Econometrics Journal (2002), volume **5**, pp. 319–344.

Modelling methodology and forecast failure

MICHAEL P. CLEMENTS[1] AND DAVID F. HENDRY[2]

[1]*Department of Economics, University of Warwick, Coventry, CV4 7AL, UK*
E-mail: m.p.clements@warwick.ac.uk

[2]*Nuffield College, Oxford, OX1 1NF, UK*
E-mail: david.hendry@nuffield.ox.ac.uk

Received: March 1999

Summary We analyse by simulation the impact of model-selection strategies (sometimes called pre-testing) on forecast performance in both constant- and non-constant-parameter processes. Restricted, unrestricted and selected models are compared when either of the first two might generate the data. We find little evidence that strategies such as general-to-specific induce significant over-fitting, or thereby cause forecast-failure rejection rates to greatly exceed nominal sizes. Parameter non-constancies put a premium on correct specification, but in general, model-selection effects appear to be relatively small, and progressive research is able to detect the mis-specifications.

Keywords: *Model selection, General-to-specific, Forecast failure.*

1. INTRODUCTION

Forecast failure is defined as significant mis-forecasting relative to the previous record (in-sample, or earlier forecasts), whereas poor forecasting is judged relative to some standard, either absolute (perhaps because of a policy requirement for accuracy), or relative to a rival model. Forecasts may be poor simply because the phenomenon is largely unpredictable—little of the variation in the variable to be forecast is explained either in-sample or over the forecast period—but forecast failure is characterized by poor forecasts relative to anticipated performance.

In this paper, we explore whether the strategy adopted during modelling may induce forecast failure: specifically, can the sequential testing inherent in 'general-to-specific' modelling strategies lead to forecast failure? The logic of such a claim might be that the modelling strategy induces 'over-fitting' by retaining variables that are spuriously significant, hence under-estimating the residual standard deviation. Forecast errors will on average then exceed in-sample errors, and thus lead to model rejection. Nevertheless, many simplified models have residual standard deviations not much smaller than their unrestricted parents, suggesting that other factors may account for empirical occurrences of forecast failure, such as structural breaks: see Clements and Hendry (1999, Chapters 3 and 4). For example, when breaks occur in various time series, inclusion of irrelevant variables in a model can induce failure, which is removed by dropping those variables from the regressor set. Consequently, simplification procedures would be highly beneficial if they achieved that. Conversely, if they spuriously retained variables that later changed, the resulting forecasts could be very poor.

There are two main alternatives—unless one believes in omniscient investigators who 'know' the correct model at the outset. The first is to use the initial unrestricted model: this delivers an unbiased estimate of the innovation variance, so should not be prone to failure in constant-parameter worlds, but may perform less well under structural breaks. Indeed, highly over-parametrized models do not seem to have a great forecasting reputation, although this may be a problem of poor, as against failed, forecasts. The second is to use *a priori* restricted models, which will do well when the restrictions are valid, but could perform badly otherwise. Thus, empirical selection seems an inevitable component of econometric modelling.

The effects of model-selection strategies are difficult to obtain analytically, especially when there are several testing steps and multiple selection criteria. Consequently, while there seems to be a perception that selection strategies induce 'overfitting', and therefore misleading inference, there is little substantive evidence for this view. The recent explosion in computer power has made simulation-based studies of selection strategies feasible, and that is the approach we take in this paper. A precusor using simulation is Hoover and Perez (1999), who find that a suitably amended 'general-to-specific' (denoted Gets) selection strategy is reasonably successful at identifying the correct model. Our focus is on the implications of selection for forecasting. A number of discussants of Hoover and Perez (1999) (including Granger and Timmermann (1999), Hand (1999), and Hansen (1999)) suggested that because time-series econometric models are commonly used for forecasting, it might be informative to judge success in terms of forecast performance.[†] The difficulty of obtaining analytical results is compounded by a tendency to confound related, but conceptually distinct, notions, such as overfitting, data-mining, lack of parsimony, etc., inhibiting a clear picture of the implications of selection for forecast failure from emerging. We explain the relationships between these notions, and we simulate selection strategies in a number of scenarios, chosen to isolate the factors that might contribute to forecast failure. Where analytical results can be obtained which serve as a check on the simulations, these are presented. Thus, we report power calculations for the test that a variable is insignificant in-sample, but rely on simulation to calculate the implications of these decisions for forecast failure. Finally, Gets is typically viewed in the context of a 'progressive research strategy' (henceforth PRS), so the simulations are designed to explore what an investigator would learn as more information became available: in our context, this amounts to allowing the forecast origin to move through time.

A related paper to ours is Clark (2000), who considers whether 'out-of-sample forecast comparisons can help prevent data mining-induced overfitting'. In a simulation study, Clark considers two situations. In the first, the out-of-sample forecast performance of the true model is compared against that of models chosen by selection—on those occasions when the selected model retains extraneous variables—to calculate the proportion of times that the selected variables are found to have predictive power (size). The tests chosen appear to be reasonably sized. Secondly, to calculate power, an out-of-sample forecast test is performed between the selected model—on those occasions that it is the true model—and a restricted benchmark. The tests have reasonable power, despite the 'data mining'. The conditioning on either an extraneous variable being included, or on the selected model being the DGP, precludes drawing general conclusions on the effects of selection on relative forecast accuracy. Nevertheless, Clark finds that 'data mining' (or selection) is relatively benign, in the sense that tests of out-of-sample performance still have good size and power.

[†] Success at forecasting and identifying the correct model are obviously closely related, but a model with the correct variables may have estimated coefficients far from the population values, for instance. Moreover, the correct in-sample model may not yield the best out-of-sample forecasts because of parameter non-constancies.

The plan of the paper is as follows. Section 2 briefly discusses testing for forecast failure. Section 3 disentangles the various ways in which model-selection effects might either attenuate or exacerbate forecast failure. To illustrate the likely importance of these effects, Section 4 reports a set of simulation experiments on a simple static model, designed to focus on the false inclusion and exclusion of variables when there are breaks in-sample and during the forecast period. Section 5 attempts to mimic a relatively high-dimensional general-to-specific empirical modelling exercise, tracking both the outcomes of forecast-failure tests, and the impacts on MSFEs over the progressive rounds of simplification for a dynamic process. Finally, Section 6 concludes with an assessment of the evidence on the likely impact of model-selection strategies on forecast failure.

2. TESTING FOR FORECAST FAILURE

Forecast failure contrasts with the notion of forecast (in)accuracy, where forecasts may be 'poor' measured in some metric, such as (root) mean squared forecast error, but are nonetheless in keeping with past performance. We use the notion of forecast failure to assess the effects of model selection for a number of reasons: it allows a model's performance to be judged without reference to external benchmarks, such as that of other models' forecast accuracy; it is a simple and widely used notion; and its occurrence is felt to be sufficiently important in the profession to call into question whole classes of models (such as Keynesian income-expenditure models after the first oil crisis in the early 1970s). We discuss the methodological implications of forecast failure elsewhere: see Clements and Hendry (1998).

We use the Chow (1960) test statistic as the test for forecast failure. This can be written as:

$$Q = \frac{\widehat{\nu}^2_{T+1|T}}{\widehat{\sigma}^2_v f_{T+1}} \tag{1}$$

for 1-step forecasts, where $\widehat{\nu}_{T+1|T}$ is the forecast error in the OLS prediction of the regressand at period $T+1$, say, y_{T+1} , based on parameters estimated from the sample $1, \ldots, T$, $\widehat{\sigma}^2_v$ is an unbiased estimator of the model's error variance, and $f_{T+1} = (1 + \mathbf{x}'_{T+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{T+1})$, where $\mathbf{x}_{T+1}$ is a p-dimensional vector of explanatory variables at period $T+1$, and $\mathbf{X}$ is a $T \times p$ matrix of observations on the p explanatory variables for periods $1, \ldots, T$. In a static regression with strongly exogenous stochastic regressors, the test is exact, and under the null of constant parameters $Q \sim \mathsf{F}^1_{T-p}$. The statistic is no longer exact when the explanatory variables contain lags of the dependent variable, but Kiviet (1986) shows by simulation that it has good size properties, and compares favourably with other asymptotically equivalent statistics. This will be relevant in Section 5 where we consider vector autoregressive models.

3. MODEL-SELECTION EFFECTS

It is sometimes suggested that a combination of overfitting, lack of parsimony, data-mining, pretesting and related notions plague general-to-specific (Gets) modelling strategies (see e.g. Hess *et al.* (1998)). In this Section, we try to disentangle some of these closely related strands of argument, and discuss both the ways in which they might affect a progressive research strategy, and whether they might contribute to forecast failure. In the introduction, we argued that in practice an investigator is faced with three main alternatives: using a general unrestricted model, *a priori* restrictions, or a sequential-testing procedure. Here we consider the possible pitfalls with these approaches. We exclude the infeasible optimum that the truth is known at the outset, but compare how well these other strategies do using the DGP as a baseline.

3.1. Overfitting and lack of parsimony

Todd (1990, p. 217) refers to overfitting as: fitting 'not only the most salient features of the historical data, which are often the stable, enduring relationships' but also 'features which often reflect merely accidental or random relationships that will not recur'. This is referred to as 'sample dependence' in Hendry (1995b). A lack of parsimony entails having models with too many parameters. The two notions are distinct, in that a few carefully chosen variables could induce overfitting (such as suitably transformed data functions which act as dummies for 'outliers' that are simply large shocks), whereas using many badly chosen variables may leave a poor fit (assuming fit is measured in units that are adjusted for degrees of freedom). Of course, overfitting may also arise if an investigator uses a 'generously parametrized model', but it is unclear how deleterious the consequences for either modelling or forecasting will be in that case, since the resulting estimates are not specifically biased by the irrelevant variables in a constant-parameter process.

From a modelling perspective, overfitting is a transient problem in a PRS, in that the accidental nature of the erroneously included variables will become apparent as more information accrues. From a forecasting perspective, although the resulting forecasts may be inaccurate, systematic forecast failure will only occur if the data properties of the incorrectly included variables change during the forecast period. Clements and Hendry (1999, Chapter 4) provide the following analytic example: as will be seen, the outcome is a consequence of the data properties—not of how the forecasting model was selected—so applies to all approaches. Later, we consider whether modelling strategies are likely to lead to a selection that is prone to this problem.

