Evaluating a Model by Forecast Performance

Michael P. Clements and David F. Hendry*
Department of Economics, University of Warwick
and
Department of Economics, University of Oxford

August 12, 2003

Abstract

Although out-of-sample forecast performance is often deemed to be the 'gold standard' of evaluation, it is not in fact a good yardstick for evaluating models. The arguments are illustrated with reference to a recent paper by Carruth, Hooker and Oswald (1998), who suggest that the good dynamic forecasts of their model support the efficiency-wage theory on which it is based.

Journal of Economic Literature classification: C53.
Keywords: Dynamic forecasts, model evaluation.

1 Introduction

Out-of-sample forecast performance is often viewed as the acid test of an econometric model. When that model is based on a well-articulated economic theory, a 'good' out-of-sample performance is then assumed to provide support for the theory. 'Good' can be assessed both in comparison to rival (often naive) forecasts, and relative to in-sample performance. The sentiment that a good forecasting performance constitutes a 'seal of approval' to the empirical model, and therefore of the theory on which the model is based, is general. Two examples suffice. First:

> any inflation forecasting model based on some hypothesized relationship cannot be considered a useful guide for policy if its forecasts are no more accurate than such a simple atheoretical forecast (namely, next year's inflation will equal last year's). Atkeson and Ohanian (2001)

Secondly, in a study of US unemployment:

> If a dynamic modeling approach is to be convincing, it needs to say something about the behavior of unemployment out of sample. Carruth, Hooker and Oswald (1998, p. 626)

*We thank Andrew Oswald and James Stock for helpful comments, and Alan Carruth for kindly providing the data used by Carruth, Hooker and Oswald (1998). Computations were performed using Givewin 2 and PcGive 10: see Doornik and Hendry (2001), and code written in the Gauss Programming Language.

While explanations for forecast success are rarely sought, beyond claiming corroboration of the underlying theory, a poor performance usually requires rationalization. Forecast failure is defined here as a significant deterioration in forecast performance relative to its anticipated outcome, usually based on historical performance. Explanations for forecast failure might point to the occurrence of extraneous or atypical events in the forecast period, or might lead to a deeper questioning of a model, such as that occasioned by the failure of the 1970's Keynesian income-expenditure models to account for the simultaneous occurrence of high inflation and unemployment rates.

Despite the intuitive appeal of these views, six dichotomies intrude on any forecast evaluation exercises:

1. unconditional versus conditional, models;
2. internal versus external standards;
3. checking constancy versus adventitious significance;
4. ex ante versus ex post evaluation;
5. 1-step versus multi-horizon forecasts; and
6. in-sample fixed coefficients versus continuous updating.

These six dichotomies relate to the type of model, method of forecasting, and method of forecast evaluation. Whichever branch one chooses will in general have important implications for what can be learnt from the forecasting exercise about the validity or usefulness of the model and the theory on which it is based. Moreover, there may be important interactions, as we note below following a detailed consideration of their implications in section 2. At least as important will be a dichotomy relating to the nature of the economic environment: whether the environment is stationary or non-stationary (in the sense of changing moments). For example, Stock and Watson (1996) show that instability is a feature of many macroeconomic time series and time-series relationships, and Clements and Hendry (1999) examine the effects of structural shifts on forecast performance.

The point of our paper is that the combined impact of the six dichotomies, and the underlying instability inherent in many economic relationships, vitiates any simplistic claims about the supremacy of forecast performance, or forecast tests, in model evaluation.

The structure of the paper is as follows. Section 2 briefly reviews forecasting in stationary versus non-stationary environments, and considers the impact of the six dichotomies on judging a model by its forecasting performance. Then, in the light of that analysis, section 3 considers the forecasting exercise in Carruth et al. (1998) in some detail. Section 4 provides an illustrative Monte Carlo study of the role of cointegration (one aspect of the empirical study) in explaining 'good' out-of-sample forecast performance. Section 5 discusses the implications of our analysis, and section 6 concludes.

2 Six dichotomies

We consider the six dichotomies in turn, but first note some of the implications of non-stationarity.

In stationary processes, as shown in Miller (1978) and Hendry (1979), forecasts on average will be as accurate unconditionally as expected. As a consequence, internal standards of

comparison (section 2.2) may lack power.[1] If the forecast-period data happen to be quiescent, a lack of failure is not persuasive evidence of a sound theory: for example, interest rates may be a crucial omitted variable in an investment equation, but happen not to change over the sample, so the false model is not rejected. Idiosyncratic factors could be such that the good performance of a model lends little support to any given theory.

Conversely, poor models may not be rejected. If the process is highly non-stationary (with changing moments), a lack of failure may merely reflect the use of an adaptive tracking device with a large forecast variance, such that systematic mis-forecasting cannot happen with most individual outcomes being consistent with the forecasts (even though the forecasts may be 'poor' in the sense of having low correlation with the outcomes). Such models may do well on internal standards (and even relative to rival models), but have no significant economic predictive power. Worse still, 'good' models may fare poorly in terms of out-of-sample performance when there are location shifts (such as in an equilibrium mean) near the forecast origin (see, e.g., Clements and Hendry, 1999, pp.307–9, for a brief review).

2.1 Unconditional versus conditional models

The first dichotomy relates to whether or not all variables are modeled. Unconditional models endogenize all variables (as in a vector autoregression, denoted VAR), whereas conditional models treat some variables as given. For statistical inference, only weak exogeneity is required to sustain the conditioning on the non-modeled variables, but for forecasting more than one step ahead, strong exogeneity is essential (see Engle, Hendry and Richard, 1983, and Ericsson, 1992 for an exposition). The distortionary effects of invalid conditioning on multi-step forecast performance are compounded when non-modeled dynamics partially substitute for modeled-variable dynamics (see section 2.5 a) such that the performance of the model forecasts may appear more accurate than they actually are.

One way round the difficulties entailed by conditioning is to require that all the variables are 'endogenized' in any forecasting exercise. However, this amounts to requiring that a theory should specify generating mechanisms for the complete vector of variables under study: otherwise the forecast performance of the system as a whole depends upon equations for variables which are not specified by the theory. That would leave little room for the large class of theory models which are partial, in the sense of making conditional statements relating economic variables. Consequently, we focus on conditional models hereafter, and return to the issue of exogeneity in the context of the empirical example.

2.2 Internal versus external standards

Internal standards denote that forecasts are evaluated relative to the estimated model (e.g., against its forecast standard errors), whereas external standards involve a comparison with other, usually mechanistic, forecasts. Internal evaluations—assuming that inferences are in fact validly based under the null—tend to be tests of parameter constancy.

[1] Consistent with the view that corroboration may be too easy, Mayo and Spanos (2000) quote: 'mere supporting instances are as a rule too cheap to be worth having; they can always be had for the asking; thus they cannot carry any weight; and any support capable of carrying weight can only rest upon ingenious tests, undertaken with the aim of refuting our hypothesis, if it can be refuted' (Popper, 1983, p. 130).

External evaluation is more in the nature of an encompassing test against a rival specification as in the above quote from Atkeson and Ohanian (2001), who deem failure against an atheoretical contender decisive. As ever, the issue is rather less clear-cut in practice: forecast failure could, but need not, impugn a policy model; could, but need not, be 'camouflaged' by a variety of devices (of which the best know are intercept corrections); and a 'naive' model could, but need not, be robust to location shifts which are pernicious for an econometric model (see e.g., Hendry and Mizon, 2000). Thus, we consider both internal and external evaluation below.

2.3 Checking constancy versus adventitious significance

The third dichotomy concerns the purpose of the evaluation exercise. One common aim is to check whether selected variables are really relevant or just significant by chance. A rather different objective is to test parameter constancy over a 'forecast period' (the nature of which will be addressed in the next sub-section). We take these in turn.

First, sub-sample evaluation is often used as a 'hold-back check' against spurious significance, as in the data-mining literature (see e.g., Hoover and Perez, 1999). This approach is based on the idea that an estimated coefficient is unlikely to be significant by chance in two sub-samples as well as the whole sample. Thus, various authors have proposed selecting which variables to include in a model by a rule of consistent significance across sub-samples. Doing so certainly lowers the implicit significance level of the selection procedure, because under the null, significance in several sub-samples is indeed less likely to occur by chance. However, it also lowers the power to retain those variables that genuinely matter. Lynch and Vital-Ahuja (1998) and Krolzig and Hendry (2003) show that a higher power at any desired significance level can in fact be achieved by a single full-sample procedure, so 'hold-back' is an inefficient device for evaluating a model for 'data mining'.

