Journal of Economic Methodology 7:2, 231–264 2000

Revisiting data mining: 'hunting' with or without a license

Aris Spanos

Abstract The primary objective of this paper is to revisit a number of empirical modelling activities which are often characterized as data mining, in an attempt to distinguish between the problematic and the non-problematic cases. The key for this distinction is provided by the notion of error-statistical severity. It is argued that many unwarranted data mining activities often arise because of inherent weaknesses in the Traditional Textbook (TT) methodology. Using the Probabilistic Reduction (PR) approach to empirical modelling, it is argued that the unwarranted cases of data mining can often be avoided by dealing directly with the weaknesses of the TT approach. Moreover, certain empirical modelling activities, such as diagnostic testing and data snooping, constitute legitimate procedures in the context of the PR approach.

Keywords: data mining, severity, use novelty, predesignationist stance, misspecification testing, data snooping

1 INTRODUCTION

It is widely recognized that economic theorists display a lot of scepticism when faced with empirical evidence which cannot be accounted for by their theories; although they are often eager to present their own empirical evidence, or refer to an empirical study, that 'confirms' (in some sense) their theory. When asked to explain such scepticism, they usually reply that the data mining activities other econometricians often indulge in, discredit the trustworthiness of such evidence (see Mayer 2000, Hoover and Perez 2000).

The purpose of this paper is to revisit the various searching activities that are often considered as data mining and attempt to distinguish between problematic and non-problematic cases. The key to being able to make this distinction is provided by the notion of severity suggested by Mayo (1996), a philosopher of science. It is argued that many unwarranted data mining activities often arise because of inherent weaknesses in the Traditional Textbook (TT) methodology. Using the Probabilistic Reduction (PR) approach to empirical modelling, it is argued that the unwarranted cases of data mining can often be often avoided by dealing directly with the weaknesses of the TT approach. Moreover, certain empirical modelling activities, such as diagnostic

Journal of Economic Methodology ISSN 1350-178X print/ISSN 1469-9427 online © 2000 Taylor & Francis Ltd
http://www.tandf.co.uk/journals

testing and data snooping, which are often characterized as data mining by TT modellers, when viewed in the context of the PR approach, they constitute legitimate procedures that contribute significantly to the reliability of empirical evidence.

2 WHAT CONSTITUTES DATA MINING?

The term 'data mining' is usually used derisively to describe a group of activities in empirical modelling that are considered to lie outside the norms of 'proper' modelling; see Hoover and Perez (2000). This presupposes the existence of a well-defined scientific tradition in econometrics that establishes the norms of scientific research. At this stage there is no clearly-articulated framework for empirical modelling over which the majority of econometricians will agree represents a well-defined methodological tradition. The closest to such a tradition is the so-called TT, which is often discussed briefly in introductory chapters of econometric textbooks; see Gujarati (1995) and Intriligator (1978). The problem with such accounts is that no practising econometrician is likely to admit that s/he belongs to this empirical tradition. In methodological discussions the critics of the TT approach (see Leamer 1978, 1983, Hendry 1993, 1995, Spanos 1988, 1995b) are often accused of criticizing a 'straw man'; see Granger (1990) for a collection of such papers. This situation renders the issue of discussing data mining difficult because it cannot be viewed as a clear problem of norm undermining. An attempt, however, will be made to uncover the feature(s) that these activities have in common and identify on what grounds they are objectionable.

2.1 Data mining, use-novelty and predesignation

Let us begin with a list of activities which are often considered as data mining:

1. The selection of the observed data (time series, cross-section, panel), the sample period or population, frequency (annual, quarterly, etc.) and their measurement.
2. The selection of regressors (hunting for a certain correlation structure).
3. Respecification: changing the assumptions of the model.
4. Diagnostic testing.
5. Data snooping.

Looking at this list carefully suggests that, perhaps, the feature that all these activities share is that paraphrasing Mayo (1996, pp. 316–7):

> the data are being utilized for double-duty, to arrive at a claim (a model, an estimator, a test or some other inference proposition) in such a way that the claim is constrained to satisfy some criteria (e.g., fit) but the same data is regarded as supplying evidence in support of the claim arrived at.

Assuming that this is the common thread that links all these activities, the primary problem raised by data mining is closely related to the widely discussed issue of the violation of ‘use novelty’ in philosophy of science. Use-novelty is understood as the data not having been used to formulate a hypothesis for which the same data will be considered as evidence (see Mayo 1996: ch. 6–9). In the present context the violation of use-novelty appears to be narrower than the general case where data are used both to arrive at and provide grounds for a claim. It arises primarily when some form of preliminary searching through various claims takes place before the final assertion (confirmation of a hypothesis or statistically significant result) is made; hence the labels ‘mining’, ‘hunting’, ‘fishing’, ‘shopping’ and ‘snooping’.

In order to avoid such misleading inferences it was thought that in hypothesis testing (the Neyman-Pearson (NP) procedure) the modeller should adhere to the *predesignationist* requirement that demands that the modeller specifies the hypothesis in question prior to ‘investigating’ the data. This condition is designed to guard against uninformative tests or fabrication. Intuitively, if one views testing as analogous to target shooting, the predesignation stance amounts to ensuring that the target is predetermined in order to avoid the scenario of shooting at a blank wall and then drawing the bull’s eye around the hole made by the bullet. The primary objective of predesignation is to ensure that by performing the test, the modeller can learn something about the underlying phenomenon of interest, using the observed data in conjunction with the theory in question. In the target shooting example, nothing is learned about the shooting skills of the person when the target is drawn after the shooting occurs. This is clearly articulated by Lovell (1983, p. 1):

> The art of fishing over alternative models has been partially automated with stepwise regression programs. While such advances have made it easier to find high $\bar{R}^2s$ and ‘significant’ t-coefficients, it is by no means obvious that reductions in the costs of data mining have been matched by a proportional increase in our knowledge of how the economy actually works.

2.2 Data mining and severity

The question that naturally arises at this stage is ‘what is wrong with the final assertion if arrived at via a *postdesignationist* searching procedure?’ According to Mayo (1996: ch. 8), a careful examination of the rationale for preferring use-novelty shows that the real aim of the requirement to avoid double-duty for the data is the desire to ensure that such claims pass a ‘severe test’. The real problem with data mining, according to Mayo, arises when its utilization gives rise to zero or low ‘error-severity’ tests. Intuitively, a low severity test is like playing tennis with the net down; one learns very little about the tennis-playing skills of the players in question. The notion of error-severity in the context of the NP formulation is defined in terms of two conditions. The first condition ensures that the hypothesis of interest *H* 'fits the data **x**', and the second condition requires that 'there is a very high probability that the test procedure *T* would *not* yield such a passing result, if *H* is false' (see Mayo 1996: p. 180). Returning to the shooting example, where the bull's eye was drawn around the hole made by the bullet, one can illustrate the severity notion by noting the following:

> (i) good fit: the bullet close to bull's eye is in accordance with what would be expected if the shooter were skilled at hitting the target. However,
> (ii) the probability that this procedure yields so good a fit, even if it is assumed that the shooter is not killed, is one; hence zero severity.[1]

Therefore, one cannot use this 'test' to discriminate between skilled and unskilled shooters, and thus 'passing' this test fails to provide evidence of skill. Severity provides an assessment of the test's probativeness and is designed to guard against misleading inferences based on 'soft' evidence, which make it too easy to regard the observed data as evidence in support of a given claim.

Mayo (1996: ch. 9), however, argues that use-novelty is not necessary for good tests nor well grounded inferences, so perhaps this can help us in distinguishing problematic from non-problematic cases of data mining in econometrics. While it is true that data mining, as with all violations of use novelty, by definition, ensures that a claim (a model, an estimator, a test or some other inference proposition) reached, will accord with certain criteria (e.g. fit) with data **x**, it does not follow that this claim has not passed a severe test with data **x**. Even though applying a procedure of data mining is assured of 'passing' the model or claim it arrives at, it may actually be very improbable that the claim is incorrect, false or guilty of a specified error. While there are cases where this improbability can be made mathematically rigorous, at other times it corresponds to a more informal assessment, by which the ways the given claim can be in error have been well probed. Once one understands the 'goal' as doing a good job of ruling out the error of interest (so that one can learn something about the phenomenon of interest), one can begin to understand why certain cases of data mining are (a) justifiably deemed problematic while (b) others are not problematic. Mayo (1996) gives an example of such a non-problematic case: testing the assumptions of the postulated statistical model:

> Examples of NP procedures that violate the predesignation – by violating use-novelty – are those involved in checking the assumptions of an experimental test. The same data may lead to constructing a hypothesis – say, that the trials are not independent – and at the same time may be used to test that hypothesis . . . In checking if a particular data satisfies assumptions, such a double use of data is likely to offer a better test than looking to the data of some new experiment.
>
> (p. 295)

Common sense suggests that in certain cases of checking for possible departures from the underlying probabilistic assumptions, nothing problematic arises from not adhering to the predesignationist stance, as long as severity is not ignored. Mayo (1996: pp. 278–93) considers, for example, how in testing Einstein's predicted deflection effect in 1919, one set of eclipse data was used to suggest, develop, as well as test the hypothesis so that it was spoiled by distortions of the telescope mirror, due to the sun's heat. Thanks to reliable procedures for distinguishing mirror distortions from others, the argument for discrediting this data was sufficiently severe. As Mayo (1996: pp. 184–5), argues, the rationale for 'learning from error' in this way is as follows:

> (a) If one finds no problem or error despite a highly probative search (one with a very high probability of uncovering the problem if it is present), then this provides strong evidence for the absence of the proble or error; e.g. failing to find a tumour despite many sensitive diagnostic tests.
>
> (b) If one finds a problem that one cannot explain away despite a trenchant search for how it could be wrong to suppose the problem is real, then there is strong evidence for its existence; e.g. the eclipse data example above.

2.3 Unwarranted data mining

With this discussion providing the backdrop, we can proceed to consider the various cases of data mining mentioned above in an attempt to distinguish between the problematic and non-problematic cases. A tentative definition of data mining for the problematic cases might be:

Unwarranted data mining[2] denotes a group of searching procedures:

(1) seeking an accordance between data $\boldsymbol{x}$ and a proposition M (model, hypothesis or claim) arrived at by either:
   (a) using the data to search for a good fit between $\boldsymbol{x}$ and M, or;
   (b) searching for data $\boldsymbol{x}$ which accords with a prespecified M;
(2) regarding data $\boldsymbol{x}$ as evidence in support of M, even though,
(3) the procedure in (1) results in M failing to pass a reliable or severe test with $\boldsymbol{x}$.