We suppose that the stationary DGP comprises two blocks:

$$\begin{pmatrix} \mathbf{y}_{1,t} \\ \mathbf{y}_{2,t} \end{pmatrix} = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} + \begin{pmatrix} \mathbf{\Pi}_{11} & 0 \\ 0 & \mathbf{\Pi}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{y}_{1,t-1} \\ \mathbf{y}_{2,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} \tag{2}$$

but the forecaster uses the unrestrictedly estimated forecasting model:

$$\widetilde{\mathbf{y}}_{T+1} = \widetilde{\phi} + \widetilde{\mathbf{\Pi}}\mathbf{y}_T. \tag{3}$$

Although the resulting forecasts will be inefficient, in a constant-parameter world, they will be (essentially) unbiased. Let:

$$\mathsf{E}[\mathbf{y}_t] = \varphi = \phi + \mathbf{\Pi}\varphi = (\mathbf{I} - \mathbf{\Pi})^{-1}\phi,$$

denote the equilibrium (or long-run) mean, then:

$$\mathsf{E}[\mathbf{y}_{T+1} - \widetilde{\mathbf{y}}_{T+1}] = \varphi - \mathsf{E}[\widetilde{\phi} + \widetilde{\mathbf{\Pi}} y_T] \simeq \varphi - \phi - \mathbf{\Pi}\varphi = \mathbf{0}.$$

When the intercept ϕ_2 of the second block in (2) changes at T_1 to ϕ_2^*, the correctly specified model for $\mathbf{y}_1$, using restricted estimates:

$$\widehat{\mathbf{y}}_{1,T+1} = \widehat{\phi}_1 + \widehat{\mathbf{\Pi}}_{11}\mathbf{y}_{1,T},$$

will forecast as anticipated. However, for T somewhat larger than T_1, when using (3):

$$\begin{aligned}\widetilde{\mathbf{y}}_{1,T+1} &= \widetilde{\phi}_1 + \widetilde{\mathbf{\Pi}}_{11}\mathbf{y}_{1,T} + \widetilde{\mathbf{\Pi}}_{12}\mathbf{y}_{2,T} \\ &= \widetilde{\varphi}_1 + \widetilde{\mathbf{\Pi}}_{11}(\mathbf{y}_{1,T} - \varphi_1) + \widetilde{\mathbf{\Pi}}_{12}(\mathbf{y}_{2,T} - \varphi_2),\end{aligned}$$

where:

$$\widetilde{\varphi}_1 = \widetilde{\phi}_1 + \widetilde{\mathbf{\Pi}}_{11}\varphi_1 + \widetilde{\mathbf{\Pi}}_{12}\varphi_2,$$

so the resulting forecasts could be poor as $\mathsf{E}[\mathbf{y}_{2,T}] \neq \varphi_2$. This form of forecast failure reflects overfitting combined with a deterministic shift in the falsely included irrelevant variables, and conforms to the adage that 'forecast failure out, needs forecast failure in'. The joint model of (2) would also fail here, even when using the correctly restricted model for the first block, revealing that some shift in the system had occurred.

The extent to which the 'non-parsimonious' unrestricted model (3) will be prone to this effect depends on the magnitude of $\widetilde{\mathbf{\Pi}}_{12}$ as well as the shift. Consequently, in-sample selection strategies probably offer little additional protection, since they will eliminate (retain) $\mathbf{y}_{2,T}$ when it has a small (large) effect in (3). Nevertheless, a PRS should gradually move towards the correct specification as the sample grows, since after the break, $\mathbf{y}_{2,T}$ will usually be eliminated (see e.g. White (1990)). These ideas are explored in Section 4.3.

3.2. Pre-testing

Using in-sample data evidence to reduce a general model to a more parsimonious representation via successive rounds of simplifications (as in Gets) raises the possibilities of both pre-test bias and overfitting (or exacerbating any overfitting already apparent in the general model). On the last of these two, despite commencing with an 'over-parametrized representation', Gets modelling need not lead to overfitting: simplification could either attenuate or exacerbate sample dependence. The former could occur if genuinely irrelevant factors were eliminated, whereas the latter could happen if the influences of accidental aspects were captured more 'significantly' by being retained in a smaller parametrization (again, a transient problem). The Monte Carlo results in Hoover and Perez (1999) suggest that Gets, extended as they suggest to incorporate both diagnostic tests and inter-model encompassing checks, often delivers a favourable outcome, in that finally selected equations are close to the ones which generated their data.

If Gets induced overfitting by retaining variables that were spuriously significant, it could have similar effects in forecasting to pre-testing. Pre-testing can result in the estimated equation standard error understating the true standard error, so that forecast failure could result from forecast errors being on average larger than in-sample. However, the opposite is also possible, as relevant variables may be omitted on a pre-test and thereby inflate the estimated equation standard error. When the process is non-constant, such omissions induce forecast failure.

3.3. Data mining

The final issue relates to the impact on forecasting of the approach known as 'data mining'—a prejudiced search for supportive evidence (see e.g. Leamer (1978))—which may lead to either over or under fitting depending on the objective of the study. Gilbert (1986) distinguishes weak data mining—corroborating a prior belief—from strong data mining—ignoring or even hiding conflicting evidence. The latter will prove detrimental to forecasting when the selected model incorporates variables that should not be present, or omits those that should, and either class changes later.

Overall, in the realistic setting when the correct specification is not known *a priori*, selection procedures may improve or worsen the forecast performance of a model relative to the unrestricted alternative. The next section tries to evaluate the likely magnitudes of such effects in a simple setting.

4. MODEL SELECTION AND FORECAST FAILURE

The next two subsections examine possible pre-test induced forecast failure in a process without structural breaks, first by potentially including an irrelevant variable (and thereby overfitting), then by potentially excluding a relevant variable (due to in-sample model-selection tests). Since the data properties are unchanged between the sample and the forecast periods, and the tests employed are exact, the only source for over-rejection on a forecast-failure test is a model-selection effect, which is seen to be small here even in very small samples. Departures of test sizes from their nominal levels, dynamics, and non-stationarity all might affect the generality of this finding, but as both the unrestricted and selected models will be distorted, the outcome could reinforce or reduce the conclusion.

Then we examine the empirically more relevant case where there are deterministic shifts during the estimation or forecast periods. Now, falsely excluding a variable which changes in the forecast period will induce forecast failure. But as time passes and the shift comes within the forecaster's information set, the hypothesis that z_t should be excluded will be rejected more often, and the pre-test model forecast failure rejection frequency (RF) will decline, as will that of the restricted model. For falsely including a variable, a similar pattern emerges. As we move through time, a forecast-period shift becomes an estimation-period shift, so the incorrect model is selected less often, and the forecast-failure rejection rate of the pre-test model is closer to the nominal on the extended sample. However, the forecast-failure rejection rate of the pre-test model can be higher than that of the unrestricted model, even though the latter always includes the 'irrelevant' variable that shifts.

Contrasting an estimation-period shift to the forecast-period shift brings out the impact of breaks on model discovery and forecast performance in a PRS. As time passes, and the hitherto forecast-period phenomenon comes within an 'extended' within-sample period, so the relative abilities of the modellers to detect model-specification and find forecast failure alter.

4.1. False inclusion without breaks

Our first set of simulations is based on the static model:

$$y_t = \mu + \gamma z_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{IN}[0, \sigma_\varepsilon^2] \tag{4}$$

where $\sigma_\varepsilon^2 = 1$ so that $\{\varepsilon_t\}$ is a sequence of *I*ndependent *N*ormal (IN) random variables, and z_t is strongly exogenous for (μ, γ), generated by $z_t \sim \mathsf{IN}[0, \sigma_z^2]$ with z_t and ε_t independent. The test for forecast failure is invariant to σ_z, which is set to unity.

The possibility of forecast failure may arise because of pre-testing, whereby 'spuriously significant' regressors that enhance in-sample fit are retained, but the out-of-sample performance is not similarly improved, resulting in a relative deterioration in forecast accuracy, and hence forecast failure. Thus, we set $\gamma = 0$ in (4) so that z_t is an extraneous variable.[†] The forecast performance of three models is assessed. The 'general' or unrestricted (U) model is (4), the first restricted (R) model is the DGP ($y_t = \mu + \varepsilon_t$), and a pre-test (P) strategy chooses one or the other based on a pre-test of the significance of z_t in the general model. Let:

$$\begin{aligned} \widehat{y}_{T+1} &= \widehat{\mu} + \widehat{\gamma} z_{T+1} & \text{if} \quad |\mathsf{t}_{\gamma=0}| &> c_\alpha \\ \widetilde{y}_{T+1} &= \widetilde{\mu} & \text{if} \quad |\mathsf{t}_{\gamma=0}| &\leq c_\alpha \end{aligned}$$

and:

$$\mathsf{p}_\alpha = \mathsf{P}(|\mathsf{t}_{\gamma=0}| > c_\alpha) = \alpha,$$

so the modeller forecasts $(1 - \mathsf{p}_\alpha)$ of the time by $\widetilde{y}_{T+1}$ and by $\widehat{y}_{T+1}$ on the remainder, represented by the indicator function $\mathbf{1}_{|\mathsf{t}_{\gamma=0}|>c_\alpha}$ on any given replication, with:

$$\mathsf{E}[\mathbf{1}_{|\mathsf{t}_{\gamma=0}|>c_\alpha}] = \mathsf{p}_\alpha.$$

Thus, the actual forecast is:

$$\begin{aligned} \overline{y}_{T+1} &= \mathbf{1}_{|\mathsf{t}_{\gamma=0}|>c_\alpha} \widehat{y}_{T+1} + (1 - \mathbf{1}_{|\mathsf{t}_{\gamma=0}|>c_\alpha}) \widetilde{y}_{T+1} \\ &= \widetilde{y}_{T+1} + \mathbf{1}_{|\mathsf{t}_{\gamma=0}|>c_\alpha} (\widehat{y}_{T+1} - \widetilde{y}_{T+1}) \\ &= \widetilde{\mu} + \mathbf{1}_{|\mathsf{t}_{\gamma=0}|>c_\alpha} (\widehat{\mu} - \widetilde{\mu}) + \mathbf{1}_{|\mathsf{t}_{\gamma=0}|>c_\alpha} \widehat{\gamma} z_{T+1}. \end{aligned}$$

Notice that

$$|\mathsf{t}_{\gamma=0}| > c_\alpha \qquad \text{implies} \quad \widehat{\gamma}^2 > c_\alpha^2 \mathsf{V}\,[\widehat{\gamma}] = \frac{c_\alpha^2 \widehat{\sigma}_\varepsilon^2}{T \widehat{\sigma}_z^2},$$

so one keeps z_{T+1} only when $\widehat{\gamma}$ is 'large' relative to its standard error. We assume $\widehat{\mu} \simeq \widetilde{\mu}$ for simplicity. Then, the forecast error when z_{T+1} is retained, is:

$$\overline{\varepsilon}_{T+1} = y_{T+1} - \overline{y}_{T+1} = (\mu - \widetilde{\mu}) + \varepsilon_{T+1} - \widehat{\gamma} z_{T+1},$$

as $\gamma = 0$, so:

$$\mathsf{V}[\overline{\varepsilon}_{T+1}] = \mathsf{V}[\widetilde{\mu}] + \mathsf{V}[\varepsilon_{T+1}] + \sigma_z^2 \mathsf{E}[\widehat{\gamma}^2 \mid |\mathsf{t}_{\gamma=0}| > c_\alpha].$$

Since $\widehat{\gamma}$ is normally distributed:

$$\mathsf{E}[\widehat{\gamma} \mid \widehat{\gamma} > \frac{c_\alpha \sigma_\varepsilon}{\sqrt{T} \sigma_z}] = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{c_\alpha^2}{2T}\right)}{1 - \Phi\left(\frac{c_\alpha}{\sqrt{T}}\right)}, \tag{5}$$

where $\Phi(\cdot)$ is the cumulative normal, and the last expression sets $\sigma_\varepsilon = \sigma_z = 1$.

