ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

A TALE OF 3 CITIES: MODEL SELECTION IN OVER-, EXACT, AND UNDER-SPECIFIED EQUATIONS

Jennifer L. Castle and David F. Hendry

Number 523
January 2011

Manor Road Building, Oxford OX1 3UQ

A Tale of 3 Cities: Model Selection in Over-, Exact, and Under-specified Equations

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

December 20, 2010

Abstract

Model selection from a general unrestricted model (GUM) can potentially confront three very different environments: over-, exact, and under-specification of the data generation process (DGP). In the first, and most-studied setting, the DGP is nested in the GUM, and the main role of general-to-specific (*Gets*) selection is to eliminate the irrelevant variables while retaining the relevant. In an exact specification, the theory formulation is precisely correct and can always be retained by 'forcing' during selection, but is nevertheless embedded in a broader model where possible omissions, breaks, non-linearity, or data contamination are checked. The most realistic case is where some aspects of the relevant DGP are correctly included, but some are omitted, leading to under-specification. We review the analysis of model selection procedures which allow for many relevant effects, but inadvertently omit others, yet irrelevant variables are also included in the GUM, and exploit the ability of automatic procedures to handle more variables than observations, and consequentially tackle perfect collinearity. Considering all of the possibilities–where it is not known which one obtains in practice–reveals that model selection can excel relative to just fitting a prior specification, yet has very low costs when an exact specification is correctly postulated initially.

JEL classifications: C51, C22.
Keywords: Model selection; Congruence; Mis-specification; Impulse-indicator saturation; *Autometrics*

Preface by David F. Hendry

It is a great pleasure to contribute a chapter on econometric modelling to a *Festschrift* in honour of Professor Lord Meghnad Desai. Meg was one of my mentors at LSE in 1966 when I first became an MSc student, and was closer in spirit to the students than the faculty, although he was already a lecturer. We interacted most in the Quantitative Economics Seminar run by Denis Sargan and Bill Phillips whose sometimes arcane debates Meg helped translate into operational terms. We both lived in Islington or nearby for most of the 1970s, then a lively area of North London yet relatively close to the School. We also became close companions in the LSE faculty Cricket team, and discussed econometrics on the train journeys to and from the ground at Berrylands, as well as statistics, philosophy of science and economic history with other team members. This was consistent with Meg's eclectic interests, spanning all aspects

[*]This research was supported in part by grants from the Open Society Institute and the Oxford Martin School. Contact details: jennifer.castle@magd.ox.ac.uk and david.hendry@nuffield.ox.ac.uk.

from econometric modelling, empirical analyses, macroeconomics and money, Marxian economics and the history of economic thought, later including globalisation and global governance, all well reflected in the contents of this volume. In an era when specialization has been a dominant force, his many and diverse contributions are a welcome beacon of genuine multi-disciplinarity, and a leading indicator of a recent recognition of the benefits of drawing on a range of skills and knowledge.

When I wrote *Autoreg*, the computer system for econometric modelling which included the precursor to *PcGive*, he was an avid and enthusiastic user, fostering its development, albeit describing it as an engine for destroying economic hypotheses. For example, Desai and Weber (1988) explicitly adopts both a general-to-specific (*Gets*) strategy and rigorous testing of the selected models for mis-specification and predictive failure. Although we worked on similar applied topics, we somehow never managed to be coauthors, perhaps reflecting our substitutability in empirical modelling rather than Meg's complementarity to almost all his colleagues. The methodology of empirical econometric modelling was inchoate in the early days, mainly fitting economic-theory derived specifications to time-series data and puzzling over the many test rejections such models tended to accrue. But group discussions, visitors and many seminars and workshops helped clarify the key issues, leading to the advances recorded in Mizon (1995) and Hendry (2003). In these, John Denis Sargan played a pivot role, and Meg's editing of Sargan (1988) has ensured some of his main contributions have been recorded (also see Maasoumi, 1988, for reprints of many of Sargan's papers). Our contribution here is to describe developments since the late 1990s: Campos, Ericsson and Hendry (2005) provide a comprehensive review prior to then, together with reprints of many of the most important papers on model selection and econometric modelling.

1 Introduction

Model selection from a general unrestricted model (GUM) can potentially confront three very different environments, where the GUM may be an over-, exact, or under-specification of the data generation process (DGP). In the first, and most studied setting, the DGP is nested in the GUM, and the main role of selection is to eliminate the irrelevant variables while retaining the relevant. In an exact specification, the theory formulation is precisely correct, but is embedded in a broader model to check for possible omitted variables, non-linearities, breaks or data contamination. The most realistic case is where some aspects of the relevant DGP are correctly included, some irrelevant variables are also included in the GUM, but some relevant variables are omitted, leading to both over- and under-specification. We review the analysis of model selection procedures which allow for many relevant effects as well as irrelevant variables being included in the GUM, and exploit the ability of such procedures to handle perfect collinearity and more candidate variables, N, than observations, T. Reviewing all of the possibilities, where it is not known in advance which one obtains, reveals that model selection can excel relative to just fitting a prior specification, yet has very low costs when an exact specification was indeed correctly postulated initially.

In economics, it is essentially impossible to specify any model that nests the DGP: the high dimensionality, non-stationarity, and unknown non-linearity of economies entail large, complicated and evolving DGPs. Rather, empirical investigations consider small subsets of variables, $\{\mathbf{x}_t\}$ say, suggested by theoretical analyses, which represent reductions of the DGP. For every choice of $\mathbf{x}_t$, there exists a local DGP, denoted LDGP, which is the joint density over the available sample, $\mathsf{D}_{\mathsf{x}}(\mathbf{x}_1 \dots \mathbf{x}_T | \boldsymbol{\theta}_T^1)$, where $\boldsymbol{\theta}_T^1 (= \theta_1, \dots, \theta_T)$ is its parametrization: see e.g., Hendry (2009) for a recent discussion. Such LDGPs can be close to, or far from, the process that actually generated $\{\mathbf{x}_t\}$ depending on the reductions needed to map from the DGP to the resulting LDGP. Good theoretical analyses hopefully guide empirical studies towards LDGPs that are useful for their intended purposes, be those modelling data to understand its properties, testing theories, forecasting, or policy advice. Thus, the choice of which set of variables to analyze is fundamental to the success of a study. Unfortunately, there cannot be generic advice on how

to achieve a good initial formulation, as that depends on the unknown DGP, and hence on the unknown reductions implicit in postulating the LDGP through the choice of $\{\mathbf{x}_t\}$.

Even given a good choice of the set $\{\mathbf{x}_t\}$, there remains the key issue of modelling $\mathsf{D}_{\mathsf{x}}(\mathbf{x}_1 \dots \mathbf{x}_T|\boldsymbol{\theta}^1_T)$, since however excellent an economic theory may be, many aspects must be data based. First, a functional form may be suggested by theory, but usually only over a wide class (e.g., monotonically non-decreasing; or embodying relative risk aversion; or not linear, etc.). Secondly, the time period of decisions can rarely be specified theoretically: a one-period lag may be a minute, day, week, month or year, and so on, and whatever it is, need not match the available frequency of observatons (e.g., quarterly). Thirdly, the theory usually requires–often unstated–*ceteris paribus* conditions on effects not included in the analysis: but the wide-sense non-stationary nature of economic data make such conditions vacuous in practice. This is especially true of the neglect of special events that cause shifts in DGPs, which theories perforce ignore. Further, macroeconomic theories rarely address the heterogeneity of behaviour across agents, although varying endowment distributions can make aggregate parametrizations non-constant. Next, assumptions about the exogeneity of some of the 'givens' also cannot be based on prior reasoning alone. Finally, though this list is illustrative rather than exhaustive, the data may be inaccurately measured or even contaminated over sub-periods. Thus, even if a brilliant theory delivered a 'correct' specification of the LDGP–thereby conflating the DGP and LDGP for $\{\mathbf{x}_t\}$–that is only the start: a major modelling exercise inevitably remains in jointly addressing the main issues in empirical model specification of the complete set of determining variables, their dynamics, functional forms, and parameter constancies.

To highlight recent progress, we will consider three cases. First, an *exact specification* defined by a theory-based joint density $\mathsf{D}_{\mathsf{x}}(\mathbf{x}_1 \dots \mathbf{x}_T|\boldsymbol{\theta}^1_T)$ that is indeed the DGP, where the specified model correctly represents that joint density. We outline an approach such that the theory variables are always retained, despite commencing from a much larger GUM within which it is nested, yet the theory specification is not imposed. This is designed to check the validity and completeness of the theory. Secondly, we will consider a setting where an investigator correctly includes all the relevant variables in $\{\mathbf{x}_t\}$, and also many irrelevant variables, not knowing which elements are relevant and which irrelevant. Thus, the GUM is over-specified, but still nests the DGP. Since which variables are relevant empirically is unknown, the theory variables cannot be forced to be retained in this setting, differentiating it importantly from the first. Finally, we consider the case where only some of the determinants of the DGP are included in the LDGP. Thus, other substantive effects are inadvertently omitted, leading to an under-specified model which does not nest the DGP, but also contains variables that would be irrelevant were the DGP correctly nested in the GUM. This seems the most likely scenario empirically, especially as many outliers, breaks and data mistakes will not be known in advance.

