A HIERARCHICAL APPROACH TO DAILY STOCK MARKET FORECASTING

Heydar Pourian, University of Missouri at Saint Louis

SAS is the best software currently available in offering a "complete" package of stock market forecasting on a day-to-day basis. Due to the richness of its DATA, OUTPUT and a variety of PROCEDURES, SAS is capable of providing a workable forecasting tool both for the individual investor and the brokerage firm.

Such a complete package of stock market forecasting can be secured with a simple SAS application of the "Analytical Hierarchy Process" and pairwise comparison (see Saaty, 1980). The hierarchical approach to forecasting the daily path of stock prices can be generally accomplished by all or some of the following steps.

I. Variable Identification
   1. DATA Step, including PROC CITIBASE/
      Xll, a priori Transformations, and
      ARRAYS
   2. Preliminary PROC PLOT/TIMEPLOT
   3. PROC CORR/RSQUARE/RSREG/TTEST

II. Estimation and Forecast
   1. Adjusted and Transformed DATA Step
   2. Adjusted (for forecasting) PROC REG/
      STEPWISE/RIGGE REG/GLM
   3. "OUTPUT Step"
   4. PROC FORECAST/AUTOREG/ARIMA/
      STATESPACE/SYSREG/SIMLIN
   5. Combinations of II.2 and II.4: for
      instance, "ARIMA+REG"
   6. Usage of MACROS for processing a
      variety of variables, dates, models,
      and methods

III. Comparison, Evaluation and Decision
   1. PROC GPLOT
   2. PROC SUMMARY/MEANS
   3. PROC CALENDAR/COMPUTAB

This paper exhibits some of the above steps used in generating daily forecasts of a stock-price measure. The results of the paper are based on the ex ante forecasts of a "representative" measure of stock prices which has a medium level of aggregation. Based on a "memory" of six to seven months, daily forecasts are generated by use of the latest available information. The measure chosen is the most-watched Dow Jones Average of 30 Industrial stock prices (DJI).

The "estimation" period is the six-month period from Monday, August 16, 1982 to Friday, February 11, 1983. Each model, therefore, has a memory of at least six months (initially, N=117).

The "forecast" period is the one-month period from Monday, February 14, 1983 to Monday, March 14, 1983. It should be noted, however, that each day data were updated and new estimates were obtained, upon which the forecasts are based. Thus, each daily forecast is based on a memory from August 16, 1982 to "yesterday." It can be seen that the estimation and forecast processes themselves followed an "iterative" process:

estimation, forecast, ..., update,
reestimation, forecast, ...

This process continued throughout the "forecast" period, utilizing the latest information available up to yesterday inclusive (N=146 for final estimates). In general, the Principle of Parsimony or Simplicity in model building is obeyed (cf. Box and Jenkins, 1970). It should also be noted that the paper is a report of a "case study" of generating forecasts through various procedures.

SAS provides a variety of econometric and time-series techniques for forecasting purposes. Perhaps the most accurate and most popular of all methods is the simple and broad category of ARIMA Method. The performances of three univariate ARIMA models are reported here. Moreover, the ARIMA procedure is used to forecast "daily" values of the Money Supply "M1" and the Monetary Base contemporaneous to the DJI which is being forecast. This latter procedure was used in Model 4. This method is, in effect, an "ARIMA+REG" procedure. The following table (next page) summarizes the methods reported here with an overview of the specification of the models.

II. ON METHODS AND PROGRAMMING

A. Consider the following general model, which as a representative model will be used to exhibit the processes involved in using various DATA-, OUTPUT-, and PROC-steps in order to generate day-to-day ex ante forecasts of the Dow-Jones Industrial Average (DJI). In the language of PROC REG Model Statement, Model 1 can be specified as:

\[
\text{UNO: MODEL DJI = DJI1 DJI2 DJI3 DJI4 DJI5 DJI6 DJI7 DJI8 DJI9 DJI10 DJI11 DJI12 DJI13 DJI14 DJI15 DJI16 DJI17 DJI18 DJI19 DJI20 DJI21 DJI22 DJI23 DJI24 DJI25 DJI26 DJI27 DJI28 DJI29 DJI30 DJI31 DJI32 DJI33 DJI34 DJI35 DJI36 DJI37 DJI38 DJI39 DJI40 DJI41 DJI42 DJI43 DJI44 DJI45 DJI46 DJI47 DJI48 DJI49 DJI50 DJI51 DJI52 DJI53;}
\]
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Method</th>
<th>Variables in the Model</th>
<th>Mean Daily Forecast Error</th>
<th>Mean Daily Absolute Forecast Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>Classical Regression (OLS)</td>
<td>Latest 3 Days of INDUSTRIAL, TRANSPORTATION, UTILITIES, &amp; SIXTY-FIVE Stocks</td>
<td>2.28</td>
<td>8.98</td>
</tr>
<tr>
<td>Model 2</td>
<td>Ridge Regression (k = 0.17)</td>
<td>ditto</td>
<td>2.14</td>
<td>8.91</td>
</tr>
<tr>
<td>Model 3</td>
<td>Classical Regression</td>
<td>Money Supply, Short-term Interest &amp; Long Term Treasury Bond Price</td>
<td>2.07</td>
<td>9.97</td>
</tr>
<tr>
<td>Model 4</td>
<td>Classical Regression</td>
<td>INDUSTRIAL, M. BASE, M. Multiplier, &amp; Treasury BOND</td>
<td>2.30</td>
<td>8.45</td>
</tr>
<tr>
<td>Model 5</td>
<td>ARIMA (0, 1, 2)</td>
<td>INDUSTRIAL</td>
<td>-1.06</td>
<td>8.65</td>
</tr>
<tr>
<td>Model 6</td>
<td>ARIMA (0, 2, 1)</td>
<td>ditto</td>
<td>-0.19</td>
<td>8.98</td>
</tr>
<tr>
<td>Model 7</td>
<td>ARIMA (1, 0, 0)</td>
<td>ditto</td>
<td>-0.35</td>
<td>8.53</td>
</tr>
<tr>
<td>Model 8</td>
<td>Simultaneous Equation System</td>
<td>Endogenous</td>
<td>1.00</td>
<td>6.73</td>
</tr>
<tr>
<td></td>
<td>With Two-Stage Least Squares</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where DJT = Transportation index, DJU = Utilities index, and D65 = Composite Sixty-Five stocks. The suffix numbers designate the daily lags. For instances, DJI1 = LAG(DJI).

B. For Model 1 a procedure proposed by Durbin was employed in detecting serial correlation. In this model neither the Durbin-Watson "DW" test nor the Durbin "h" test is relevant. The former is not relevant, because of the existence of lagged dependent variable among the explanatory variables, while the second cannot be used due to large number of observations and large variance of the estimates of the lagged variable in our case study.

In terms of our model below, if the estimated coefficient of "RESID" is not statistically significant in model "RESUNO," then the existence of serial correlation in model "UNO" above is not a problem. Notice that the first observation is dropped.

PROC REG DATA=DOW NOPRINT;
UNO: MODEL DJI = DJI1 DJI2 DJI3 DJT1 DJT2 DJT3 DJU1 DJU2 D651 D652 D653;
OUTPUT OUT=UNO R=RESID;
DATA DURBIN;
MERGE DOW UNO; BY T;
IF T = 1;
RESID = LAG(RESID);
PROC REG DATA=DURBIN;

RESUNO: MODEL RESID = RESID1 DJI1 DJI2 DJT1 DJT2 DJT3 DJU1 DJU2 D651 D652 D653;

C. One "ridge" regression (Model 2) is used to "correct" for a high degree of multicollinearity present in Model 1 by introducing artificial orthogonalization in regression (see, e.g., Vinod, 1978). Although the existence of (imperfect) multicollinearity is not a problem in forecasting, a ridge regression was applied to test whether the forecast errors could be reduced. The performance of Model 2 is only slightly better than Model 1.

D. "Day-to-day" is defined here as one-day ahead, ex ante forecast, which is similar to specifying, e.g., in PROC ARIMA's Forecast statement:

PROC ARIMA;
... FORECAST LEAD=1 ...;

This day-to-day forecast must be, however, incorporated to some other PROC steps by some modifications. For instance, PROC REG does not generate forecasts on an ex ante basis. But with a minor modification, as shown below, it is capable to generate ex ante forecasts.

PROC REG DATA=DOW OUT=UNO NOPRINT;
UNO: MODEL DJI = DJI1 DJI2 DJI3 DJT1
DATA XUNOA; RENAME=(OJT1=Il DJT1=T1 OJU1=U1 DJU1=U1 OJT2=T2 DJT2=T2 OJU2=U2 DJU2=U2 OJT3=T3 DJT3=T3 OJU3=U3 DJU3=U3 INTERCEPT=A1);

SET XUNO;
DATA XUNOB; SET DO'; IF T= & LAST;
DATA XUNOC; MERGE XUNOA XUNOB;

REGI = A + DJT1*T1 + DJT2*T2 + DJT3*T3 + DJU1*U1 + DJU2*U2 + DJU3*U3 + D651*S1 + D652*S2 + D653*S3;

Where 

"&LAST" is the macro code which only allows the forecast statement to "see" the latest observations relevant to REGI equation.