That is, the combination of (1) and (2) is unwarranted when (3) is the case. As already shown informally, there are certainly contexts where carrying out (1)-(a) and even (1)-(b) do not result in a lack of severity. Other cases are spared from being labelled data mining because there is no claim that (2) holds. Often, the mining procedures are problematic precisely because they prevent even an approximate assessment of the overall reliability or severity of a particular inference or test. Unless one can show the claim in question is not vitiated by lack of severity (e.g. high error probabilities), one cannot provide the positive grounds needed to sustain the inference. Moreover, reporting on the searching does not automatically spare if (2) is still asserted; see Mayo (1996: p. 297).

The methodological principle that 'unwarranted data mining should be avoided' can be justified in terms of the meta-methodology espoused in Mayo (1996: pp. 148–50) where any methodological rule, principle, requirement, etc., should be appraised according to whether or not it violates (or contributes to) the reliability[3] of the overall inference. In acknowledging this, the discussion is intended to have two positive payoffs:

(1) It will help to see clearly why certain data mining procedures of interest to economists increase rather than diminish this overall reliability.
(2) By understanding just how certain common types of data mining procedures may influence and considerably raise the test's overall error probabilities, we may be in a better position to see just what a researcher would need to show in order to circumvent the problem.

3 TWO ALTERNATIVE APPROACHES: A BRIEF SUMMARY

The discussion that follows relates the various data mining activities to norm undermining in the context of two alternative approaches to econometric modelling: the Traditional Textbook (TT) and the Probabilistic Reduction (PR) as expounded by Spanos (1986, 1988, 1989, 1995b).

3.1 Traditional Textbook (TT) approach

The Traditional Textbook (TT) approach is summarized by one of the classic econometric textbooks as follows:

> Standard econometric practice for a long time was to (i) formulate a model on the basis of theory or previous econometric findings, (ii) estimate the parameters of the model using what relevant sample data one could obtain, and (iii) inspect the resultant estimates and associated statistics to judge the adequacy of the specified model. The inspection typically focused on the overall fit, the agreement of the signs of the coefficients with a priori expectation, the statistical significance of the coefficients, and a test of autocorrelation in the disturbances. If the model were deemed 'satisfactory' on these criteria, a new equation would be added to the literature and might well be used to make predictions for data points outside the time scale or empirical range of the sample.
>
> (Johnston and Dinardo 1997: p. 112)

The same textbook proceeds to discuss how data mining arises in practice:

> If the estimated model were deemed 'unsatisfactory', the investigator would engage in a specification search, trying out different reformulations

> in an attempt to reach a 'satisfactory' equation. That search process went largely unreported, for it smacked of *data mining*, which was held to be reprehensible, and also because it was practically impossible to determine correct *P*-values and confidence coefficients for the final statistics.
>
> (Johnston and Dinardo 1997: p. 112).

In an attempt to explain how some of the unwarranted data mining activities have arisen, the main modelling stages of the TT approach, as described in the above quotation from Johnston and Dinardo (1997), are schematically shown in Figure 1; see Intriligator (1978) for a similar diagram.

As argued by Spanos (1995b), the traditional textbook approach constitutes an adaptation of the experimental design modelling framework where the statistical model and the design of the experiment are two faces of the same coin, differing only by white-noise errors. This adaptation often turns out to be inappropriate (and misleading) for the statistical analysis of observational (non-experimental) data, where the gap between the theory and the observed data is usually sizeable and can rarely be bridged using just white-noise errors. Moreover, the TT approach leaves little room for assessing the reliability of empirical evidence.

3.2 The Probabilistic Reduction (PR) approach

In the context of the Probabilistic Reduction (PR) approach the gap between the theory and the observed data is explicitly acknowledged by distinguishing between theory and statistical information (and the corresponding models), and devising statistical procedures, which take the probabilistic structure of the data explicitly into consideration when postulating statistical models. Built into the approach is the overriding concern to follow procedures that ensure the reliability of empirical evidence. In the context of the PR approach the primary objective of econometric modelling is 'the systematic study of economic phenomena using observed data' (see Spanos 1986, pp. 670–1) and

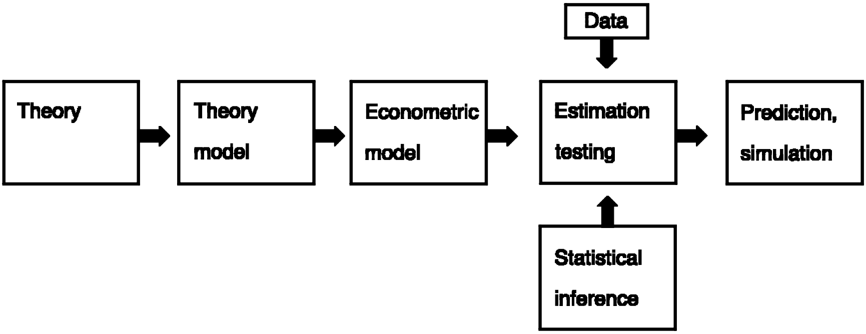

Figure 1 The traditional textbook approach to econometric modelling

not ‘the quantification of theoretical relationships’ as in the TT approach. The ‘learning from the observed data’ is primarily accomplished by modelling procedures that give rise to ‘reliable empirical evidence’; evidence that can be used to assess the theory (or theories) in question. Hence, any activity that contributes to the enhancement of the reliability of empirical evidence, and thus promotes learning from the observed data, should be strongly encouraged. Indeed, the justification of statistical methods and models is found in their ability to provide systematic strategies for learning from observed data, see Mayo (1996).

In summary, the PR approach begins by distinguishing between the observable phenomenon of interest (actual Data Generating Mechanism (DGM)), the ‘theory’ that purports to explain it, the ‘theoretical model’ specified in terms of the behaviour of economic agents and its ‘estimable’ form given the available data; it often comes in the form of the observational implications of a theory[4] (see Figure 2).

A particularly important concept in the context of the PR approach is that of a ‘statistical model’, viewed as an idealized description (codified exclusively in terms of probabilistic concepts) of the stochastic mechanism that gave rise

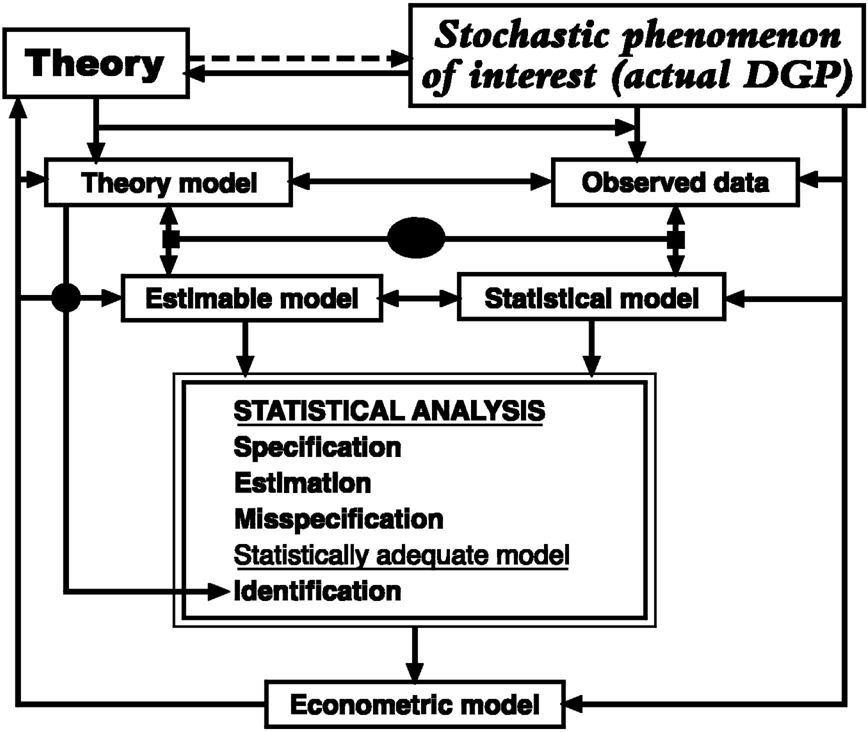

Figure 2 The Probabilistic Reduction approach

to the observed data; a statistical DGM. The statistical model is specified with a view to 'capturing' the probabilistic structure of the observed data; see Spanos (1986: pp. 20–1). Formally 'a statistical model' is defined as a set of internally consistent probabilistic assumptions, which purports to describe the observable stochastic phenomenon of interest. Its primary objective is to model the systematic attributes of the observable phenomenon of interest via the systematic (statistical) information in the observed data. Intuitively, statistical information is any pattern that the modeller can capture via probabilistic concepts (distribution, dependence, heterogeneity) without having to resort to any information from the theory in question; to discern the statistical information the modeller does not need to know what the series actually measures. The success of the empirical modelling depends crucially on being able to detect these 'chance regularity patterns' and then choose the appropriate probabilistic concepts (in the form of probabilistic assumptions) in order to capture this information.

'Specification' refers to the actual choice of a statistical model based on the information provided by the theoretical model and the probabilistic structure of the observed data in question. 'Misspecification' refers to informal graphical assessment and the formal testing of the assumptions underlying the statistical model. 'Respecification' refers to the choice of an alternative statistical model when the original choice is found to be inappropriate for the data in question. This process from specification to misspecification testing and respecification will be repeated until a statistically adequate model is found. 'Identification' constitutes the last stage of empirical modelling at which the theoretical model is related to the statistically adequate estimated statistical model.

By utilizing key features of the PR approach in conjunction with Mayo's notion of severity, one will be in a position to separate licenced from unwarranted data mining activities.

4 SPECIFICATION/RESPECIFICATION

4.1 The selection of the observed data

The selection of the observed data as it relates to their nature (time series, cross-section, panel), the sample period or population, frequency (annual, quarterly etc.), and their measurement, are often considered as part of the unwarranted data mining activities because in the context of the TT approach the choice of the data is viewed as an afterthought. Looking at Figure 1, we can see that the econometric (statistical) model is specified before the observed data are chosen, giving the impression that the former should be invariant to the choice of the data. The classic example of such an attitude is provided by the history of the Absolute Income Hypothesis (AIH) consumption function ($C = \alpha_0 + \alpha_1 Y^D$), where the differing coefficient estimates of (α_0, α_1) arising from being estimated using time series as opposed to cross-section data, was considered a paradox in need of explanation, see Wallis (1973). Even in cases where the nature of the data, as well as the sample period are specified, say time series for the period 1947–1990, TT modellers often indulge in searching activities, which involve trying out several series loosely connected to the theory variables (C, Y^D). In practice several series under the general rubric 'consumption' and 'income' measuring real, nominal as well as per capita series are often tried out; there are several possible candidates such as consumer 's expenditure (with or without durables) and personal disposable income, GDP etc.