Nevertheless, conditional on having selected a model, a few post-sample observations can be dramatically more helpful than a large number of in-sample data points in discriminating between substantive and adventitious significance. This occurs because conditional on a significant outcome being observed on the full sample, no split of the sample can discriminate between that being due to chance or because the null is really false. However, out-of-sample, t-tests on explanatory variables with non-zero coefficients have increasing non-centralities as the number of new observations grows, so their significance increases on average, whereas t-tests on the estimated coefficients of irrelevant variables converge towards zero.

The key difference between these two uses of the sub-samples is that the first is employed to select the model, whereas the second tests an already selected model on previously unseen data. Since we are concerned with evaluating models by their forecast performance, we will now focus on the use of sub-samples to check parameter constancy, and turn to doing so using existing or new data.

2.4 Ex ante versus ex post evaluation

The fourth dichotomy contrasts whether the 'forecasts' are genuinely made before the outcomes have occurred, and evaluated at a later stage when the outcomes are known (ex ante), or are evaluated against a sub-set of the originally available data 'retained for in-sample forecasts' (ex post).

Ex ante evaluation has high face validity, and could constitute a Neyman–Pearson 'quality control' test with high content validity: a few observations after a model is developed can provide definitive evidence on its performance. Nevertheless, considerable care is needed in interpreting outcomes of either success or failure. First, Clements and Hendry (1998) use the analogy of a serious forecast failure when a spacecraft to the moon is knocked off course by a meteor, to emphasize that there need be no implications for the underlying theory—here that based on Newton's laws. Indeed, it need not even reflect badly on the soundness of the forecasting algorithms, which probably quickly corrected their forecasts. Secondly, Hendry (1996) shows that ex ante forecast failure is consistent with ex post parameter constancy in some situations (particularly when there changes in data measurement: see also Patterson, 2003). Thirdly, we have already alluded to the fact that the evaluation may lack power. Finally, ex ante evaluation is only possible in conditional models—the focus here—by using a sequence of 1-step forecasts.

Ex post evaluation tests the key attribute of parameter constancy, and although in-sample tests can be conducted using the Lagrange-multiplier approach, a formal hold-back sample is often used. Such ex post evaluation usually assumes a division of the sample into an in-sample and out-of-sample period. Suppose there are T observations in total, indexed by $t = 1, \ldots, T$, and that the initial estimation period is $t = 1, \ldots, R$, with H 1-step ahead forecasts for $R + 1$ through to $R + H$ (so $R + H = T$). West (1996), for example, discusses fixed, recursive and rolling forecasting schemes, depending on whether model parameters are estimated on data up to R, and then held fixed for the calculation of all the forecasts, estimated on an expanding window of data ($t = 1$ to R, $t = 1$ to $R + 1$, etc), or estimated on a fixed window that moves through the sample ($t = 1$ to R, $t = 2$ to $R + 1$, etc).We discuss recursive forecasting (simply referred to as updating) versus fixed schemes in section 2.6.

In ex post evaluation, 'data-snooping' is hard to exclude: an investigator may have checked the performance of a variety of models on the 'hold-back' sample, and only reported those that happen to be constant. For example, Boughton (1992) and Hendry and Starr (1993) show that constancy can be designed as part of the process of modelling. That potential problem is almost certainly why investigators deem ex ante evaluation to be preferable, taking us rapidly in a circle. Subject to the caveats just made about careful interpretation, the conclusion is perhaps to support the use of 'blind hold-back', as in computer learning competitions, where some of the sample is not available to the modeler, which then jointly evaluates adventitious significance and parameter constancy. That raises the issue of precisely how to use the information in such a sample, which is addressed in the next two sub-sections, now further restricting our analysis to ex post evaluation of conditional models.

2.5 1-step versus multi-horizon forecasts

1-step ex post forecast evaluation in a conditional model takes the regressors at their observed values and compares the forecasts with realized values, using internal standards. Appropriate tests with powers against different alternatives are widely available, and some at least seem to have good operating characteristics: see inter alia, Chow (1960), Andrews (1993).

However, there are a number of reasons why multi-step ex post 'forecast' performance (dynamic simulation), may not be a good guide to the credence to be attached to a model:

a. The first was stressed by Chong and Hendry (1986) – 'what dynamic simulation tracking accuracy mainly reflects is the extent to which the explanation of the data is attributed

to non-modelled variables' (italics in original). In many instances, the researcher has some latitude in choosing the mix of lags of dependent and explanatory variables. If the latter are unmodeled, and so replaced by their actual values in dynamic 'forecasting' or simulation exercises, apparent performance may be improved by reducing the role of own-variable dynamics.[2]

b. The second reason relates specifically to equilibrium-correction (EC) models, which embody long-run relations suggested by economic theory, where a good forecast performance—especially over long horizons—may be thought to lend support to the theory embodied in the EC term. Analytical calculations in Clements and Hendry (1995) show that, in general, this assertion is incorrect, in that long-horizon forecasts of the changes in the variables are no more accurate than forecasts from models which omit the EC terms. Given the importance often attributed to EC terms, especially for long-horizon forecasts, section 4 considers this issue in the context of forecasting the US unemployment rate.

c. The third reason is that in any particular instance there may be idiosyncratic factors at work, such that the good dynamic performance of the model actually lends little support to the theory: this is also addressed in section 4.

d. Finally, the appropriate variance matrix for evaluating multi-step ex post forecasts is a function of the 1-step, and there is no real benefit in the exercise when properly conducted: see Pagan (1989).

Sections 3 and 4 show how these explanations – of why multi-step ex post forecast performance may have little bearing on the validity of an economic theory – cast doubt on the conclusions drawn from the results of the out-of-sample forecast exercise reported by Carruth et al. (1998). In section 6 we also comment on the role of non-stationarity, more extensively discussed in Clements and Hendry (1999).

Accurate dynamic simulations many steps ahead are tantamount to owning a crystal ball: the ability to forecast unemployment a decade or more ahead more accurately than its unconditional distribution (even if that was stationary) entails knowledge of a vast range of events, laws and circumstances that were undreamt of at the time the forecast was supposed to be made (especially from 1978Q2).

2.6 Fixed coefficients versus updating

Finally, an investigator can choose between a number of forecasting schemes. Estimated coefficients could be held fixed at their in-sample values for the whole ex post forecast evaluation horizon, or updated recursively as the period unfolds. The nature of their multi-step evaluation constrains Carruth et al. (1998, p. 626) to use in-sample fixed coefficients. However, the outcomes of an empirical forecast comparison exercise can depend on whether model coefficients are continuously updated or are held fixed at in-sample values, especially when there are nonconstancies. Models that are robust to location shifts will have a relative advantage for fixed coefficients (see Eitrheim, Husebø and Nymoen, 1999): continuously updating may blunten this

[2] There is no suggestion that considerations of this sort played a part in the specification of the efficiency-wage model discussed below.

edge, consistent with the success of re-selecting the model specification and re-estimating as new information accrues (see e.g., Phillips, 1994, 1995, 1996, and Swanson and White, 1997). Although Carruth et al. (1998, p. 626) use a fixed-coefficients scheme, we will also look at the impact of updating.

In general, the interactions between these dichotomies also matter: for example, conditional multi-horizon forecasts require known future values of unmodeled variables and fixed coefficients; ex post evaluation on purely internal criteria at best checks for parameter constancy, rather than forecast performance; and so on.

3 An out-of-sample forecasting exercise

An implication of the efficiency-wage model developed by Carruth et al. (1998, p. 626) (henceforth CHO) is that in the long-run, or static equilibrium, unemployment should respond to input prices. For input prices, they consider real oil prices and the real rate of interest. They establish that the unemployment rate and the two input price variables are individually integrated of order one, but that a linear combination of these three variables is integrated of order zero, that is, the variables are cointegrated. The signs of the estimated coefficients in the cointegrating combination are consistent with their theory: higher input prices lead to a higher unemployment rate in the long run. An out-of-sample forecasting exercise over a sixteen year period from 1979 to 1995 is viewed as an 'exceptionally hard examination' (CHO, p.626) and a dynamic equilibrium-correction model is judged to fare reasonably well.

