Reformulating Empirical Macro-econometric Modelling

David F. Hendry
Economics Department, Oxford University

and

Grayham E. Mizon*
Economics Department, Southampton University.

1 Introduction

The policy implications derived from any estimated macro-econometric system depend on the formulation of its equations, the methodology used for the empirical modelling and evaluation, the approach to policy analysis, and the forecast performance. Drawing on recent results in the theory of forecasting, we question the role of 'rational expectations' in the first stage; criticise the present approach to testing economic theories prevalent in the profession; show that impulse-response methods of evaluating the policy implications of models are seriously flawed; and doubt the mechanistic derivation of forecasts from econometric systems. In their place, we propose that expectations should be treated as instrumental to agents' decisions; suggest a powerful new approach to the empirical modelling of econometric relationships; offer viable alternatives to studying policy implications; and discuss modifications to forecasting devices that can enhance their robustness to unanticipated structural breaks. We first sketch the arguments underlying our critical appraisals, then briefly describe the constructive replacements, before presenting more detailed analyses of these four issues.

Our approach builds on extensive research that has radically altered our understanding of the causes of forecast failure, the occurence of which was one of the driving forces behind the so-called 'rational expectations revolution' that replaced 'Keynesian' models. Forecast failure is a significant deterioration in forecast performance relative to the anticipated outcome, usually based on the historical performance of a model: systematic failure is the occurrence of repeated mis-forecasting. In modelling weakly-stationary processes (where means and variances are constant over time), three important results can be established. First, causal variables will outperform non-causal (i.e., variables that do not determine the series being forecast), both in terms of fit and when forecasting. Secondly, a model that in-sample fully exploits the available information (called congruent) and is at least as good as the alternatives (encompassing) will dominate in forecasting at all horizons. Thirdly, forecast failure will rarely occur, since the sample analyzed is 'representative' of the sample that needs to be forecast. Such theorems can be extended to non-stationary processes only when the model coincides with the data generation process (DGP). Unfortunately, the systematic mis-forecasting and forecast failure that has periodically blighted macroeconomics highlights a large discrepancy between such theory and practice, which is also visible in other disciplines: see e.g., Fildes and Makridakis (1995) and Makridakis and Hibon (2000). The taxonomy of sources of forecast errors in Clements and Hendry (1998, 1999a) has revealed that forecast-period shifts in deterministic factors (such as equilibrium means) are a dominant cause of systematic failure. Thus, forecast failure is not primarily due to model mis-specification, poor data, collinearity, a lack

*Financial support from the U.K. Economic and Social Research Council under grant L138251009 is gratefully acknowledged. We are pleased to acknowledge helpful comments from Chris Allsopp, Mike Clements, Jurgen Doornik, Bronwyn Hall, John Muellbauer and Bent Nielsen.

of parsimony, inconsistent estimation, or the Lucas (1976) critique of changing parameters: by themselves, none of these factors induces systematic failure. The key problem is the inherent non-stationary nature of economic data, even after differencing and cointegration transforms have removed unit roots. In a high-dimensional and evolving world, it is impossible to imagine an empirical model coinciding with the DGP. Consequently, one can disprove the basic theorem that forecasts based on causal variables will dominate those from non-causal. Our first critique follows – since 'rational expectations' claim to embody the actual conditional expectations, they do not have a sound theoretical basis in an economy subject to deterministic shifts. Further, in the presence of unmodelled deterministic shifts, models embodying previously-rational expectations will not forecast well in general.

Secondly, tests of economic theories based on whole-sample goodness of fit comparisions can be seriously misled by unmodelled deterministic shifts. This occurs because such shifts can be proxied by autoregressive dynamics, making lagged information from other variables appear irrelevant, so tests of theories – particularly of Euler equations – can be badly distorted. Further, cointegration often fails in the face of such breaks, so long-run relationships – often viewed as the statistical embodiment of economic theory predictions – then receive no support. Thus, our second critique follows: so long as the degree of non-congruence of a model is unknown, false theories can end being accepted, and useful ones rejected.

A necessary condition for both economic theories and macro-economic models to be of practical value is that their parameters remain constant over the relevant horizon and for the admissible range of policy changes to be implemented. Many structural breaks are manifest empirically, and so were easy to detect; deterministic shifts are a salient example. The class of breaks that are easy to detect comprises shifts in the unconditional expectations of non-integrated (denoted I(0)) components. However, it transpires that a range of parameter changes in econometric models cannot be easily detected by conventional statistical tests. This class includes changes that leave unaltered the unconditional expectations, even when dynamics, adjustment speeds, and intercepts are radically altered: illustrations are provided in Hendry and Doornik (1997). This leads to our third critique – impulse-response methods of evaluating the policy implications of models are dependent on the absence of such 'undetectable breaks', and so can be mis-leading in both sign and magnitude when non-deterministic shifts have occurred.

There is ample evidence that forecasts from econometric systems can err systematically in the face of deterministic shifts, so that they perform worse than 'naive' methods in forecasting competitions – and now theory to explain why has been developed: see e.g., Clements and Hendry (1999c). The implementation of cointegration may in practice have reduced the robustness of econometric-model forecasts to breaks, by ensuring they adjust back to the pre-existing equilibrium, even when that equilibrium has shifted. Mechanistic econometric-model based forecasts, therefore, are unlikely to be robust to precisely the form of shift that is most detrimental to forecasting. It is well known that devices such as intercept corrections can improve forecast performance (see e.g., Turner, 1990), but manifestly do not alter policy implications; and conversely, that time-series models with no policy implications might provide the best available forecasts. Hence our fourth critique – it is inadvisable to select policy-analysis models by their forecast accuracy: see Hendry and Mizon (2000).

The existence of these four problems implies that many macro-econometric models are incorrectly formulated and wrongly selected, with policy implications derived by inappropriate methods. Whilst we suspect that some amelioration arises in practice as a result of most macro-forecasters continuing to use intercept corrections to improve forecasts, the almost insupperable problems confronting some approaches to macro-economics remain. Fortunately though, effective alternatives exist.

First, since expectations are instrumental to

the decisions of economic agents, not an end in themselves, the devices that win forecasting competitions – which are easy to use and economical in information – suggest themselves as natural ingredients in agents' decision rules (possibly 'economically-rational expectations': see Feige and Pearce, 1976). We show that is the case, with the interesting implication that the resulting rules may not be susceptible to the Lucas (1976) critique, thus helping to explain its apparent empirical irrelevance: see Ericsson and Irons (1995).

Next, stimulated by Hoover and Perez (1999), Hendry and Krolzig (1999) investigate econometric model selection from a computer-automation perspective, focusing on general-to-specific reduction approaches, embodied in the program *PcGets* (**ge**neral–**t**o–**s**pecific: see Krolzig and Hendry, 2000). In Monte Carlo experiments, *PcGets* recovers the DGP with remarkable accuracy, having empirical size and power close to what one would expect if the DGP were known, suggesting that search costs are low. Thus, a general-to-specific modelling strategy that starts from a congruent general model and requires congruence and encompassing throughout the reduction process offers a powerful method for selecting models. This outcome contrasts with beliefs in economics about the dangers of 'data mining'. Rather, it transpires that the difficult problem is to retain relevant variables, not eliminate spurious ones. In addition, the existence of *PcGets* makes software available for modellers to take advantage of this model-selection strategy.

Thirdly, there is a strong case for using open, rather than closed, macro-econometric systems, particularly those which condition on policy instruments. Modelling open systems has the advantage that amongst their parameters are the dynamic multipliers which provide estimates of the responses of targets to policy changes. Further, it is difficult to build models of variables such as interest rates, tax rates, and exchange rates, which are either policy instruments or central to the determination of targets. Since many policy decisions entail shifts in the unconditional means of policy instruments, corresponding shifts in the targets' unconditional means are required for policy to be effective. Whenever there is co-breaking between these means (see Clements and Hendry, 1999a, ch. 9), reliable estimates of the policy responses can be obtained from the model of the targets conditioned on the instruments, despite the probable absence of weak exogeneity of policy instruments for the parameters of interest in macro-models due to mutual dependence on previous disequilibria. The existence of co-breaking between the means of the policy instruments and targets is testable, and moreover, is anyway necessary to justify impulse-response analysis (see Hendry and Mizon, 1998).

Finally, there are gains from separating policy models – to be judged by their ability to deliver accurate advice on the responses likely from policy changes – from forecasting models, to be judged by their forecast accuracy and precision. No forecast can be robust to unanticipated events that occur after its announcement, but some are much more robust than others to unmodelled breaks that occurred in the recent past. Since regime shifts and major policy changes act as breaks to models that do not embody the relevant policy responses, we discuss pooling robust forecasts with scenario differences to avoid both traps.

We conclude that the popular methodologies of model formulation, modelling and testing, policy evaluation, and forecasting may prejudice the accuracy of implications derived from macro-econometric models. Such a danger confronted earlier generations of macro-models: the use of dynamic simulation to select systems was shown in Hendry and Richard (1982) to have biased the choice of models to ones which over-emphasized the role of unmodelled ('exogenous') variables at the expense of endogenous dynamics, with a consequential deterioration in forecast performance, and mis-leading estimates of speeds of policy responses. As in that debate, we propose positive antidotes to each of the major lacunae in existing approaches. The detailed analyses have been presented in other publications: here we seek to integrate and explain their implications for macro-econometric modelling.

The structure of the paper is as follows. Since

we attribute a central role to forecast-period shifts in deterministic factors as the cause of forecast failure, section 2 first reviews its determinants. Section 3 derives the implications for 'rational expectations', and suggests alternatives that are both feasible and more robust to breaks. Then section 4 describes model-selection procedures, and contrasts these with tests of theory-based propositions. Section 6 turns to model selection for forecasting, then section 7 considers policy analyses based on impulse responses and on dynamic multipliers, and justifies the role of congruent modelling. Section 8 concludes.