Viewed in terms of the definition of unwarranted data mining, it is clear that this activity constitutes a case of (1)-(b). The modeller follows a search procedure where several data sets (differing in nature, sample period, frequency or what they are actually measuring) are tried out, in an attempt to discover an estimated model (say, a linear regression), which can be considered as a verification of the theory in question when judged in terms of certain criteria; see quotation from Johnston and Dinardo (1997) above. This data mining activity prevents even an approximate assessment of the overall reliability or severity of the claim, and the criteria utilized are not enough to show that the claim in question has 'passed' a severe test. Hence, the positive grounds needed to sustain the claim are not furnished. Moreover, full reporting on the search does not allow the assessment of the error probabilities because, as shown below, there is no comprehensive statistical model which includes the tried out models as special cases.

4.1.1 The PR approach: observed data and statistical models

When viewed in the context of the PR approach, the apparent leeway that allows for data mining, involving 'trying out' dozens of data series in search of a good fit for a theory model, is a consequence of a major weakness of the TT approach: the way the statistical model is assumed to coincide with the theory model, apart from some white-noise error terms, leaves no room to deal effectively with the gap between theory concepts and observed data; see Spanos (1986, 1988, 1995a). In the context of the PR approach there is nothing arbitrary about the choice of the observed data and the latter are inextricably bound up with the choice of a statistical model. The decision regarding what data are appropriate depends crucially on the nature of the primary (theoretical) questions of interest and the relationship between theory concepts and the available data. The statistical model is specified directly in terms of the observable random variables that gave rise to the observed data in question, and the modeller should ensure that any gap between the theory concepts and the observable variables is judiciously bridged. This is the motivation behind the introduction of the 'estimable model' notion. Often, there are no data which correspond directly to the theory concepts, and the modeller has to consider the question of 'what form of the theory model is estimable given this data?'; see Spanos (1986, 1995a) for an extensive discussion on the demand–supply model. In this sense the issue of selecting the observed data, as well as their measurement, do not constitute a statistical inference problem. It's fundamentally an issue concerning the primary questions of interest and how they are related to the nature of the observed data: measurement information (see Spanos 1986: ch. 26). In view of the fact that economic data are rarely the result of designed experiments, bridging the gap between the theory concepts and the available data constitutes one of the most difficult tasks awaiting the modeller; brushing it under the carpet does not make this problem go away.

In the context of the PR approach, the statistical model, which provides the foundation of the overall statistical inference, presupposes that the above choices concerning the data (measurement information) have already been made with an appropriate bridging of the gap between theory concepts and observed data. Statistical inference proper commences at the point where the 'observed data' $Z{:} = (z_1, z_2, \ldots, z_T)$ are chosen, in view of the theory or theories in question. In an attempt to delineate the statistical from the theoretical issues, the informational universe of discourse for statistical inference is demarcated by the joint distribution (see Spanos 1989) of the 'vector stochastic process':

$$\{Z_t, t \in \mathbb{T}\}, Z_t \in \mathbb{R}^m_Z, \mathbb{T} \text{ being an index set,} \tag{1}$$

underlying the observed data chosen, the Haavelmo distribution:

$$D(Z_1, Z_2, \ldots, Z_T; \varphi), \forall (z_1, z_2, \ldots, z_T) \in \mathbb{R}^{mT}_Z. \tag{2}$$

Note that in order to emphasize the distinction one denotes the random vector by Z_t and its value by Z_t. Statistical models, such as the linear regression, the dynamic linear regression and the Vector Autoregression (VAR), are viewed as reductions from (2). That is, denoting the set of all possible statistical models defined in terms of (2) by $\mathcal{P}$, the chosen (parametric model $P_\theta \in \mathcal{P}$, constitutes just one element, which can be viewed as arising by imposing certain probabilistic (reduction) assumptions on the process (1). This enables the modeller to view the model, not in isolation, but in the context of a broader modelling framework, which can help to delineate a number of statistical issues such as misspecification testing and respecification; see below.

In an attempt to illustrate the above procedure let us consider how the Normal/Linear Regression (NLR) model can be viewed as a the reduction from (2). The reduction assumptions are:

$\{Z_t, t \in \mathbb{T}\}$ is Normal (N), Independent (I), and Identically Distributed (ID).

The assumptions of IID give rise to the simplification:

$$D(\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_T;\varphi) \overset{\text{I}}{=} \Pi_{t=1}^{T} D_t(\mathbf{Z}_t;\varphi_t), \text{ where } \mathbf{Z}_t^{\mathrm{T}}:=(y_t,\boldsymbol{X}_t^{\mathrm{T}})^{\mathrm{T}}, \forall \boldsymbol{z}_t \in \mathbb{R}_Z^m,$$
$$\overset{\text{IID}}{=} \Pi_{t=1}^{T} D(\mathbf{Z}_t;\varphi) = \Pi_{t=1}^{T} D(y_t|X_t;\varphi_1)\cdot D(\boldsymbol{X}_t;\varphi_2). \tag{3}$$

The NLR model is specified exclusively in terms of $\Pi_{t=1}^{T} D(y_t \mid \boldsymbol{X}_T;\varphi_1)$, using the weak exogeneity of X_t with respect to φ_1, which follows from the normality reduction assumption; see Spanos (1986). The model is often specified in terms of the statistical Generating Mechanism (GM):

$$y_t = \beta_0 + \beta_1^{\mathrm{T}}\boldsymbol{x}_t + u_t,\ t \in \mathbb{T}, \tag{4}$$

where model parameters $\phi := (\beta_0, \beta_1, \sigma^2)$, associated with the conditional process $\{(y_t \mid \boldsymbol{X}_t), t \in \mathbb{T}\}$, are related to the primary parameters:

$$\psi := (\mu, \Sigma),\ \mu = \mathrm{E}(\mathbf{Z}_t) := \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \mathrm{Cov}(\mathbf{Z}_t) := \begin{pmatrix} \sigma_{11} & \sigma_{21}^{\mathrm{T}} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix},$$

associated with the original process $\{\boldsymbol{Z}_t, t \in \mathbb{T}\}$ via:

$$\beta_0 = \mu_1 - \beta_1^{\mathrm{T}}\mu_2, \qquad \beta_1 = \Sigma_{22}^{-1}\sigma_{21}, \qquad \sigma_2 = \sigma_{11} - \sigma_{21}^{\mathrm{T}}\Sigma_{22}^{-1}\sigma_{21}. \tag{5}$$

The (testable) model assumptions [1]–[5] (see Spanos 1986 for further details):

[1] Normality:	$D(y_t \mid \boldsymbol{x}_t;\psi)$ is normal,
[2] Linearity:	$E(y_t \mid \boldsymbol{X}_t = \boldsymbol{x}_t) = \beta_0 + \beta_1^{\mathrm{T}}\boldsymbol{x}_t$, linear in $\boldsymbol{x}_t$,
[3] Homoskedasticity:	$Var\,(y_t \mid \boldsymbol{X}_t = \boldsymbol{x}_t) = \sigma^2$, free of $\boldsymbol{x}_t$,
[4] *t*-homogeneity:	$(\beta_0, \beta_1, \sigma^2)$ are not functions of $t \in \mathbb{T}$,
[5] 'temporal' independence:	$\{(y_t \mid \boldsymbol{X}_t = \boldsymbol{x}_t), t \in \mathbb{T}\}$ is an independent process. (6)

An important dimension of specification in the context of the PR approach is the relationship between the reduction and model assumptions, see Table 1.

The NLR model (specified by the statistical GM in conjunction with

Table 1 Probabilistic assumptions

Reduction: $\{\boldsymbol{Z}_t, t \in \mathbb{T}\}$		Model: $\{(y_t \mid \boldsymbol{X}_t = \boldsymbol{x}_t), t \in \mathbb{T}\}$
N	→	[1],[2],[3]
I	→	[5]
ID	→	[4]

assumptions [1]–[5]) specify a statistical DGM that defines a stochastic mechanism of which the observed data are interpreted as a realization. The ultimate objective of empirical modelling is to be able to learn something about the observable phenomenon of interest via the utilization of both theory and statistical information. Statistical inference has an important role to play in this learning process only when appropriate statistical procedures are followed. Statistical inference is often viewed as the quintessential form of 'induction': using a set of data (specific) to draw conclusions about the stochastic phenomenon (general) that gave rise to the observed data. However, it is often insufficiently recognized that this inductive procedure has a fundamentally deductive argument embedded within it. If one ignores the first step (the observed data) and the last step (using inference results to draw conclusions about the underlying DGM), the procedure from the postulated model to the inference propositions (estimators, tests, predictors, etc.) is essentially 'deductive'; no data are used in deriving inference propositions concerning the optimality of estimators, tests etc.; estimators and tests are pronounced 'optimal' based on purely deductive reasoning. The deductive component of statistical reasoning takes the form:

> If certain premises are assumed valid, certain conclusions necessarily follow.

In this sense, statistical inference depends crucially on the validity of the premises: the statistical model. Assuming this, one proceeds to derive statistical inference propositions (estimators, test statistics, predictors, etc.) using mathematical deduction. However, in empirical modelling one needs to establish the validity of the premises in order to ensure the reliability of the overall inference based on the premises. Statistical adequacy is established by testing the assumptions that make up the model in question for possible departures from the underlying assumptions [1]–[5] *vis-a-vis* the observed data, and if no departures are detected after an intense but judicious probing, one can use the estimated model as ensuring the reliability of statistical inference results concerning the primary question of interest; see Spanos (1986, 1999) for several misspecification tests for each of the assumptions [1]–[5].

4.1.2 Searching for data to confirm a theory: unwarranted data mining

Let us return to the data mining activity of searching through several data sets, in an attempt to discover an estimated model (say, a linear regression (4)) which instantiates a theory M, on the basis of the TT commonly used criteria. Looking at (4), (5) and (6), it becomes obvious that the observed data ($z_t := (y_t, \boldsymbol{x}_t)$, $t = 1, 2, \ldots, T$) and the associated statistical model are inextricably bound up to the extent that any changes in the data for will change the statistical model; different coefficients are estimated. Any attempt to formalize the above data mining search, in order to keep track of the actual error probabilities is doomed because there is no comprehensive statistical model within which every other model can be nested. It is apparent, however, that the final model M fails to pass a reliable or severe test with Z because the criteria being utilized provide only 'soft' evidence for M in the sense that 'the probability that the followed test procedure would yield so good an accordance result when M is false, is very high'.