The plots of the data are shown in figure 1, panels a and b. The real oil price and unemployment rate variables match quite closely over the post-War period. The coefficients of the cointegrating regression are not readily interpretable because the data are in levels (not logs), but evaluating the elasticities at the sample means gives 0.36 and 0.06 for real oil prices and real interest rates respectively, indicating that 10% increases in the real input price variables will result in increases of 3.6% and 0.6% in the unemployment rate in the long run.[3] The greater magnitude of the response to the real oil price reflects the greater visual match between the series.

Figure 2 shows dynamic forecasts of the annual change in the unemployment rate based on the estimation period 1955Q4–1978Q4, reproducing CHO Figure 2A. The model estimates are given in table 2 in the Appendix. The forecasts use actual values of all the explanatory variables save for the lagged dependent variables, which are replaced by recursively computed predictions. We assume the lagged dependent variable component of the EC term is known, corresponding to CHO Approach A. The forecasts appear to track the actual course of annual changes reasonably well. In addition to the point forecasts, we display error fans which allow for both error and parameter estimation uncertainty. These show that the periods for which CHO identify large forecast errors generally correspond to the actual values lying outside a two-sided 95% forecast interval.

Although not reported by CHO, the long-run coefficient on the oil price in the equilibrium-correction term is quite different when the estimation period ends in 1978: see tables 2 and 3. The estimated coefficient is nearly three times larger than the full-sample estimate, implying an elasticity (evaluated at the sub-sample data means) of 0.94, so that close to a 10% change

[3] Table 3 records the full-sample estimates.

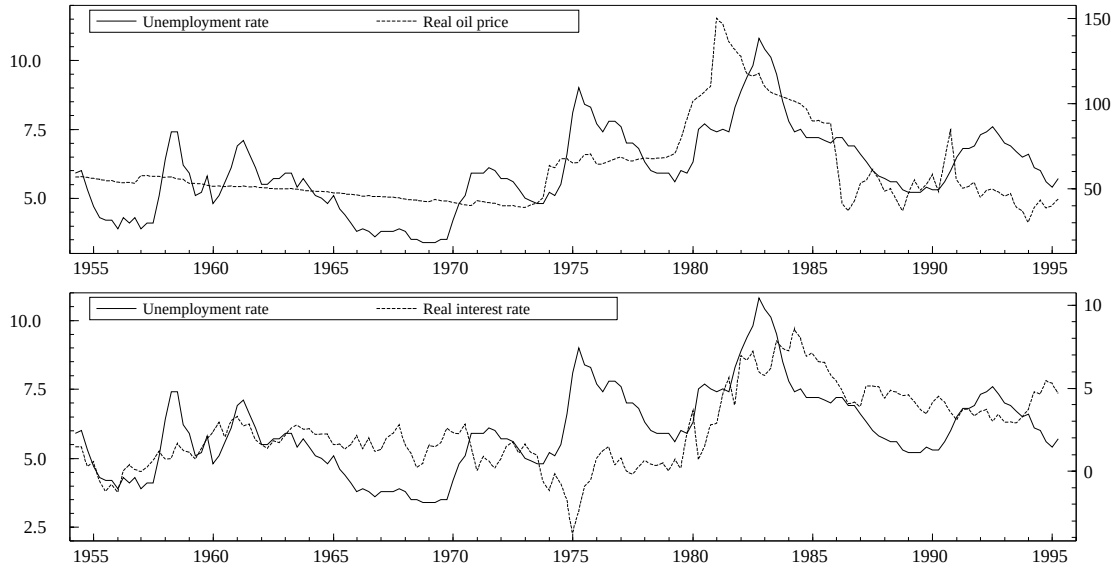

Figure 1: Time series plots of the unemployment rate and real input price series. Uses the CHO data. The unemployment rate series is reproduced in each panel for ease of comparison. The input price series are plotted against the right-hand vertical axis

in the unemployment rate results from a 10% change in the oil price in the long run. The magnitude of the change in the long-run elasticity (between the two samples) suggests that it may be possible to improve upon the assumed linear relationship between oil prices and the unemployment rate, as in the recent literature relating output growth to oil price changes, but that is beyond our immediate concern of what can be learnt from the out-of-sample forecasting exercise.[4] Figure 3 records the 'equilibrium errors' based on the full and sub-sample estimates of the cointegrating regressions. They mainly differ at the times of high levels of oil prices in the early eighties and nineties.

Using the full-sample estimates of the cointegrating relationship to produce the forecasts results in figure 4, which otherwise match figure 2. The model estimates are recorded in table 3. The forecasts based on the full-sample estimates of the long run are more accurate, as might be expected, because by construction the full-sample estimates reflect the long-run relationship over the forecast period. The first two rows of table 1 record summary forecast-error statistics for the fixed-coefficient dynamic forecasts displayed in figures 2 (first row of the table) and 4 (second row, headed CHO_{FS}, to denote the use of full-sample estimates of the long-run relationship), and show an approximate 40% reduction in the RMSE. The use of the forecast period to estimate the long run is of course illegitimate from the perspective of an out-of-sample forecasting exercise (as is CHO's use of the full sample to obtain the model specification) but

[4]Hooker (1996) shows that the linear relationship between oil prices and output growth of Hamilton (1983) does not appear to hold from 1973 onwards. Hamilton (1996) suggests that output should be related to the net increase in oil prices over the previous year. Given sustitution in production and consumption as input prices change, a constant coefficient relation seems unlikely.

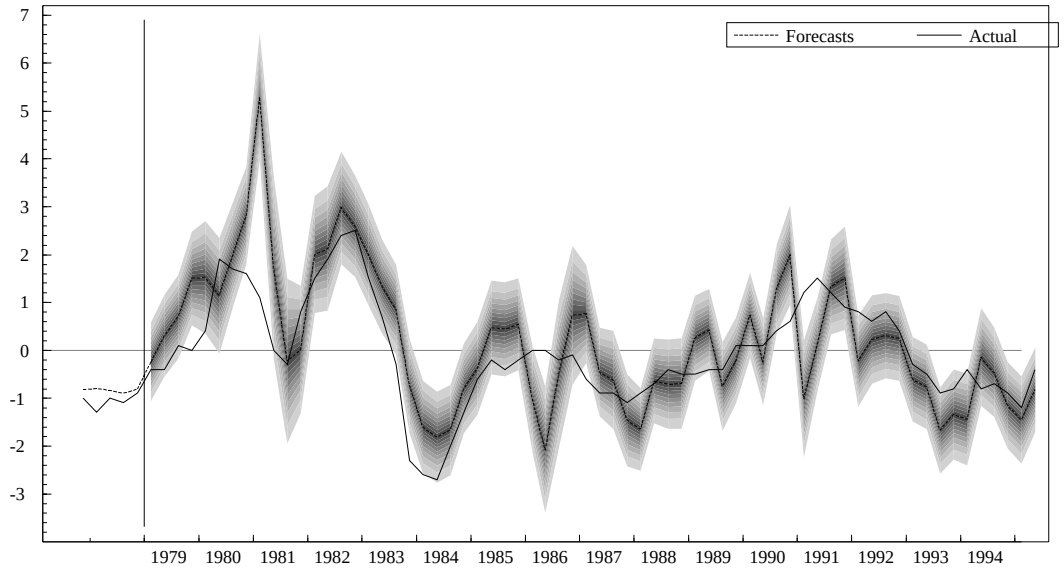

Figure 2: Forecasts of the annual change in the unemployment rate from col. 2 of Table 3 of CHO, with the ECM estimated on the sub-sample. The forecasts are generated using CHO Approach A, i.e., actual values of the real oil, interest rates and equilibrium correction variables are used for the explanatory variables, though lagged unemployment rate changes are predictions. The error fans allow for parameter estimation uncertainty

