RMM Vol. 2, 2011, 146–178
Special Topic: Statistical Science and Philosophy of Science
Edited by Deborah G. Mayo, Aris Spanos and Kent W. Staley
http://www.rmm-journal.de/

Aris Spanos

Foundational Issues in Statistical Modeling: Statistical Model Specification and Validation[*]

Abstract:

Statistical model specification and validation raise crucial foundational problems whose pertinent resolution holds the key to learning from data by securing the reliability of frequentist inference. The paper questions the judiciousness of several current practices, including the theory-driven approach, and the Akaike-type model selection procedures, arguing that they often lead to unreliable inferences. This is primarily due to the fact that goodness-of-fit/prediction measures and other substantive and pragmatic criteria are of questionable value when the estimated model is statistically misspecified. Foisting one's favorite model on the data often yields estimated models which are both statistically and substantively misspecified, but one has no way to delineate between the two sources of error and apportion blame. The paper argues that the *error statistical* approach can address this Duhemian ambiguity by distinguishing between statistical and substantive premises and viewing empirical modeling in a piecemeal way with a view to delineate the various issues more effectively. It is also argued that Hendry's general to specific procedures does a much better job in model selection than the theory-driven and the Akaike-type procedures primary because of its error statistical underpinnings.

1. Introduction

A glance through the recent *International Encyclopedia of Statistical Science* (see Lovric 2010) reveals that there have been numerous noteworthy developments in frequentist statistical methods and techniques since the late 1930s. Despite these impressive technical advances, most of the foundational problems bedeviling the original Fisher-Neyman-Pearson (F-N-P) model-based approach to frequentist modeling and inference remained largely unresolved (see Mayo 2005).

Foundational problems, like the abuse and misinterpretations of the accept/reject rules, the p-values and confidence intervals, have created perpetual con-

[*] Paper presented at the conference on *Statistical Science & Philosophy of Science: Where Do/Should They Meet in 2010 (and Beyond)?* London School of Economics, June 21–22, 2010. Many thanks are due to my colleague, Deborah Mayo, for numerous constructive criticisms and suggestions that improved the paper considerably. Special thanks are also due to David F. Hendry for valuable comments and suggestions.

fusions in the minds of practitioners. This confused state of affairs is clearly reflected in the startling dissonance and the numerous fallacious claims made by different entries in the same encyclopedia (Lovric 2010).

In the absence of any guidance from the statistics and/or the philosophy of science literatures, practitioners in different applied fields, including psychology, sociology, epidemiology, economics and medicine, invented their own ways to deal with some of the more pressing foundational problems like statistical vs. substantive significance. Unfortunately, most of the proposed 'solutions' added to the confusion instead of elucidating the original foundational problems because they ended up misusing and/or misinterpreting the original frequentist procedures (see Mayo and Spanos 2011).

Along side these well-known foundational problems pertaining to frequentist inference, there have been several modeling problems that have (inadvertently) undermined the reliability of inference and largely derailed any learning from data. The two most crucial such problems pertain to *statistical model specification* [where does a statistical model $\mathcal{M}_\theta(\mathbf{z})$ come from?] and *model validation* [are the probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{z})$ valid for data $\mathbf{z}_0$:=$(z_1,...,z_n)$?]. These two foundational problems affect, not just frequentist inference, but all forms of modern statistical analysis that invoke the notion of a statistical model. It turns out that underlying every form of *statistical analysis* (estimation, testing, prediction, simulation) of a *scientific* (substantive) model $\mathcal{M}_\varphi(\mathbf{z})$ is a distinct *statistical model* $\mathcal{M}_\theta(\mathbf{z})$ (often implicit). Moreover, all statistical methods (frequentist, Bayesian, nonparametric) rely on an underlying statistical model $\mathcal{M}_\theta(\mathbf{z})$ whose form might be somewhat different. Statistical model validation is so crucial because the presence of statistical misspecification plagues the reliability of frequentist, Bayesian and likelihood-based (Sober, 2008) inferences equally badly. To be more specific, a misspecified $\mathcal{M}_\theta(\mathbf{z})$ stipulates an invalid distribution of the sample $f(\mathbf{z};\theta)$, and thus a false likelihood $L(\theta;\mathbf{z}_0)$, which in turn will give rise to erroneous error probabilities (frequentist), incorrect fit/prediction measures (Akaike-type procedures) and wrong posterior distributions $\pi(\theta|\mathbf{z}_0)$=$\pi(\theta)L(\theta;\mathbf{z}_0)$ (Bayesian).

Despite its crucial importance for the reliability of inference, statistical model specification and validation has been neglected in the statistical literature:

> "The current statistical methodology is mostly model-based, without any specific rules for model selection or validating a specified model." (Rao 2004, 2)

The main argument of the paper is that the *error statistical* approach addresses these two modeling problems by proposing specific rules for model specification and validation that differ considerably from the criteria used by traditional approaches like theory-driven modeling and Akaike-type model selection procedures. In particular, the key criterion for securing the reliability of inference is *statistical adequacy* [the probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{z})$ valid for data $\mathbf{z}_0$]. Goodness-of-fit/prediction and other substantive and pragmatic criteria are neither necessary nor sufficient for that.

Mayo (1996) has articulated an “error statistical philosophy” for statistical modeling in science/practice that includes the interpretation and justification for using the formal frequentist methods in learning from data. As she puts it, she is supplying the statistical methods with a statistical philosophy. What does this philosophy have to say when it comes to model specification and model validation? Where do the contemporary foundational debates to which Mayo, in this volume, calls our attention meet up with the issues of modeling? That is what I will be considering as a backdrop to my discussion of the meeting grounds of statistical and substantive inference.

To offer a crude road map of the broader context for the discussion that follows, a sum-up of the foundational problems that have impeded learning from data in frequentist inference is given in *section 2*. In *section 3* the paper offers a brief sketch of the common threads underlying current practices as they pertain to model specification and validation. *Section 4* uses the inability of the theory-driven approach to secure the reliability of inference to motivate some of the key elements of the error statistical approach. The latter approach distinguishes, *ab initio*, between *statistical* and *substantive premises* and creates the conditions for addressing the unreliability of inference problem, as discussed in *section 5*. This perspective is then used in *section 6* to call into question the reliability of models selected using Akaike-type model selection procedures. The latter procedures are contrasted with Hendry’s general to specific procedure that is shown to share most of its key features with the error statistical perspective.

2. Foundational Problems in Frequentist Statistics

Fisher (1922) initiated a change of paradigms in statistics by recasting the then dominating *descriptive statistics paradigm*, relying on large sample size (n) approximations, into a frequentist *model-based induction*, grounded on *finite sampling distributions* and guided by the relevant *error probabilities*.

The cornerstone of frequentist inference is the notion of a (parametric) statistical model is formalized in purely *probabilistic* terms (Spanos 2006b):

$$\mathcal{M}_\theta(\mathbf{z})=\{f(\mathbf{z};\theta),\ \theta\in\Theta\},\ \mathbf{z}\in\mathbb{R}^n_Z,\ \text{for}\ \theta\in\Theta\subset\mathbb{R}^m,\ m<n,$$

where $f(\mathbf{z};\theta)$ denotes the (joint) *distribution of the sample* $\mathbf{Z}$: $=(Z_1,...,Z_n)$. It is important to emphasize at the outset that the above formulation can accommodate highly complex models without any difficult. For the discussion that follows, however, we will focus on simple models for exposition purposes.

Example. The quintessential statistical model is *the simple Normal*:

$$\mathcal{M}_\theta(\mathbf{z}):\ Z_k \backsim \mathsf{NIID}(\mu,\sigma^2),\ \theta:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\ k\in\mathbb{N}:=(1,2,...n,...), \tag{1}$$

where ‘NIID’ stands for ‘Normal, Independent and Identically Distributed’ and $\mathbb{R}_+:=(0,\infty)$ denotes the positive real line.

The distribution of the sample $f(\mathbf{z};\theta)$ provides the relevant error probabilities based the sampling distribution $F_n(t)$ of any statistic $T_n{=}g(\mathbf{Z})$ (estimator, test or predictor) via:

$$F(t;\theta){:=}\mathbb{P}(T_n \leq t;\theta) = \underbrace{\int\int\cdots\int}_{\{\mathbf{z}:\ g(\mathbf{z})\leq t;\ \mathbf{z}\in\mathbb{R}^n_Z\}} f(\mathbf{z};\theta)d\mathbf{z}. \tag{2}$$

Example. In the case of the simple Normal model (1), one can use (2) to derive the following sampling distributions (Cox and Hinkley 1974):

$$(\overline{Z}_n{=}\tfrac{1}{n}\textstyle\sum_{k=1}^n Z_k) \backsim \mathsf{N}(\mu,\tfrac{\sigma^2}{n}),\quad (n{-}1)s^2{=}\textstyle\sum_{k=1}^n(Z_k{-}\overline{Z}_n)^2 \backsim \sigma^2\chi^2(n{-}1), \tag{3}$$

where '$\chi^2(n{-}1)$' denotes the chi-square distribution with $(n{-}1)$ degrees of freedom. In turn, (3) can be used to derive the sampling distribution(s) of the test statistic $\tau(\mathbf{Z}){=}\sqrt{n}(\overline{Z}_n{-}\mu_0)/s$ for testing the hypotheses:

$$H_0\text{:}\ \mu{=}\mu_0, \text{ vs. } H_1\text{:}\ \mu > \mu_0,\quad \mu_0 \text{ - prespecified},$$

$$\text{[i] } \tau(\mathbf{Z}) \overset{H_0}{\backsim} \mathsf{St}(n{-}1),\quad \text{[ii] } \tau(\mathbf{Z}) \overset{H_1(\mu=\mu_1)}{\backsim} \mathsf{St}(\delta_1;n{-}1),\ \delta_1{=}\sqrt{n}(\mu_1{-}\mu_0)/\sigma,\ \mu_1 > \mu_0,$$

where '$\mathsf{St}(n{-}1)$' denotes the Student's t distribution with $(n{-}1)$ degrees of freedom. The sampling distribution in [i] is used to evaluate both the type I error probability:

$$\mathbb{P}(\tau(\mathbf{Z}) > c_\alpha;\ H_0) = \alpha,$$

as well as the p-value: $\mathbb{P}(\tau(\mathbf{Z}) > \tau(\mathbf{z}_0);\ H_0) = p(\mathbf{z}_0)$, where $\mathbf{z}_0{:=}(z_1,...,z_n)$ denotes the observed data. The sampling distribution in [ii] is used to evaluate both the type II error probability:

$$\mathbb{P}(\tau(\mathbf{Z}) \leq c_\alpha;\ H_1(\mu{=}\mu_1)){=}\beta(\mu_1), \text{ for all } \mu_1 > \mu_0,$$

as well as the power of the test: $\mathbb{P}(\tau(\mathbf{Z}) > c_\alpha;\ H_1(\mu{=}\mu_1)){=}\pi(\mu_1)$, for all $\mu_1 > \mu_0$. The mathematical apparatus of frequentist statistical inference was largely in place by the late 1930s. Fisher (1922; 1925; 1935), almost single-handedly, created the current theory of 'optimal' point estimation and formalized significance testing based on the p-value reasoning. Neyman and Pearson (1933) proposed an 'optimal' theory for hypothesis testing, by modifying/extending Fisher's significance testing. Neyman (1937) proposed an 'optimal' theory for interval estimation analogous to N-P testing. However, its philosophical foundations concerned with the proper form of the underlying *inductive reasoning* were left in a state of muddiness (see Mayo 1996). The last exchange between these pioneers took place in the mid 1950s (see Fisher 1955; Neyman 1956; Pearson 1955) and left the philosophical foundations of frequentist statistics in a state of befuddlement, raising more questions than answers.

The foundational problems bedeviling frequentist inference since the 1930s can be classified under two broad categories.

A. Modeling

[a] *Model specification*: how does one select the prespecified statistical model $\mathscr{M}_\theta(\mathbf{z})$?