In the light of the above discussion of statistical adequacy, in addition to the severity problem, the inference results of this procedure are likely to be unreliable because the statistical adequacy of the estimated models has been ignored. Lack of statistical adequacy will render 'statistical criteria', such as the R^2, the t-ratios for the significance of the coefficients, as well as 'theory-oriented criteria' such as the signs and magnitudes of the estimates, not only misleading but also 'unreliable'. The only model assumption that has been partly assessed is assumption [5] using the DW test; the latter is a misspecification test of limited scope (see Spanos 1986). Departures from any of the other assumptions are likely to call into question the reliability of the criteria being utilized. Moreover, the severity of the search procedure itself is seriously suspect because it was designed to avoid situations where M might be false, utilizing the same (potentially) unreliable criteria.

4.2 The selection of regressors: hunting for correlation

This data mining issue is concerned with the activity of changing the variables in X_t in order to achieve a certain predetermined correlation structure (significant t-ratios etc.) in the context of (4). According to Mayer (2000, p. 2):

> Data mining occurs because most economic hypotheses do not have a unique empirical interpretation but allow the econometrician much leeway in selecting conditioning variables, lags, functional forms, and sometimes the sample.

This involves running several regressions using a variety of combinations of regressors and reporting as statistically significant those that are nominally significant, say at the 0.05 level, i.e. those that would reach 0.05 significance if they were the only (predesignated) hypotheses. It can be shown that the actual significance level, that is the overall probability of committing a type I error, differs from and may be much greater than the nominal level. So reporting the nominal level, as the actual one, results in regarding an effect as unusual under the null, even though it is not unusual under the null (e.g. rule of thumb in Lovell (1983: p. 3), also Mayo (1996: ch. 9). Honest hunters, as Mayo calls them, either set the significance levels suitably low or report the number of 'misses' so as to calculate, at least approximately, the overall significance level. In what follows one argues that being an honest hunter, although very important, is not enough to ensure the reliability of the overall inference.

4.2.1 Selection of regressors: the PR perspective

When this activity is considered in the context of the PR perspective, it becomes apparent that one needs to distinguish between adding new conditioning variables and including lags of the existing conditioning variables. This is because the latter leaves $D(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_T; \varphi)$ unchanged but the former does not.

A closer look at the PR specification of the NLR model (4) reveals that both the parameterization $(\beta_0, \beta_1, \sigma^2)$ as well as the underlying distribution $D(y_t \mid \boldsymbol{X}_t; \varphi_1)$ are inextricably bound up with a particular $(y_t, \boldsymbol{X}_t)$. This suggests that by changing the variables in $\boldsymbol{X}_t$, the statistical model itself changes rendering the associated statistics for different models non-comparable; see the discussion on the omitted variables argument in Spanos (1986: pp. 418–21). Hence, m different choices of regressors $(\boldsymbol{X}_{1t}, \boldsymbol{X}_{2t}, \ldots, \boldsymbol{X}_{mt})$ where:

$$\boldsymbol{X}_{it} \neq \boldsymbol{X}_{jt}, \text{ for } i \neq j \text{ and } \boldsymbol{X}_{it} \subset \boldsymbol{X}_t,\ i, j = 1, 2, \ldots, m,$$

give rise to m different regression models:

$$y_t = \beta_{0k} + \beta_{1\mathrm{k}}^{\mathrm{T}} \boldsymbol{x}_{1t} + u_{kt},\ t \in \mathbb{T}, \quad \text{for} \quad k = 1, 2, \ldots, m, \tag{7}$$

with m different parameterizations $((\beta_{0k}, \beta_{1k}, \sigma_k^2), k = 1, 2, \ldots, m,)$ based on m different conditional distributions $(D(y_t \mid \boldsymbol{X}_{kt}; \varphi_{1k}), k = 1, 2, \ldots, m)$. From the PR persepctive, the main conclusion relating to the data mining problem of changing the regressors is that each one of the estimated specifications constitutes a different statistical model, which renders the 'tried out' models non-comparable as they stand.

The only statistically coherent way to render these 'tried out' models comparable, is to specify a comprehensive NLR model which all these models are nested. In the above case the comprehensive NLR model chooses X_t as the union of all regressors: $\cup_{k=1}^{m} \boldsymbol{X}_{kt} = \boldsymbol{X}_t$. Indeed, this constitutes the primary motivation underlying the 'general-to-specific approach' associated with Hendry (1993, 1995). When the tried out models are viewed as arising from 'exclusion restrictions' on the comprehensive model (4), they can be tested by assessing the validity of their respective restrictions. It goes without saying that care should be taken to keep track of the error probabilities in cases of sequential or multiple testing; see Lovell (1983). However, adjusting the error probabilities is not sufficient to render the inference reliable because this presupposes that each of the estimated models (linear regressions) is statistically adequate; otherwise, not only the nominal, but what is considered to be the actual significance level, is likely to be misleading. This is because under several forms of misspecification (such as departures from assumption [5]) the t-test and F-test for significance have very different type I and type II errors than those under statistical adequacy.

Hence, for the general-to-specific procedure to be reliable, the statistical adequacy of the comprehensive model should be established first. Beyond passing the exclusion restrictions tests, the chosen model must also be statistically adequate, before it can be used to test any primary hypothesis of interest reliably. If all ‘tried out’ models turn out to be statistically inadequate, there is no choice. If one is statistically adequate and the others are not, then the choice on statistical adequacy grounds is straightforward. When more than one model turns out to be statistically adequate, then the choice has to be made on other statistical criteria such as parsimonious encompassing (Mizon and Richard 1986), predictive ability, parsimony and robustness or on theoretical grounds. However, theory congruence by itself does not suffice for valid inference. If an estimated model is going to be used for any form of statistical inference, its statistical adequacy is what ensures the reliability of the overall inference.

4.2.2 ‘Hunting’ for statistical significance: unwarranted data mining

Estimating and comparing several ‘tried out’ models using the commonly used TT criteria mentioned above, amounts to an unwarranted data mining activity, which lacks both reliability and severity as well as coherence. The lack of reliability and severity arises for the same reasons as the search through several different series for the variables involved, discussed above. The lack of coherence arises from the fact that the comparison between such models is tantamount to comparing ‘bad apples and soild oranges’. In summary, proper sequential (and multiple) testing requires honest hunters who also keep an eye on the trail: the statistical adequacy of the estimated models. If the statistical adequacy problems is ignored, the end result will often be spurious correlation; the honest hunter will be led astray.

4.3 Spurious correlation: exploiting chance?

The discussion of establishing a certain correlation structure often focuses on ‘goodness of fit’, t-ratios, etc. which, of course, ignore the problem of ‘statistical adequacy’. Indeed such criteria constitute the primary victims of statistical inadequacy. When inadequate attention is paid to the problem of statistical adequacy, the modeller is likely to utilize meaningless statistics, as if there were proper t-ratios and goodness of fit measures such as the R^2, to arrive at a certain correlation structure, deluding him/herself into thinking that the choices were guided by proper statistical arguments. For example, in the case where a set of time series data $Z := (z_1, z_2, \ldots, z_T)$ exhibit mean-heterogeneity (such as trends), the usual sample correlations (cross-correlation and autocorrelation), defined by:

$$\widehat{Corr}\,(z_{it}, z_{j(t\text{-}k)}) = \frac{\Sigma_{t=\kappa+1}^{T}(z_{it} - \bar{z}_i)(z_{j(t-\kappa)} - \bar{z}_j)}{\sqrt{\left[\Sigma_{t=1}^{T}(z_{it} - \bar{z}_i)^2\right]\left[\Sigma_{t=1}^{T}(z_{j(t-\kappa)} - \bar{z}_j)^2\right]}},\ i, j = 1, ..., m,\ k = 0, 1, ..., T - 1, \quad (8)$$

are likely to be very misleading. This is because one of the (implicit) assumptions underlying these statistics is that the mean in constant and can be estimated consistently using the sample means: $z_i = (1/T)\,\Sigma_{k=1}^{T}\, z_{it}$, $i = 1, 2, \ldots, m$. These spurious correlation results are due to the statistical misspecification of mean-heterogeneity, which renders the correlation numbers meaningless. The same is true for both the t-ratios and the R^2 when such data are used in the context of a linear regression model (4). This can be easily seen in the simple one regressor case where:

$$R^2 = 1 - \frac{\Sigma_{t=1}^{T}\hat{u}_t^2}{\Sigma_{t=1}^{T}(y_t - \bar{y})^2},\quad \tau(y) = \frac{\hat{\beta}_1}{s\sqrt{\Sigma_{t=1}^{T}(x_t - \bar{x})^2}}, \quad (9)$$

where $\Sigma_{t=1}^{T}\,(y_t - \bar{y})^2$ and $\Sigma_{t=1}^{T}\,(x_t - \bar{x})^2$ are likely to be artificially inflated by taking deviations from a constant mean instead of a proper time-changing one. This, in turn, will give rise to artificially high R^2 and t-ratio $\tau(\mathrm{y})$. Once the mean heterogeneity is captured and the above statistics are evaluated using deviations from appropriate means, the above spurious correlation problems disappear.

In the same class of problems one can include the well-known 'spurious regression' problem highlighted by Granger and Newbold (1974) and explained by Phillips (1986). In the case where the stochastic processes $\{y_t,\ t \in \mathbb{T}\}$, $\{x_t,\ t \in \mathbb{T}\}$ are two uncorrelated Wiener processes, the static regression:

$$y_t = \beta x_t + u_t,\ t \in \mathbb{T},$$

will be misspecified, giving rise to the spurious regression problem, because it ignores the temporal structure of the processes involved. On the other hand, when the static regression is respecified in the form of the Dynamic Linear Regression (DLR) model (in order to capture the temporal structure):

$$y_t = \alpha_0 x_t + \alpha_1 x_{t-1} + \alpha_2 y_{t-1} + u_t,\ t \in \mathbb{T},$$

no spurious regression problems arises and no new 'fancy' sampling distribution results are needed. Hence, the ingenuity of Phillips's analytical explanation notwithstanding, the real issue underlying the spurious regression problem is one of misspecification, not inappropriate use of sampling distribution results. The statistical adequacy of the postulated statistical model relative to the observed data chosen is of overriding importance for any form of valid statistical inference.

The connection between spurious regression (correlation) and statistical adequacy is relevant for the problem known as ‘exploiting chance’. Hoover and Perez (2000) argue that:

> Data-mining is considered reprehensible largely because the world is full of accidental correlations, so that what a search turns up is thought to be more a reflection of what we want to find than what is true about the world.