For $T = 3$ and $c_{0.05} = 3.2$, evaluating (5) yields:

$$\mathsf{E}[\widehat{\gamma} \mid \widehat{\gamma} > 1.85] \simeq 2.24,$$

inducing:

$$\mathsf{V}[\overline{\varepsilon}_{T+1}] = \left(1 + \frac{1}{3} + 2.25^2\right) \simeq 6.4,$$

[†] This example was suggested to us by Paul Ruud in a private communication.

Table 1. Forecast test rejection frequencies.

	5% forecast-failure test				1% forecast-failure test		
T	R	U	P(5,5)	P(1,5)	R	U	P(1,1)
3	5.00	5.10	8.12	5.89	1.06	1.02	1.71
6	4.98	4.98	6.26	5.46	0.98	1.03	1.24
9	5.02	4.98	5.76	5.33	1.02	1.03	1.18
12	4.93	4.96	5.45	5.12	1.02	1.00	1.09
18	5.00	4.96	5.30	5.10	1.01	1.00	1.06
24	5.00	5.04	5.24	5.08	1.03	1.05	1.06
30	4.88	4.91	5.06	4.95	0.95	0.94	0.98

The columns headed R, U and P are the Restricted, Unrestricted and Pre-test cases.

on 5% of trials. Consequently, there is a large inflation in the forecast-error variance relative to 1.333. Since $c_{0.01} = 5.8$ if $T = 3$, then $\mathsf{E}[\widehat{\gamma}|\widehat{\gamma} > 3.35]$ is even larger, but happens only 1% of the time, so the test rejections must be close to those for the restricted model. Finally, when $T = 21$, then $c_{0.05} = 2.1$ and evaluating (5) gives 1.11, so that:

$$\mathsf{V}[\bar{\varepsilon}_{T+1}] = \left(1 + \frac{1}{21} + 1.11^2\right) \simeq 2.3,$$

inducing a smaller increase: thus, there should be little excess rejection.

Each model/strategy is subjected to the forecast failure test, and rejection frequencies are calculated from $M = 100\,000$ Monte Carlo replications, for $T = 3, 6, 9, \ldots, 30$, using common random numbers over T.[†] The results for a selection of sample sizes are reported in Table 1, where P(α,β) denotes an in-sample pre-test at α% and a forecast-failure test conducted at β%, using 5% and 1% nominal sizes.

The rejection frequencies for the general and the restricted models are close to the nominal size for all T and both critical values. The test for the restricted model is exact, so that the differences between the rejection frequencies and the test size reflect Monte Carlo error. The standard error is $\sqrt{p(1-p)/M}$, where p is the size (as a percentage), so standard errors are approximately 0.07% and 0.03% for the 5% and 1% levels. Thus, a 95% confidence interval for the 1% level test is approximately (0.94, 1.06).

The rejection frequency is inflated to around 8% for the pre-test(5, 5) strategy for a sample size of $T = 3$, but this is all but removed if the more stringent 1% level is used for the pre-test. In both cases, the excess rejections are close to those expected from the above analysis. Indeed, the pre-test(1, 5) strategy delivers close to 5% forecast-failure rejections at all the sample sizes, and pre-test(1, 1) is near 1% for $T \geq 9$. Consequently, there is no evidence of substantive failure induced by pre-testing in this simple set-up.

Notice that the coefficient estimates are not biased here by pre-testing, since although only significant estimates are retained, they are symmetrically distributed around zero, as is the corresponding t-test. Consequently, forecasts are unbiased, but have excess variance relative to the restricted model when z_t is irrelevant.

[†] All the simulations reported in this paper were coded in the Gauss Programming Language, Aptech Systems Inc, and used the RNDNS random number generator.

4.2. False exclusion without breaks

We also investigated pre-test biases when the DGP includes z_t: specifically, we kept disturbance term variances at unity, and considered combinations of $\gamma = \{0.15, 0.3, \ldots, 1.5\}$ and $T = \{3, 6, \ldots, 30\}$. The restricted and unrestricted model forecast-failure RFs were close to the nominal sizes, and hence are not reported, while the outcomes for selection are similar to those shown in Table 1 (the same experiment with $\gamma = 0$). For the nominal 1% level tests, the RFs never exceed 2% and fall in T. For the 5% level tests, the RFs fall in T from 9% at $T = 3$ to 5% at $T = 30$, and in both cases vary little with γ. Thus the effect of selection on forecast failure is largely invariant to whether the candidate variable belongs in the DGP for this simple example. Equally, the restricted and unrestricted models have similar RFs in these constant processes.

We also experimented with allowing $\{z_t\}$ to be autoregressive, namely, a zero-mean first-order process with autocorrelation of 0.9, keeping $\sigma_z^2 = 1$. Then, the Chow test of forecast failure in the restricted model will no longer be exact, because the model's errors are autocorrelated, and also of interest is the impact of this on the pre-test strategy. The results suggest relatively little effect on the restricted model rejection frequencies, as might be expected given the evidence in Kiviet (1986) that the test is relatively insensitive to dynamic mis-specification—the most notable discrepancy is a rejection rate of around 4% (at a nominal size of 5%) for large T and γ. There is also little effect on the pre-test Chow test rejections: these are just over 8% for all values of γ at $T = 3$, but are less than 6% for $T > 20$.

4.3. Structural breaks

To investigate the interaction between selection and structural breaks on tests of forecast failure, we undertook experiments with both post-sample and in-sample breaks.

False exclusion and post-sample breaks. First, for forecast-period shifts, we set $T = 20$, and allowed γ to take values from 0.15 to 1.5, where z_t is given by:

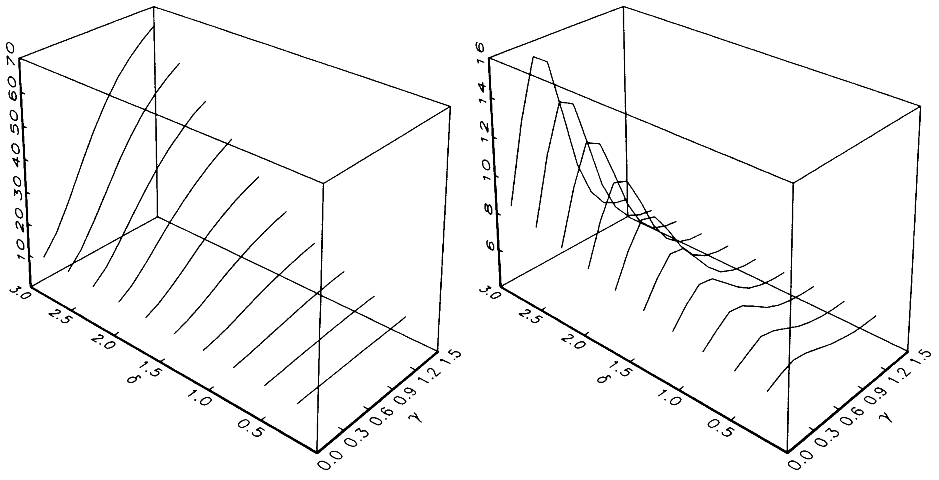
$$z_t = \delta \times 1_{t \geq \tau} + \zeta_t, \qquad \zeta_t \sim \mathsf{IN}[0, 1],$$

where $1_{t \geq \tau}$ is unity when the subscript is true, with δ ranging from 0.3 to 3. We set $\tau = 21$. Thus, y_t now depends on z_t, but the mean of z_t shifts by δ at $t = 21$. The estimation period is $t = 1, \ldots, 20$, followed by a 1-step ahead forecast test for period $t = 21$. When z_t is excluded, the forecast error $\widehat{\varepsilon}_{T+1} = y_{T+1} - \widehat{y}_{T+1}$:

$$\widehat{\varepsilon}_{T+1} = \varepsilon_{T+1} + (\mu - \widehat{\mu}) + \gamma\delta + \gamma\zeta_{T+1},$$

is increasing in $\gamma\delta$, whereas both the forecast period 'noise' and the likelihood of retaining z_t are increasing in γ. Thus, for $\gamma \neq 0$, increases in δ will affect forecast failure much more.

Figure 1 shows that the RFs for the restricted model are increasing in γ and δ, as the forecast-period data properties diverge increasingly from the in-sample properties in both these parameters. The RFs for the selected model (see Figure 1) are smaller than those from the restricted model. For large γ, pre-testing will seldom lead to z_t being omitted, so that the size of δ is immaterial. For small γ, such that z_t is often omitted, the RFs increase sharply in δ. In this setting, the unrestricted model does well, as it is correctly specified.