[b] the role of *substantive* (subject matter) *information* in statistical modeling (Lehmann 1990; Cox 1990),

[c] the nature, structure and role of the notion of a *statistical model* $\mathscr{M}_\theta(\mathbf{z})$, $\mathbf{z} \in \mathbb{R}^n_Z$,

[d] *model adequacy*: how to assess the adequacy of a statistical model $\mathscr{M}_\theta(\mathbf{z})$ *a posteriori*, and

[e] *model re-specification*: how to respecify a model $\mathscr{M}_\theta(\mathbf{z})$ when found *misspecified*.

B. Inference

[f] the role of *pre-data* vs. *post-data error probabilities* (Hacking 1965),

[g] safeguarding frequentist inference against:

(i) the *fallacy of acceptance*: interpreting accept H_0 [no evidence against H_0] as evidence for H_0; e.g. the test had low power to detect existing discrepancy,

(ii) the *fallacy of rejection*: interpreting reject H_0 [evidence against H_0] as evidence for a particular H_1; e.g. conflating statistical with substantive significance (Mayo 1996; Mayo and Spanos 2010; 2011), and

[h] a *frequentist interpretation of probability* that provides an adequate foundation for frequentist inference (Spanos 2011).

The present paper focuses almost exclusively on problems [a]–[e], paying particular attention to model validation that secures the error reliability of inductive inference (for extensive discussions pertaining to problems [f]–[h] see Mayo and Cox 2006; Mayo and Spanos 2004; 2006 and Spanos 2000; 2007). These papers are relevant for the discussion that follows because, when taken together, they demarcate what Mayo (1996) called the 'error statistical approach' that offers a unifying inductive reasoning for frequentist inference. This is in direct contrast to widely propagated claims like:

> "this statistical philosophy [frequentist] is not a unified theory rather it is a loose confederation of ideas." (Sober 2008, 79)

The *error statistical* perspective provides a coherent framework for all frequentist methods and procedures by bringing out the importance of the alternative forms of reasoning underlying different inference methods, like estimation (point and interval), testing and prediction; the unifying thread being the relevant error probabilities.

In this sense, error Statistics can be viewed as a refinement/extension of the Fisher-Neyman-Pearson (F-N-P) motivated by the call for addressing the above foundational problems (Mayo and Spanos 2011). In particular, error statistics aims to:

[A] *refine* the F-N-P approach by proposing a broader framework with a view to secure *statistical adequacy*, motivated by the foundational problems [a]–[e], and

[B] *extend* the F-N-P approach by supplementing it with a *post-data severity* assessment with a view to address problems [f]–[g] (Mayo and Spanos 2006).

In error statistics probability plays two interrelated roles. *Firstly*, $f(\mathbf{z};\theta)$, $\mathbf{z}\in\mathbb{R}^n_Z$ attributes probabilities to all legitimate events related to the sample $\mathbf{Z}$. *Secondly*, it furnishes all relevant error probabilities associated with any statistic $T_n{=}g(\mathbf{Z})$ via (2). Pre-data these error probabilities quantify the generic capacity of any inference procedures to discriminate among alternative hypotheses. That is, error probabilities provide the basis for determining whether and how well a statistical hypothesis—a claim about the underlying data generating mechanism, framed in terms of an unknown parameter θ—is warranted by data $\mathbf{x}_0$ at hand (see Spanos 2010d). Post-data error probabilities are used to establish the warranted discrepancies from particular values of θ, using a post-data severity assessment. For the error statistician probability arises, post-data, not to measure degrees of confirmation or belief in hypotheses, but to quantify how well a statistical hypothesis has passed a test. There is evidence for a particular statistical hypothesis or claim just to the extent that the test that passes such a claim with $\mathbf{x}_0$ is *severe*: that with high probability the hypothesis would *not* have passed so well as it did if it were false, or specific departures were present (see Mayo 2003).

3. Model Specification/Validation: Different Approaches

As observed by Rao (2004) in the above quotation, modern empirical modeling is model-based, but the selection and validation of these models is characterized by a cacophony of ad hoc criteria, including statistical significance, goodness-of-fit/prediction, substantive meaningfulness and a variety of pragmatic norms like simplicity and parsimony, without any discussion of the conditions under which such criteria are, indeed, appropriate and relevant.

A closer look at different disciplines reveals that empirical models in most applied fields constitute a blend of statistical and substantive information, ranging from a solely *data-driven formulation* like an ARIMA(p,d,q) model, to entirely *theory-driven formulation* like a structural (simultaneous) equation model (e.g. DSGE model) (see Spanos 2006a; 2009b). The majority of empirical models lie someplace in between these two extremes, but the role of the two sources of information has not been clearly delineated. The limited discussion of the role of the two sources of information in the statistical literature is often combined with the varying objectives of such empirical models [description, prediction, explanation, theory appraisal and policy assessment] to offer a variety of classifications of models in such a context (see Lehmann 1990; Cox 1990). These classifications, however, do not address the real question of interest: the respective roles of statistical and substantive information in empirical modeling and how they could be properly combined to learn from data.

Despite the apparent patchwork of different ways to specify, validate and use empirical models, there are certain common underlying threads and invoked criteria that unify most of these different approaches.

The most important of these threads is that if there is some form of substantive information pertaining to the phenomenon being modeled, the selected model should take that into account somehow. Where some of the alternative approaches differ is how one should take that information into account. Should such information be imposed on the data at the outset by specifying models that incorporate such information, or could that information be tested before being imposed? If the latter, how does one implement such a procedure in practice? If not the theory, where does a statistical model come from? It is also important to appreciate that substantive information varies from specific subject matter information to generic mathematical approximation knowledge that helps narrow down the possible models one should be entertaining when modeling particular phenomena of interest. For example, the broad ARIMA(p,d,q) family of models stems from mathematical approximation theory as it relates to the Wold decomposition theorem (see Cox and Miller 1968).

Another commonly used practice is to use goodness-of-fit/prediction as criteria for selecting, and sometimes validating models. Indeed, in certain disciplines such measures are used as the primary criteria for establishing whether the estimated model does 'account for the regularities in the data' (see Skyrms 2000).

The key weakness of the overwhelming majority of empirical modeling practices is that they do not take the statistical information, reflected in the probabilistic structure of the data, adequately into account. More often than not, such probabilistic structure is imposed on the data indirectly by tacking unobservable (white-noise) error terms on structural models, and it's usually nothing more than an afterthought. Indeed, it is often denied that there is such a thing as 'statistical information' separate from the substantive information.

In the next few sections we discuss several different approaches to model specification and validation with a view to bring out the weaknesses of current practice and make a case for the error statistical procedure that can ensure learning from data by securing the reliability of inductive procedures.

4. Theory-driven Modeling: The CAPM Revisited

4.1 The Pre-Eminence of Theory (PET) Perspective

The Pre-Eminence of Theory (PET) perspective, that has dominated empirical modeling in economics since the early 19th century, views model specification and re-specification as based exclusively on substantive information. The empirical model takes the form of *structural model*—an estimable form of that theory in light of the available data. The data play only a subordinate role in availing the quantification of the structural model after some random error term is tacked on to transform it into a statistical model. In a certain sense, the PET perspective denies the very existence of statistical information separate from any substantive dimension (see Spanos 2010a).

In the context of the PET perspective the appropriateness of an estimated model is invariably appraised using three types of criteria:

[i] statistical (goodness-of-fit/prediction, statistical significance),
[ii] substantive (theoretical meaningfulness, explanatory capacity),
[iii]) pragmatic (simplicity, generality, elegance).

An example of such a *structural model* from financial economics is the *Capital Asset Pricing Model (CAPM)* (Lai and Xing 2008):

$$\left(r_{kt}-r_{ft}\right)=\beta_k(r_{Mt}-r_{ft})+\varepsilon_{kt},\ k{=}1,2,...,m,\ t{=}1,...,n, \tag{4}$$

where r_{kt}—returns of asset k, r_{Mt}—market returns, r_{ft}— returns of a risk free asset. The error term ε_{kt} is assumed to satisfy the following probabilistic assumptions:

$$\begin{array}{l}\text{(i) } E(\varepsilon_{kt})=0,\ \text{(ii) } Var(\varepsilon_{kt})=\sigma_k^2-\beta_k^2\sigma_M^2=\sigma_{\varepsilon k}^2,\\ \text{(iii) } Cov(\varepsilon_{kt},\varepsilon_{\ell t}){=}v_{k\ell},\ k{\neq}\ell,\ k{,}\ell{=}1,...,m,\\ \text{(iv) } Cov(\varepsilon_{kt},r_{Mt}){=}0,\ \text{(v) } Cov(\varepsilon_{kt},\varepsilon_{ks}){=}0,\ t{\neq}s,\ t{,}s{=}1,...,n.\end{array} \tag{5}$$

The *economic principle* that underlies the CAPM is that financial markets are *information efficient* in the sense that prices 'fully reflect' the available information. That is, prices of securities in financial markets must equal fundamental values, because arbitrage eliminates pricing anomalies. Hence, one cannot consistently achieve returns $\left(r_{kt}-r_{ft}\right)$ in excess of average market returns $(r_{Mt}-r_{ft})$ on a risk-adjusted basis. The key structural parameters are the betas $(\beta_1,\beta_2,\cdots,\beta_m)$, where β_k measures the sensitivity of asset return k to market movements, and $\sigma_k{=}\sqrt{\sigma_{\varepsilon k}^2+\beta_k^2\sigma_M^2}$, the total risk of asset k, where $\beta_k^2\sigma_M^2$ and $\sigma_{\varepsilon k}^2$ denote the systematic and non-systematic components of risk; $\sigma_{\varepsilon k}^2$ can be eliminated by diversification but $\beta_k^2\sigma_M^2$ is non-diversifiable.

Another important feature of the above substantive model is that its probabilistic structure is specified in terms of unobsevable error terms instead of the observable processes $\{y_{kt}{:=}(r_{kt}-r_{ft}),\ k{=}1,2,...,m,\ X_t{:=}(r_{Mt}-r_{ft}),\ t{\in}\mathbb{N}\}$. This renders the assessment of the appropriateness of the error assumptions (i)–(v) at the specification stage impossible.

To test the appropriateness of the CAPM the structural model (4) is usually embedded into a *statistical model* known as the *stochastic Linear Regression*:

$$y_{kt}=\alpha_k+\beta_kX_t+u_{kt},\ k{=}1,..,m,\ t{=}1,...,n, \tag{6}$$

subject to the restrictions on $\beta_0{:=}(\alpha_1,\alpha_2,\cdots,\alpha_m){=}0$, which can be testing using the hypotheses:

$$H_0\text{: }\beta_0{=}\mathbf{0},\ \text{vs. } H_1{:}\beta_0{\neq}\,\mathbf{0}. \tag{7}$$

Empirical example (Lai and Xing 2008, 72–81). The relevant data $\mathbf{Z}_0$ come in the form of *monthly log-returns* $(\Delta \ln P_t)$ of 6 stocks, representing different sectors [Aug. 2000 to Oct. 2005] (n=64): Pfizer Inc. (PFE)—pharmaceuticals, Intel Corp. (INTEL)—semiconductors, Citigroup Inc. (CITI)—banking, American Express (AXP)—consumer finance, Exxon-Mobil Corp. (XOM)—oil and gas, General Motors (GM)—automobiles, the market portfolio is represented by the SP500 index and the risk free asset by the 3-month Treasury bill (3-Tb) rate.

For illustration purposes let us focus on one of these equations for Citigroup Inc. Estimating the statistical model (6) yields:

$$\begin{array}{ll} \text{Structural model:} & y_t = \alpha + \beta X_t + \varepsilon_t,\ \varepsilon_t \backsim \mathsf{NIID}(0,\sigma^2), \\ \text{Estimated (CITI):} & (r_{3t}-\mu_{ft}) = \underset{(.0033)}{.0053} + \underset{(.089)}{1.137}(r_{Mt}-\mu_{ft}) + \underset{(.0188)}{\widehat{\varepsilon}_{3t}}, \\ & R^2 = .725,\ s = .0188,\ n = 64. \end{array} \tag{8}$$

On the basis of this estimated model, the typical assessment will go something like:

(a) the signs and magnitudes of the estimated coefficients are in accordance with the CAPM (α=0 and $\beta > 0$),

(b) the beta coefficient β_3 is statistically significant, on the basis of the t-test:

$$\tau(\mathbf{z}_0;\beta_3) = \tfrac{1.137}{.089} = 12.775[.000],$$

where the p-value is given in square brackets,

(c) the CAPM restriction α_3=0 is *not rejected* by the data, at a 10% significance level, using the t-test:

$$\tau(\mathbf{z}_0;\alpha_3) = \tfrac{.0053}{.0033} = 1.606[.108],$$

(d) the goodness-of-fit is reasonably high (R^2=.725), providing additional support for the CAPM.

