Price Forecasting Methods and Evaluation Procedures

by

Jon A. Brandt and David A. Bessler

PRICE FORECASTING METHODS AND EVALUATION PROCEDURES

Jon A. Brandt and David A. Bessler*

Introduction

The uncertainty of future price, production, and consumption levels make market strategy and investment planning difficult. Low demand price elasticities for most agricultural products, coupled with frequently large seasonal changes in production, provide the setting for rather large price fluctuations. In such an environment, a detailed listing of various courses of action, their consequences under alternative outcomes, and a statement of well-being under each is, of course, recommended. Forecasts of commodity price changes given specified market conditions provide information necessary to carry out the marketing or investment planning process.

As suggested above, forecasts may be used for different reasons. The intended use of the forecasts, in large degree, influences or dictates the type of model selected to generate the forecast and the

*The authors are assistant professors in the Department of Agricultural Economics at Purdue University. This paper is Journal Paper No. 8785 with the Purdue Agricultural Experiment Station.

criteria used to evaluate the forecasting technique. The objective of avoiding extremely large forecast errors might suggest a forecast approach different from that associated with the goal of predicting price turning points in the market. Similarly, short-run marketing strategies would require a different set of predictions than long-run investment planning.

In this paper, we have three objectives. First, we briefly review alternative methods of forecast generation. As we have discussed this point elsewhere (Bessler and Brandt), we do not spend much time on it here. Second, we discuss a shopping list of alternative evaluation procedures. Within this list we present a rather new procedure for testing the significance of average squared error evaluations. Finally, our third objective is to apply some of these methods of forecasting and evaluation to quarterly U.S. hog prices.

Methods of Forecast Generation

The literature contains numerous approaches to short-term forecasting. Following the distinction made by the late Jacob Marschak we can classify an approach as either structural or nonstructural. The former uses estimates of parameters identified (determined to be tentatively non-zero) by way of a general theory or structural model; the latter uses just historical observations on the series (and perhaps a few related series) to be forecasted. Models of both types have demonstrated success in applied studies.[1]

Forecasting models can also be classified as mechanical or nonmechanical depending on whether human judgment is required to arrive at a set of forecasts. Mechanical models, once built, can be maintained by merely "reading in" new empirical observations--the generated forecasts being used without the requirement of human judgment. The mechanical distinction can also extend to model construction. The general loss function models for fitting univariate or bivariate autoregressive processes, as discussed by Akaike and Amemiya and applied by Hsiao and Bessler and Binkley (1980, 1981), are examples of models which require little human judgment at the model construction stage. The reader may want to contrast these approaches to the rather involved judgmental requirements of say Box and Jenkins' methods.

By nonmechanical forecasts, we mean to imply forecasts which rely on human judgment. Most econometric models fall into this category. Here generous amounts of judgment are used at both the model construction and forecast generation stage. Sims quite well describes this judgmental element in usual econometric model construction (of a demand relationship):

> It is a common and reasonable practice to make shrewd aggregations and exclusion restrictions so that our small partial equilibrium system omits most of the many prices we know enter the demand relation in principle and possibly includes a shrewdly selected set of exogenous variables we expect to be especially important in explaining variations in . . . demand (p. 2).

The widely known three-step procedures of Box and Jenkins also fall into this nonmechanical typology. Here the analyst must use generous amounts of judgment and experience in identifying the autoregressive integrated moving average (ARIMA) orders of a particular series. Extensions to multivariate processes (Tiao and Box) require much human intervention also.

Forecasts also often rely on judgment. Once the model is built, the outputted forecasts are often "calibrated" to take into account information not captured by the model. For example, Fair (p. 285) in evaluating short-run forecasting models notes, "The current practice of most model proprietors who issue regular forecasts is to adjust the forecasts from their models before the forecasts are released." Similar arguments regarding model calibration have been made by Klein and Burmeister.

Of course the distinction between mechanical and nonmechanical is not always clear--there are degrees of mechanization in the range of forecast methods in use today. For instance, at Purdue we have a commodity outlook staff who base their forecasts entirely on their knowledge of the industry. They make no explicit use of formal models. Others, however, use either econometric or time series models, calibrated by their own judgments. We do not seem to have any forecasters who use just mechanical models (although we suspect these may improve forecast performance in some cases).