Our objective is to examine the role of automatic model selection in each setting, and contrast its performance with that of simply estimating a pre-specified theory model. We will show that:
(a) when the theory model is the DGP and is forced to be retained within a much larger initial model, which could even have more candidate variables that observations, **selection has no effect on the estimated parameter distributions**, so these are the same as directly estimating the correct and complete theory model;
(b) when the GUM is an under-specification of the DGP, selection can deliver estimates with smaller mean-square errors (MSEs) around their DGP values compared to just estimating a theory-model that is an under-specification of the DGP;
(c) when the GUM nests the DGP, but it is not known which variables are relevant and which irrelevant, the costs of selection are small.
Thus, in all three settings, spanning the range of possibilities in empirical research, selection either dominates, or is equivalent to, estimation of a theory-based specification.

Such a finding runs counter to widespread folklore about model selection, which is usually deemed at best to be a necessary evil and at worst, a pernicious practice that distorts parameter estimates. Criti-

cisms inclue pre-test bias (see e.g., Judge and Bock, 1978), over-fitting by data mining (see e.g., Lovell, 1983), repeated testing that undermines the validity of inferences (see e.g., Leamer, 1974, 1983), with results that are dependent on the path searched (see e.g., Pagan, 1987), and consitute 'measurement without theory' (see e.g., Koopmans, 1947), so have a high probability of delivering garbage. Hendry (2000) discusses the origins of these beliefs, and shows that they lack substance as generic claims. Nevertheless, some model selection algorithms do have such properties: indeed, the most common empirical approach of fitting many models (often covertly) and picking the 'best' one has all these problems in spadefuls, compounded by it being impossible to assess the true uncertainty about the reported choice. Recent developments of automatic model selection algorithms based on the 'LSE methodology' of general-to-specific (*Gets*) modelling, after testing for a congruent initial GUM, ensure that all are avoided: bias correction after selection can be based on the known properties of the selection procedure (see e.g., Hendry and Krolzig, 2005); the selection provides a near unbiased estimate of fit measured by the equation standard error, so there is no over-fitting (see Hendry and Krolzig, 2005); repeated testing difficulties are avoided by only selecting variables, and not models as in some approaches, setting the significance level to control retention of chance significant irrelevant variables at the desired rate (see e.g., Castle, Doornik and Hendry, 2010); path dependence is avoided by exploring all feasible simplifications (see e.g., Hoover and Perez, 1999, and Doornik, 2009); theory models are embedded in the selection to be retained if they are valid (see e.g., Hendry and Johansen, 2010); so selection will not deliver garbage, and yet will retain the relevant variables with almost the same probabilities as if the DGP had been the initial specification (see e.g., Hendry and Krolzig, 2005, and Castle *et al.*, 2010).

The reason we refer to the above eminent authorities' claims as folklore is not just that massive improvements in the theory and practice of automatic model selection have rendered their analyses otiose (but unfortunately not forgotten), they all implicitly presume that the initial theory model is correct, complete and immutable, which is totally unrealistic. Once it is admitted that theory models are at best incomplete, rough and evolving guides to some of the dependencies in economies, and many features of reality are not covered by any theories, it becomes obvious that selection is inevitable. Consequently, it is essential to analyze the many extant approaches, from imposing theory models on data, through covertly running 'hundreds of regressions' and only reporting a few that the investigator 'liked', to using other selection devices (such as AIC from Akaike, 1973, or SIC from Schwarz, 1978, or the LASSO from Tibshirani, 1996). Without a structured and controlled approach based on congruent and encompassing models, inferences are hazardous. Indeed approaches that do not commence from a formulation that nests the DGP (or even the LDGP) are bound to end with an incorrect specification; those that use expanding (rather than contracting) searches can miss key combinations of variables; those that do not search comprehensively for breaks will often conclude with non-constant relationships; those that do not specify congruent encompassing models cannot rely on their inferences; and those without a theory analysis cannot interpret their findings. Thus, all the ingredients of our approach seem essential.

Our review focuses on the linear regression context, although most of the results can be generalized to non-linear equations (see Castle and Hendry, 2010a), or to systems (sketched in Hendry and Krolzig, 2005). We explicitly confront the problems posed by omitted variables interacting with structural breaks, specifically location shifts where previous unconditional means change at some points in the sample, as in Castle, Doornik and Hendry (2009). The resulting procedures invariably lead to more candidate variables, N, than observations, T, so that issue is also addressed, based on Hendry, Johansen and Santos (2008) and Johansen and Nielsen (2009) as implemented in *Autometrics* (see Doornik, 2009, and Hendry and Doornik, 2009). Such an approach seeks to locate the DGP for the chosen set of variables, building on Hoover and Perez (1999) and Hendry and Krolzig (2005): its properties are documented in Castle *et al.* (2010) and Castle and Hendry (2010a) *inter alia*.

The various cases are illustrated using an extension of the original *PcGive* artificial data set of 159 observations (mimicking quarterly data, 1953(3)–1992(3): see e.g., Doornik and Hendry, 1994), de-

signed in 1984 to simulate the impact of the 1970's Oil Crisis on a 'DHSY' consumption function (as in Davidson, Hendry, Srba and Yeo, 1978: for a recent update, see Hendry, 2010). The four variables involved are consumers' expenditure c, income i, inflation Δp and aggregate output q, where the Oil Crisis induced a sharp unanticipated rise in inflation and a concomitant fall in output following a location shift in 1973(3). The consumption function in fact related c_t, to i_t, Δp_t, c_{t-1}, and i_{t-1}. The database has since been extended by Jurgen Doornik adding 20 irrelevant variables, denoted $z_{0,t}, \ldots, z_{19,t}$, the lags of which are also irrelevant, enabling all three settings to be illustrated. Figure 1 shows time-series plots of the DGP variables in the four panels labelled *a*, *b*, *c*, *d*.

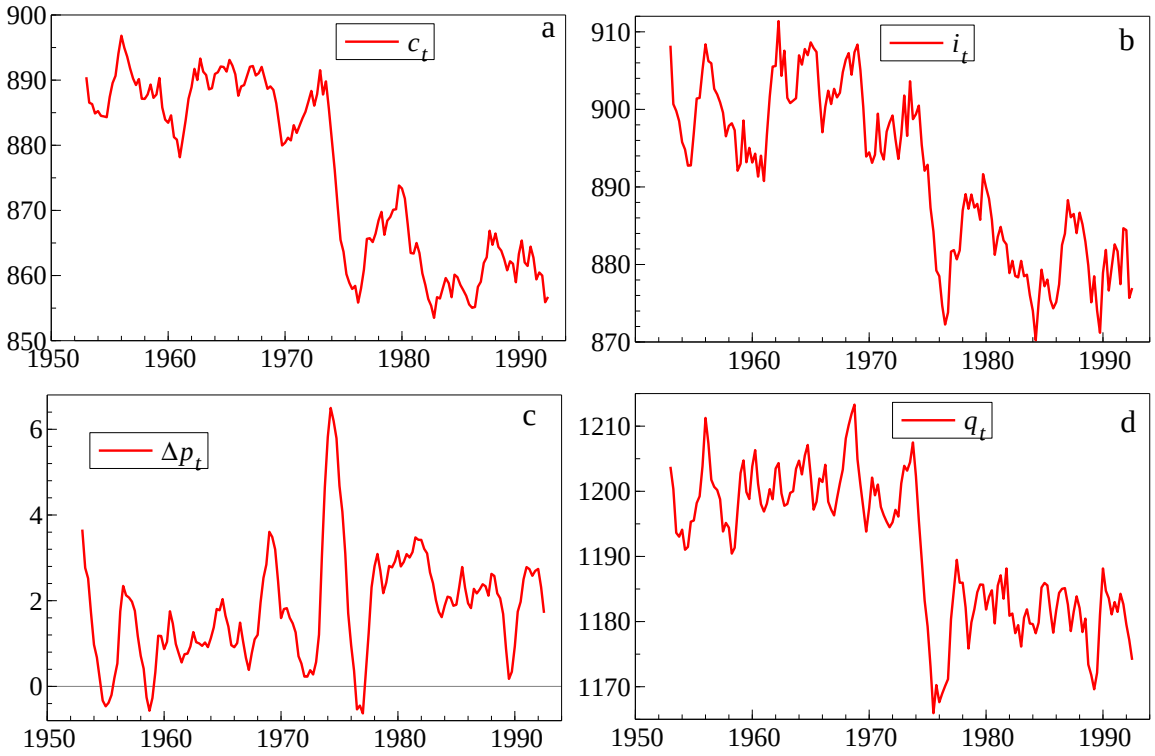

Figure 1: Four artificial data series

The structure of the chapter is as follows. Section 2 considers the case where the model is an over-specification of the DGP, allowing for some substantively relevant variables, as well as many irrelevant effects. Section 3 draws on Hendry and Johansen (2010) who show that when the theory model is precisely correct and is 'forced', selection leaves the distribution of the estimated parameters of interest unchanged relative to simply fitting the DGP. Section 4 draws on Castle and Hendry (2010c) to highlight the manifest advantages of selection from a GUM relative to fitting an incorrect equation, even though the former omits the same subset of relevant variables. Section 5 concludes.