Taken together (a)–(d) are regarded as providing *good evidence for* the CAPM.

Concluding that the above data confirm the CAPM, however, is unsubstantiated because the underlying statistical model has *not* been validated vis-à-vis data $\mathbf{z}_0$, in the sense that no evidence has been provided that the statistical premises in (5) are valid of this data. This is necessary because the reliability of the above reported inference results relies heavily on the validity of the statistical premises (5) implicitly invoked by these test procedures as well as the R^2. Hence, unless one validates the probabilistic assumptions in (5) invoked by the inferences in question, the reliability of any inductive inference based on the estimated model is, at best, unknown. It is often claimed that the error assumptions in (5) are really innocuous and thus no formal testing is needed. As shown below, these innocuous looking error assumptions imply a number of restrictive probabilistic assumptions pertaining to the observable process $\{\mathbf{Z}_t{:=}(\mathbf{y}_t, X_t),\ t{\in}\mathbb{N}\}$ underlying data $\mathbf{z}_0$, and the validity of these assumptions vis-à-vis $\mathbf{z}_0$ is ultimately what matters for the trustworthiness of the above results.

Any departures from these probabilistic assumptions—statistical misspecification—often induces a discrepancy between *actual* and *nominal error probabilities*—stemming from the 'wrong' $f(\mathbf{z};\theta)$ via (2)—leading inferences astray. The surest way to draw an invalid inference is to apply a .05 (nominal) significance level test when its actual type I error probability is closer to .99 (see Spanos 2009a for several examples). Indeed, any inferential claim, however informal, concerning the sign, magnitude and significance of estimated coefficients, as well as goodness-of-fit, is likely to be misleading because of the potential discrepancy between the actual and nominal error probabilities due to statistical misspecification. What is even less well appreciated is that, irrespective of the modeler's stated objectives, the above criteria [i]–[iii] are undermined when the estimated model is statistically misspecified.

4.2 The Duhem problem

This neglect of statistical adequacy in statistical inference raises a serious philosophical problem, known as the *Duhem problem* (Chalmers 1999; Mayo 1997):

> "It is impossible to reliably test a substantive hypothesis in isolation, since any statistical test used to assess this hypothesis invokes auxiliary assumptions whose cogency is unknown."

That is, theory-driven modeling raises a critical problem that has devastated the trustworthiness of empirical modeling in the social sciences. When one imposes the substantive information (theory) on the data at the outset, the end result is often a statistically and substantively misspecified model, but one has no way to delineate the two sources of error:

(I) the substantive information is false, or
(II) the inductive premises are mispecified, (9)

and apportion blame with a view to address the unreliability of inference problem.

The key to circumventing the *Duhem problem* is to find a way to disentangle the statistical from the substantive premises (Spanos 2010c).

5. The Error-Statistical Approach

The error statistical approach differs from the other approaches primarily because, *ab initio*, it distinguishes clearly between statistical and substantive information as underlying two (ontologically) different but prospectively related models, the statistical $\mathscr{M}_\theta(\mathbf{z})$ and the structural $\mathscr{M}_\varphi(\mathbf{Z})$, respectively.

5.1 Statistical vs. Substantive Premises

The CAPM, a typical example of a structural model $\mathscr{M}_\varphi(\mathbf{Z})$:

$$\left(r_{kt}-r_{ft}\right)=\beta_k(r_{Mt}-r_{ft})+\varepsilon_{kt},\ k{=}1,2,...,m,\ t{=}1,...,n, \tag{10}$$

is naturally viewed by the PET adherents as the sole mechanism (premises) assumed to underlie the generation of data such as $\mathbf{Z}_0$. What often surprises these advocates is the idea that when data $\mathbf{Z}_0$ exhibit 'chance regularity patterns' one can construct a statistical mechanism (premises) that could have given rise to such data, without invoking any substantive information. This idea, however, is not new. Spanos (2006b) argues that it can be traced back to Fisher (1922), who proposed to view the initial choice (specification) of the statistical model as a response to the question:

> "Of what population is this a random sample?" (313), emphasizing that: "the adequacy of our choice may be tested a posteriori." (314)

What about the theory-ladeness of observation? Doesn't the fact that data $\mathbf{Z}_0$ were selected on the basis of some substantive information (theory) render it *theory-laden*? If anything, in fields like economics, the reverse is more plausible: theory models are data-laden so that they are rendered estimable. Theories aiming to explain the behavior of economic agents often need to be modified/adapted in order to become estimable in light of available data. However, such data in economics have been invariably gathered for purposes other than assessing the particular theory under consideration. They are usually gathered by government and other agencies for their own, mostly accounting purposes, and do not often correspond directly to any theory variables (see Spanos 1995). Moreover, the policy makers are interested in understanding the behavior of the actual economy, routinely giving rise to such data, and not the potential behavior of certain idealized and abstracted economic agents participating in an idealized economy.

5.2 Is There Such a Thing as 'Statistical' Information?

Adopting a *statistical perspective*, one views data $\mathbf{Z}_0$ as a realization of a generic (vector) stochastic process $\{\mathbf{Z}_t{:=}(\mathbf{y}_t,X_t),\ t{\in}\mathbb{N}\}$, regardless of what the variables $\mathbf{Z}_t$ measure substantively, thus separating the 'statistical' from the 'substantive' information. This is in direct analogy to Shannon's information theory based on formalizing the informational content of a message by separating 'regularity patterns' in strings of 'bits' from any substantive 'meaning':

> "Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semanticspects of communication are irrelevant to the engineering problem." (Shannon 1948, 379)

Analogously, the statistical perspective formalizes statistical information in terms of probability theory, e.g. probabilistic assumptions pertaining to the process $\{\mathbf{Z}_t,\ t{\in}\mathbb{N}\}$ underlying data $\mathbf{Z}_0$, and the substantive information (meaning) is irrelevant to the purely statistical problem of validating the statistical premises. The point is that there are crucially important distinctions between both the nature and warrant for the appraisal of substantive, in contrast to the statistical, adequacy of models. Indeed, this distinction holds the key to untangling the Duhem conundrum.

The construction of the relevant statistical premises begins with a given data $\mathbf{Z}_0$, separate from the theory or theories that led to the particular choice of $\mathbf{Z}_0$. Indeed, once selected, data $\mathbf{Z}_0$ take on 'a life of their own' as a particular realization of an underlying stochastic process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$. The connecting bridge between the real world of data $\mathbf{Z}_0$ and the mathematical world of the process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ is provided by the key question:

> 'what probabilistic structure pertaining to the process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$
> would render $\mathbf{Z}_0$ a *truly typical realization* thereof?'

That presupposes that one is able to glean the various chance regularity patterns exhibited by $\mathbf{Z}_0$ using a variety of graphical techniques as well as relate such patterns to probabilistic assumptions (see Spanos 1999). A pertinent answer to this question provides the relevant probabilistic structure of $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$, and the statistical premises $\mathcal{M}_\theta(\mathbf{z})$ are constructed by parameterizing that in a way that enables one to relate $\mathcal{M}_\theta(\mathbf{z})$ to the structural model $\mathcal{M}_\varphi(\mathbf{z})$.

To shed some light on the notions of *statistical information*, *chance regularities*, *truly typical realizations* and how they can be used to select $\mathcal{M}_\theta(\mathbf{z})$ in practice, let us consider the t-plots of different data in *figures 1–4*. Note that no substantive information about what these data series represent is given. The chance regularities exhibited by the data in *figure 1*, indicate that they can be realistically viewed as a typical realization of a NIID process. In this sense, the simple Normal model in (1) will be an appropriate choice. In practice, this can be formally confirmed by testing the NIID assumptions using simple Mis-Specification (M-S) tests (see Spanos 1999).

In contrast, the data in *figures 2–4*, exhibit chance regularities that indicate a number of different departures from the NIID assumptions. Hence, if one adopts the simple Normal model in (1) for any of the data in *figures 2–4*, the estimated model will be *statistically misspecified*; this can be easily verified using simple M-S tests (see Mayo and Spanos 2004). In particular, the data in *figure 2* exhibit a distinct departure from Normality since the distribution chance regularity indicates a highly skewed distribution. The data in *figure 3* exhibit a trending mean and variance; a clear departure from the ID assumption. The data in *figure 4* exhibit irregular cycles which indicate positive t-dependence; a clear departure from the Independence assumption.

Having chosen the appropriate probabilistic structure for $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$, the next step is to *parameterize* it in the form of the statistical model $\mathcal{M}_\theta(\mathbf{z})$ in such a way so as to nest (embed parametrically) the structural model, i.e. $\mathcal{M}_\varphi(\mathbf{z})\subset\mathcal{M}_\theta(\mathbf{z})$ and the embedding takes the general form of implicit restrictions between the statistical (θ) and structural (φ) parameters, say $\mathbf{G}(\theta,\varphi)=\mathbf{0}$. The technical details on how one goes from $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ to the statistical model, via probabilistic reduction, are beyond the scope of this paper (but see Spanos 1995; 2006a).

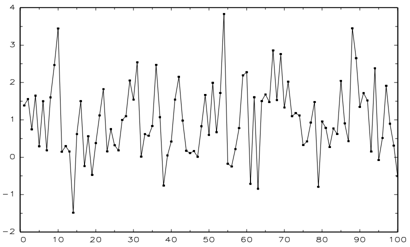

Figure 1: A typical realization of a NIID process

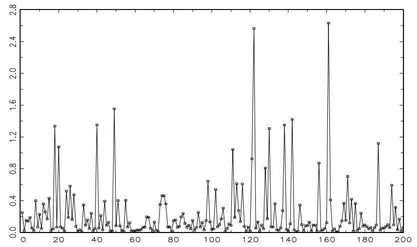

Figure 2: A typical realization of a Log-Normal IID process

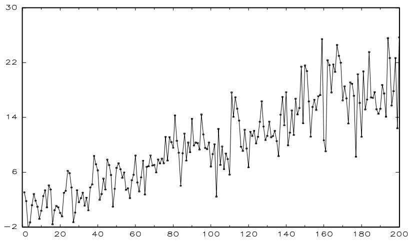

Figure 3: A typical realization of a NI, but t-heterogeneous process

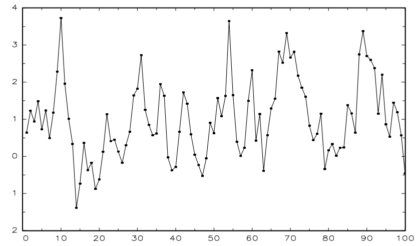

Figure 4: A typical realization of a Normal, Markov, Stationary process

Example. Consider a situation where all m data series exhibit chance regularities similar to *figure 1*, and one proceeds to assume that the vector process $\{\mathbf{Z}_t{:=}(Z_{1t},Z_{mt},...Z_{mt}),\ t{\in}\mathbb{N}\}$ has the following probabilistic structure:

$$\mathbf{Z}_t \backsim \mathsf{NIID}(\mu,\Sigma),\ \Sigma>0,\ t{\in}\mathbb{N}.$$

The statistical model $\mathcal{M}_\theta(\mathbf{Z})$ in *table 1* can be viewed as a parameterization of $\{\mathbf{Z}_t,\ t{\in}\mathbb{N}\}$. In practice the parameterization associated with $\mathcal{M}_\theta(\mathbf{Z})$ is selected to meet two interrelated aims:

(A) to account for the chance regularities in data $\mathbf{Z}_0$, in a way so that

(B) $\mathcal{M}_\theta(\mathbf{z})$ nests (parametrically) the structural model $\mathcal{M}_\varphi(\mathbf{z})$.

