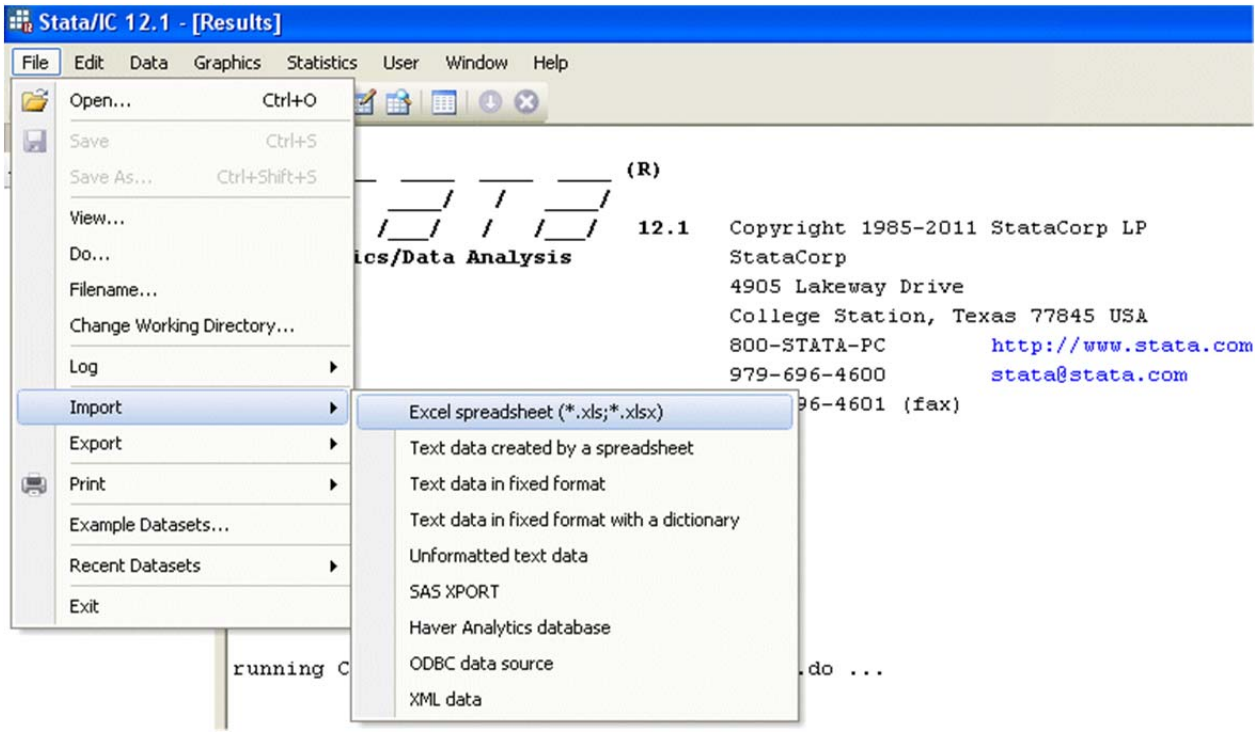
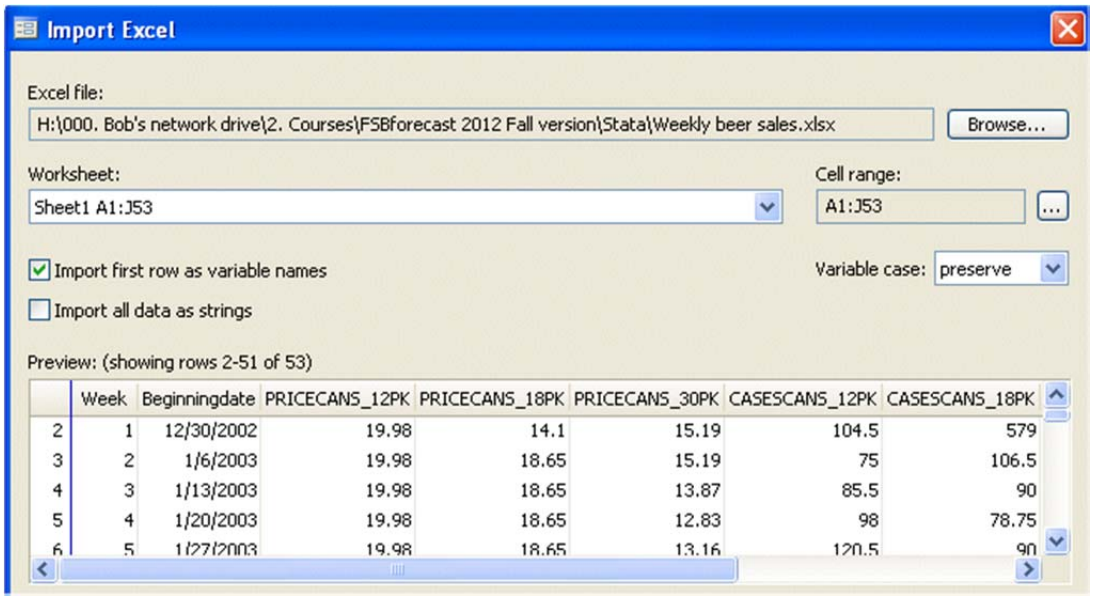
Data analysis and regression in Stata

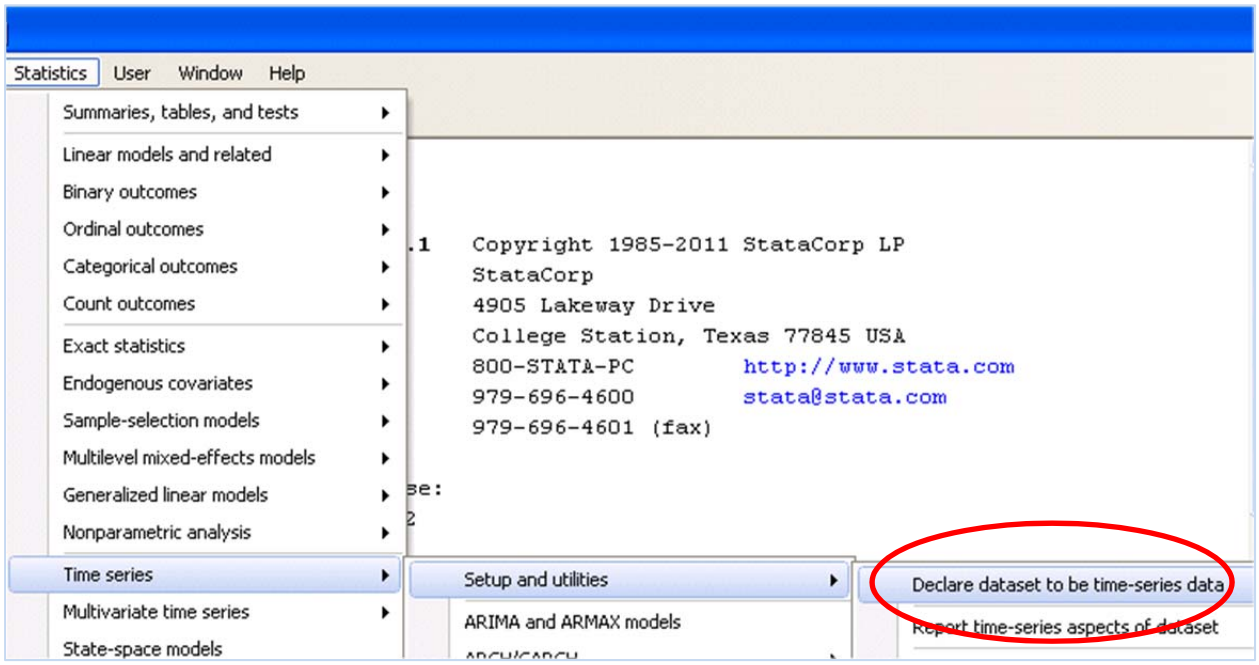
This handout shows how the weekly beer sales series might be analyzed with Stata (the software package now used for teaching stats at Kellogg), for purposes of comparing its modeling tools and ease of use to those of FSBForecast. To analyze the weekly beer sales series, the first step is to import the data from the Excel file. Any statistical software package can import Excel files easily.



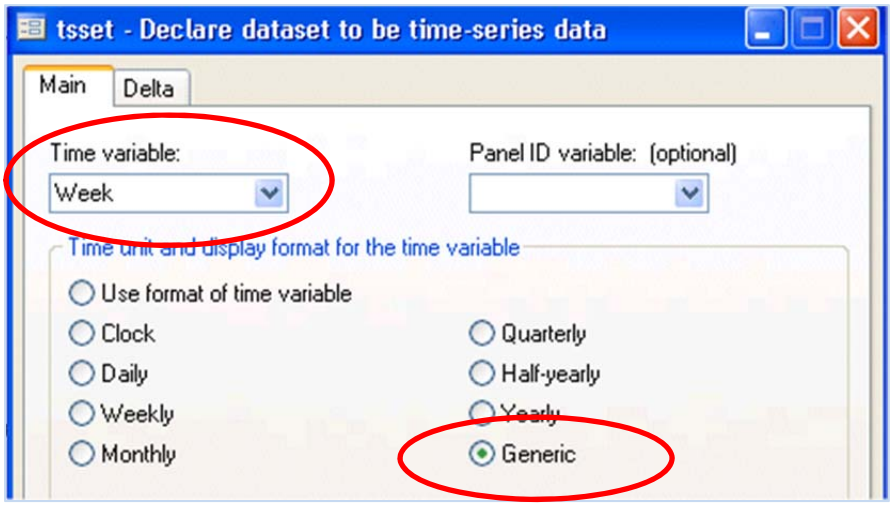
The dialog box for importing the Excel file offers the option of reading the variable names from the first row, which is also standard. So, the same data file that worked for FSBForecast will work here.



