Staff Paper P84-22

August 1984

Staff Papers Series

SEQUENTIAL REGRESSION: A FLEXIBLE TOOL FOR TIME SERIES MODELING

by

Paul L. Fackler
Research Assistant

and

Robert P. King
Associate Professor

Department of Agricultural and Applied Economics

University of Minnesota
Institute of Agriculture, Forestry and Home Economics
St. Paul, Minnesota 55108

Staff Paper P84-22 August 1984

SEQUENTIAL REGRESSION: A FLEXIBLE TOOL FOR TIME SERIES MODELING*

by

Paul L. Fackler
Research Assistant

and

Robert P. King
Associate Professor

Department of Agricultural and Applied Economics
University of Minnesota

* Selected paper presented at the annual meeting of the American Agricultural Economics Association, Ithaca, New York, August 1984. This research was funded by the Minnesota Agricultural Experiment Station and the Graduate School of the University of Minnesota.

Staff Papers are published without formal review within the Department of Agicultural and Applied Economics.

The analysis of time series is of continuing interest to agricultural economists. The areas of price and market analysis, supply response, and policy evaluation all make considerable use of various econometric and statistical approaches to modeling with time series data. Furthermore, there is an ongoing effort to specify accurate forecasting models for policy and extension purposes, particularly as our profession has begun to wrestle with the issues raised when risk is incorporated into economic and decision models.

Unfortunately, many of the statistical and econometric methods economists typically encounter in their training are not suited to this task. Most developed out of a statistical theory that necessitates standards of experimental design that are impossible to maintain and are inapproriate in this study of time series. While there is a growing recognition of this problem in our field, the available literature on time series analysis requires a considerable investment of time and energy to even begin to master.

A further problem in this area is the lack of simple, usable software for developing and analyzing time series models, particularly software that is micro compatable and thereby usable in field extension offices for up-to-date modeling and analysis. Models that make use of nonlinear estimation procedures, in particular, suffer in this regard.

In this paper we show how sequential regression methods (Harvey) can be used to address these issues. In the sections which follow, we first outline an accurate, computationally efficient sequential regression algorithm. We then discuss the applicability of the sequential regression approach to standard econometric and ARIMA modeling. Next we present results from a modeling experiment involving price analysis. In the final section, we summarize the strengths and weaknesses of the approach we outline, with particular emphasis on its applicability in the areas of forecasting and risk analysis.

A Sketch of the Computational Algorithm

Sequential regression is a procedure by which the parameters of a model are recomputed iteratively as each new observation is added. We will only discuss least squares estimators, for which simple recursion algorithms are easily found. A number of these exist, but the one provided by Gentleman, which is simple, accurate and computationally efficient, deserves further attention.

Gentleman's algorithm is a modification of the Givens method of QR orthogonal decomposition. This decomposition may be developed as follows. Let X (TxK) be a matrix of regressors and y (Tx1) be a vector of regressands. Define

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}, \quad D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}, \quad R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}, \text{ and } \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

such that $Q'Q = I_T$, D is diagonal and R_1 is unit upper triangular. Q_{11}, D_1 and R_1 are each (KxK) and θ_1 is (Kx1). Furthermore, let $X = QD^{\frac{1}{2}}R$ and $y = QD^{\frac{1}{2}}\theta$. The least squares normal equations are given by

$$(1) \quad (X'X)^{-1}\beta = R'D^{\frac{1}{2}}Q'QD^{\frac{1}{2}}R\beta = R_1'\,D_1R_1\beta = R_1'\,D_1\theta = R'D^{\frac{1}{2}}Q'QD^{\frac{1}{2}}\theta = X'y.$$

With X full rank, R_1 and D_1 are nonsingular and we may reduce (1) to $R_1\beta = \theta_1$, which is easily solved by back substitution.

The modified Givens algorithm provides a way to implement this decomposition one observation at a time. At each iteration, the current values of x_t and y_t are incorporated into D_1, R_1 and θ_1. x_t may be discarded, while y_t will be transformed to a step ahead forecast error. While there is not room here to explore this algorithm in more depth, it is easily implemented on a microcomputer. Furthermore, Gentleman has shown that its numerical accuracy is comparable to the best known computational methods (see appendix).