(a) Restricted model (5% level)

(b) Pre-test (5,5)

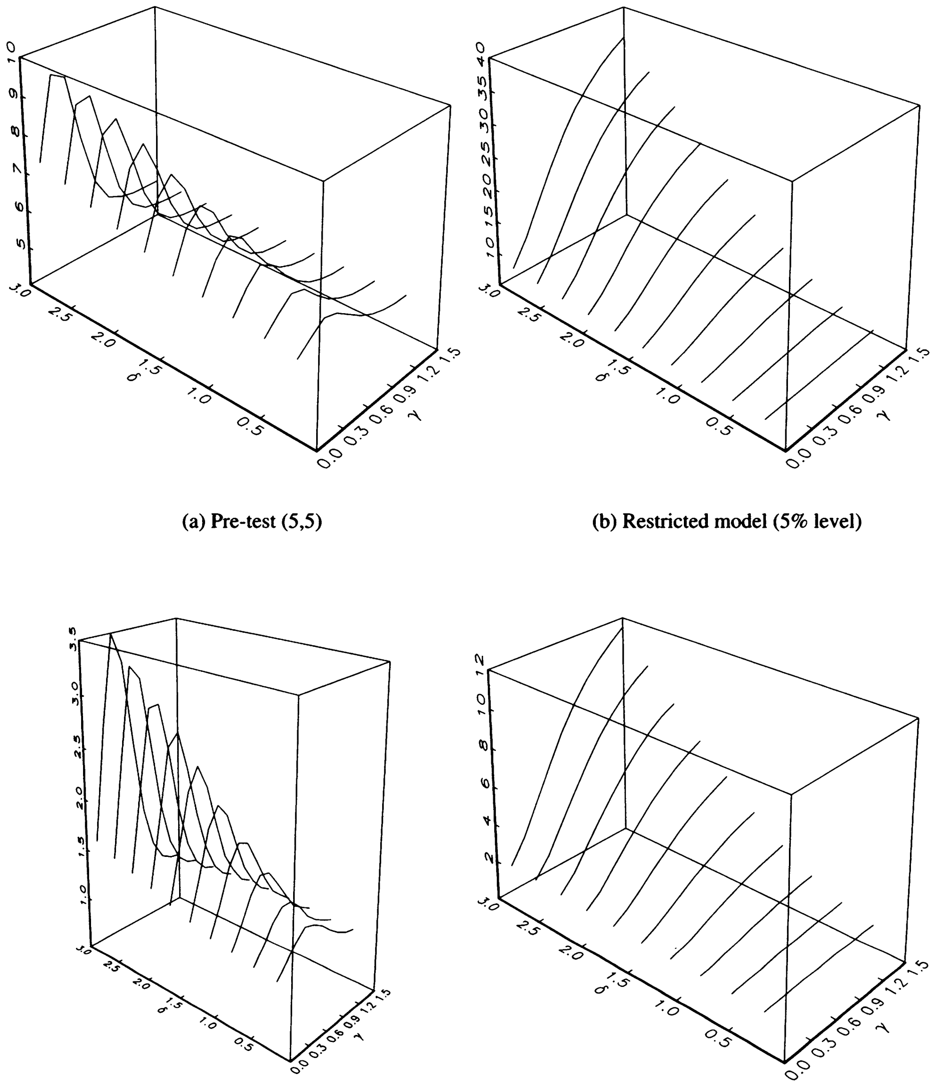
Figure 1. RFs of restricted and pre-test models, forecast-period shift.

False exclusion and in-sample breaks. We now set $\tau = 19$. The results for the restricted model and for pre-testing are shown in Figure 2 at 5% and 1% nominal sizes. The pre-test RFs initially increase in γ (for a given δ) before peaking and then declining towards their nominal size, following the pattern for the forecast-period shift. Note that for small degrees of non-constancy (small δ), model mis-specification (i.e. z_t excluded) does not matter much, but forecast failure increases as the degree of non-constancy increases.

The forecast-failure RFs for the restricted and selected models are less than for the forecast-period shift. The in-sample occurrence of the break increases the denominator of the forecast-failure statistic leading to fewer rejections. In the limit, as τ approaches 1 the surfaces for the restricted and pre-test strategy forecast failures would flatten and resemble those from the first set of experiments, as the influence of the structural break diminishes. The impact of structural breaks on forecast-failure tests is greatest when they are purely forecast-period events which do not affect in-sample observations. Conversely, the rejection frequencies of the in-sample test, H_0: $\gamma = 0$, are higher when the break occurs within-sample. Table 2 records the simulated rejection frequencies at the 1% level, underlying Figure 2c, and Table 3 presents analytical approximations based on the formulae in the Appendix, which are reasonably similar. Thus, as time passes and the shift comes within the forecaster's information set, the model that falsely excludes z_t will be rejected more often, and the excess rejections of both the restricted model and the pre-test model will decline.

False inclusion and post-sample breaks. Preliminary experimentation based on breaks in an irrelevant variable (extending (4)) revealed little of interest,[†] hence we decided to examine cases where the conditional model was mis-specified for the behavioural relation due to cross-equation

[†] For example, when $\gamma = 0$ and δ ranges from 0.3 to 3, the forecast-period break in z_t leads to unrestricted model rejection frequencies of only a little over 5 (generally less than 5.5) and pre-testing rejection frequencies range from 5.5 to around 6.5.



(a) Pre-test (5,5)

(b) Restricted model (5% level)

(c) Pre-test (1,1)

(d) Restricted model (1% level)

Figure 2. Pre-test and restricted model RFs for an in-sample break.

correlation. The experimental design is now given by:

$$
\begin{aligned}
y_t &= \mu + \gamma z_t + \varepsilon_t, \\
z_t &= \delta \times 1_{t \geq \tau} + \zeta_t + \lambda \varepsilon_t. \qquad (6)
\end{aligned}
$$

In (6), ε_t and ζ_t are both IN[0, 1], distributed independently of each other. Also, $T = 20$, $\gamma = 0$, $\tau = 21$ (so the structural break is solely a forecast-period phenomenon), and we consider

Table 2. Rejection frequencies of $\mathsf{H}_0 : \gamma = 0$ at the 1% level.

	In-sample break			
δ	$\gamma = 0.15$	$\gamma = 0.60$	$\gamma = 1.05$	$\gamma = 1.5$
0.3	2.44	42.12	89.84	99.18
1.2	2.71	47.12	92.79	99.58
2.1	3.16	57.92	96.64	99.87
3.0	3.94	71.35	99.03	99.99
	2.47	41.58	89.62	99.16

Table 3. Test power calculations ($\mathcal{P}$) at 1% for break in regressor.

γ	0.15	0.3	0.45	0.6	0.75	0.9	1.05	1.2	1.35	1.5
					Forecast-period shift ($\tau = 21$)					
	3.03	10.7	27.0	52.6	78.4	94.0	99.0	99.9	100.0	100.0
					In-sample shift ($\tau = 19$)					
$\delta = 1.5$	3.55	13.1	34.4	64.5	88.3	98.0	99.8	100.0	100.0	100.0
$\delta = 3$	5.20	22.1	56.4	87.9	98.8	100.0	100.0	100.0	100.0	100.0

The $\mathcal{P}$ values are obtained by linear interpolation of χ^2's with integer degrees of freedom.

combinations of $\delta = \{3, 6, \ldots, 30\}$ and $\lambda = \{0.1, 0.2, \ldots, 1\}$. The correlation between the errors of the y_t and z_t equations induces false inclusion of z_t in a conditional model, even when $\gamma = 0$ (see the Appendix). We assume knowledge of z_{T+1}, but not of the break in that process.

Analytic calculations (see the Appendix for details) recorded in Table 4 indicate that the null hypothesis $\gamma = 0$ will be rejected 7% for $\lambda = 0.1$ up to 90% of the time for $\lambda = 1$.

Although these calculations are approximate, they are close to the simulated rejection frequencies recorded in Table 5 for selected values of λ. Because the process for z_t shifts in the forecast period, the false inclusion of z_t will result in forecast failure in the unrestricted model, even though the restricted model is constant (the simulation confirms that the RFs of the latter are close to the nominal). Figure 3a depicts the RFs for the unrestricted model. These increase in δ and λ, and are up to 95% for $\{\delta, \lambda\} = \{30, 1\}$. Figure 3b plots the ratio of the pre-test to unrestricted model RFs (5% in-sample and forecast-failure tests). Screening by pre-testing for the significance of z_t offers no benefit, in that the pre-test forecast-failure RFs exceed the unrestricted model RFs almost everywhere. This occurs because the event of rejecting on the in-sample test is highly correlated with the event of forecast failure when z_t is included. Table 5 provides some insight into this finding.

Consider $\lambda = 0.1$. The unrestricted model forecast-failure RF approaches 7% for $\delta = 30$. The selected model retains z_t 7.1% of the time (column 1): from column 4, retaining z_t and rejecting on the forecast-failure test occurs 6.2% of the time, so the forecast-failure test rejects $(6.2/7.1) \times 100\% = 87\%$ of the time that z_t is retained. This is much higher than if the tests of forecast failure and retaining z_t were independent (column 3). So when z_t is re-

Table 4. Test power calculations ($\mathcal{P}$) for break in extraneous variable.

λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
					Forecast-period shift ($\tau = 21$)					
	7.29	13.9	24.0	36.4	50.1	62.4	73.2	80.6	86.3	90.2
					In-sample shift ($\tau = 19$)					
$\delta = 3$	6.21	9.79	15.5	22.9	31.5	41.1	51.1	60.2	68.3	75.1
$\delta = 15$	5.10	5.39	5.88	6.56	7.44	8.50	9.74	11.2	12.7	14.5
$\delta = 30$	5.03	5.11	5.23	5.41	5.63	5.91	6.24	6.62	7.04	7.52

The $\mathcal{P}$ values are obtained by linear interpolation of χ^2's with integer degrees of freedom.

Table 5. Pre-tests and out-of-sample tests as filtering devices.