In order to be able to use time transformation options later on, it is necessary to declare the variables to be time series. To do this, don't go to the "Data" menu, which is where most data-definition operations are performed. Instead, go to the "Statistics" menu and look under the" time series" options there.



For simplicity, let's just use the week number as the "generic" time index:

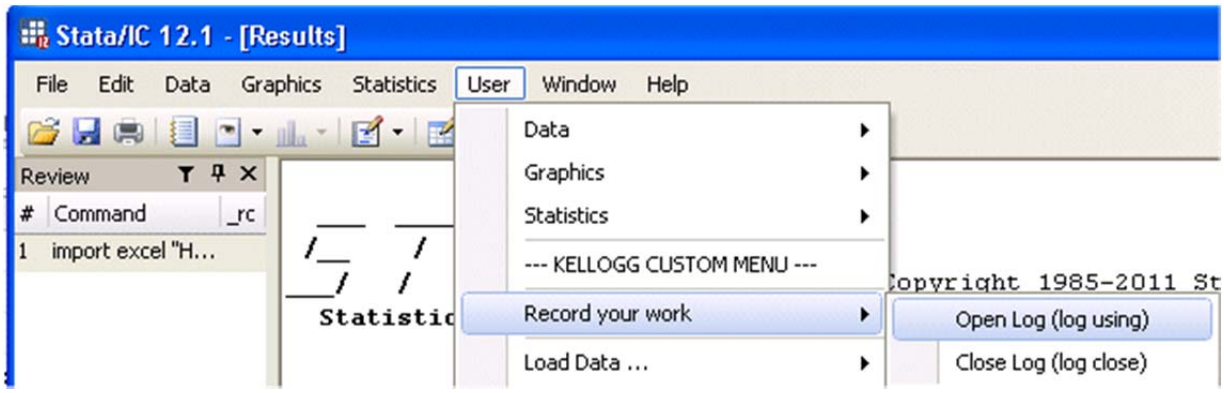


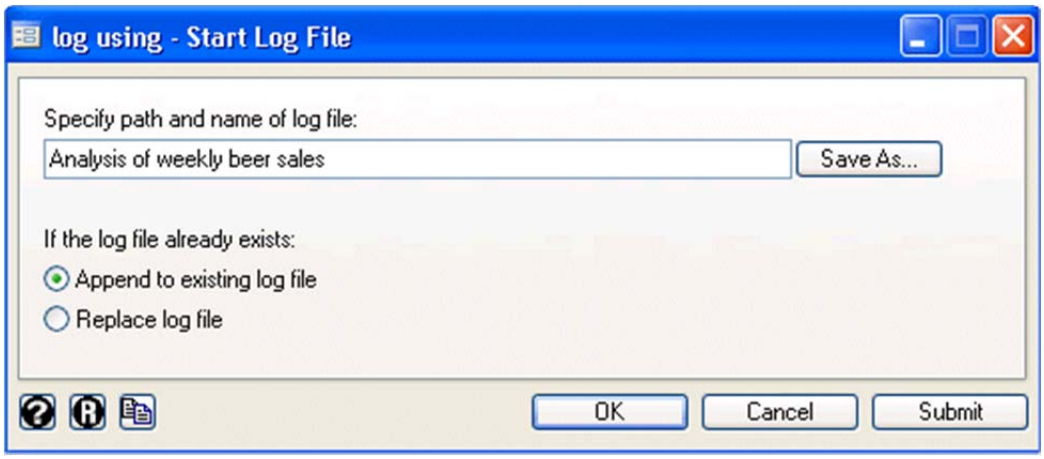
This executes the following command which is shown in the output window:

```
. tsset Week, generic
        time variable:  Week, 1 to 52
                delta:  1 unit
```

Under the hood this is a command-language program, as are SPSS and SAS. Choosing options from the menu causes the appropriate code to be generated and executed. Most serious users of programs like Stata write their code directly rather than letting a menu system do it for them.

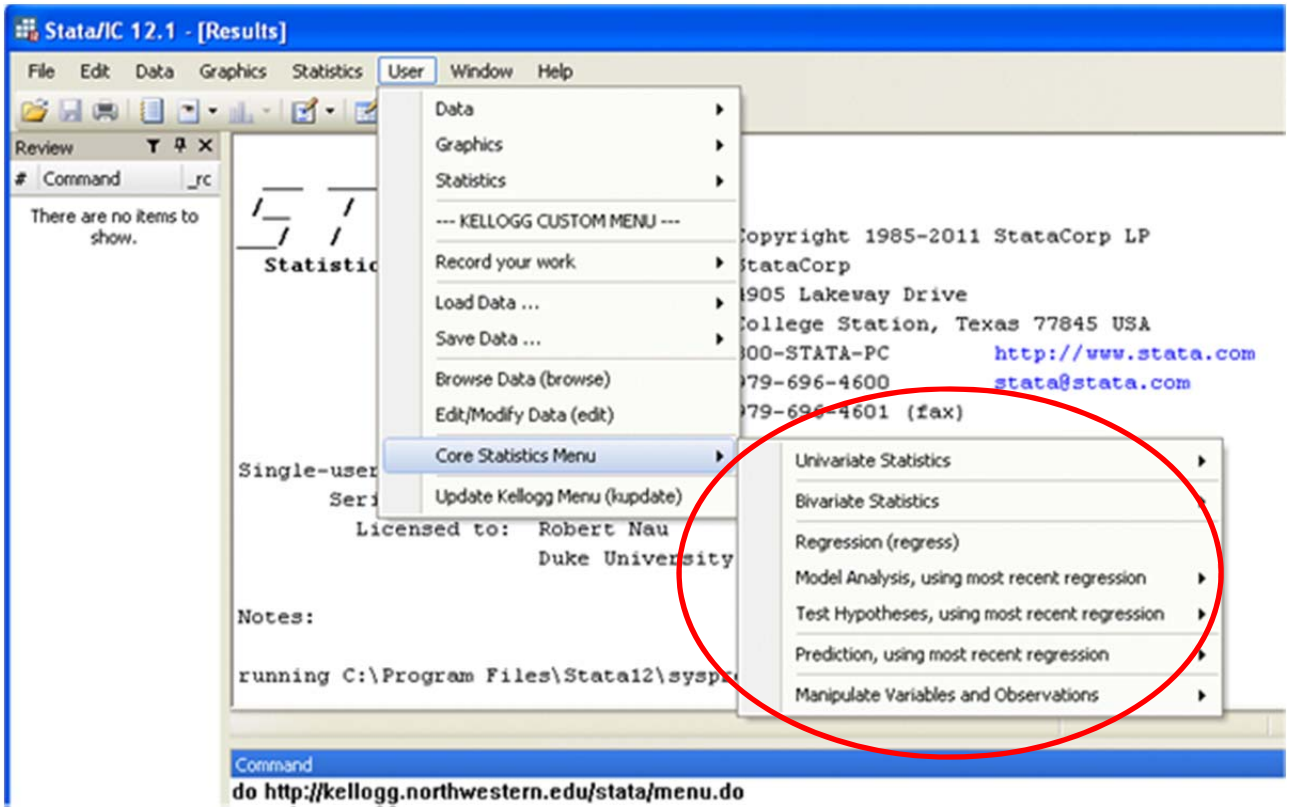
The numerical results of your analysis will be written to the output window along with the code that created them, in the form of a single scrolling log file. Before proceeding with your analysis, you need to open and assign a name to the log file for your session, so that you can save your results later:



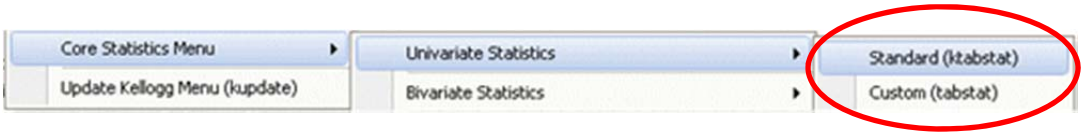


The log file is a plain text file that contains the output that you see scrolling by while doing your analysis. It contains only the text output, not the graphs. It can be opened and edited later with Microsoft Word or other text-editing software, and you can use the "append" option to re-open the log file and add more analysis to it later. If you open it in Microsoft Word, you should change the font size to 8 points to avoid wrapping the lines and use a fixed-width font such as Courier so that table contents will line up.