It has been shown elsewhere, that where alternative models are available, the most appropriate action for one seeking to make the best forecast is not necessarily to determine and use the best individual forecasting method. Indeed, just as a risk averse economic agent gains in well-being through diversification, so too can a forecast user gain by forming a composite forecast. Given the vast amount of information (some quantitative, much nonquantitative) available, which may influence future economic variables, it is our feeling that consideration of alternative forecasts is not only practical but also wise. In fact, formal combinations of structural and nonstructural forecasts with either mechanical or nonmechanical forecasts have been shown to outperform individual model forecasts. We do not review methods to combine forecasts here. Such a review and an empirical demonstration can be found in Granger and Newbold (1977, pp. 269-278) and Brandt and Bessler, respectively.

The Evaluation of Forecasts

Here we present alternative methods of forecast evaluation. Most of what we say concerning evaluation is with respect to alternative forecasting models (comparing or ordering forecasts obtained from two or more forecasting methods). This is so because it is difficult to say much of anything about a forecast method when viewed in isolation.

We also suggest that forecast evaluation should (whenever possible) be based on out-of-sample forecasting experience. That is, the period over which the evaluation is constructed should not be the same as that over which the parameters of the model were fit. It is generally well accepted that it is easier to find a model which fits better than another than to find a model which predicts better. In fact, it is possible to fit a high order polynomial in time to time ordered observations and obtain a perfect fit; however, actual out-of-sample forecasting will most likely show such profligate models perform poorly. The aim is to obtain simple models which actually predict well in out-of-sample tests.

Here we review a number of out-of-sample forecasting evaluation methods. Prime in this discussion is mean squared error and the distribution of mean squared error differences. The mean squared error provides a measure of the size of individual forecast errors from the actual values. Because the error is squared, large errors detract substantially more from the performance of the forecasting method than do small errors. The premium is placed on generating forecasts which do not differ greatly from the true values.[2]

The mean squared error (or equivalently root mean squared error) has been used extensively in applied and theoretical works on forecasting. In the applied work it is common to present mean squared error calculations for alternative models, and explicitly or implicitly to pick the method with lowest MSE. Little or no attention has been given to the stochastic nature of these MSE predictions. Indeed, in some of our own work we merely present MSE's on alternative models and imply that better performance is obtained by the method yielding the lowest MSE. A preferred approach for analysis would involve the statistical test of significant differences among MSE's of alternative models. Our reading to date suggests that only Ashley, Granger, and Schmalensee take this preferred approach.

In formulating a test of mean squared error differences, it should be noted at the outset that no simple test is possible. For example, one can not rely on usual tests for equality of means found in statistics books (see Mood, Graybill and Boes, p. 432) because our sample observations (squared errors in T periods) are not likely to be independent. That is, the forecasted error series are likely to be cross-correlated and autocorrelated due to specification error, sampling error, and possible structural changes. The test offered by Ashley, Granger, and Schmalensee takes these possible interdependencies into account directly. The test procedure is outlined in Appendix 1. This test is applied to various forecasts (including composite forecasts) of U.S. hog prices in the empirical application section of this paper.

Two other single variables error evaluations include the mean forecast error and the mean absolute forecast error. The mean forecast error is determined by taking the average of the difference between the summation of the over predictions and the summation of the under predictions. A negative sign would indicate that the average forecast series is above the mean of the actual series; a positive sign suggests an average forecast which is low. The mean absolute forecast error is simply the average of the absolute values of the forecast errors. These measures also provide the researcher with an indication of the goodness of the particular model.

Another group of performance indicators include tracking measures, which might be used to track the movements of actual and forecast price series. The number of turning points missed or falsely predicted compared with those correctly forecast are particularly useful, if the forecaster or users of the forecasts are interested in knowing when a series is likely to turn upward or downward from its current pattern. These measures will not indicate which forecasting method most closely approximates the actual series. In addition to calculating turning point measures, an evaluation of the number of under- or over-predictions would provide the forecaster with further performance evaluation information.

Dhrymes et al., (p. 315) also suggest comparing econometric forecasts with the forecasts from other methods including various "naive" forecasts, forecasts based on "judgmental," "consensus," or other non-econometric, and other econometric forecasts. This type of evaluation provides the forecaster with a measure of the relative goodness of his model compared with other available techniques. As with the other non-parametric tests, this measure provides information to the model builder which increases or decreases his confidence in the forecasting ability of the particular technique.