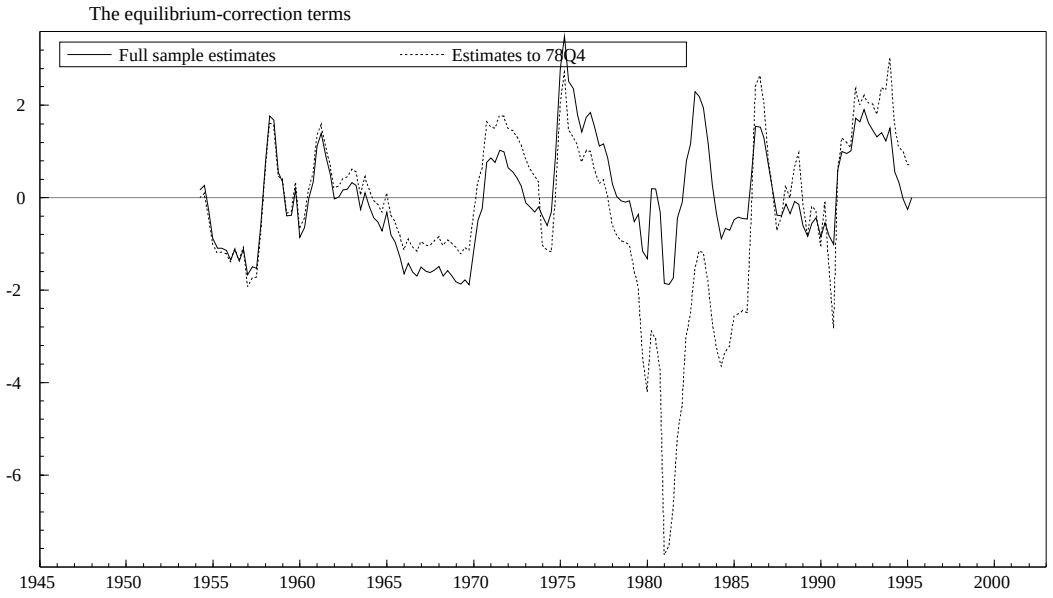

Figure 3: The equilibrium error based on full sample estimates and estimates up to 78Q4

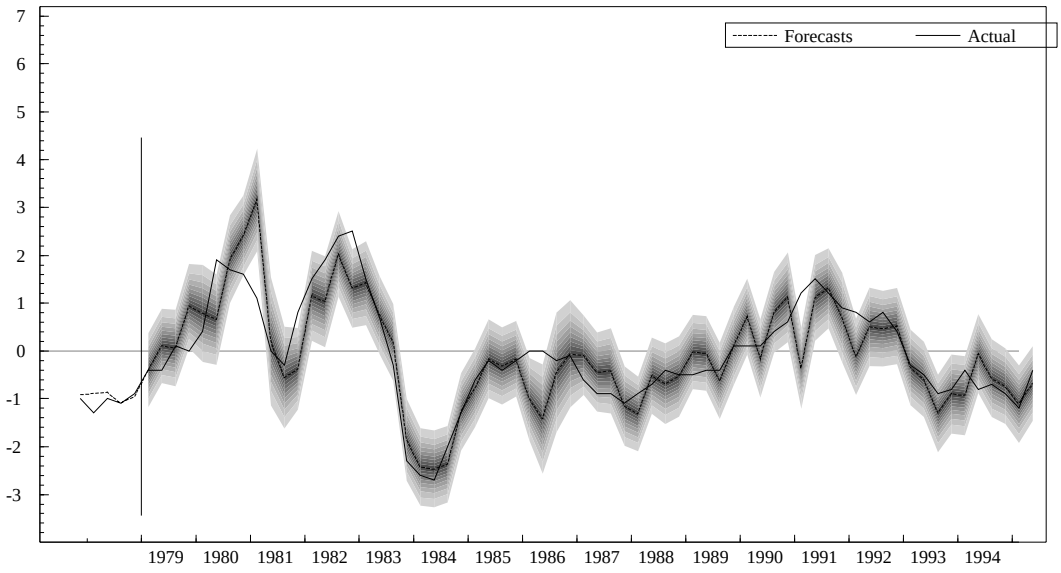

Figure 4: Forecasts of the annual change in the unemployment rate using full-sample estimates of the long-run relationship. The forecasts are generated using CHO Approach A, i.e., actual values of the real oil, interest rates and equilibrium correction variables are used for the explanatory variables, though lagged unemployment rate changes are predictions. The error fans allow for parameter estimation uncertainty

by improving the forecast performance of the model would appear to work in CHO's favour, strengthening the support for the theory.

A comparison of figures 2 and 4 shows that using full-sample estimates of the long-run approximately halves the 4 point over-prediction of the unemployment rate in 1981Q1. The 'proper' sub-sample model (CHO) has a dynamic contemporaneous response to oil price changes which is nearly twice as large as in CHO_{FS}, and the 48% increase in the real oil price in 81Q1 on a year earlier leads to a predicted increase in the unemployment rate far in excess of that which materialized.[5]

A closer examination of the CHO_{FS} specification on the first sub-sample suggests some simplifications of the model may be possible, and we consider the effects of these on forecast performance. The two lagged dependent variable terms are not significant on the sub-period, and the small magnitude of the estimated coefficients suggests their effect on the forecasts is minimal. Also, the EC term enters with a coefficient close to unity at lag 1, and in excess of -1 at lag 4. An F-test restricting these two coefficients to 1 and -1 respectively yielded a p-value of 0.0645, and if in addition we simultaneously restrict the coefficients on the lagged dependent variables to zero, we obtain an $F_{4,86}$ with a p-value of 0.0324 under the null. This restriction is not significant at the 1% level, indicating that such a restricted model might be a reasonable approximation to the CHO model. Figure 5 records the forecasts from this model. A visual comparison between figures 4 and 5 suggests that the models are indeed similar, and this is borne out by the forecast-error summary statistics reported in table 1 (compare the rows CHO_1–Eqn. (1)–and CHO_{FS}). The form of the restricted model is revealing. Algebraically:

$$\Delta_4 U_t = \quad + \alpha_{RO}\Delta_4 RO_t + \alpha_{RR}\Delta_4 RR_t + \Delta_3 \left(U_{t-1} - \beta_{RO} RO_{t-1} - \beta_{RR} RR_{t-1}\right) + error \qquad (1)$$

where L is the lag operator, $\Delta_j = 1 - L^j$; U, RO and RR are the unemployment rate, real oil price, and real interest rate variables, respectively, and the cointegrating vector is $[1 : -\beta_{RO} : -\beta_{RR}]'$, with the cointegrating regression estimates such that $\beta_{RO} > 0$ and $\beta_{RR} > 0$. This model ostensibly relates the annual change in the unemployment rate to annual changes in the input price variables and EC terms that include lagged values of the unemployment rates. But the lag polynomial on U_t is $\left(1 - L^4 - L\left(1 - L^3\right)\right) = (1 - L)$, so that we obtain:

$$\Delta U_t = \quad + \alpha_{RO}\Delta_4 RO_t + \alpha_{RR}\Delta_4 RR_t - \Delta_3 \left(\beta_{RO} RO_{t-1} + \beta_{RR} RR_{t-1}\right) + error. \qquad (2)$$

The quarterly change in unemployment depends only on input prices, and using $\Delta_4 U_t = \sum_{s=0}^{3} \Delta U_{t-s}$, the restricted model predictions of the annual changes in the unemployment rate are also determined solely by contemporaneous and lagged values of variables other than the dependent variable. Therefore, although CHO stress that "in neither approach are the lagged dependent variables updated with actual unemployment" (CHO, p. 626) – where approach A uses the actual EC values, and approach B replaces the lagged dependent variable in the EC by a prediction – this claim lacks force. We have shown that in a restricted version of their model, which has a superior performance in terms of forecast bias and RMSE, there are no lagged

[5] CHO_{FS} exhibits a smaller long-run response as well, but the impact of this is hard to disentangle because the EC terms enter with a positive coefficient at lag one, and a negative (and larger in absolute magnitude) coefficient at lag four. The dynamic and long-run effects of oil on unemployment vary in a complicated, interrelated fashion in the CHO model between the estimation and full-sample periods, but imposing the full-sample long-run estimates and estimating the model up to 78Q4, as in CHO_{FS}, reduces the general dependence on oil and produces more accurate forecasts.

dependent variables. This means that actual values of all the explanatory variables are being used. Because the contemporaneous values of real oil price and interest rates are explanatory variables, the 'dynamic' forecasts are not even proper 1-step ahead forecasts, given that the forecast of period t uses information dated period t.