The world is full of ‘apparent’ chance correlations, but a closer look is likely to reveal that the overwhelming majority are ‘spurious correlations’. An efficient filtering device for separating ‘real’ statistical correlation from ‘spurious’ correlation is provided by a comprehensive misspecification testing when applied to the underlying statistical model. In the case of the linear regression model, real statistical significance can be distinguished from spurious significance by testing for departures from assumptions [1]–[5]. Given the restrictiveness of these assumptions, one can go as far as to suggest that any empirical regression that survives a thorough misspecification testing (especially the t-invariance of the parameters) is worth another theoretical look (is there a theory justification for such a correlation?) because it captures something that appears to persist and is invariant over the whole of the sample period. It is well known that numerous discoveries in science, including penicillin and radioactivity, were gainful cases of ‘exploiting chance’. Despite the accidental discovery of a regularity, on closer examination they turned out to be real phenomena and not spurious, because they were replicable. What is often not reported in books on the history of science are the numerous discoveries that turned out to be spurious after attempts to replicate them failed. Examples of such spurious results that failed to be confirmed are the well-known cold fusion and numerous spurious rays, such as the ‘black light’ and N-rays, that were proposed after the discovery of radioactivity in the early 20th century; see Kragh (1999). These examples suggest that when an empirical regularity persists and withstands resolute attempts to detect departures from the underlying probabilistic assumptions, the modeller should consider the possibility that a real effect might have been detected.

In light of the above discussion, one can also raise some doubts about the informativeness of the Monte Carlo simulation results reported in Hoover and Perez (2000). Using Monte Carlo simulated data they examine the effectiveness of the ‘general-to-specific’ procedure. The results of such an exercise can be misleading as a guide to what happens in actual empirical modelling because the statistically adequacy issue is completely sidestepped. The Monte Carlo simulations, assuming their design is not at fault, ensure the statistical adequacy of all the statistical models; a highly unlikely scenario for such a data mining activity when modelling with real data. Given the statistical adequacy of all the models involved, however, the discussion above suggests that it should not be surprising to discover that the general-to-specific procedure has superior performance properties; it provides a coherent statistical framework for comparing the various sub-models, using nested testing.

4.4 The PR approach and lags: respecification

At this point, it is important to return to the distinction between adding new regressors to a linear regression model and introducing lags. When lags on the original variables $\{Z_t, t \in \mathbb{T}\}$ are introduced into the regression, the situation is very different from adding new conditioning variables, because the lags do not change the original (statistical) informational universe of discourse. It can be shown that extending the original linear regression (4) by including lags $(Z_{t-k}, k = 1, 2, \ldots, t-1)$ can be rationalized in the context of the PR approach as a legitimate respecification of the original model in an attempt to account for departures from the temporal independence assumption [5] above.

In the context of the PR approach, respecification (schematically) takes the form of tracing the results of the misspecification tests back to the reduction assumptions, using the relationship shown in Table 1, and then changing the reduction assumptions judiciously to account for the sources of detected departures to choose a more appropriate statistical model. Evidence of any form of misspecification is interpreted as suggesting that the postulated model (as a whole) is invalid and the modeller should (schematically at least) return to the 'drawing board' − the Haavelmo distribution. Testing the individual assumptions amounts to assessing the symptoms of an inappropriate choice and thus any form of respecification that deals with the symptoms is likely to be misleading; more often than not the commonly used TT 'patching up' of the original model does not even constitute a set of internally consistent set of probabilistic assumptions; see Spanos (1995a). In the context of the PR approach one traces the symptoms (departures from individual model assumptions) back to the source (reduction assumptions) and an alternative choice of reduction assumptions gives rise to an alternative (and hopefully a more appropriate) statistical model; see Spanos (1986, 1999).

Returning to the NLR model (6), let us assume that, after thorough misspecification testing, the source of the problem has been traced back to the reduction assumption of independence for $\{\mathbf{Z}_t\ t \in \mathbb{T}\}$. A possible respecification scenario suggests the replacement of the assumption of independence with that of Markov (ℓ) dependence, i.e. replace $D(\mathbf{Z}_1, \ldots, \mathbf{Z}_T; \varphi) \overset{\text{I}}{=} \prod_{t=1}^{T} D_t(\mathbf{Z}_t; \varphi_t)$, with:

$$D(\mathbf{Z}_1, \ldots, \mathbf{Z}_T; \varphi) \overset{\text{M}}{=} D(\mathbf{Z}_1; \phi_1) \prod_{t=2}^{T} D_t(\mathbf{Z}_t \mid \mathbf{Z}_{t-1}^{\ell}; \phi_t),$$

where $Z_{t-1}^{\ell}\ :+\ (Z_{t-1}, Z_{t-2}, \ldots, Z_{t-\ell})$. Inevitably one needs to replace the reduction assumption of ID with stationarity in order to allow for the homogeneity of the covariances. By changing the reduction assumptions of Normality, Independence and Identically Distributed for $\{\boldsymbol{Z}_t\, t \in \mathbb{T}\}$ to Normality, Markovness and Stationarity, the reduction (3) becomes:

$$D(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_T; \varphi) \overset{\mathrm{M}(\ell)\&\mathrm{S}}{=} D(\boldsymbol{Z}_1; \phi_1) \Pi_{t=2}^{T} D(\boldsymbol{Z}_t | \boldsymbol{Z}^{\bar{}\ell}_{t-1}; \phi), \forall(z_1, z_2, \ldots, z_T) \in \mathbb{R}_Z^{mt}$$

.

$$= D(\boldsymbol{Z}_1; \varphi_1) \Pi_{t=2}^{T} D(y_t | \boldsymbol{X}_t, \boldsymbol{Z}^{\ell}_{t-1}; \varphi_1) \cdot D(\boldsymbol{X}_t | \boldsymbol{Z}^{\ell}_{t-1}; \varphi_2) \tag{10}$$

This reduction gives rise to two well-known statistical models in econometrics.

First, imposing normality, the first line of the reduction gives rise to the well-known VAR (ℓ) model with a statistical Generating Mechanism (GM):

$$\boldsymbol{Z}_t = \boldsymbol{a}_0 + \Sigma_{k=1}^{\ell} \boldsymbol{A}_k^{\mathrm{T}} \boldsymbol{Z}_{t-k} + \boldsymbol{u}_t, \quad t \in \mathbb{T} \tag{11}$$

with ‘model assumptions’ specified in terms of $D(\boldsymbol{Z}_t | \boldsymbol{Z}^{0}_{t-1}; \varphi)$:

[1] Normality: $D(\boldsymbol{Z}_t | \boldsymbol{Z}^{0}_{t-1}; \phi)$ is Normal
[2] Linearity: $E(\boldsymbol{Z}_t | \sigma(\boldsymbol{Z}^{0}_{t-1})) = \mathbf{a}_0 + \Sigma_{k=1}^{\ell} \boldsymbol{A}_k^{\mathrm{T}} \boldsymbol{Z}_{t-k}$, linear in $\boldsymbol{Z}^{\ell}_{t-1}$,
[3] Homoskedasticity: $Cov(\boldsymbol{Z}_t | \sigma(\boldsymbol{Z}^{0}_{t-1})) = \Omega$, free of $\boldsymbol{Z}^{0}_{t-1}$,
[4] t-homogeneity: $(\alpha_0,. \boldsymbol{A}_1, \ldots, \boldsymbol{A}_\ell, \Omega)$ are not functions of $t \in \mathbb{T}$,
[5] Martingale difference: $\{(\boldsymbol{u}_t \mid \boldsymbol{Z}^{0}_{t-1}),\ t \in \mathbb{T}\}$ is a martingale difference process. (12)

Second, the second line in (10), which involves a further reduction, gives rise to the Dynamic Linear Regression (DLR) model statistical GM:

$$y_t = \alpha_0 + \alpha_1^{\mathrm{T}} \boldsymbol{x}_t + \Sigma_{k=1}^{\ell} \gamma_k^{\mathrm{T}} \boldsymbol{Z}_{t-k} + \varepsilon_t, \quad t \in \mathbb{T}, \tag{13}$$

with model assumptions relating to the conditional distribution $D(y_t | \boldsymbol{X}_t, \boldsymbol{Z}^{\ell}_{t-1}; \varphi_1)$, which are analogous to (12); see Spanos (1986: ch. 23) for further details. It should be noted that in actual empirical modelling this respecification procedure is only schematic in the sense that these results need to be derived once; subsequent modellers can then simply conjecture what reduction assumptions are plausible in his/her case. For instance, if the normal distribution is replaced with the student’s *t* in (3), the resulting model is the student’s *t* Linear Regression model with quadratic heteroskedasticity; see Spanos (1994).

4.4.1 Adding lags: warranted data mining

Returning to our original question of respecification, one can see how a (statistical) respecification of the NLR model (4) gives rise to the DLR model (13), under certain conditions. Such a respecification is called for by statistical adequacy considerations and there is nothing ad hoc about it on statistical grounds. It's considered ad hoc from the viewpoint of the TT approach because the statistical model is viewed narrowly as coinciding with the theoretical model and respecification is (misleadingly) considered a theoretical issue. Having respecified the NLR model (4) into the DLR model (13), the modeller cannot proceed on the assumption that the latter is statistically adequate; this has to be established by testing its assumptions which are different from those of the NLR model, for possible departures.

4.5 How to data mine if you must!

In conclusion, the unwarranted data mining activity of appending and/or omitting regressors until a certain specification looks acceptable on certain grounds (R^2, t-ratios, signs and magnitude of the estimated coefficients) but ignoring statistical adequacy, often amounts to a sequence of (statistically) arbitrary decisions based on non-comparable entities and is usually devoid of any formal statistical justification. As argued above, this data mining activity is often a symptom of the substantial gap between theory concepts and observed data. In such cases it can be effectively dealt with by asking the modeller to justify his/her choice of the observed data *vis-a-vis* their connection to the theoretical concepts in terms of which the theory model is defined. Instead, this gap is often misinterpreted as allowing the modeller the leeway to indulge in the data mining activities of (a) searching through a set of potential data series with different sample characteristics; and (b) appending and/or omitting certain data series. This is a poor substitute for the real problem of bridging the gap between theory concepts and observed data by addressing the issue of 'what is the ideal data?' for the theory in question and 'how the available data differ from the ideal?', as well as 'how that difference can be bridged?'. It is no accident that very few applied papers (if any) provide convincing arguments concerning their choice of the particular data set; and there is nothing coincidental about the practice of 'passing the bucket' on the choice of the data to a previously published paper.

In cases where the theory is indeed very vague as to the relevant explanatory variables, the modeller should be explicit about the objective of the empirical search for potentially relevant variables and try to do a good job of the search. In such cases the modeller has, often, little option but to indulge in data mining activities and it is imperative to follow a more systematic procedure by taking into consideration the following statistical issues: (a) do the search in a systematic way by nesting all statistical models of interest into a comprehensive model; (b) pay particular attention to the statistical adequacy of the comprehensive and any selected sub-model; and (c) keep track of the actual significance level in sequential and multiple testing.