Other measures could be examined which provide further tests of performance of the forecasting models. One includes a "face value" examination of the models over the "fit" period. That is, for the econometric model, the signs of the coefficients and their corresponding t-statistics could be used as indicators of how well the model fits the data prior to forecasting. This provides no guarantee that the model will perform well over the forecast period; however, it does suggest where improvements in the specification might be made and provides some evidence of what could be expected assuming the data remained within the current range.

Empirical Applications

In order to illustrate empirically several of the forecasting techniques and evaluation measures discussed above, models for predicting hog prices in the United States were specified and estimated. We do not wish to imply that any of the models or techniques used to forecast hog prices is necessarily the best available. They do, however, allow us to examine alternative methods of combining forecasts and to evaluate forecasting performance through selected criteria or measures.

Quarterly data were used to estimate an econometric model and an autoregressive integrated moving average (ARIMA) process over the period extending from the first quarter of 1960 (6001) through the fourth quarter of 1975 (7504). The estimated coefficients, variables used in the models, and selected summary statistics are shown in Table 1. The econometric estimation is a single equation in reduced-form of a structural model. Each of the estimated coefficients is large relative to its standard error. The R^2 and Durbin-Watson statistics corresponding t-statistics could be used as indicators of how well the model fits the data prior to forecasting. This provides no guarantee that the model will perform well over the forecast period; however, it does suggest where improvements in the specification might be made and provides some evidence of what could be expected assuming the data remained within the current range.

Empirical Applications

In order to illustrate empirically several of the forecasting techniques and evaluation measures discussed above, models for predicting hog prices in the United States were specified and estimated. We do not wish to imply that any of the models or techniques used to forecast hog prices is necessarily the best available. They do, however, allow us to examine alternative methods of combining forecasts and to evaluate forecasting performance through selected criteria or measures.

Quarterly data were used to estimate an econometric model and an autoregressive integrated moving average (ARIMA) process over the period extending from the first quarter of 1960 (6001) through the fourth quarter of 1975 (7504). The estimated coefficients, variables used in the models, and selected summary statistics are shown in Table 1. The econometric estimation is a single equation in reduced-form of a structural model. Each of the estimated coefficients is large relative to its standard error. The R^2 and Durbin-Watson statistics

Table 1. Forecasting Models Used to Predict Hog Prices, 7601-8004

Econometric[a]

$$HP_t = -168.61 + \underset{(2.64)}{46.91}\ell nY_t - 8.44SF_{t-2} - \underset{(.68)}{3.94}SF_{t-3} - \underset{(5.76)}{44.91}CS_{t-1}$$

$$- \underset{(.74)}{3.84}CLST_{t-1} - \underset{(7.01)}{43.33}HTCH_{t-1}$$

$$R^2 = .93 \qquad D.W. = 1.84$$

ARIMA[b]

$$(1-B)HP_t = (1-\underset{(.14)}{.43}B^5)\hat{e}_y \quad R^2 = .87 \quad \text{Ljung-Box Q-statistic} = 21.65$$

Adaptive[c]

$$HP_t = a_t[\text{Economic forecast}_t] + (1-a_t)[\text{ARIMA forecast}_t]$$

where

$$a_{N+1} = \frac{\sum\limits_{j=N-2}^{N} \hat{e}^2_{2j}}{\sum\limits_{i=1}^{2} \sum\limits_{j=N-2}^{N} \hat{e}^2_{ij}}$$

[a] The econometric model is continuously updated by re-estimation as data from subsequent periods become available. The variables include: HP-price of all barrows and gilts at seven terminal markets; ℓnY-logarithm of total disposable income; SF-number of sows farrowing in fourteen states; CS-cold storage of pork in U.S.; CSLT-pounds of meat from U.S. commercial cattle slaughter; HTCH-number of broiler-type chicks hatched in U.S.

[b] The observed error from the model is designed as ê. The critical Q-statistic for 23 degrees of freedom at the 5 percent level of significance is 35.17.