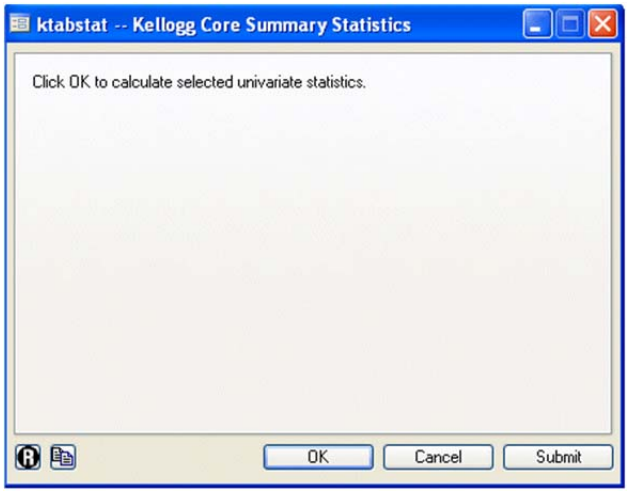
Here is Kellogg's custom menu for their core statistics class, which can be loaded by typing the "do" statement shown in the command window at the very bottom of the screen:



The “univariate statistics” command provides summary statistics of some or all variables:



If you choose the standard summary statistics report, which shows pre-selected statistics for all variables in the file, you must click through this screen:



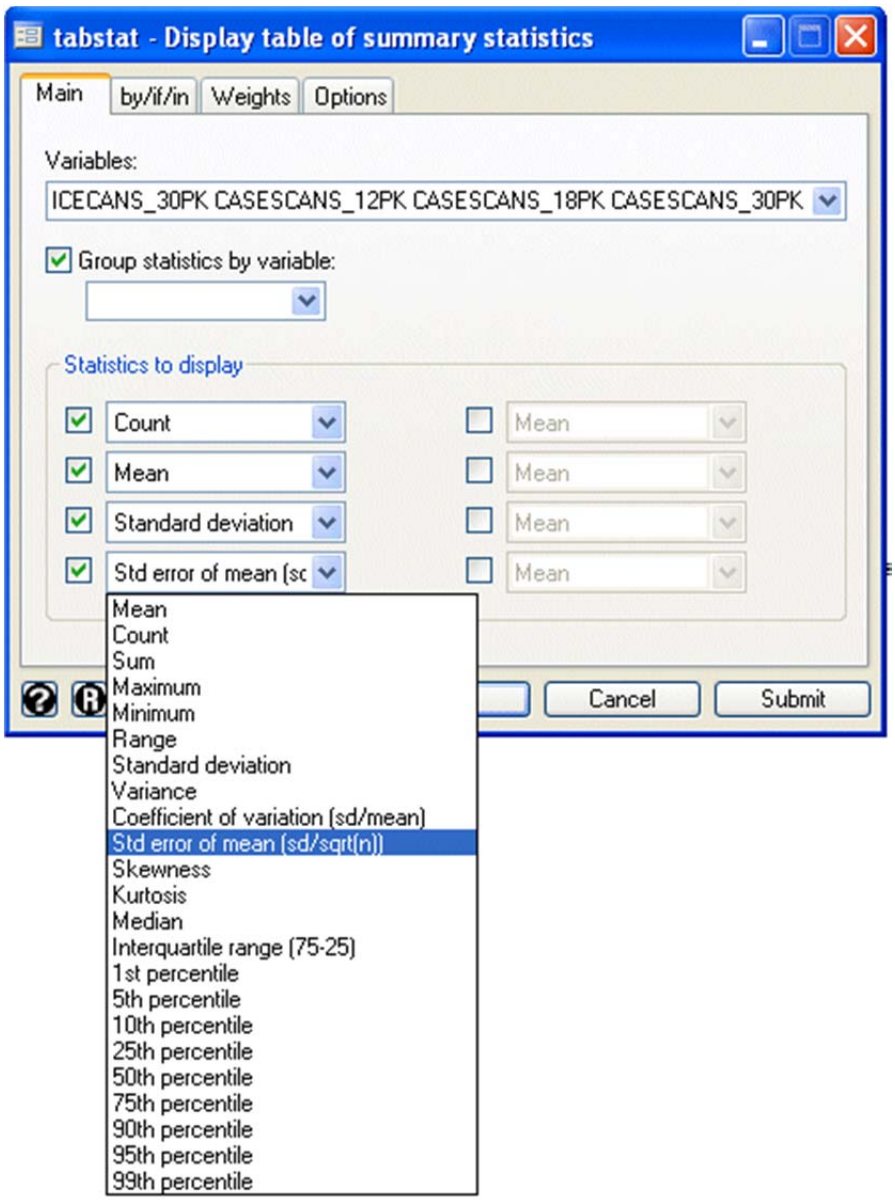
You then get the following output:

```
. ktabstat
preserve
destring, replace force
tabstat _all, s(mean sd semean min median max range skewness kurtosis count)
```

stats	Week	Beginn~e	PRIC~2PK	PRIC~8PK	PRIC~OPK	CASE~2PK	CASE~8PK	CASE~OPK	CASEST~L	PRICEA~E
mean	26.5	15882.5	19.08769	16.72462	14.37923	95.25	189.476	265.4567	550.1827	15.2425
sd	15.15476	106.0833	2.088128	2.411076	.8057924	63.66853	174.3693	179.5747	239.2266	.9546499
se(mean)	2.101587	14.71111	.2895712	.3343561	.1117433	8.829236	24.18068	24.90253	33.17475	.1323861
min	1	15704	14.33	13.26	12.83	14	18.75	41.25	120	14.07
p50	26.5	15882.5	19.98	18.65	14.395	78	90.375	219.375	542.875	15.03
max	52	16061	21.28	19.5	15.19	334.5	639	837.5	1119.5	17.81
range	51	357	6.95	6.24	2.36	320.5	620.25	796.25	999.5	3.74
skewness	0	0	-1.39919	-.2582795	-.323973	1.888871	1.058442	1.026605	.0516941	1.041537
kurtosis	1.799112	1.799112	3.34019	1.163823	1.628442	6.68313	2.766275	3.887676	2.322529	3.358092
N	52	52	52	52	52	52	52	52	52	52

Apparently there is no way around the default abbreviation of variable names on some reports. Best to use short names (8 characters or less)!

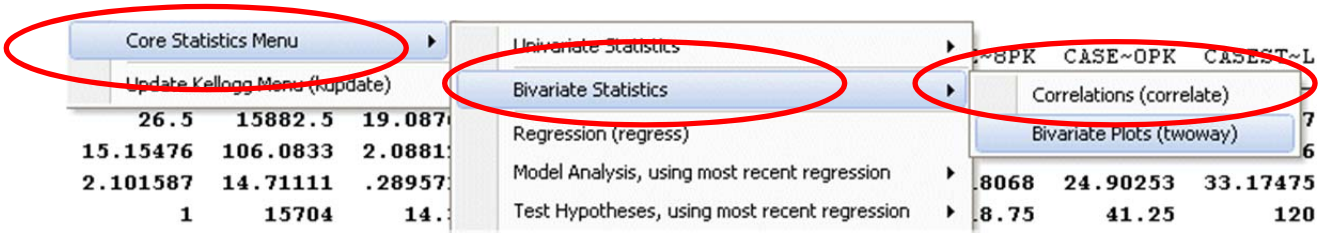
In a custom univariate statistics report, you can choose a subset of variables to analyze, and you can choose up to 8 stats by a sequence of steps in which you use separate pull-down menus:



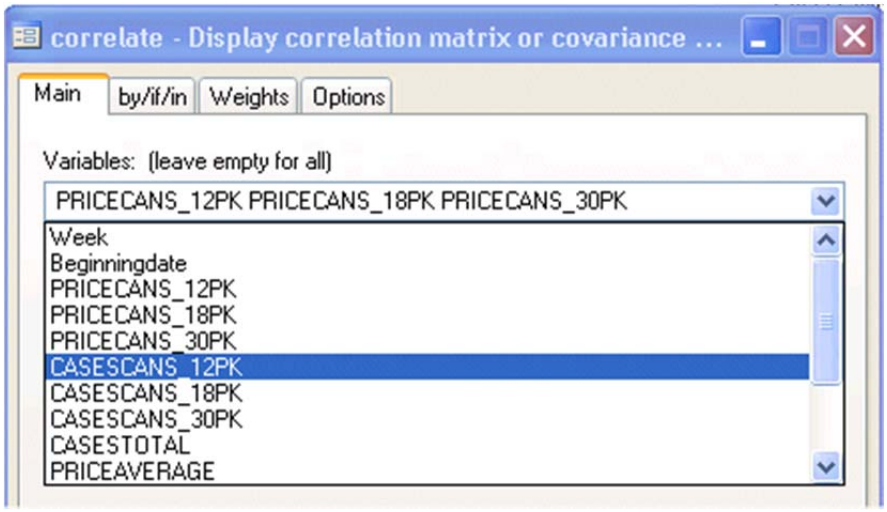
Here is the output of this particular custom analysis, which shows 4 stats for 6 variables:

stats	PRIC~2PK	PRIC~8PK	PRIC~OPK	CASE~2PK	CASE~8PK	CASE~OPK
N	52	52	52	52	52	52
mean	19.08769	16.72462	14.37923	95.25	189.476	265.4567
sd	2.088128	2.411076	.8057924	63.66853	174.3693	179.5747
se(mean)	.2895712	.3343561	.1117433	8.829236	24.18068	24.90253

How to generate a correlation matrix:



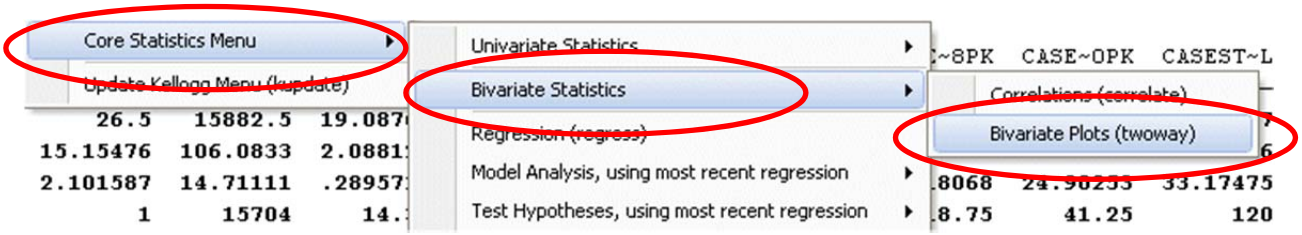
This command opens a dialog box in which you can choose a list of variables by clicking on them. As each one is clicked, it is added to the list in the window, which is typical of all procedures in Stata that operate on multiple variables.

