ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

ON NOT EVALUATING ECONOMIC MODELS BY FORECAST OUTCOMES

Jennifer L. Castle and David F. Hendry

Number 538
February 2011

Manor Road Building, Oxford OX1 3UQ

On Not Evaluating Economic Models by Forecast Outcomes

Jennifer L. Castle[†] and David F. Hendry[*]
[†]Magdalen College and Institute for Economic Modelling,
Oxford Martin School, University of Oxford, UK
[⋆]Economics Department and Institute for Economic Modelling,
Oxford Martin School, University of Oxford, UK

February 28, 2011

Abstract

Even in scientific disciplines, forecast failures occur. Four possible states of nature (a model is good or bad, and it forecasts well or badly) are examined using a forecast-error taxonomy, which traces the many possible sources of forecast errors. This analysis shows that a valid model can forecast badly, and a poor model can forecast successfully. Delineating the main causes of forecast failure reveals transformations that can correct failure without altering the 'quality' of the model in use. We conclude that judging a model by the accuracy of its forecasts is more like fools' gold than a gold standard.

JEL classifications: C18, C52.
KEYWORDS: Model evaluation; Forecast failure; Model selection.

1 Introduction

There are four main purposes for building an econometric model: describing the evidence, testing economic theories, policy analysis, and forecasting. Several steps are involved in a model's construction. Empirical models need to be formulated, usually on the basis of subject-matter theory; selected from a potentially large class of 'representative' models, according to some criteria (possibly including forecasting performance on subsamples of the available data); estimated by an appropriate statistical method; and evaluated, where the success of the finally chosen model is sometimes judged by its 'end-of-sample forecasting performance'. When the sole purpose of a model is to make *ex ante* forecasts of future outcomes, the 'accuracy' thereof as judged by some loss function can provide one useful evaluation criterion (see e.g., Granger and Pesaran, 2000, and Granger, 2001). However, to quote Hendry (1986):

> Judging econometric models by their forecasting success seems such a natural procedure that it might occasion surprise to question its usefulness.

As we will show below, the ability of forecasting to reveal even gross mis-specifications of a model depends both on the properties of the data processes and on the structure of the model: good models can forecast poorly and bad models can forecast well. Of course, it is also true that good models can forecast well and bad models can forecast poorly, which makes it clear that forecast performance is not a discriminating criterion between good and bad models.

[*]We are grateful to Michael P. Clements, Neil R. Ericsson and Grayham E. Mizon for helpful comments on an earlier draft, and to Ed Nelson for helping to clarify the nature of the change in UK interest rates in 1984. This research was supported in part by grants from the Open Society Institute and the Oxford Martin School. Contact details: jennifer.castle@magd.ox.ac.uk and david.hendry@nuffield.ox.ac.uk.

Nevertheless, beliefs in the efficacy of judging a model, and the underlying theory, by its forecasts clearly persist. It is often claimed that *ex ante* forecasts are the strongest test of the validity of an estimated econometric model, and consequently, the failure of economists to forecast the large changes worldwide from the 2007–2010 financial crisis and ensuing recession signals the lack of scientific status of economics. A typical example is Gideon Rachman, *Financial Times*, September 6, 2010 citing Joe Stiglitz in support.[1] Clements and Hendry (2005) provide two examples of such unsubstantiated claims. 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 ... a simple atheoretical forecast ... Atkeson and Ohanian (2001)

That statement is despite the analysis in Hendry and Mizon (2000) showing that one should never select policy analysis models by forecast accuracy, nor necessarily reject their policy implications because of poor forecasts (notwithstanding the critique in Lucas, 1976; see Aldrich, 1989, for the history; and Ericsson and Irons, 1995, for a counter-critique). Second:

> 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)

Clements and Hendry (2005) consider the second paper in detail and conclude:

> Out-of-sample forecast performance is not a reliable indicator of whether an empirical model offers a good description of the phenomenon being modelled, nor therefore of the economic theory on which the model is based.

An empirical model that produces forecasts for a future time period that transpire to be 'accurate' for the later measured outcomes seems to deserve epithets like 'credible' or 'good'. However, even finessing the contentious issue of how to measure the 'accuracy' of forecasts (see Clements and Hendry, 1993a, and section 4), using some new settings, we reiterate that there need be no connection between the validity, or verisimilitude, of a model, in terms of the 'goodness' of its representation of the economy, and any reasonable measure of its forecast accuracy.

The opening quote from Hendry (1986) also makes clear that this issue has been a concern for some time–but Hendry was far from the first to discuss the problem. Mills (2010) has recently reminded us that Bradford Bixley Smith (1926) saw many of the key issues that would confront the analysis of non-stationary time series. Just before the onset of the 1929 financial crisis and Great Depression, Smith wrote presciently about the problems of economic forecasting, in particular, see Smith (1927) and Smith (1929). Somehow his research contributions to both modelling and forecasting were completely forgotten, despite being published in the *Journal of the American Statistical Association*. Hendry and Richard (1982) showed the dangers of evaluating estimated models by dynamic simulation outcomes, which are, of course, conditional multi-period ahead forecasts. Nevertheless, the issue remains both important and misunderstood, so we re-analyze the relation between the 'quality' of a model and the 'quality' of its forecasts.