Applications to Standard Regression Problems

Standard OLS estimates can be derived using Gentleman's algorithm. By performing a back substitution after each observation has been processed, a sequence of coefficient estimates is obtained. These estimates can be interpreted as the best (in the least squares sense) given the data available up to that point in time.

This sequence of estimates can be a useful tool in checking for model misspecification. In particular, given usual assumptions, the coefficient estimates should converge to stable values as the sample period grows. However, if structural change or trends are occurring that are not explicitly modeled, this will be reflected by sudden jumps and/or trends in these coefficients. In the former case, the coefficient sequence will be useful in pinpointing the dates when significant events occurred. These can then be analyzed more thoroughly.

In addition, forecasts of future dependent variable values can be generated at each period. Forecast errors can be derived from these as new observations become available. Indeed the step ahead prediction errors, defined as $v_t = y_t - x_t\beta_{t-1}$ where β_{t-1} is the vector of coefficient estimates derived last period, as well as the recursive residuals,

$r_t = v_t/(1 + x_t'(X_{t-1}'X_{t-1})^{-1}x_t)^{\frac{1}{2}}$, are essentially by-products of the computational algorithm.

As discussed in Harvey, these recursive residuals have some useful properties which derive from the fact that they are linear unbiased scalar (LUS) residuals. Theil has shown that given the classical OLS assumptions, a set of T-K residuals, u, is linear in y and unbiased with scalar covariance matrix if there exists a T x (T-K) matrix C such that $u = C'y$, $C'X = 0$ and $C'C = I$. It may be shown that $r = D_2^{\frac{1}{2}}\theta_2 = (Q'_{12}Q'_{22})y$ and, hence, are LUS residuals. Because of this property, they may be used to devise exact tests for such problems as serial correlation, non-normality, and structural change. Furthermore, they need not sum to zero and can exhibit properties quite different from those of the OLS residuals. Harvey has suggested that plots of these residuals can be a useful graphical tool for detecting misspecification. One final feature of these residuals is that the sum of their squares is equal to the sum of the squared OLS errors.

The forecast errors, themselves, may also be of analytical value. One use for a historical series of forecast errors is in assessing risk. Young discusses a measure of objective risk as the mean squared step ahead forecast error and gives a number of criteria for judging whether a forecast process is suitable for calculating such a measure. The sequential regression technique satisfies all of these criteria, including the use of only past information, frequent updating, incorporation of new information as it becomes available, and ease of computation. The property that recent information should be given higher weight can be achieved by using a discounting technique that is easily incorporated into the algorithm.

Thus the sequential regression approach yields a richer set of results while computing OLS estimates at little extra cost. Extensions can be made to models which fail to meet classical OLS assumptions. This offers a potentially fruitful area for future study.

Applications To ARIMA Type Models

Increasingly, economists are turning to time series models such as the ARIMA model suggested by Box and Jenkins. While purely autoregressive models present few estimation problems, the same cannot be said for models that contain moving average components. Two problems arise with these models. First, estimators are intrinsically non-linear in the dependent variable. Most approaches make use of a search or gradient method that seeks parameters which minimize the sum of squared errors. These methods tend to be computationally costly and may exhibit convergence problems. The second problem concerns specifying initial values for error components. Box and Jenkins advocate the use of backcasting pre-sample dependent variable values with pre-sample errors set to zero. This allows within sample errors to be calculated over the entire sample period.

An alternative estimation procedure is to set initial forecast errors equal to the errors of a naive (no change) forecast. Sequential regression can then be used to calculate the step ahead forecast error which can, in turn, be used as regressors in subsequent iterations. Because the influence of moving average components dies out after a limited number of periods, these estimates can be expected to converge to estimates made using more sophisticated techniques.

The advantage of this approach is that computations may be done simply and quickly, without recourse to non-linear methods. In addition, the advantages cited in the preceeding section also apply in this setting. A set of sequential coefficients is generated, which may be used to evaluate the stability of the process and to identify periods of change. The forecast errors are, again, of interest, both for evaluating the suitability of the model specification and for use risk analyses.

Finally, with little extra effort, the sequential regression approach could be applied to models that combine both standard regression and ARIMA features. Indeed there is often little reason for the separate treatment of these types of models.