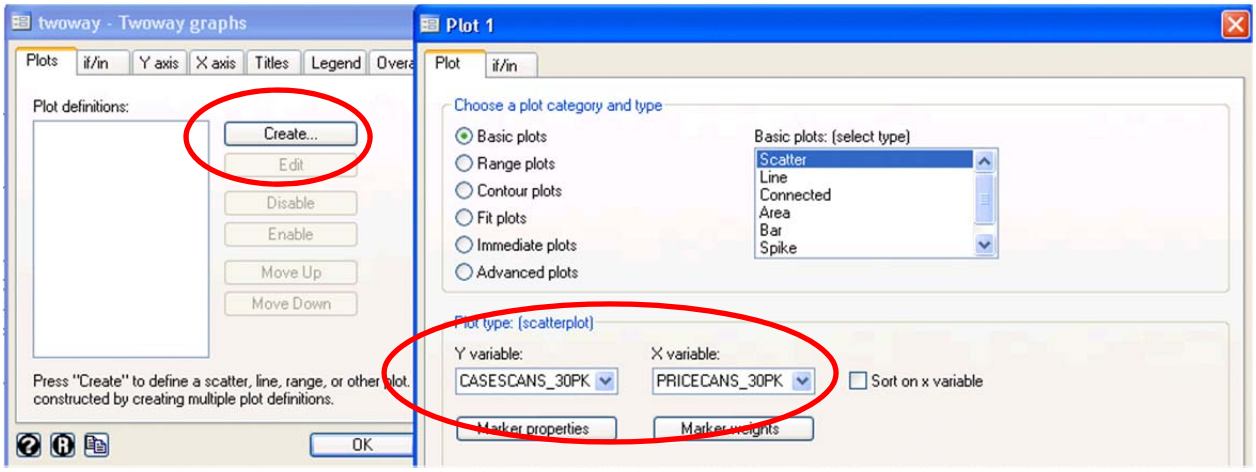


This 6-variable correlation matrix is small enough to fit in the Stata output window. A larger matrix would be broken into pieces when it was displayed, and I do not think it would be possible to copy it to a spreadsheet or other document where you could see it as a single triangular array. Again, it is best to use short variable names.

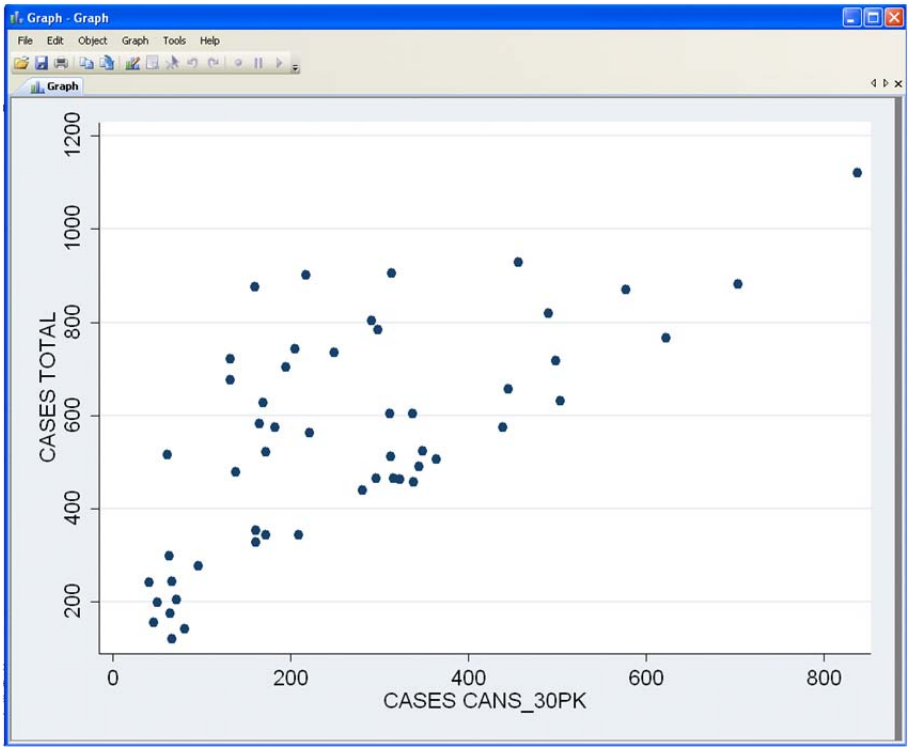
```
. correlate PRICECANS_12PK PRICECANS_18PK PRICECANS_30PK CASESCANS_12PK
(obs=52)

             | PRIC~2PK PRIC~8PK PRIC~0PK CASE~2PK CASE~8PK CASE~0PK
-------------+------------------------------------------------------
PRICECAN~2PK |   1.0000
PRICECAN~8PK |  -0.0836   1.0000
PRICECAN~0PK |  -0.3635  -0.2515   1.0000
CASESCAN~2PK |  -0.6253   0.2963   0.0186   1.0000
CASESCAN~8PK |   0.2168  -0.8173   0.1329  -0.2091   1.0000
CASESCAN~0PK |   0.2823   0.3123  -0.8718   0.1971  -0.1492   1.0000
```

How to get a single scatterplot:

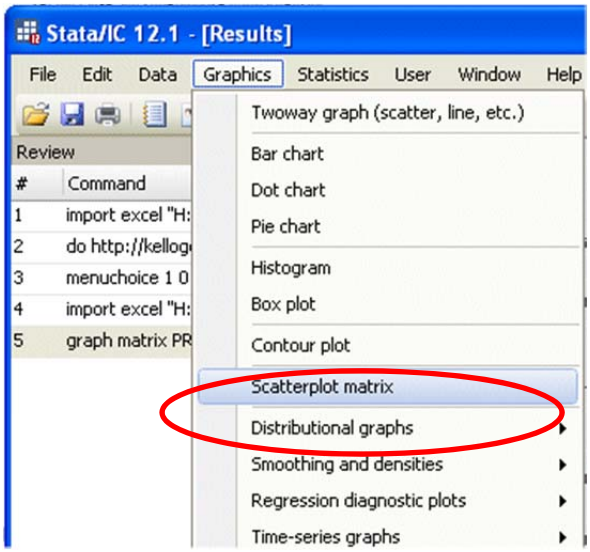




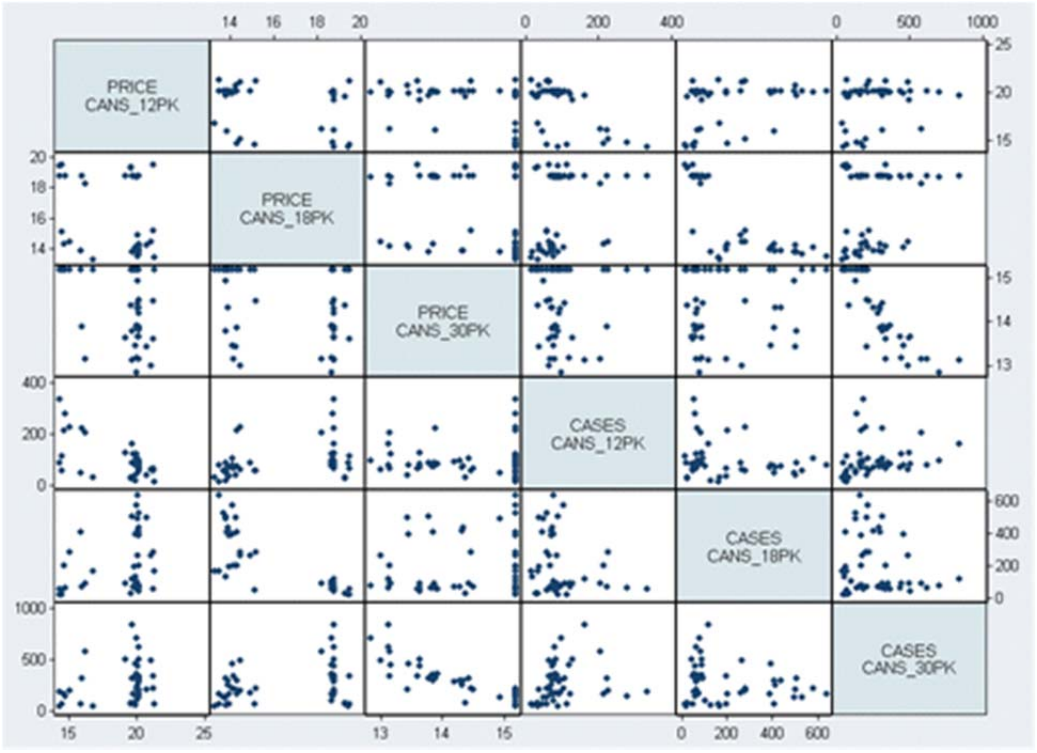
The graph is displayed in a separate graph windows that opens up:



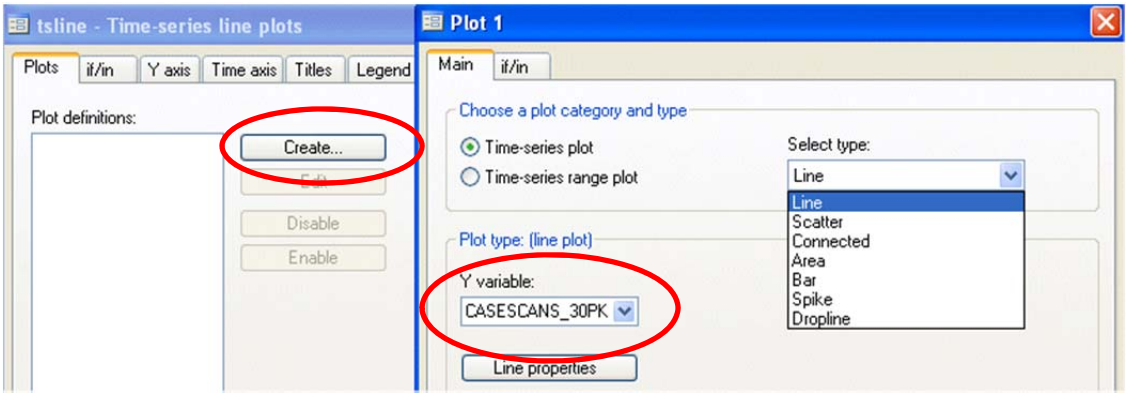
The main graphics menu has other plot options, including a scatterplot matrix:

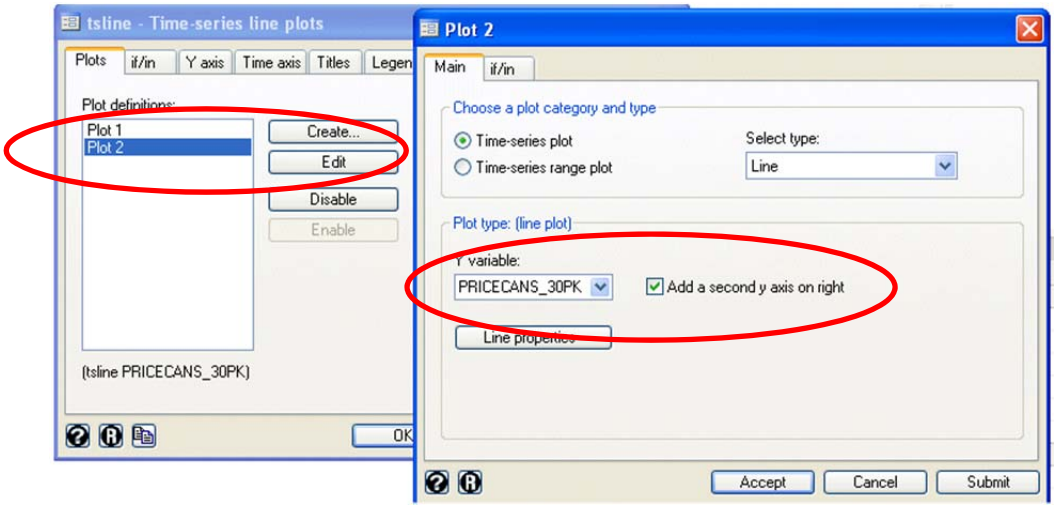


The new graph replaces the old one in the graph window. You can't get multiple open graph windows (i.e., more than one graph visible at a time) using only menu commands. You can do it by writing code, though. The graph that is currently in the window can be directly copied and pasted to other documents like this Word file. This is a nicely formatted plot, although (as with a correlation matrix) you have to read across and down to determine which variables are shown in a given plot, as well as their axis scale numbers. Axis scale numbers are provided on each separate scatterplot in the matrix in FSBforecast, along with the correlations and their squared values or regression slope coefficients.

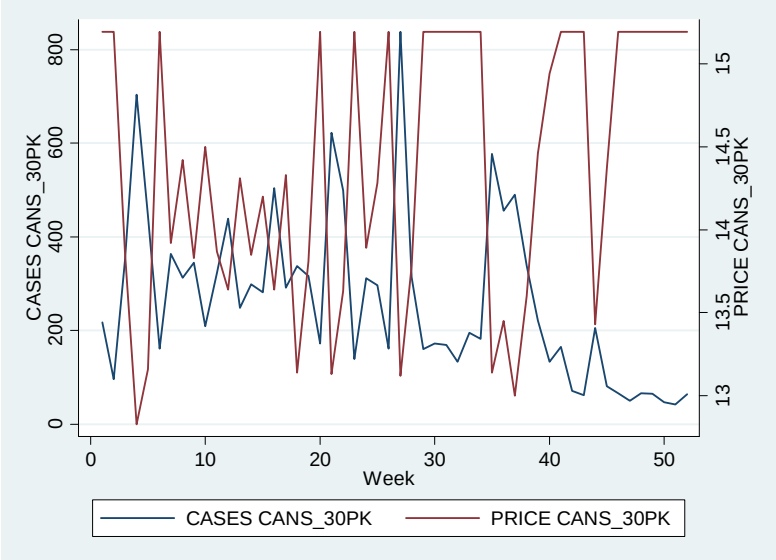


How to plot 2 time series on the same chart by using the “time series graphs” option on the main graphics menu:



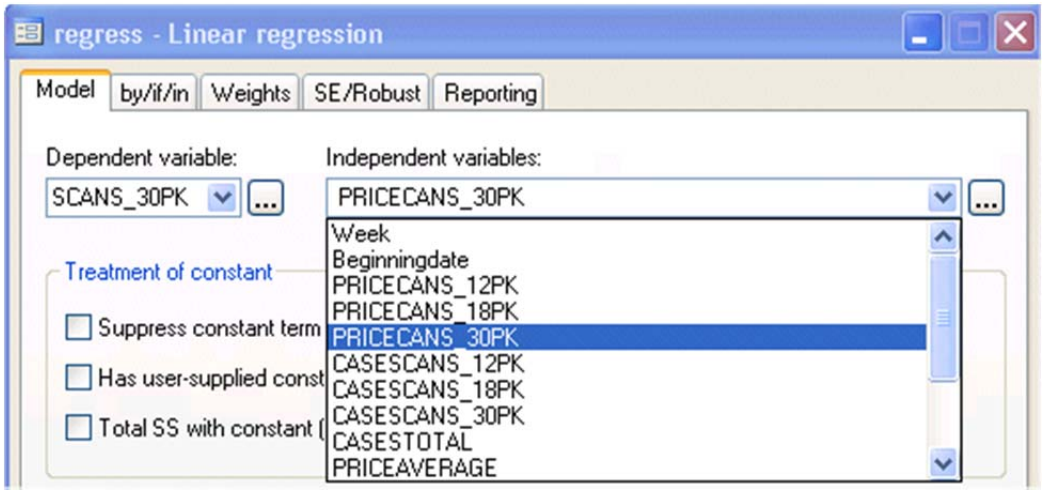


Here you can try to judge how the peaks and valleys in the two series line up:



The multiple regression procedure:

Select the dependent and independent variables from pull-down lists (best to use short variable names if you want to see the full name of the dependent variable here):



Here is the regression output that you get by default:

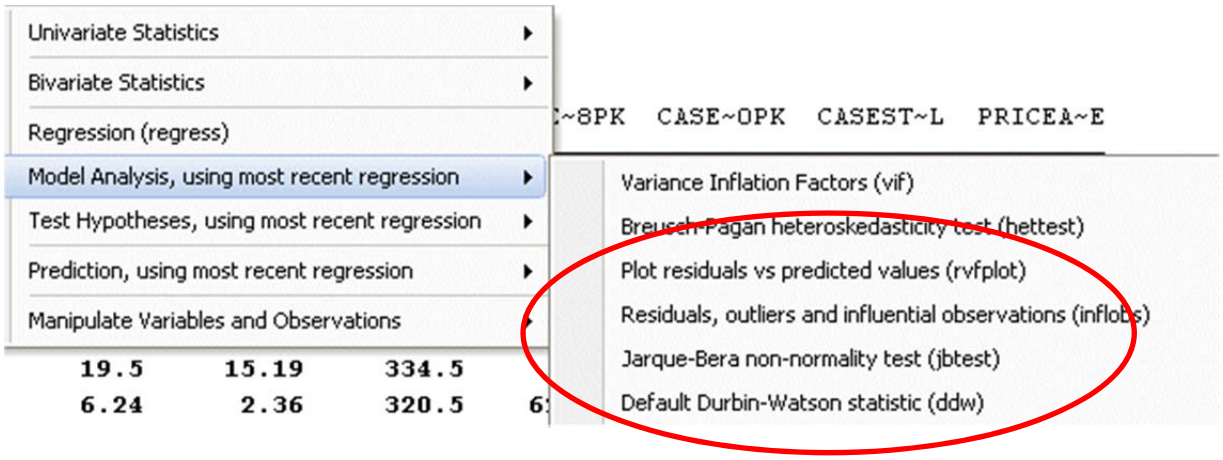
```
. regress CASESCANS_30PK PRICECANS_30PK
```

Source	SS	df	MS
Model	1250056.55	1	1250056.55
Residual	394543.541	50	7890.87082
Total	1644600.09	51	32247.0606