2 Causes of forecast failure

In a constant-parameter, stationary world, forecast failure should rarely occur: the in-sample and out-of-sample fits will be similar because the data properties are unchanged. As discussed in Miller (1978), stationarity ensures that, on average (i.e., excluding rare events), an incorrectly-specified model will forecast within its anticipated tolerances (providing these are correctly calculated). Although a mis-specified model could be beaten by methods based on correctly-specified equations, it will not suffer excessive forecast failure purely because it is mis-specified. Nevertheless, since a congruent, encompassing model will variance-dominate in-sample, it will continue to do so when forecasting under unchanged conditions. Thus, adding causal variables will improve forecasts on average; adding non-causal variables (i.e., variables that do not enter the DGP) will only do so when they proxy for omitted causal variables.

Empirical models are usually data-based (selected to match the available observations), which could induce some overfitting, but should not produce systematic forecast failure (see Clements and Hendry, 1999b). Conversely, when the data properties over the forecast horizon differ from those in-sample – a natural event in non-stationary processes – forecast failure will result. The latter's regular occurrence is strong evidence for pandemic non-stationarity in economics, an unsurprising finding given the manifest legislative, social, technological and political changes witnessed over modern times (and indeed through most of history).

Once such non-stationarity is granted, many 'conventional' results that are provable in a constant-parameter, stationary setting change radically. In particular, since the future will not be like the present or the past, two important results can be established in theory, and demonstrated in practice: the potential forecast dominance of models using causal variables by those involving non-causal variables; and of in-sample well-specified models by badly mis-specified ones. Clements and Hendry (1999a) provide several examples. Together, such results remove the theoretical support for basing forecasting models, and hence agents' expectations formation, on the in-sample conditional expectation given available information. We develop this analysis in section 3. Moreover, these two results potentially explain why over-differencing and intercept corrections – both of which introduce non-causal variables into forecasting devices – could add value to model-based forecasts: this aspect is explored in section 6. Finally, a failure to model the relevant non-stationarities can distort in-sample tests, and lead to incorrect inferences about the usefulness or otherwise of economic theories: that is the topic of section 4.

Not all forms of non-stationarity are equally pernicious. For example, unit roots generate stochastic trends in data series, which thereby have changing means and variances, but nevertheless seem relatively benign. This form of non-stationarity can be removed by differencing or cointegration transformations, and often, it may not even matter greatly whether or not those transforms are imposed (see e.g., Sims, Stock and Watson, 1990, for estimation, and Clements and Hendry, 1998, for forecasting). Of course, omitting dynamics could induce 'nonsense regressions', but provided appropriate critical values are used, even that hypothesis is testable – and its rejection entails cointegration.

Further, while structural changes in a model's

parameters also reflect data non-stationarity, many such changes do not induce forecast failure, whereas others lead to serious problems. To understand why, we must dissect the ingredients of econometric models. In general, these have three main components: deterministic terms, namely variables whose future values are known; observed stochastic variables with unknown future values; and unobserved errors all of whose values are unknown. Most relationships in models involve all three components, and these could be: misspecified; poorly estimated; based on inaccurate data; selected by inappropriate methods; involve collinear variables with non-parsimonious formulations; and suffer structural breaks. Given the complexity of modern economies, most of these possible causes of forecast failure will be present in any empirical macro-model, and will reduce forecast performance by increasing inaccuracy and imprecision. However, and somewhat surprisingly, most combinations do not induce systematic forecast failure: analysis implicates shifts in deterministic terms as the primary cause of forecast failure. The documentation supporting these claims is presented in Clements and Hendry (1998, 1999a).

Consider an h-step ahead forecast made at time T, denoted $\widehat{\mathbf{y}}_{T+h|T}$, for a vector of n_y variables $\mathbf{y}_{T+h}$. The difference between the eventual outcomes and the forecast values is the vector of forecast errors $\mathbf{e}_{T+h} = \mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h|T}$, and this can be decomposed into the various mistakes and unpredictable elements. Doing so delivers a forecast-error taxonomy, partitioned appropriately into deterministic, observed stochastic, and innovation-error influences. Although the decomposition is not unique, it can be expressed in nearly-orthogonal effects corresponding to influences on forecast-error means and variances respectively. The former involve all the deterministic terms; the latter the remainder. We now briefly consider these major categories of error, commencing with mean effects, then turn to variance components.

2.1 Deterministic factors

Systematic forecast-error biases derive from deterministic factors being mis-specified, mis-estimated, or non-constant. The simplest example is omitting a trend; or when a trend is included, under-estimating its slope; or when the slope is correct, experiencing a shift in the growth rate. Any of these will lead to a systematic, and possibly increasing, divergence between outcomes and forecasts. However, there is an important distinction between the roles of intercepts, trends etc., in models, and any resulting deterministic shifts, as we will now explain. To clarify the roles of deterministic, stochastic, and error factors, we first consider a static equation with a mean shift, then turn to a scalar regression with a changed parameter, and finally investigate an autoregression where the dynamics alter.

Let the in-sample DGP, for $t = 1, \ldots, T$, be:

$$y_t = \mu + \epsilon_t \text{ where } \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right], \qquad (1)$$

when $\mathsf{IN}\left[0, \sigma_\epsilon^2\right]$ denotes an independent normal error having mean zero and constant variance σ_ϵ^2, so that $\mathsf{E}[y_t] = \mu$ (the equilibrium mean) and $\mathsf{V}[y_t] = \sigma_\epsilon^2$ (unconditional variance). The empirically-relevant case is when the variable labelled y_t is actually a log-difference, so μ defines the mean growth rate. Consider using the estimated DGP (1) as the forecasting model. For simplicity, we assume unbiased parameter estimates, so $\mathsf{E}\left[\widehat{\mu}\right] = \mu$, and neglect estimation uncertainty. Then, with an exactly-measured forecast origin at time T, (7) produces the h-step ahead forecast sequence:

$$\widehat{y}_{T+h|T} = \widehat{\mu}. \qquad (2)$$

However, over the forecast period, $h = 1, \ldots, H$, there is a shift in the parameters of the process, so that in fact:

$$y_{T+h} = \mu^* + \epsilon_{T+h},$$

where $\epsilon_{T+h} \sim \mathsf{D}[0, (\sigma_\epsilon^*)^2]$. Thus, the DGP generates the future outcomes by:

$$y_{T+h} = \mu^* + \epsilon_{T+h}.$$

The resulting sequence of forecast errors $e_{T+h|T} = y_{T+h} - \widehat{y}_{T+h|T}$ is:

$$e_{T+h|T} = (\mu^* - \mu) + \epsilon_{T+h}. \tag{3}$$

From (3):

$$\mathsf{E}_{T+h}\left[e_{T+h|T}\right] = \mu^* - \mu,$$

which is non-zero if and only if $\mu^* \neq \mu$. Such a 'deterministic shift' is pernicious when $\mu^* - \mu$ increases by several σ_ϵ, but is then usually easy to detect. Notice that most econometric growth-rate equations have $\mu \simeq \sigma_\epsilon$ (often around 1%), so the requirement that $\mu^* - \mu = 2\sigma_\epsilon$ is actually very strong: e.g., a doubling of the trend rate of growth. Consequently, even moderate trend shifts can be hard to detect till quite a few periods have elapsed.

Next, consider a regression as the in-sample DGP:

$$y_t = \beta x_t + \epsilon_t \text{ where } \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right], \tag{4}$$

where x_t has mean zero and variance σ_x^2, with known future values. As before, let β change to β^* over the forecast horizon. The resulting sequence of forecast errors, using $\widehat{y}_{T+h|T} = \beta x_{T+h}$ (4) as the forecasting equation, is $e_{T+h|T} = y_{T+h} - \widehat{y}_{T+h|T}$ so:

$$\begin{aligned} e_{T+h|T} &= \beta^* x_{T+h} + \epsilon_{T+h} - \beta x_{T+h} \\ &= (\beta^* - \beta)\, x_{T+h} + \epsilon_{T+h}. \end{aligned} \tag{5}$$

Unlike (3), both terms in (5) have zero expectations, so no forecast bias results. There is an increase in the variance, from σ_ϵ^2 to $(\beta^* - \beta)^2 \sigma_x^2 + \sigma_\epsilon^2$, and the detectability (or otherwise) of the break depends on how much the variance increases. For $(\beta^* - \beta)^2 = 4\sigma_\epsilon^2$ (say), then the ratio is $1 + 4\sigma_x^2$ which can be difficult to detect against the background noise (see e.g, Hendry and Doornik, 1997, for simulation illustrations).

Finally, we combine these two effects in a scalar autoregression:

$$y_t = \gamma + \rho y_{t-1} + \epsilon_t \text{ where } \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right], \tag{6}$$

with $|\rho| < 1$ to ensure in-sample stationarity. Then $\mathsf{E}[y_t] = \gamma/(1-\rho) = \mu$ (the equilibrium mean) and $\mathsf{V}[y_t] = \sigma_\epsilon^2/(1-\rho^2)$ (long-run variance) so (6) can be written as the homogeneous difference relation:

$$(y_t - \mu) = \rho\,(y_{t-1} - \mu) + \epsilon_t. \tag{7}$$

For simplicity, we assume unbiased parameter estimates, so $\mathsf{E}[\widehat{\gamma}] = \gamma$ and $\mathsf{E}[\widehat{\rho}] = \rho$ (a symmetric error distribution would justify such approximations), and neglect estimation uncertainty. Then, with an exactly-measured forecast origin at time T, (7) produces the h-step ahead forecast sequence:[1]

$$\widehat{y}_{T+h|T} = \widehat{\gamma} + \widehat{\rho}\widehat{y}_{T+h-1|T} \simeq \gamma + \rho\widehat{y}_{T+h-1|T}. \tag{8}$$

Rewrite (8) as we did for (7):

$$\widehat{y}_{T+h|T} - \mu = \rho\left(\widehat{y}_{T+h-1|T} - \mu\right),$$

and recursively solve backwards to the forecast origin:

$$\widehat{y}_{T+h|T} - \mu = \rho^h\left(y_T - \mu\right). \tag{9}$$

Then (9) shows how the h-step ahead forecasts are generated.