Results of Modeling Experiments

In order to demonstrate the usefulness of sequential regression we have extended Bessler and Brandt's analysis of quarterly hog and cattle prices. In their study, using data available through 1975(IV), they specified econometric and ARIMA models for both price series. The econometric models used past values of sows farrowing, pounds of cattle slaughtered, broiler eggs hatched and, for the hog model, cured and frozen pork storage. The log of consumer disposable income was also used as a regressor in both models. The ARIMA models were both first order moving average models, with the hog model having a non-zero parameter at the fifth lag and the cattle model at the first lag. Bessler and Brandt used these models to generate step ahead forecasts for the period 1976(I) - 1979(II). In deriving their econometric model forecasts, Bessler and Brandt performed a sequential updating of model coefficients. They did not update the estimates for their ARIMA models. By way of contrast, tables 1 and 2 present forecast errors for both types of models

Table 1. Hog Price Models

Date	Price	Naive Errors	Econometric Model Errors		ARIMA Model Errors		Mixed Model
			No Updating	With Updating	No Updating 1/	With Updating	
76 (I)	47.90	-4.24	-0.65	-0.65	-0.64	1.39	-0.67
76 (II)	49.15	1.25	2.39	2.51	0.10	0.38	2.51
76 (III)	43.53	-5.62	-4.05	-4.23	-3.28	-2.94	-4.24
76 (IV)	34.16	-9.37	-9.69	-9.53	-4.96	-3.83	-9.57
77 (I)	38.96	4.80	-7.32	-6.36	3.69	3.77	-6.10
77 (II)	40.76	1.80	-5.61	-4.06	1.52	2.57	-3.71
77 (III)	43.67	2.91	-3.07	-1.33	2.95	3.12	-1.01
77 (IV)	41.30	-2.37	-8.80	-6.54	-3.82	-3.97	-6.27
78 (I)	47.43	6.13	-3.13	-0.30	3.94	4.09	-0.21
78 (II)	47.85	0.42	-2.76	-1.16	2.05	2.50	-0.02
78 (III)	48.59	0.74	-0.50	1.81	1.41	2.13	1.89
78 (IV)	50.03	1.44	-4.62	-2.05	2.75	3.11	-1.97
79 (I)	51.79	1.76	-0.42	2.10	0.07	-0.34	2.29
79 (II)	43.07	-8.72	-9.85	-7.94	-6.98	-6.58	-7.98
mean squared error:		21.34	30.33	20.87	10.92	10.87	20.29

1/ Adapted from Bessler and Brandt.

Table 2. Cattle Price Models

Date	Price	Naive Errors	Econometric Model Errors: No Updating	Econometric Model Errors: With Updating	ARIMA Model Errors: No Updating[1/]	ARIMA Model Errors: With Updating	Mixed Model
76 (I)	38.68	-7.43	-5.02	-3.02	-7.77	-7.76	-3.90
76 (II)	41.41	2.73	0.38	0.83	0.02	0.56	1.47
76 (III)	37.29	-4.12	-6.98	-6.84	-4.11	-3.96	-6.94
76 (IV)	38.98	1.69	-2.19	-1.54	0.26	0.54	-0.27
77 (I)	37.87	-1.11	-5.98	-5.12	-1.02	-0.95	-5.23
77 (II)	40.75	2.88	-3.84	-2.62	2.52	2.60	-1.75
77 (III)	40.45	-0.30	-6.24	-5.05	0.58	0.48	-4.58
77 (IV)	42.41	1.96	-4.04	-2.39	2.16	2.10	-1.50
78 (I)	45.78	3.37	-2.09	-0.16	4.12	3.99	-0.00
78 (II)	55.12	9.34	5.46	7.37	10.77	10.47	7.40
78 (III)	53.70	-1.42	1.96	3.12	2.34	0.87	1.97
78 (IV)	54.75	1.05	1.46	2.16	1.86	1.23	2.17
79 (I)	65.16	10.41	11.40	12.07	11.06	10.66	11.88
79 (II)	72.46	7.30	14.43	12.57	11.16	9.31	10.85
mean squared error:		25.67	40.22	35.41	34.19	29.78	31.54

1/ Adapted from Bessler and Brandt.

using coefficients based on data up through 1975(IV) and those using sequential updating of coefficients. The sequential ARIMA model used naive (no change) forecast errors to initialize the error process, as was discussed in the previous section.