An example of such a statistical model is given in table 1 in terms of a statistical Generating Mechanism (GM) and the probabilistic assumptions [1]–[5]. The nesting restrictions take the form of $\beta_0{=}\,\mathbf{0}$, where $\beta_0{:=}(\alpha_1,\alpha_2,...,\alpha_m)$, and can be formally tested using the hypotheses in (7).

In relation to the specification in *table 1*, it is important to highlight the fact that assumptions [1]–[5] define a complete set of internally *consistent* and *testable* assumptions (statistical premises) in terms of the observable process $\{\mathbf{Z}_t,\ t{\in}\mathbb{N}\}$, replacing an incomplete set of assumptions pertaining to an unobservable error term:

$$\{(\mathbf{u}_t|X_t) \backsim \mathsf{NIID}(\mathbf{0},\mathbf{V})\},$$

Statistical GM:	$\mathbf{y}_t = \beta_0 + \beta_1 X_t + \mathbf{u}_t,\ t\in\mathbb{N}$
[1] Normality:	$\mathbf{Z}_t{:=}(\mathbf{y}_t, X_t) \backsim \mathsf{N}(.,.)$
[2] Linearity:	$E(\mathbf{y}_t \mid \sigma(X_t)) = \beta_0 + \beta_1 X_t,$
[3] Homosk/city:	$Var(\mathbf{y}_t \mid \sigma(X_t)) = \mathbf{V} > 0,$
[4] Independence:	$\{\mathbf{Z}_t, t\in\mathbb{N}\}$ independent process,
[5] t-invariance:	$\theta{:=}(\beta_0, \beta_1, \mathbf{V})$ do not change with t.
$\beta_0 = E(\mathbf{y_t}) - \beta_1 E(X_t),$	$\beta_1 {=} \frac{Cov(X_t, \mathbf{y}_t)}{Var(X_t)}, \mathbf{V} = Var(\mathbf{y_t}) - \beta_1 Cov(\mathbf{y}_t, X_t)$

Table 1: Stochastic Normal/Linear Regression model

as given in (5). This step is particularly crucial when the statistical model is only implicitly specified via the probabilistic assumptions pertaining to the error term of structural (substantive) model. In such cases one needs to derive the statistical model by transferring the error probabilistic assumptions onto the observable process $\{(\mathbf{y}_t|X_t),\ t\in\mathbb{N}\}$ with a view to ensure a complete and internally consistent set of assumptions because the error assumptions are often incomplete and sometimes internally inconsistent. Indeed, one of the most crucial assumptions, [5], is only implicit in (5) and is rarely validated in practice. Moreover, the traditional way of specifying such regression models interweaves the statistical and substantive premises in ways that makes it impossible to untangle the two (see Spanos 2010c). The quintessential example of this muddle is the assumption of *no omitted variables*, which clearly pertains to substantive adequacy and has nothing to do with statistical adequacy. Attempting to secure both the statistical and substantive adequacy simultaneously is a hopeless task in practice (see Spanos 2006c).

5.3 A Sequence of Interconnected Models

In any scientific inquiry there are primary questions of interest pertaining to the phenomenon of interest, and secondary ones that pertain to how to address the primary questions adequately. Spanos (1986, 12) suggested that a most effective way to bridge the gap between the phenomenon of interest and one's explanations or theories is to use a sequence of interlinked models [theory, structural (estimable), statistical], linking actual data to questions of interest. An almost identical idea was independently proposed by Mayo (1996) using a different terminology for the various models; primary, experimental and data models.

As Mayo (1996) emphasizes, splitting up the inquiry into levels and models is not a cut and dried affair. However, because of the way in which models 'below' have to be checked, and given how questions 'above' shape the variables of interest in the structural or statistical models, as the ways in which those questions determine the relevant criteria for scrutinizing the outputs of the statistical analysis, there is a back and forth process that constrains the modeling. Despite the fact that one might have started out from a different entry point or conjectured model, there is a back and forth multi-stage process of model specification, misspecification testing, respecification that is anything but arbitrary.

It is a mistake, however, to regard such an interconnected web of constraints as indicative of a Quinan-style "web of beliefs" (Quine and Ullian 1978), where models confront data in a holist block. Instead, taking seriously the piece-meal error statistical idea, we can distinguish, for a given inquiry, the substantive and statistical appraisals.

5.4 Statistical Adequacy and M-S Testing

A prespecified model $\mathscr{M}_\theta(\mathbf{z})$ is said to be *statistically adequate* when its assumptions (the statistical premises) are valid for data $\mathbf{Z}_0$. The question is 'How can one establish the statistical adequacy of $\mathscr{M}_\theta(\mathbf{z})$?' The answer is by applying thorough *Mis-Specification (M-S) testing* to assess the validity of the statistical premises vis-à-vis data $\mathbf{Z}_0$. For an extensive discussion of how one can ensure the thoroughness and reliability of the misspecification diagnosis see Mayo and Spanos 2004. As mentioned above, the substantive information plays no role in the purely statistical problem of validating the statistical premises. This enables one to assess the validity of the statistical premises before the probing of the substantive information, providing the key to circumventing Duhemian ambiguities.

The crucial role played by statistical adequacy stems from the fact that such a model constitutes *statistical knowledge* (similar to Mayo's experimental knowledge) that demarcates the empirical regularities that need to be explained using the substantive information. That is, data $\mathbf{Z}_0$ determine to a very large extent what kinds of models can reliably be used to learn about a phenomenon of interest. This is radically different from attaining a mere 'good fit', however, you measure the latter! It is also crucial to emphasize that the information used to infer an adequate/inadequate statistical model with severity is separate and independent of the parameters of the statistical model that will be used to probe the substantive questions of interest (see Mayo and Spanos 2004; Spanos 2010b).

Having secured the statistical adequacy of $\mathscr{M}_\theta(\mathbf{z})$, a necessary first step in securing the *substantive adequacy* of a structural model $\mathscr{M}_\varphi(\mathbf{z})$—parametrically nested within $\mathscr{M}_\theta(\mathbf{z})$—is to test the validity of the $p{=}(m{-}r) > 0$ restrictions in $\mathbf{G}(\theta,\varphi) = \mathbf{0}$, where the number of structural parameters (φ) r, is less than the number of statistical parameters (θ) m. The p restrictions imply a reparameterization/restriction of the generating mechanism described by $\mathscr{M}_\theta(\mathbf{z})$ with a view to transform the statistical knowledge into *substantive knowledge* that sheds additional light on the phenomenon of interest and the underlying mechanism. Questions of confounding factors and deep structural parameters should arise at this stage and not before. Let us illustrate some of these issues using the above empirical model.

5.5 Statistical Misspecification and Its Implications

For the CAPM the relevant nesting restrictions $\mathbf{G}(\theta,\varphi) = \mathbf{0}$, relating the statistical model $\mathscr{M}_\theta(\mathbf{z})$ in *table 1* with the structural model $\mathscr{M}_\varphi(\mathbf{z})$ in (4), takes the form of the statistical hypotheses in (7). The appropriate F-test yields: $F(\mathbf{z}_0;\alpha)$=2.181[.058], which does not reject H_0 at .05 level (Lai and Xing 2008). This result, however, is reliable only if $\mathscr{M}_\theta(\mathbf{z})$ in *table 1* is statistically adequate for the above data.

Mis-Specification (M-S) testing. Several simple M-S test results, based on simple auxiliary regressions, are reported in *table 2* (see Spanos and McGuirk 2001 for the details of the reported M-S tests). The small p-values associated with the majority of the tests indicate clear *departures from model assumptions* [1] and [3]–[5]! That is, $\mathscr{M}_\theta(\mathbf{z})$ is clearly misspecified, calling into question the reliability of all inferences reported above in (a)–(d), (mis)interpreted as confirming the CAPM.

[1] Normality:	Small(12)=46.7[.000]*
[2] Linearity:	$F(6,55)=7.659$[.264]
[3] Homoskedasticity:	$F(21,43)$=55.297[.000]*
[4] Independence:	$F(8,56)=55.331$[.021]*
[5] t-homogeneity:	$F(12,52)$=2.563[.010]*

Table 2: System Mis-Specification (M-S) tests

For expositional purposes let us focus our discussion on one of these estimated equations for CITI (r_{3t}), where the numbers in brackets below the estimates denote the standard errors. Not surprisingly, the single equation M-S results largely reflect the same misspecifications as those for the whole system of equations.

$$\boxed{\begin{array}{c}(r_{3t}-\mu_{ft})=\underset{(.0032)}{.0053}+\underset{(.089)}{1.137}(r_{Mt}-\mu_{ft})+\underset{(.0188)}{\widehat{u}_{3t}}\ ,\\ R^2{=}.725,\ s{=}.0188,\end{array}}$$

Mis-Specification (M-S) tests

[1] Normality:	$S{-}W=0.996$[.098]
[2] Linearity:	$F(1,61)=.468$[.496]
[3] Homoskedasticity:	$F(2,59)=4.950$[.010]*
[4] Independence:	$F(1,59)=6.15$[.016]*
[5] t-homogeneity:	$F_\beta(2,60)=4.611$[.014]*

A less formal, but more intuitive way to construe statistical adequacy is in terms of the *non-systematicity* (resulting from assumptions [1]–[5]) of the residuals from the estimated model. When $\mathscr{M}_\theta(\mathbf{Z})$ is statistically adequate, the systematic component defined by $E(y_t|\sigma(X_t))$=$\beta_0+\beta_1 X_t$ 'captures' the systematic (recurring) statistical information in the data, and thus the residuals $u_t=y_t-\widehat{\beta}_0-\widehat{\beta}_1 X_t$ are non-systematic in the sense of being an instantiation of a particular type of a 'white-noise' process; formally it is a 'martingale difference' process resulting from assumptions [1]–[5]; see Spanos (1999). The t-plot of the residuals from the estimated equation in (8), shown in *figure 5*, exhibit systematic information in the form of a trend and irregular cycles.

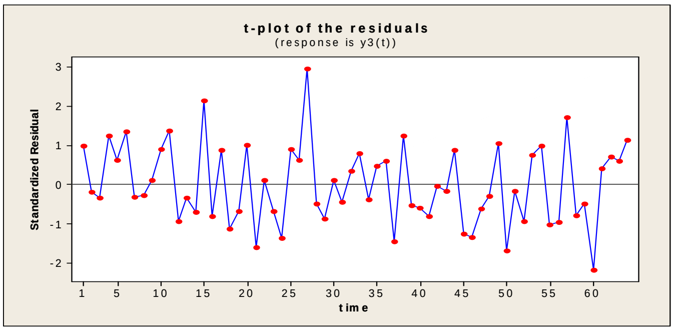

Figure 5: t-plot of the residuals from (8)

Of crucial interest is the departure from the t-invariance of the parameter β_1 parameter. An informal way to demonstrate that this assumption is invalid for the above data is to plot the recursive and 25-window estimates of β_1 shown in *figures 6–7* (see Spanos 1986). The non-constancy of these estimates indicates that [5] is invalid. These departures stem primarily from the t-heterogeneity of the sample mean and variance exhibited by data $\mathbf{z}_0$, shown in *figures 8–9*. Both t-plots exhibit a distinct quadratic trend in the mean and a decrease in the variation around this mean after observation t=33. In light of the fact that the statistical parameters relate to the mean and variance of $\mathbf{Z}_t$:=(y_t,X_t) via:

$$\beta_0{=}E(y_t)-\beta_1 E(X_t), \quad \beta_1{=}\frac{Cov(X_t,y_t)}{Var(X_t)}, \quad \sigma^2{=}Var(y_t){-}\beta_{1k}Cov(y_t,X_t)$$

the estimates of these parameters exhibit the non-constancy observed in *figures 6–7*.

Statistical misspecification and model evaluation criteria. The question that naturally arises is:

What does the above misspecification results imply for the traditional criteria: [a] statistical, [b] substantive and [c] pragmatic, used to evaluate empirical models? It is clear that the presence of statistical misspecification calls into question, not only the formal t and F tests invoked in (8) assessing the validity of the substantive information, but also the informal evaluations of the sign and magnitude of the estimated coefficients, as well as the goodness-of-fit/prediction measures. In light of that, any claims pertaining to theoretical meaningfulness and explanatory capacity are clearly unwarranted because they are based on inference procedures of questionable reliability; the invoked nominal error probabilities are likely to be very different from the actual ones! In general:

> No evidence *for* or *against* a substantive claim (theory) can be secured on the basis of a statistically misspecified model.

