Philosophy of Econometrics

Aris Spanos[1] [aris@vt.edu]
Virginia Tech, USA

April 2021

"I fully agree with you about the significance and educational value of methodology as well as history and philosophy of science. So many people today—and even professional scientists—seem to me like somebody who has seen thousands of trees but has never seen a forest. Knowledge of the historic and philosophical background gives that kind of independence from prejudices of his generation from which most scientists are suffering. This independence created by philosophical insight is—in my opinion—the mark of distinction between a mere artisan or specialist and a real seeker after truth." (Einstein to Thornton, 7 December 1944, Einstein Archives, 61–574).

1 Introduction

The above quotation from Einstein's reply to Robert Thornton, a young philosopher of science who began teaching physics at the university level in 1944, encapsulates succinctly the importance of examining the methodology, history and the philosophical foundations of different scientific fields to avoid missing the forest for the trees. The history of a scientific field gives researchers a glimpse of its roots and evolution, but more importantly, it provides researchers with a balanced perspective on the current 'paradigm' they find themselves engaged in, as well as its potential growth and development. Broadly speaking, a paradigm is a conceptual framework that includes theories, beliefs, values, research methods, objectives, the professional and educational structure of a scientific field, as well as standards for what constitutes legitimate contributions to a field. Philosophy of science emphasizes skills that are often absent from the training of scientists in most fields, including attentiveness to conceptual clarification and coherence, vigilance against equivocation, the accuracy of expression and weak links in arguments, the capacity to detect gaps in traditional arguments and devise novel perspectives, as well as the ability to frame alternative conceptual perspectives. Successful scientific fields, such as physics, chemistry, astronomy, and biology, have repeatedly redefined their conceptual framework over time, along with their goals, methods, and tools. Such conceptual revisions are the result of long periods of reflection revolving around the incessant dialogue between theory and data and guided by the systematic re-examination of the current methodology, and philosophical foundations.

The field of interest in the discussion that follows is modern econometrics whose roots

[1] Thanks are due to Julian Reiss for several comments and suggestions that improved the discussion substantially.

can be traced back to the early 20th century; see Morgan (1990), Qin (1993), Spanos (2006a). Econometrics is primarily concerned with the systematic study of economic phenomena employing observed data in conjunction with statistical models and substantive subject matter information. The philosophy of econometrics relates to *methodological* issues concerning to the effectiveness of econometric methods, and procedures used in empirical inquiry, as well as *ontological* issues concerned with the worldview of the econometrician; see Hoover (2006). Hence, its success should be evaluated with respect to its effectiveness in enabling practitioners to 'learn from data' about such phenomena, i.e. the extent to which econometric modeling and inference gives rise to trustworthy evidence that transforms tentative substantive conjectures into reliable knowledge about economic phenomena. One transforms tentative conjectures into real knowledge by testing the cogency of the substantive information using observable data given rise to by the phenomenon of interest. From this perspective econometric modeling and inference provides a statistical framework with a twofold objective: to account for the chance regularity patterns in data and to construct 'provisional' substantive models that shed adequate light (explain, describe, predict) economic phenomena of interest.

When assessed on such grounds current econometric methodology would be judged to be an inauspicious failure, or so it is argued in what follows. That makes the task of a meta-level appraisal of the methods, procedures, and strategies employed in studying economic phenomena using econometrics all the more urgent. It is often forgotten that scientific fields do not have a methodology written in stone with well-defined objectives and a fixed conceptual framework, even though it might look that way for newcomers to the field. The history of science teaches us that all these components evolve toward (hopefully) better science, sometimes after long digressions.