[c] N refers to the number of observations of the estimation or sample period. Weighting the forecasts based on the squared forecast errors of the two periods prior to the prediction period reflects the optimal adaptive scheme for hog prices. In this case, e_2 is the error associated with the ARIMA forecast, e_1 with the economic forecast.

Table 1 (Continued)

Minimum Variance[d]

$$HP_t = b \cdot [\text{Economic forecast}_t] + (1-b)[\text{ARIMA forecast}_t]$$

where

$$b = \frac{\sigma_2^2 - \rho_{12}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2}$$

Simple Average

$$HP_r = .5\ [\text{Econometric forecast}_t] + .5[\text{ARIMA forecast}_t]$$

[d] σ_1 and σ_2 are the standard deviations of errors for the econometric and ARIMA forecast models, respectively, over the period of model estimation (6001-7504). ρ_{12} is the correlation coefficient between the observed errors.

suggest the variables explain the movement in hog prices over time rather well without any indication of serial correlation in the errors. Similarly, the ARIMA model indicates a good fit of the data and the Q-statistic is well under the critical value suggesting the hypothesis that the autocorrelations are based on white noise residuals could not be rejected.

Using the econometric and ARIMA models, one quarter ahead forecasts of hog prices were generated over a 20-quarter period (7601-8004).[3] The forecasts of the individual models were then combined using three different methods to form composite forecasts. These composite techniques and the methods for determining their weighting schemes are illustrated in Table 1. The optimal weighting pattern for the adaptive composite forecasting procedure was based on the errors of the two most recent forecasted periods. The composite technique described as minimum variance is based on the prediction errors of the econometric and ARIMA models over the estimation or fit period (6001-7504). The weight assigned to the econometric model is based on the variance of the errors and correlation coefficient between the errors of the two individual models. This weight, once determined, remains constant over the forecasting period. Finally, and for comparison, a simple average of the forecasts of the two models is generated.

The weights of the adaptive composite model for the econometric forecasts ranged from a low of .029 to a high of .998. The average weight was .408 indicating that the squared errors of the econometric forecasts were larger than those of the ARIMA process. The weight for the econometric model forecasts based on the minimum variance criterion was .735. This is consistent with the information in Table 1, which suggests that the estimated econometric model fits the data better than the ARIMA model over the estimation or sample period, 6001-7504. Recall that the minimum variance weights are based only on the sample estimation period, whereas the adaptive weights change during the out-of-sample forecast period.

In Table 2, forecasts of the five techniques and actual barrow and gilt prices for seven terminal markets are given. Over the forecast period, actual prices ranged from a high of $51.98/cwt. in

Table 2. Forecasts of Hog Prices from Individual and Composite Techniques, 7601-8004

Period	Actual prices[a]	Econometric	ARIMA	Adaptive	Minimum variance	Simple average
		(Dollars per hundredweight)				
7601	47.99	48.442	48.801	48.573	48.537	48.622
7602	49.19	46.380	49.064	46.537	47.092	47.722
7603	43.88	47.476	46.728	46.785	47.278	47.102
7604	34.25	43.561	39.772	40.836	42.556	41.667
7701	39.08	45.146	35.327	38.069	42.542	40.237
7702	40.87	44.614	39.429	40.804	43.239	42.022
7703	43.85	44.992	40.816	41.824	43.885	42.904
7704	41.38	47.823	45.075	46.240	47.094	46.449
7801	47.44	47.603	43.755	45.094	46.583	45.679
7802	47.84	47.866	45.826	46.634	47.325	46.846
7803	48.52	46.607	47.220	46.608	46.770	46.914
7804	50.05	52.121	47.215	50.212	50.820	49.668
7901	51.98	49.580	51.640	50.506	51.126	50.610
7902	43.04	50.704	50.395	50.533	50.622	50.550
7903	38.52	47.786	42.173	44.736	46.300	44.980
7904	36.39	47.023	37.961	40.843	44.620	42.492
8001	36.74	44.019	35.170	35.822	41.672	39.595
8002	31.18	40.281	36.594	36.700	39.303	38.438
8003	46.23	40.810	34.344	35.570	39.095	37.577
8004	47.38	46.548	47.801	47.045	46.880	47.175

[a] Actual prices are the quarterly average of the monthly prices for all barrows and gilts in seven terminal markets reported by USDA.