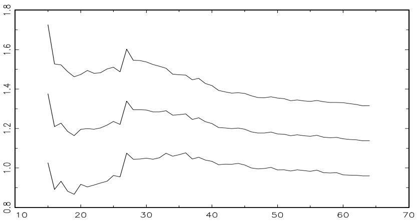

Figure 6: Recursive estimates of β_1

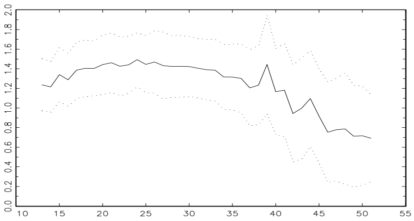

Figure 7: 25-window estimates of β_1

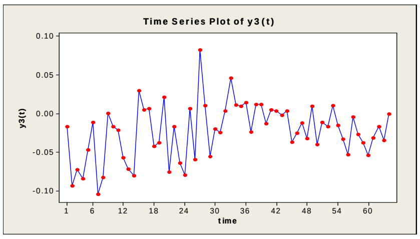

Figure 8: t-plot of CITI excess returns

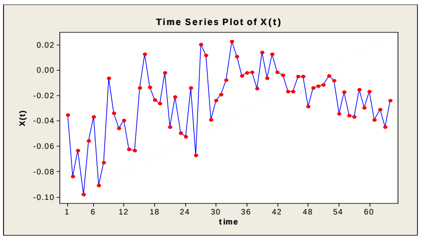

Figure 9: t-plot of market excess returns

In this sense, statistical adequacy provides a precondition for assessing *substantive adequacy*: establishing that the structural model $\mathcal{M}_\varphi(\mathbf{x})$ constitutes an adequate explanation of the phenomenon of interest. Without it the reliability of any inference procedures used to assess the substantive information is at best unknown; As argued in Spanos (2010a), a statistically adequate model $\mathcal{M}_\theta(\mathbf{z})$ gives data $\mathbf{z}_0$ 'a voice of its own' in the sense that any adequate explanation stemming from $\mathcal{M}_\varphi(\mathbf{x})$ should, at the very least, account for the empirical regularities demarcated by $\mathcal{M}_\theta(\mathbf{z})$.

What about pragmatic criteria like *simplicity* and *parsimony*? A statistical model $\mathcal{M}_\theta(\mathbf{z})$ is chosen to be as elaborate as necessary to secure *statistical adequacy*, but no more elaborate. Claims like 'simple models predict better' should be qualified to read: simple, but statistically adequate models, predict better than (unnecessarily) overparameterized models. Without statistical adequacy pragmatic criteria, such as simplicity, generality and elegance, are vacuous if such models will be used as a basis of inductive inference; they impede any learning from data (Spanos 2007).

What about pragmatic criteria like *goodness-of-fit/prediction* ? Perhaps the most surprising implication of statistical inadequacy is that it calls into question the most widely used criterion of model selection, the *goodness-of-fit/prediction* measures like:

$$R^2=1-\frac{\sum_{t=1}^n(y_t-\widehat{y}_t)^2}{\sum_{t=1}^n(y_t-\overline{y})^2}, \qquad MSPE=\sum_{t=n+1}^{n+p}(y_t-\widehat{y}_t)^2,$$

where $\widehat{y}_t{=}\widehat{\alpha}+\widehat{\beta}X_t,\ t{=}1{,}2{,}...{,}n$, denote the fitted values. Intuitively, what goes wrong with the R^2 is that, in the presence of t-heterogeneity in the mean and variance of $\mathbf{Z}_t{:=}(y_t{,}X_t)$, the statistics:

$$\tfrac{1}{n}\textstyle\sum_{t=1}^{n}(y_t{-}\widehat{y}_t)^2 \text{ and } \tfrac{1}{n}\textstyle\sum_{t=1}^{n}(y_t{-}\overline{y})^2$$

constitute unreliable (inconsistent) estimators of the conditional $[Var(y_t|X_t)]$ and marginal variance $[Var(y_t)]$, respectively. As argued in Spanos 2007, goodness-of-fit/prediction is neither necessary nor sufficient for statistical adequacy. This is because such criteria rely on the smallness of the residuals instead of their non-systematicity. Residuals can be small but systematically different from white-noise, and large but non-systematic.

This would seem totally counter-intuitive to theory-driven modelers whose intuition would insist that there is something *right-headed* about the use of such goodness-of-fit/prediction measures. This erroneous intuition stems from conflating statistical and substantive adequacy. In a case where a structural model $\mathscr{M}_{\varphi}(\mathbf{x})$ is data-acceptable, in the sense that its overidentifying restrictions $\mathbf{G}(\theta,\varphi)\mathbf{=0}$ are valid vis-à-vis a statistically adequate model $\mathscr{M}_{\theta}(\mathbf{x})$, such criteria become relevant for substantive adequacy. They measure a model's comprehensiveness (explanatory capacity/predictive ability) vis-à-vis the phenomenon of interest. It should be re-iterated that when goodness-of-fit/prediction criteria are used without securing statistical adequacy, they are vacuous and potentially highly misleading. Statistical adequacy does *not* ensure that. It only ensures that the actual error probabilities of any statistical inference procedures based on such a model approximate closely the nominal ones. That is, statistical adequacy sanctions the credibility of the inference procedures invoked by the modeler, including probing the substantive adequacy of a model.

5.6 When Probing for Substantive Adequacy is a Bad Idea

To illustrate what can go wrong is attempting to assess *substantive adequacy* when the estimated model is *statistically misspecified*, let us return to the above estimated model (8) and ask whether $(r_{6(t\text{-}1)}{-}\mu_{f(t\text{-}1)})$, the previous period excess returns of General Motors, constitute a relevant variable in explaining $(r_{3t}{-}\mu_{ft})$—excess returns of Citibank. Estimating the augmented model yields:

$$(r_{3t}{-}\mu_{ft}){=}\underset{(.0032)}{.0027}{+}\underset{(.087)}{1.173}(r_{Mt}{-}\mu_{ft})\ \boxed{\underset{(.048)}{\text{-.}119}(r_{6(t\text{-}1)}{-}\mu_{f(t\text{-}1)})}\ +\underset{(.0181)}{\widehat{v}_{3t}},$$
$$R^2{=}.753,\ s{=}.0181,$$

Hence, the answer is *yes* if the t-test ($\tau(\mathbf{z}_0){=}\frac{.119}{.048}{=}2.479[.017]$) is taken *at face value*! However, this is misleading because any variable with a certain trending structure is likely to appear significant when added to the original model, including generic trends and lags:

$$(r_{3t}{-}\mu_{ft}){=}\underset{(.0116)}{.0296}{+}\underset{(.119)}{1.134}(r_{Mt}{-}\mu_{ft})\ \boxed{\underset{(.065)}{\text{-.}134t}+\underset{(.083)}{.168t^2}}\ +\underset{(.0184)}{\widehat{v}_{3t}},$$
$$R^2{=}.745,\ s{=}.0184,$$

$$(r_{3t}-\mu_{ft})=\underset{(.003)}{.0023}+\underset{(.099)}{1.251}(r_{Mt}-\mu_{ft})\ \boxed{\underset{(.065)}{-.134}(r_{3(t\text{-}1)}-\mu_{f(t\text{-}1)})}+\underset{(.0182)}{\widehat{v}_{3t}}.$$
$$R^2=.750,\ s=.0182.$$

That is, statistical misspecifications are likely to give rise to highly unreliable inferences concerning, not only when probing for omitted variables, but any form of probing for substantive adequacy (see Spanos 2006c).

5.7 Addressing Duhemian Ambiguities

Viewing empirical modeling as a piecemeal process that relies on distinguishing between the statistical $\mathscr{M}_\theta(\mathbf{x})$ vs. substantive premises $\mathscr{M}_\varphi(\mathbf{x})$, and proceeds by securing statistical adequacy before any probing of the substantive premises, enables one to circumvent the Duhemian ambiguities that naturally arises in the PET approach discussed above. By insisting that the warranted inference be related to the particular error that might arise to impede learning from data, the error statistical framework distinguishes the following two questions:

(a) is model $\mathscr{M}_\theta(\mathbf{x})$ inadequate for accounting for the chance regularities in data $\mathbf{x}_0$?

(b) is model $\mathscr{M}_\varphi(\mathbf{x})$ inadequate as an explanation (causal or otherwise) of the phenomenon of interest?

That is, statistical models need to be justified as: (a) valid for the data $\mathbf{x}_0$, and (b) relevant for learning from data about phenomena of interest. It is important to emphasize that (b) does not necessarily coincide with finding a 'true' substantive (structural) model $\mathscr{M}_\varphi(\mathbf{x})$. One can learn a lot about a particular phenomenon of interest without requiring that $\mathscr{M}_\varphi(\mathbf{x})$ is a substantively 'true' model, whatever that might mean.

As argued in the next section, the modeling problems raised above are not unique to economics. These problems also arise in the context of two other approaches that seem different because they are more statistically oriented. It turns out, however, that their primary difference is that they often rely on alternative forms of substantive information.

6. Akaike-type Model Selection Procedures

Akaike-type procedures, which include the *Akaike Information Criterion* (AIC), the Bayesian (BIC), the Schwarz (SIC), the Hannan-Qinn (HQIC) and the Minimum Description Length (MDL), as well as certain forms of Cross-Validation; (Rao and Wu 2001; Burnham and Anderson 2002; Konishi and Kitagawa 2008), are widely used in econometrics, and other applied disciplines, as offering *objective methods* for selecting parsimonious models because they rely on maximizing the likelihood function subject to certain parsimony (simplicity) constraints.

A closer look at the Akaike-type model selection procedures reveals two major weaknesses. First, they rely on a misleading notion of objectivity in inference.

Second, they ignore the problem of statistical adequacy by taking the likelihood function at face value.

6.1 Objectivity in Inference

The traditional literature seems to suggest that 'objectivity' stems from the mere fact that one assumes a statistical model (a likelihood function), enabling one to accommodate highly complex models. Worse, in Bayesian modeling it is often misleadingly claimed that as long as a prior is determined by the assumed statistical model—the so called *reference prior*—the resulting inference procedures are objective, or at least as objective as the traditional frequentist procedures:

> "Any statistical analysis contains a fair number of subjective elements; these include (among others) the data selected, the model assumptions, and the choice of the quantities of interest. Reference analysis may be argued to provide an 'objective' Bayesian solution to statistical inference in just the same sense that conventional statistical methods claim to be 'objective': in that the solutions only depend on model assumptions and observed data." (Bernardo 2010, 117)

This claim brings out the unfathomable gap between the notion of 'objectivity' as understood in Bayesian statistics, and the error statistical viewpoint. As argued above, there is nothing 'subjective' about the choice of the statistical model $\mathcal{M}_\theta(\mathbf{z})$ because it is chosen with a view to account for the statistical regularities in data $\mathbf{z}_0$, and its validity can be objectively assessed using trenchant M-S testing. Model validation, as understood in error statistics, plays a pivotal role in providing an 'objective scrutiny' of the reliability of the ensuing inductive procedures.

Objectivity does NOT stem from the mere fact that one 'assumes' a statistical model. It stems from establishing a *sound link* between the process generating the data $\mathbf{z}_0$ and the assumed $\mathcal{M}_\theta(\mathbf{z})$, by securing statistical adequacy. The *sound* application and the *objectivity* of statistical methods turns on the *validity* of the assumed statistical model $\mathcal{M}_\theta(\mathbf{z})$ for the particular data $\mathbf{z}_0$. Hence, in the case of 'reference' priors, a misspecified statistical model $\mathcal{M}_\theta(\mathbf{z})$ will also give rise to an inappropriate prior $\pi(\theta)$.