2 Descriptive statistics and induction

The problem of induction in the sense of justifying an inference from particular instances to realizations yet to be observed, has been bedeviling the philosophy of science since Hume's (1748) discourse on the problem. In its simplest form, *induction by enumeration* boils down to justifying the *straight-rule:* if the proportion of red marbles from a sample of size n is (m/n), infer that approximately a proportion (m/n) of all marbles in the urn are red"; see Salmon (1967), p. 50. The key feature of inductive inference is that it is *ampliative* in the sense that it goes beyond the observed data (m/n) to the unknown $\theta = \mathbb{P}(R)$ – that reflects the proportion of red (R) marbles in the urn – enhancing our knowledge about the underlying set-up that gave rise to the observed data. Numerous attempts to justify this inductive rule have failed, and the problem of induction is still unresolved in philosophy of science; see Henderson (2020), Reiss (2013, 2015) inter alia.

A case can be made that Karl Pearson's approach to descriptive statistics (Yule, 1916), can be viewed as a more sophisticated form of *induction by enumeration*. The approach is data-driven in search of a model in the sense that one would begin with the raw data $\mathbf{x}_0 := (x_1, \ldots, x_n)$, and in step 1 one would summarize $\mathbf{x}_0$ using a histogram with $m \geq 10$ bins. In step 2 one would select a frequency curve $f(x; \boldsymbol{\theta})$, $x \in \mathbb{R}_X \subset \mathbb{R}$ -real line, from *the Pearson family* whose members are generated by:

$$[d\ln f(x;\boldsymbol{\theta})/dx]=[(x-\theta_1)/(\theta_2+\theta_3x+\theta_4x^2)],\ x\in\mathbb{R}, \tag{1}$$

aiming to describe the data even more succinctly in terms of four unknown parameters $\boldsymbol{\theta}:=(\theta_1,\theta_2,\theta_3,\theta_4)$; note that (1) includes several well-known distributions, such as the Normal, the Student's t, the Beta, the Gamma, etc. This is achieved in step 3 by estimating $\boldsymbol{\theta}$ using the first four data raw moments, $\hat{\mu}_k=\frac{1}{n}\sum_{t=1}^{n}x_t^k,\ k=1,2,3,4,$and solving a system of four equations stemming from (1) for $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0)$; see Spanos (2019), p. 551. In step 4 one would use the estimates $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0)$ to select a member of this family $f(x;\widehat{\boldsymbol{\theta}})$ that 'best' describes the data. In step 5 one would evaluate the 'appropriateness' of $f(x;\widehat{\boldsymbol{\theta}})$ using Pearson's goodness-of-fit chi-square test, based on the difference $(\widehat{\boldsymbol{\theta}}(\mathbf{x}_0)-\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ denotes the selected $f(x;\boldsymbol{\theta}_0),\ x\in\mathbb{R}_X$ known parameters. Pearson's justification of his statistical analysis was based on the fact that the chosen frequency curve $f(x;\widehat{\boldsymbol{\theta}}),x\in\mathbb{R}_X,$ is the 'best' on goodness-of-fit grounds, and that could justify going beyond the data in hand $\mathbf{x}_0$. Although his approach to statistical induction was Bayesian in spirit, his use of uniform priors routinely enhanced the role of $f(x;\widehat{\boldsymbol{\theta}}),x\in\mathbb{R}_X$.

Similarly, Pearson's approach to correlation and regression amounts to curve-fitting guided by goodness-of-fit with a view to describe succinctly the association between data series, say $\mathbf{z}_0:=\{(x_t,y_t),\ t=1,2,\ldots,n\}$. The conventional wisdom underlying the Pearson-type statistics is summarized by Mills's (1924) who distinguishes between 'statistical description vs. statistical induction'. In statistical description measures such as the 'sample' mean $\overline{x}=\frac{1}{n}\sum_{t=1}^{n}x_t$, variance $s_x^2=\frac{1}{n}\sum_{t=1}^{n}(x_t-\overline{x}_n)^2$, and correlation coefficient:

$$r=\left[\left(\sum_{t=1}^{n}(x_t-\overline{x}_n)(y_t-\overline{y}_n)\right)/\sqrt{[\sum_{t=1}^{n}(x_t-\overline{x}_n)^2][\sum_{t=1}^{n}(y_t-\overline{y}_n)^2]}\right], \tag{2}$$

'provide just a summary for the data in hand' and "may be used to perfect confidence, as accurate descriptions of the given characteristics" (p. 549). However, when the results are to be extended *beyond* the data in hand - statistical induction - their validity depends on certain inherent *a priori* stipulations, such as (a) the 'uniformity' for the *population* and (b) the 'representativeness' of the *sample* (pp. 550-2). That is, statistical description does not invoke the validity of any assumptions, but if the same data are used to go beyond the data in hand (inductive inference), one needs to invoke (a) and (b).

