes are small.

Similar to the one sample binomial problems, there is an exact procedure for two way tables that should be used when the normal approximation doesn't hold. In the 2×2 table case, its known as Fisher's exact test. It can be easily performed in Stata.

Idea of Fisher's exact test

Assume that the margins for each table are fixed

```
-> major = B

           |        Status
    Gender |  Admitted   Rejected |     Total
-----------+----------------------+----------
       Men |       353        207 |       560
-----------+----------------------+----------
     Women |        17          8 |        25
-----------+----------------------+----------
     Total |       370        215 |       585

          Pearson chi2(1) =   0.2537   Pr = 0.614
           Fisher's exact =                 0.677
```

In a 2 by 2 table, with fixed margins, once you set one cell in the table, the others are determined exactly. For example lets set the value for the Rejected Women cell at 6. Then the other cells must be 351, 209, and 19.

Under the null hypothesis, the probability of any possible table is given by the hypergeometric distribution

a	b	U
c	d	V
S	T	N

The probability of this table is given by

$$\frac{\binom{U}{a}\binom{V}{c}}{\binom{N}{S}}$$

Fisher's exact tests sums these probabilities over all tables considered as or more extreme than the observed table.