Moreover, there is nothing subjective or arbitrary about the 'choice of the data and the quantities of interest' either. The appropriateness of the data is assessed by how well data $\mathbf{z}_0$ correspond to the theoretical concepts underlying the substantive model in question. Indeed, one of the key problems in modeling observational data is the pertinent bridging of the gap between the theory concepts and the available data $\mathbf{z}_0$ (see Spanos 1995). The choice of the quantities of interest, i.e. the statistical parameters, should be assessed in terms of the statistical adequacy of the statistical model in question and how well these parameters enable one to pose and answer the substantive questions of interest.

For error statisticians, *objectivity* in scientific inference is inextricably bound up with the *reliability* of their methods, and hence the emphasis on thorough probing of the different ways an inference can go astray (see Cox and Mayo 2010). It is in this sense that M-S testing to secure statistical adequacy plays a pivotal role in providing an *objective scrutiny* of the reliability of error statistical procedures.

In summary, the well-rehearsed claim that the only difference between frequentist and Bayesian inference is that they both share several subjective and arbitrary choices but the latter is more honest about its presuppositions, constitutes a lame excuse for the ad hoc choices in the latter approach and highlights the huge gap between the two perspectives on modeling and inference. The appropriateness of every choice made by an error statistician, including the statistical model $\mathscr{M}_\theta(\mathbf{z})$ and the particular data $\mathbf{z}_0$, is subject to independent scrutiny by other modelers.

6.2 'All models are wrong, but some are useful'

A related argument—widely used by Bayesians (see Gelman, this volume) and some frequentists—to debase the value of securing statistical adequacy, is that statistical misspecification is inevitable and thus the problem is not as crucial as often claimed. After all, as George Box remarked:

> "All models are false, but some are useful!"

A closer look at this locution, however, reveals that it is mired in confusion.

First, in what sense 'all models are wrong'?

This catchphrase alludes to the obvious simplification/idealization associated with any form of modeling: it does not represent the real-world phenomenon of interest in all its details. That, however, is very different from claiming that the underlying statistical model is unavoidably misspecified vis-à-vis the data $\mathbf{z}_0$. In other words, this locution conflates two different aspects of empirical modeling:

(a) the *realisticness* of the substantive assumptions comprising the structural model $\mathscr{M}_\varphi(\mathbf{z})$ (*substantive premises*), vis-à-vis the phenomenon of interest, with

(b) the *validity* of the probabilistic assumptions comprising the statistical model $\mathscr{M}_\theta(\mathbf{z})$ (*statistical premises*), vis-à-vis the data $\mathbf{z}_0$ in question.

It's one thing to claim that a model is not an exact picture of reality in a substantive sense, and totally another to claim that this statistical model $\mathscr{M}_\theta(\mathbf{z})$ could *not* have generated data $\mathbf{z}_0$ because the latter is statistically misspecified. The distinction is crucial for two reasons. To begin with, the types of *errors* one needs to probe for and guard against are very different in the two cases. *Substantive adequacy* calls for additional probing of (potential) errors in bridging the gap between theory and data. Without securing *statistical adequacy*, however, probing for substantive adequacy is likely to be misleading. Moreover, even though good fit/prediction is neither *necessary* nor *sufficient* for statistical adequacy, it *is* relevant for *substantive adequacy* in the sense that it provides a measure of the structural model's comprehensiveness (explanatory capacity) vis-à-vis the phenomenon of interest (see Spanos 2010a). This indicates that part of the confusion pertaining to model validation and its connection (or lack of) to goodness-of-fit/prediction criteria stem from inadequate appreciation of the difference between substantive and statistical information.

Second, how wrong does a model have to be to *not* be useful?

It turns out that the full quotation reflecting the view originally voiced by Box is given in Box and Draper (1987, 74):

> "[. . .] all models are wrong; the practical question is how wrong do they have to be to not be useful."

In light of that, the only criterion for deciding when a misspecified model is or is *not* useful is to evaluate its potential unreliability: the implied *discrepancy* between the relevant actual and nominal error probabilities for a particular inference. When this discrepancy is small enough, the estimated model can be useful for inference purposes, otherwise it is not. The onus, however, is on the practitioner to demonstrate that. Invoking vague *generic robustness* claims, like 'small' departures from the model assumptions do not affect the reliability of inference, will not suffice because they are often highly misleading when appraised using the error discrepancy criterion. Indeed, it's not the discrepancy between models that matters for evaluating the robustness of inference procedures, as often claimed in statistics textbooks, but the discrepancy between the relevant actual and nominal error probabilities (see Spanos 2009a).

In general, when the estimated model $\mathscr{M}_{\widehat{\theta}}(\mathbf{z})$ is statistically misspecified, it is practically useless for inference purposes, unless one can demonstrate that its reliability is adequate for the particular inferences.

6.3 Trading Goodness-of-fit/prediction against Simplicity

These Akaike-type *model selection* procedures aim to address the choice of a prespecified model by separating the problem into two stages. In stage 1, a broader family of models $\{\mathscr{M}_{\varphi_i}(\mathbf{z}),\ i{=}1,...m\}$ is selected using *substantive information*. It is important to emphasize that substantive information comes in a variety of forms including mathematical approximation theory. In stage 2, a best model $\mathscr{M}_{\varphi_k}(\mathbf{z})$ within this family is chosen by trading goodness-of-fit/prediction against parsimony (simplicity). In philosophy of science such modeling selection procedures are viewed as providing a pertinent way to address the curve fitting problem (see Forster and Sober 1994).

Example. Consider the case where the broader family of models $\{\mathscr{M}_{\varphi_i}(\mathbf{z}),$ $i{=}1,2,...m\}$ is the *Gauss-Linear model* (Spanos 2010b):

$$y_t = \textstyle\sum_{i=0}^{m} \alpha_i \phi_i(x_t) + \varepsilon_t,\ \varepsilon_t \backsim \mathsf{NIID}(0, \sigma^2(m)), \tag{11}$$

where $\varphi_i{=}(\alpha_0, \alpha_1, ..., \alpha_i,\ \sigma^2(i))$, and $\phi_i(x_t)$ $i{=}1,2,...m$, are known functions; often orthogonal polynomials . This family of models is often selected using a combination of substantive subject matter information and mathematical approximation theory.

The stated objective is motivated by the *curve-fitting perspective* and the selection is guided by the principle of trading *goodness-of-fit* against *overfitting*. The rationale is that when the goal is goodness-of-fit, the key problem in selecting the optimal value of m is thought to be *overfitting*, stemming from the fact that one can make the error $\varepsilon(x_t;m)=y_t-\sum_{i=0}^{m}\alpha_i\phi_i(x_t)$ as small as desired by increasing m. Indeed, it is argued that one can make the approximation error equal to *zero* by choosing $m=n-1$ (see Skyrms 2000). That is, the parameter estimates $\widehat{\varphi}_i$ are 'fine-tuned' to data-specific patterns and not to the generic recurring patterns. Hence, as this argument goes, goodness-of-fit cannot be the sole criterion for 'best'. To avoid *overfitting* one needs to supplement goodness-of-fit with pragmatic criteria such as *simplicity* (parsimony, which can be justified on prediction grounds, since simpler curves enjoy better predictive accuracy (see Forster and Sober 1994; Sober 2008).

Akaike's Information Criterion (AIC) is based on penalizing goodness-of-fit, measured by the log-likelihood function (-2ln$L(\theta)$), using the number of unknown parameters (K) in θ:

$$\mathsf{AIC} = \overbrace{-2\ln L(\widehat{\theta};\mathbf{z}_0)}^{\text{goodness-of-fit}} + \overbrace{2K}^{\text{penalty}}.$$

For the model in (11) the AIC takes the particular form:

$$\mathsf{AIC} = n\ln(\widehat{\sigma}^2)+2K, \quad \text{or} \quad \mathsf{AIC}_n = \ln(\widehat{\sigma}^2)+\tfrac{2K}{n}, \tag{12}$$

where $\widehat{\sigma}^2=\frac{1}{n}\sum_{t=1}^{n}(y_t-\sum_{i=0}^{m}\widehat{\alpha}_i\phi_i(x_t))^2$ and $K=m+2$.

Attempts to improve the AIC criterion gave rise to several modifications/extensions of the penalty function $g(n,K)$. Particular examples based on (11) are:

$$\mathsf{BIC}_n=\ln(\widehat{\sigma}^2)+\tfrac{K\ln(n)}{n}, \quad \mathsf{HQIC}_n=\ln(\widehat{\sigma}^2)+\tfrac{2K\ln(\ln(n))}{n}, \quad \mathsf{MDL}_n=(\mathsf{BIC}_K/2). \tag{13}$$

6.4 What Can Go Wrong with Akaike-type Procedures

As argued above, goodness-of-fit/prediction criteria are neither necessary nor sufficient for securing statistical adequacy, and the latter provides the only criterion for when a statistical model $\mathcal{M}_\theta(\mathbf{z})$ 'accounts for the regularities in data'. Where does this leave these Akaike-type model selection procedures?

These procedures (AIC, BIC, HQIC, MDL, etc.) are particularly vulnerable to statistical misspecification, because they take the likelihood function at face value, assuming away the problem of model validation (see Lehmann 1990). When the prespecified family $\{\mathcal{M}_{\varphi_i}(\mathbf{z}),\ i=1,2,...m\}$ is statistically misspecified, the very notion of goodness-of-fit/prediction is called into question because the likelihood function is incorrect and these procedures will lead to erroneous choices of a 'best' model with probability one.

To illustrate what goes wrong with the goodness-of-fit/prediction criteria let us return to the above example in (11) and assume that the Normality assumption is false, and instead, the underlying distribution is *Laplace*:

$$f(y_t;\theta)=\tfrac{1}{2\sigma}\exp\{\{-|y_t-\textstyle\sum_{i=0}^{m}\alpha_i\phi_i(x_t)|/\sigma\}, \quad \theta\mathbf{:=}(\alpha,\sigma)\in\mathbb{R}^{m+1}\times\mathbb{R}_+,\ y_t\in\mathbb{R}.$$

In this case the the likelihood function based on the Normal distribution yields:

$$-2\ln L_N(\widehat{\theta};\mathbf{z}_0) \quad = n\ln(\tfrac{1}{n}\textstyle\sum_{t=1}^{n}(y_t - \sum_{i=0}^{m}\widehat{\alpha}_i\phi_i(x_t))^2),$$

where $(\widehat{\alpha}_i,\ i=1,2,...,m)$ denote the *least-squares* estimators, will provide the *wrong* measure. The *correct* measure of goodness-of-fit stemming from the Laplace distribution is:

$$-2\ln L_L(\widehat{\theta};\mathbf{z}_0) \quad = 2n\ln(\tfrac{1}{n}\textstyle\sum_{t=1}^{n}\left|(y_t - \sum_{i=0}^{m}\widetilde{\alpha}_i\phi_i(x_t))\right|),$$

where $(\widetilde{\alpha}_i,\ i=1,2,...,m)$ denote the *least absolute deviation* estimators (see Shao 2003). Similarly, in the case where the true distribution is *Uniform*:

$$f(y_t;\theta) = \tfrac{1}{2\sigma}, \quad -\sigma \le [y_t - \textstyle\sum_{i=0}^{m}\alpha_i\phi_i(x_t)] \le \sigma,$$

the correct likelihood-based goodness-of-fit measure will take the form:

$$-2\ln L_U(\widehat{\theta};\mathbf{z}_0) \quad = 2n\ln([\max_{t=1,..,n}[y_t - \textstyle\sum_{i=0}^{m}\breve{\alpha}_i\phi_i(x_t)],$$

where $(\breve{\alpha}_i,\ i=1,2,...,m)$ denote the *minimax* estimators (see Spanos 2010b).

This suggests that there is nothing 'natural' about defining the goodness-of-fit/prediction criterion in terms of the sum of squares of the errors, as often claimed in the literature (see Forster and Sober 1994; Sober 2008). The 'naturalness' depends crucially on the assumed distributional and other probabilistic assumptions underlying $\mathcal{M}_\theta(\mathbf{z})$. Hence, when $\mathcal{M}_\theta(\mathbf{z})$ is statistically misspecified, not only is the reliability of the inference procedures called into question, but also the validity of the goodness-of-fit/prediction criteria.