What Pearson and Mills did not appreciate sufficiently is that, even for descriptive purposes, going from the raw data $\mathbf{x}_0$ to the histogram invokes the assumptions of Independence and Identically Distributed (IID). When these assumptions are invalid the histogram will provide an inappropriate description of $\mathbf{x}_0$, and the frequency curve that is chosen on goodness–of-fit grounds will be highly misleading. Similarly, correlation and regression assume that the data $\mathbf{z}_t:=(x_t,y_t),\ t=1,2,\ldots,n,$ are IID over the ordering t. When these assumptions are invalid, the summary statistics will be spurious; see Spanos (2019).

3 Model-based statistical modeling and inference

3.1 Model-based statistical induction

Fisher's (1922) recasting of statistics open the door for the standpoint that data $\mathbf{x}_0 := (x_1, \ldots, x_n)$ can be viewed as a typical realization of a stochastic processes $\{X_t,\ t \in \mathbb{N}\}$ to be integrated into modern statistics properly, although most of the statistical models introduced by Fisher were based random samples (IID). The way he recast modern statistics was to turn Pearson's approach on its head. Instead of commencing with the raw data $\mathbf{x}_0 := (x_1, \ldots, x_n)$ in search of a statistical model, he would view data as a typical realization of a prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ (he called a 'hypothetical infinite population') and answer the question: "Of what population is this a random sample?" (p. 313). This is not just a re-organization of Pearson's approach, but a complete reformulation of statistical induction from generalizing observed 'events' described by summary statistics to unobserved data events, to modeling the underlying 'process' in the form of a stochastic mechanism $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ that gave rise to data $\mathbf{x}_0$, and not to summarize/describe $\mathbf{x}_0$.

Modern model-based frequentist inference revolves around a prespecified *parametric statistical model*, generically defined by:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) = \{f(\mathbf{x}; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m\},\ \mathbf{x} \in \mathbb{R}^n_X,\ n > m, \tag{3}$$

where $f(\mathbf{x}; \boldsymbol{\theta}),\ \mathbf{x} \in \mathbb{R}^n_X$ denotes the joint distribution of the sample $\mathbf{X} := (X_1, \ldots, X_n)$, $\mathbb{R}^n_X$ denotes the sample space and Θ the parameter space. This represents a statistical generating mechanism specified in terms of the observable stochastic process $\{X_t,\ t \in \mathbb{N} := (1,2, \ldots, n, \ldots)\}$ underlying data $\mathbf{x}_0 := (x_1, \ldots, x_n)$. The unknown parameters $\boldsymbol{\theta}$ are viewed as *constants* and the interpretation of probability is frequentist, firmly anchored on the Strong Law of Large Numbers (SLLN). As argued in Spanos (2013), some of the criticisms of the frequentist interpretation of probability, including (i) the circularity of its definition, (ii) its reliance on 'random samples', (iii) its inability to assign 'single event' probabilities, and (iv) the 'reference class' problem (Salmon, 1967, Hajek,(2007), stem from conflating the model-based frequentist interpretation anchored on the SLLN with the von Mises (1928) interpretation. In Bayesian statistics, by contrast, $\boldsymbol{\theta}$ is viewed as a random variable (vector) and probability is interpreted as 'degrees of belief'.