7901 to a low of $31.18/cwt. in 8002, only five quarters later. None of the models appears to have been very accurate over the rather volatile period of 7902-8003 although the ARIMA model's forecast appears to react the quickest to the price change signals.

A closer examination of the forecast errors is afforded by the statistics in Table 3. The first column of numbers gives the mean of the 20 forecast errors for each of the techniques. Clearly, the ARIMA model has the mean error closest to zero of any of the models. The composite model mean errors lie between those of the econometric and ARIMA errors. Because of its ability to change weights as forecast errors change, the adaptive procedure has the next lowest mean error. The high weight for the econometric forecasts resulted in the rather large mean error of the minimum variance technique.

The mean of the absolute errors suggests the same ordering forecasting techniques but the size of the means are now much closer. This only tends to verify the bias in the econometric forecasting technique where the errors from the ARIMA process tended to cancel out over the twenty periods (resulting in the mean error close to zero). Interestingly, little difference is seen among the variances of the forecast errors. In fact, the simple average procedure produced the lowest error variance.

The mean squared error (MSE) criterion is commonly used to rank or compare forecasting procedures. Based on the statistics shown in Table 3, the ARIMA model was superior, followed closely by the adaptive and simple average methods. The minimum variance procedure had a mean squared error closest to the ARIMA model than to that of the econometric model but clearly suffered due to the high weight placed on the econometric forecasts.

From statistics, we know that the mean squared error consists of

Table 3. Performance Measures for Out of Sample Forecasts of Hog Prices, 7601-8004

Forecasting technique	Forecast Errors			
	Mean	Absolute mean	Variance	Mean squared error
Econometric	-3.18	4.52	21.82	31.93
ARIMA	.03	3.16	17.44	17.45
Composite:				
Adaptive	-.71	3.17	17.52	18.02
Minimum variance	-2.33	3.80	18.12	23.54
Simple average	-1.57	3.31	16.38	18.85

two components - the variance and the bias squared. Squaring the mean of the errors gives the bias squared. Thus, it is easy to see that although the variances of the forecasts of the five procedures were similar in magnitude, the biases of the econometric and minimum variance techniques put them at a severe disadvantage. The relative low magnitude of the simple average MSE was due to the low variance offsetting the rather large mean error.

In an earlier section, a procedure for testing whether the mean squared errors of two models were different was described. This procedure allows researchers to evaluate or compare alternative forecasting models in a more rigorous fashion than has heretofore been available. This test was applied to the original forecast errors of the respective techniques which can be calculated from Table 2.

Based on Table 3, we expected that the mean squared error of the econometric model would be significantly greater than those of the ARIMA model and each of the composite techniques. We did obtain these results. However, somewhat curiously, the mean squared error of the econometric model was statistically greater than the MSE's of the composites at far lower significance levels (.1%) than it was compared to the ARIMA MSE (10%). Recall that the largest difference in the MSE's occurred between the econometric and ARIMA forecast errors.

Further analysis of the errors suggested that the ARIMA model had an unusually large error in period 8003. Without this observation, the standard error of the regression (Equation (3)) is reduced and the likelihood of obtaining this observation is one in 500. Following a suggestion by R. Ashley, the observation was deleted and Equation 3 was estimated. The results of these regressions are summarized in Table 4.

The F-statistics in Table 4 suggest that the mean squared error from the econometric forecasts is significantly larger than those of the ARIMA and composite model forecasts. Although there are shown to be statistical differences between the MSE's of the composite and ARIMA forecasts, the significance levels of these differences are higher, thereby indicating less statistical significance. These results were anticipated, because since the mean squared errors of all forecasting approaches with the exception of

Table 4. Summary Statistics on Tests of Mean Square Errors Differences

Mean square[a] error comparisons	t Statistic	F Statistic	Significance level, %[b]	Durbin-Watson
Econometric > ARIMA		3.52	1.33	1.73[c]
Econometric > Adaptive		3.57	1.28	1.57[c]
Econometric > Minimum Variance		14.79	.1	1.65
Econometric > Simple Average		11.99	.1	1.57
Minimum Variance > Simple Average		9.53	.1	1.47
Minimum Variance > Adaptive	2.06[d]		3.00	1.30[c,e]
Simple Average > Adaptive	1.75[d]		5.00	1.26[c,e]
Minimum Variance > ARIMA		2.69	2.43	1.66[c]
Simple Average > ARIMA	2.19[d]		2.00	1.62[c]
Adapative > ARIMA		2.21	3.51	1.65

[a] The statistics test whether the mean square error of the forecasting technique to the left of the > sign is significantly greater than the mean square error of the technique to the right of the sign. The regression is based on Equation (3) in the text.