2 Over-specification

We consider two canonical cases. First, in subsection 2.1, we discuss the null model where all the variables under consideration are irrelevant. This serves to establish that even in such an extreme case, which nevertheless satisfies the assumption that the GUM nests the DGP, when setting significance levels appropriate to the problem under analysis there is only a small chance of retaining several irrelevant variables whose coefficient estimates are adventitiously significant. Many of the issues about model

selection questioned in the literature, as noted above, can be resolved in this setting. For example, it is obvious that selecting a model by goodness of fit will not recover the DGP, and that tighter significance levels raise the probability of locating the correct specification of the null model.

In the second case, discussed in subsection 2.2, there are some relevant and many irrelevant variables, which extends §2.1 by including some non-zero parameters in the DGP. Now the key difficulty, so to speak, is sorting the wheat from the chaff. All estimated coefficients have sampling distributions, so some relevant variables (those with non-zero parameters) may be insignificant by chance, and some irrelevant significant by chance. Now there is a trade off between ensuring retention of relevant variables and elimination of irrelevant as the significance level changes. To clarify concepts, we define the *gauge* as the average retention frequency of irrelevant variables, and the *potency* as the average retention frequency of relevant variables. While close to size and power respectively, the concepts differ importantly as we also require models to be congruent (see section 2.1.2) so variables can be retained when they are insignificant to offset what might otherwise be a failure on a mis-specification test. False retention of insignificant irrelevant variables can occur under the null of no mis-mispecification: see Castle *et al.* (2010).

The aim of selection is to choose a final specification that is as close to the DGP commencing from the GUM as would be found commencing from the DGP itself, using the same decision rules. When that can be achieved, the costs of search (selection) are clearly small. However, the DGP itself may be retained rarely even when commencing from it if some relevant variables' estimated parameters would have small t-statistics in the available sample. The costs of inference–not keeping relevant effects–apply even when the DGP specification is correctly postulated, but, absent omniscience, the model is not known to be the DGP, so inference must be conducted to determine what variables are significant. Part of the confusion in earlier analyses of model selection was failing to draw this crucial distinction between the costs of inference—which are inevitable and unavoidable in any non-exact science—and the costs of search, which are additional due to commencing from a GUM which nests, but is larger than, the DGP. Thus, a failure by a search algorithm to locate the DGP may simply reflect that the DGP would not be retained even if it were the initial model.

2.1 The null model

The first canonical case is one in which all N variables are irrelevant, so potential 'over-fitting' is the main problem. Consider a constant-parameter linear regression with $N << T$ mutually orthogonal but irrelevant regressors for $t = 1, \ldots, T$:

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \text{ where } \epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2] \tag{1}$$

where $T^{-1} \sum_{t=1}^{T} z_{i,t} z_{j,t} = \lambda_i \delta_{i,j} \ \forall i, j$, where $\delta_{i,j} = 1$ if $i = j$ and is zero otherwise, with $\{\epsilon_t\}$ independent of all $\{z_{i,t}\}$, when $T >> N$. In the GUM given by (1), all aspects of its specification are correct and known to be correct, except it is not known that all regressors have $\beta_i = 0$. Full-sample least-squares estimation of the GUM is feasible here as $N << T$, and yields $(\widehat{\beta}_1 \ldots \widehat{\beta}_N)$, all of which are unbiased estimators (of zero), and from orthogonality and normality:

$$\widehat{\beta}_i \sim \mathsf{N}\left[0, \sigma_\epsilon^2 \left(T^{-1} \sum_{t=1}^{T} z_{i,t}^2\right)^{-1}\right] \tag{2}$$

Also, the squared residual standard deviation $\widehat{\sigma}_\epsilon^2$ provides an unbiased estimate of the squared equation standard error σ_ϵ^2, a useful baseline to monitor for possible over-fitting. Then:

$$\mathsf{t}_{\widehat{\beta}_i} = \frac{\widehat{\beta}_i}{\widehat{\sigma}_{\widehat{\beta}_i}} \text{ where } \widehat{\sigma}_{\widehat{\beta}_i} = \widehat{\sigma}_\epsilon \left(T^{-1} \sum_{t=1}^{T} z_{i,t}^2 \right)^{-1/2}.$$

When t-testing for the significance of each regressor at significance level α, corresponding to a critical value c_α, a decision to retain the ith regressor is made if $|\mathsf{t}_{\widehat{\beta}_i}| > c_\alpha$. The complete set of well-known probabilities of rejections and non-rejections under the null for (1) are shown in Table 3.

event	probability	number retained
$\mathsf{P}(\lvert\mathsf{t}_i\rvert < c_\alpha,\ \forall i = 1, \dots N)$	$(1-\alpha)^N$	0
$\mathsf{P}(\lvert\mathsf{t}_i\rvert \geq c_\alpha \mid \lvert\mathsf{t}_j\rvert < c_\alpha,\ \forall j \neq i)$	$N\alpha\,(1-\alpha)^{N-1}$	1
$\vdots$	$\vdots$	$\vdots$
$\mathsf{P}(\lvert\mathsf{t}_i\rvert < c_\alpha \mid \lvert\mathsf{t}_j\rvert \geq c_\alpha,\ \forall i \neq j)$	$(N-1)\,N\alpha^{(N-1)}\,(1-\alpha)$	$N-1$
$\mathsf{P}(\lvert\mathsf{t}_i\rvert \geq c_\alpha,\ \forall i = 1, \dots N)$	α^N	N

(3)

Then, the average number of null variables retained from Table 3 is given by the binomial sum:

$$m = \sum_{i=0}^{N} i \frac{N!}{i!\,(N-i)!} \alpha^i\,(1-\alpha)^{N-i} = N\alpha. \tag{4}$$

The key determinants, when the tests are indeed independent and distributed as t, are N and α, so in principle any value of m is possible. However, sensible decision rules must link these two decision variables, and one simple rule is $\alpha = k/N$ for a small integer $k = 1, 2, 3$ say. Thus, when $N = 100$, which is 'large' relative to most time-series models, then for $k = 1$, one would set $\alpha = 0.01$ which yields $m = 1$. Consequently, 99 out of the 100 irrelevant regressors would be eliminated on average, and just one retained. A great deal is learned about what does not matter, effecting a massive reduction from an initial 100 candidate regressors to one or a few 'spuriously significant' variables that are adventitiously retained, although the conventional measure of the 'size' of the procedure is:

$$1 - (1-\alpha)^N = 1 - (1-0.01)^{100} \simeq 0.63,$$

suggesting such a procedure is not useful. The conventional significance level of 5% is fine for a single, or 1-off, test, but is not helpful once multiple tests are required. The cost of shifting to a tighter significance level like 1% is that c_α increases, making it harder to retain relevant variables when they are present in a large set of irrelevance.

2.1.1 Illustrating regressions with no relevant variables

Using the extended *PcGive* artificial data set, we formulated a GUM for one of the irrelevant variables, $z_{0,t}$, dependent on its first 2 lags, a constant and on current and 2 lags of $z_{1,t}, \dots, z_{19,t}$, which made 60 variables for the remaining 157 observations. Selecting by Autometrics at $\alpha = 0.01$, so $m = 0.6$, duly delivered the null model, which matched the null DGP. Thus, over-fitting did not occur despite $N = 100$ irrelevant regressors in the GUM.

However, before considering cases with relevant variables in subsection 2.2, there are seven important considerations: congruence, normality, bias corrections, different significance levels for different groups of decisions, 'forcing' retention of variables, goodness of fit, and selecting variables, not models. We take these in turn.