The *primary objective* of frequentist inference is to use the sample information, as summarized by $f(\mathbf{x}; \boldsymbol{\theta}),\ \mathbf{x} \in \mathbb{R}^n_X$, in conjunction with data $\mathbf{x}_0$, to *narrow down* Θ as much as possible, ideally, to a single point:

$$\mathcal{M}^*(\mathbf{x}) = \{f(\mathbf{x}; \boldsymbol{\theta}^*)\},\ \mathbf{x} \in \mathbb{R}^n_X,$$

where $\boldsymbol{\theta}^*$ denotes the 'true' value of $\boldsymbol{\theta}$ in Θ; 'the true value of a parameter $\boldsymbol{\theta}$', in this context, is shorthand for saying that the generating mechanism specified by $\mathcal{M}^*(\mathbf{x})$ could have generated data $\mathbf{x}_0$. In practice, this ideal situation is unlikely to be reached, except by happenstance, but that does not preclude learning from $\mathbf{x}_0$. Learning from data about $\boldsymbol{\theta}^*$ is often referred to as accurate 'identification' of the generating mechanism $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ that could have given rise to $\mathbf{x}_0$.

Example 1. The *simple Normal model* is specified by:

$$X_t \backsim \text{NIID}(\mu, \sigma^2), \boldsymbol{\theta} := (\mu, \sigma^2) \in \Theta := (\mathbb{R} \times \mathbb{R}_+),\ x_t \in \mathbb{R},\ t \in \mathbb{N}\}, \tag{4}$$

where $\mu = E(X_t)$, $\sigma^2 = Var(X_t)$, and NIID are the assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

Example 2: The simple Bernoulli model:

$$X_k \backsim \text{BerIID}(\theta,\theta(1-\theta)),\ x_k=0{,}1,\ E(X_k)=\theta\in[0{,}1],\ Var(X_k)=\theta(1-\theta),\ k\in\mathbb{N}. \quad (5)$$

The initial choice (specification) of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is based on rendering $\mathbf{x}_0$ a typical realization thereof, or equivalently, the probabilistic assumptions selected for the stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ would reflect the chance regularity patterns exhibited by data $\mathbf{x}_0$. The search for patterns is not as unyielding as it might seen at first sight because that there are three broad categories of chance regularity patterns and corresponding probabilistic assumptions: distribution, dependence and heterogeneity; see Spanos (2006b). It is worth noting that the simple Normal model in (4) has one probabilistic assumption from each category, and the same applies to all statistical models in the model-based ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$) approach.

What is particularly interesting from the philosophy of science perspective is that Fisher's specification process echoes Charles Saunders Peirce's process of *abduction* " ... there are but three elementary kinds of reasoning. ... The first, which I call abduction ... consists in examining a mass of facts and in allowing these facts to suggest a theory. In this way we gain new ideas; but there is no force in the reasoning." (8.209).[2] "Abduction is the process of forming an explanatory hypothesis. It is the only logical operation which introduces any new idea; for induction does nothing but determine a value, and deduction merely evolves the necessary consequences of a pure hypothesis." (5.172)

One can make a strong case that the specification of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ relates directly to Peice's abduction in the sense that 'examining a mass of facts' comes in the form of detecting the chance regularity patterns exhibited by data $\mathbf{x}_0$, and abduction suggests an explanatory hypothesis in the form of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ that comprises the probabilistic assumptions aiming to account for these regularities. Also, the next step of validating the initial choice is in sync with that of Fisher when Peirce argues that: " A hypothesis adopted by abduction could only be adopted on probation, and must be tested." (7.202). Hence, the crucial role of Mis-Specification (M-S) testing; testing the validity of the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$. Related to that is another important insight about induction from Peirce: "... [inductive] reasoning tends to correct itself, and the more so the more wisely its plan is laid. Nay, it not only corrects its conclusions, it even corrects its premises." (5.575); see Mayo (1996). That is, the key to model-based statistical induction consists in 'selecting $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ wisely' to account for all the chance regularities in data $\mathbf{x}_0$, combined with validating its premises.