Why is successful out-of-sample performance often regarded as the 'gold standard' for model evaluation? On the one hand, fears about 'data mining' and over-fitting a given sample are taken to entail that in-sample evaluation is inadequate. We concur that models can be designed to be well-specified, in the sense of satisfying all the relevant tests for mis-specification, and that can even be achieved by 'camouflaging' problems. On the other hand, 'new' data may be thought more of a challenge to fit, as it has occurred after formulating, selecting and estimating a model, so is less open to 'manipulation'. In

[1] www.ft.com/cms/s/0/93d9ff2a-b9e1-11df-8804-00144feabdc0.html.

stationary processes, systematic relative forecast out-performance supports using the associated model compared to its competitors. Unfortunately, economies are distinctly non-stationary, and that greatly limits what can be learnt from either successful or unsuccessful forecasting.

The structure of the paper is as follows. Using examples from disciplines that are avowedly scientific, section 2 shows that forecast failure still occurs, so cannot entail any implications about scientific status. Section 3 proposes an analytic framework based on the four possible states of nature created by the combinations that a model is good or bad, and it forecasts well or badly. An explicit, if simple, example traces the many possible forecast errors that can occur. Section 4 shows how difficult it is to even measure the 'accuracy' of forecasts, then section 5 shows that a valid model can unfortunately forecast badly. Section 6 considers the conditions under which a poor model can nevertheless forecast successfully. Section 7 delineates the main causes of forecast failure, and section 8 considers transformations that correct such forecast failure, without altering the 'quality' of the model in use. Section 9 concludes.

2 No discipline should be judged by its forecasts

If a failure to forecast entailed a lack of scientific status, that same charge could also be laid against a number of other disciplines that most people would doubtless class as sciences. As two examples, geologists and oceanographers failed to predict the 2004 Indian Ocean tsunami; and NASA failed in its prediction that Apollo 13 would get to the Moon. These two examples highlight key insights. The former was due to a lack of both the pertinent information to forecast the undersea earthquake that caused the tsunami, and measuring instruments that could record its devastating progress, a lacuna now corrected by a system of satellites. The latter exemplifies ever increasing 'forecast failure', since the lunar module has still not arrived at its destination. Yet who would reject Newton's laws of gravity, or even NASA's forecasting algorithms, because of this outcome? Rather, the unanticipated explosion of an oxygen cylinder precipitated the problem–an accident outside the 'universe' of events considered when formulating the forecast.

These are two of the many reasons why the financial crisis and its consequences were not forecast in advance: most of the information was lacking at the required time, and some of the causes (such as the bankruptcy of Lehman Brothers) were unanticipated till they happened. By itself, such a forecast failure tells us nothing more about the quality of economic reasoning or the entailed models than does Apollo 13 about Newton's laws. As we have shown in previous research (see Clements and Hendry, 1998, 1999), the forecasts from a given econometric model can be good or bad depending on how they are used, almost independently of the verisimilitude of the model.

3 An analytic framework

There are four possible states of nature, as follows.
(I) A model forecasts successfully, and is indeed a valid representation of the relevant data-generating process (DGP) in-sample and over the forecast horizon.
(II) A model forecasts successfully, but in fact is invalid as a representation of even the in-sample DGP.
(III) A model deservedly suffers forecast failure because it is a poor approximation to the in-sample DGP.
(IV) A model suffers forecast failure, yet is a correct in-sample representation of the DGP.
If all four cases can occur, then it should be clear that forecast performance cannot differentiate good from bad models. And all four can occur, as we now discuss.

The first is obvious, but impossible to establish from the success of the forecasts alone: whether or not a model is a good representation of the DGP needs to be judged on that basis, and not on whether it can forecast the future.

The second is sufficiently misunderstood that we devote section 6 to analyzing the many settings in which a model can forecast quite successfully, but in fact is an invalid representation of the DGP. Surprisingly, a model which uses no causal variables can be the ‘best’ forecasting device available.

The third seems relatively obvious, but again with the *caveat* that whether or not it is a bad representation of the DGP has to be established independently of the quality of the forecasts.

Finally, in section 5, we analyze the many possible causes of an outcome in which a good model of the DGP nevertheless suffers forecast failure.

The analytical example below helps illustrate these four cases. It is an extension of that in Hendry (2011b) (see Clements and Hendry, 2008, for an exposition, and the excellent discussion in Ericsson, 2008). Consider a stationary scalar autoregressive DGP with an additional exogenous regressor z_t such that:

$$y_t = \alpha + \rho y_{t-1} + \besta z_t + v_t \text{ where } v_t \sim \mathsf{IN}\left[0, \sigma_v^2\right] \text{ with } |\rho| < 1, \tag{1}$$

and $\mathsf{IN}[0, \sigma_v^2]$ denotes an independent normal distribution with mean, $\mathsf{E}[v_t] = 0$, and variance $\mathsf{V}[v_t] = \sigma_v^2$. Equation (1) can be rewritten in the equilibrium-correction form:

$$\Delta y_t = (\rho - 1)(y_{t-1} - \mu) + \beta(z_t - \kappa) + v_t \tag{2}$$

where $\mathsf{E}[z_t] = \kappa$ and:

$$\mathsf{E}[y_t] = \mu = \frac{\alpha + \beta\kappa}{(1 - \rho)} \tag{3}$$

When $\{v_t\}$ is an innovation process, (1) is constant, z_{T+1} is known and the data are all correctly measured, then the ‘best’ possible forecast is based on knowing the DGP, so that:

$$\widetilde{y}_{T+1|T} = \alpha + \rho y_T + \beta z_{T+1} \tag{4}$$

with the forecast error:

$$\widetilde{v}_{T+1|T} = y_{T+1} - \widetilde{y}_{T+1|T} = v_{T+1} \sim \mathsf{IN}\left[0, \sigma_v^2\right] \tag{5}$$

which has a mean of zero and a variance of σ_v^2. Consequently, a ‘good forecast’ $\widehat{y}_{T+1|T}$ is presumably one where $\mathsf{E}[y_{T+1} - \widehat{y}_{T+1|T}] \simeq 0$ and $\mathsf{V}[y_{T+1} - \widehat{y}_{T+1|T}] \simeq \sigma_v^2$. Unfortunately, the assumptions needed for (4) to produce the ‘perfect’ outcome in (5) are unrealistic, so let us ‘deconstruct’ the sources of possible forecast errors. The general case is considered in Hendry and Mizon (2011a).

When (1) is the DGP, the parameters α, ρ, κ and β are constant in sample (already a strong assumption) and estimated over $t = 1, \ldots, T$, using $\widehat{\mu} = (\widehat{\alpha} + \widehat{\beta}\widehat{\kappa})/(1 - \widehat{\rho})$ from (3), then the forecast for $T + 1$ from $\widehat{y}_T$ (an estimate of the forecast origin) using a forecast $\widehat{z}_{T+1}$ for z_{T+1} is:

$$\widehat{y}_{T+1|T} = \widehat{\mu} + \widehat{\rho}\left(\widehat{y}_T - \widehat{\mu}\right) + \widehat{\beta}\left(\widehat{z}_{T+1} - \widehat{\kappa}\right). \tag{6}$$

Denote all the expected values of the in-sample estimates by (e.g.) $\mathsf{E}[\widehat{\mu}] = \mu_e$ etc. The 1-step ahead forecast error is $\widehat{v}_{T+1|T} = \overline{y}_{T+1} - \widehat{y}_{T+1|T}$ where $\overline{y}_{T+1}$ is the ‘flash’ or initial estimate of y_{T+1} against which the forecast will be evaluated at time $T + 1$. Let $\overline{v}_{T+1} = \left(\overline{y}_{T+1} - y_{T+1}\right)$ be the initial measurement error, and allow every forecast-period parameter to have changed from its in-sample value (denoted *, where from (3), the new equilibrium mean is $\mu^* = (\alpha^* + \beta^*\kappa^*)/(1 - \rho^*)$ with $|\rho^*| < 1$). Then $\widehat{v}_{T+1|T}$ leads to the forecast-error taxonomy in (7):[2]

[2] This is deliberately simplified by dropping the interaction terms $(\beta_e - \widehat{\beta})(\kappa - \widehat{\kappa}) + (\rho_e - \widehat{\rho})(\mu - \widehat{\mu}) - (\rho_e - \widehat{\rho})(y_T - \widehat{y}_T) - (\beta_e - \widehat{\beta})(z_{T+1} - \widehat{z}_{T+1})$, which are usually of a smaller order in probability: the interested reader can add their implications to the analysis below.

$$
\begin{aligned}
\widehat{v}_{T+1|T} \simeq\; & (1-\rho^*)(\mu^*-\mu) - \beta_e(\kappa^*-\kappa) && \text{[A]} \\
& + (\rho^*-\rho)(y_T-\mu) + (\beta^*-\beta)(z_{T+1}-\kappa^*) && \text{[B]} \\
& + (1-\rho)(\mu-\mu_e) - \beta(\kappa-\kappa_e) && \text{[C]} \\
& + (\rho-\rho_e)(y_T-\mu) + (\beta-\beta_e)(z_{T+1}-\kappa^*) && \text{[D]} \\
& + (1-\rho_e)(\mu_e-\widehat{\mu}) - \beta_e(\kappa_e-\widehat{\kappa}) && \text{[E]} \\
& + (\rho_e-\widehat{\rho})(y_T-\mu) + \left(\beta_e-\widehat{\beta}\right)(z_{T+1}-\kappa) && \text{[F]} \\
& + \overline{v}_{T+1} + \rho_e(y_T-\widehat{y}_T) + \beta_e(z_{T+1}-\widehat{z}_{T+1}) + v_{T+1}. && \text{[G]}
\end{aligned}
\qquad (7)
$$

The first two rows correspond respectively to terms arising from changes in means in [A] and slopes in [B], mean mis-specification in [C] and slope mis-specification in [D], mean and slope estimation in [E] and [F], and measurement mistakes and errors in [G].

First, if there is no parameter change, rows [A] & [B] will be zero. If there is no mis-specification, [C] & [D] will be zero. If the estimation sample is very large, rows [E] & [F] will be negligible. And if the data are accurate and z_{T+1} is known, row [G] reduces to v_{T+1} so then $\widehat{y}_{T+1|T}$ will deliver a 'good forecast'. Thus, a good model can forecast well. But, as we will show, it need not.

From (7), the 16 possible individual problems and all their joint occurrences create far too many cases to consider even for first moment mistakes (see Ericsson, 2001, for a discussion of the general issue of forecast uncertainty), but the next two sections discuss a number of salient settings. However, that huge number of different ways in which a forecast can deviate from the 'optimum' in (4), despite the simplicity of this setting, signals why judging a model by its *ex ante* forecast performance may not be straightforward.