2.1.2 Congruence

First, inferences based on conventional statistics and critical values are valid only if the GUM is congruent, which here requires a constant-parameter linear regression with no omitted relevant variables, accurate data, and errors ϵ_t that are distributed as $\mathsf{IN}[0, \sigma_\epsilon^2]$ independently of the regressors. Any violations of congruence can induce false selections, which cannot necessarily be rectified by heteroskedastic-autocorrelation consistent estimators of parameter standard errors (HACSEs, as in e.g., White, 1980, and Andrews, 1991), both because these do not reflect the selection decisions made *en route* to the chosen specification, and because they rely on the untested *non sequitur* that the problem manifest in the residuals is due to the assumed solution for the errors. For example, residual autocorrelation could be due to an unmodelled break, and HACSEs will not 'correct' that. Thus, for an estimable GUM, the first step is to test for congruence, which if it is accepted, then there exists a simplification path to a congruent final model (which may be the GUM if there are no valid reductions). In *Autometrics*, five such tests are standard, namely autocorrelation, heteroskedasticity, non-normality, non-linearity, non-constancy, close to the set used in *Autoreg* (see Hendry and Srba, 1980, and e.g., Desai and Weber, 1988): these are delineated in §2.4.1. To control the overall null rejection frequency, determined by any one mis-specification test rejecting, we set $\delta = 0.01$ for each test, yielding $1 - (1 - 0.01)^5$ from Table 3, or about 5% overall.

2.1.3 Normality

Secondly, although normality is an aspect of congruence, it also plays a separate role. When the distributions of tests are close to the normal, with what Denis Sargan called 'thin tails' in Sargan (2001), critical values increase slowly with decreases in α in the tails. Table 5 records these changes, rounded.

$$\begin{array}{ccccccc} \alpha & 0.05 & 0.01 & 0.005 & 0.0025 & 0.001 \\ c_\alpha & 2.0 & 2.6 & 2.8 & 3.0 & 3.3 \end{array} \tag{5}$$

For example, even at 0.25%, $c_{0.0025} = 3.0$, just requiring a t-value of 3.0, rather than the famous 2.0, to reject the null. Yet, if 0.25% is applied to (4), $m = 100 \times 0.0025 = 0.25$ so only one irrelevant variable out of 100 would be retained once every four trials, with none retained on average on the remaining three trials. While ts of 2 are conventional, moving to values around 3 (at $\alpha = 0.0025$) would entail almost never retaining irrelevant variables even after starting with $N = 100$ candidates. Of course, such an analysis places a premium on having approximate normality, and we address that in subsection 2.4. Indeed, the next consideration also requires approximate normality, making it doubly important.

2.1.4 Bias corrections

Thirdly, bias corrections were developed in Hendry and Krolzig (2005) and analyzed by Castle *et al.* (2010). As only 'significant' estimates are retained, this corresponds to a decision rule where:

$$\begin{array}{ll} \widetilde{\beta}_i = \widehat{\beta}_i & |\mathsf{t}_{\widehat{\beta}_i}| > c_\alpha \\ \widetilde{\beta}_i = 0 & |\mathsf{t}_{\widehat{\beta}_i}| \leq c_\alpha \end{array} \tag{6}$$

when the final retained estimates are denoted $\widetilde{\beta}_i$. The distribution of $\widetilde{\beta}_i$ for only those regressors that are retained by $|\mathsf{t}_{\widehat{\beta}_i}| > c_\alpha$ is called the conditional distribution; the complete, or unconditional, distribution also includes all the zero values assigned by (6). Selection by $|\mathsf{t}_{\widehat{\beta}_i}| > c_\alpha$ induces a doubly-truncated t distribution where the central part, namely values between $\pm c_\alpha$, is discarded, and only the tails are retained. It is convenient to approximate this by a doubly-truncated normal distribution where all the formulae are well known (see e.g., Johnson, Kotz and Balakrishnan, 1994).

When $\beta_i = 0$, the resulting estimates are unbiased, as $\mathsf{E}[\widehat{\beta}_i] = 0$, but for $\beta_i \neq 0$, the retained $\widetilde{\beta}_i$ need to be corrected for selection, since from (6) only significant estimates are retained, so:

$$\mathsf{E}\left[\widetilde{\beta}_i \mid |\mathsf{t}_{\widehat{\beta}_i}| > c_\alpha\right] \neq \beta_i. \tag{7}$$

However, the truncation point c_α is known, so one can correct $\widetilde{\beta}_i$ after selection, denoted $\widetilde{\widetilde{\beta}}_i$, such that:

$$\mathsf{E}\left[\widetilde{\widetilde{\beta}}_i \mid |\mathsf{t}_{\widehat{\beta}_i}| > c_\alpha\right] \simeq \beta_i. \tag{8}$$

Bias correction as in (8) leads to some increase in the mean-square errors (MSEs) of the estimated coefficients of relevant variables, namely where $\beta_i \neq 0$, and exacerbates the downward bias in the unconditional estimates due to setting some $\widetilde{\beta}_i = 0$. There is no impact on the bias of estimated parameters of irrelevant variables, as their $\beta_i = 0$, but there is a marked *decrease* in their MSEs—essentially a 'free lunch'—since most bias correction occurs for $|\mathsf{t}_{\widehat{\beta}_i}|$ near c_α, and that is the most likely outcome under the null, driving the resulting $\widetilde{\widetilde{\beta}}_i$ near to zero. This result applies at loose α, so even if many irrelevant variables were retained, their estimated parameters would be small on average after bias correction, suggesting a possible approach when selecting forecasting models, providing estimates of the equation standard error were also bias corrected.

2.1.5 Significance levels differing by decisions

The fourth consideration is to use different significance levels for different groups of decisions, which will lead into the fifth, namely 'forcing' the retention of some variables. If a subset of variables is deemed substantively important, then it could be selected at a loose significance level, say 25%, whereas other variables that are thought to be less important, but still may be relevant, are selected at a much tighter level such as 0.5%, so need strong evidence for their retention. In particular, selection and mis-specification testing are often conducted at different significance levels, as are linear and non-linear reactions.

2.1.6 'Forcing' retention of variables

The fifth consideration is 'forcing', which relates to always retaining some variables, while others are subject to selection. In effect, the first set uses a 100% significance level: that is what retaining the GUM entails, of course. 'Forcing' allows a theory model to be embedded in a GUM where only the additional variables are selected. While this guarantees that the theory-based variables are retained, they may not be significant nor have their anticipated signs, and other variables may transpire to be the important determinants of the dependent variable. It is often advisable to force retention of the intercept, as it can be insignificant early in a simplification yet highly significant in the final model. We return to the distributional properties of parameter estimates for forced variables in section 3.

2.1.7 Goodness of fit

The crucial aspect about goodness of fit is that it was **not** considered by the decision rule. For a known model where all regressors are relevant, choosing parameters by goodness of fit (or maximum likelihood) has important justifications. For selecting which variables to include, goodness of fit has no justification, and is most unlikely to choose a model that is a close approximation to the DGP. Implicitly, the selection of α affects the resulting goodness of fit, as measured by $\widetilde{\sigma}_\epsilon$ in the chosen representation, but that information is not directly used to select. Nevertheless, while such a result only holds in the present null model setting, the probability of selecting the DGP *rises* here as the significance level becomes tighter–the worst fitting model, namely one with no variables, matches the DGP.

2.1.8 Selecting variables not models

The final issue is that the calculations in Table 3 are based on selecting variables commencing from a congruent GUM, not selecting models. There are N variables, but 2^N models, one of which must certainly maximize goodness of fit, or 'penalized' versions thereof (like AIC). There are so many sub-models within the set of 2^N that many will be non-congruent, and most will not be useful approximations to the DGP. For the above example of $N = 100$, $2^{100} \simeq 10^{30}$, some of which will surely be 'garbage'. Indeed, it is difficult to imagine usable values of α such that the retention of spurious models from 2^N could be controlled unless N was very small. To understand the difference between selecting models and selecting variables, consider a procedure that not only tested the relevance of each regressor in (1), but also tested all possible pairs, all triples, etc., right up to all combinations of $N - 1$. That would augment Table 3 with a vast array of probabilities of rejecting on F-tests for every possible combination. There are bound to be many combinations 'significant' by chance.

We conclude from the above analysis that eliminating irrelevant variables is not a fundamental difficulty for variable selection–even when there is a large number of potential candidate regressors. Despite its many special characteristics, and the fact that the complete null model is not an empirically relevant case in economics, it is important to know that when significance levels are set sensibly, over-fitting holds few terrors. We now need to explore the impact of selecting when there are relevant variables in (1), then extend the analysis to selecting in relevant economic settings with collinearity, breaks, non-normality, dynamics, omitted variables, non-linearity and probably measurement errors, perhaps leading to $N > T$.