As before, there is a shift in the parameters of the processover the forecast period, so that:

$$y_{T+h} = \gamma^* + \rho^* y_{T+h-1} + \epsilon_{T+h},$$

where $\epsilon_{T+h} \sim \mathsf{D}[0, (\sigma_\epsilon^*)^2]$, or:

$$y_{T+h} - \mu^* = \rho^*\left(y_{T+h-1} - \mu^*\right) + \epsilon_{T+h},$$

with $\mu^* = \gamma^*/(1-\rho^*)$. Thus, the DGP generates the future outcomes by:

$$y_{T+h} - \mu^* = (\rho^*)^h\left(y_T - \mu^*\right) + \sum_{i=0}^{h-1}(\rho^*)^i\,\epsilon_{T+h-i}.$$

The resulting sequence of forecast errors $e_{T+h|T} = y_{T+h} - \widehat{y}_{T+h|T}$ is:

$$\begin{aligned} e_{T+h|T} &= \mu^* + (\rho^*)^h\left(y_T - \mu^*\right) \\ &\quad + \sum_{i=0}^{h-1}(\rho^*)^i\,\epsilon_{T+h-i} \\ &\quad - \left[\mu + \rho^h\left(y_T - \mu\right)\right], \end{aligned}$$

[1] Forecast origin mis-measurement can be pernicious, because the incorrect starting level is 'carried forward' in dynamic models.

so collecting terms:

$$\begin{aligned} e_{T+h|T} &= \left(1-(\rho^*)^h\right)(\mu^*-\mu) + \\ &\quad \left((\rho^*)^h-\rho^h\right)(y_T-\mu) \\ &\quad +\sum_{i=0}^{h-1}(\rho^*)^i\,\epsilon_{T+h-i}. \end{aligned} \tag{10}$$

Of the three terms in (10), the last two have expectations of zero, so do not contribute to the mean, leaving:

$$\mathsf{E}_{T+h}\left[e_{T+h|T}\right] = \left(1-(\rho^*)^h\right)(\mu^*-\mu).$$

This first term is non-zero if and only if $\mu^* \neq \mu$, irrespective of whether or not ρ changes (for $\rho^* \neq 1$, as assumed, to keep the DGP I(0)). Indeed, when $y_T \simeq \mu$, the contribution to e_{T+h} from a change in ρ will be negligible from the second term; and if $|\rho^*| < |\rho|$, the error accumulation will be slower than anticipated from the third term, although shifts in σ_ϵ^2 could offset that. A strong, and corroborated, prediction from (10) is that shifts in both γ^* and ρ^* which leave μ unchanged will not induce forecast failure, and tests will be relatively powerless to detect that anything has changed. Indeed, the situation where μ is constant is precisely the same as when all the means are zero: Hendry and Doornik (1997) and section 7.2 below provide the details, and lead to the conclusion that the deterministic term of relevance is μ, not the original intercept γ, and that shifts in the equilibrium mean μ to μ^* are the primary source of forecast failure.

To illustrate that an 'incorrect' model can outperform the in-sample DGP in forecasting, we return to the simplest case in (1). When μ shifts to μ^*, (3) shows that the expected forecast error sequence from using the in-sample DGP will be $(\mu^*-\mu)$. That remains true when the forecast origin moves through time to $T+1$, $T+2$ etc.: because the forecasting model remains unchanged, so do the average forecast errors. Consider, instead, using the 'mis-specified' predictor $\widetilde{y}_{T+h|T} = y_T$. The resulting sequence of forecast errors will be similar to $e_{T+h|T}$ when the origin is T: unanticipated shifts after forecasting are bound to be pernicious for all methods. However, when forecasting from time $T+1$ onwards, a very different result ensues for $\widetilde{e}_{T+h|T+1} = y_{T+h} - \widetilde{y}_{T+h|T+1}$ as:

$$\begin{aligned} \widetilde{e}_{T+h|T+1} &= y_{T+h} - y_{T+1} \\ &= \mu^* + \epsilon_{T+h} - \mu^* - \epsilon_{T+1} \\ &= \Delta_{h-1}\epsilon_{T+h}, \end{aligned} \tag{11}$$

which has a mean of zero, despite the deterministic shift. Thus, on a bias criterion, $\widetilde{y}_{T+h|T+1}$ outperforms the in-sample DGP (and would win on mean-square error if the forecast-error variance of $2\left(\sigma_\epsilon^*\right)^2$ on (11) was smaller than $\left(\sigma_\epsilon^*\right)^2 + (\mu^*-\mu)^2$). Dynamics make the picture more complicated, but the principle still applies.

An interesting, and much studied, example of a deterministic shift concerns forecast failure in a model of narrow money (M1) after the Banking Act of 1984, which permitted interest payments on current accounts in exchange for all interest payments being after the deduction of 'standard rate' tax. The own rate of interest (R_o) changed from zero to near the value of the competitive rate (R_c: about 12 per cent per annum at the time) in about 6 quarters, inducing very large inflows to M1. Thus, a large shift occurred in the mean opportunity cost of holding money, namely a deterministic shift from R_c to $R_c - R_o$. Pre-existing models of M1 – which used the outside rate of interest R_c as the measure of opportunity cost – suffered marked forecast failure, which persisted for many years after the break. Models that correctly re-measured the opportunity cost by $R_c - R_o$ continued to forecast well, once the break was observed, and indeed had the same estimated parameter values after the break as before. However, methods analogous to $\widetilde{y}_{T+h|T}$ also did not suffer forecast failure: see Clements and Hendry (1999c) for details and references.

In general, let $\mathsf{E}_t\left[\mathbf{y}_t|\mathbf{Y}_{t-1}\right]$ be the conditional expectation (where that exists) formed at time t, given the history of the process, denoted by $\mathbf{Y}_{t-1}$. The key drivers of forecast failure are mis-specification of, uncertainty in, or changes to $\mathsf{E}_{T+h}\left[\mathbf{y}_{T+h}|\mathbf{Y}_T\right]$ relative to the mean forecast $\mathsf{E}_T\left[\widehat{\mathbf{y}}_{T+h|T}|\mathbf{Y}_T\right]$, where $\mathbf{Y}_T$ is

information available at the forecast origin. Then, $\mathsf{E}_{T+h}\left[\mathbf{y}_{T+h}|\mathbf{Y}_T\right] - \mathsf{E}_T\left[\widehat{\mathbf{y}}_{T+h|T}|\mathbf{Y}_T\right]$ will differ from zero when $\widehat{\mathbf{y}}_{T+h|T}$ is a biased estimator of $\mathsf{E}_T\left[\mathbf{y}_{T+h}|\mathbf{Y}_T\right]$, or when E_{T+h} shifts unexpectedly relative to E_T. Because forecast failure is usually judged relative to in-sample behaviour, the latter is the dominant cause. However, misestimation of coefficients of deterministic terms could be deleterious to forecast accuracy when $\mathsf{E}_{T+h}\left[\mathbf{y}_{T+h}|\mathbf{Y}_T\right] - \widehat{\mathbf{y}}_{T+h|T}$ is large by chance.

2.2 Stochastic factors

As demonstrated in section 7.2, shifts in the coefficients of zero-mean variables have a surprisingly small impact on forecasts (measured by the inability of parameter-constancy tests to detect the break). Thus, such shifts seem an unlikely explanation for observed forecast failure.

Omitting zero-mean stochastic components is unlikely to be a major source of forecast failure, but could precipitate failure if stochastic mis-specification resulted in deterministic shifts elsewhere in the economy affecting the model. Equally, the false inclusion of zero-mean stochastic variables is a secondary problem, whereas wrongly including regressors which experienced deterministic shifts could have a marked impact on forecast failure as the model mean shifts although the data mean does not.

Estimation uncertainty in the parameters of stochastic variables also seems to be a secondary problem, as such errors add variance terms of $\mathsf{O}(1/T)$ for stationary components. Neither collinearity nor a lack of parsimony *per se* seem likely culprits, although interacting with breaks occurring elsewhere in the economy could induce problems.

Finally, better-fitting models have smaller error accumulation, but little can be done otherwise about the contribution from that source.

That concludes the 'scene setting' analysis, summarized as: deterministic shifts of the data relative to the model are the primary source of forecast failure. Monte Carlo evidence presented in several papers bears out the analytics: parameter non-constancy and forecast-failure tests reject for small changes in equilibrium means, but not for substantial changes in dynamics, or in all parameters when that leave equilibrium means unaltered (all measured as a proportion of σ_ϵ).

3 'Rational expectations'

When unanticipated deterministic shifts make an economy non-stationary, the formation of 'rational expectations' requires agents to know:

- all the relevant information ($\mathbf{Y}_{t-1}$ above);
- how every component enters the joint data density (i.e, $\mathsf{E}_{\cdot}\left[\cdot|\mathbf{Y}_{t-1}\right]$);
- the changes at each point in time (i.e, $\mathsf{E}_t\left[\cdot|\cdot\right]$).

In terms of our scalar example, the model error $e_{T+h|T}$ in (10) equals the 'rational expectations' error $\sum_{i=0}^{h-1}\left(\rho^*\right)^i \epsilon_{T+h-i}$ if and only if every other term is zero. Yet most shifts, and many of their consequences, cannot be anticipated: assuming knowledge of current and *future* deterministic shifts is untenable. Otherwise, the resulting forecasting device can be dominated by methods which use no causally-relevant variables. Thus, it ceases to be rational to try and form expectations using the current conditional expectation when that will neither hold in the relevant future, nor forecast more accurately than other devices. Agents will learn that they do better forming expectations from 'robust forecasting rules' – which adapt rapidly to deterministic shifts. These may provide an example of 'economically-rational expectations' as suggested by Feige and Pearce (1976), equating the marginal costs and benefits of improvements in the accuracy of expectations: Hendry (2000b) provides a more comprehensive discussion.