5 AD HOC MODIFICATIONS VERSUS PROPER RESPECIFICATION

5.1 Changing the functional form of a model

In empirical modelling the gap between the theory concepts and the observed data is often interpreted as allowing the modeller some leeway to try different functional forms in an attempt to find something that ‘fits’ the data in some sense. This is an unwarranted data mining activity, which can be viewed as ad hoc modification to save a theoretical model from anomalous evidence. Theory T suggests a theoretical model M, but if the data do not fit M (and thus T), the modeller searches the data so as to design $M'(Z)$ which gives a good fit with data Z. Viewing this in light of the definition of unwarranted data mining, one can see that it constitutes a case where an accordance between data Z and a proposition $M'(Z)$ (an ad hoc modification of M) is sought in such a way that the latter has not passed a severe test for Z, i.e. the probability of so good a fit, under the assumption that $M'(Z)$ is false, is very high.

5.1.1 Theoretical versus statistical relationships: the PR perspective

A cursory look at empirical modelling in econometrics reveals that the overwhelming majority of theoretical models are linear equations but the bulk of estimated empirical models are log-linear (multiplicative in the original variables). How can one explain this discrepancy? It is true that linearity is often a convenient fiction and thus when a theoretical model is specified in terms of a linear equation, there is some leeway to consider linearity as a first approximation of a possibly more complicated relationship. This by itself, however, does not justify the ad hoc modification of the functional form to ‘fit’ the data Z without ensuring that the claim has passed a severe test for data Z. In an attempt to ensure reliability and severity, the PR approach distinguishes between theoretical and statistical relationships with the former defined in terms of the theory concepts but the latter in terms of the observable random variables underlying the observed data. For simplicity let us assume that the relationship between the theoretical variables (ς, ξ) takes the form:

$$\varsigma = a_0 + a_1\xi, \quad \varsigma \in \mathbb{R}_\varsigma, \quad \xi \in \mathbb{R}_\xi, \tag{14}$$

where the coefficients (a_0, a_1) enjoy a clear theoretical interpretation. In contrast, a statistical relationship, such as a regression function between two observable random variables (Y_t, X_t), defined by:

$$E(Y_t \mid X_t = x_t) = h(x_t), x_t \in \mathbb{R}_X, \tag{15}$$

is a very different entity. Their differences stem from the fact that the form of (14) is determined by the theory information but that of (15) is determined exclusively by the statistical information contained in the observed data $(z_1, z_2, \ldots, z_T)$, where $z_t := (y_t, x_t)$. More precisely, the form of the joint distribution $D(Y_t, X_t; \varphi)$ determines the functional form of (15) via the conditional distribution defined by:

$$D(Y_t \mid X_t; \phi) = \left[\int_{y_t \in \mathbb{R}_Y} D(Y_t, X_t; \varphi)\, dy_t\right]^{-1} D(Y_t, X_t; \varphi);$$

see Spanos (1986: pp. 124–5) for several examples. For instance, in the case where $D(Y_t, X_t; \varphi)$ is normal:

$$D(Y_t \mid X_t = x_t) = \int_{y_t \in \mathbb{R}} y_t D(Y_t \mid X_t; \phi)\, dy_t = \beta_0 + \beta_1 x_t,\ x_t \in \mathbb{R}_X. \tag{16}$$

It is important to emphasize that the statistical parameters have a clear probabilistic interpretation of the form:

$$\beta_0 = E(Y_t) - \beta_1 = E(X_t),\ \ \beta_1 = \frac{Cov(Y_t, X_t)}{Var(X_t)}.$$

In the case where the joint distribution is exponential with parameter :

$$E(Y \mid X = x) = \frac{(1+\theta+\theta x)}{(1+\theta x)^2},\ \ Var(Y \mid X = x) = \frac{\{(1+\theta+\theta x)^2 - 2\theta^2\}}{\{1+\theta x\}^4},\ \ x \in \mathbb{R}_+,\ \ \theta > 0.$$

Hence, the functional form in the context of a statistical model is a statistical adequacy issue which depends crucially on the nature of the joint distribution $D(y_t, X_t; \varphi)$ and can only be decided on statistical information grounds using judicious graphical techniques (such as scatter plots) and comprehensive misspecification testing; see Spanos (1999: pp. 316–34). No theory, however sophisticated, can rescue a statistically misspecified model; misspecification is, by its very nature, a departure from probabilistic assumptions and the 'cure' can only be based on statistical information; whether the statistical misspecification can be rationalized in terms of some theory is a very different issue. Ideally, the modeller would like (15) to coincide with (14) but the modelling should allow for the possibility that they do not. Comprehensive misspecification testing, which includes tests for possible departures from the postulated functional form, ensures that the claim has passed a severe test for data *Z*.

5.2 The selection of the error term assumptions

5.2.1 The TT approach

Another important implication of ignoring the gap between theoretical concepts and the observed data, is to presume that the theoretical and statistical models can only differ by a white-noise error term. For instance, in the case of the simple linear regression model:

$$y_t = \beta_0 + \beta_1 x_t + u_t,\ t \in \mathbb{T}, \tag{17}$$

where the error process $\{u_t, t \in \mathbb{T}\}$ satisfies the assumptions:

$$\left.\begin{array}{lll} (1) & \text{zero mean:} & E(u_t) = 0, \\ (2) & \text{constant variance:} & E(u_t^2) = \sigma^2, \\ (3) & \text{no autocorrelation:} & E(u_t u_s) = 0,\ t \neq s, \\ (4) & \text{normality:} & u_t \sim \mathrm{N}(.\,,.), \end{array}\right\} \quad t, s \in \mathbb{T}. \tag{18}$$

In cases where the observed data indicate departures from these assumptions (e.g. low Durbin-Watson (DW) statistic), the traditional econometrician feels free to change the error assumptions and, as long as the systematic component $(\beta_0 + \beta_1 x_t)$, remains the same, the modelling is considered legitimate because it is confined within the boundaries demarcated by the theory. The modification of the error assumptions amounts to changing the original model by allowing a different probabilistic structure for the error. The most popular such modification is to replace (3) with the AR(1) formulation:

$$u_t = \rho u_{t-1} + \varepsilon_t,\ |\rho| < 1,\ \varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2),\ t \in \mathbb{T}. \tag{19}$$

This is considered as a ‘solution’ to the original misspecification problem detected by the DW test. Moreover, the same TT modeller who considers this modification legitimate, often accuses the adherents to the general-to-specific procedure, who, in the case of a low DW statistic, replace the original model with the Dynamic Linear Regression (DLR(1)) model:

$$y_t = \alpha_0 + \alpha_1 x_t + \alpha_2 x_{t-1} + \alpha_3 y_{t-1} + \varepsilon_t,\ t \in \mathbb{T}, \tag{20}$$

as indulging in data mining.

5.2.2 The PR perspective

When this problem is viewed in the context of the PR approach, several things become apparent. First, adopting the alternative in a misspecification test as a ‘cure’ for the detected departure, without any further testing, constitutes another form of unwarranted data mining because it clearly violates severity; this is elaborated on in the next section. Second, it turns out that, under certain conditions discussed above, (20) constitutes a legitimate respecification of the statistical model (17), whose appropriateness could be established via misspecification testing. Third, (19) constitutes a special case of (20), since by substituting out u_{t} yields the statistical GM:

$$y_t = \beta_0 (1 - \rho) + \beta_1 x_t - \rho\beta_1 x_{t-1} + \rho y_{t-1} + \varepsilon_t, \; t \in \mathbb{T}, \quad (21)$$

which is a special case of (20) under the common factor restrictions:

$$\alpha_2 + \alpha_1\alpha_3 = 0, \quad (22)$$

(see Hendry and Mizon 1978). As shown in Spanos (1987) this restriction is highly unlikely to be valid in practice because it requires that the stochastic processes $\{y_t\, t \in \mathbb{T},\}$ and $\{X_t\, t \in \mathbb{T},\}$ have almost identical temporal structures. In this sense the TT modeller restricts the statistical specification in a way which often makes matters worse. What is more important for statistical purposes, the same modeller considers that departures from assumption (3) have only a minor effect on the properties of the OLS estimators of (β_0, β_1), since it retains both unbiasedness and consistency. As shown in Spanos (1986), this argument is valid only when the departure from (3) is of the form (19); i.e. the common factor restrictions (22) do hold. In cases where the common factor restrictions are not valid, the OLS estimators are both 'biased' and 'inconsistent'; see Spanos (2000).

5.2.3 Ad hoc modifications to save a model: unwarranted data mining

Viewing the TT modification of the error assumptions in an attempt to find a model that fits the data in the context of the definition of unwarranted data mining, it is apparent that it constitutes another example of ad hoc modification of the statistical model, which fails to ensure that this claim has passed a severe test. Rejecting the null 'no autocorrelation' assumption with data *Z* via the DW test is taken, not just as a rejection of the null, but as evidence for a specific alternative (21). In the context of the PR approach, severity is ensured by viewing the DLR(1) model as a possible statistical respecification of (16) (in an attempt to account for the temporal dependence when the error autocorrelation assumption (3) above is invalid), which has to be tested as part of misspecification testing for the DLR(1) model. Let us consider this in some more detail by shedding some light on the nature of testing.

6 MISSPECIFICATION TESTING: THE PR APPROACH

In the context of the TT approach diagnostic testing and data snooping are considered as data mining activities, which raise fundamental issues of pretest bias, multiple testing and the overall significance level. A typical view of misspecification testing from the TT approach perspective is given by Kennedy (1998):

> Extensive use of diagnostic tests/checks is not universally applauded. Goldberger (1986) claims a recent empirical study reported more

> diagnostic tests statistics than number of observations in the data set. (p. 88)

Kennedy (1998) goes on to list a number of complaints and warnings concerning diagnostic testing including:

> (2) it may be replacing one kind of data mining with another; . . .
> (6) sequences of tests distort things like the probability of a type I error;
> (7) most of the tests used are not independent of one another;
> (8) the properties of pre-test estimators are not well understood.
> (p. 89)

In this section, it is argued that when diagnostic testing is viewed as proper misspecification testing in the context of the PR approach, none of these problems arise. The key to unraveling the confusion concerning misspecification testing lies with the nature of such testing as it differs from the 'predesignationist' NP testing procedure. In the context of the PR approach two types of tests are distinguished: tests of primary (theory-based) hypotheses and misspecification tests. For instance, a *t*-test for the significance of a coefficient in a regression constitutes a test of a primary hypothesis but the DW test for error autocorrelation is a misspecification test. These two types of tests are both different in nature and as well as in their claims. As argued below, Kennedy's warnings (6)–(8) suggest that there is a confusion in the minds of the critics of misspecification testing with regard to the error of concern in this context; the crucial error is not that of type I.