[b] From Abramowitz and Stegun, the significance level of an F-statistic with 2 n degrees of freedom is given by $[n/(n+2F)]^{n/2}$.

[c] These statistics were based on the Cochrane-Orcutt iterative procedure for generalized least squares estimation.

[d] One β coefficient was negative but insignificant. The t value of the positive β coefficient is reported.

[e] Inconclusive test for the presence of serial correlation of the error term at the five percent level of significance.

those from the econometric model are rather close in magnitude.

Another measure of forecasting performance commonly used is turning point statistics. They indicate how well the forecasting models track changes in the data series over time. Good performance is indicated by predicting changes in the direction of price movements when they occur and predicting no change in their movements when they do not occur. Thus, high numbers in the (1,1) and (2,2) diagonal elements are preferred.

Table 5 indicates that only the ARIMA model generated the correct directional signals more than fifty percent of the time. The minimum variance technique had the poorest performance. The simple average method predicted correct directional movement one-half of the time. The econometric and adaptive methods were correct only seven times out of 18 potential changes.

Conclusions

Several techniques for generating forecasts and various procedures measuring performance were described and illustrated in this paper. These by no means exhaust the list of alternatives. They do, however, lead to several suggestions for builders of forecasting models and users of forecasts.

For forecast users, the evidence suggests that little or no significant difference is found between the mean squared error of the better individual forecasting (ARIMA) model and those of composite

Table 5. Tracking Measures for Evaluating Hog Price Forecasts - Turning Point Statistics

Forecast direction of price movements		Actual direction of price movements	
		Change	No change
Econometric	Change	5	8
	No change	3	2
ARIMA	Change	5	3
	No change	3	7
Adaptive	Change	4	7
	No change	4	3
Minimum variance	Change	3	8
	No change	5	2
Simple average	Change	2	3
	No change	6	7

models. However, significant improvement was shown moving from the poorer individual forecasting (econometric) model to the composite models. This is consistent with earlier results that suggest the use of composite models can reduce the forecast errors which may be evident in individual models. The forecast user who selects the wrong (i.e., poor) forecasting model for making marketing and investment decisions is not likely to be in business long. Somewhat surprising, the simple average composite fared quite well in forecasting performance compared with the forecasts of more sophisticated composite procedures.

Forecast model builders are urged to evaluate the output of their models in a variety of ways. The procedure chosen to determine the accuracy or acceptability of the model depends, in part, on its ultimate intended use. The methods described in this paper are but a few of those available for performance evaluation. Model builders should also compare their forecasts with those of other models, ranging from naive "no change" models to more sophisticated approaches. However, forecast model builders should not be satisfied that their model simply performs better than another. Model builders must continuously seek to improve the accuracy and consistency of their forecasts.

Finally, we would be remiss in our task of examining forecasting methods and evaluation procedures if we failed to suggest areas of needed research. Important on this list, in our consideration, is more attention to the economic significance of forecasts instead of simply evaluating their statistical properties. We are also guilty of this. Useful to decision-makers such as producers, packers, handlers, processors, and distributors of agricultural products would be comparative evaluations of the economic outcomes of decisions based on alternative forecasting procedures. While the historical performance would not necessarily be indicative of future performance, it would perhaps make model builders more keenly aware of the importance of their forecasts and the advice to decision-makers that is implicitly contained in those forecasts.

Footnotes

[1] Recent work on multivariate time series suggests this structural-nonstructural distinction is perhaps not as clear as in Marschak's day. For example, the general multivariate ARIMA model can be shown to be of the same form as the structural econometric model (Granger and Newbold, 1977, Chapter 7). However, whereas the structural model builder presumes to know where to place zeros in the general ARIMA specification (based on theory), the nonstructural model builder is more willing to allow the data to specify such zeros.