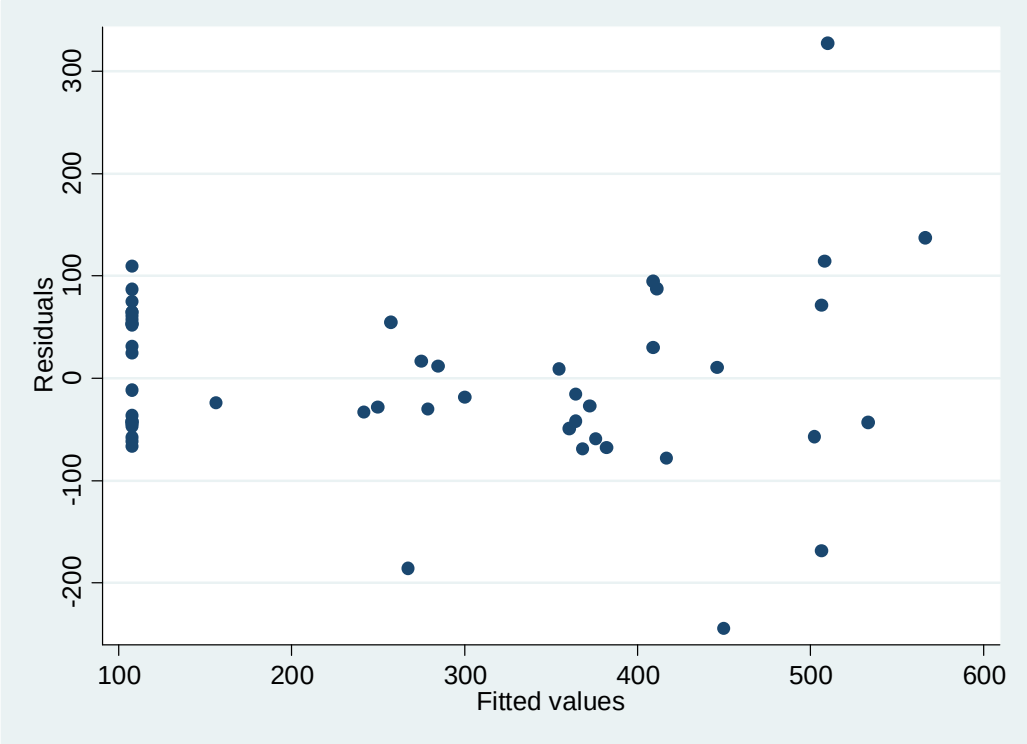
```
Number of obs =      52
F(  1,     50) =  158.42
Prob > F      =  0.0000
R-squared     =  0.7601
Adj R-squared =  0.7553
Root MSE      =  88.831
```

CASESCANS_30PK	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
PRICECANS_30PK	-194.2927	15.43669	-12.59	0.000	-225.2983	-163.2872
_cons	3059.237	222.3093	13.76	0.000	2612.716	3505.758

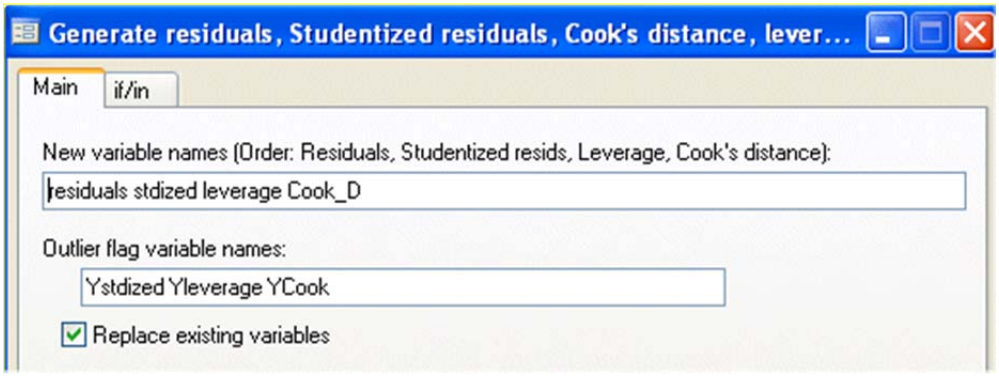
You need to run some separate procedures to get additional model stats and charts:

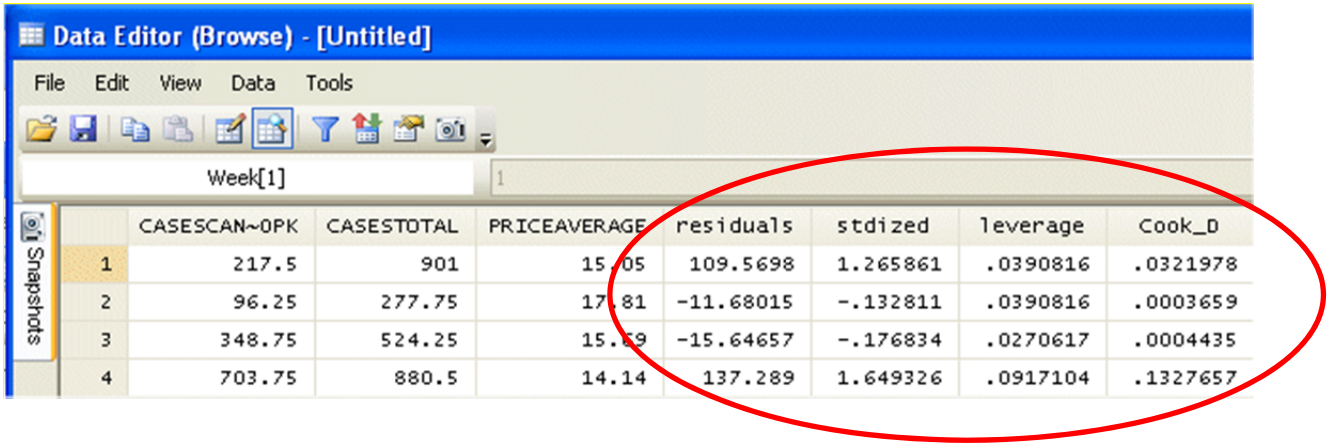


The residual vs. predicted plot is the only residual plot that is included on this menu. Other types of residual plots can be generated from the "Graphics" menu. (More about this later.)



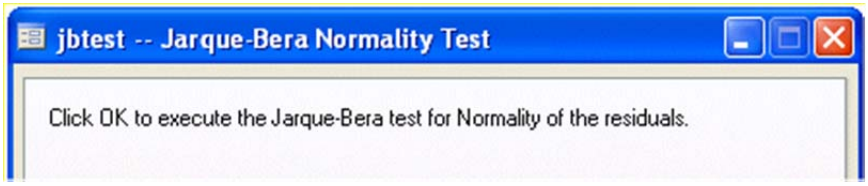
The "Residuals, outliers, and influential observations" command stores the residuals and standardized residuals and a few other stats (leverage, Cook's D) on the data worksheet. Before clicking through its dialog box you might want to edit the new variable names if you are saving the results of different models in the same file—e.g., "Model_1_residuals".





	CASESCAN~OPK	CASESTOTAL	PRICEAVERAGE	residuals	stdized	leverage	Cook_D
1	217.5	901	15.05	109.5698	1.265861	.0390816	.0321978
2	96.25	277.75	17.81	-11.68015	-.132811	.0390816	.0003659
3	348.75	524.25	15.69	-15.64657	-.176834	.0270617	.0004435
4	703.75	880.5	14.14	137.289	1.649326	.0917104	.1327657

The Jarque-Bera normality test is a test that is based only on the skewness and kurtosis coefficients of the residuals, unlike the Anderson-Darling or Kolmogorov-Smirnoff tests which are based on the entire cumulative distribution.

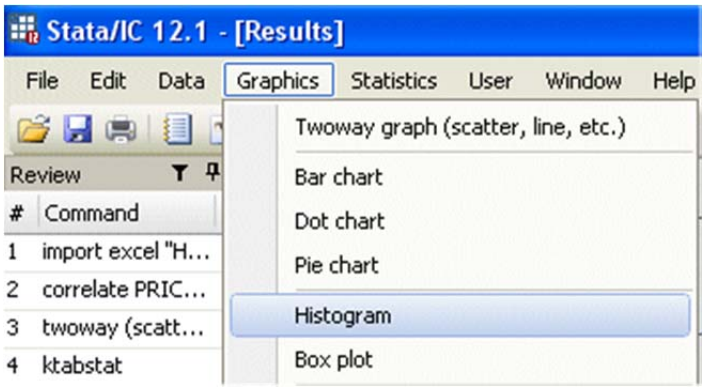


The result of this test is written back to the output window and looks like this:

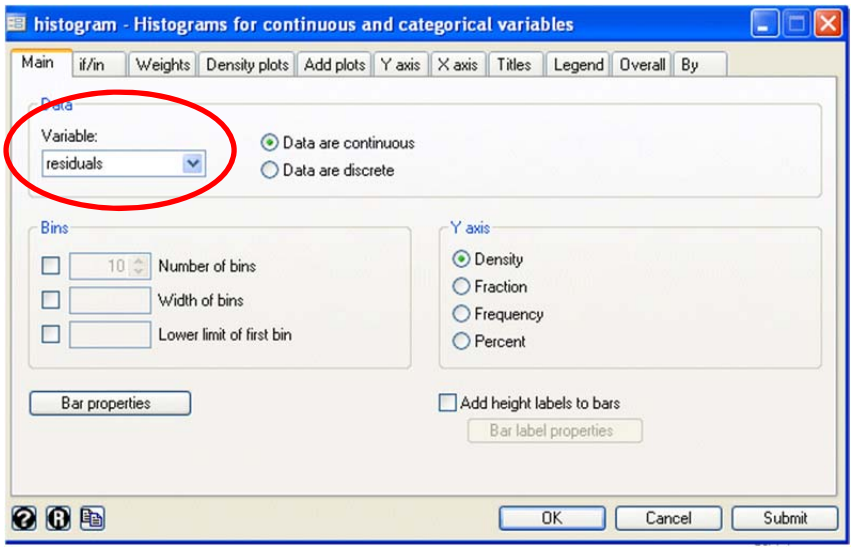
```
. jbtest
Jarque-Bera normality test:  22.87 Chi(2)  1.1e-05 p-value
Jarque-Bera test for Ho: normality of residuals
```

In this case the p-value of 1.1e-05 indicates that the normality hypothesis is strongly rejected.

