Time Series: An Example and a Re-Introduction to Stata[1]

Because ECO 250 and ECO 255 now both use Stata, we will not be doing a full fledged introduction to Stata. These notes instead gather some time series, discuss some of the issues involved with collecting them, and then demonstrate some time series parts of Stata.

As an illustrative example, I will focus on the following question following question: Most U.S. recessions are preceded by a noticeable decline in housing starts. The housing sector along with other variables such as output, inflation, etc. is part of a jointly determined economic system. To untangle the causal relationships among these variables, I collect data on the variables included in such a system.

Among the many advantages of being a macroeconomist (all macroeconomists are among the world's kindest, bravest, warmest, most wonderful human beings you'll ever know in your life.), macroeconomic data is usually pretty easy to get. Motivated by the previous example, I collected the following time series from the St. Louis Fed's FRED database:

1. Total: New Privately Owned Housing Units Started. This time-series is available monthly and I gather it beginning in April 1974. Housing starts are of interest because they catch an important component of GDP (residential investment) in its early stages and because they known asga good leading economic indicator. I could have chosen housing permits instead. The two variables are highly correlated, however, so it doesn't really matter which I pick. Including both would introduce multicolinearity, blowing up by error bands.

2. The Consumer Price Index for All Urban Consumers. There are a few issues here. First, because food and energy prices are considered largely exogenous, many studies use a CPI that excludes food and energy. Other studies use other price indices, such as a Producer Price Index. Ultimately, it comes down to which time series best captures what your specific project is trying to measure. I know that the CPI will be non stationary. So, I will use the annualized rate of change, which is just a measure of inflation.

3. U-3 unemployment. This will be my measure of the business cycle. I pick it instead of GDP because that is only available on a quarterly basis. I could also have used industrial production, a variable that is correlated with GDP but which is available monthly.

[1]These are undergraduate lecture notes. They do not represent academic work. Expect typos, sloppy formatting, and occasional (possibly stupefying) errors.

4. The one year Treasury Bond yield. 15 years ago, we thought we knew how to best measure the current state of monetary policy: just use the Federal Funds rate or its target. But the latter did not change significantly between 2009 and 2015 despite the many dramatic policies the Federal Reserve has initiated. Many macroeconomists, however, believe that the 2 year Treasury Rate is the best measure of the state of monetary policy because it capture a time frame over which more agents actually borrow than an overnight rate. But there are shortcomings as well. This rate is affected by many things besides monetary policy so it is problematic to interpret a change in it as a change in monetary policy. I choose to go with the 1-year Treasury Bond yield mostly because the 2-year is missing for a while on FRED.

5. The 30-year average fixed mortgage rate. I expect that this variable is important for the housing sector.

6. U.S. Population. Housing starts should probably be adjusted for population growth. To do so, I include this series. I include it as a separate file so I can demonstrate the merge command.

Stata and Data Organization

I now have two data files. Ex1.dta is the file containing most of my variables. Pop.dta includes the population variable. I also have Ex1.do, my do file although in class I will enter the commands one at a time.

Merging files in Stata is easy. Here, I need a 1-1 merge, the simplest type because each data file includes the same time periods:

merge 1:1 month using $C:\backslash\backslash Users\backslash pshea\backslash Documents\backslash Pop.dta$, keep(match);

An important part of working with time series in Stata is that you need to designate which variable designates time. I do this with the following command:

tsset month;

Forgetting this step will cause the time series commands that we use not to work.

Notice that my data includes two time variables, *date* and *month*, that also measures the time period. Having had too many problems with Stata failing to detect the format of the time variable, I have started creating a new variable that just starts at 1. Perhaps you will have better luck.

The *gen* command is used to create new variables. We typically work with the natural logs of continuous variables that are not already measured as a percentage.[2] Doing so allows us to eventually interpret regression coefficients in terms of a percentage change instead of measuring the change in units. It is easy to take the log of a variable in Stata.

gen log_hs=ln(hs)

This step is not, however, always necessary. I can, for example, just include the log directly in a regression. Below is the world's dumbest regression:

reg hs ln(hs) *log_hs*, r;

Here I regress housing starts on its own log. Note that Stata just throws out one of the logged GDP terms that I included as regressors. Because they are the same variable, I have perfect multicolinearity. In my day, the program would have crashed. Kids today are spoiled. Recall that the r option chooses robust standard errors.

We will often work with lags. Here I create the first and second lags of GDP:

gen $lag1\_hs = l.hs$;

gen $lag2\_hs = l2.hs$;

The cool kids use a shortcut which I show in the following third order autoregression. So I am really just using this as an excuse to remind you of the *gen* command.

reg hs l.hs l2.hs l3.hs, r;

Most empirical papers report descriptive statistics, which can easily be done with the *summarize* command:

summarize hs my30 ty1 ue pi_cpi pop, detail;

Table 1: Mean and Variance

	hs	my30	ty1	ue	pi_cpi
Mean	1458.536	8.490979	5.62023	6.442418	4.119578
Std. Dev.	443.8869	3.052795	3.573935	1.552301	3.282934
Min	478	3.35	.1	3.8	-2.9
Max	2494	18.45	16.72	10.8	18.4

Reporting descriptive statistics may seem like it is just pro-forma. But they are important. Examining them is a way to see if our data make sense. Suppose, for example, that we see a minimum value of -99. This wouldn't make sense and would probably indicate that some missing observations were coded in this way.