3.1 Evaluating the forecasts

In terms of the framework introduced above, now consider the six dichotomies.

1. Unconditional versus conditional models.

The model is conditional on two regressors, and the forecasts on their known future values. The validity of conditioning on these variables for carrying out statistical inference rests upon the weak exogeneity of these variables, which is certainly not unreasonable even if not guaranteed. But valid forecasting requires strong exogeneity. One aspect of strong exogeneity is easily tested, namely, Granger non-causality from unemployment to oil prices and interest rates (see Granger (1969)). Granger non-causality is a necessary though not sufficient condition for strong exogeneity, such that rejection would reject the validity of the conditioning for forecasting. Testing causality in a VAR(5) for the three variables yields non-rejection; allowing contemporaneous unemployment to affect interest rates, however, leads to rejection of the null (but does not for oil prices).

2. Internal versus external standards.

As an internal standard of assessment, the model based on known future input price variables tracks the actual evolution of annual changes in the unemployment rate 'reasonably well' as judged by the graph. As an external standard, we follow the long-standing tradition in the forecast evaluation literature of comparing econometric model forecasts to those from time-series models of the Box–Jenkins type, see, for example, Nelson (1972) and Granger and Newbold (1975). Comparisons of this sort ask whether the explanatory variables contribute to more accurate forecasts than forecasts based on the history of the variable alone. Because the economic model forecasts are at best 1-step ahead forecasts, we generate a sequence of 1-step forecasts from a second-order autoregressive model of $\Delta_4 U_t$, using fixed coefficients. These are depicted in figure 6. They compare favorably with the econometric model forecasts. From table 1 (the row labelled AR(2)) the forecast bias of the AR model is similar to that of CHO_1, the restricted CHO_{FS} model, and the RMSE is some 20% smaller.

3. Checking constancy versus adventitious significance.

The main use of the sub-samples in CHO was to check parameter constancy, rather than the significance of the selected variables in sub-samples, so we have focused on that aspect.

4. Ex post and ex ante evaluation.

All the evaluations offered are ex post rather than ex ante. It would now be possible to undertake an ex ante evaluation using the data for the second half of the 1990s that has become available subsequent to the development of the efficiency-wage model. Care would need to be taken to ensure that the results were not affected by data revisions (e.g., Patterson (2003)).

5. 1-step versus multi-step.

The first and third explanations for why dynamic forecast performance does not lend much support to the theory are relevant and reinforce each other here. Despite the initial appearance of the model, the model forecasts are not truly 'dynamic': the model can be restricted to a version in which lagged dependent variables do not appear. This aspect is the 'idiosyncratic

explanation', because it will not in general be true. When coupled with generating 'dynamic forecasts' by replacing the unmodeled variables by their actual values, then in the restricted version of the model, CHO_1, the variation in the dependent variable is explained wholly by unmodeled variables whose future values are assumed known in the forecasting exercise. Thus, the CHO_1 forecasts are 1-step, in that they make use of information on the unemployment rate in the period immediately prior to that being forecasted (as well as the values of input prices in the same period as that being forecasted).

6. Updating versus fixed coefficients.

To isolate the impact of updating the model parameter estimates, we first produce forecasts from a variant of CHO_1 which provides a more suitable benchmark. This is denoted in table 1 as CHO_1^*. Recall that the estimates of the long-run parameters that feed into CHO_1 were based on the full-sample up to 1995. CHO_1^* has the same explanatory variables as CHO_1, but the estimates are obtained by freely estimating the model on the initial sample period. The parameters are held fixed at these values for forecasting, so that the statistics recorded in the table for CHO_1^* are for a fixed forecasting scheme. The relative gain of the AR(2) over the efficiency-wage model (now represented by CHO_1^*) is much reduced, but still apparent. The rows labelled $\mathrm{AR}(2)_u$ and CHO_{1u}^* report the results of producing forecasts from these two models based on continuously updating the parameter estimates. The AR model is barely affected (there are no apparent differences to two decimal places), although there is a marked improvement in the accuracy of the efficiency-wage model. Clearly, the way in which the forecasting exercise is conducted can give rise to markedly different results in terms of the support accorded to the theory on which the model is based.

In section 2.6, we noted that models which are robust to location shifts will have a relative advantage for fixed coefficients, and that updating may blunten this edge. An explanation for our empirical findings is that the AR(2) enjoys a certain amount of robustness, in part due to its specification in (fourth) differences. The efficiency-wage model is less robust, but by effectively being a model in differences (the EC term is absent from the restricted version of the CHO model, which nevertheless closely approximates their model), this too is more robust to locations shifts than would be models with important EC terms. Paradoxically, therefore, the improved forecasting success of this version now detracts from the economic theoretical basis of all the models considered.

4 Monte Carlo analysis of the relevance of EC terms

The efficiency-wage theory implies an equilibrium relationship between the unemployment rate and input prices. Is it the case, therefore, that cointegration, as embodied in the EC terms, explains the 'good' out-of-sample forecast performance, especially a decade and a half out? Clements and Hendry (1995) show that cointegration is not an important determinant of long-horizon forecast performance, assessed by RMSE, unless one is interested in forecasting the stationary equilibrium combination of the variables.

To demonstrate the force of that argument here, a small Monte Carlo study was undertaken. First, we use the CHO_{SS} model given in col. 2 of their Table 3 as the data generating process. Because CHO's analysis is single-equation, we condition on the actual values of the oil and

Table 1: Multi-step dynamic forecast error summary statistics 1979Q1 to 1995Q2 for $\Delta_4 U$

	Mean	RMSE
CHO	−0.22	0.97
CHO_{FS}	0.04	0.59
$\text{CHO}_{\mathtt{1}}$ (Eq.(1))	0.02	0.60
$\text{CHO}^*_{\mathtt{1}}$	0.03	0.47
$\text{CHO}^*_{\mathtt{1}u}$	−0.04	0.38
AR(2)	−0.02	0.45
AR(2)_u	−0.02	0.45

Notes.
CHO is the Carruth, Hooker and Oswald (1998) model, from col. 2 of their Table 3, with forecasts obtained using their Approach A.
CHO_{FS} is as CHO but using full-sample estimates of the cointegrating relationship.
$\text{CHO}_{\mathtt{1}}$ is eqn. (1), i.e., the restricted version of CHO_{FS}.
$\text{CHO}^*_{\mathtt{1}}$ has the same regressors as $\text{CHO}_{\mathtt{1}}$ but with freely estimated parameters (but fixed over the forecast period).
$\text{CHO}^*_{\mathtt{1}u}$ has the same regressors as $\text{CHO}_{\mathtt{1}}$ but the forecasts are generated by a recursive-updating scheme.
AR(2) is the second-order autoregression in the annual change in unemployment.
AR(2)_u is as AR(2), but based on recursively-updated parameter estimates. The AR(2) model forecasts are 1-step ahead.