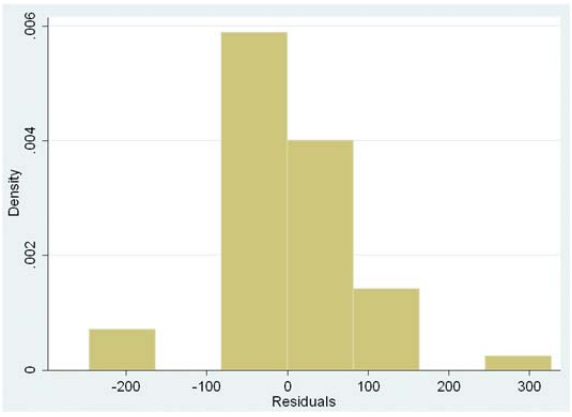
For a closer look at the error distribution, a residual histogram chart can be generated by going to the main graphics menu and using the histogram procedure with the newly-created "residuals" variable as the input:



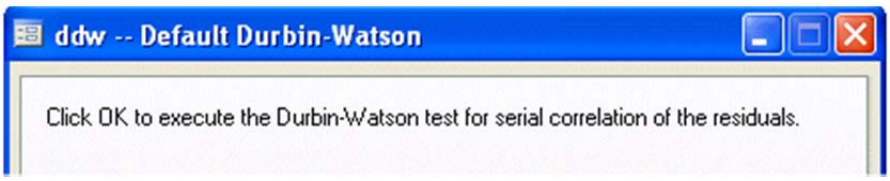
Here is the dialog box for the histogram procedure:



The default bin specifications are pretty coarse—here is what you get in this case. You might want to fiddle with the bin settings in order to show more fine detail.



The Durbin-Watson statistic (which requires executing another separate menu command in order to be reported) is a test for autocorrelation at lag 1 in the residuals. Click through this dialog box:



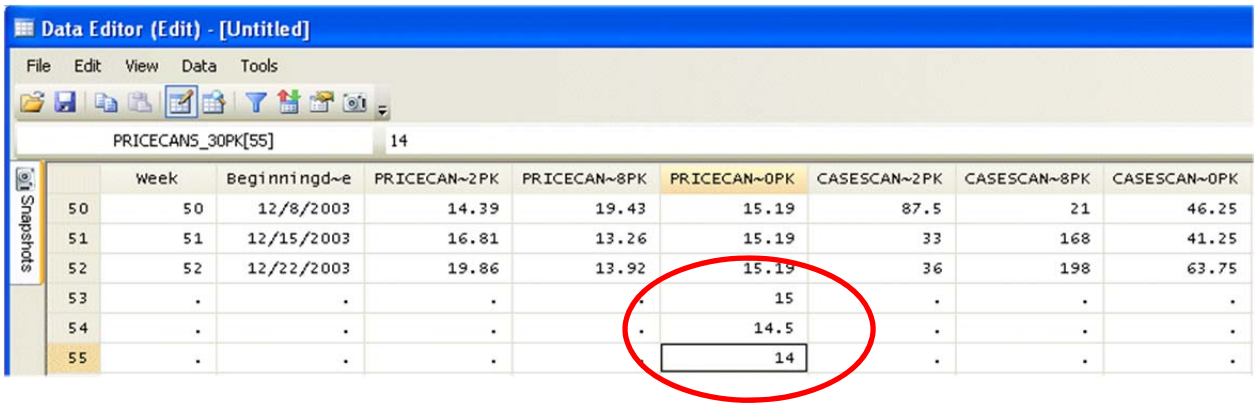
The resulting report of the DW stat looks like this. The user needs to know whether a value of 1.4 is significant, because no p-value is reported for it.

```
. ddw
        time variable:   __000000, 1 to 52
                delta:   1 unit

Durbin-Watson d-statistic(  2,     52) =  1.402001
```

In general the DW statistic is approximately equal to $2(1-r_1)$ where r_1 is the lag 1 residual autocorrelation. Its range is from 0 to 4 and it approaches 2 when the lag 1 autocorrelation approaches 0. The lag-1 residual autocorrelation for this model is 0.281, which is the test statistic that is shown in FSBForecast output. FSBForecast also reports the residual autocorrelations for lags 2 through 7 and lag 12, along with their 95% significance limits. The 95% significance limit for testing the lag-k autocorrelation is 2/SQRT(n-k), where n is the sample size, which works out to be 0.280 for lag 1 in this model, so this is a borderline-significant value.

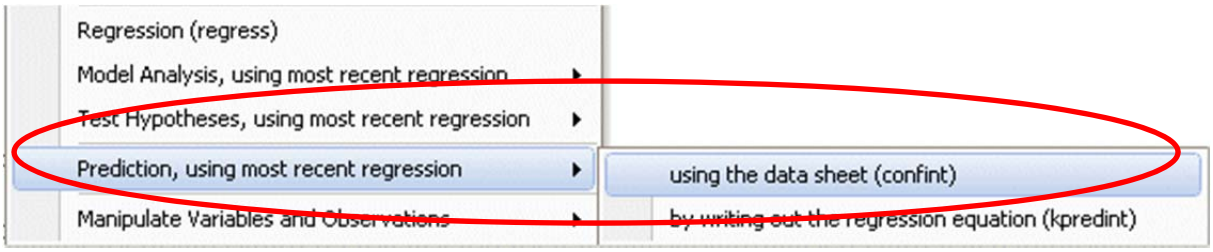
Forecasting from a regression model:

Stata generates forecasts in a manner similar to FSBForecast. If values for the independent variable(s) are typed or already exist in additional rows at the bottom of the data set, the "Prediction, using most recent regression, using the data sheet" command will cause the corresponding forecasts to be computed. Here some additional price values were typed in at the bottom of the data worksheet:

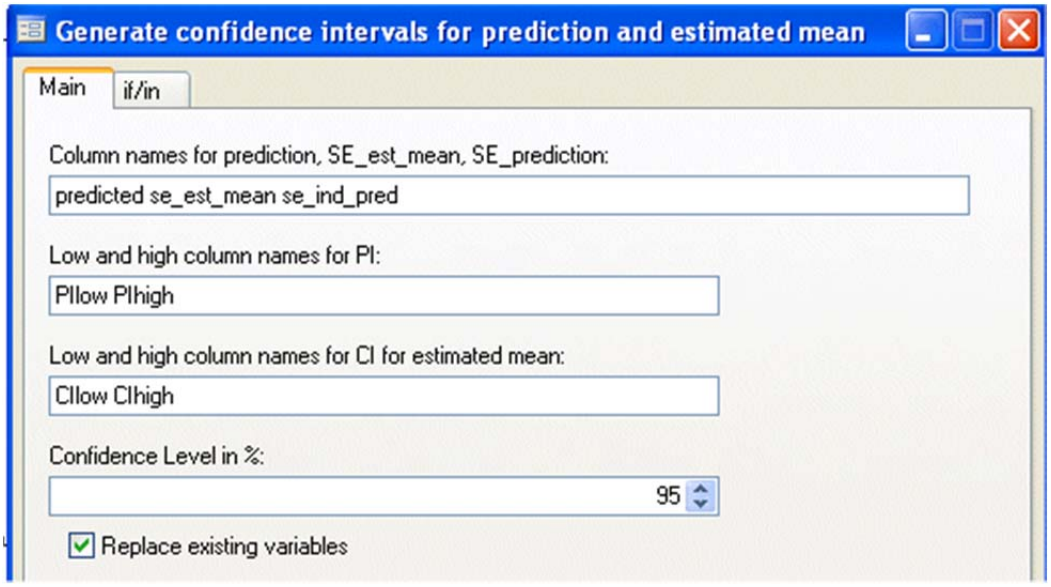


	Week	Beginningd~e	PRICECAN~2PK	PRICECAN~8PK	PRICECAN~0PK	CASESCAN~2PK	CASESCAN~8PK	CASESCAN~0PK
50	50	12/8/2003	14.39	19.43	15.19	87.5	21	46.25
51	51	12/15/2003	16.81	13.26	15.19	33	168	41.25
52	52	12/22/2003	19.86	13.92	15.19	36	198	63.75
53	.	.	.	.	15	.	.	.
54	.	.	.	.	14.5	.	.	.
55	.	.	.	.	14	.	.	.