4 Measuring forecast 'accuracy'

The difficulty of simply measuring forecast 'accuracy' is emphasized by Clements and Hendry (1993a), prompting discussant's comments that are longer than the original paper (also see those authors' reply in Clements and Hendry, 1993b, and an update in Clements and Hendry, 2011a): Ericsson (1992) provides a clear introduction. One problem is the lack of invariance of conventional measures like mean-square forecast errors, (MSFEs), to admissible linear transformations of the variable being forecast, to whether a single or several variables are under consideration, and whether one or multi-step ahead forecasts are being evaluated.

Comparisons based on MSFEs can yield inconsistent rankings as apparently innocuous changes are made, as Figure 1 illustrates. In this simulated-data graph, when forecasting by $\widehat{y}_{T+h}$ the level of a variable $y_{T+h}, h = 1, \ldots, H$ from a forecast origin at $T = 1975(i)$ in the left-hand columns, forecaster *a* is apparently less accurate than $\widetilde{y}_{T+h}$ by forecaster *b* in the lower left panel, and indeed has a larger MSFE. Yet when forecasting the change Δy_{T+h} in the right-hand panels, $\Delta\widehat{y}_{T+h}$ is clearly more accurate than $\Delta\widetilde{y}_{T+h}$, and now has a smaller MSFE. Worse still, since the forecast-origin value y_T is known, the levels derived from the changes forecast by *a* can be cumulated as $\Delta\widehat{y}_{T+1} + y_T$ etc., to deliver much more accurate levels forecasts than the original (an illustration of the potential efficacy of an intercept correction).

There are three important lessons from this simple example. First, without an explicit agreed measure, evaluation is simply not unique. Secondly, even a given set of forecasts like $\widehat{y}_{T+h}$ is potentially subject to amendment by devices like intercept corrections: is one to evaluate the forecast or the model that made the forecast? Finally, while forecast evaluation could use a specific loss function, from which

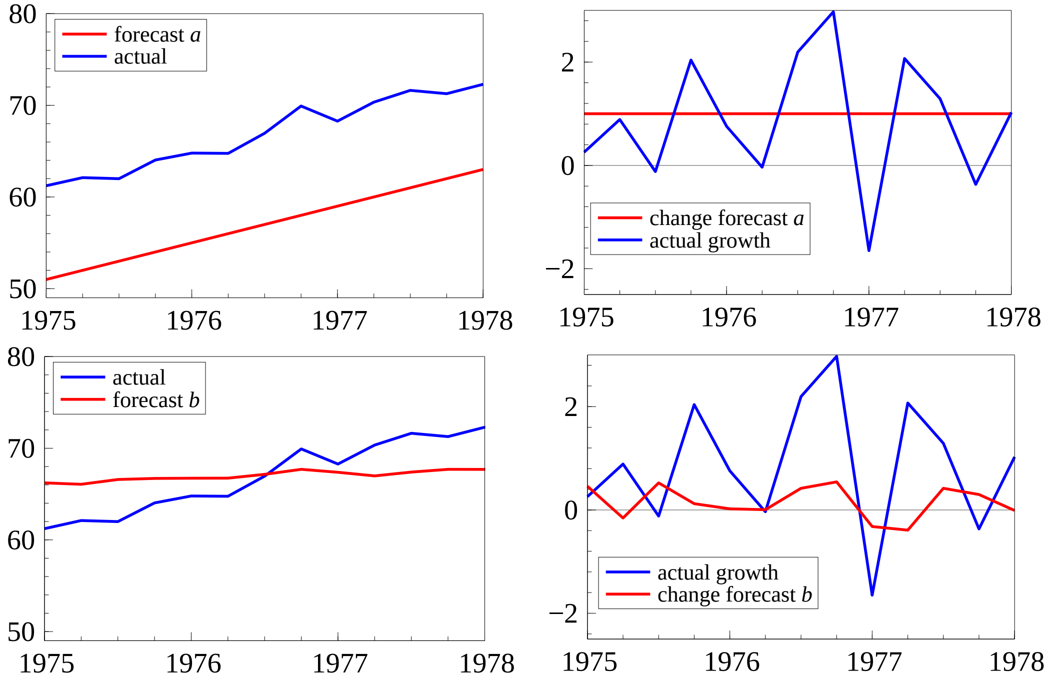

Figure 1: Who wins: forecaster *a* or *b*?

an optimal predictor is derived (as in Granger and Pesaran, 2000, and Granger, 2001), the outcome now depends on the choice of that loss function, and well-defined mappings between forecast errors and costs are anyway not typical in macroeconomics (and could still not be invariant to admissible transformations as in Figure 1). It seems odd that a 'gold standard' should lack a unique measure of the quality of the gold.

5 A valid model can forecast inaccurately

There are four known situations under which an estimated correct specification of a DGP, with constant parameters both before and after forecasting, can nevertheless forecast inaccurately relative to its in-sample performance. These are:
(i) a change in the collinearity of unmodelled exogenous variables, even when their future values are known;
(ii) measurement errors at the forecast origin;
(iii) mis-measured flash estimates of the forecast outcome that will be revised later; and
(iv) a changed measurement system.
We address these in turn. It is well known that forecasts can be inaccurate in an absolute sense because of a low signal-to-noise ratio, but the cases here concern *relative* inaccuracy of forecasts as against fit.