Robust forecasting rules need not alter with changes in policy. Of course, if agents fully understood a policy change and its implications, they would undoubtedly be able to forecast better: but that would require the benefits of doing so to exceed the costs. The problem for agents is compounded by the fact that many major policy changes occur in turbulent times, precisely when it

most difficult to form 'rational expectations', and when robust predictors may outperform. Thus, many agents may adopt the adaptive rules discussed above, consistent with the lack of empirical evidence in favour of the Lucas (1976) critique reported in Ericsson and Irons (1995). Consequently, if an econometric model used Δx_t as a replacement for the expected change $\Delta x^e_{t+1|t}$ when agents used robust rules, then the model's parameters need not change even after forecast failure occurred. Alternatively, the unimportant consequences for forecasting of changes in reaction coefficients, rather than their absence, could account for the lack of empirical evidence that the critique occurs, but anyway reduces its relevance.

4 Empirical model selection

We now discuss the costs of search, distinguish them from the (unavoidable) costs of inference, explain why a general-to-specific modelling strategy – as implemented in *PcGets* – is able to perform so well despite the problem of 'data mining', and suggest that the practical problem is to retain relevant variables, not eliminate spurious ones.

Statistical inference is always uncertain because of type I and type II errors (rejecting the null when it is true; and failing to reject the null when it is false respectively). Even if the DGP were derived *a priori* from economic theory, an investigator could not *know* that such a specification was 'true', and inferential mistakes will occur when testing hypotheses about it. This is a 'pretest' problem: beginning with the truth and testing it will sometimes lead to false conclusions. 'Pretesting' is known to bias estimated coefficients, and may distort inference (see *inter alia*, Judge and Bock, 1978). Of course, the DGP specification is never known in practice, and since 'theory dependence' in a model has as many drawbacks as 'sample dependence', data-based model-search procedures are used in practice, thus adding search costs to the costs of inference. A number of arguments point towards the advantages of 'general-to-specific' searches.

Statistical analyses of repeated testing provide a pessimistic background: every test has a non-zero null rejection frequency ('size'), so type I errors accumulate. Size could be lowered by increasing the significance levels of selection tests, but only at the cost of reducing power to detect the influences that really matter. The simulation experiments in Lovell (1983) suggested that search had high costs, leading to an adverse view of 'data mining'. However, he evaluated outcomes against the truth, compounding costs of inference with costs of search. Rather, the key issue for any model-selection procedure is: how costly is it to search across many alternatives relative to commencing from the DGP? As we now discuss, it is feasible to lower size and raise power simultaneously by improving the search algorithm.

First, White (1990) showed that with sufficiently-rigorous testing and a large enough data sample, the selected model will converge to the DGP, so selection error is a 'small-sample' problem, albeit a difficult and prevalent one. Secondly, Mayo (1981) noted that diagnostic testing was effectively independent of the sufficient statistics from which parameter estimates are derived, so would not distort the latter. Thirdly, since the DGP is obviously congruent with itself, congruent models are the appropriate class within which to search. This argues for commencing the search from a congruent model. Fourthly, encompassing – explaining the evidence relevant for all alternative models under consideration– resolves 'data mining' (see Hendry, 1995) and delivers a dominant outcome. This suggests commencing from a general model that embeds all relevant contenders. Fifthly, any model-selection process must avoid getting stuck in search paths that inadvertently delete relevant variables, thereby retaining many other variables as proxies. The resulting approach of sequentially simplifying a congruent general unrestricted model (GUM) to obtain the maximal acceptable reduction, is called general-to-specific (*Gets*).

To evaluate the performance of *Gets* modelling procedures, Hoover and Perez (1999) reconsidered the Lovell (1983) experiments, searching for a sin-

gle conditional equation (with 0 to 5 regressors) from a large macroeconomic database (containing up to 40 variables, including lags). By following several reduction search paths – each terminated by either no further feasible reductions or significant diagnostic test outcomes – they showed how much better the structured *Gets* approach was than any method Lovell considered, suggesting that modelling *per se* need not be bad. Indeed, the overall 'size' (false null rejection frequency) of their selection procedure was close to that expected without repeated testing, yet the power was reasonable.

Building on their findings, Hendry and Krolzig (1999) and Krolzig and Hendry (2000) developed the Ox (see Doornik, 1999) program *PcGets* which first tests the congruency of a GUM, then conducts pre-selection tests for 'highly irrelevant' variables at a loose significance level (25% or 50%, say), and simplifies the model accordingly. It then explores many selection paths to eliminate statistically-insignificant variables on F- and t-tests, applying diagnostic tests to check the validity of all reductions, thereby ensuring a congruent final model. All the terminal selections resulting from search paths are stored, and encompassing procedures and information criteria select between the contenders. Finally, sub-sample significance is used to assess the reliability of the resulting model choice. In Monte Carlo experiments, *PcGets* recovers the DGP with power close to what one would expect if the DGP were known, and empirical size often below the nominal, suggesting that search costs are in fact low. In the 'classic' experiment in which the dependent variable is regressed on 40 irrelevant regressors, *PcGets* correctly finds the null model about 97% of the time for the Lovell database.

Some simple analytics proposed in Hendry (2000a) suggest why *PcGets* performs well, even though the following analysis ignores pre-selection, search paths and diagnostic testing (all of which improve the algorithm). An F-test against the GUM using critical value c_γ would have size $\mathsf{P}(\mathsf{F} \geq c_\gamma) = \gamma$ under the null if it were the only test implemented. For k regressors, the probability of retaining no variables from t-tests at size α is:

$$\mathsf{P}(|\mathsf{t}_i| < c_\alpha \; \forall i = 1, \ldots, k) = (1-\alpha)^k, \quad (12)$$

where the average number of variables retained then is:

$$n = k\alpha. \quad (13)$$

Combined with the F-test of the GUM, the probability π of correctly selecting the null model is no smaller than:

$$\pi = (1-\gamma) + \gamma(1-\alpha)^k. \quad (14)$$

For $\gamma = 0.05$ and $\alpha = 0.01$, when $k = 40$, then $\pi = 0.98$ and $n = 0.4$. Although falsely rejecting the null on the F-test signals that spurious significance lurks, so (13) will understate the number of regressors then retained, nevertheless, eliminating adventitiously-significant (spurious) variables is not the real problem in empirical modelling.

Indeed, the focus in earlier research on 'overfitting' – reflecting inferior algorithms – has misdirected the profession's attention. The really difficult problem is retaining the variables that matter. Consider an equation with six relevant regressors, all with (absolute) t-values of 2 on average (i.e., $\mathsf{E}[|\mathsf{t}_i|] = 2$). The probability in any given sample that each observed $|\widehat{\mathsf{t}}_i| \geq c_\alpha = 2$ (say) is approximately 0.5, so even if one began with the DGP, the probability of retaining all six is:

$$\begin{aligned} \mathsf{P}\left(|\widehat{\mathsf{t}}_i| \geq c_\alpha \; \forall i = 1, \ldots, k \mid |\mathsf{t}_i| = 2\right) &= 0.5^6 \\ &\simeq 0.016. \end{aligned}$$

Using 1% significance lowers this to essentially zero. Surprisingly, even if every $\mathsf{E}[|\mathsf{t}_i|] = 3$, the chances of keeping the DGP specification are poor:

$$\mathsf{P}\left(|\widehat{\mathsf{t}}_i| \geq c_\alpha \; \forall i \mid |\mathsf{t}_i| = 3\right) = 0.84^6 \simeq 0.35.$$

Thus, the costs of inference are high in such full-sample testing, and will lead to under-estimating model size. An alternative, block-testing, approach discussed in Hendry (2000a) seems able to improve the power substantially.

Nevertheless, many empirical equations have many regressors. This is probably due to the high average t-values found in economics:

$$\mathsf{P}\left(|\widehat{\mathsf{t}}_i| \geq c_\alpha \; \forall i \mid |\mathsf{t}_i| = 5\right) \simeq 0.989^6 \simeq 0.935,$$

(so almost all will always be retained), and not to selection biases as shown above. Even selecting by t-testing from 40 candidate regressors at 5% would only deliver 2 significant variables on average. We conclude that models with many significant variables correctly represent some of the complexity of aggregate economic behaviour and not 'overfitting'.

5 Model selection for theory testing

Although not normally perceived as a 'selection' issue, tests of economic theories based on whole-sample goodness of fit comparisions involve selection, and can be seriously misled by deterministic shifts. Three examples affected by unmodelled shifts are: lagged information from other variables appearing irrelevant, affecting tests of Euler equation theories; cointegration failing so long-run relationships receive no empirical support; and tests of forecast efficiency rejecting because of residual serial correlation induced *ex post* by an unpredictable deterministic shift. We address these in turn.

The first two are closely related, so our illustration concerns tests of the implications of the Hall (1978) Euler-equation consumption theory when credit rationing changes, as happened in the UK (see Muellbauer, 1994). The log of real consumer's expenditure on non-durables and services (c) is not cointegrated with the log of real personal disposable income (y) over 1962(2)–1992(4): a unit-root test using 5 lags of each variable, a constant and seasonals delivers $\mathsf{t}_{ur} = 0.97$ so does not reject (see Banerjee and Hendry, 1992, and Ericsson and MacKinnon, 1999, on the properties of this test). Nevertheless, the solved long-run relation is:

$$c = \underset{(0.99)}{-\ 0.53} \ + \ \underset{(0.10)}{0.98}\ y + \textit{Seasonals}. \qquad (15)$$

Lagged income terms are individually ($max\ \mathsf{t} = 1.5$) and jointly ($\mathsf{F}(5, 109) = 1.5$) insignificant in explaining $\Delta_4 c_t = c_t - c_{t-4}$. Such evidence appears to support the Hall life-cycle model, which entails that consumption changes are unpredictable, with permanent consumption proportional to fully-anticipated permanent income. As fig. 1a shows for annual changes, the data behaviour is at odds with the theory after 1985, since consumption first grows faster than income for several years, then falls faster – far from smoothing. Moreover, the large departure from equilibrium in (15) is manifest in panel b, resulting in a marked deterioration in the resulting (fixed-parameter) 1-step forecast errors from the model in Davidson, Hendry, Srba and Yeo (1978) after 1984(4) (the period to the right of the vertical line in fig. 1c). Finally, an autoregressive model for $\Delta\Delta_4 c_t = \Delta_4 c_t - \Delta_4 c_{t-1}$ produces 1-step forecast errors which are smaller than average after 1984(4), consistent with a deterministic shift around the mid 1980s (see Hendry, 1994, and Muellbauer, 1994, for explanations based on financial deregulation inducing a major reduction in credit rationing), which neither precludes the *ex ante* predictability of consumption from a congruent model, nor consumption and income being cointegrated. The apparent insignificance of additional variables may be an artefact of misspecifying a crucial shift, so the 'selected' model is not valid support for the theory. Conversely, non-causal proxies for the break may seem significant. Thus, models used to test theories should first be demonstrated to be congruent and encompassing.