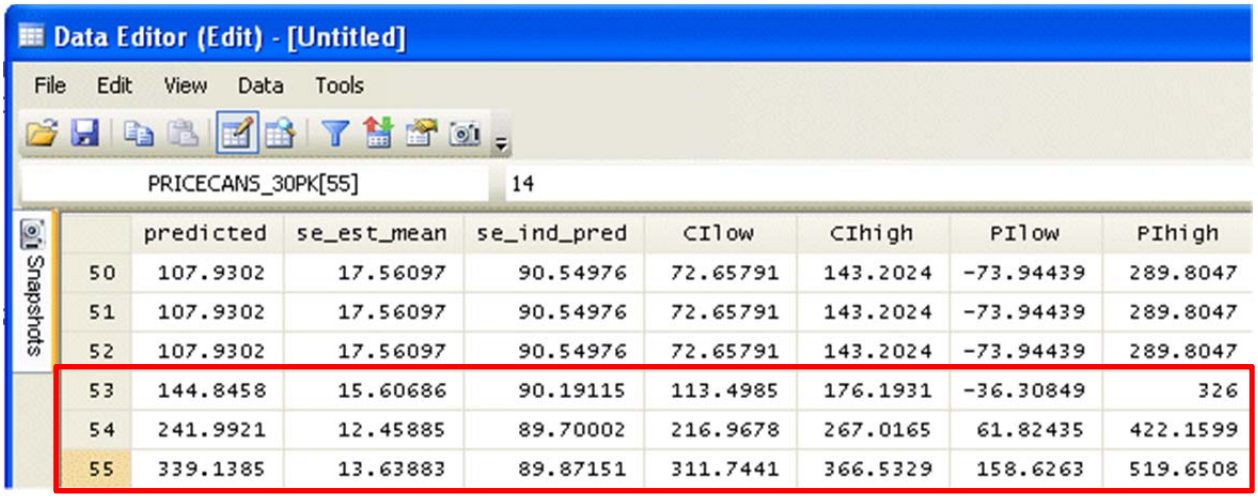
The prediction-using-data sheet option was chosen next:



You then have to click through this box in which you can edit the names of the forecast statistics that that will be shown . Here too you might want to edit the names of the variables to be created, if you are fitting more than 1 model using the same file:

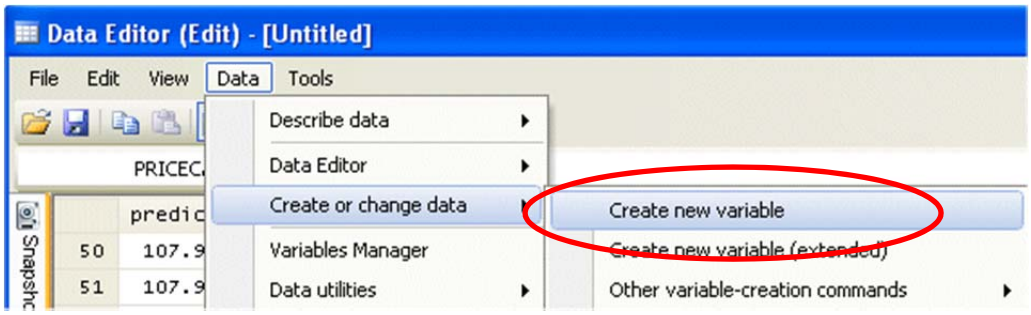


The forecasts and their standard errors and confidence limits are written back to the data spreadsheet, but not plotted:

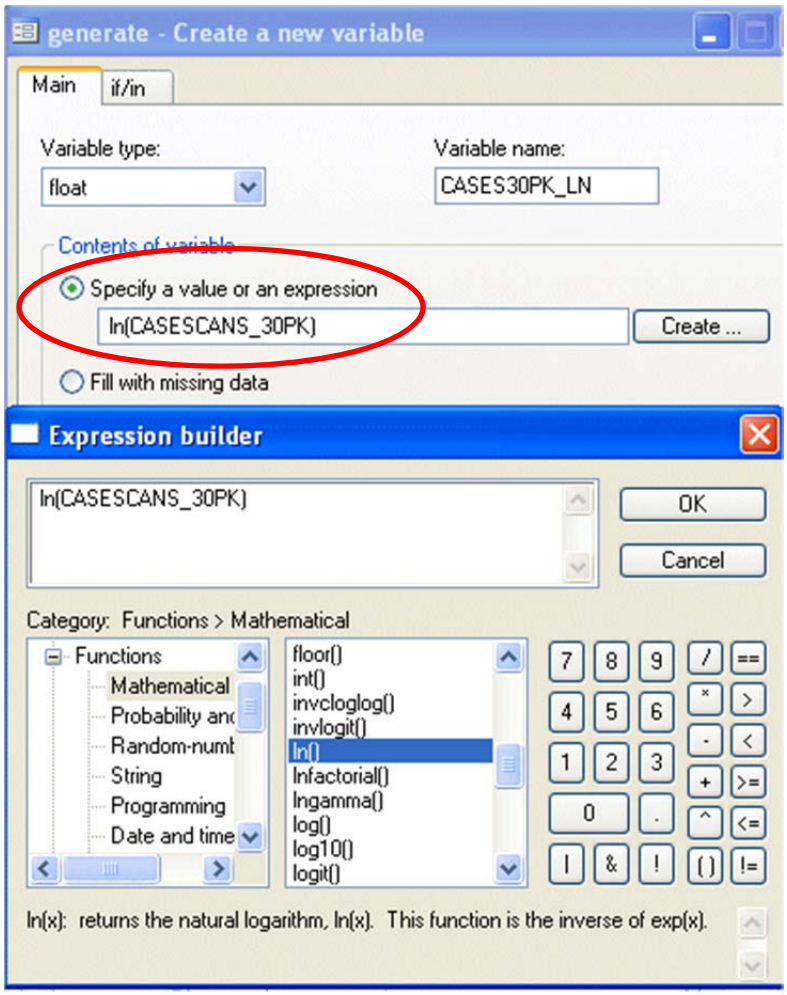


	predicted	se_est_mean	se_ind_pred	CIlow	CIhigh	PIlow	PIhigh
50	107.9302	17.56097	90.54976	72.65791	143.2024	-73.94439	289.8047
51	107.9302	17.56097	90.54976	72.65791	143.2024	-73.94439	289.8047
52	107.9302	17.56097	90.54976	72.65791	143.2024	-73.94439	289.8047
53	144.8458	15.60686	90.19115	113.4985	176.1931	-36.30849	326
54	241.9921	12.45885	89.70002	216.9678	267.0165	61.82435	422.1599
55	339.1385	13.63883	89.87151	311.7441	366.5329	158.6263	519.6508

Mathematical transformations can be applied with the "create new variable" option on the Data menu:

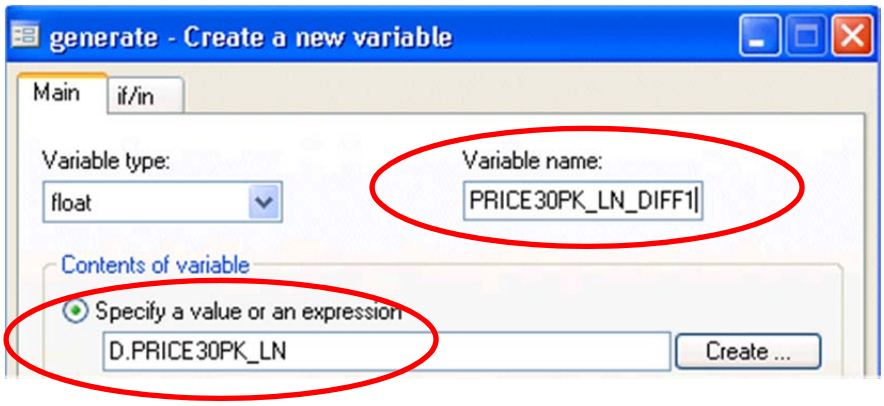


For example, here you can apply the natural log transformation. You need to type a name for the new variable and then you need to type the formula to compute it. There is also an "expression builder" dialog box that you can bring up by hitting the "create" button. It shows the list of available transformations and can type their names for you if you click on them.



However, there are no time transformations on the function list: no lag or difference or difference-of-natural-log or percentage-difference relative to previous observations.

You can get a difference-of-natural-log-from-one-period-ago transformation by going back to the top level of the create-new variable procedure and then applying the difference operator, whose syntax is "D.", to the logged variable. Here again you must type a name for the new variable as well as the mathematical expression that creates it.



The code that was executed by the create-variable procedure in the process of applying the log and difference transformations is shown below, along with the results of fitting a regression model to the logged-and-differenced variables. This is the same as Model 3 that was fitted to the same data set with FSBForecast.

```
. generate CASES30PK_LN = ln(CASESCANS_30PK)

. generate PRICE30PK_LN = ln(PRICECANS_30PK)

. generate CASES30PK_LN_DIFF1 = D.CASES30PK_LN
(1 missing value generated)

. generate PRICE30PK_LN_DIFF1 = D.PRICE30PK_LN
(1 missing value generated)

. regress CASES30PK_LN_DIFF1 PRICE30PK_LN_DIFF1
```

Source	SS	df	MS
Model	16.6001561	1	16.6001561
Residual	4.3365367	49	.088500749
Total	20.9366928	50	.418733856

Number of obs	51
F(1, 49)	187.57
Prob > F	0.0000
R-squared	0.7929
Adj R-squared	0.7886
Root MSE	.29749

CASES30PK_LN_DIFF1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
PRICE30PK_LN_DIFF1	-8.874448	.6479757	-13.70	0.000	-10.1766	-7.572292
_cons	-.0240633	.041657	-0.58	0.566	-.1077763	.0596496

From here you can go on and construct the rest of the regression output one piece at a time as shown earlier.