It is also interesting to examine how the variables are correlated with each other. This may be reported in a correlation matrix:

correlate hs my30 ty1 ue pi_cpi pop l.hs;

by including $l.hs$, I am also reporting logged housing start's first order autocorrelation.

hs	1.0000						
$my30$	0.1482	1.0000					
$ty1$	0.2522	0.9407	1.0000				
ue	-0.5105	0.2044	-0.0029	1.0000			
pi_cpi	0.0694	0.5839	0.6701	0.0681	1.0000		
pop	-0.4508	-0.6856	-0.7351	-0.0455	-0.5953	1.0000	
$hs.l1$	0.9669	0.1564	0.2717	-0.5375	0.0974	-0.4510	1.0000

That the diagonals equal one says that each variable is perfectly correlated with itself, which is tautological. Housing starts and unemployment are strongly negatively correlated. Do not take this as evidence of a causal relationship.

Note that housing starts are positively and highly autocorrelated. This is unsurprising.

Some Incompetent Econometrics

We now take a first pass at testing whether housing starts affect unemployment. we shall fail miserably. We start by regressing unemployment on the other variables :

[2]There are exceptions. This would not work for a variable that is often negative.

reg ue log_hs my30 pi_cpi, r;

which yields the following result:

$$ue = \underset{(0.970)}{25.11} + \underset{(0.135)}{-2.80}\, log\_hs + \underset{(0.0246)}{0.207}\, my30 + \underset{(.020)}{-0.050}\, pi\_cpi + u_t \quad (1)$$

All the regression coefficients are statistically significant at over or close to the 99% confidence level. So might we conclude that more housing starts lowers unemployment while higher mortgage rates increases unemployment? No. Our regression is so horrendously misspecified that it would earn a grade of F.[3] Partly to review and partly to preview, we con consider some of these sources of misspecification:

1. We are ignoring the temporal nature of the data generating process. Think back to ECO 103. Shocks to monetary policy, or other variables, often affect the business cycle (measured here using unemployment) with a lag. Perhaps housing starts affect the business cycle 12 months out, but not immediately. The system is almost surely dynamic, that is the state of the world in period t tells us something about $t+1$. Because we rely on no time series methods, we are not even trying to capture the relationships that exist among variables in different time periods. We will soon turn our attention to developing these time series techniques.

2. Some of our variables may be non-stationary. Recall that a stationary variable tends to revert back to a time independent mean. Because we have not yet adjusted for population, housing starts are likely non-stationary. Interest rates have been trending downwards for decades, thus the mortgage rate is also a candidate. When we include non-stationary variables in an OLS regression, the Gauss-Markov conditions do not hold. Our results are biased. Here the bias is horrendous. The course's next topic is how to test for and deal with non-stationarity. It is not obvious whether the 2 year Treasury yield is also non-stationary.

3. Endogeneity. Do housing starts cause unemployment.? Does unemployment cause housing starts? Does something else cause both of them? Macroeconomic theory strongly suggests that all of the variables in this system are jointly and simultaneously determined. Simultaneity is a form of endogeneity that also is a violation of Gauss-Markov. This is yet another source, also probably serious, of bias.

4. Omitted Variable Bias. All econometric models are misspecified and omitted variable bias is always present. Our goal is thus to minimize it. In this case, I worry that because we have

[3]But at least the standard errors are robust.

not yet included short-term interest rates to capture monetary policy, omitted variable bias is problematic. So to get my grade up to an $F+$, I throw those in:

$$ue = \underset{(1.05)}{17.67} + \underset{(0.136)}{-2.03}\, log\_hs + \underset{(0.0493)}{0.768}\, my30 + \underset{(.020)}{0.070}\, pi\_cpi + \underset{(.042)}{-0.596}\, ty1 + u_t \qquad (2)$$

Equation (2) finds a negative coefficient on one-year Treasury yield, statistically significant at 99% confidence level. Omitted variable bais was an issue in (1).

Suppose we are worried that our earlier observations are inappropriate for some reason. It is easy to throw them out.

reg ue log_hs my30 ty1 pi_cpi if $month \geq 72$, r;

This was just an excuse to show you the *if* statement. I don't care enough about the results to include them.

Two final comments:

1. The R^2 for these specifications are 0.43-0.63. Recall that this implies that 53-63% of the variation of output is explained by deviations in the independent variables. R^2 is a useful metric. But we should take great care not to care too much about it. In time-series, too high of an R^2 can indicate a problem, such as including non-stationary variables. Also, if we include lags, R^2 will usually be quite high. But this is not always a good thing.

Also note that adding variables can only increase R^2. But overfitting is a serious mistake. The goal of a regression is not to maximize R^2 and we should take care not to judge our results by this metric.

2. Stata automatically does a t-test for each regressor. It is easy to use the *test* command to do an f-test for the joint significance of more than one variable. here is an example:

test log_hs my30;

Note that including just a single variable returns a t-test.