We must stress that our example is not an argument against econometric modelling. While $\Delta\Delta_4 c_{t-1}$ may be a more robust forecasting device than the models extant at the time, it is possible in principle that the appropriate structural model – which built in changes in credit markets – would both have produced better forecasts and certainly better policy. For example, by 1985, building society data suggested that mortgages were available on much easier terms than had been the case historically, and 'housing-equity withdrawal' was already causing concern to policy makers. Rather, we are criticising the practice of 'testing theories' without first testing that the model used is a congruent and undominated representation, precisely because 'false' but robust predictors exist, and deterministic shifts appear to occur intermittently.

The same data illustrate the third mistake: re-

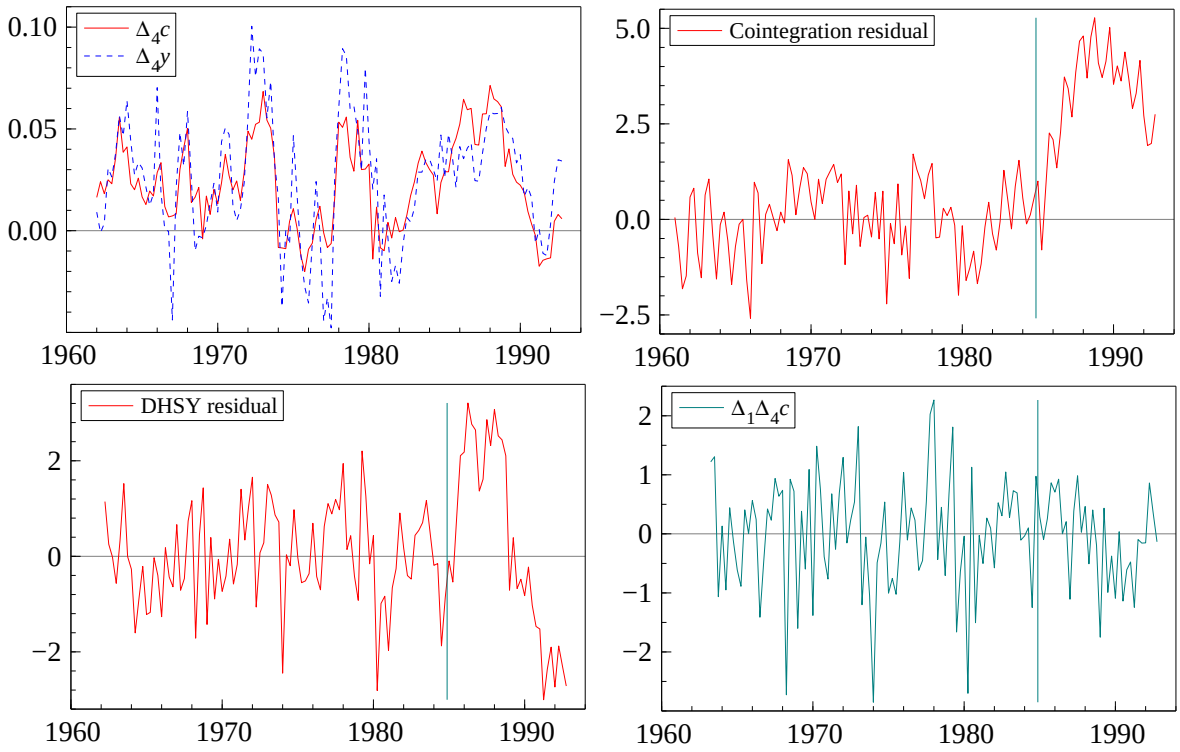

Figure 1 UK real consumers' expenditure and income with model residuals.

jecting forecast efficiency because of residual serial correlation induced *ex post* by an unpredictable deterministic shift. A model estimated prior to such a shift could efficiently exploit all available information; but if a shift was unanticipated *ex ante*, and unmodelled *ex post*, it would induce whole-sample residual serial correlation, apparently rejecting forecast efficiency. Of course, the results correctly reject 'no mis-specification'; but as no-one could have outperformed the in-sample DGP without prescience, the announced forecasts were not *ex ante* inefficient in any reasonable sense.

6 Model selection for forecasting

Forecast performance in a world of deterministic shifts is not a good guide to model choice, unless the sole objective is short-term forecasting. This is because models which omit causal factors and cointegrating relations, by imposing additional unit roots, may adapt more quickly in the face of unmodelled shifts, and so provide more accurate forecasts after breaks. We referred to this above as 'robustness' to breaks.

The admissible deductions on observing either the presence or absence of forecast failure are rather stark, particularly for general methodologies which believe that forecasts are the appropriate way to judge empirical models. In this setting of structural change, there may exist non-causal models (i.e., models none of whose 'explanatory' variables enter the DGP) that do not suffer forecast failure, and indeed may forecast absolutely more accurately on reasonable measures, than previously congruent, theory-based models. Conversely, *ex ante* forecast failure may merely reflect inappropriate measures of the inputs, as we showed with the example of 'opportunity-cost' affecting M1: a model that suffers severe forecast failure may nonetheless have constant parameters on *ex post* re-estimation. Consequently, neither relative success nor failure in forecasting is a reliable basis for selecting between models – other than for forecasting purposes. Apparent failure on forecasting need have no implications for the goodness of a model, nor its theoretical underpinnings, as it may arise from incorrect data, that are later corrected.

Some forecast failures will be due to model

mis-specification, such as omitting a variable whose mean alters; and some successes to having well-specified models that are robust to breaks. The problem is discriminating between such cases, since the event of success or failure *per se* is insufficient information. Because the future can differ in unanticipated ways from the past in non-stationary processes, previous success (failure) does not entail the same will be repeated later. That is why we have stressed the need for 'robust' or adaptable devices in the forecasting context. If it is desired to use a 'structural' or econometric model for forecasting, then there are many ways of increasing its robustness, as discussed in Clements and Hendry (1999a). The most usual are 'intercept corrections', which adjust the fit at the forecast origin to exactly match the data, and thereby induce the differences of the forecast errors that would otherwise have occurred. Such an outcome is close to that achieved by modelling the differenced data, but retains the important influences from disequilibria between levels. Alternatively, and closely related, one could update the equilibrium means and growth rates every period, placing considerable weight on the most recent data, retaining the in-sample values of all other reaction parameters.

In both cases, howsoever the adjustments are implemented, the policy implications of the underlying model are unaltered, although the forecasts may be greatly improved after deterministic shifts. The obvious conclusion, discussed further below, is that forecast performance is also not a good guide to policy-model choice. Without the correction, the forecasts would be poor; with the correction they are fine, but the policy recommendation is unaffected. Conversely, simple time-series predictors like $\Delta\Delta_4 c_{t-1}$ have no policy implications. We conclude that policy-analysis models should be selected on different criteria, which we now discuss.

7 Model selection for policy analysis

We consider three main issues: using forecast performance to select a policy model; investigating only closed models, where every variable is endogenous; and analyzing policy in open models, which condition on some policy variables.

7.1 Forecast performance and policy models

A statistical forecasting system is one having no economic-theory basis, in contrast to econometric models for which economic theory is the hallmark. Since the former system will rarely have implications for economic-policy analysis – and may not even entail links between target variables and policy instruments, being the 'best' available forecasting device is insufficient to ensure any value for policy analysis. Consequently, the main issue is the converse: does the existence of a dominating forecasting procedure invalidate the use of an econometric model for policy? Since forecast failure often results from factors unrelated to the policy change in question, an econometric model may continue to characterize the responses of the economy to a policy, despite its forecast inaccuracy. Further, when policy changes are implemented, forecasts from a statistical model may be improved by combining them with the predicted policy responses from an econometric model. The forecasting model may nevertheless remain distinct from that policy model, for reasons explained by Hendry and Mizon (2000).

The rationale for this analysis follows from the taxonomy of forecast errors in section 2 which recorded that deterministic shifts were the primary source of systematic forecast failure in econometric models. Devices like intercept corrections can robustify forecasting models against breaks which have occurred prior to forecasting (see e.g., Clements and Hendry, 1999a). As stressed above, while such 'tricks' may mitigate forecast failure, they do not alter the reliability of the policy implications of the resulting models.

Nevertheless, post-forecasting policy changes will induce breaks in models that do not embody the relevant policy links, whereas econometric systems that do so need not experience any policy-regime shift. Consequently, when both structural breaks and regime shifts occur, neither econo-

metric nor time-series models alone are adequate: this suggests that they should be combined, and Hendry and Mizon (2000) provide an empirical illustration of doing so.

7.2 Impulse-response analyses

The finding that shifts in the parameters of dynamic reactions are not readily detectable is potentially disastrous for 'impulse-response' analyses of economic policy based on closed systems, usually VARs. Since changes in VAR intercepts and dynamic coefficient matrices may not be detected – even when tested for – but the full-sample estimates are a weighted average across different regimes, the resulting impulse responses need not represent the policy outcomes that will in fact occur. Indeed, this problem may be exacerbated by specifying VARs in first differences, since deterministic factors play a small role in such models.