(i) Breaks in the marginal process alter the collinearities between explanatory variables, and this is important for its impact on the most collinear combination, which thereby has the smallest eigenvalue (say λ_n) in the $n \times n$ data second-moment matrix (see Castle, Fawcett and Hendry, 2010). In general, breaks reduce collinearities, and hence increase the smallest eigenvalue in the forecast period relative to its in-sample value (to say $\lambda_n^* > \lambda_n$). Thus, despite an increase in the information content of the data, there is an adverse impact on the MSFE when collinearity changes. Worse, that impact on MSFE depends on λ_n^*/λ_n, so can be very large, and is unavoidable because deleting the collinear variables does not help unless they are actually irrelevant. However, immediate updating of the parameter estimates in the next period can attenuate that effect. Nevertheless, a very bad forecast can result from a well-specified, constant model in that setting. This result is 'hidden' in [E] and [F] in (7), which relates to first moments only.

(ii) As briefly discussed above, measurement errors can be large at the forecast origin, and can deliver inaccurate forecasts despite the model being excellent. Given the correct specification and constant parameters in (7), but introducing mis-estimation of the forecast origin, so the model is correct but the data are not, forecast failure can occur: a good model can appear to forecast badly. It may not be known at the time of the forecast that the origin is badly measured, so for a while the forecast will be judged a failure. Later revisions will of course correct that mis-perception, but the point remains: forecast accuracy is an unreliable guide to model validity.

(iii) As noted, this setting is similar to (ii), but now the mis-measurement occurs for the forecast outcome, initially suggesting that failure occurred when compared to an incorrect flash value for y_{T+1}. Again, although later revisions may reverse that judgement, the point remains that an apparently bad forecast need not entail that the model is invalid. Castle, Fawcett and Hendry (2009) consider differentiating between breaks and measurement errors at the forecast origin.

(iv) Entire measurement systems can change. Judged against a new system, a good forecast can appear to be very bad. A 'classic' example is the major shift in the measurement of the opportunity cost of holding narrow money (M1) in the UK in 1984 quarter 3. This was induced by the introduction of tax deduction before payment of bank deposit interest earnings and compulsory reporting thereof to the tax authorities,[3] when previously interest income had been paid pre-tax and left to individual reporting (see Hendry, 1985, and Hendry and Ericsson, 1991). Forecasts based on a model where the opportunity cost is measured by the outside interest rate are wildly inaccurate over the horizon after 1984, when high rates of interest were paid on checking accounts. However, they are as accurate as in-sample fitted values when a model with *identical* estimated coefficients is used but with the correct (learning-adjusted) differential interest rate measure: see Castle, Fawcett and Hendry (2011) for a recent update.

Thus, these four examples illustrate how a good model can forecast poorly. The next section considers cases in which an invalid model can nevertheless forecast accurately.

6 An invalid model can forecast accurately

In this section, we consider three cases where the forecasting model does not coincide with the DGP and yet this does not lead to forecast failure. First, in a stationary, ergodic world, empirical models whose parameters are estimated by least-squares must be consistent for their associated conditional expectations (when second moments exist). Under stationarity, sample second moments converge to their population counterparts, and so continuous functions thereof do so as well. Consequently, forecasts from such models must on average attain their expected accuracy unconditionally. Mathematical analyses are provided by Miller (1978) and Hendry (1979). For the DGP in (1) under constancy, the main possible in-sample mis-specification is not knowing that z_t should be included. When the DGP is constant and the data are correctly measured with means of zero, but z_t is incorrectly omitted, then $\widehat{\beta} = \beta_e = 0$ in (7). However, that mistake adds very little additional cost to forecasting, as $\mathsf{E}[\widehat{v}_{T+1|T}] \simeq 0$ when all variables have mean zero, although there is an additional variance term from $\beta^2 \mathsf{V}[z_{T+1}]$. More importantly, the in-sample error variance estimate is 'inflated' by as much as the forecast-error variance, so the mis-specified model forecasts as well as is anticipated. A key reason for this state of affairs has been known since ancient times concerning 'saving the appearances'. A classic example is the epicycle system of Ptolemy (see Harré, 1985, p. 86), where the stationarity of the solar system's behaviour entailed that even an incorrect

[3] Finance Bill (no. 2) 1984, clause 43: see http://www.legislation.gov.uk/ukpga/1984/43/schedule/8/enacted.

model thereof would forecast approximately as accurately as it fitted.[4] In economics, if some simple data transformations (such as to growth rates) produced time series that were stationary and ergodic, forecasts on average would have the same error variances as within-sample fits and hence would reject α% of the time on a properly calibrated test with a theoretical rejection rate of α%.

Next, in a correctly specified model of a stationary process, large-sample forecast accuracy only depends on the variance of the innovation process. By construction, that variance cannot be reduced without either extending the information set or reducing the implicit discrete-time measurement interval. Thus, the innovation variance provides an irreducible lower bound to forecast error variances if the observation period and the ex-ante information set are fixed. When some component has a 'large' variance, then poor forecasts will result even from the 'correct' model, whereas, if the error variance is 'small', one can obtain accurate forecasts despite using an inappropriate information set. Moreover, it can be proved that:
a] the conditional expectation is the unbiased minimum MSFE predictor;
b] a dominant, encompassing, model in-sample will provide the minimum MSFE forecasts in large samples.
Unfortunately, such theorems are cold comfort in processes subject to unanticipated shifts.