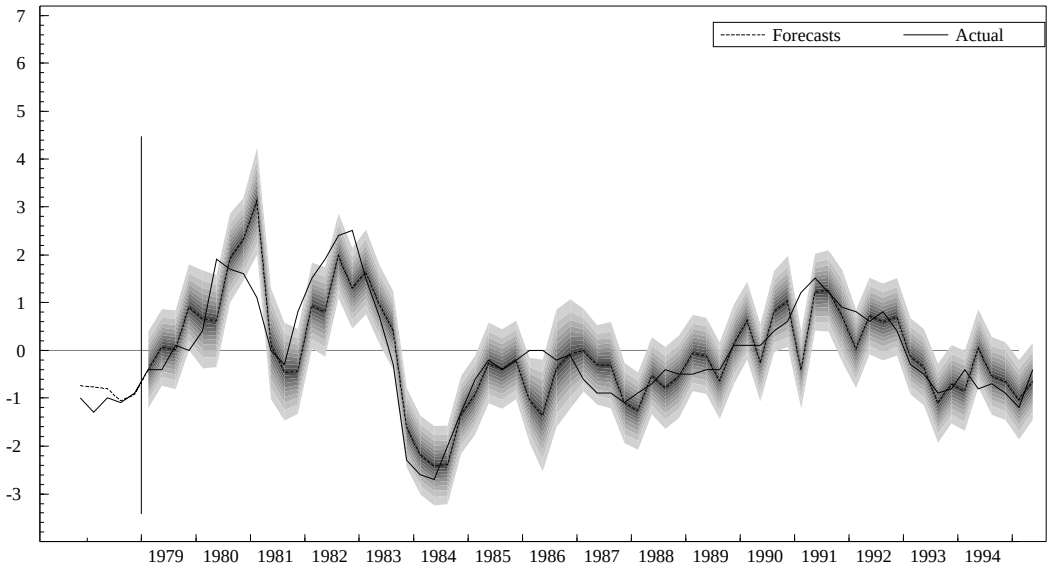

Figure 5: Forecasts of the annual change in the unemployment rate based on the model given by equation 1. The forecasts are generated as in figure 4. The error fans allow for parameter estimation uncertainty

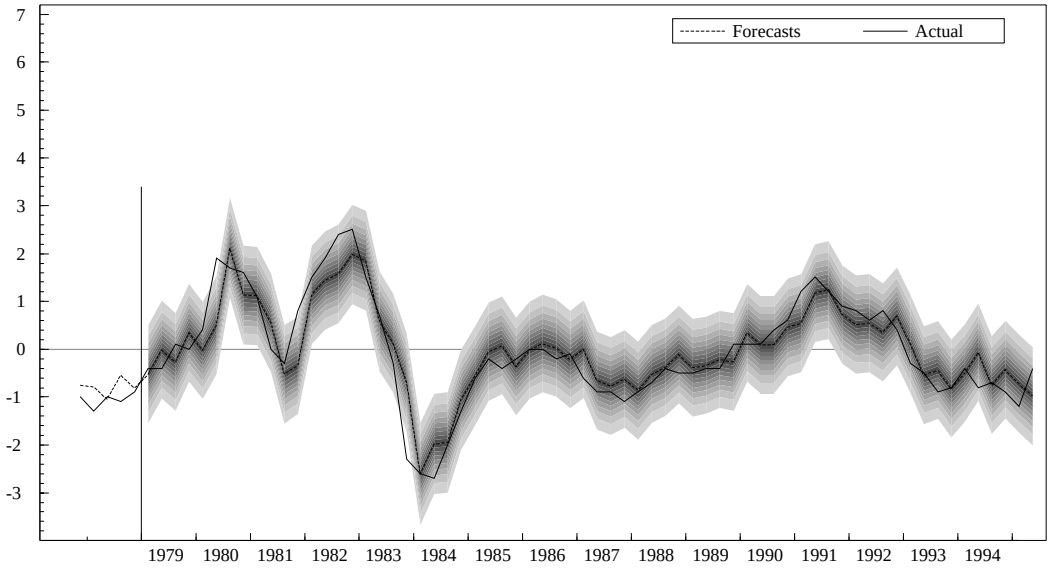

Figure 6: Forecasts of the annual change in the unemployment rate from a second-order autoregression for the annual change in the unemployment rate. The forecasts are 1-step ahead, so that the explanantory variables – the first two lags of the dependent variable – are replaced by actual values. The error fans allow for parameter estimation uncertainty

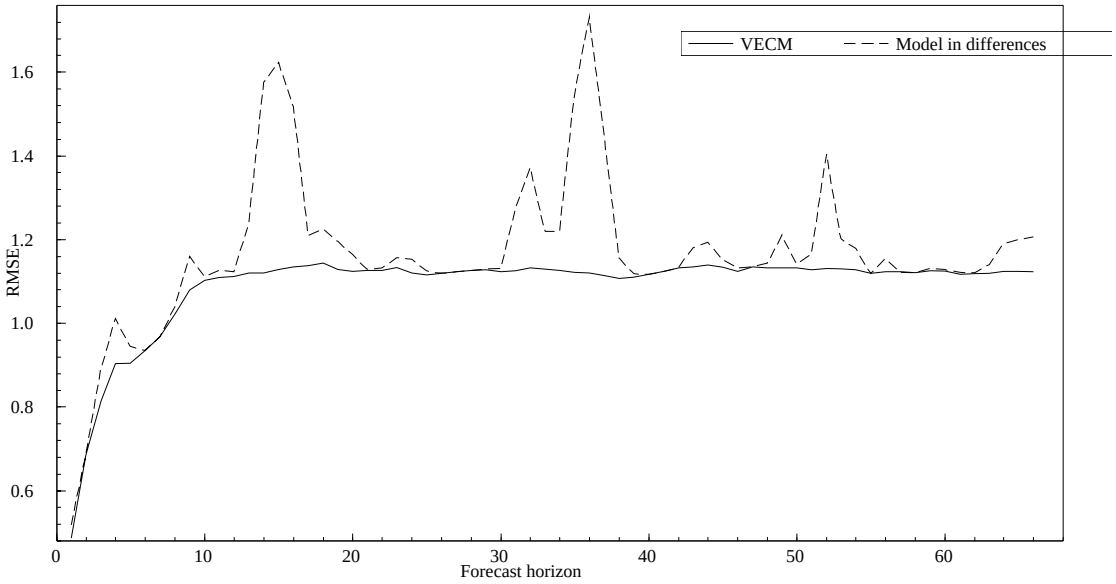

Figure 7: RMSEs obtained by Monte Carlo for the CHO model and a model where the equilibrium-correction term is switched off

interest rate variables rather than specifying equations for these variables, so in all replications, the actual values of these variables are used. Based on the data up to period 1978Q4, we simulate the next 66 values of the unemployment rate, replacing lagged values of the unemployment rate (including in the EC terms) by simulated values as appropriate. The disturbances are given by pseudo-random Gaussian variables with a standard error of 0.49, thereby ignoring the autocorrelation in the estimated model's errors. We then consider forecasting with two models. The first is the CHO model, which uses the correct coefficients and the actual values of the non-modelled variables, but replaces lagged unemployment rate terms (including those in the EC terms) by predictions. There is no model mis-specification nor estimation uncertainty. The second model is expressed entirely in (annual) differences. Estimating a model of this sort on the data to 1978Q4 would not be appropriate because the DGP for the (simulated) forecast data differs from the estimation period actual data, to the extent that it ignores the residual correlation in the fitted empirical EC model. Nor can we simulate a representative sample of data on which to estimate the model in differences, because we do not have equations for the unmodeled variables. Our solution to this problem was as follows. Use the same coefficients as in the EC model (and the DGP), but replace the EC variables by their estimation sample means. Thus the difference model does not exhibit equilibrium-correcting behavior, but in all other respects it matches the EC model.

Series of 1 to 66-step ahead forecasts were generated and compared to the simulated ('actual') values of annual unemployment rate changes on each of 10,000 replications, over which the RMSEs for each step ahead were calculated and are plotted in figure 7. There are some odd departures between the RMSEs of the two models at specific horizons – attributable to the odd dynamics of the model – but overall the EC terms do not appreciably improve the forecasts.

5 Implications

What, therefore, can be learned from forecast performance? Clearly, one learns how well the given model actually forecasts over the specific historical period, absolutely, relative to its earlier (in-sample) behavior, and in comparison to other forecasting devices. But as adverts by financial institutions now always warn: 'past performance is not necessarily a good guide to future performance'. More generally, corroboration is not definitive and rejection is not final in any progressive science: either successful or failed forecast performance is but one item of information in the gestalt needed to appraise both models and theories.

The CHO model may be a useful partial description of the economic relations of interest, even though it fails to predict well. Taking instrumentalism to be the view that 'theories can never be considered to be true or false but merely as instruments of prediction', as in Lawson (1989, p.238), a poor forecast performance is damaging. Evidently, we have a poor instrument. But in a non-stationary environment, that instrument may still be the best available for other purposes. Equally, the realist view would be that the model describes a 'tendency' that in any instance may not be fulfilled because of the interplay of other influences in an 'open system': here the claimed tendency is for unemployment to rise as the real oil price increases, but at high levels of the real oil price other non-modelled forces come to bear (e.g., firms substitute from machinery, now expensive to run, to relatively cheaper labor). Whether or not the model provides a useful partial description of the forces behind the evolution of the unemployment rate is a moot point. Certainly it is unlikely to be a useful description outside the historical period on which it was estimated. We note that unemployment in the US was about the same in the last half of the 19th century as now, but oil prices then were irrelevant.