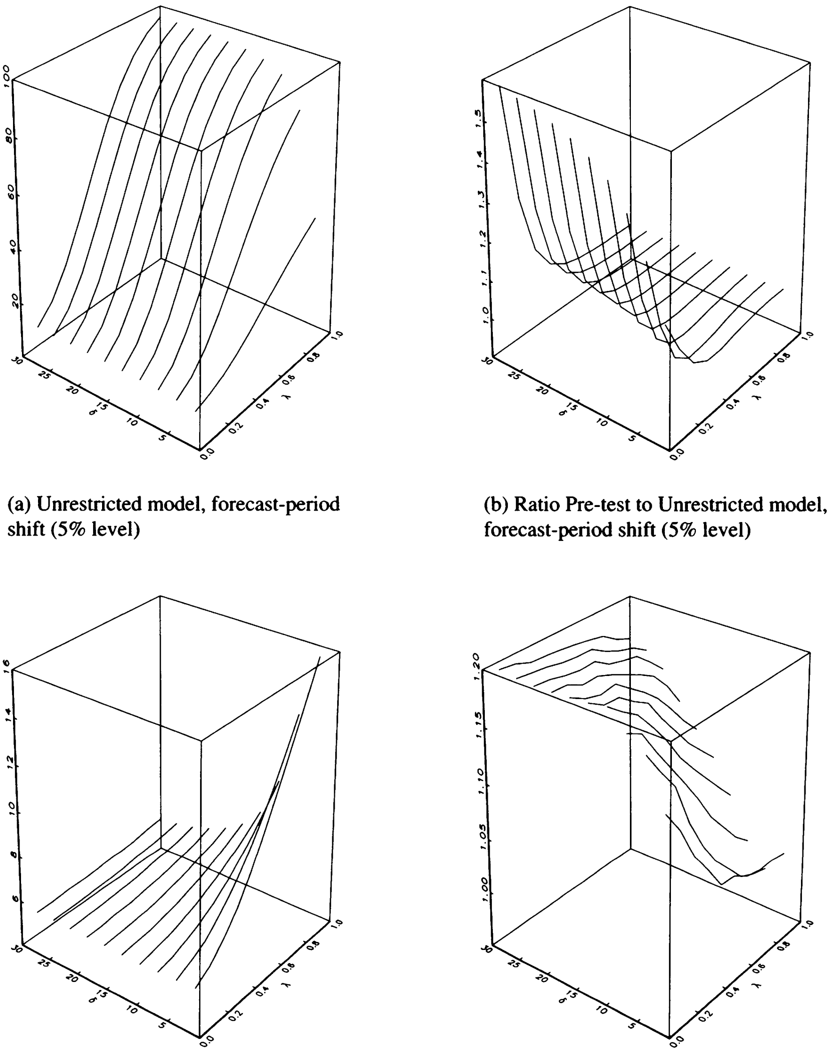
δ	$\lambda = 0.1$				$\lambda = 0.4$				$\lambda = 0.7$				$\lambda = 1.0$			
3	7.08	5.67	0.40	1.84	37.60	13.81	5.19	10.01	78.98	27.87	22.01	26.06	96.12	41.20	39.61	40.79
12	7.08	6.70	0.47	4.95	37.60	33.45	12.58	29.73	78.98	72.32	57.12	69.77	96.12	92.46	88.87	91.77
21	7.08	6.91	0.49	5.84	37.60	36.09	13.57	33.48	78.98	76.75	60.61	74.79	96.12	95.05	91.36	94.47
30	7.08	6.98	0.49	6.23	37.60	36.86	13.86	34.87	78.98	77.82	61.46	76.32	96.12	95.65	91.94	95.16
	92.92	5.06	4.70	4.71	62.40	5.06	3.16	3.51	21.02	5.06	1.06	1.60	3.88	5.06	0.20	0.39

The experiment is for pre-tests and forecast-failure tests at the 5% level. The first of each set of four columns in the top panel is the percentage of time z_t is retained as a result of the pre-test; the second is the % of times the forecast-failure test of the unrestricted model rejects; the third column is the product of columns (1) and (2); the fourth is the % of times the variable is retained and the forecast-failure test rejects. In the second panel there is only one row because the findings are invariant to δ. In this panel, the first of each set of four columns is the percentage of times z_t is dropped, the second is the RF of the restricted (correct) model, the third is the product of the first two columns, and the fourth is the percentage of times that z_t is dropped and the restricted model fails.

By construction, summing the entries in columns 4 across panels gives the RFs for the pre-test model. Summing columns 3 gives the pre-test RF that would result if the in-sample and forecast failure tests were independent: then the pre-test is a weighted average of the restricted and unrestricted model forecast-failure RFs.

tained, the pre-test forecast-failure RFs are only a little below the overall rejection of the unrestricted model. Additionally, the pre-test model also rejects some of the time on the forecast-failure test when z_t is dropped—this is recorded in the last row (panel 2) of the table. While columns 3 and 4 are similar, nevertheless the pre-test rejects 4.7% of the time when z_t is dropped, so that the total forecast failure RF of 10.9% is more than 1.5 times as large as the unrestricted forecast-failure RF of 7%. However, the factor of 1.5 is misleading: from Figure 3 this only holds at low λ, when the forecast-failure RFs of both models are small in absolute terms. As the forecast-failure RFs of the unrestricted model increase, the ratio of the pre-test to the unrestricted falls to unity. For large λ, the pre-test model RFs approach those of the unrestricted model.

Intuitively, the reason the pre-test strategy is ineffective in countering the 'excess' forecast-test rejections of the unrestricted model is that on iterations of the Monte Carlo when the realizations of the variables are such that false inclusion of z_t is unlikely to be harmful to forecast performance, it is likely to be excluded, whereas when it is likely to be harmful, it is likely to be included. In terms of model discovery, though, this is a strength rather than a weakness. From a



(a) Unrestricted model, forecast-period shift (5% level)

(b) Ratio Pre-test to Unrestricted model, forecast-period shift (5% level)

(c) Unrestricted model, in-sample shift (5% level)

(d) Ratio Pre-test to Unrestricted model, in-sample shift (5% level)

Figure 3. RFs of pre-test and unrestricted models: false inclusion.

modelling perspective, both the in-sample pre-test and the test of forecast failure can be viewed as filtering devices for weeding out mis-specified models. The times when the false model survives the pre-test (i.e. the null that $\gamma = 0$ is rejected) are precisely those occasions when the model with the irrelevant variable is rejected on the out-of-sample test. This effectiveness of the combined test matches the findings in Hoover and Perez (1999).

False inclusion and in-sample breaks. To investigate the effects of structural shifts in extraneous variables that occur within sample, we kept the above formulation, but set $\tau = 19$.[†] As the Appendix shows, when the occurrence of a break is unknown, the whole-sample linear model relating y_t to z_t can be written as:

$$y_t = \mu + \phi z_t + v_t \qquad \text{where} \quad \phi = \frac{\lambda}{1 + \lambda^2 + 2\delta^2/T}, \tag{7}$$

so a test for including z_t is a test of $\phi = 0$. That null is now rejected less often, and as (7) suggests, the rejection rate falls as δ increases for a given value of λ. This change, relative to $\tau = 21$, occurs because of the failure of the y_t and z_t series to 'co-break' (see Hendry (1995c))—the in-sample shift in z_t is not matched by a shift in y_t—so that the coefficient on z_t is reduced towards zero.

Table 4 records analytical results for selected values of δ (based on the calculations set out in the Appendix), and these are close to the simulated RFs (see Table 6), confirming the value of the approximations. The results for the unrestricted model and the selected model forecast-failure RFs are shown in Figure 3. Table 6 details the simulation results for this experiment.

For a given δ, the unrestricted model forecast-failure RF is increasing in λ (as for the forecast-period shift), but now for a given λ, the forecast-failure RF is decreasing in δ, because as δ gets large, γ is driven to zero, and the model approaches the (correctly specified) restricted model. The forecast failure RF of the pre-test model is also decreasing in δ, the size of the break, because the number of times that z_t is retained falls as δ increases. From a comparison of Figure 3c and d, it is evident that the ratio of pre-test to unrestricted-model RFs is lower when the shift occurs in-sample.

From a PRS perspective, Table 4 confirms that the correct model (which excludes z_t) is rejected less often as time passes and the 'forecast shift' occurs within sample, and moreover, the forecast-failure RFs of the pre-test model are less inflated (compare Tables 5 and 6).

4.4. *Overview*

When the correct specification is unknown and the DGP is prone to structural shifts, the three strategies of unrestricted, restricted and selected models have disparate behaviour on tests for forecast failure. Naturally, the correct specification invariably does best, but switches from the unrestricted to the restricted depending on the unknown state of nature: across states, always using either one does badly. The usual empirical practice of selecting the model in the light of the sample evidence fares reasonably in two senses: first it rarely costs much to select against the first-best alternative, and can greatly dominate the other option; secondly, as part of a progressive research strategy, forecast mistakes at one point in time induce better selection later. Generally, the simple examples considered in detail here do not suggest that over-fitting is a primary cause of forecast failure. But while the models are simple, selection nevertheless induces fairly complicated effects, so that some of the outcomes are at first sight surprising, although the sizes of

[†] By conditioning on $1_{t \geq \tau}$, that is, knowing the timing and occurrence of a break, as well as on z_t, then for $t \geq \tau$:

$$\mathsf{E}[y_t \mid z_t, 1_{t \geq \tau}] = \mu + \frac{\lambda}{1 + \lambda^2} z_t - \frac{\lambda\delta}{1 + \lambda^2} 1_{t \geq \tau},$$

so that the 'unrestricted' model of y_t on a constant and z_t falsely *excludes* $1_{t \geq \tau}$, rather than falsely including z_t. However, we view our example as investigating false inclusion where the break is unknown. We are grateful to a referee for bringing this interpretation to our attention.

Table 6. Pre-tests and out-of-sample tests as filtering devices.

δ	$\lambda = 0.1$				$\lambda = 0.4$				$\lambda = 0.7$				$\lambda = 1.0$			
3	5.915	5.155	0.3049	1.012	22.50	7.163	1.612	3.409	54.64	11.05	6.037	8.487	82.22	15.54	12.77	14.35
6	5.281	5.087	0.2686	1.138	11.85	6.412	0.7596	2.428	26.63	9.058	2.412	5.244	47.61	12.81	6.097	9.565
9	5.090	5.041	0.2566	1.122	8.247	5.759	0.4749	1.807	15.57	7.275	1.133	3.304	26.99	9.514	2.568	5.659
12	5.025	4.980	0.2502	1.127	6.751	5.452	0.3681	1.566	11.10	6.389	0.7091	2.444	17.68	7.770	1.374	3.817
15	4.972	5.012	0.2492	1.153	6.085	5.279	0.3212	1.421	8.799	5.908	0.5198	1.989	13.19	6.872	0.9066	2.907
18	4.955	5.020	0.2487	1.154	5.732	5.167	0.2962	1.330	7.600	5.627	0.4277	1.755	10.66	6.350	0.6770	2.392
21	4.937	5.018	0.2477	1.150	5.491	5.129	0.2816	1.280	6.877	5.442	0.3742	1.600	9.050	5.984	0.5416	2.047
24	4.938	5.020	0.2479	1.149	5.343	5.120	0.2736	1.258	6.420	5.344	0.3431	1.502	8.055	5.729	0.4615	1.851
27	4.949	5.014	0.2481	1.153	5.261	5.088	0.2677	1.224	6.079	5.271	0.3204	1.424	7.375	5.547	0.4091	1.691
30	4.955	5.014	0.2484	1.166	5.192	5.072	0.2633	1.206	5.841	5.197	0.3036	1.359	6.878	5.453	0.3751	1.597
3	94.09	5.059	4.760	4.794	77.50	5.059	3.920	4.117	45.36	5.059	2.295	2.747	17.78	5.059	0.8995	1.341
6	94.72	5.059	4.792	4.807	88.15	5.059	4.460	4.531	73.38	5.059	3.712	3.956	52.39	5.059	2.651	3.078
9	94.91	5.059	4.801	4.826	91.75	5.059	4.642	4.684	84.43	5.059	4.271	4.387	73.01	5.059	3.693	3.931
12	94.98	5.059	4.805	4.829	93.25	5.059	4.717	4.741	88.90	5.059	4.498	4.554	82.32	5.059	4.165	4.305
15	95.03	5.059	4.807	4.827	93.92	5.059	4.751	4.768	91.20	5.059	4.614	4.648	86.81	5.059	4.392	4.481
18	95.05	5.059	4.808	4.829	94.27	5.059	4.769	4.776	92.40	5.059	4.675	4.699	89.34	5.059	4.520	4.568
21	95.06	5.059	4.809	4.824	94.51	5.059	4.781	4.787	93.12	5.059	4.711	4.729	90.95	5.059	4.601	4.634
24	95.06	5.059	4.809	4.825	94.66	5.059	4.789	4.796	93.58	5.059	4.734	4.752	91.95	5.059	4.651	4.690
27	95.05	5.059	4.809	4.820	94.74	5.059	4.793	4.800	93.92	5.059	4.751	4.761	92.63	5.059	4.686	4.711
30	95.05	5.059	4.808	4.815	94.81	5.059	4.796	4.802	94.16	5.059	4.764	4.765	93.12	5.059	4.711	4.728

