Estimating and Forecasting the Stock Market[1]

These notes introduce new concepts such as autoregressions, autoregressive distributed lag models, and simple forecasting. All of this material will be covered through the example of stock prices using daily (excluding weekends and other non-trading days) S&P 500 data.

The Data

I gathered two time series. The first is from *Yahoo Finance* and is the daily S&P 500 stock index price from 7/1/1954 through 9/25/2015. The second is the daily Effective Federal Funds rate, taken from the St. Louis fed, for the same period. The former series excludes weekends and holidays where the U.S. stock market is closed. I discard such observations.

The Stata do file for these notes shows how to easily merge them together.

Motivation

The significance of stock price forecastibility is hopefully quite clear. Were we able to get strong and correct results, we could potentially use these results to profit by trading on the market. But to know how to correctly specify our model, we need some theoretical background. The simple Dividend Asset Pricing model predicts that a stock price should equal:

$$S_t = \gamma_t + E_t \left[\frac{S_{t+1}}{1 + i_t} \right] \tag{1}$$

This simple model sets the price of a stock equal to its dividend (γ_t) plus the expected discounted present value of its future sales price. Taking logs of (1) yields:

$$s_t = \gamma_t + E_t[s_{t+1}] - ln(i + i_t) \tag{2}$$

Note that for sufficiently small values of i_t, $ln(i + i_t) \approx i_t$.

$$s_t = \gamma_t + E_t[s_{t+1}] - i_t \tag{3}$$

We can also note that if agents are risk averse, they may also insist on an equity premium. We can add this to (4) in an admittedly ad-hoc way:

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

$$s_t = \gamma_t + E_t[s_{t+1}] - i_t + EP_t \tag{4}$$

Under rational expectations, $E_t[s_{t+1}] = s_{t+1} + v_t$ where v_t is mean-zero white noise. In words, fully rational agents do not make systematic errors, on average they are right. Their errors are also not autocorrelated otherwise their mistakes would not be unforcastable.

$$s_t = \gamma_t + s_{t+1} - i_t + EP_t + v_t \tag{5}$$

Finally, we can consolidate the equity premium, expectationial error, and dividend into a constant and a mean zero error term:

$$s_t = \alpha^{'} + s_{t+1} - i_t + u_t^{'} \tag{6}$$

Re-arranging and re-dating:

$$s_t = \alpha + s_{t-1} + i_{t-1} + u_t \tag{7}$$

For expectations to be rational, u_t must also be mean-zero white noise.

Differencing (6) yields:

$$\Delta s_t = \alpha + i_{t-1} + u_t \tag{8}$$

This simple model thus predicts that stock prices follow a random walk, which entails an AR(1) coefficient equal exactly to one. It also predicts that the first difference of stock prices is white noise plus a constant.

Dickey Fuller

The first step is to check both of our time-series for non-stationarity. We, as before, will use Augmented Dickey Fuller tests. We determine lag length using the $dfgls$ command in Stata and use the $dfuller$ command to test for a trend. I find the following:

1. The Federal Funds Rate is I(1) and has no significant trend.

2. The S&P 500 has an upward trend and is also I(1).

I thus de-trend stock prices and first difference both variables.

A Caution

By using daily data, we have far more observations than with our money-income causality example that relies on quarterly data. These extra observations allow us to exploit short term (less than monthly) variation and thus obtain more precise point estimates and standard errors. In other ways, however, these extra observations do not help us. One example of this is in testing for unit roots. Suppose that we have a stochastic process that is measured annually:

$$x_t = 0.95x_{t-1} + u_t \tag{9}$$

If we have enough observations, we will reject the null hypothesis of a unit root. With a smaller sample, however, we may fail to reject a unit root. Suppose that we are able to increase our sample size by obtaining daily data. Then the process for (9), may be rewritten:

$$x_t = 0.95^{\frac{1}{365}} x_{t-1} + u_t = 0.999859x_{t-1} + u_t \tag{10}$$

For x_{t+1} to equal 95% (in expectation) of x_t, then the daily AR(1) coefficient must be much closer to one. Because the daily AR(1) coefficient is closer to a unit root, it is harder (for the same number of observations) to reject the hypothesis of a unit root. It can be shown that the two effects cancel each other out.

In time series, it is thus better to expand your sample by expanding the time frame as opposed to expanding the sample by increasing the frequency of the data. Only the former will improve the accuracy of unit root tests.

Autoregressions

An autoregression is a regression that simply regresses a variable on lags of itself. Our first pass will impose structure on the specification by including only one lag of stock prices. An autoregression is run in Stata using the *var* command:

$$\Delta s_t = \overset{.0002638}{(.000081)} + \overset{.0263753}{(.018927)}\ \Delta s_{t-1} + u_t \tag{11}$$

The constant is statistically significant at the 99% confidence level. The p-value for the lag, however, is only 0.163. We thus cannot reject the prediction of the Dividend Pricing Model that the first difference of stock prices is white noise.