[2] By denoting the series to be forecast as X_t and its forecast with method j as P_t^j, the mean squared error (MSE) for T out-of-sample forecasts is given as:

$$D_j^2 = \sum_{t=1}^{T} (X_t - P_t^j)^2/T.$$

This is an estimate of the expected forecast error $E(e_{jt}^2) = E(X_t - P_t^j)^2$. Clearly D_j^2 increases with $|e_{jt}|$ and the natural ordering suggests method j be preferred to method k if $D_j^2 < D_k^2$.

[3] The econometric model was re-estimated in each quarter after data became available. By the end of the twenty quarter forecasting period, however, the estimated coefficients did not vary much from their initial values. The ARIMA process was not adjusted from its initial estimation; however, forecasts changed as new prices were observed.

References

Abramowitz, M. and I. Stegun. Handbook of Mathematical Functions. Dover, 1972.

Akaike, H. "Fitting Autoregressive Models for Prediction." Annals of the Institute of Statistical Mathematics 21(1969):243-247.

Amemiya, T. "Selection of Regressors." International Economic Review 21 (1980):331-354.

Ashley, R., C. W. J. Granger, and R. Schmalensee. "Advertising and Aggregate Consumption: An Analysis of Causality." Econometrica 48 (1980):1149-1167.

Bessler, D. A. and J. K. Binkley. "Autoregressive Filtering of Some Economic Data Using PRESS and FEP." Proceedings American Statistical Association, Business and Economic Statistics Section (1980):261-265.

Bessler, D. A. and J. A. Brandt. Composite Forecasting of Livestock Prices: An Analysis of Combining Alternative Forecasting Methods. Purdue University Agr. Exp. Sta. Bul. 265, 1979.

Binkley, J. K. and D. A. Bessler. "An Analysis of Short Run Behavior in Bulk Ocean Shipping." Purdue University, 1981.

Box, G. E. P. and G. M. Jenkins. Time Series Analysis, revised edition, San Francisco: Holden-Day, 1976.

Brandt, J. A. and D. A. Bessler. "Composite Forecasting: An Application With U.S. Hog Prices." American Journal of Agricultural Economics 63(1981):135-140.

Dhrymes, P. J., et al. "Criteria for Evaluation of Econometric Models." Annals of Economic and Social Measurement 1 (1972): 291-324.

Fair, Ray C. "An Evaluation of a Short-Run Forecasting Model." International Economic Review 15(1974):285-303.

Granger, C. W. J. and Paul Newbold. Forecasting Economic Time Series. New York: Academic Press, 1977.

Hsiao, C. "Autoregressive Modeling of Canadian Money and Income Data." Journal of the American Statistical Association 74 (1979):553-56.

Klein, L. R.and E. Burmeister. Econometric Model Performance. Philadelphia: University of Pennsylvania Press, 1976.

Marschak, Jacob. "Economic Measurement for Policy and Prediction." in Hood, W. C. and T. C. Koopmans editors, Studies in Econometric Method, New Haven: Yale University Press, 1953.

Mood, A. M., F. A. Graybill, and D. C. Boes. Introduction to the Theory of Statistics. New York: McGraw-Hill Book Company, 1974.

Sims, C. "Macroeconomics and Reality." Econometrica 48(1980): 1-48.

Tiao, G. C. and G. E. P. Box. "An Introduction to Applied Multiple Time Series." University of Wisconsin, Department of Statistics Technical Report No. 582, 1979.

U.S. Department of Agriculture. Livestock and Meat Statistics. Statistical Bulletin 522 and Annual Supplements, ERS, Washington, D.C.

Appendix A

Given two alternative methods of forecasting X_t, the problem is to determine whether method 1 (giving T forecasts P_t^1) is better than method 2 (giving T forecasts P_t^2). The evaluation is done in terms of out-of-sample mean squared errors. From the definition of mean squared error as the variance plus the bias squared, the difference between two mean squared errors can be written as:

$$[1] \quad MSE(e_1) - MSE(e_2) = [S(e_1)^2 - S(e_2)^2] + [m(e_1)^2 - m(e_2)^2],$$

where e_i is the error vector of out-of-sample errors $e_1' = [e_{i1}, E_{i2}, E_{i3}, \ldots, e_{iT}]$, $i = 1,2$; $S(e_i)^2$ is the sample variance of out-of-sample errors from forecast method i; and $m(e_i)$ is the sample mean of out-of-sample forecast errors from method e_i. Define

$$\Delta_t = [e_{1t} - e_{2t}]$$

and

$$\Sigma_t = [e_{1t} + e_{2t}].$$

Equation [1] can be re-expressed as

$$[2] \quad MSE(e_1) - MSE(e_2) = [Cov(\Delta,\Sigma)] + [m(e_1)^2 - m(e_2)^2],$$

where Cov represents the covariance between the difference of errors (Δ), and the sum of errors (Σ) over the out-of-sample period. The hypothesis test of interest is then that both on the righthand side of [2] equal zero.