The experiment is for pre-tests and forecast failure tests at the 5% level. The first of each set of four columns in the first panel is the percentage of time z_t is retained as a result of the pre-test; the second is the % of times the forecast failure-test of the unrestricted model rejects; the third column is the product of columns (1) and (2); the fourth is the % of times the variable is retained and the forecast failure test rejects. In the second panel the first of each set of four columns is the percentage of times z_t is dropped, the second is the RF of the restricted (correct) model, the third is the product of the first two columns, and the fourth is the percentage of times that z_t is dropped and the restricted model fails.

By construction, summing the entries in columns 4 across panels gives the RFs for the pre-test model. Summing columns 3 gives the pre-test RF that would result if the in-sample and forecast failure tests were independent: then the pre-test is a weighted average of the restricted and unrestricted model forecast failure RFs.

the effects are small. However, the suspicion may remain that multiple choices could exacerbate any potential failings from selection in more general models, notwithstanding the findings of Hoover and Perez (1999) which suggest that this does not occur. The next section investigates the performance of a more general model in a realistic setting as a check on the generality of the simple, static model results.

5. A SIMULATION STUDY OF A DYNAMIC, MULTIVARIATE MODEL

A number of problems arise in designing realistic simulation experiments to assess the impact of model selection strategies, such as Gets, on forecast failure and forecast accuracy. The DGPs have to be high dimensional to mimic the applied researcher's task of obtaining a useful model when faced by a range of potential explanatory variables with unknown lag structures. The high dimension of the problem implies that many parameters (coefficients, standard errors, covariances etc.) need to be specified. To reduce the specificity of the conclusions from one set of parameter values, other values might need to be considered, but the dimensionality of the problem entails a large number of combinations of candidate sets of parameter values. We also wish to explore the effects of model selection when the initial model (the general model) is both over- and under-parametrized relative to the DGP, so a wide range of DGPs and models needs to be entertained.

Our approach is not immune to these problems, but is designed to address the key issues. To ensure their empirical relevance, the DGPs are estimated VARs based on the set of variables used by a number of authors in modelling UK M1.[†] The data are quarterly and seasonally adjusted, transformed as follows: the first difference of the log of real M1 (nominal money divided by the total expenditure deflator), the first difference of the log of real total final expenditure (TFE), the second difference of the log of the TFE deflator (i.e. the change in the rate of inflation), and the first difference of the 3-month local-authority interest rate. The cited literature posits two long-run equilibrium relationships between the variables, but these are ignored here, and the DGPs are n-th order VARs of the above four variables, where $n = 1, \dots, 4$. The 'general models' of the sequential simplification strategies are the equations from the VAR for the first variable, of lag order m, where $m = 1, \dots, 4$, so that there are 4^2 combinations of DGPs and general models. Hence we have correctly , under- and over-specified general models.[‡] Two simplification strategies are examined. On each round of model simplification, the first approach simultaneously deletes all those regressors whose coefficients are not individually significant at the 5% level on the basis of one-off t-tests ('multiple simplification'). The second recognizes that the explanatory variables are likely to be highly correlated, and on each round deletes the one with the smallest (absolute) t-value less than the critical value ('single simplification'). While sequential simplification is one aspect of the Gets modelling strategy, there are others which are beyond the scope of the present simulation exercise, so our results specifically relate to model simplification by variable deletion, rather than Gets modelling (e.g. prior orthogonalization, encompassing, multiple-path searches, and diagnostic checking are not incorporated: see Hoover and Perez (1999), and Hendry and Krolzig (1999)). After testing for the significance of the regressors in each round, the restricted model is estimated, forecasts 1-step ahead, and a forecast-failure test is calculated. This has the form of equation (1), except that for predicting $T + 1$, $\mathbf{X}$ contains data up to and including $T - 1$, and $\mathbf{x}_T$ replaces $\mathbf{x}_{T+1}$ in the denominator of the statistic, so that only lagged information is used. As remarked in Section 2, because of the presence of lagged dependent variables, the test is no longer exact.

The MSFE of the forecast is also calculated, and this is expressed as a ratio to the DGP money-equation error variance (population value). The 'excess MSFE' then measures forecast accuracy, while the forecast-failure test compares the out-of-sample squared error to the average in-sample squared errors.

The notional number of simplification rounds in each case is 16. This accommodates the model order $m = 4$, but for lower-order models (in particular), implicit stopping rules bite much earlier. For example, if on round r ($r < 16$) none of the explanatory variables is insignificantly different from zero, then the values of the MSFE and forecast-failure test for $r + 1, \dots, 16$ are

[†] See, *inter alia*, Hendry (1979), Hendry and Ericsson (1991), Hendry and Mizon (1993), Hendry (1996) and Hendry and Doornik (1994).

[‡] DGPs with large n may contain many parameters that are close to zero.

set equal to their round r values. Similarly, if on round r, none of the explanatory variables (other than the constant) is retained.

Two sample sizes are used in the study: $T = 88$ and $T = 30$. In both cases the data start in 1964. This creates 32 experiments. We also replicate all the above allowing for an unmodelled shift in the money-demand equation of 2% points affecting the last eight estimation periods and the forecast-period observation. Empirically, this is in line with the effect of the financial innovation that occurred in the mid 1980s, when the payment of interest on checking accounts became legal. In the simulations, it allows us to assess the impact of model selection interacting with parameter non-constancy in a dynamic, high-dimensional setting.

For each of the 64 simulation experiments, we perform 20 000 replications. The data are generated from the VAR(n) model using pseudo-random numbers from the Gauss function `RNDN`, transformed to have a covariance matrix in accordance with that for the VAR(n) estimated on the empirical data.

5.1. Results

To condense the information presented, we average over the cases for which the general model is correctly specified [i.e. $\{m, n = (1, 1), (2, 2), (3, 3), (4, 4)\}$], the model lag order is over-parametrized by one [i.e. $\{m, n = (2, 1), (3, 2), (4, 3)\}$] etc. For the multiple strategies, we present eight rounds of simplification, but there are generally no changes after the third. For the single-strategy results, there are redundant parameters when $m > n$, and since these can only be omitted one at a time, in some cases, the results change down to round 16.

The multiple-simplification strategy (see Table 8) leads to fairly modest increases in the forecast-failure test rejection frequencies above the nominal 5% level, in no case exceeding 11%. When the general model has the correct order ($m = n$), successive rounds of simplification lead to a rejection rate of just over 7% for $T = 30$—this rises to 9% when there is a shift, and the $T = 30$ results are little affected by the increase in the sample size to $T = 88$.

Under-parametrization ($m < n$) inflates 'general model' (i.e. round 0) rejection frequencies when there is a shift, relative to over-parametrization. However, after the successive rounds of simplifications, there is little to choose between under and over-specification when $T = 88$, whether there is a shift or not. Under-specification inflates 'final model' rejection frequencies at $T = 30$.

Single simplifications (Tables 7 to 11) lead to more excess rejections. For $T = 30$, over-parametrization is worse than under-parametrization. In the extreme case of $m = n + 3$, the rejection rate goes up to 20%. For $T = 88$, a similar pattern emerges to the multiple case: for the no-shift experiments, the excess rejections problem is alleviated, and for the shift case, although under-parametrization is worse than over for the 'general model', by the time the final model is reached there is little to choose between the two.

In summary, then, over-parametrization in conjunction with a small sample size might inflate rejection frequencies to as high as 20% when a single simplification strategy is adopted. Otherwise, rejection frequencies are generally little higher than 10%. Since none of the standard tests in dynamic models are exact, the last is not a notable departure.

What of forecast accuracy, as measured by the excess of the MSFE over the in-sample (population) error variance? First, when $m = n$, there are sharp drops between the general and final model MSFEs for $T = 30$, and these are greater for the multiple than the single simplification strategy, since the final model MSFEs for the former are lower. These drops are exacerbated as

Table 7. Forecast-failure rejection frequencies: 'single', $T = 30$ without shift.

Round	$m = n$	$m = n + 1$	$m = n + 2$	$m = n + 3$	$m = n - 1$	$m = n - 2$	$m = n - 3$
0	0.046	0.048	0.048	0.045	0.049	0.050	0.058
1	0.062	0.066	0.067	0.071	0.061	0.060	0.069
2	0.078	0.084	0.091	0.100	0.072	0.070	0.078
3	0.092	0.099	0.112	0.129	0.082	0.078	0.087
4	0.101	0.114	0.131	0.154	0.089	0.082	0.089
8	0.117	0.132	0.161	0.203	0.097	0.090	0.089
12	0.114	0.125	0.145	0.181	0.097	0.090	0.089
16	0.118	0.130	0.150	0.192	0.097	0.090	0.089

Table 8. Forecast-failure rejection frequencies: 'multiple' strategy.