2.2 Regressions with relevant variables

We next consider (1) when some of the $\beta_i \neq 0$, where some $\beta_i = 0$ as before, but it is not known for certain which regressors are relevant. In an orthogonal case like (1), Castle *et al.* (2010) show that the selection decision can be made in '1-cut': rank every $|\mathsf{t}_{\widehat{\beta}_i}|$ and retain (discard) those which exceed (are smaller than) c_α. Thus, there is no repeated testing even when $N = 1000$ (say, which is the case they consider). The probability of retaining relevant variables depends on the non-centrality, ψ_i, of their $|\mathsf{t}_{\beta_i}|$, where $\psi_i = 0$ when $\beta_i = 0$. Naturally, this probability falls as c_α increases, and in any case is quite low even for conventional significance levels, as (say) $\mathsf{p}[\mathsf{t}_{\beta_i} > 2|\psi_i = 2] \simeq 0.5$, but rises exponentially as ψ_i increases, so that $\mathsf{p}[\mathsf{t}_{\beta_i} > 3|\psi_i = 6] > 0.99$.

Castle *et al.* (2010) also show that *Autometrics* applied to the same setting as '1-cut' tends to outperform it in MSEs, and they also present supporting Monte Carlo simulation evidence on the performance of *Autometrics* for a range of autoregressive-distributed lag models.

2.3 Perfect collinearity

Consider as a simple example the DGP:

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t \tag{9}$$

where in fact $\beta_1 = -\beta_2$, but that is not known. When the GUM is specified as

$$y_t = \gamma_1 x_{1,t} + \gamma_2 x_{2,t} + \gamma_3 (x_{1,t} - x_{2,t}) + v_t \tag{10}$$

using a comprehensive multi-path search, then one path will delete $(x_{1,t} - x_{2,t})$ and (for sufficiently large test non-centralities) retain $x_{1,t}$ and $x_{2,t}$; a second path will eliminate $x_{2,t}$ and should retain $(x_{1,t} - x_{2,t})$ but also drop $x_{1,t}$ as now insignificant; and similarly for the third path commencing from first dropping $x_{1,t}$. Thus, an advantage of such procedures in dynamic specification searches is that they allow many forms of possible lag response to be included, such as $x_{1,t}$, $x_{1,t-1}$, $\Delta x_{1,t} (= x_{1,t} - x_{1,t-1})$, $(x_{1,t} + x_{1,t-1})$

and so on, where the relevant subset is retained. Campos and Ericsson (1999) discuss the importance of the choice of the initial linear transformations of the regressors in the GUM for the final selection when testing is only for null hypotheses like $\gamma_i = 0$.

The potential cost of doubling the number of variables by perfect collinearity is that the procedure will now retain approximately $2\alpha n$ irrelevant regressors. On the one hand, it could be argued that should not occur because there are still only n separate regressors, hence null retentions will be unchanged. On the other hand, many linear combinations of variables are being included and in a t-test based approach, such combinations could be significant (still under the null) even when the components would not be. For example, $x_{1,t}$ and $x_{1,t-1}$ might have t-test values less than c_α, yet $(x_{1,t} + x_{1,t-1})$ have a significant coefficient and so be retained by chance.

Given the difficult analytic nature of trying to establish which of the two arguments holds, we have undertaken a number of Monte Carlo simulation studies. The simplest is close to the model in (10). The DGP is:

$$y_t = \epsilon_t \text{ where } \epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2] \tag{11}$$

for $t = 1, \ldots, T = 100$ where:

$$\mathbf{x}_t \sim \mathsf{IN}_2[\mathbf{0}, \mathbf{I}] \tag{12}$$

We first consider a case with no collinearity. The GUM has the form in (13) as a baseline to check that the gauge $g \simeq \alpha$ with 5 irrelevant regressors:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t-1} + \beta_4 x_{2,t-1} + \epsilon_t \tag{13}$$

Setting $\alpha = 0.01$, with diagnostic tests for congruence also at 1%, $M = 10000$ replications delivered $g = 0.014$ which is a little 'over-gauged' as anticipated from conducting diagnostic tests. Thus, on average $5g = 5 \times 0.014 = 0.07$ irrelevant variables were retained per replication. Without diagnostic checking, $g = 0.0098$ so the algorithm is calibrated to deliver the correct significance level under the null when there are no diagnostic checks.

If we now include perfectly collinear variables, repeating these simulations, but with the GUM specified as:

$$\begin{aligned} y_t = \gamma_0 + \gamma_1 x_{1,t} + \gamma_2 x_{2,t} + \gamma_3 (x_{1,t} - x_{2,t}) + \gamma_4 x_{1,t-1} + \gamma_5 x_{2,t-1} \\ + \gamma_6 (x_{1,t-1} - x_{2,t-1}) + \gamma_{76} \Delta x_{1,t} + \gamma_8 \Delta x_{2,t} + v_t \end{aligned} \tag{14}$$

where (14) has 9 irrelevant regressors with four collinearities, again at $\alpha = 0.01$, now $M = 10000$ replications delivered $g = 0.0057$ so that $9g = 9 \times 0.0057 = 0.05$ irrelevant variables were retained per replication. This is actually slightly smaller than anticipated, but is consistent with the argument that adding perfectly collinear variables does not increase the retention rate.

Finally we examine the case where $N \gg T$ and there is perfect collinearity. We augment (14) with $\sum_{j=1}^{100} \delta_j u_{j,t}$ where:

$$\mathbf{u}_t \sim \mathsf{IN}_{100}[\mathbf{0}, \mathbf{I}] \tag{15}$$

resulting in 109 regressors for 100 observations. Selection at $\alpha = 0.01$ with $M = 10000$ replications delivered $g = 0.0105$, so $109g = 1.14$ irrelevant variables were retained per replication despite including over 100 irrelevant variables. Undertaking selection at $\alpha = 0.005$ yields 0.44 variables retained per replication, demonstrating that inclusion of many irrelevant variables does not lead to over-fitting.

However, since much of the analysis depends on having approximate normality, we now turn to how that might be obtained, and also address the issues of multiple structural breaks, outliers, and data contamination, leading to a special case where $N > T$.

2.4 Impulse-indicator saturation

One of the simplest cases where more candidate variables than observations occurs is adding an impulse indicator for every observation to the set of possible explanatory variables, a procedure called impulse-indicator saturation and denoted IIS below (see Hendry *et al.*, 2008, and Johansen and Nielsen, 2009). When $N > 0$, using IIS creates $N + T > T$. At first sight, such a setting seems quite problematic, but it is not. As shown by Salkever (1976), the Chow (1960) test includes an indicator variable for every observation in the forecast period, and tests for them being significantly different from zero. Recursive estimation can be interpreted as having an impulse indicator for every observation in the later period, sequentially removing them one by one. Rolling windows put in blocks, first in the future and then in both the future and the past when moving through the sample. Consequently, many existing methods can be interpreted as using indicators for every observation, but in different ways. Indeed, reversing recursive estimation involves implicitly entering more indicators than data points.

In the simplest IIS theory in Hendry *et al.* (2008) and Johansen and Nielsen (2009), indicators are added in two blocks of $T/2$, significant outcomes in each block being recorded then that block omitted while the other is included, again recording significant outcomes, then the two sets of significant indicators are added together in the final specification. Unequal and multiple splits are also analyzed, and *Autometrics* uses a general block algorithm, and searches across many such splits: no matter how many splits are tried, an outlier will only be found to match an indicator if it is there.

Under the null that there are no outliers, IIS will retain αT indicators by chance: thus, for example, when $T = 100$ and $\alpha = 0.01$, one indicator will be significant by chance, in effect reducing the available sample from 100 to 99 as an observation is 'dummied out'. Viewed as a robust estimator, therefore, IIS is 99% efficient. Of course, there is little point in using IIS when the null is true. Rather, the aim of IIS is to check at every observation whether there has been an outlier, same-signed contiguous blocks of outliers which would reveal a location shift or data contamination in a subsample. Castle *et al.* (2009) conduct an extensive Monte Carlo simulation study of IIS under the alternative for many break forms including multiple breaks.

Here, α must be the appropriate significance level for the underlying distribution, the form of which is rarely known. In practice, α is usually chosen for the normal, so the question arises as what happens when IIS is applied to a non-normal distribution. Castle *et al.* (2009) consider a Student-t distribution with 3 degrees of freedom, denoted t_3, and show that impulse indicators capture much of the non-normality in this fat-tailed distribution. Many indicators are retained, of course. Despite the critical value, c_α, used for selection being incorrect, after IIS the resulting distribution is sufficiently near normal that the null retention frequency of other irrelevant variables is close to, but slightly larger than, α. The retention of relevant variables is improved relative to ignoring the fat-tail problem.