A Monte Carlo simulation illustrates the problem, using the unrestricted I(0) VAR:

$$\begin{aligned} y_{1,t} &= \lambda_1 + \phi_{11} y_{1,t-1} + \phi_{12} y_{2,t-1} + \epsilon_{1,t} \\ y_{2,t} &= \lambda_2 + \phi_{21} y_{1,t-1} + \phi_{22} y_{2,t-1} + \epsilon_{2,t} \end{aligned}$$

where $\epsilon_{i,t} \sim \mathsf{IN}[0, \sigma_{ii}]$, with $\mathsf{E}[\epsilon_{1,t}\epsilon_{2,s}] = 0 \; \forall t, s$. The $y_{i,t}$ are to be interpreted as I(0) transformations of I(1) variables. We consider breaks in the $\Phi = (\phi_{ij})$ matrix maintaining constant unconditional expectations of zero ($\mathsf{E}[y_{i,t}] = 0$) with $\boldsymbol{\lambda} = \mathbf{0}$. The full-sample size is $T = 120$, with a single break at $t = 0.5T$, and $\sigma_{ii} = 0.01$ (1% in a log-linear model). The unrestricted VAR with intercept and one lag is estimated, and then tested for breaks. The critical values for the constancy tests are those for a *known* break point, which delivers the highest possible power for the test used. We consider a large parameter shift, from:

$$\boldsymbol{\Phi} = \begin{pmatrix} 0.50 & -0.20 \\ -0.20 & -0.25 \end{pmatrix}, \tag{16}$$

to:

$$\boldsymbol{\Phi}^* = \begin{pmatrix} 0.50 & 0.20 \\ 0.20 & 0.25 \end{pmatrix}, \tag{17}$$

so the sign is altered on all but one repsonse. 1000 replications are used, and rejection frequencies at both 0.05 and 0.01 nominal test sizes are recorded (standard errors about 0.007 and 0.003 respectively). The rejection frequencies are reported graphically, for both p values. The graphs serve to illustrate the outcomes visually, showing that rejection frequencies are everywhere low in most cases, but confirming that the highest power is immediately before the break.

7.2.1 Test size

As fig. 2 reveals, the null rejection frequencies in the I(0) baseline data are reassuring: with 1000 replications, the approximate 95% confidence intervals are (0.036, 0.064) and (0.004, 0.016) for 5% and 1% nominal, and these are shown on the graphs as dotted and dashed lines respectively. The actual test sizes are close to their nominal levels.

7.2.2 I(0) dynamic shift

The detectability of a shift in dynamics is low when the DGP is an I(0) VAR. The first element of $\boldsymbol{\Phi}^*$ is left constant simply to highlight the changes in the other impulses. The break delivers the constancy-test graph in fig. 3. The highest power is less than 25%, even though the change constitutes a major structural break for the model economy. This may be an explanation for the lack of evidence supporting the Lucas (1976) critique: shifts in zero-mean reaction parameters are relatively undetectable, rather than absent.

7.2.3 Misleading impulse responses

Finally, we record the impulse responses from the averages of pre- and post- break models, and the model fitted across the regime shifts in fig. 4. The contrast is marked: despite the near undetectability of the break, the signs of most of the impulses have altered, and those obtained from the fitted model sometimes reflect one regime, and sometimes the other. Overall, mis-leading policy advice would follow.

7.3 Open-model policy analysis

Many of the problems in analyzing the responses of targets to changes in instruments noted above

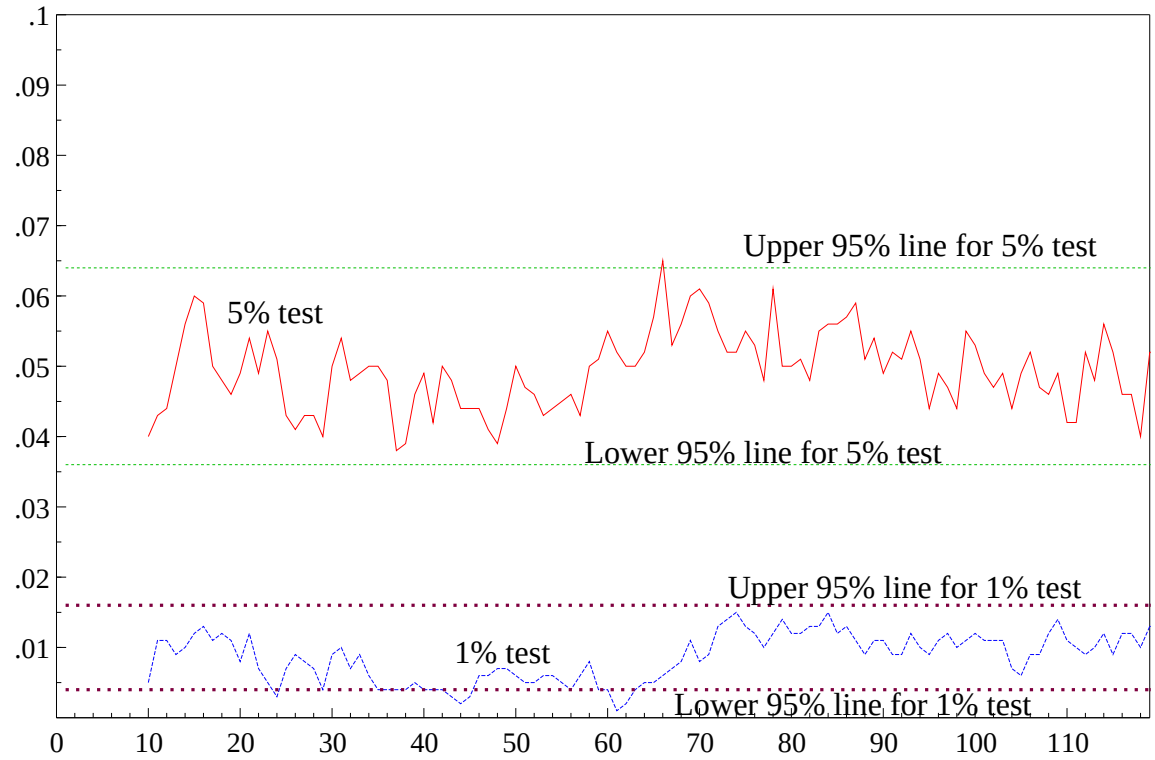

Figure 2 Constancy-test rejection frequencies for the I(0) null.

are absent when the modelling is conditional on the instruments, leading to an open model. Since it is often difficult to model the instruments, particularly in high-dimensional systems, conditioning on them is much easier and is certainly preferable to omitting them from the analysis. For economic policy analysis, another advantage of modelling n_y target variables $\mathbf{y}_t$ conditionally on n_z instrument variables z_t, is that the required policy responses $\partial \mathbf{y}_{t+h}/\partial \mathbf{z}_t'$ are directly estimable analytically or via simulation. Although the weak exogeneity of $\mathbf{z}_t$ is required for there to be no loss of information in making inferences on the parameters of interest, in practice it is likely that reliable estimates of policy responses will be obtained even when $\mathbf{z}_t$ is not weakly exogenous. Note that the fact that the $\{\mathbf{z}_t\}$ process is under the control of a policy agency does not ensure that $\mathbf{z}_t$ are exogenous variables. Whether the policy involves a regime shift, or a shock, the instruments must be super exogenous for the parameters of interest if modelling with an open econometric model is to result in no loss of information. However, if there is a policy regime shift, then co-breaking between the targets and instruments ensures that the policy is effective, and that the response of $\mathbf{y}_t$ can be reliably estimated (efficiently, when $\mathbf{z}_t$ is weakly exogenous for the response parameters).

7.3.1 I(0) Case

This section draws on some results in Ericsson, Hendry and Mizon (1998a), and is presented for completeness as a preliminary to considering the $\mathsf{I}(1)$ case in the next section. Modelling the conditional distribution for $\mathbf{y}_t$ given $\mathbf{z}_t$ and any relevant lags will yield efficient inference on the parameters of interest if $\mathbf{z}_t$ is weakly exogenous for those parameters. In addition, the conditional model will provide reliable estimates of the response in $\mathbf{y}_t$ to policy changes in $\mathbf{z}_t$ when its parameters are invariant to the policy change. When these conditions are satisfied it is relevant to use the conditional model to derive impulse responses and dynamic multipliers for assessing the effects of policy. However, Ericsson *et al.* (1998a) showed that, in general, the weak exogeneity status of conditioning variables is not invariant to transformations such as orthogonalizations or identified 'structures'.