Round	$m = n$	$m = n + 1$	$m = n + 2$	$m = n + 3$	$m = n - 1$	$m = n - 2$	$m = n - 3$
$T = 30$: no shift							
0	0.0463	0.0477	0.0483	0.0449	0.0487	0.0499	0.0581
1	0.0718	0.0684	0.0691	0.0711	0.0739	0.0794	0.0858
2	0.0718	0.0678	0.0683	0.0683	0.0742	0.0803	0.0876
4	0.0715	0.0673	0.0676	0.0679	0.0741	0.0802	0.0878
8	0.0715	0.0672	0.0675	0.0677	0.0741	0.0802	0.0878
$T = 30$: shift							
0	0.0533	0.0478	0.0453	0.0438	0.0608	0.0713	0.0760
1	0.0907	0.0852	0.0886	0.0885	0.0958	0.0987	0.105
2	0.0915	0.0867	0.0887	0.0888	0.0953	0.0988	0.106
4	0.0911	0.0864	0.0881	0.0885	0.0954	0.0988	0.106
8	0.0911	0.0864	0.0880	0.0886	0.0954	0.0988	0.106
$T = 88$: no shift							
0	0.0488	0.0489	0.0489	0.0500	0.0524	0.0529	0.0551
1	0.0617	0.0651	0.0633	0.0638	0.0647	0.0620	0.0623
2	0.0626	0.0656	0.0634	0.0637	0.0652	0.0628	0.0625
4	0.0625	0.0654	0.0634	0.0635	0.0653	0.0625	0.0625
8	0.0625	0.0654	0.0634	0.0635	0.0653	0.0625	0.0625
$T = 88$: shift							
0	0.066	0.067	0.071	0.072	0.081	0.091	0.094
1	0.091	0.095	0.103	0.101	0.102	0.106	0.108
2	0.093	0.096	0.104	0.101	0.103	0.106	0.109
4	0.093	0.096	0.104	0.101	0.103	0.106	0.109
8	0.093	0.096	0.104	0.101	0.103	0.106	0.109

Table 9. Forecast-failure rejection frequencies: 'single', $T = 30$ with shift.

Round	$m = n$	$m = n + 1$	$m = n + 2$	$m = n + 3$	$m = n - 1$	$m = n - 2$	$m = n - 3$
0	0.053	0.048	0.045	0.044	0.061	0.071	0.076
1	0.070	0.066	0.066	0.069	0.075	0.083	0.088
2	0.086	0.085	0.090	0.097	0.088	0.095	0.099
3	0.010	0.104	0.111	0.125	0.099	0.101	0.109
4	0.111	0.120	0.131	0.152	0.107	0.106	0.111
8	0.132	0.144	0.166	0.204	0.120	0.111	0.111
12	0.132	0.140	0.159	0.188	0.121	0.111	0.111
16	0.138	0.148	0.173	0.212	0.121	0.111	0.111

Table 10. Forecast-failure rejection frequencies: 'single', $T = 88$ without shift.

Round	$m = n$	$m = n + 1$	$m = n + 2$	$m = n + 3$	$m = n - 1$	$m = n - 2$	$m = n - 3$
0	0.049	0.049	0.049	0.050	0.052	0.053	0.055
1	0.052	0.052	0.052	0.054	0.055	0.056	0.058
2	0.055	0.055	0.056	0.057	0.058	0.058	0.060
4	0.059	0.060	0.062	0.063	0.062	0.061	0.062
8	0.063	0.065	0.066	0.073	0.065	0.062	0.062
12	0.064	0.067	0.067	0.071	0.066	0.062	0.062
16	0.064	0.067	0.068	0.082	0.066	0.062	0.062

Table 11. Forecast-failure rejection frequencies: 'single', $T = 88$ with shift.

Round	$m = n$	$m = n + 1$	$m = n + 2$	$m = n + 3$	$m = n - 1$	$m = n - 2$	$m = n - 3$
0	0.066	0.067	0.071	0.072	0.081	0.091	0.094
1	0.070	0.071	0.075	0.076	0.085	0.094	0.098
2	0.074	0.074	0.079	0.080	0.089	0.098	0.104
3	0.077	0.078	0.083	0.084	0.092	0.102	0.107
4	0.080	0.082	0.087	0.088	0.095	0.103	0.108
8	0.087	0.090	0.097	0.101	0.100	0.106	0.108
12	0.090	0.093	0.101	0.103	0.102	0.106	0.108
16	0.090	0.094	0.104	0.124	0.102	0.106	0.108

the extent of model over-parametrization increases. For $T = 30$, over-parametrization is more costly than under at the general model stage, though the final models of the over-parametrized specifications invariably have smaller MSFEs. For $T = 88$, the under-parametrized general and final models have larger MSFEs than their over-parametrized counterparts. The shift cases have markedly higher MSFEs than the constant-parameter cases, as expected.

Table 12. MSFE for 'multiple' strategy.

Round	$m = n$	$m = n + 1$	$m = n + 2$	$m = n + 3$	$m = n - 1$	$m = n - 2$	$m = n - 3$
$T = 30$: no shift							
0	2.05	2.28	2.75	3.37	1.76	1.63	1.64
1	1.63	1.49	1.49	1.52	1.66	1.76	1.89
2	1.62	1.47	1.46	1.47	1.66	1.77	1.92
4	1.61	1.46	1.46	1.47	1.66	1.77	1.92
8	1.61	1.46	1.46	1.47	1.66	1.77	1.92
$T = 30$: shift							
0	2.87	3.18	3.77	4.96	2.50	2.34	2.22
1	2.30	2.19	2.25	2.36	2.30	2.33	2.40
2	2.30	2.18	2.22	2.32	2.30	2.35	2.43
4	2.30	2.18	2.22	2.32	2.30	2.35	2.43
8	2.30	2.18	2.22	2.32	2.30	2.35	2.43
$T = 88$: no shift							
0	1.15	1.19	1.22	1.27	1.28	1.37	1.55
1	1.21	1.23	1.20	1.18	1.35	1.42	1.60
2	1.23	1.24	1.20	1.18	1.36	1.43	1.60
4	1.23	1.24	1.20	1.18	1.36	1.43	1.60
8	1.23	1.24	1.20	1.18	1.36	1.43	1.60
$T = 88$: shift							
0	1.40	1.47	1.55	1.61	1.67	1.89	2.17
1	1.56	1.59	1.64	1.58	1.81	2.01	2.28
2	1.58	1.61	1.65	1.58	1.83	2.01	2.28
4	1.59	1.61	1.65	1.58	1.83	2.01	2.28
8	1.59	1.61	1.65	1.58	1.83	2.01	2.28

Table 13. MSFEs for 'single' strategy: $T = 30$ without shift.

Round	$m = n$	$m = n + 1$	$m = n + 2$	$m = n + 3$	$m = n - 1$	$m = n - 2$	$m = n - 3$
0	2.05	2.28	2.75	3.37	1.76	1.63	1.64
1	2.05	2.27	2.74	3.35	1.76	1.64	1.65
2	2.05	2.26	2.72	3.33	1.77	1.66	1.73
4	2.01	2.18	2.61	3.20	1.76	1.68	1.86
8	1.85	1.91	2.19	2.67	1.69	1.72	1.86
12	1.75	1.74	1.89	2.15	1.68	1.72	1.86
16	1.83	1.85	2.05	2.46	1.68	1.72	1.86

Table 14. MSFEs for 'single' strategy: $T = 30$ with shift.

Round	$m = n$	$m = n + 1$	$m = n + 2$	$m = n + 3$	$m = n - 1$	$m = n - 2$	$m = n - 3$
0	2.87	3.18	3.77	4.96	2.50	2.34	2.22
1	2.87	3.17	3.76	4.95	2.50	2.34	2.21
2	2.86	3.15	3.74	4.91	2.48	2.34	2.25
3	2.83	3.11	3.69	4.82	2.47	2.34	2.35
4	2.79	3.05	3.61	4.68	2.45	2.32	2.39
8	2.61	2.70	3.08	3.86	2.38	2.34	2.39
12	2.50	2.49	2.72	3.10	2.38	2.34	2.39
16	2.63	2.69	3.05	3.76	2.38	2.34	2.39

Table 15. MSFEs for 'single' strategy: $T = 88$ without shift.

Round	$m = n$	$m = n + 1$	$m = n + 2$	$m = n + 3$	$m = n - 1$	$m = n - 2$	$m = n - 3$
0	1.15	1.19	1.22	1.27	1.28	1.37	1.55
1	1.15	1.19	1.22	1.27	1.28	1.37	1.56
2	1.16	1.19	1.22	1.27	1.29	1.38	1.57
4	1.17	1.19	1.21	1.27	1.31	1.41	1.60
8	1.19	1.19	1.19	1.24	1.33	1.42	1.60
12	1.20	1.21	1.18	1.19	1.34	1.42	1.60
16	1.20	1.22	1.21	1.43	1.34	1.42	1.60

Table 16. MSFEs for 'single' strategy: $T = 88$ with shift.

Round	$m = n$	$m = n + 1$	$m = n + 2$	$m = n + 3$	$m = n - 1$	$m = n - 2$	$m = n - 3$
0	1.40	1.47	1.55	1.61	1.67	1.89	2.17
1	1.41	1.47	1.55	1.61	1.68	1.89	2.18
2	1.42	1.47	1.55	1.61	1.69	1.92	2.22
4	1.44	1.46	1.55	1.60	1.72	1.96	2.27
8	1.48	1.48	1.53	1.58	1.78	1.99	2.27
12	1.51	1.52	1.55	1.54	1.79	1.99	2.27
16	1.52	1.53	1.59	1.87	1.79	1.99	2.27

6. CONCLUSIONS

The simulations based on the simple static models suggest that model selection or pre-testing is unlikely to matter greatly in a constant-parameter world. Whether we include an irrelevant variable (and thereby overfit), or exclude a relevant variable, or use model selection, rejection rates of forecast-failure tests will generally be close to the nominal levels, providing the data properties are unchanged between the sample and the forecast periods. Model mis-specification will not be flagged by inflated forecast-failure rejection frequencies, and model-selection effects would appear to be small even in quite small samples.

Non-constancies in the form of deterministic shifts first occur in the forecast period, then enter a later in-sample information set. Falsely excluding a variable which changes in the forecast period will induce forecast failure: to the extent that the selection strategy results in the variable being dropped from the model, the pre-test rejection rate will exceed the nominal. But as time passes and the shift comes within the forecaster's information set, the hypothesis that the variable should be excluded will be rejected more often, and the pre-test model forecast-failure rejection frequency will decline, as the variable is retained more often. For falsely including a variable, a similar pattern emerges. As we move from the forecast-period shift to an estimation-period shift, the correct model is rejected less often, and the rejection rate of the pre-test model is closer to the nominal. But the pre-test forecast-failure rejection rate will generally exceed that of the unrestricted model, because the extraneous variable will tend to be retained on precisely those occasions when its presence contributes most to forecast failure. This dependence between the in-sample and forecast tests tempers the usefulness of pre-testing as a screening device to reject irrelevant variables, at least from a forecasting perspective.