Thus, IIS enables normality to be a reasonable assumption for inference and bias correction. Section 2.6 on non-linearity depends on near normality to avoid spurious outcomes. Moreover, the form of analysis showing that IIS can be feasible—despite more indicators plus regressors than observations—applies to cases with $N > T$ due to more candidate regressor variables than observations, as we now discuss.

2.4.1 Illustrating regressions with no relevant variables using IIS

Repeating the exercise in 2.1.1 with IIS, which creates $N + T = 217$, and now selecting at $\alpha = 0.0025$, so $m \simeq 0.5$, again delivered the null model. In both illustrations, there was a probability of just under a half of locating the null DGP, and the second demonstrates the practical implementation of $N > T$ using block searches. The crucial issue, however, is whether the original non-null DGP can be located.

2.4.2 Illustrating regressions with relevant variables using IIS

The *PcGive* artificial DGP for the consumption equation is:

$$c_t = 0.85c_{t-1} + 0.5i_t - 0.35i_{t-1} - \Delta p_t \tag{16}$$

and direct estimation yields:

$$\begin{aligned} c_t = & \underset{(0.022)}{0.84}\ c_{t-1} + \underset{(0.027)}{0.49}\ i_t - \underset{(0.032)}{0.33}\ i_{t-1} - \underset{(0.085)}{0.95}\ \Delta p_t \\ \widehat{\sigma} = 1.10 \quad & \mathsf{F}_{\mathsf{ar}}(5,148) = 1.09 \quad \mathsf{F}_{\mathsf{arch}}(4,149) = 0.78 \\ \chi^2_{\mathsf{nd}}(2) = 0.48 \quad & \mathsf{F}_{\mathsf{het}}(8,148) = 1.05 \quad \mathsf{F}_{\mathsf{reset}}(2,151) = 3.67^* \end{aligned} \tag{17}$$

R^2 is the squared multiple correlation, and $\widehat{\sigma}$ is the residual standard deviation, with coefficient standard errors shown in parentheses. The diagnostic tests are of the form $\mathsf{F}_{\mathsf{j}}(k, T-l)$ which denotes an approximate F-test against the alternative hypothesis j for: k^{th}-order serial correlation (F_{ar}: see Godfrey, 1978), k^{th}-order autoregressive conditional heteroskedasticity ($\mathsf{F}_{\mathsf{arch}}$: see Engle, 1982), heteroskedasticity ($\mathsf{F}_{\mathsf{het}}$: see White, 1980); the RESET test ($\mathsf{F}_{\mathsf{reset}}$: see Ramsey, 1969); and a chi-square test for normality ($\chi^2_{\mathsf{nd}}(2)$: see Doornik and Hansen, 2008).

Including all four DGP variables and the 20 irrelevant $z_{01,t}, \ldots, z_{19,t}$, plus an intercept, with the same lag length of 2 as before and current-dated regressors, then $N = 72$, and $T = 157$ after creating the 2 lags. Applying IIS and setting $\alpha = 0.0025$, then $\alpha(N+T) = 0.5725$ so again the probability of retaining one adventitiously-significant irrelevant variable is just over a half, with a negligible probability of retaining more than one irrelevant variable or indicator. *Autometrics* finds the DGP equation plus one indicator with an estimate of 3.39 (1.08); that the relevant variables were retained follows from their large non-centralities seen in (17). When $\alpha = 0.001$, the DGP equation (17) is found precisely.

2.5 More candidate regressor variables than observations

First consider the case where $N = T$, so that a split-half approach–as with IIS—is feasible. The same logic applies, but now αN irrelevant variables will be retained under the null, each of which costs a degree of freedom, rather than a data point, spreading the 'cost' over the whole sample. For $N > T$, multiple blocks are needed to handle all the searches.

However, such block searches require expanding as well as contracting searches, so are no longer strictly general-to-specific. A key aspect is to use large blocks. Early selection algorithms, such as stepwise, added variables one at a time, so their significance could be masked by not jointly including other variables with the opposite net effect on the regressor (e.g., negatively when the current candidate has a positive effect). As shown in Castle *et al.* (2010), the MSEs of estimated parameters of relevant variables then increase linearly as N increases from $N << T$, through $N = T$ to $N >> T$, so there is no 'jump' in the neighbourhood of $N = T$.

2.5.1 Illustrating regressions with relevant variables when $N > T$

To take a relatively extreme case, we will use a lag length of 20, so including an intercept and current-dated regressors $N = 504$, with $T = 139$ after creating 20 lags, but no IIS. Setting $\alpha = 0.001$, then $\alpha N = 0.504$ and again the probability of retaining one adventitiously-significant irrelevant variable is just over a half. Indeed, here *Autometrics* finds the exact DGP equation starting from almost any over-specified candidate set, with up to 20 lags, and without or with also undertaking IIS (there are no substantial outliers or breaks in the conditional model although there are in the DGP).

Including IIS makes $N + T = 643$, so a large excess of variables over observations has to be confronted, yet *Autometrics* again delivers the DGP equation (17) at $\alpha = 0.001$. The probability of retaining no irrelevant variables was just 0.36, so the outcome that none were retained was slightly 'lucky'. Importantly, such calculations are feasible prior to selection, and tend to be borne out by both simulations and artificial data modelling.

2.6 Models with non-linear variables

Castle and Hendry (2010b) test for non-linearity by forming the principal components $\mathbf{w}_t$, say, of the original n regressors $\mathbf{x}_t$, and use second and third powers and exponentials of the $w_{i,t}$. Since the $\mathbf{w}_t$ are generally linear combinations of all the $\mathbf{x}_t$, their powers include many squares, cubics and up to triple interactions between the $x_{i,t}$. Their approach still applies when N would exceed T from adding up to cubic polynomials in the $\mathbf{x}_t$ (even without IIS), such that $N = n + n(n+1)(2n+4)/6 > T$ although $n << T$, since there would only be $4n$ variables in total. Applying that test to (17) yields $\mathsf{F}(12, 141) = 1.05$, so does not reject. A 'complete' test of all those non-linear functions in the $x_{i,t}$ would have added 48 variables even for $n = 4$.

When the test rejects and apparently entails a non-linear specification, it is essential to also apply IIS to avoid a spurious match between outliers and some of the non-linear terms as discussed in Castle and Hendry (2010a). With $n = 23$ regressors excluding the dependent variable, we generate the demeaned squares and cubics of the corresponding principal components, $w_{i,t}$, and augment the GUM in §2.4.2 with $\sum_{j=0}^{2} \sum_{i=1}^{23} w_{i,t-j}^{k}$ for $k = 1, 2, 3$, resulting in $N = 279$ with a further $T = 157$ indicators from IIS. Although there are many perfectly collinear relationships, undertaking expanding and contracting path searches enables a non-singular representation to be obtained. Applying selection at $\alpha = 0.001$ results in the exact DGP specification being retained, so augmenting the GUM with many additional non-linear terms does not result in over-fitting: we would expect to retain 0.44 of a variable on average under the null of 436 irrelevant variables, regardless of whether the variables are standard regressors, principal components, or non-linear factors.

3 Exact-specification

Hendry and Johansen (2010) show that forcing retention of a correct theory-based set of variables (distinct from imposing their coefficient values) leaves unchanged the distributions of the parameter estimates in a GUM with many irrelevant variables as compared to direct estimation of the theory model. That result also holds for models with endogenous variables when there are adequate instrumental variables to viably estimate the GUM, and even when $N > T$.

While perhaps astonishing at first sight that selection has no impact on estimator distributions in such a procedure, the explanation is that the irrelevant variables can be orthogonalized relative to the correct theory variables, which does not alter the theory parameters, and it is well known that the inclusion or omission of orthogonal variables does not alter estimator distributions. Nevertheless, that result should profoundly alter attitudes towards model selection–it has no effect when the theory model is correct, and as we now show, can be hugely beneficial when the theory is incomplete or incorrect.

3.1 Forcing the correct specification

To demonstrate assume that the correct specification (16) were known and therefore embedded in a GUM with 2 lags of all variables and IIS by forcing c_{t-1}, i_t, i_{t-1} and Δp_t to be retained in selection. As in §2.4.2 there are $N + T = 229$ regressors, 4 of which are forced, and $\alpha = 0.0025$. The resulting selected model is identical to the case where the theory variables are not forced, with one additional indicator

retained. As the theory variables are highly significant, they would have been retained even if selected over and so forcing has no effect here. Likewise, if a subset of the relevant variables had been forced, the selected model would have been the same.