6.1 The nature of misspecification testing

With the exception of the adoption of the alternative associated with the DW tests, the data mining activities discussed so far are concerned with arriving at and 'passing' (primary) inferences or models using a particular set of data and should be contrasted with procedures for checking the validity or adequacy of the postulated model itself. The cardinal objective of misspecification testing, in the context of the PR approach, is to ensure the reliability of the overall inference, such as testing the primary hypotheses, by ensuring that the postulated statistical model is statistically adequate for the data in question. In misspecification testing, one is also reaching claims but they will be assertions about how well the actual realized data accord with the various assumptions of the statistical model, upon which the (primary) statistical inference is based. Since misspecification testing concerns questions about the realized data, only aspects of the realized data can serve to answer them and thus violating use-novelty is necessary. Given that the violation of use-novelty is necessary for misspecification testing, the question arises: if one uses the data to search for inadequacies or to arrive at a more adequate characterization of the data that one is engaged in unwarranted data mining? The answer is 'no' in cases where the search is accomplished without violating severity.

6.1.1 Misspecification testing versus the NP procedure

As shown in Spanos (1998, 1999), misspecification testing differs from the NP procedure in one important respect: misspecification accords with the Fisher approach to testing. How is the latter different from the NP testing? In the familiar NP formulation, the null and alternative hypotheses take the form:

$$H_0 : f(x) \in \Phi_0 \text{ against } H_1 : f(x) \in \Phi_1, \Phi = \Phi_0 \cup \Phi_1, \Phi_0 \cap \Phi_1 = \varnothing, \qquad (23)$$

where Φ denotes the postulated statistical model. Primary hypotheses can be formulated as restrictions on the parameter space of and thus they can be tested using the NP procedure. In contrast, misspecification testing is concerned with establishing the adequacy of the postulated model itself and thus the alternative hypothesis is by definition the non-null, i.e.:

$$H_0 : f(x) \in \Phi \text{ against } \bar{H}_0 : f(x) \in \mathcal{P} - \Phi, \qquad (24)$$

where $\mathcal{P}$ denotes the set of all possible statistical models that can be specified in terms of the sample $\boldsymbol{X}$: = $(X_1, X_2, \ldots, X_n)$ underlying the observed data $\boldsymbol{x}$: = $(x_1, x_2, \ldots, x_n)$. For example, in the case of the LR model as specified in (6), misspecification testing amounts to assessing the appropriateness of assumptions [1]–[5] by probing beyond the boundaries of the model as demarcated by these assumptions. This suggests that NP testing is testing 'within' and misspecification testing is testing 'without' the boundaries of the postulated model Φ. An important implication of this is that, assuming that is statistically adequate for the data $\boldsymbol{x}$, the null and the alternative hypotheses exhaust all possibilities. Without the statistical adequacy, however, this is no longer true because both hypotheses can be false. Hence, in the context of the PR approach the statistical adequacy of the postulated model should be established before testing the primary hypothesis of interest; otherwise the reliability of the NP test is not assured. When a TT modeller pronounces that the estimated coefficients are significant and have the expected signs/magnitudes, s/he utilizes statistical inference arguments whose validity should be established first.

In addition to the difference in testing being within and without the postulated statistical model, the NP and the Fisher procedures also differ with respect to their objective. Fisher testing is inferential in nature in the sense that the end result concerns the level of accordance of the observed data with the statistical model described by the null hypothesis; more specifically, it's based on a 'measure of accordance' between the sampling distribution of a test statistic $\tau(X)$ under H_0, say $f(\tau; H_0)$ and the observed value $\tau(x)$. One such measure is the p-value, which takes the form of the tail probability:

$$\mathbb{P}(\tau(X) \geq \tau(x); H_0 \text{ is valid}) = p. \qquad (25)$$

In the context of the PR approach, misspecification testing is viewed as primarily ‘destructive’ in nature, in the sense that the inference sought concerns indications of departures from the null hypothesis being tested. Hence, no questive of adopting the non-null hypothesis, when such departures are detected, arises. This is because the generality of the non-null $\bar{H}_0 : f(x) \in \mathcal{P} - \Phi$ necessitates probing the alternative models in $\mathcal{P} - \Phi$ before such a construction inference can be drawn. A misspecification test differs from a NP test in so far as the non-null is not a choice because it does not constitute a proper statistical model; it’s the set of all possible alternative models, which is often infinite. This issue becomes even more apparent in the case where the misspecification testing is performed in a piece-meal fashion by testing individual or groups of assumptions making up the postulated statistical model. Typically, misspecification tests probe in specific directions, as determined by the implications of a particular form of alternative from the null, seeking to establish the presence of such a type of departure. In view of the fact that $\bar{H}_0$ can take a (possibly) infinite number of forms, deriving a test requires the modeller to provide a more restrictive (operational) form, say $\bar{H}_0(h)$, where $\bar{H}_0(h) \subset \bar{H}_0$ but $\bar{H}_0 - \bar{H}_0(h) \neq \varnothing$. Detection of departures from the null in the direction of $\bar{H}_0(h)$ is sufficient to consider the null as false but not to consider $\bar{H}_0(h)$ as true. For example, in the case of the NLR model (assumptions [1]–[5]) a misspecification test of the linearity assumption [2] can be tested using:

$$H_0 : E(Y_t \mid X_t = x_t) = \beta_0 + \beta_1 x_t, \qquad \bar{\mathrm{H}}_0\!: E(Y_t \mid X_t = x_t) = \alpha_0 + \alpha_1 x_t \; + \alpha_1 x_t^2$$

Detecting a departure from H_0 in the in the direction of $\bar{H}_0(h)$ does not call for the conclusion that $\bar{H}_0(h)$ is true; it suggests only that there is evidence of a discrepancy from H_0. Indeed, if the ‘true’ regression function is, say:

$$H_* : E(Y_t \mid X_t = x_t) = \frac{(1+\theta+\theta x_t)}{(1+\theta x_t)^2}\;, H_* \in [\bar{H}_0 - \bar{H}_0(h)],$$

the misspecification test based on $\bar{H}_0(h)$ is likely to detect departures from H_0, despite the fact that $\bar{H}_0(h)$ differs greatly from the true regression function. This is because $\bar{H}\,_0(h)$ lies beyond the postulated statistical model in the direction of H_*. Hence, when the null hypothesis is rejected, (without any further testing) the modeller can only infer that the postulated statistical model is misspecified because it does not account for the systematic information in the direction that the particular test is probing. The probability that $\bar{H}_0(h)$ fits Z, even though $\bar{H}_0(h)$ is false, might be very high and not low as severity demands.

6.1.2 Error-autocorrelation and the TT approach revisited

Returning to the TT approach, one can consider the appropriateness of assumption (3) (of the linear regression model) using the misspecification testing formulation:

$$H_0 : E(u_t u_s) = 0, t \neq s, \quad \text{against} \quad \bar{H}_0 : E(u_t u_s) \neq 0, t \neq s, t, s \in \mathbb{T}. \tag{26}$$

In the above case, to operationalize $\bar{H}_0$ the TT approach uses parametric models for the error, such as the AR(1) (19), which reformulates the problem into:

$$H_0(h) : \rho = 0, \quad \text{against} \quad \bar{H}_0(h) : \rho \neq 0, \text{ where } |\rho| < 1. \tag{27}$$

Evidence against the null, based on a low DW test statistic, cannot be considered as sufficient to conclude that the true temporal dependence is of the form (19). This is because the set $[H_0(h) \cup \bar{H}_0(h)]$ does not exhaust all possibilities; the illusion that it does, arises from the fact that it's viewed as testing within. This is a clear case of unwarranted data mining because the decision to adopt the alternative did not pass a reliable or severe test.

The primary goal of misspecification/respecification in the context of the PR approach is finding a statistically adequate model and so the error of concern at this point is that data *Z* will be taken as indicating model *M* when in fact *M* has not passed a severe or probative test. The inference relating to (21) had no chance of uncovering the ways it could be in error. So, in general it has this flaw, but also, in reaching the specific model (21), it is seen that there is a high probability of reaching an erroneous attribution of *Z*. The DLR model (20), by contrast, is arrived at after passing its own misspecification tests, which can be said to have had its possible errors well-probed and found absent. A false respecification would mean, not statistically adequate, or one that fails to capture the statistical information in the data adequately.

6.2 Misspecification testing and data mining

The question that naturally arises at this point is: whether the utilization of several misspecification tests for assessing the validity of each of the assumptions comprising the postulated statistical model, constitutes another form of unwarranted data mining? The quotation from Kennedy (1998) above suggests that the sequential and multiple nature of misspecification testing renders the actual significance level very different from the nominal.

The basic idea behind misspecification testing, in the context of the PR approach, is that the modeller would like to argue that, on the basis of the test results, statistical adequacy is established when a comprehensive misspecification testing is applied and no departures from the underlying assumptions were detected, despite a highly probative search. This enables the modeller to infer that, in cases where the tests have a very high probability of detecting the departures if they were present, the negative misspecification test results provide strong evidence for the absence of any such departures. Moreover any additional checks which agree with the original finding can only fortify (not weaken) the original evidence; see Mayo (1996: pp. 184–5).

Given that the primary objective of misspecification testing is (inferentially) 'destructive', the problem facing the modeller is to ensure that the battery of the tests applied is effective in detecting departures from H_0: the postulated model is statistically adequate for the observed data $\mathbf{Z}$. How does a modeller guard against the different types of errors? A misspecification test can err in two ways. Type I error (reject H_0 when valid), the test is hypersensitive in directions of departures very close to H_0, or to spurious departures. Type II error (accept H_0 when invalid), the test probes in directions different from those of the actual departures in the data.

Given that a misspecification test is considered 'good' if it can discriminate H_0 from $\bar{H}_0$ effectively, the modeller can guard against both types of error by ensuring that the probability of detecting any departures if present is very high. Keeping in mind that the results of the misspecification tests constitute a set of overlapping evidence concerning the validity of the postulated statistical model as a whole, this can be accomplished by following two complimentary strategies:

(i) Utilize a comprehensive battery of misspecification tests probing in as many potential directions of departure as possible; and
(ii) Exploit any reliable information concerning the direction of possible departures.

In contrast to primary hypothesis testing using the NP procedure, both types of error are reduced by utilizing more not less tests. Each misspecification test has its own 'null' and 'implicit alternative hypotheses' with which it constitutes an optimal test (in terms of power or sensitivity). The implicit alternative constitutes a specific direction of departure from the null. In the absence of information as to the nature of possible departures from the assumption being tested, the modeller should consider different misspecification tests in order to probe the observed data for possible departures in different directions, as well as ensure that any detected departures are not artifacts but systematic statistical information. Different tests are often derived using alternative probabilistic assumptions and they usually enjoy robustness with respect to different departures. Because of this, the use of a mixture of parametric and non-parametric tests is strongly encouraged in this context.