In principle, one would want to close the system, and model all the forces that have a bearing, and ex post one might be able to make progress in this direction, but failure to do so does not invalidate the theory from a realist perspective.

Finally, it has emerged that whether we adopt fixed coefficients or update the estimates can be decisive. If the model's parameters are constant, more information (as in a recursive scheme) would be expected to provide more precise parameter estimates and improved forecasts, although Clements and Hendry (1998) suggest that improvements emanating from this source are likely to be of secondary importance. The gains of the size we observed from updating point to the non-constancy of the efficiency-wage model parameters. But if the model's parameters are not constant, what are we to conclude about the underlying theory on which the model is based? Since adaptability is key to successful forecasting, updating will improve forecasts even if parameters are not constant and the theory is not relevant. And in terms of what we can learn from forecast exercises, would it not be preferable to check the model's parameters directly using recursive methods? Figure 8 reports Chow tests for the CHO_1^* model where the parameter estimates are estimated recursively over the forecast period. The '1-step' statistics indicate several periods of parameter non-constancy, although the more quiescent later data period offsets this in the break-point tests.

6 Conclusions

Out-of-sample forecast performance is not a reliable indicator of the validity of an empirical model, nor therefore of the economic theory on which the model is based. This is despite the

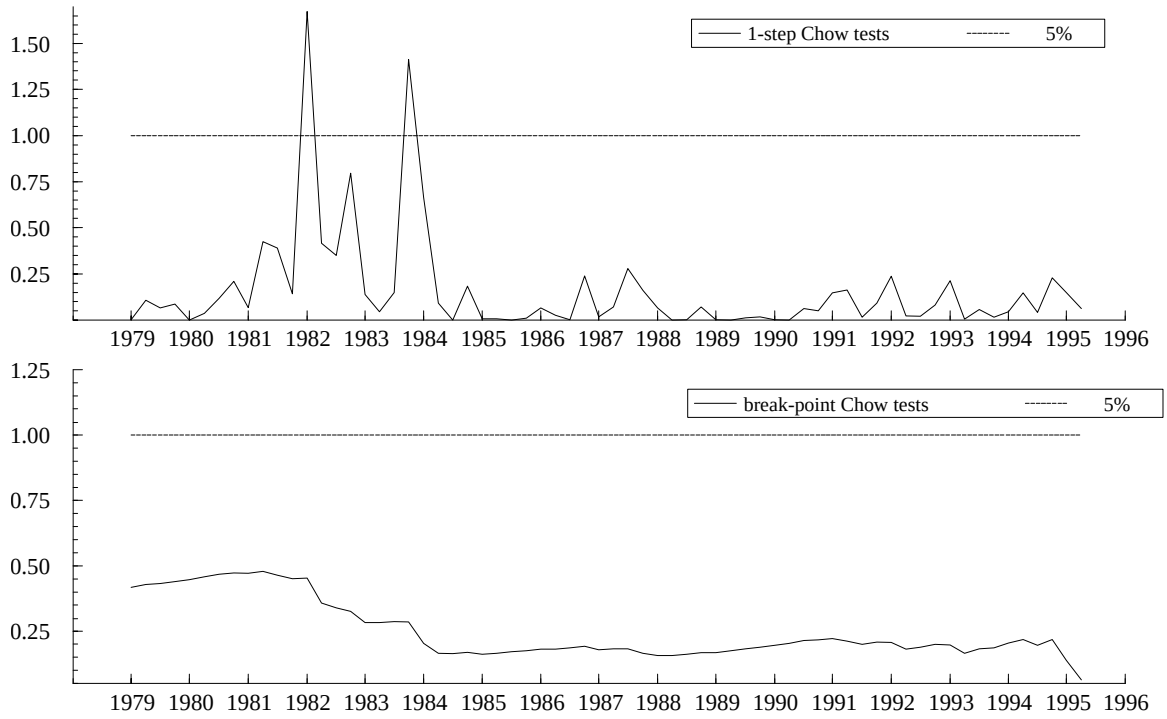

Figure 8: Recursive Chow statistics for the CHO_1^* version of the efficiency-wage model

apparently widespread belief in the economics profession that a good out-of-sample performance conveys strong support for a model. The arguments are illustrated with an efficiency-wage model of post-War US equilibrium unemployment. That model appears to have a good forecast performance, but we show that this does not establish the validity of the model. The efficiency-wage model may well explain the post-War course of US unemployment, but the standard assessments of out-of-sample performance typically reported do not lend much credence to the claim, and indeed the original study's authors provide empirical evidence in support of their model from other sources, such as Granger causality tests.

In the present context, one implication of the belief that out-of-sample forecast performance validates a model, and the theory on which it is based, would be that the annual change in the unemployment rate is an AR(2) process with no impact from any other factors in the economy. This follows because we have shown that a model using only lagged actual values of unemployment rate changes performs better than a model based on the actual values of contemporaneous and lagged oil and interest rate variables. If we allow the model estimates to be recursively updated, then the economic model fares better than the AR(2). But this is cold comfort for advocates of the use of forecast performance to evaluate models. It highlights an instance where one of our 'dichotomies' is decisive in determining whether or not the model receives support, but the literature rarely pays attention to these 'dichotomies'. Put more starkly, do the forecast evaluation differently, and you get different results. It also raises the issue of the support offered to a theory by a model whose forecasts are improved by allowing for non-constant parameters.

It may be that a combination of the economic model forecasts and AR forecasts would be an improvement on either alone, as would be the case if neither model forecast encompasses the other.[6] This may be the best way to proceed for practical forecasting, given that our restricted version of the CHO_{FS} model shows no role for lagged unemployment rate terms, but again what support that such a finding accords to a theory is not obvious.

We also show that the belief that the equilibrium-correction terms can be responsible for a good long-horizon performance is incorrect. Finally, we mentioned briefly in the introduction the argument of Clements and Hendry (1999) that location shifts mitigate the usefulness of out-of-sample forecast performance for model evaluation. Without examining this aspect in detail in the present context, a comparison of figures 5 and 6 is illuminating. Figure 5 depicts 1-step ahead 'forecasts' from the restricted model relating unemployment to the input price variables, and figure 6 reports the 1-step AR(2) model forecasts. The success of the AR(2) forecasts derives from their closely tracking the time series of annual unemployment rate changes, resulting in small forecast errors. The economic model records large errors in the early 1980s when the real oil price variable shifts to unprecedentedly high levels. The long-run relationship which held prior to this period no longer appears a reliable guide at levels of the real oil price far in excess of the values historically observed (compare the EC terms in tables 2 and 3, specifically the estimated oil price coefficients). This suggests that the poor forecast performance of the economic model relative to the AR model is due to the shift in the oil price, and not to the merits of the model itself. Models are approximations to the local DGP (namely, the joint density of the variables under analysis), so that expecting them to characterize the relationships between variables over ranges that bear little resemblance to the historical sample period may

[6] For explanations and further references on forecast combination (or pooling) and forecast encompassing, see Newbold and Harvey (2002).

be unreasonable. Thus, the poor forecast performance in the early 1980s by itself need not invalidate the efficiency-wage hypothesis, but suggests that the response of unemployment to very large oil price changes is more muted than indicated by the linear model with pre-1980 coefficients. In effect, rejecting a model of a theory also does not disconfirm that theory.

The main message is that one must be very careful when using a forecasting exercise as an evaluation device in non-stationary processes.

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. Econometrica, **61**, 821–856.

Atkeson, A., and Ohanian, L. (2001). Are Phillips curves useful for forecasting inflation?. Federal Reserve Bank of Minneapolis, Quarterly Review, **25**, 2–11. (1).

Boughton, J. M. (1992). The demand for M1 in the United States: A comment on Baba, Hendry and Starr. Economic Journal, **103**, 1154–1157.

Carruth, A. A., Hooker, M. A., and Oswald, A. J. (1998). Unemployment equilibria and input prices: Theory and evidence from the United States. Review of Economics and Statistics, **80**, 621–628.