E. Serial correlation detected in Model 3 was corrected through the Hildreth-Lu procedure through the following statements. Note that the model is a new one as specified in the "MUNO" Model statement. The new symbols used are:

G=Hildreth-Lu Rho, CMR=Call-Money Rates of Interest, TBDP=Treasury-Bond Price, Ml=Money Supply "M1."

Since this procedure employs two lags, notice the "DO" statements used for the first two observations. (The constants, however, are particular to our (DOW) data.)

G = 0.9745; G2 = G**2; GA = SQRT 1-G**2;
LOJI = LOG (OJI); LOJII = LAG (LOJI); LOJIH = LOJI - G*LOJI;
IF _ N=1 THEN DO;
LDJH = GA*(6.675);
END;
CMRI = LAG (CMRI); LCMRI = LOG (CMRI);
LTM1 = LAG (TBDP1); LTBDP1 = LOG (TBDP1);
LTM1H = LTM1 - G*LTM1;

F. The following is the simultaneous-equation model based on both stock and money market measures. It uses the 2SLS to simulate DJI.

The new variables are: CDR=Certificate of Deposit Rate of Interest, MISA=Money Supply "M1" Seasonally Adjusted.

G. There is a great deal of controversy surrounding the choice of the statistic which best evaluates a forecast process (cf., e.g., Granger and Newbold, 1977). For the stock market, however, probably the most relevant forecast statistic is the one which gets closest to the actual value, while avoids the "cover up" of averaging of underestimates and overestimates. Presumably the Mean Absolute Forecast Error appears to be the proper forecast statistic. Two difference forecast-evaluation measures are reported in the above table. Below, we also show the specification of three more evaluation criteria.

The following statements specify and measure five different errors of simulation or forecast-evaluation criteria:

(1) MEAN FORECAST ERROR: ME1
(2) MEAN ABSOLUTE FORECAST ERROR: MAE1
(3) PERCENT ROOT MEAN SQUARE FORECAST ERROR: PRMSI
(4) ROOT MEAN SQUARE FORECAST ERROR: RMSI
(5) THEIL'S INEQUALITY MEASURE U2: UREGI

PROC SUMMARY is then used to sum and generate a new data set "OUSUM." Finally, data "EVALUATE" provides the desired forecast/simulation errors. The variable "A" is the actual DJI, while as a representative "REGI" is the ex ante forecast generated by the "UNO" model above.

DATA RESULTS;
INPUT A REGI ... ;
EREGI = A - REGI; AEREGI = ABS (EREGI);
RMSI = SQRT (EREGI**2);
UREGI = SQRT (OOREGIS/OAS);
where "&FN" is the macro code for the number of forecast periods (FN=20 in our case).

Different models rank differently under alternative evaluation criteria. On a more deterministic basis, however, Model 4 slightly outperforms other models. The method employed is "ARIMA+REG" with the following model specification:

DJ1 = constant + Yesterday's DJI + Monetary Base as forecast by ARIMA(0,1,1) + Money Multiplier as forecast by ARIMA(0,1,1) + Yesterday's Treasury Bond Price.
** DAILY FORECASTS OF DOW JONES INDUSTRIAL **
* PROC REG AND PROC RIDGREG FORECASTS *

ACTUAL -- REG --- RIDGREG ------

---

1150
1140
1130
1120
1110
1100
1090
1080
1070
1060
1050

D J

14FEB83 19FEB83 24FEB83 01MAR83 06MAR83 11MAR83 16MAR83

EXCHANGE DAYS
** DAILY FORECASTS OF DOW JONES INDUSTRIAL **
* PROC ARIMA AND PROC ARIMA+REG FORECASTS *

ACTUAL --- ARIMA --- ARIMA+REG ---

EXCHANGE DAYS

14FEB83 19FEB83 24FEB83 01MAR83 06MAR83 11MAR83 16MAR83
forecast errors, can be used to choose between competing methods and models.

FOOTNOTES

1. This paper is based on various procedures employed in an extended theoretical and statistical work (Pourian, 1983) available from the author.
2. Similar "day-to-day" modification is required (as in II.1.b above) in order to generate ex ante forecasts.
3. In fact, among approximately 100 models attempted by the author, this "ARIMA+REG" method performs the best in terms of the mean absolute forecast error.

REFERENCES


The author is assistant professor of economics at the Department of Economics, University of Missouri, Saint Louis, Missouri 63121, USA, 314/553-5665. He is indebted to Larry Pickett for the first SAS lesson.