This insight has not been heeded by modern statisticians, even though the early pioneers were clear about the importance of validating $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; see Neyman (1952), p. 27.

Example 1 (continued). In the case of (4), the NIID assumptions need to be validate vis-a-vis data $\mathbf{x}_0$.

Diagram 1 depicts the form of model-based induction described above, with $\mathbb{Q}(\boldsymbol{\theta};\mathbf{x})$ denoting inferential propositions, such as optimal (i) estimators, (ii) confidence intervals and (iii) tests, that are derived *deductively* from $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

Example 1 (continued). $\mathbb{Q}(\boldsymbol{\theta};\mathbf{x}),\ \mathbf{x}\in\mathbb{R}_X^n$ could refer to the estimators $\overline{X}_n=\frac{1}{n}\sum_{i=1}^n X_i,\ s^2=\frac{1}{n-1}\sum_{i=1}^n(\overline{X}_n-X_i)^2$ being optimal because they satisfy certain properties, such as unbiasedness, consistency, efficiency and sufficiency; these properties stem from their sampling distributions (Lehmann and Romano, 2005):

[2] All references to Peirce are to his Collected Papers, and are cited by volume and paragraph number; see Burks (1958).

$$\overline{X}_n \backsim \mathrm{N}(\mu, \frac{\sigma^2}{n}),\ [\frac{(n-1)s^2}{\sigma^2}] \backsim \chi^2(n-1), \tag{6}$$

where $\chi^2(m)$ denotes a chi-square distribution with m degrees of freedom.

Fisher's enduring contributions to model-based induction includes devising a general way to 'operationalize' the reliability of inference by (a) *deductively* deriving error probabilities from $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, and (b) providing a measure of the procedure's 'effectiveness' in learning from data about $\boldsymbol{\theta}^*$. The form of induction envisaged by Fisher and Peirce is one where the reliability of the inference is stemming from the 'trustworthiness' of the inference procedure – how often it errs; see Mayo (1996).

$lrclrlStatisticalModel: \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}) = \{f(\mathbf{z};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m\},\ \mathbf{x} \in$

$\mathbb{R}^n_X 1c\Uparrow$ ABDUCTION$lData: \mathbf{z}_0 = (z_1, z_2, ..., z_n)c^{\text{DEDUCTION}}$ Inferential propositions: $\mathbb{Q}(\boldsymbol{\theta};\mathbf{z}) \longleftarrow$

Inference:

$lReal-worldphenomenonofinterest\overset{\Longrightarrow}{\swarrow}$ $\mathbb{Q}(\boldsymbol{\theta};\mathbf{z}_0)$ $1cDiagram1: Model-$

$basedstatisticalINDUCTION$

The inferential propositions in $\mathbb{Q}(\boldsymbol{\theta};\mathbf{z})$ are *deductively valid* when the truth of the premises ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$) ensures the truth of the conclusions ($\mathbb{Q}(\boldsymbol{\theta};\mathbf{x})$), assured by valid mathematical derivations. $\mathbb{Q}(\boldsymbol{\theta};\mathbf{z})$ is rendered *sound* by securing the *statistical adequacy* of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ – the validity of its probabilistic assumptions vis-a-vis data $\mathbf{x}_0$ – which can be established by thorough misspecification testing. Statistical adequacy in turn guarantees the statistical reliability of the inference results $\mathbb{Q}(\boldsymbol{\theta};\mathbf{x}_0)$ based on $\mathbf{x}_0$; see Spanos (2019).

Example 1 (continued). $\mathbb{Q}(\boldsymbol{\theta};\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^n_X$ represents sound inferential propositions only when the NIID assumptions are valid; see Spanos (2019).

Fisher (1922) identified the 'problems of statistics' to be: (1) **specification**, (2) **estimation** and (3) **distribution,** and emphasized that addressing (2)-(3) depends crucially on dealing with (1) successfully first. That is, the key to learning from data is an apropos specification: 'how appropriate' (or wise per Peirce) the initial selection of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is. Fisher (1922, 1925) laid the foundations of an optimal theory of (point) estimation introducing most of the desirable properties. Under distribution Fisher included all forms of inferential propositions based on sampling distributions of estimators and test statistics, including "statistics designed to test the validity of our specification." (p. 8).