In all cases, forecasts based on sequentially updated coefficients had lower mean squared errors over the 14 quarter test period than did those without updating. This was particularly true of the econometric models. An examination of the sequence of coefficients revealed strong trends in their values, which may help explain this result. In the hog model, the coefficients associated with number of sows farrowing, pounds of cattle slaughtered, and broiler eggs hatched all were negative and became absolutely larger over time in an almost monotonic fashion. The coefficient associated with disposable income was positive and rose from the late 1960s through the mid-1970s. Also some sharp jumps in coefficient values were noticeable in the 1973-74 period. In the cattle model the coefficients associated with pounds of cattle slaughtered and income also showed similar strong trends. Again, sudden jumps in the 1973-74 period were present.

All the coefficients in the hog model had large t-statistics (>5). In the cattle model, the cattle slaughter and income variables had t-statistics greater than 5 while the other regressors had t-statistics less than 1. Thus, in both models the coefficients of variables with significant explanatory power were also the ones that exhibited trends. While it is not our intention to attempt an analysis of the economic factors, if any, causing these results, we do, however, feel they suggest that such an analysis would be worthwhile.

Turning to the ARIMA models, the improvement was slight in the hog model and somewhat larger in the cattle model. Again, an examination of the coefficient sequences proved instructive. With the hog model, the coefficient on the fifth lagged forecast error ranged between -0.46 and -0.55 in the 1976(I) - 1979(II) period and showed no discernible trend. With the cattle model, on the other hand, the initial value of the coefficient associated with the first lagged forecast error was -0.30, but it began declining sharply, in an absolute sense, after 1977(IV) and was estimated to be -0.08 in 1979(II). This change occurred during a time when cattle prices increased by over 70 percent.

Neither model exhibited clearly discernible trends in the values of coefficient estimates over the whole sample period. However, as in the econometric models, there were some instances of sharp jumps in estimate levels, including especially the 1973-74 period. It is encouraging to note that the coefficient estimates for 1975(IV) of -0.50 and -0.30, for the hog and cattle models respectively, were reasonably close to the values of -0.44 and -0.35 obtained by Bessler and Brandt using usual ARIMA estimation methods. This lends at least preliminary support to the claim that the sequential approach outlined earlier is a reasonably good method for deriving estimates for models with moving average components.

In one final experiment we estimated the parameters of the econometric models, but included the most recent forecast error as an additional regressor. The results of this experiment are shown in tables 1 and 2 under the "Mixed Model" column heading. This estimator is suggested as an alternative for models exhibiting first order autocorrelation in their error terms. It simultaneously estimates the autocorrelation coefficient and the coefficients associated with the regressors.

For the hog model, this addition had little impact on mean square forecast error and the value of the estimated autocorrelation coefficient was generally close to zero. For the cattle model, however, this addition did reduce the mean squared error by over 10 percent. Investigating the matter further, it was found that a test based on the Durbin-Watson statistic would have been inconclusive at the 5 percent signficance level with the sample period extending up until 1975(IV). However, with the full sample through 1979(II), this test would have rejected the hypothesis of zero autocorrelation at the 2.5 percent significance level. The estimated values of the autocorrelation coefficient for these two periods were 0.20 and 0.29.

Summary and Conclusions

In this paper, the sequential regression approach to estimation has been shown to have a number of advantages over more familar estimation procedures. The existence of an efficient, accurate sequential regression algorithm, which can be implemented on a microcomputer, makes this approach feasible for a wide range of practical applications. In addition to providing OLS estimates this approach makes it easy to derive a rich set of results from which tests and graphical checks for misspecification can be made. The extension to weighted regression to correct for heteroscedasticity is straightforward and an alternative sequential approach to estimation when errors exhibit autocorrelation has been suggested and implemented empirically in this paper.

A sequential approach was also suggested for ARIMA models. In addition to providing a sequence of coefficients and forecast errors based only of past data, the approach presented here provides a simple way to estimate the parameters associated with moving average components and to update those estimates as new data becomes available.