Third, in processes where all the variables have zero means and no trends, the forecast error taxonomy in Clements and Hendry (1998) shows that changes in the parameters can be very difficult to detect (also see Hendry, 2000), a result that can be extended to include mis-specification of the model in use for the DGP. In (7), consider $\alpha^* = \alpha = 0$ and $\kappa^* = \kappa = 0$, so $\mu^* = \mu = 0$, entailing that all variables have mean zero, and that is known, so that rows [A], [C] and [E] are zero. Given a large estimation sample size, correct specification, and accurate data, then rows [D], [F] and [G] in (7) are also approximately zero, other than v_{T+1}, with $\mathsf{E}[v_{T+1}] = 0$ and $\mathsf{V}[v_{T+1}] = (\sigma^*)^2_v$. Then:

$$\begin{aligned}\mathsf{E}[\widehat{v}_{T+1|T}] &\simeq \mathsf{E}[(\rho^* - \rho)\, y_T + (\beta^* - \beta)\, z_{T+1}] + \mathsf{E}[v_{T+1}] \\ &= (\rho^* - \rho)\, \mathsf{E}[y_T] + (\beta^* - \beta)\, \mathsf{E}[z_{T+1}] = 0,\end{aligned}$$

so the forecast is unbiased despite any changes in the parameters. Taking the special case where $\beta^* = \beta$ and ignoring second-order effects for simplicity:

$$\mathsf{V}[\widehat{v}_{T+1|T}] \simeq (\rho^* - \rho)^2\, \mathsf{V}[y_T] + (\sigma^*)^2_v$$

Since $|\rho| < 1$ and $|\rho^*| < 1$, $(\rho^* - \rho)^2$ will generally be small: e.g., for $\rho^* = 0.8$ and $\rho = 0.4$, then $(\rho^* - \rho)^2 = 0.16$. Thus, only a small fall in $(\sigma^*)^2_v$ is needed to leave $\mathsf{V}[\widehat{v}_{T+1|T}] \simeq \sigma^2_v$. Despite the dynamic feedback coefficient having doubled, so the model is now a poor representation of 'reality', the forecast remains good on the usual criteria: a 'changed' model can even forecast quite well.

Combining these mean-zero cases, so the model is mis-specified by omitting z_t, and all the other parameters change, still entails a similar conclusion: $\mathsf{E}[\widehat{v}_{T+1|T}] \simeq 0$, and $\mathsf{V}[\widehat{v}_{T+1|T}] \simeq \mathsf{V}[\widehat{v}_t]$, so the forecasts look 'fine', as Monte Carlo simulations confirm. Most of these conclusions also hold for multi-step forecasts.

7 The causes of forecast failure

When the DGP entails a non-zero equilibrium mean, $\mu \neq 0$, either because $\alpha \neq 0$ or $\kappa \neq 0$, then any change in parameters will induce systematic mis-forecasting, even if the model is correctly specified for

[4] Spanos (2007) shows that the Ptolemaic system can be rejected against the Copernican by the larger, and systematic nature of the, errors in the former.

the DGP in-sample and accurately estimated. Setting all rows [B]–[G] to have expectations of zero, row [A] alone delivers:

$$\mathsf{E}[\widehat{v}_{T+1|T}] \simeq (1-\rho^*)(\mu^*-\mu) - \beta(\kappa^*-\kappa) \neq 0. \tag{8}$$

Furthermore, (2) ensures that such a forecast error will persist because the DGP will correct towards μ^* whereas the model reverts towards $\widehat{\mu}$. Written in equilibrium-correction form:

$$\text{DGP}: \ \Delta y_{T+h} = (\rho^*-1)(y_{T+h-1}-\mu^*) + \beta^*(z_{T+h}-\kappa^*) + v_{T+h} \tag{9}$$

$$\text{Model}: \ \Delta \widehat{y}_{T+h|T} = (\widehat{\rho}-1)\left(\widehat{y}_{T+h-1|T}-\widehat{\mu}\right) + \widehat{\beta}(z_{T+h}-\widehat{\kappa}). \tag{10}$$

Since economic data are indices and can have arbitrary units (such as trillions versus billions), parameters like μ and κ are arbitrary, yet determine the extent of systematic forecast failure after a shift.

Section 6 showed that the mis-specification of incorrectly omitting z_t hardly changed the apparent cost of forecast errors when all means were zero. A very different outcome emerges when intercepts are non-zero, even if they stay constant, so $\alpha^* = \alpha$ and $\kappa^* = \kappa$. Now, even if $\alpha = 0$, and despite the model being correctly specified with $\beta^* = \beta$, $\rho = \rho_e$ and $\beta = \beta_e$ etc., so only the dynamic feedback parameter ρ shifts, then from (7):

$$\mathsf{E}[\widehat{v}_{T+1|T}] \simeq (1-\rho^*)(\mu^*-\mu) = \beta\kappa\frac{(\rho^*-\rho)}{(1-\rho)} \neq 0,$$

which can be extremely large depending on the magnitude of $\beta\kappa$. Even more surprising, in such a setting, this failure is little affected by not including z_t in the model, as (11) shows:

$$\mathsf{E}[\widehat{v}_{T+1|T}] \simeq (1-\rho^*)(\mu^*-\mu) + \beta\mathsf{E}[z_{T+1}-\kappa] = (1-\rho^*)(\mu^*-\mu). \tag{11}$$

Consequently, it is the existence of a non-zero mean for z_t that causes forecast failure after the shift in ρ, not the goodness of the specification.