In an attempt to address the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, error statistics *refines* the Fisher's approach to frequentist inference by separating *the modeling* from the *inference facet.* The modeling facet includes the *specification*, *estimation*, *M-S testing*, and *respecification* with a view to arrive at a statistically adequate model. This is because the *inference* facet presumes the validity of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ when posing substantive questions of interest to the data; see Mayo and Spanos (2004), Spanos (2018). This ensures that the inference procedures enjoy the optimal properties invoking the validity of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; see Spanos (2019).

The effectiveness and reliability of inference procedures is evaluated using ascertainable *error probabilities* stemming from the sampling distribution $f(y_n;\boldsymbol{\theta})$, of statistics (estimator, test, predictor) of the form $Y_n = g(X_1, X_2, ..., X_n)$ derived via:

$$F_n(Y_n \leq y) = \underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ h(\mathbf{x})\leq y\}} f(\mathbf{x};\boldsymbol{\theta})d\mathbf{x},\ \forall y \in \mathbb{R}, \tag{7}$$

The value of $\boldsymbol{\theta}$ in (7) is always prespecified taking two different forms stemming from the underlying reasoning:

(i) **Factual** (estimation and prediction): the true value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^*$, whatever that happens to be in $\boldsymbol{\Theta}$. Confidence Intervals (CIs) are derived under $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

(ii) **Hypothetical** (hypothesis testing): various hypothetical scenarios based on $\boldsymbol{\theta}$ taking different prespecified values under H_0: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ vs. H_1: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, where $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1 = \boldsymbol{\Theta},\ \boldsymbol{\Theta}_0 \cap \boldsymbol{\Theta}_1 = \varnothing$; the relevant error probabilities include the type I and II, the power as well as the p-value; see Spanos (2019).

The effectiveness of frequentist inference is defined in terms of the optimal properties of a statistic (estimator, test, predictor) $Y_n = g(\mathbf{X})$, and framed in terms of its sampling distribution $f(y_n;\boldsymbol{\theta}),\ y_n \in \mathbb{R}$. These optimal properties, however, assume that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is statistically adequate: its probabilistic assumptions are valid for $\mathbf{x}_0$.

Unreliability of inference. When any of these assumptions are invalid, $f(\mathbf{x};\boldsymbol{\theta})$ will be erroneous, the *optimality* of the statistic $Y_n = g(\mathbf{X})$ and the *reliability* of any inference based on it –the approximate equality of the actual error probabilities with the nominal ones – will be undermined. Applying a $.05$ significance level test when the actual type I error (due to a misspecified $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$) is closer to $.97$, will lead that inference astray by inducing *inconsistency* in estimators and/or sizeable *discrepancies* between the actual and nominal (assumed) error probabilities (type I, II, p-values).

Simulation example - Spanos and McGuirk (2001). To get some idea how misleading the inferences can be when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is misspecified, consider the case of the Linear Regression (LR) model:

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t,\ (\varepsilon_t|X_t = x_t) \backsim \text{NIID}(0,\sigma^2),\quad t \in \mathbb{N}, \tag{8}$$