Irrespective of the exogeneity status of $\mathbf{z}_t$,

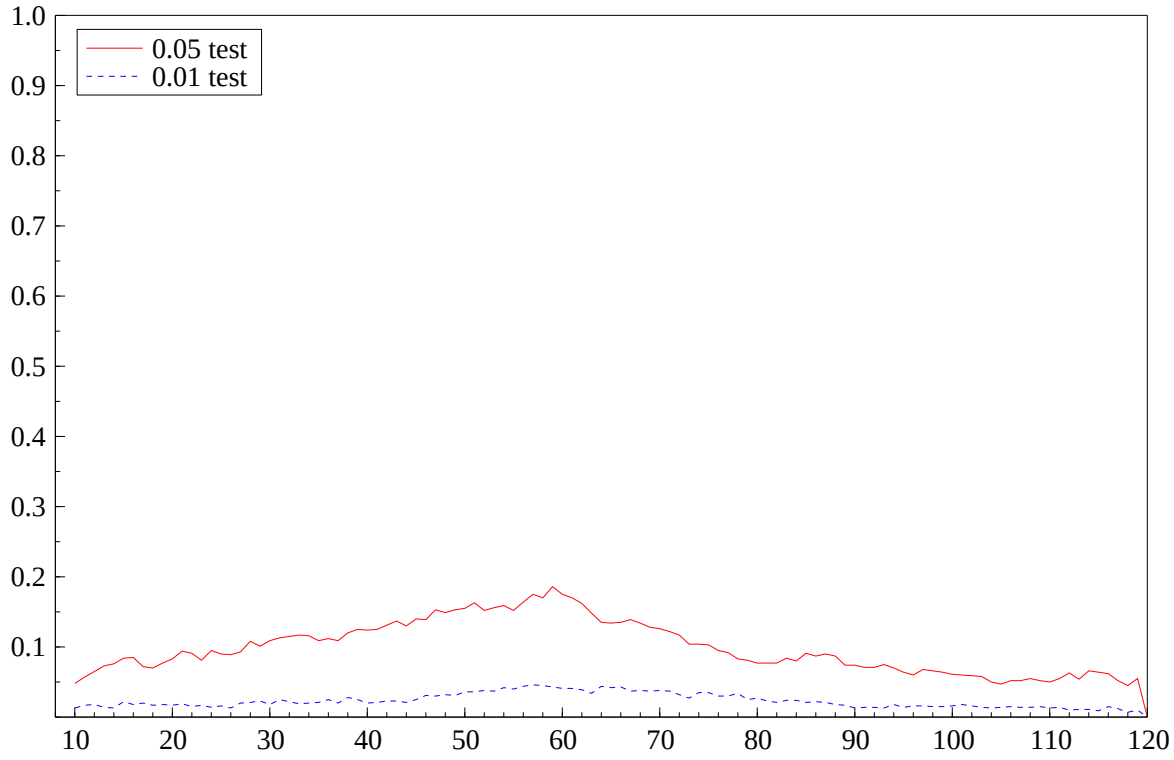

Figure 3 Constancy-test rejection frequencies for the I(0) structural break.

modelling the conditional distribution alone will result in very different impulse response matrices $\partial \mathbf{y}_{t+h}/\partial \boldsymbol{\varepsilon}_t'$ and dynamic multipliers $\partial \mathbf{y}_{t+h}/\partial \mathbf{z}_t'$, as a result of the latter taking into account the effect from contemporaneous and lagged $\mathbf{z}_t$. Thus, the response of $\mathbf{y}_t$ to an impulse in the innovation $\boldsymbol{\varepsilon}_t$ of the conditional model is not the relevant response for assessing the effects of policy changes in the $\{\mathbf{z}_t\}$ process.

7.3.2 I(1) Case

In the $\mathsf{I}(1)$ case, the class of open models will be equilibrium-correction systems conditional on the current growth-rate of the policy instruments, $\Delta \mathbf{z}_t$, assuming that the $\mathbf{z}_t$ are $\mathsf{I}(1)$, and are included in some of the r cointegration relations $\boldsymbol{\beta}' \mathbf{x}_{t-1}$. For there to be no loss of information, this analysis requires that $\mathbf{z}_t$ be weakly exogenous for both the long-run parameters $\boldsymbol{\beta}$ and any short-run dynamic response parameters (for both lagged $\mathbf{y}_{t-s}$ and lagged $\mathbf{z}_{t-k}$), all of which parameters should be invariant to the policy change. Under these conditions, it is possible to estimate the responses in the growth rates $\Delta \mathbf{y}_t$ and the disequilibria $\boldsymbol{\beta}' \mathbf{x}_t$ to particular choices of the instruments $\mathbf{z}_t$ even when the latter are $\mathsf{I}(1)$. To derive the impact on (say) $\Delta \mathbf{y}_{t+h}$ from a change in $\mathbf{z}_t$ requires a specification of the future path of $\mathbf{z}_{t+i}$ in response to $\mathbf{z}_t$ over $t = 1$ to $t + h$; implicitly, the model must be closed. This provides a link to the 'policy rules literature' (see e.g., Taylor, 1993, 2000), where alternative mappings of the policy instruments onto past values of disequilibria are evaluated. Nevertheless, the outcomes obtained can differ substantially from impulse-response analysis based on a (cointegrated) VAR when the policy rule does not coincide with the historical description of policy responses.

Partitioning the disequilibria $\boldsymbol{\beta}' \mathbf{x}_t = \boldsymbol{\beta}_y' \mathbf{y}_t + \boldsymbol{\beta}_z' \mathbf{z}_t$ reveals that $\boldsymbol{\beta}_y' \mathbf{y}_t$ are feasible target variables in this context despite $\mathbf{y}_t$ and $\boldsymbol{\beta}_y' \mathbf{y}_t$ being $\mathsf{I}(1)$. In particular, when $\boldsymbol{\beta}$ is invariant to changes in the instruments $\mathbf{z}_t$, a policy which keeps $\boldsymbol{\mu}$ constant but changes $\mathbf{z}_t$ must result in $\mathbf{y}_t$ changing: in this sense, such a policy is effective. However, an implication of this analysis is that very special conditions are required for policy that changes a single instrument $z_{i,t}$ (e.g., the minimum lending rate) to successfully target a single target variable $y_{j,t}$

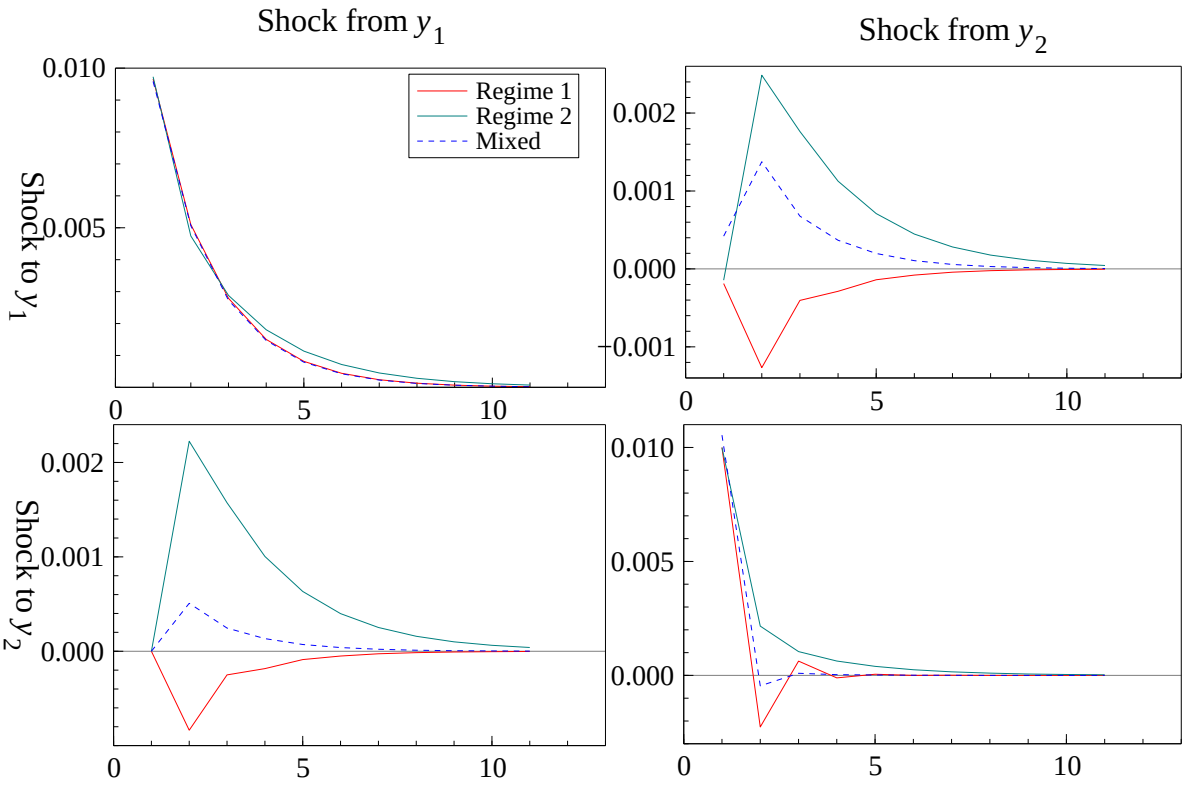

Figure 4 Impulse response comparisons in an I(0) VAR.

(e.g., inflation) when these variables are $\mathsf{I}(1)$. Conversely, Johansen and Juselius (2000) demonstrate that if a policy that targets an $\mathsf{I}(1)$ variable is successful, then the target will be rendered $\mathsf{I}(0)$.

Also, there are only $n_y + n_z - r$ unconstrained stochastic trends when r is equal to the number of cointegrating vectors, so the growth rates $\boldsymbol{\pi}_y$ of $\mathbf{y}_t$ and $\boldsymbol{\pi}_z$ of $\mathbf{z}_t$ are linked by $\boldsymbol{\beta}_y'\boldsymbol{\pi}_y + \boldsymbol{\beta}_z'\boldsymbol{\pi}_z = \mathbf{0}$. However, when there is co-breaking between $\Delta\mathbf{y}_t$ and $\Delta\mathbf{z}_t$, then a change in $\boldsymbol{\pi}_z$ will result in a corresponding change in the unconditional mean $\boldsymbol{\pi}_y$ of $\Delta\mathbf{y}_t$. Hence, just as in Hendry and Mizon (1998, 2000), linkages between deterministic terms are critical for policy to be effective when it is implemented via shifts in deterministic terms in the instrument process. Moreover, co-breaking here requires that $\Delta\mathbf{y}_t$ responds to contemporaneous and/or lagged changes in $\mathbf{z}_t$.