More broadly, viewing the Akaike-type model selection procedures from the *error statistical perspective* (Mayo 1996), calls for:

(a) delineating the notion of a 'best' model as it relates to the various objectives [description, prediction, explanation, theory/policy appraisal, etc.] associated with using $\mathcal{M}_{\varphi_k}(\mathbf{z})$, and

(b) probing all the different ways the final inference: $\mathcal{M}^*_{\varphi_k}(\mathbf{z})$ is the 'best' model within the prespecified family $\{\mathcal{M}_{\varphi_i}(\mathbf{z}),\ i=1,2,...m\}$, might be in error.

In a closer look at the various different objectives, one thing stands out: all these objectives invoke, directly or indirectly, some form of *inductive inference*! Hence, minimally, a 'best' model should be statistically adequate, otherwise these objectives are imperiled.

The *first potential error* arises when the prespecified family $\{\mathcal{M}_{\varphi_i}(\mathbf{z}),\ i=1,...m\}$ does *not* include an adequate model. This will invariably lead one astray because the Akaike-type procedures will select an 'erroneous' model with probability one. What about circumventing this problem by securing the statistical adequacy of $\{\mathcal{M}_{\varphi_i}(\mathbf{z}),\ i=1,...m\}$ using trenchant Mis-Specification (M-S) testing? That will automatically address the problem of selecting a particular model within this family (stage 2), rendering these Akaike-type procedures redundant (Spanos 2010b). Selecting a statistical model on statistical adequacy grounds effectively annuls the use of simplicity to circumvent the problem of overfitting. Statistical adequacy addresses overfitting because the latter induces artificial systematic information in the residuals which can be detected using discerning M-S testing. Moreover, statistical adequacy calls into question the pertinence of 'expected' predictive success as a selection criterion because, as argued above, good prediction should be measured in terms of the prediction errors being non-systematic, not 'small'!

A *second potential error* arises when the family $\{\mathcal{M}_{\varphi_i}(\mathbf{z}),\ i{=}1,...m\}$ *does* include a statistically adequate model, say $\mathcal{M}_{\varphi_j}(\mathbf{z})$, but is different from the selected model $\mathcal{M}_{\varphi_k}(\mathbf{z})$, $j{\neq}k$. This error is inherent to the Akaike-type procedures because they ignore the relevant error probabilities associated with their stage 2 selection of a 'best' model with the particular family. Spanos (2010b) shows that the ranking of the different models within the family $\{\mathcal{M}_{\varphi_i}(\mathbf{z}),\ i{=}1,...m\}$ is equivalent to formal Neyman-Pearson (N-P) testing comparisons among these models, with one crucial difference: there is no 'controlling' of the relevant error probabilities. Moreover, different model selection procedures (AIC, BIC, etc.) select different models primarily because the (implicit) relevant error probabilities differ from one method to another. Indeed, it is shown that the type I error probability implicitly invoked by these procedures is often closer to $\alpha{=}.20$ than to the usual $\alpha{=}.05$. In addition, securing statistical adequacy also circumvents the *overfitting* problem these selection procedures allegedly address (Spanos 2007). Moreover, conflating the statistical and substantive premises, as these procedures do, will invariably raise the Duhem problem that cannot be addressed in the Akaike-type context.

In summary, the Akaike-type procedures are vulnerable to two serious problems stemming from the fact that they:

[a] assume away the problem of *model validation*, and that undermines the credibility of the likelihood function as the relevant measure of goodness-of-fit/prediction,

[b] ignore the *relevant error probabilities* associated with their selection of the 'best' model with the broader family. It is ironic that some modelers claim that the primary advantage of Akaike-type procedures over Neyman-Pearson methods on objectivity grounds is that the minimization upon which such procedures are based avoids the 'arbitrariness' of choosing a pre-data significance level (see Sober 2008).

Contrary to the conventional wisdom that Akaike-type procedures are needed to address inadequacies in frequentist methods, the above discussion shows that the only way to bring an objective scrutiny to those model selection procedures is to keep track of the relevant error probabilities; not ignore them under the pretense of some irrelevant trade-off between goodness-of-fit/prediction and simplicity!

7. Hendry's *General to Specific* Procedure

To shed further light on the weaknesses of the above theory-driven and Akaike-type procedures, let us compare them to *Hendry's general to specific procedure* (Campos et al. 2003) which shares some of its key features with the error statistical approach, including 'controlling' the relevant error probabilities and the due emphasis on statistical adequacy.

To make the discussion more specific, let us consider the model selection problem within the family of *Normal, linear regression* models:

$$\mathcal{M}_m: \quad y_t = \beta_0 + \textstyle\sum_{i=1}^{m} \beta_i x_{it} + u_t,\ u_t \backsim \mathsf{NIID}(0, \sigma^2),\ t\in\mathbb{N}, \tag{14}$$

whose statistical adequacy has been secured. The primary objective is to chose the substantively relevant subset $\mathbf{x}_{1t} \subset \mathbf{x}_t$ of the explanatory variables $\mathbf{x}_t$:=$(x_{1t}, x_{2t}, ..., x_{mt})$, but without sacrificing statistical adequacy. In the context of such probing goodness-of-fit/prediction measures, theoretical meaningfulness and simplicity, can play an important role, but *not* at the expense of statistical adequacy. For securing the latter one needs a more complete specification of the statistical premises in terms of probabilistic assumptions pertaining to the observables (*table 3*).

$$\left.\begin{array}{ll}
\text{Statistical GM:} & y_t = \beta_0 + \beta_1^{\top}\mathbf{x}_t + u_t,\ \ t\in\mathbb{N}:=(1,2,..,n,...) \\
\text{[1] Normality:} & (y_t \mid \mathbf{X}_t{=}\mathbf{x}_t) \backsim \mathsf{N}(.,.), \\
\text{[2] Linearity:} & E(y_t \mid \mathbf{X}_t{=}\mathbf{x}_t) = \beta_0 + \beta_1^{\top}\mathbf{x}_t, \\
\text{[3] Homoskedasticity:} & Var(y_t \mid \mathbf{X}_t{=}\mathbf{x}_t) = \sigma^2, \\
\text{[4] Independence:} & \{(y_t \mid \mathbf{X}_t{=}\mathbf{x}_t),\ t\in\mathbb{N}\} \text{ indep. process}, \\
\text{[5] t-invariance:} & \theta{:=}(\beta_0, \beta_1, \sigma^2) \text{ are } \textit{not} \text{ varying with } t,
\end{array}\right\}\ t\in\mathbb{N}.$$

$$\beta_0{=}\left[\mu_1 - \beta_1^{\top}\mu_2\right]\in\mathbb{R},\ \beta_1{=}\left[\Sigma_{22}^{-1}\sigma_{21}\right]\in\mathbb{R},\ \sigma^2 = \left[\sigma_{11} - \sigma_{21}^{\top}\Sigma_{22}^{-1}\sigma_{21}\right]\in\mathbb{R}_+,$$
$$\mu_1{=}E(y_t),\ \mu_2{=}E(\mathbf{X}_t),\ \sigma_{11}{=}Var(y_t),\ \sigma_{21}{=}Cov(\mathbf{X}_t, y_t),\ \Sigma_{22}{=}Cov(\mathbf{X}_t).$$

Table 3: Normal/Linear Regression model

Controlling the error probabilities. In contrast to Akaike-type procedures, Hendry's (1995) *general to specific* keeps track of the relevant error probabilities by framing the selection problem as a descending Neyman-Pearson (N-P) sequential selection procedure (Anderson 1962), and without neglecting statistical adequacy. In the case of the family of models in (14), the hypotheses of interest are arranged in a descending order beginning with the most general model and ending with the most specific:

$$\begin{array}{ll}
H_0^{(m+1)}: & \text{the unconstrained model} \\
H_0^{(m)}: & \beta_m = 0, \\
H_0^{(m-1)}: & \beta_m = \beta_{m-1} = 0, \\
\vdots & \vdots \\
H_0^{(1)}: & \beta_m = \beta_{m-1} = \cdots = \beta_1 = 0.
\end{array}$$

In view of the fact that $(\beta_0, \beta_1, ..., , \beta_m) \in \Theta := \mathbb{R}^{m+1}$, we can see that the successive hypotheses define a *decreasing sequence of subsets* of the parameter space Θ. This order ensures that the validity of a particular hypothesis implies the validity of all the proceeding ones; this is because: $H_0^{(m)} \subset H_0^{(m-1)} \subset \cdots \subset H_0^{(1)}$. The sequential N-P formulation of the null and alternative hypotheses takes the form:

$$\begin{array}{rl}
H_0^{(m)} \text{ vs. } H_0^{(m-1)} \Big\} & \left\{ \begin{array}{l} \text{if } H_0^{(m-1)} \text{ rejected, end and accept } H_0^{(m)}, \\ \text{if } H_0^{(m-1)} \text{ accepted, test next hypothesis:} \end{array} \right. \\
H_0^{(m-2)} \text{ vs. } H_0^{(m-1)} \Big\} & \left\{ \begin{array}{l} \text{if } H_0^{(m-2)} \text{ rejected, end and accept } H_0^{(m-1)}, \\ \text{if } H_0^{(m-2)} \text{ accepted, test next hypothesis:} \end{array} \right. \\
\vdots & \vdots \\
H_0^{(1)} \text{ vs. } H_0^{(2)} \Big\} & \left\{ \begin{array}{l} \text{if } H_0^{(1)} \text{ rejected, end and accept } H_0^{(2)}, \\ \text{if } H_0^{(1)} \text{ accepted, stop.} \end{array} \right.
\end{array}$$

This sequential testing continues until a null hypothesis $H_0^{(k)}$ is rejected and thus $H_0^{(k+1)}$ is accepted. For each k the hypotheses of interest are:

$$H_0^{(k)} : \beta_{k-1}{=}0, \text{ vs. } H_0^{(k+1)} : \beta_{k-1}{\neq}0, \ k = 1, 2, ...m$$

and the test statistic, based on the assumptions [1]–[5] in *table 3*, is either a two-sided Student's t or equivalently, an F test based on:

$$F_k(\mathbf{y}) {=} \left[\left(\frac{\widehat{\sigma}_k^2 - \widehat{\sigma}_{k+1}^2}{\widehat{\sigma}_{k+1}^2} \right) \left(\frac{n-k-2}{1} \right) \right] \overset{H_0^{(k)}}{\backsim} F(1, n{-}k{-}2), \ k{=}1, 2, ..., m,$$

where $\widehat{\sigma}_k^2 {=} \frac{1}{n} \sum_{t=1}^n (y_t - \widehat{\beta}_0 - \sum_{i=1}^k \widehat{\beta}_i x_{it})^2$. A sequence of F-tests is defined in terms of the rejection regions:

$$C_1^{(k)} {:=} \{F_i(\mathbf{y}) > c_{\alpha_k}\}, \ k{=}1, 2, ..., m,$$

where α_k denotes the *significance level* in testing hypothesis $H_0^{(k)}$ and c_{α_k} the corresponding threshold. It can be shown that this sequence of tests is *Uniformly Most Powerful Unbiased* (UMPU).

A major advantage of this general to specific procedure is that, under $H_0^{(k)}$ the statistics $F_m(\mathbf{y})$, $F_{m-1}(\mathbf{y}), \cdots F_k(\mathbf{y})$ are mutually independent of $F_{k-1}(\mathbf{y})$, $F_{k-2}(\mathbf{y})$, $\cdots F_1(\mathbf{y})$ (see Anderson 1962). This property enables us to derive the type I error probability at each stage of the testing sequence as follows:

$$\begin{array}{rl}
\mathbb{P}(\text{reject } H_0^{(k)}; H_0^{(k)} \text{ is true }) & = 1 - \mathbb{P}(\text{accept } H_0^{(k)}; H_0^{(k)} \text{ is true}) \\
& = 1 - \prod_{i=k}^m (1 - \alpha_i), \text{ for } k{=}1, 2, ..., m.
\end{array}$$

In contrast, if one were to arrange the hypotheses of interest in an ascending order, from specific to general, this cannot be achieved because the test statistic $F_k(\mathbf{y})$ is no longer independent of $(F_{k-1}(\mathbf{y}),F_{k-2}(\mathbf{y}),\cdots,F_1(\mathbf{y}))$; one needs to invoke crude upper bounds such as the Bonferroni (see Wassermann 2006). The crudeness of such upper bounds often defeats the whole purpose of 'controlling' the relevant error probabilities.