In general, though, these simulation results coupled with the demand-for-money example give little support to assertions that model-selection strategies are an important cause of forecast failure empirically. This finding is consistent with Clements and Hendry (1999, Chapters 3 and 4), where our analysis of the causes of forecast failure led us to the conclusion that the apparent success/failure in forecasting of a model depended primarily on what happened over the forecast horizon relative to the in-sample model, not on the degree of congruence/non-congruence of a model in-sample, which is neither necessary nor sufficient for forecasting success or failure. Thus, for forecasting, the route by which a final model is arrived at may matter less than is sometimes believed.

ACKNOWLEDGEMENTS

Financial support from the UK Economic and Social Research Council under research grants L116251015 and L138251009 is acknowledged by both authors. Helpful comments by two anonymous referees and Neil Shephard are also gratefully acknowledged.

REFERENCES

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

Clark, T. E. (2000). Can out-of-sample forecast comparisons help prevent overfitting? Research Division, Federal Reserve Bank of Kansas City.

Clements, M. P. and D. F. Hendry (1998). *Some Methodological Implications of Forecast Failure*. Mimeo, Institute of Economics and Statistics, University of Oxford.

Clements, M. P. and D. F. Hendry (1999). *Forecasting Non-Stationary Economic Time Series*. Cambridge, Mass.: MIT Press. The Zeuthen Lectures on Economic Forecasting.

Gilbert, C. L. (1986). Professor Hendry's econometric methodology. *Oxford Bulletin of Economics and Statistics 48*, 283–307.

Granger, C. W. and A. Timmermann (1999). Data mining with local model specification uncertainty: a discussion of Hoover and Perez. *Econometrics Journal 2*, 220–5.

Hand, D. J. (1999). Discussion contribution on 'Data mining reconsidered: encompassing and the general-to-specific approach to specification search' by Hoover and Perez. *Econometrics Journal 2*, 241–3.

Hansen, B. E. (1999). Discussion of 'Data mining reconsidered'. *Econometrics Journal 2*, 192–201.

Hendry, D. F. (1979). Predictive failure and econometric modelling in macro-economics: The transactions demand for money. In P. Ormerod (ed.), *Economic Modelling*, pp. 217–42. London: Heinemann.

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

Hendry, D. F. (1995b). Econometrics and business cycle empirics. *Economic Journal 105*, 1622–36.

Hendry, D. F. (1995c). *A Theory of Co-breaking*. Mimeo, Nuffield College, University of Oxford.

Hendry, D. F. (1996). On the constancy of time-series econometric equations. *Economic and Social Review 27*, 401–22.

Hendry, D. F. and J. A. Doornik (1994). Modelling linear dynamic econometric systems. *Scottish Journal of Political Economy 41*, 1–33.

Hendry, D. F. and N. R. Ericsson (1991). Modeling the demand for narrow money in the United Kingdom and the United States. *European Economic Review 35*, 833–86.

Hendry, D. F. and H.-M. Krolzig (1999). *On the Properties of General-to-simple Modelling*. Mimeo, Oxford Institute of Economics and Statistics, Oxford.

Hendry, D. F. and G. E. Mizon (1993). Evaluating dynamic econometric models by encompassing the VAR. In P. C. B. Phillips (ed.), *Models, Methods and Applications of Econometrics*, pp. 272–300. Oxford: Basil Blackwell.

Hess, G. D., C. S. Jones and R. D. Porter (1998). The predictive failure of the Baba, Hendry and Starr model of M1. *Journal of Economics and Business 50*, 477–507.

Hoover, K. D. and S. J. Perez (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal 2*, 167–91.

Johnson, N. L. and S. Kotz (1970). *Continuous Univariate Distributions—2*. New York: John Wiley.

Kiviet, J. F. (1986). On the rigor of some mis-specification tests for modelling dynamic relationships. *Review of Economic Studies 53*, 241–61.

Leamer, E. E. (1978). *Specification Searches. Ad-Hoc Inference with Non-Experimental Data*. New York: John Wiley.

Todd, R. M. (1990). Improving economic forecasting with Bayesian vector autoregression. In C. W. J. Granger (ed.), *Modelling Economic Series*, pp. 214–34. Oxford: Clarendon Press.

White, H. (1990). A consistent model selection. In C. W. J. Granger (ed.), *Modelling Economic Series*, pp. 369–83. Oxford: Clarendon Press.

7. APPENDIX

7.1. Test power calculations

Break in an included variable. The DGP is given by (4) and $z_t = \delta \times 1_{t \geq \tau} + \zeta_t$, $\zeta_t \sim \mathsf{IN}[0, 1]$, for $T = 20$. Let $\pi = 1 - (\tau - 1)/T$, so $\pi = 1$ when the break occurs at the first observation of the estimation sample ($\tau = 1$) and $\pi = 0$ when $\tau = T + 1$.

Under the null that $\gamma = 0$ in (4), the square of the t-statistic has the central $\mathsf{F}^1_T(0)$ distribution, which can be approximated by the central $\chi^2(1)$ distribution, i.e.:

$$\mathsf{t}^2 = \frac{T(\widehat{\gamma} - \gamma)^2 (T^{-1}\mathbf{z}'\mathbf{z})}{\widehat{\sigma}^2_\varepsilon} \sim \mathsf{F}^1_T(0) \;{}_{\widetilde{app}}\; \chi^2(1),$$

where $\mathbf{z} = (z_1, \ldots, z_T)'$. Hence:

$$\mathsf{t}^2_{\gamma=0} = \frac{T\widehat{\gamma}^2(T^{-1}\mathbf{z}'\mathbf{z})}{\widehat{\sigma}^2_\varepsilon} \,\widetilde{app}\, \chi^2(1, \psi^2),$$

where ψ^2 is the non-centrality parameter:

$$\psi^2 = T\gamma^2\sigma_\varepsilon^{-2}T^{-1}\mathsf{E}[\mathbf{z}'\mathbf{z}] = T\gamma^2(1 + \pi\delta^2),$$

because $\mathsf{E}[\mathbf{z}'\mathbf{z}] = T(\sigma^2_\zeta + \pi\delta^2)$ and $\sigma^2_\varepsilon = \sigma^2_\zeta = 1$. We approximate the non-central χ^2 distribution by a proportion of the central χ^2 distribution, see e.g. Johnson and Kotz (1970) and Hendry (1995a, p. 475):

$$\mathsf{t}^2_{\gamma=0} = \gamma^2 T(1 + \pi\delta^2) \sim \chi^2(1, \psi^2) \simeq h\chi^2(m),$$

where:

$$h = 1 + \frac{\psi^2}{1 + \psi^2} \to 2, \qquad m = 1 + \frac{\psi^4}{1 + 2\psi^2} \to \infty,$$

as $T \to \infty$. Using this approximation, the power of the test under the alternative, H_T, is given by:

$$\mathsf{P}(\mathsf{t}^2_{\gamma=0} > c_\alpha \mid \mathsf{H}_T) = \mathsf{P}(\chi^2(1, \psi^2) > c_\alpha) \simeq \mathsf{P}(\chi^2(m, 0) > h^{-1}c_\alpha).$$

Break in an extraneous variable. The experimental design in Section 4.3 is:

$$\begin{aligned} y_t &= \mu + \gamma z_t + \varepsilon_t, \\ z_t &= \delta \times 1_{t \geq \tau} + \zeta_t + \lambda\varepsilon_t. \end{aligned} \tag{8}$$

where ε_t and ζ_t are both $\mathsf{IN}[0, 1]$, and distributed independently of each other. Also, $T = 20$, $\gamma = 0$, and either $\tau = 21$ or $\tau = 19$ (the structural break is solely a forecast-period phenomenon, or occurs in-sample, respectively), and we consider combinations of $\delta = \{3, 6, \ldots, 30\}$ and $\lambda = \{0.1, 0.2, \ldots, 1\}$. In the text, we use:

$$\mathsf{E}[\varepsilon_t z_t] = \lambda\mathsf{E}[\varepsilon^2_t] = \lambda,$$

and:

$$\begin{aligned} \mathsf{E}[z^2_t] &= (1 + \lambda^2) \qquad \text{for} \quad t < \tau \\ \mathsf{E}[z^2_t] &= (1 + \lambda^2 + \delta^2) \qquad \text{for} \quad t \geq \tau, \end{aligned}$$

so:

$$\begin{aligned} \mathsf{E}[\mathbf{z}'\mathbf{z}] &= (\tau - 1)(1 + \lambda^2) + (T - \tau + 1)(1 + \lambda^2 + \delta^2) \\ &= T(1 + \lambda^2 + \pi\delta^2), \end{aligned}$$

where π is the proportion of post-break observations, and hence:

$$(\mathsf{E}[\mathbf{z}'\mathbf{z}])^{-1}\mathsf{E}[\varepsilon'\mathbf{z}] = (1 + \lambda^2 + \pi\delta^2)^{-1}\lambda = \phi,$$

which defines ϕ. Thus, we can re-parametrize the estimation equation as:

$$y_t = \mu + \phi z_t + v_t,$$

where $\mathsf{E}[\mathbf{v}'\mathbf{z}] = 0$. Notice that $\phi = \lambda/(1 + \lambda^2)$ for a post-sample break, and $\phi = \lambda/(1 + \lambda^2 + 2\delta^2/T)$ for a break at $\tau = T - 1$. Then:

$$\frac{T(\widehat{\phi} - \phi)^2(T^{-1}\mathbf{z}'\mathbf{z})}{\widehat{\sigma}_v^2} \sim \mathsf{t}^2,$$

and so, under H_0, the square of the t-test on $\phi = 0$ yields:

$$\mathsf{t}^2_{\phi=0} = \frac{T\widehat{\phi}^2(T^{-1}\mathbf{z}'\mathbf{z})}{\widehat{\sigma}_v^2} \underset{app}{\sim} \chi^2(1, \psi^2),$$

where:

$$\psi^2 = T\phi^2(T^{-1}\mathsf{E}[\mathbf{z}'\mathbf{z}]) = T\lambda^2(1 + \lambda^2 + \pi\delta^2)^{-1}.$$

Again using $\chi^2(1, \psi^2) \simeq h\chi^2(m)$, the power of the test under H_T is given by:

$$\mathsf{P}(\mathsf{t}^2_{\phi=0} > c_\alpha \mid \mathsf{H}_T) = \mathsf{P}(\chi^2(1, \psi^2) > c_\alpha) \approx \mathsf{P}(\chi^2(m, 0) > h^{-1}c_\alpha),$$

where h and m are as defined above.