Appendix A

Given two alternative methods of forecasting X_t, the problem is to determine whether method 1 (giving T forecasts P^1_t) is better than method 2 (giving T forecasts P^2_t). The evaluation is done in terms of out-of-sample mean squared errors. From the definition of mean squared error as the variance plus the bias squared, the difference between two mean squared errors can be written as:

$$[1] \quad MSE(e_1) - MSE(e_2) = [S(e_1)^2 - S(e_2)^2] + [m(e_1)^2 - m(e_2)^2],$$

where e_i is the error vector of out-of-sample errors $e_1' = [e_{i1}, E_{i2}, E_{i3}, \ldots, e_{iT}]$, i = 1,2; $S(e_i)^2$ is the sample variance of out-of-sample errors from forecast method i; and $m(e_i)$ is the sample mean of out-of-sample forecast errors from method e_i. Define

$$\Delta_t = [e_{1t} - e_{2t}]$$

and

$$\Sigma_t = [e_{1t} + e_{2t}].$$

Equation [1] can be re-expressed as

$$[2] \quad MSE(e_1) - MSE(e_2) = [Cov(\Delta,\Sigma)] + [m(e_1)^2 - m(e_2)^2],$$

where Cov represents the covariance between the difference of errors (Δ), and the sum of errors (Σ) over the out-of-sample period. The hypothesis test of interest is then that both on the righthand side of [2] equal zero.

The test can be performed using ordinary least squares regression. Consider the regression equation:

$$[3] \quad \Delta_t = \beta_0 + \beta_1[\Sigma_t - m(\Sigma)] + U_t$$

where $m(\Sigma) = \sum_{i=1}^{T} \Sigma_t/T$, and U_t is a white noise disturbance. The least squares operation obtains:

$$\beta_0 = \sum_{t=1}^{T} e_{1t}/T - \sum_{t=1}^{T} e_{2t}/T$$

$$\beta_1 = \frac{\sum_{t=1}^{T} (e_{1t}+e_{2t}-m(\Sigma))(e_{1t}-e_{2t})}{\sum_{t=1}^{T} (e_{1t}+3_{2t}-m(\Sigma))^2}$$

Then, for $\sum_{t=1}^{T} e_{1t}/T$, $\sum_{t=1}^{T} e_{2t}/T > 0$, the joint test that both bracketed elements on the righthand side of Equation [2] = 0 is equivalent to testing the null hypothesis that $\beta_0 = \beta_1 = 0$ against the alternative that they both are nonnegative and at least one is positive.

For $\sum_{t=1}^{T} e_{1t}/T < 0$ and $\sum_{t=1}^{T} e_{2t}/T > 0$, all errors from forecast method 1 must be multiplied by a minus 1, and the regression Equation [3] is performed with the new error series $e^*_{1t}(e^*_{1t} = -1e_{1t})$. This will, of course, require new vectors Δ^*_t and Σ^*_t formed in an analogous manner. Finally, if both error series have negative means, both series must be multiplied by minus one before performing the hypothesis test.

If either of the two least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, is significantly negative, then the null hypothesis is not rejected and one must conclude that no significant difference exists between the two MSE's. If one estimate is negative but not significantly different from zero, a one-tailed t-test on the other estimate can be used. Finally, if both estimates are positive an F-test of the null hypothesis that both population coefficients are zero can be performed. However, the usual F-test is four-tailed, it does not take into account the sign of the coefficients. Under independence of β_0 and β_1 (which is true in this case), the probability of obtaining an F-statistic greater than F_0 and having both estimates positive is equal to one-fourth the significance level of F_0. (The authors express appreciation to R. Ashley for helpful discussion in regard to this test.)