The one case where the mis-specification of incorrectly omitting z_t differs from including it is when the only shift is in κ to κ^*. Now, from (7), in the latter case, $\mathsf{E}[\widehat{v}_{T+1|T}] \simeq 0$ whereas omission entails $\widehat{\beta} = \beta_e = 0$ so leads to:

$$\mathsf{E}[\widehat{v}_{T+1|T}] \simeq \beta(\kappa^*-\kappa_e) = \beta\kappa^* \tag{12}$$

as $\kappa_e = 0$ when exclusion occurs. This is perhaps the case that may suggest that forecast evaluation is useful in model choice, since the failure is uniquely induced by the incorrect specification, but that is just one of dozens of possible sources of forecast failure.

However, these results highlight a deeper problem discussed by Hendry and Mizon (2010) (see Hendry and Mizon, 2011b, for a non-mathematical explanation): the very theorems that underpin intertemporal analyses in economics also fail when unanticipated shifts occur. Specifically, conditional expectations formed today for an outcome tomorrow are neither unbiased nor minimum MSFE, and the law of iterated expectations (namely the expectation today of the conditional expectation tomorrow equals the unconditional expectation tomorrow), no longer holds as the integrals required to prove the conventional result must be over the same distribution, but are now over different distributions. Forecast failure entails analytical failure.

8 Correcting forecast failure

Is all lost? Let us return to the Apollo 13 example: why are NASA's forecasting algorithms not thrown into doubt despite a large forecast failure? The answer probably lies partly in their generic basis in Newton's laws, which remain indubitable for such purposes, and partly in the fact that the same algorithms correctly forecast the new trajectory of the module almost immediately after the explosion, and indeed during the astronauts' return to Earth. Thus, surely one should judge the model successful in the same way if a forecasting model had that property–namely that despite a failure from an unanticipated location shift it correctly forecasts shortly after.

Unfortunately, forecast failure is relatively easily 'hidden' (or 'fixed') one period after such a location shift. Consider a one-step forecast from (10) at time $T+1$ when the break occurred at T, so that:

$$\Delta\widehat{y}_{T+2|T+1} = \left(\widehat{\rho}-1\right)\left(y_{T+1}-\widehat{\mu}\right)+\widehat{\beta}\left(z_{T+2}-\widehat{\kappa}\right). \tag{13}$$

As (13) shows, the initial forecast error will be essentially repeated, and a sequence of systematic, typically same-signed, forecast errors will occur. To avoid that problem, difference equation (13) instead of using (13) itself to produce:

$$\begin{aligned}\Delta\widehat{y}_{T+2|T+1} &= \Delta y_{T+1}+\left(\widehat{\rho}-1\right)\Delta\left(y_{T+1}-\widehat{\mu}\right)+\widehat{\beta}\Delta\left(z_{T+2}-\widehat{\kappa}\right)\\ &= \Delta y_{T+1}+\left(\widehat{\rho}-1\right)\Delta y_{T+1}+\widehat{\beta}\Delta z_{T+2},\end{aligned} \tag{14}$$

taking the now known lagged value Δy_{T+1} to the right-hand side. Compare that to the outcome of the equivalent operation on the DGP:

$$\Delta y_{T+2} = \Delta y_{T+1}+\left(\rho^{*}-1\right)\Delta y_{T+1}+\beta^{*}\Delta z_{T+2}+\Delta v_{T+2}. \tag{15}$$

The difference between (14) and (15) is $\widetilde{v}_{T+2|T+1} = \Delta y_{T+2}-\Delta\widehat{y}_{T+2|T+1}$, which is:

$$\begin{aligned}\widetilde{v}_{T+2|T+1} &= \Delta y_{T+1}+\left(\rho^{*}-1\right)\Delta y_{T+1}+\beta^{*}\Delta z_{T+2}+\Delta v_{T+2}-\Delta y_{T+1}-\left(\widehat{\rho}-1\right)\Delta y_{T+1}-\widehat{\beta}\Delta z_{T+2}\\ &= \left(\rho^{*}-\widehat{\rho}\right)\Delta y_{T+1}+\left(\beta^{*}-\widehat{\beta}\right)\Delta z_{T+2}+\Delta v_{T+2}.\end{aligned} \tag{16}$$

Hence, $\mathsf{E}[\widehat{v}_{T+2|T+1}]$ has a small value. Castle *et al.* (2011) show how effective (14) can be at forecasting directly after a break. Other methods of offsetting shifts like those in (9) include intercept correction.

An alternative interpretation is that forecasting growth rates makes judging failure much harder, which is indeed the case, but not necessarily for the reason in (16) (see Clements and Hendry, 1993a). A useful check is on the accuracy of the level outcome, which cumulates the forecast errors of the differences, and can reveal that, while the sequence of forecasts of (say) 1%, 1%, . . ., 1% never lie significantly outside their *ex ante* forecast intervals when the outcomes are 2%, 2%, . . ., 2%, the cumulative error is 1%, 2%, . . ., k% for k-steps ahead, which is eventually detectable as statistically unsatisfactory.

9 Conclusion

Judging the 'validity' of a model by the accuracy of its *ex ante* forecasts seems closer to fools' gold than the gold standard that many seem to believe it is. We have shown how many factors determine the goodness or otherwise of a forecast, and how few depend on the verisimilitude of the model relative to the data-generating process.

Not only does such an evaluation not discriminate good models from bad, selecting a model by its forecasting performance over a short period also places too much weight on a small set of data points designated the 'forecast period'.

How can one select then evaluate an empirical model? That is another, rather longer, story, to which a possible answer is offered in Hendry (2011a).

References

Aldrich, J. (1989). Autonomy. *Oxford Economic Papers*, **41**, 15–34.