The sequential regression approach can be used for a wide range of time series modeling applications. It is in the area of forecasting, however, that this approach is particularly useful. It has long been recognized that the model with the smallest within sample sum of squares may be virtually worthless as a forecasting model. By using a sequential estimation approach a series of actual forecast errors is generated. These can be used to more properly assess the forecasting ability of the model. They are the basis for construction of an empirical cumulative distribution, which can be evaluated relative to the distribution of other forecast models using any number of alternative utility (loss) functions.

In a related application, this set of forecast errors can be used to derive measures of riskiness for both analytical and extension purposes. The sequential approach satisfies all the criteria that Young suggested for such measures. Since it is flexible in regard to model specification, however, this approach permits the incorporation of far more information than do the generally ad hoc approaches reviewed by Young.

Sequential regression is, then, a flexible tool for modeling time series. Further research is needed in three broad areas. First, the possibilities for new types of model specification need to be explored in greater depth. Second, the sampling properties of estimators that incorporate forecast errors as regressors need to be determined. Finally, our understanding of how sequential regression results can be used for hypothesis testing and decision analysis need to be broadened.

REFERENCES

Bessler, D.A. and J.A. Brandt. "Composite Forecasting of Livestock Prices: An Analysis of Combining Alternative Forecasting Methods." Purdue University, Department of Agricultural Economics, Station Bulletin 265. 1979.

Box, G.E.P., and G.M. Jenkins. Time Series Analysis: Forecasting and Control, revised ed. San Francisco: Holden Day, Inc. 1976.

Gentleman, W.M. "Least Squares Computations by Givens Transformations Without Square Roots." J. Inst. Maths. Applics. 12(1973):329-36.

Harvey, A.C. The Econometric Analysis of Time Series. New York: John Wiley and Sons, Inc. 1981.

Theil, H. Principles of Econometrics. New York: John Wiley and Sons, Inc., 1971.

Young, D.L. "Evaluating Procedures for Computing Objective Risk from Historical Time Series," in Risk Analysis in Agriculture: Research and Educational Developments. University of Illinois, Department of Agricultural Economics. AE-4492. PP. 1-21, 1980.

APPENDIX

```
      SUBROUTINE GIVENS(N,M,DELTA,D,R,X,Y,THETA,DISCOUNT)
      DIMENSION D(*),R(*,*),X(*),Y(**),THETA(*,**)
      DO 30 I=1,N
        DNEW=D(I)+DELTA*X(I)**2
        IF(X(I).EQ.0.OR.DELTA.EQ.0)GOTO 30
        C=D(I)/DNEW
        S=DELTA*X(I)/DNEW
        II=I+1
        DO 10 J=II,N
          TEMP=R(I,J)
          R(I,J)=C*TEMP+S*X(J)
10        X(J)=X(J)-TEMP*X(I)
        DO 20 J=1,M
          TEMP=THETA(I,J)
          THETA(I,J)=C*TEMP+S*Y(J)
20        Y(J)=Y(J)-TEMP*X(I)
        DELTA=C*DELTA
30      D(I)=DNEW*DISCOUNT
       RETURN
```

Remarks:

N=# of regressors
M=# of equations
DISCOUNT=a factor to exponentially discount past observations (0 < DISCOUNT < 1) ; DISCOUNT=1 typically
X=an N vector of regressor values
Y=an M vector of dependent variable values
D,R,THETA,DELTA,C,S=variables as defined in Gentleman's article
DNEW=D' of Gentleman's article

Input includes current values of X and Y, and the accumulated values of D,R,and THETA. Delta is input as 1 for OLS and as some weighting value in a weighted regression. In a heteroskedastic situation Var(u(t)) is inversely proportional to this weight, where u(t) is the "true" error term.

Output includes updated values of D,R, and Theta. Y will contain step ahead forecast errors (ie. e(t)=y(t)-X(t)b(t-1), where b(t-1) is the coefficient vector calculated in the previous period). DELTA will contain a constant proportional to the variance of the forecast error such that Y/(DELTA**.5) yields the recursive residual.

Notice that Y may be scalar (M=1) or vector (M>1). This allows for straightforeward estimation of reduced form or vector autoregressive models.

Also notice that intermediate coefficient estimates need never be calculated if sequential errors and/or final coefficient values are all that are desired.