Chong, Y. Y., and Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. Review of Economic Studies, **53**, 671–690. Reprinted in Granger, C. W. J. (ed.) (1990), Modelling Economic Series. Oxford: Clarendon Press.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. Econometrica, **28**, 591–605.

Clements, M. P., and Hendry, D. F. (1995). Forecasting in cointegrated systems. Journal of Applied Econometrics, **10**, 127–146. Reprinted in T. C. Mills (ed.) Economic Forecasting. The International Library of Critical Writings in Economics, Edward Elgar.

Clements, M. P., and Hendry, D. F. (1998). Forecasting Economic Time Series. Cambridge: Cambridge University Press.

Clements, M. P., and Hendry, D. F. (1999). Forecasting Non-stationary Economic Time Series. Cambridge, Mass.: MIT Press.

Doornik, J. A., and Hendry, D. F. (2001). GiveWin: An Interface to Empirical Modelling. London: Timberlake Consultants Press.

Eitrheim, Ø., Husebø, T. A., and Nymoen, R. (1999). Equilibrium-correction versus differencing in macroeconometric forecasting. Economic Modelling, **16**, 515–544.

Engle, R. F., Hendry, D. F., and Richard, J.-F. (1983). Exogeneity. Econometrica, **51**, 277–304. Reprinted in Hendry, D. F., Econometrics: Alchemy or Science? Oxford: Blackwell Publishers, 1993, and Oxford University Press, 2000; and in Ericsson, N. R. and Irons, J. S. (eds.) Testing Exogeneity, Oxford: Oxford University Press, 1994.

Ericsson, N. R. (1992). Cointegration, exogeneity and policy analysis: An overview. Journal of Policy Modeling, **14**, 251–280.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, **37**, 424–438.

Granger, C. W. J., and Newbold, P. (1975). Economic forecasting: The atheist's viewpoint. In Renton, G. A. (ed.), Modelling the Economy. London: Heinemann Educational Books.

Hamilton, J. D. (1983). Oil and the Macroeconomy since World War II. Journal of Political Economy, **91**, 228–248.

Hamilton, J. D. (1996). This is what happened to the oil price-macroeconomy relationship. Journal of Monetary Economics, **38**, 215–220.

Hendry, D. F. (1979). The behaviour of inconsistent instrumental variables estimators in dynamic systems with autocorrelated errors. Journal of Econometrics, **9**, 295–314.

Hendry, D. F. (1996). On the constancy of time-series econometric equations. Economic and Social Review, **27**, 401–422.

Hendry, D. F., and Mizon, G. E. (2000). On selecting policy analysis models by forecast accuracy. In Atkinson, A. B., Glennerster, H., and Stern, N. (eds.), Putting Economics to Work: Volume in Honour of Michio Morishima, pp. 71–113. London School of Economics: STICERD.

Hendry, D. F., and Starr, R. M. (1993). The demand for M1 in the USA: A reply to James M. Boughton. Economic Journal, **103**, 1158–1169.

Hooker, M. A. (1996). Whatever happened to the oil price-macroeconomy relationship?. Journal of Monetary Economics, **38**, 195–213.

Hoover, K. D., and Perez, S. J. (1999). Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. Econometrics Journal, **2**, 167–191.

Krolzig, H.-M., and Hendry, D. F. (2003). Assessing subsample-based model selection procedures. Working paper, Economics Department, Oxford University.

Lawson, T. (1989). Realism and instrumentalism in the development of econometrics. Oxford Economic Papers, **41**, 236–258.

Lynch, A. W., and Vital-Ahuja, T. (1998). Can subsample evidence alleviate the data-snooping problem? A comparison to the maximal R^2 cutoff test. Discussion paper, Stern Business School, New York University.

Mayo, D. G., and Spanos, A. (2000). A post-data interpretation of Neyman–Pearson methods based on a conception of severe testing. Working paper, Department of Philosophy, Virginia Tech.

Miller, P. J. (1978). Forecasting with econometric methods: A comment. Journal of Business, **51**, 579–586.

Nelson, C. R. (1972). The prediction performance of the FRB-MIT-PENN model of the US economy. American Economic Review, **62**, 902–917.

Newbold, P., and Harvey, D. I. (2002). Forecasting combination and encompassing. In Clements, M. P., and Hendry, D. F. (eds.), A Companion to Economic Forecasting, pp. 268–283: Oxford: Blackwells.

Pagan, A. R. (1989). On the role of simulation in the statistical evaluation of econometric models. Journal of Econometrics, **40**, 125–139.

Patterson, K. D. (2003). An integrated model of the data measurement and data generation processes with an application to consumers' expenditure. International Journal of Forecasting, **19**, 177–197.

Phillips, P. C. B. (1994). Bayes models and forecasts of Australian macroeconomic time series. In Hargreaves, C. (ed.), Non-stationary Time-series Analysis and Cointegration. Oxford: Oxford University Press.

Phillips, P. C. B. (1995). Automated forecasts of Asia-Pacific economic activity. Asia-Pacific Economic Review, **1**, 92–102.

Phillips, P. C. B. (1996). Econometric model determination. Econometrica, **64**, 763–812.

Popper, K. (1983). Realism and the Aim of Science. Totowa, New Jersey: Rowman and Littlefield.

Stock, J. H., and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. Journal of Business and Economic Statistics, **14**, 11–30.

Swanson, N. R., and White, H. (1997). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. International Journal of Forecasting, **13**, 439–462.

West, K. D. (1996). Asymptotic inference about predictive ability. Econometrica, **64**, 1067–1084.

7 Appendix

Table 2: CHO model estimated 1955Q4 to 1978Q4

$$\begin{aligned}\Delta_4 U_t \;=\;& \underset{(0.26)}{0.272}\ \Delta_4 U_{t-1} \;-\; \underset{(0.11)}{0.108}\ \Delta_4 U_{t-2} \;+\; \underset{(0.057)}{0.0137} \\ &+ \underset{(0.018)}{0.0752}\ \Delta_4 RO_t \;+\; \underset{(0.064)}{0.0768}\ \Delta_4 RR_t \;+\; \underset{(0.22)}{0.727}\ ECM_{t-1} \\ &- \underset{(0.21)}{0.918}\ ECM_{t-4}\end{aligned}$$

The above equation corresponds to CHO Table 3, col. 2 p.625. It gives the estimates obtained on the sample to 1978Q4, using the estimated EC term recorded below (estimated 1954Q2 to 1978Q4). The figures in parenthesis below the parameter estimates are heteroscedasticity and autocorrelation consistent standard errors. The estimated standard error of the equation is 0.49, and the p-values of an LM test for serial correlation up to fifth order, heteroscedasticity, and normality, are respectively, 0.000, 0.013 and 0.017.

$$U_t \;=\; 0.187 \;+\; 0.0974\ RO_t \;+\; 0.115\ RR_t$$

Table 3: CHO model estimated 1955Q4 to 1978Q4 with full-sample estimates of EC term

$$\begin{aligned}\Delta_4 U_t \;=\;& \underset{(0.23)}{0.0405}\ \Delta_4 U_{t-1} \;-\; \underset{(0.11)}{0.0757}\ \Delta_4 U_{t-2} \;-\; \underset{(0.050)}{0.00719} \\ &+ \underset{(0.009)}{0.0438}\ \Delta_4 RO_t \;+\; \underset{(0.060)}{0.121}\ \Delta_4 RR_t \;+\; \underset{(0.17)}{1.02}\ ECM_{t-1} \\ &- \underset{(0.17)}{1.13}\ ECM_{t-4}\end{aligned}$$

The above equation corresponds to CHO Table 3, col. 2 p.625 but using the EC term based on the data through to 1995Q2. (The EC term below matches that of CHO Table 2, p. 625.) The figures in parenthesis below the parameter estimates are heteroscedasticity and autocorrelation consistent standard errors. The estimated standard error of the equation is 0.46, and the p-values of an LM test for serial correlation up to fifth order, heteroscedasticity, and normality, are respectively, 0.000, 0.075 and 0.001.

$$U_t \;=\; 3.51 \;+\; 0.0356\ RO_t \;+\; 0.136\ RR_t$$