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

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.

Castle, J. L., Fawcett, N. W. P., and Hendry, D. F. (2009). Nowcasting is not just contemporaneous forecasting. *National Institute Economic Review*, **210**, 71–89.

Castle, J. L., Fawcett, N. W. P., and Hendry, D. F. (2010). Forecasting with equilibrium-correction models during structural breaks. *Journal of Econometrics*, **158**, 25–36.

Castle, J. L., Fawcett, N. W. P., and Hendry, D. F. (2011). Forecasting Breaks and During Breaks. in Clements, and Hendry (2011b), Ch. 11. Forthcoming.

Clements, M. P., and Hendry, D. F. (1993a). On the limitations of comparing mean squared forecast errors. *Journal of Forecasting*, **12**, 617–637.

Clements, M. P., and Hendry, D. F. (1993b). On the limitations of comparing mean squared forecast errors: A reply. *Journal of Forecasting*, **12**, 669–676.

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.

Clements, M. P., and Hendry, D. F. (2005). Evaluating a model by forecast performance. *Oxford Bulletin of Economics and Statistics*, **67**, 931–956.

Clements, M. P., and Hendry, D. F. (2008). Economic forecasting in a changing world. *Capitalism and Society*, **3, 2, 1**, 1–18.

Clements, M. P., and Hendry, D. F. (2011a). Forecasting from Mis-specified Models in the Presence of Unanticipated Location Shifts. in *Oxford Handbook of Economic Forecasting* (2011b), Ch. 10. Forthcoming.

Clements, M. P., and Hendry, D. F. (eds.)(2011b). *Oxford Handbook of Economic Forecasting*. Oxford: Oxford University Press.

Ericsson, N. R. (1992). Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extensions, and illustration. *Journal of Policy Modeling*, **14**, 465–495.

Ericsson, N. R. (2001). Forecast uncertainty in economic modeling. in Hendry, and Ericsson (2001), pp. 68–92.

Ericsson, N. R. (2008). Comment on ‘Economic forecasting in a changing world’ (by Michael Clements and David Hendry). *Capitalism and Society*, **3, 2, 2**, 1–16.

Ericsson, N. R., and Irons, J. S. (1995). The Lucas critique in practice: Theory without measurement. In Hoover, K. D. (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, pp. 263–312. Dordrecht: Kluwer Academic Press.

Granger, C. W. J. (2001). Evaluation of forecasts. in Hendry, and Ericsson (2001), pp. 93–103.

Granger, C. W. J., and Pesaran, M. H. (2000). A decision-theoretic approach to forecast evaluation. In Chon, W. S., Li, W. K., and Tong, H. (eds.), *Statistics and Finance: An Interface*, pp. 261–278. London: Imperial College Press.

Harré, R. (1985). *The Philosophies of Science*. Oxford: Oxford University Press.

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. (1985). Monetary economic myth and econometric reality. *Oxford Review of Economic Policy*, **1**, 72–84.

Hendry, D. F. (1986). The role of prediction in evaluating econometric models. In *Proceedings of the Royal Society*, Vol. A407, pp. 25–33.

Hendry, D. F. (2000). On detectable and non-detectable structural change. *Structural Change and Economic Dynamics*, **11**, 45–65.

Hendry, D. F. (2011a). Empirical economic model discovery and theory evaluation. Discussion paper 529, Economics Department, Oxford University.

Hendry, D. F. (2011b). Mathematical models and economic forecasting: Some uses and mis-uses of mathematics in economics. Discussion paper 530, Economics Department, Oxford University.

Hendry, D. F., and Ericsson, N. R. (1991). Modeling the demand for narrow money in the United Kingdom and the United States. *European Economic Review*, **35**, 833–886.

Hendry, D. F., and Ericsson, N. R. (eds.)(2001). *Understanding Economic Forecasts*. Cambridge, Mass.: MIT Press.

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 Mizon, G. E. (2010). On the mathematical basis of inter-temporal optimization. Discussion paper 497, Economics Department, Oxford.

Hendry, D. F., and Mizon, G. E. (2011a). An open-model forecast-error taxonomy. Working paper, Economics Department, Oxford University.

Hendry, D. F., and Mizon, G. E. (2011b). What needs rethinking in macroeconomics?. *Global Policy*, forthcoming.

Hendry, D. F., and Richard, J.-F. (1982). On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics*, **20**, 3–33.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In Brunner, K., and Meltzer, A. (eds.), *The Phillips Curve and Labor Markets*, Vol. 1 of *Carnegie-Rochester Conferences on Public Policy*, pp. 19–46. Amsterdam: North-Holland Publishing Company.

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

Mills, T. C. (2010). Bradford Smith: An econometrician decades ahead of his time. *Oxford Bulletin of Economics and Statistics*, DOI: 10.1111/j.1468–0084.2010.00615.x.

Smith, B. B. (1926). Combining the advantages of first-difference and deviation-from-trend methods of correlating time series. *Journal of the American Statistical Association*, **21**, 55–59.

Smith, B. B. (1927). Forecasting the volume and value of the cotton crop. *Journal of the American Statistical Association*, **22**, 442–459.

Smith, B. B. (1929). Judging the forecast for 1929. *Journal of the American Statistical Association*, **24**, 94–98.

Spanos, A. (2007). Curve-fitting, the reliability of inductive inference and the error-statistical approach. *Philosophy of Science*, **74**, 1046–1066.