This general to specific procedure has been elaborated and extended in Hendry 2011 (this volume) into a systematic model selection algorithm that takes into account the different possible orderings (selection paths) of the explanatory variables $\mathbf{x}_t{:=}(x_{1t},x_{2t},...,x_{mt})$. In addition, the algorithm can be applied to highly complex models where there are more variables than observations $m > n$, and can accommodate any form of substantive information from highly specific to very vague. Its effectiveness stems primarily from the combination of adopting general specifications that deal with many potentially relevant variables, long lags, non-linearities, breaks and outliers to ensure a congruent selection, retaining the theory model, and using multi-path search constrained by encompassing and congruence (see Castle et al. 2011). This oppugns claims that traditional frequentist methods cannot handle highly complex models with numerous variables.

It is important to bring out the fact that the autometrics algorithm shares a lot of its main features with the error-statistical model specification and validation approach articulated in *section 4*. In addition to 'controlling' the relevant error probabilities and the emphasis on statistical adequacy, Hendry's model selection algorithm identifies the primary aim of empirical modeling as building models and designing probative procedures with a view to 'find things out' and learn from data about the phenomenon of interest. This does not require one to adopt a strictly *instrumentalist* stance about the nature of theories. It simply acknowledges the distinction between statistical and substantive adequacy: a *substantive model* $\mathscr{M}_\varphi(\mathbf{z})$ may always come up short in fully capturing or explaining a phenomenon of interest, but a *statistical model* $\mathscr{M}_\theta(\mathbf{z})$ could be entirely adequate to reliably test and assess the substantive questions of interest, including confounding effects, omitting relevant or admitting irrelevant variables, as well as appraising the appropriateness of $\mathscr{M}_\varphi(\mathbf{z})$ and other competing substantive models. Indeed, a statistically adequate $\mathscr{M}_\theta(\mathbf{z})$ provides a benchmark for any aspiring theory in the sense that it establishes 'what there is to explain' given data $\mathbf{z}_0$. In this respect, Hendry's model selection procedure attempts to systematize certain aspects of 'model discovery' within a prespecified family of models, by demonstrating that there are intelligent ways to carry out such probing effectively, as well as certain impertinent ways to be avoided (see also Hendry 2009).

8. Summary and Conclusions

Statistical adequacy [the probabilistic assumptions comprising the statistical model are valid for the particular data] is of paramount importance when learning from data is a key objective in empirical modeling. Without it, the reliability of any inductive inference is at best unknown, calling into question any probing of substantive questions of interest. In short, statistical adequacy can be ignored at the expense of any learning from data about the phenomena of interest.

Despite the crucial importance of securing the reliability of statistical inference, statistical adequacy has been seriously neglected in the empirical literature. Instead, the focus has been on using substantive information to specify statistical models and appraising their appropriateness using goodness-of-fit/prediction and other criteria, which are of questionable value when the estimated model is statistically misspecified. It was argued that, not only theory-driven modeling, but also the Akaike-type model selection procedures, often give rise to unreliable inferences, primarily because the probabilistic structure of the data is indirectly imposed on the data via arbitrary error terms and its appropriateness is not assessed using thorough misspecification testing. The latter also calls into question the reliance on asymptotically justifiable inference procedures. Foisting one's favorite theory on data often yields estimated models which are both statistically and substantively misspecified, but one has no way to delineate between the two sources of error and apportion blame, raising serious Duhemian ambiguities. The error-statistical approach addresses this problem by adopting a purely probabilistic construal of statistical models that enables one to separate the statistical and substantive premises, creating the conditions for securing statistical adequacy before appraising the substantive information. in contrast to the theory-driven and the Akaike-type procedures, some of the key features of the error statistical approach are shared by Hendry's general to specific procedure that does a much better job in selecting a model within a prespecified family.

References

Anderson, T. W. (1962), "The Choice of the Degree of the Polynomial Regression as a Multiple Decision Problem", *Annals of Mathematical Statistics* 33, 255–266.

Bernardo, J. M. (2010), "Bayesian Statistics", in: Lovric, M. (ed.), *International Encyclopedia of Statistical Science*, New York: Springer, 107–133.

Box, G. E. P. and N. R. Draper (1987), *Empirical Model-Building and Response Surfaces*, New York: Wiley.

Burnham, K. P. and D. R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed., New York: Springer.

Campos, J., D. F. Hendry and H.-M. Krolzig (2003), "Consistent Model Selection by an Automatic Gets Approach", *Oxford Bulletin of Economics and Statistics* 65, 803–819.

Castle, J. L., J. A. Doornik and D. F. Hendry (2011), "Evaluating Automatic Model Selection", *Journal of Time Series Econometrics* 3(1), DOI: 10.2202/1941-1928.1097.

Chalmers, A. F. (1999), *What Is This Thing Called Science?*, 3rd ed., Indianapolis: Hackett.

Cox, D. R. (1990), "Role of Models in Statistical Analysis", *Statistical Science* 5, 169–174.

— and D. V. Hinkley (1974), *Theoretical Statistics*, London: Chapman & Hall.

— and H. D. Miller (1968), *The Theory of Stochastic Processes*, London: Chapman & Hall.

— and D. G. Mayo (2010), "Objectivity and Conditionality in Frequentist Inference", in: Mayo, D. G. and A. Spanos (eds.), *Error and Inference*, Cambridge: Cambridge University Press, 276–304.

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics", *Philosophical Transactions of the Royal Society A* 222, 309–368.

— (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

— (1925b), "Theory of Statistical Estimation", *Proceedings of the Cambridge Philosophical Society* 22, 700–725.

— (1955), "Statistical Methods and Scientific Induction", *Journal of The Royal Statistical Society B* 17, 69–78.

Forster, M. and E. Sober (1994), "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions", *British Journal for the Philosophy of Science* 45, 1–35.

Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge: Cambridge University Press.

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

— (2009), "The Methodology of Empirical Econometric Modeling", in: Mills, T. C. and K. Patterson, *New Palgrave Handbook of Econometrics*, vol. 2, London: MacMillan, 3–67.

— (2011), "Empirical Economic Model Discovery and Theory Evaluation", this issue.

Konishi, S. and G. Kitagawa (2008), *Information Criteria and Statistical Modeling*, New York: Springer.

Lai, T. L. and H. Xing (2008), *Statistical Models and Methods for Financial Markets*, New York: Springer.

Lehmann, E. L. (1990), "Model Specification: The Views of Fisher and Neyman, and Later Developments", *Statistical Science* 5, 160–168.

Lovric, M. (2010), *International Encyclopedia of Statistical Science*, New York: Springer.

Mayo, D. G. (1996), *Error and the Growth of Experimental Knowledge*, Chicago: The University of Chicago Press.

— (1997), "Duhem's Problem, the Bayesian Way, and Error Statistics, or 'What's Belief Got to Do with It?'", *Philosophy of Science* 64, 222–244.

— (2003), "Severe Testing as a Guide for Inductive Learning", in: Kyburg, H. E. and M. Thalos (eds.), *Probability is the Very Guide of Life*, Chicago: Open Court, 89–118.

— (2005), "Philosophy of Statistics", in: Sarkar, S. and J. Pfeifer (eds.), *Philosophy of Science: An Encyclopedia*, London: Routledge, 802–15.

— and D. R. Cox (2006), "Frequentist Statistics as a Theory of Inductive Inference", in: Rojo, J. (ed.), *Optimality: The Second Erich L. Lehmann Symposium*, Lecture Notes-Monograph Series, vol. 49, Beachwood: Institute of Mathematical Statistics, 77–97.

— and A. Spanos (2004), "Methodology in Practice: Statistical Misspecification Testing", *Philosophy of Science* 71, 1007–1025.

— and — (2006), "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction", *British Journal for the Philosophy of Science* 57, 323–57.

— and — (2010), *Error and Inference*, Cambridge: Cambridge University Press.

— and — (2011), "Error Statistics", in: Gabbay, D., P. Thagard and J. Woods (eds.), *Handbook of Philosophy of Science, vol. 7: Philosophy of Statistics*, Amsterdam: Elsevier, 151–196.

Neyman, J. (1937), "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability", *Philosophical Transactions of the Royal Statistical Society of London A* 236, 333–380.

— (1956), "Note on an Article by Sir Ronald Fisher", *Journal of the Royal Statistical Society B* 18, 288–294.

— and E. S. Pearson (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses", *Phil. Trans. of the Royal Society A* 231, 289–337.

Pearson, E. S. (1955), "Statistical Concepts in the Relation to Reality", *Journal of the Royal Statistical Society B* 17, 204–207.

Quine, W. V. and J. S. Ullian (1978), *The Web of Belief*, 2nd ed., New York: McGraw-Hill.

Rao, C. R. (2004), "Statistics: Reflections on the Past and Visions for the Future", *Amstat News* 327, 2–3.

— and Y. Wu (2001), "On Model Selection", in: Lahiri, P. (ed.), *Model Selection*, Lecture Notes-Monograph series, vol. 38, Beachwood: Institute of Mathematical Statistics, 1–64.

Shannon, C. E. (1948), "A Mathematical Theory of Communication", *Bell System Technical Journal* 27, 379–423; 623–656.

Skyrms, B. (2000), *Choice and Chance: An Introduction to Inductive Logic*, 4th ed., Belmont: Wadsworth.

Shao, J. (2003), *Mathematical Statistics*, 2nd ed., New York: Springer.

Sober, E. (2008), *Evidence and Evolution: The Logic behind the Science*, Cambridge: Cambridge University Press.

Spanos, A. (1986), *Statistical Foundations of Econometric Modelling*, Cambridge: Cambridge University Press.

— (1995), "On Theory Testing in Econometrics: Modeling with Nonexperimental Data", *Journal of Econometrics* 67, 189–226.

— (1999), *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge: Cambridge University Press.

— (2000), "Revisiting Data Mining: 'Hunting' with or without a License", *The Journal of Economic Methodology* 7, 231–264.

— (2006a), "Econometrics in Retrospect and Prospect", in: Mills, T. C. and K. Patterson, *New Palgrave Handbook of Econometrics*, vol. 1, London: MacMillan, 3–58.

— (2006b), “Where Do Statistical Models Come From? Revisiting the Problem of Specification”, in: Rojo, J. (ed.), *Optimality: The Second Erich L. Lehmann Symposium*, Lecture Notes-Monograph Series, vol. 49, Beachwood: Institute of Mathematical Statistics, 98–119.

— (2006c), “Revisiting the Omitted Variables Argument: Substantive vs. Statistical Adequacy”, *Journal of Economic Methodology* 13, 179–218.

— (2007), “Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach”, *Philosophy of Science* 74, 1046–1066.

— (2009a), “Statistical Misspecification and the Reliability of Inference: The simple t-test in the Presence of Markov Dependence”, *The Korean Economic Review* 25, 165–213.

— (2009b), “The Pre-Eminence of Theory versus the European CVAR Perspective in Macroeconometric Modeling”, *Economics: The Open-Access, Open-Assessment E-Journal* 3, URL: http://www.economics-ejournal.org/economics/journalarticles/2009-10.

— (2010a), “Theory Testing in Economics and the Error Statistical Perspective”, in: Mayo and Spanos (2010), 202–246.

— (2010b), “Akaike-type Criteria and the Reliability of Inference: Model Selection versus Statistical Model Specification”, *Journal of Econometrics* 158, 204–220.

— (2010c), “Statistical Adequacy and the Trustworthiness of Empirical Evidence: Statistical vs. Substantive Information”, *Economic Modelling* 27, 1436–1452.

— (2010d), “Is Frequentist Testing Vulnerable to the Base-Rate Fallacy?”, *Philosophy of Science* 77, 565–583.

— (2011), “A Frequentist Interpretation of Probability for Model-based Inductive Inference”, forthcoming in *Synthese*.

— and A. McGuirk (2001), “The Model Specification Problem from a Probabilistic Reduction Perspective”, *Journal of the American Agricultural Association* 83, 1168–1176.

Wasserman, L. (2006), *All of Nonparametric Statistics*, New York: Springer.