We can also estimate a second-order autoregression by adding a second lag:

$$\Delta s_t = \underset{(.000082)}{0.000274} + \underset{(.0204)}{-.0415}\ \Delta s_{t-2} + \underset{(.0187)}{.0274}\ \Delta s_{t-1} + u_t \tag{12}$$

Surprisingly, the coefficient on the second lag is significant at the 95% confidence level. This autoregression thus suggests a small oscillatory dynamic. A 1% increase in stock prices today is expected to cause a 0.027% increae in prices tomorrow and a 0.042% decrease in two days.

This estimate does show a small amount of autocorrelation. The theoretical model predicts none. This may be taken as evidence against our theoretical model. A key question is whether any forecastability is large enough to overcome the transaction costs associated with exploiting it. Or, we may worry that our econometric specification is flawed. This may result from omitted variable bias (such as omitting the interest rate), or from not including more lags. Here, I added a third and then fourth lag, but found they did not do much.

Forecasting

Autoregressions may be used to forecast out of sample; future stock prices in this example. Calculating the point estimate for a forecast is easy, we need only re-date (12) forward. For one period, we obtain:

$$E_t[\Delta s_{t+1}] = 0.000274 - 0.0415\Delta s_{t-1} + .0274\Delta s_t \tag{13}$$

where t is the last period in the sample and $t+1$ is the first out of sample period. We could also choose to use a part in our regression of our data and form a forecast over the remainder. This would allow us to compare our forecast with the actual data.

Re-dating two periods ahead then yields:

$$E_t[\Delta s_{t+2}] = 0.000274 - 0.0415\Delta s_t + .0274E_t[\Delta s_{t+1}] \tag{14}$$

This process can then be repeated as far into the future as we wish. Note that because the AR(1) coefficient is small, these forecasts quickly converge to the neighborhood of the constant (.000274).

Forecasting in Stata is done using the $fcast$ command.

It is harder to form confidence intervals on our forecasts. Sometimes, we can calculate these analytically. Other times, we entail a process known as *bootstrapping*. It works as follows:

1. Use a random number generator to take a draw of the AR(1) coefficient from (12) using the point estimate as the mean and the estimated standard error as the standard deviation of the distribution.

2. Use a random number generator to obtain draws of u_t going as far into the future as we wish our forecast to go. We use zero as the mean and the estimated variance of the error terms as the distribution's variance. We further assume that the errors are normally distributed.

3. We repeat N times where N is large enough for the law of large numbers to take effect. Note that it actually takes Stata a minute to complete this process.

4. For any period $t+k$, the 95% confidence interval then consists of the middle 95% of simulated stick prices in period $t + k$.

Because our estimated AR(1) coefficient is small, our forecasts are not very exciting. Because stock prices declined on the last day of our sample, we expect a small decline in stock prices on the first out of sample day. Quickly, however, the forecast converges to the intercept which predicts a small but important daily increase.

Picking Lag Length for the Autoregression

We now allow the data to inform our choice of lag length for the autoregression. The approach is similar to choosing lag length for Dickey Fuller tests, we do it for different lag lengths and then choose using an information criteria. The Stata command *varsoc* does this for us.

```
varsoc d.lclose2;
```

All of the information criteria suggest an optimal lag length of 2.

Autoregressive Distributed Lag Models

So far, we have ignored interest rates even though (6) suggest that they are important. We may add them by expanding our autoregression into an autoregressive distributed lag model. This specification simply adds lags of other variables as regressors. Doing so yields:

$$\Delta s_t = \underset{(.000082)}{0.000274} + \underset{(.0204)}{-.0413}\ \Delta s_{t-2} + \underset{(.0187)}{.0276}\ \Delta s_{t-1} - \underset{(.000201)}{.000403}\ \Delta i_{t-1} + u_t \tag{15}$$

The coefficient on interest rates is negative at the 99% confidence level. This is consistent with the theoretical model's prediction of a negative coefficient.

We can test between our AR(1) and ADL model using information criteria. The Stata command *estatic* reports these for a recently run regression. Employing this command for both specifications yields ambiguous results.

Forecasting with an ADL model is harder than with a simple autoregression. The $fcast$ command in Stata will not work. To forecast, we would need to impose some process for how expectations of i_t are formed.[2] This need not be insurmountable. We may, for example, be willing to impose that the Federal Funds Rate will stay near zero in the short term. We could then re-write (15) as:

$$\Delta s_t \approx \overset{0.000274}{(.000082)} + \overset{-.0413}{(.0204)}\ \Delta s_{t-2} + \overset{.0276}{(.0187)}\ \Delta s_{t-1} + u_t \tag{16}$$

which simply drops the interest rate term which is close to zero. We could then forecast as we did with the AR(2). This forecast would only be valid, however, over the time frame in which we expect the Fed to keep interest rates near zero.

Conclusions

We now turn our attention to dealing with endogeneity. The ADL model is a good start as long as lags of stock prices and lagged interest rates are uncorrelated with the error term. This seems plausible (although remember that t dated variables can cause $t-1$ variables through expectations). But there are two issues. First, if interest rates and stock prices are jointly determined through a common data generating process, then it is more efficient to estimate them jointly. Second, ADLs generally do not allow for forecasting. We next turn our attention to system estimation through vector autoregressions (VARS).[3]

[2]Because we used the lag of interest rates and we do have i_t, we can easily make a 1 period ahead forecast.
[3]We have already used VAR coding. By including only one variable, a VAR reduces to an autoregression.