A moment's reflection suggests that the type II error is the most serious of the two. The application of several misspecification tests for each assumption ensures that if a certain test detects a departure that the other tests do not confirm, then further probing will determine whether it's real or an artifact. In Spanos (1992), a case is made for joint misspecification tests as an effective way to deal with 'spurious' departures; see McGuirk *et al.* (1993) for an application. Moreover, if the original statistical model is erroneously rejected as misspecified, the modeller should proceed to respecify (choose another statistical model) and test the assumptions of the new model and keep repeating that process until a statistically adequate model is found. In doing so, the modeller is likely to find out that the original choice deserves re-examination.

In the context of the PR approach, the most reliable information for (ii) is provided by judicious graphical techniques in conjunction with the relationship between the reduction and model assumptions. For instance, in the case of the NLR model (4), one can use Table 1 in conjunction with *t*-plots and scatter plots of the observed data in order to assess the reduction assumptions directly and thus trace any departures to the model assumptions. If, for instance, there is evidence that the data might be leptokurtic (departures from normality) the modeller will know to expect departures in model assumptions [1]–[3] which are likely to take specific forms; see Spanos (1994, 1999). Misspecification testing is rendered more efficient when combined with a judicious utilization of graphical techniques (data snooping) which are designed to detect possible directions of departures and guide the modeller in his/her choice of appropriate tests.

Returning to the original question 'whether the utilization of several misspecification tests constitutes unwarranted data mining?', it is apparent that the answer is definitely 'no'! An affirmative answer is tantamount to arguing that the diagnosis based on a sequence of medical tests (blood pressure and temperature, blood test, urine test, culture (bacteria) test) is less reliable because so many tests were performed! What is relevant in the context of misspecification testing is the smallest *p*-value (the observed significance level) of the test that detects the presence of a departure. Hence, common charges of unwarranted data mining levelled against misspecification testing are misplaced, and result from confusion over the error of concern.

The above discussion of misspecification testing in the context of the PR approach also renders the 'pre-test bias argument' misplaced. The argument attempts to formalize a situation where the result of a test will give rise to a choice between two different estimators, one estimated under the null and the other under the alternative. One can grant this is a sensible thing to do in a NP testing procedure. However, as argued above, in a misspecification test the alternative is not considered as an option and thus the formalization is not relevant in this context.

It is important to emphasize that in defending misspecification testing against charges of unwarranted data mining, there is no claim that one has overcome all of the problems and issues raised by such testing. There is still important work to be done to make explicit the departures that particular misspecification tests are capable of detecting, to check and fortify assumptions of the tests themselves; in short to set out a fully adequate and reasonably complete repertoire of misspecification testing. One can embark upon this research with clear directions, however, only when one has freed the project of charges of being illicit; see Mayo and Spanos (1999).

6.3 Data snooping: a warranted data mining activity

Once more, the above discussion suggests that in the context of the PR approach there is nothing reprehensible in using graphical techniques in either misspecification testing or respecification.

In misspecification testing the primary role of data snooping is to contribute to the judicious operationalization of the non-null $\bar{H}_0 : f(x) \in \mathcal{P} - \Phi$. The modeller is utilizing graphical techniques to render the probing for the presence or absence of predesignated statistical information more efficient. Moreover, the presence of such systematic (statistical) information cannot be fabricated post hoc because the data do or do not contain such information, irrespective of the activities of the modeller. It is important to stress that the role of data snooping in the context of misspecification testing is the same for both observational and experimental data; see Mayo (1996).

In the context of the PR approach using observational data, the cycle specification, misspecification testing, respecification until a statistically adequate model is found (Figure 2), can be markedly more effective by exploiting graphical techniques which can help narrow down the set of all possible models $\mathcal{P}$ considerably; see Spanos (1999: ch. 5–6). Graphical techniques can be utilized in conjunction with the relationship between the reduction and model assumptions, in order to render the respecification facet an informed procedure and not a 'hit-or-miss' affair. Having made an educated conjecture as to which statistical model might be appropriate, the modeller will then proceed to test the statistical adequacy of the chosen model in order to assess the validity of that conjecture; ensuring that the conjecture has 'passed' a severe test.

7 CONCLUSION

The basic objective of this paper has been to reconsider a number of activities that are often interpreted as data mining in the context of the TT approach, from the methodological perspective of the PR approach. Armed with the notion of severity (see Mayo 1996), one was able to distinguish between problematic and non-problematic cases of data mining. Some of the data mining activities were reinterpreted and legitimized in the context of the PR approach by demonstrating that severity is, indeed, adhered to. The selection of the observed data as well as the relevant explanatory variables are considered as questions primarily pertaining to bridging of the gap between the theory and the observed data. The issue of selecting a functional form for a statistical model as well as the probabilistic assumptions underlying the error term, are problems that concern primarily the notion of statistical adequacy; in the context of the PR approach, the latter is concerned exclusively with capturing all the statistical systematic information in the observed data. Finally, the charges of data mining arising from applying several misspecification tests in conjunction with data snooping are misplaced because misspecification testing is very different in nature from NP testing. What is relevant in misspecification testing is the smallest p-value of the test that detects departures from the assumptions of the statistical model.

Viewing the scepticism exhibited by theorists concerning the trustworthiness of empirical evidence in applied econometrics from the PR perspective suggests that, although certain data mining activities do contribute to the uniformativeness of the empirical findings, the crucial problem is the 'unreliability' of such evidence due to the fact that the overwhelming majority of estimated empirical models are 'statistically inadequate'. This precludes any serious dialogue between empirical evidence and theories and the theorists can (and should) ignore unreliable evidence without remorse. Hence, the ball is squarely in the econometricians court to adopt modelling procedures which yield reliable empirical evidence; the PR approach has been formulated with that objective in mind. If followers of the TT approach agree with the goal of reliable inference identified above, then the challenge for them is to show how they can avoid violations of severity and statistical adequacy when they use their procedures.

Aris Spanos
Virginia Tech,
aris@vt.edu

ACKNOWLEDGEMENTS

I am most grateful to Deborah G. Mayo for numerous constructive criticisms and suggestions.

NOTES

[1] Note that severity does not coincide with the notion of power; see Mayo (1996: ch. 6, 11).

[2] This definition has been suggested by Mayo in a private correspondence.

[3] Our intuitions here are that mere accordance is not sufficient, if such an accordance is guaranteed whether or not the model is invalid. Those who deny reliability matters, however, will be untroubled by violations due to hunting and data mining.

[4] This theory-data gap arises in all empirical modelling, not just in economics. Einstein's 1905 theoretical model explaining the Brownian motion was so successful because he formulated his theory in terms of the observable 'mean displacement' and not the unobservable 'velocity' as previous attempts to model the motion of a particle in a liquid (see Perrin 1913: pp. 109–15).

REFERENCES

Granger, C.W.J. (1990) *Modelling economic series: readings on the methodology of econometric modelling*, Oxford: Oxford University Press.

Granger, C.W.J. and Newbold, P. (1974) 'Spurious regressions in econometrics', *Journal of Econometrics* 2: 111–20.

Gujarati, D.N. (1995) *Basic Econometrics*, 3rd ed, New York: McGraw-Hill.

Hendry, D.F. (1993) *Econometrics: alchemy or science?*, Oxford: Blackwell.

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

Hendry, D.F. and Mizon G.E. (1978) 'Serial correlation as a convenient simplification not a nuisance: a comment on a study of the demand for money by the Bank of England', *Economic Journal* 88: 549–63.

Hoover, K.D. and Perez, S.J. (2000) 'Three Attitudes towards Data-mining', *Journal of Economic Methodology* 7: 195–210.

Intriligator, M.D. (1978) *Econometric Models, Techniques and Applications*, Amsterdam: North-Holland.

Jonhston, J. and Dinardo J. (1997) *Econometric Methods*, 4th edn, New York: McGraw-Hill.

Kennedy, P. (1998) *A guide to Econometrics*, 4th edn, Cambridge MA: The MIT Press.

Kragh, H. (1999) *Quantum generations*, New Jersey: Princeton University Press.

Leamer, E. (1978) *Specification searches: ad hoc inference with nonexperimental data*, New York: Wiley.

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

Lovell, M.C. (1983) 'Data Mining', *The Review of Economics and Statistics* 65: 1–12.

Mayo, D.G. (1996) *Error and the growth of experimental knowledge*, Chicago: University of Chicago Press.

Mayo, D.G. and Spanos A. (1999) 'Self-correcting statistical inferences: understanding misspecification testing', mimeo, Virginia Tech.

Mayer, T. (2000) 'Data mining: a reconsideration', *Journal of Economic Methodology* 7: 183–94.

McGuirk, A., Driscoll, P. and Alwang, J. (1993) 'Misspecification testing: a comprehensive approach', *American Journal of Agricultural Economics* 75: 1044–55.

Mizon, G.E. and Richard, J.-F. (1986) 'The encompassing principle and its application to testing non-tested hypotheses', *Econometrica* 54: 657–78.

Perrin, J. (1913) *Atoms*, trans. D.L. Hammick, Woodbridge, Connecticut: Ox Bow Press.

Phillips, P.C.B. (1986) 'Understanding spurious regression in econometrics', *Journal of Econometrics* 33: 311–40.

Spanos, A. (1986) *Statistical foundations of econometric modelling*, Cambridge: Cambridge University Press.

——— (1987) 'Error autocorrelatin revisited: the AR(1) case', *Econometric Reviews* 6: 285–94.

——— (1988) 'Towards a unifying methodological framework for econometric modelling', *Economic Notes* 17: 107–34.

——— (1989) 'On re-reading Haavelmo: a retrospective view of econometric modeling', *Econometric Theory* 5: 405–29.

——— (1992) 'Joint misspecification tests in linear regression', mimeograph, Virginia Polytechnic Institute and State University.

——— (1994) 'On modeling heteroskedasticity: the student's *t* and elliptical regression models, *Econometric Theory* 10: 286–315.

——— (1995a) 'On normality and the liner regression model', *Econometric Reviews* 14: 195–203.

——— (1995b) 'On theory testing in econometrics: modeling with nonexperimental data', *Journal of Econometrics* 67: 189–226.

——— (1998) 'Econometric testing', in *The Handbook of Economic Methodology* edited by J.B. Davis, D.W. Hands and U. Maki (eds), Edward Elgar, pp. 116–30.

——— (1999) *Probability theory and statistical inference: econometric modeling with observational data*, Cambridge: Cambridge University Press.

——— forthcoming 'Time series and dynamic models', in B. Baltagi (ed) *Companion volume in Theoretical Econometrics* Oxford: Basil Blackwell.

Wallis, K. (1973) *Topics in Applied Econometrics*, London: Gray-Mills.