3.2 Forcing an incorrect specification

If we assume that the postulated theory consists of c_{t-1}, q_t, q_{t-1} and Δp_t rather than i_t and i_{t-1}, and hence irrelevant variables are forced in selection, again commencing with $N + T = 229$ regressors for $T = 157$ observations results in:

$$\begin{aligned} c_t = & \underset{(0.022)}{0.84}\ c_{t-1} + \underset{(0.031)}{0.51}\ i_t - \underset{(0.032)}{0.32}\ i_{t-1} - \underset{(0.093)}{0.94}\ \Delta p_t - \underset{(0.029)}{0.01}\ q_t - \underset{(0.032)}{0.01}\ q_{t-1} \\ & \widehat{\sigma} = 1.09\ \ \mathsf{F}_{\mathsf{ar}}(5,146) = 1.02\ \ \mathsf{F}_{\mathsf{arch}}(4,149) = 1.20 \\ \chi^2_{\mathsf{nd}}(2) = 0.75\ \ & \mathsf{F}_{\mathsf{het}}(12,144) = 1.18\ \ \mathsf{F}_{\mathsf{reset}}(2,149) = 1.83 \end{aligned} \tag{18}$$

Both q_t and q_{t-1} are forced to be retained in the final selected model but both are insignificant. Including i_t and i_{t-1} in the regressor set resulted in them being retained through selection, so the final model is a close approximation to the DGP despite forcing irrelevant variables. The one indicator found previously is no longer retained as q_t and q_{t-1} could be picking up the outlier. Forcing theory variables does not guarantee that they will be significant or have the correct sign, but it is relatively costless as long the GUM nests the DGP variables. We now turn to the case where this doesn't hold.

4 Under-specification

If relevant variables are omitted from the GUM any resulting selected model will face the classical omitted variable bias problem. Selection cannot mitigate that, and motivates commencing from sufficiently general models to minimize the chance of excluding potentially relevant variables: the costs of commencing from an under-specified model by far outweigh the costs of commencing from an over-specified model. However, selection can be very beneficial even in under-specified models, particularly when the omitted variables are subject to breaks. Castle and Hendry (2010c) show that augmented general-to-specific model selection strategies using IIS can excel in mitigating most of the adverse effects of breaks in equations due to omitting relevant variables that suffer location shifts. To illustrate their analysis, first consider leaving inflation out of the GUM, while maintaining one lag and not selecting, which leads to:

$$c_t = \underset{(11.4)}{2.5}\ + \underset{(0.03)}{0.99}\ c_{t-1} + \underset{(0.04)}{0.50}\ i_t - \underset{(0.04)}{0.49}\ i_{t-1} \tag{19}$$

$$\begin{aligned} \mathsf{R}^2 = 0.988\ \ \widehat{\sigma} = 1.48\ \ \chi^2_{\mathsf{nd}}(2) = 7.65^{*}\ \ \mathsf{F}_{\mathsf{ar}}(5,149) = 7.82^{**} \\ \mathsf{F}_{\mathsf{arch}}(4,150) = 6.28^{**}\ \ \mathsf{F}_{\mathsf{reset}}(2,152) = 2.98\ \ \mathsf{F}_{\mathsf{het}}(6,151) = 1.09 \end{aligned}$$

Three of the mis-specification tests reject. Figure 2 shows the resulting non-constancy using recursive estimation of (19).

Further, the omission of the breaking variable, Δp_t, leads to the fitted model being essentially in first differences, and suggests the absence of any long run. Implementing that idea:

$$\Delta c_t = \underset{(0.05)}{0.23}\ \Delta c_{t-1} + \underset{(0.03)}{0.50}\ \Delta i_t \tag{20}$$

$$\begin{aligned} \widehat{\sigma} = 1.42\ \ \chi^2_{\mathsf{nd}}(2) = 1.86\ \ \mathsf{F}_{\mathsf{ar}}(5,142) = 3.19^{**} \\ \mathsf{F}_{\mathsf{arch}}(4,141) = 2.25\ \ \mathsf{F}_{\mathsf{reset}}(2,145) = 0.15\ \ \mathsf{F}_{\mathsf{het}}(5,143) = 2.46^{*} \end{aligned}$$

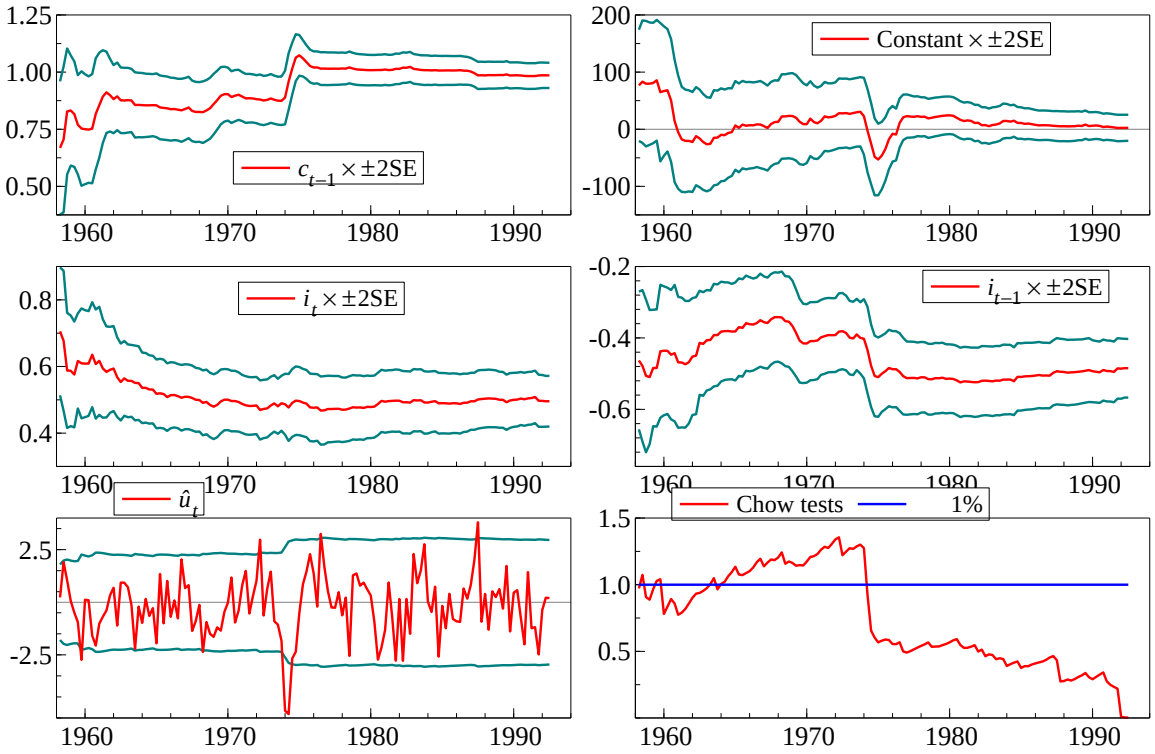

Figure 2: Incorrect artificial-data model specification

However, several mis-specification tests still reject, and the entailed solution in first differences suggests that c only responds 2/3rds to i (although the DGP has a 1–1 response).

4.1 Model selection with IIS

Now consider selection using IIS at $\alpha = 0.001$, commencing from a GUM with 2 lags, which delivers:

$$\begin{aligned} c_t = & \underset{(0.05)}{1.15}\ c_{t-1} - \underset{(0.05)}{0.17}\ c_{t-2} + \underset{(0.03)}{0.51}\ i_t - \underset{(0.03)}{0.49}\ i_{t-1} \\ & - \underset{(1.34)}{4.30}\ 1_{1974(1)} - \underset{(1.35)}{4.44}\ 1_{1974(2)} \end{aligned} \tag{21}$$

$$\begin{aligned} \widehat{\sigma} = 1.32 \quad \chi^2_{\mathsf{nd}}(2) = 0.22 \quad \mathsf{F}_{\mathsf{ar}}(5, 146) = 2.05 \\ \mathsf{F}_{\mathsf{arch}}(4, 149) = 0.91 \quad \mathsf{F}_{\mathsf{reset}}(2, 149) = 0.55 \quad \mathsf{F}_{\mathsf{het}}(14, 140) = 0.85 \end{aligned}$$

Indicators at the 'oil-crisis dates' 1974(1) and 1974(2) and the longer lag of c proxy the omission of Δp_t. That is a key benefit of *Autometrics* over conventional modelling, even when the basic set is substantively incomplete–picking up the break effects of an omitted variable that shifts is very advantageous relative to having a non-constant model. The break is only partly modelled by the dummy, and the rest by a near unit root (a typical outcome), which helps in forecasting, but misleads in policy reactions and latencies. However, no diagnostics are now significant, and the fit is closer to that of the DGP ($\sigma = 1$), as well as the solved static long-run equation for c on i having a coefficient of 0.98, albeit that a unit root cannot be rejected. The theoretical and simulation analyses in Castle and Hendry (2010c) explain such outcomes. Figure 3 shows the resulting improvement in constancy for recursive estimation of (21). The Chow test does not reject anywhere, and the large outliers in 1974 have been eliminated.