7.4 Congruent modelling

As a usable knowledge base, theory-related, congruent, encompassing econometric models remain undominated by matching the data in all measurable respects (see, e.g., Hendry, 1995). For empirical understanding, such models seem likely to remain an integral component of any progressive research strategy. Nevertheless, even the 'best available model' can be caught out when forecasting by an unanticipated outbreak of (say) a major war or other crisis for which no effect was included in the forecast. However, if empirical models which are congruent within sample remain subject to a non-negligible probability of failing out of sample, then a critic might doubt their worth. Our defence of the program of attempting to discover such models rests on the fact that empirical research is part of a progressive strategy, in which knowledge gradually accumulates. This includes knowledge about general causes of structural changes, such that later models incorporate measures accounting for previous events, and hence are more robust (e.g., to wars, changes in credit rationing, financial innovations, etc.). For example, the dummy variables for purchase-tax changes in Davidson *et al.* (1978) that at the time 'mopped up' forecast failure, later successfully predicted the effects of introducing VAT, as well as the consequences of its doubling in 1979; and the First World-War shift in money demand in Ericsson, Hendry and Prestwich (1998b)

matched that needed for the Second World War.

Since we now have an operational selection methodology with excellent properties, *Gets* seems a natural way to select models for empirical characterization, theory testing and policy analyses. When the GUM is a congruent representation, embedding the available theory knowledge of the target-instrument linkages, and parsimoniously encompassing previous empirical findings, the selection strategy described in section 4 offers scope for selecting policy models. Four features favour such a view. First, for a given null rejection frequency, variables that matter in the DGP are selected with the same probabilities as if the DGP were known. In the absence of omniscience, it is difficult to imagine doing much better systematically. Secondly, although estimates are biased on average, conditional on retaining a variable, its coefficient provides an unbiased estimate of the policy reaction parameter. This is essential for economic policy – if a variable is included, *PcGets* delivers the right response; otherwise, when it is excluded, one is simply unaware that such an effect exists.[2] Thirdly, the probability of retaining adventitiously significant variables is around the nominal size of the selection t-tests for the variables that remain after pre-selection simplification. If that is (say) even as large as 30 regressors, of which 5 actually matter, then at 1% significance, 0.25 variables will be retained on average: i.e., one additional 'spuriously-significant' variable per four equations. This seems unlikely to distort policy in important ways. Finally, the sub-sample – or more generally recursive – selection procedures help to reveal which variables have non-central t-statistics, and which central (and hence should be eliminated). Overall, the role of *Gets* in selecting policy models looks promising.

While changes to the coefficients of zero-mean variables may be difficult to detect in dynamic models, for policy models they remain hazardous: the estimated parameters would appear to be constant, but in fact be mixtures across regimes, leading to inappropriate advice. In a progressive research context (i.e., from the perspective of learning), this is unproblematic since most policy changes involve deterministic shifts (as opposed to mean-preserving spreads), hence earlier incorrect inferences will be detected rapidly – but is cold comfort to the policy maker, or the economic agents subjected to the wrong policies.

8 Conclusion

The implications for econometric modelling that result from the observance of forecast failure differ considerably from those obtained when the model is assumed to coincide with a constant mechanism. Causal information can no longer be shown to uniformly dominate non-causal. Intercept corrections have no theoretical justification in stationary worlds with correctly-specified empirical models, but in a world subject to structural breaks of unknown form, size, and timing, they serve to 'robustify' forecasts against deterministic shifts – as the practical efficacy of intercept corrections confirms. Forecasting success is no better an index for model selection than forecast failure is for model rejection. Thus, emphasising 'out-of-sample' forecast performance (perhaps because of fears over 'data-mining') is unsustainable (see, e.g., Newbold, 1993, p.658), as is the belief that a greater reliance on economic theory will help forecasting (see, e.g., Diebold, 1998), because that does not tackle the root problem.

The taxonomy of potential sources of forecast errors clarifies the roles of model misspecification, sampling variability, error accumulation, forecast origin mis-measurement, intercept shifts, and slope-parameter changes. Forecast failure seems primarily attributable to deterministic shifts. Such findings are potentially disastrous for 'impulse-response' analyses of economic policy. Since the changes in VAR intercepts and dynamic

[2]This is one of three reasons why we have not explored 'shrinkage' estimators, which have been proposed as a solution to the 'pre-test' problem, namely, they deliver biased estimators (see, e.g., Judge and Bock, 1978). The second, and main, reason is that such a strategy has no theoretical underpinnings in processes subject to intermittent parameter shifts. The final reason concerns the need for progressivity, explaining more by less, which such an approach hardly facilitates.

coefficient matrices may not be detected even when tested for, but the recorded estimates are a weighted average across the different regimes, the resulting impulse responses do not represent the policy outcomes that will in fact occur.

If the economy were reducible by transformations to a stationary stochastic process, where the resulting unconditional moments were constant over time, then well-tested, causally-relevant, congruent models which embodied valid theory restrictions would both fit best, and by encompassing, also dominate in forecasting on average. The prevalence historically of unanticipated deterministic shifts suggests that such transformations do not exist. Nevertheless, the case for continuing to use econometric systems probably depends on their competing reasonably successfully in the forecasting arena. Cointegration, cobreaking, and model-selection procedures as good as *PcGets*, with rigorous testing should help, but none of these ensures immunity to forecast failure from new breaks. An approach which incorporates causal information in a congruent econometric system for policy, but operates with robustified forecasts, merits consideration. We have not yet established that *Gets* should be used for selecting policy models from a theory-based GUM – and such a proof may not be possible, despite the relative accuracy with which the DGP is located. Nevertheless, achieving that aim represents the next step of our research program, and we anticipate that *Gets* will perform well in selecting models for policy.

References

Banerjee, A., and Hendry, D. F. (1992). Testing integration and cointegration: An overview. *Oxford Bulletin of Economics and Statistics*, **54**, 225–255.

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

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

Clements, M. P., and Hendry, D. F. (1999b). Modelling methodology and forecast failure. Unpublished typescript, Economics Department, University of Oxford.

Clements, M. P., and Hendry, D. F. (1999c). On winning forecasting competitions in economics. *Spanish Economic Review*, **1**, 123–160.

Davidson, J. E. H., Hendry, D. F., Srba, F., and Yeo, J. S. (1978). Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal*, **88**, 661–692. Reprinted in Hendry, D. F., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993, and Oxford University Press, 2000.

Diebold, F. X. (1998). The past, present and future of macroeconomic forecasting. *The Journal of Economic Perspectives*, **12**, 175–192.

Doornik, J. A. (1999). *Object-Oriented Matrix Programming using Ox* 3rd edn. London: Timberlake Consultants Press.

Ericsson, N. R., Hendry, D. F., and Mizon, G. E. (1998a). Exogeneity, cointegration and economic policy analysis. *Journal of Business and Economic Statistics*, **16**, 370–387.

Ericsson, N. R., Hendry, D. F., and Prestwich, K. M. (1998b). The demand for broad money in the United Kingdom, 1878–1993. *Scandinavian Journal of Economics*, **100**, 289–324.

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*. Dordrecht: Kluwer Academic Press.

Ericsson, N. R., and MacKinnon, J. G. (1999). Distributions of error correction tests for cointegration. International finance discussion paper no. 655, Federal Reserve Board of Governors, Washington, D.C. www.bog.frb.fed.us/pubs/ifdp/1999/655/

default.htm.

Feige, E. L., and Pearce, D. K. (1976). Economically rational expectations. *Journal of Political Economy*, **84**, 499–522.

Fildes, R. A., and Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review*, **63**, 289–308.

Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: Evidence. *Journal of Political Economy*, **86**, 971–987.

Hendry, D. F. (1994). HUS revisited. *Oxford Review of Economic Policy*, **10**, 86–106.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, D. F. (2000a). *Econometrics: Alchemy or Science?* Oxford: Oxford University Press. New Edition.

Hendry, D. F. (2000b). Forecast failure, expectations formation, and the Lucas critique. Mimeo, Nuffield College, Oxford.

Hendry, D. F., and Doornik, J. A. (1997). The implications for econometric modelling of forecast failure. *Scottish Journal of Political Economy*, **44**, 437–461. Special Issue.

Hendry, D. F., and Krolzig, H.-M. (1999). Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez. *Econometrics Journal*, **2**, 41–58.

Hendry, D. F., and Mizon, G. E. (1998). Exogeneity, causality, and co-breaking in economic policy analysis of a small econometric model of money in the UK. *Empirical Economics*, **23**, 267–294.

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 Richard, J.-F. (1982). On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics*, **20**, 3–33. Reprinted in Granger, C. W. J. (ed.) (1990), *Modelling Economic Series*. Oxford: Clarendon Press and in Hendry D. F. (1993,2000) *op. cit.*

Hoover, K. D., and Perez, S. J. (1999). Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, **2**, 1–25.

Johansen, S., and Juselius, K. (2000). How to control a target variable in the VAR model. Mimeo, European University of Institute, Florence.

Judge, G. G., and Bock, M. E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North Holland Publishing Company.

Krolzig, H.-M., and Hendry, D. F. (2000). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*. forthcoming.

Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics*, **65**, 1–12.

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.

Makridakis, S., and Hibon, M. (2000). The M3-competition: Results, conclusions and implications. Discussion paper, INSEAD, Paris.

Mayo, D. (1981). Testing statistical testing. In Pitt, J. C. (ed.), *Philosophy in Economics*, pp. 175–230: D. Reidel Publishing Co. Reprinted as pp. 45–73 in Caldwell B. J. (1993), *The Philosophy and Methodology of Economics*, Vol. 2, Aldershot: Edward Elgar.

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

Muellbauer, J. N. J. (1994). The assessment: Consumer expenditure. *Oxford Review of Economic Policy*, **10**, 1–41.

Newbold, P. (1993). Comment on 'On the limitations of comparing mean squared forecast errors', by M.P. Clements and D.F. Hendry. *Journal of Forecasting*, **12**, 658–660.

Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica*, **58**, 113–144.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie–Rochester Conference Series on Public Policy*, **39**, 195–214.

Taylor, J. B. (2000). The monetary transmission mechanism and the evaluation of monetary policy rules. Forthcoming, Oxford Review of Economic Policy.

Turner, D. S. (1990). The role of judgement in macroeconomic forecasting. *Journal of Forecasting*, **9**, 315–345.

White, H. (1990). A consistent model selection. In Granger, C. W. J. (ed.), *Modelling Economic Series*, pp. 369–383. Oxford: Clarendon Press.