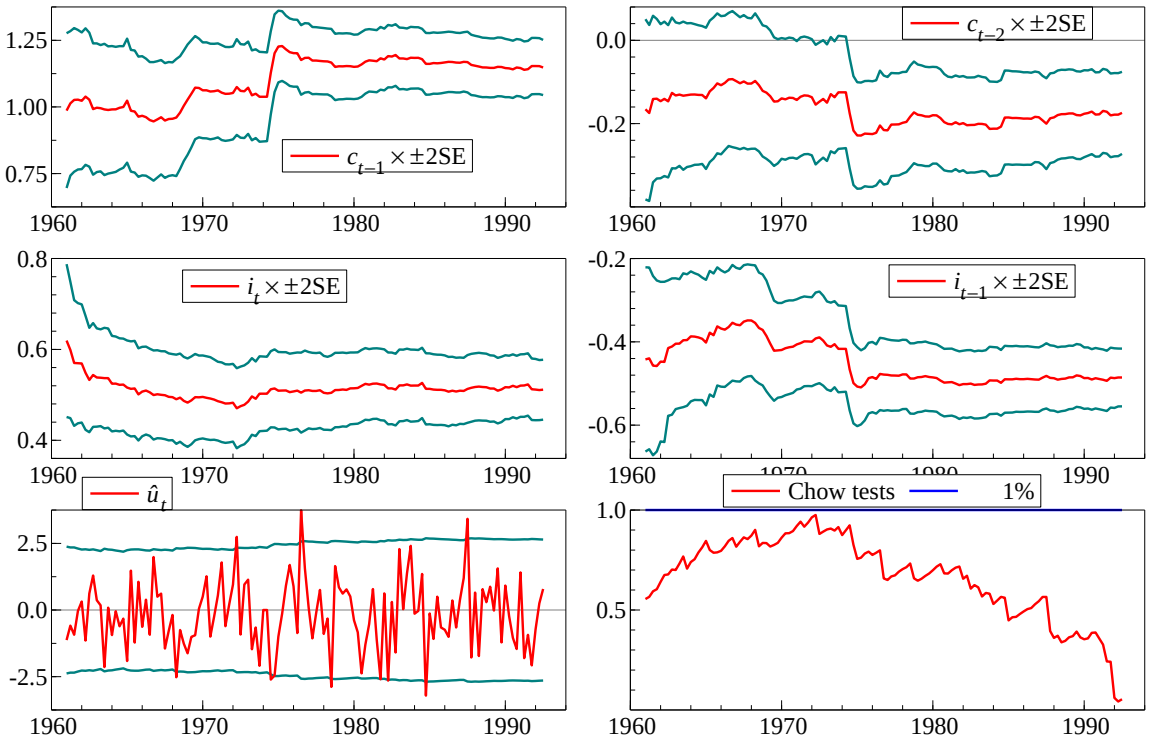

Figure 3: Incorrect artificial-data model specification with IIS

5 Conclusion

Recent developments in automatic model selection enable the real complexities of economic data modelling to be tackled, jointly addressing many candidate variables, some of which matter whereas others do not, long lag lengths, non-linearity, multiple location shifts and data contamination. The intercorrelations between economic variables, their non-stationarity, and high dimensionality necessitate handling all of these together if sustainable models are to result.

The chapter has discussed the application of such methods to the three central states of nature, where the initial model is over-specified relative to the data generation process, exactly specified and under-specified. In the first setting, the key issue is eliminating irrelevant variables while retaining relevant, and even large numbers of candidate variables hold few terrors. In the second, selection over non-theory variables is costless when the theory variables are retained. In the third, a mis-specified outcome is bound to occur, but selection can mitigate some of the problems due to location shifts in the unknowingly omitted variables. Thus, selection from a much larger initial general unrestricted model is generally beneficial relative to fitting a pre-specified equation, reversing the widely-held folklore of the economics profession that model selection is a pernicious, if unfortunately necessary, activity.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., and Csaki, F.(eds.), *Second International Symposion on Information Theory*, pp. 267–281. Budapest: Akademia Kiado.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59**, 817–858.

Campos, J., and Ericsson, N. R. (1999). Constructive data mining: Modeling consumers' expenditure in Venezuela. *Econometrics Journal*, **2**, 226–240.

Campos, J., Ericsson, N. R., and Hendry, D. F. (2005). Editors' introduction. In Campos, J., Ericsson, N. R., and Hendry, D. F.(eds.), *Readings on General-to-Specific Modeling*, pp. 1–81. Cheltenham: Edward Elgar.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2009). Model selection when there are multiple breaks. Working paper 472, Economics Department, University of Oxford.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2010). Evaluating automatic model selection. *Journal of Time Series Econometrics*, forthcoming.

Castle, J. L., and Hendry, D. F. (2010a). Automatic selection of non-linear models. In Wang, L., Garnier, H., and Jackman, T.(eds.), *System Identification, Environmental Modelling and Control*, forthcoming. New York: Springer.

Castle, J. L., and Hendry, D. F. (2010b). A low-dimension, portmanteau test for non-linearity. *Journal of Econometrics*, **158**, 231–245.

Castle, J. L., and Hendry, D. F. (2010c). Model selection in under-specified equations with breaks. Discussion paper 509, Economics Department, Oxford University.

Castle, J. L., and Shephard, N.(eds.)(2009). *The Methodology and Practice of Econometrics*. Oxford: Oxford University Press.

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

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.

Desai, M. J., and Weber, G. (1988). A Keynesian macro-econometric model of the UK: 1955–1984. *Journal of Applied Econometrics*, **3**, 1–33.

Doornik, J. A. (2009). Autometrics. In Castle, and Shephard (2009), pp. 88–121.

Doornik, J. A., and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, **70**, 927–939.

Doornik, J. A., and Hendry, D. F. (1994). *PcGive 8: An Interactive Econometric Modelling System*. London: International Thomson Publishing, and Belmont, CA: Duxbury Press.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1007.

Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica*, **46**, 1303–1313.

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

Hendry, D. F. (2003). J. Denis Sargan and the origins of LSE econometric methodology. *Econometric Theory*, **19**, 457–480.

Hendry, D. F. (2009). The methodology of empirical econometric modeling: Applied econometrics through the looking-glass. In Mills, T. C., and Patterson, K. D.(eds.), *Palgrave Handbook of Econometrics*, pp. 3–67. Basingstoke: Palgrave MacMillan.

Hendry, D. F. (2010). Revisiting UK consumers' expenditure: Cointegration, breaks, and robust fore-

casts. *Applied Financial Economics*, **21**, 19–32.

Hendry, D. F., and Doornik, J. A. (2009). *Empirical Econometric Modelling using PcGive: Volume I*. London: Timberlake Consultants Press.

Hendry, D. F., and Johansen, S. (2010). Model selection when forcing retention of theory variables. Unpublished paper, Economics Department, University of Oxford.

Hendry, D. F., Johansen, S., and Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, **33**, 317–335. Erratum, 337–339.

Hendry, D. F., and Krolzig, H.-M. (2005). The properties of automatic Gets modelling. *Economic Journal*, **115**, C32–C61.

Hendry, D. F., and Srba, F. (1980). AUTOREG: A computer program library for dynamic econometric models with autoregressive errors. *Journal of Econometrics*, **12**, 85–102.

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

Johansen, S., and Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In Castle, and Shephard (2009), pp. 1–36.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions – 1* 2nd ed. New York: John Wiley.

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.

Koopmans, T. C. (1947). Measurement without theory. *Review of Economics and Statistics*, **29**, 161–179.

Leamer, E. E. (1974). False models and post-data model construction. *Journal of the American Statistical Association*, **69**, 122–131.

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, **73**, 31–43.

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

Maasoumi, E. (ed.)(1988). *Contributions to Econometrics: John Denis Sargan*. Cambridge: Cambridge University Press.

Mizon, G. E. (1995). Progressive modelling of macroeconomic time series: The LSE methodology. In Hoover, K. D. (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, pp. 107–169. Dordrecht: Kluwer Academic Press.

Pagan, A. R. (1987). Three econometric methodologies: A critical appraisal. *Journal of Economic Surveys*, **1**, 3–24.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society B*, **31**, 350–371.

Salkever, D. S. (1976). The use of dummy variables to compute predictions, prediction errors and confidence intervals. *Journal of Econometrics*, **4**, 393–397.

Sargan, J. D. (1988). *Lectures on Advanced Econometric Theory*. Oxford: Basil Blackwell. Edited, and with an introduction, by Meghnad Desai.

Sargan, J. D. (2001). Model building and data mining. *Econometric Reviews*, **20**, 159–170.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, **B, 58**, 267–288.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817–838.