(see table 4 for more details relating to the assumptions), where data $\mathbf{z}_0 := \{(x_t, y_t),\ t = 1,\ldots,100\}$ are replicated ($N = 10000$) by simulation under two scenarios. In scenario 1, all the LR probabilistic assumptions [1]-[5] are valid, and in scenario 2 assumption [5] is invalid ([1]-[4] are valid), stemming from mean-heterogeneity exhibited by $\mathbf{z}_0$ (e.g. the term $.14t$ is missing from (8)). The estimate of β_0 is $\hat{\beta}_0 = .228(.315)$ with its standard error in brackets indicating that β_0 is statistically insignificant since $\tau_{\beta_0}(\mathbf{z}_0) = .724$ and the p-value is $p(\mathbf{z}_0) = .470$, when the true value is $\beta_0^* = 1.5$. Also the nominal error type I probability is $\alpha = .05$ but actual one is $.968$. On the other hand, $\hat{\beta}_1 = 1.989(.015),\ \tau_{\beta_0}(\mathbf{z}_0) = 298.4,\ p(\mathbf{z}_0) = .0000$, when the true value is $\beta_1^* = 0.5$, and the actual type I error probability is 1.0 – rejecting a true null hypothesis 100% of the time; see Spanos and McGuirk (2001). It is important to note that most of the published results using the LR model are likely to have more than one invalid assumption among [1]-[5]!

It is important to emphasize that when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is statistically misspecified, it will undermine not just frequentist inference, but also Bayesian since the posterior is defined by $\pi(\boldsymbol{\theta}|\mathbf{x}_0) \propto \pi(\boldsymbol{\theta}) \cdot f(\mathbf{x}_0;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta$, where $\pi(\boldsymbol{\theta})$ is the prior. It will also undermine Akaike-type model selection procedures since they depend on the likelihood $L(\boldsymbol{\theta};\mathbf{x}_0),\ \boldsymbol{\theta} \in \Theta$; see Spanos (2010a).

Modern statistical inference, as a form of induction, is based on data that exhibit inherent chance regularity patterns. They differ from deterministic regularities in so far as they cannot be accounted for (described) using mathematical equations. More specifically, chance regularities come from recurring patterns in numerical data that can be accounted for (modeled) using probabilistic assumptions from three broad categories: Distribution, Dependence and Heterogeneity; see Spanos (2019).

Model-based statistical induction differs from other forms of induction, such as induction by enumeration (Henderson, 2020), in three crucial respects.

First, the inductive premises of inference, $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, represents a stochastic generating mechanism that could have given rise to data $\mathbf{x}_0$ that provides the cornerstone for the ampliative dimension of model-based induction; see Spanos (2013). This should be contrasted with enumerative induction and Pearson's descriptive statistics which rely on the straight rule and summarizing the data $\mathbf{x}_0$. This relates to Hacking's (1965) questions Salmon's claim about the straight rule: "Salmon and Reichenbach maintain that *if long-run frequencies exist, the straight rule for estimating long-run frequencies is to be preferred to any rival estimator*. Other propositions are needed to complete their vindication of induction, but only this one concerns us. Salmon claims to have proved it. This is more interesting than mere academic vindications of induction; practical statisticians need good criteria for choosing among estimators, and, if Salmon were right, he would have very largely solved their problems, which are much more pressing than Hume's." (p. 261).

Example 2 (continued). Viewing the straight rule in the context of model-based statistics, where $\mathcal{M}_{\theta}(\mathbf{x})$ is the simple Bernoulli model in (5), where with $\mathbb{P}(X_k = 1) = \theta$ and $\mathbb{P}(X_k = 0) = (1 - \theta)$, the straight rule ration $(m/n) = \frac{1}{n}\sum_{k=1}^{n} x_k$, and thus $\widehat{\theta}(\mathbf{x}_0) = (m/n)$ constitutes an estimate of θ, the observed value of the estimator $\widehat{\theta}(\mathbf{X}) = \frac{1}{n}\sum_{k=1}^{n} X_k$ when evaluated at the data point $\mathbf{x}_0$. Hence, in the context of $\mathcal{M}_{\theta}(\mathbf{x})$ in (5), $\widehat{\theta}(\mathbf{X})$ is the Maximum Likelihood estimator of θ and enjoys all optimal properties, including unbiasedness, full efficiency, sufficiency and consistency; see Spanos (2019). Regrettably, the optimality of any estimator $\widehat{\theta}(\mathbf{X})$ does not entail the claim $\widehat{\theta}(\mathbf{x}_0) \simeq \theta^*$, for a large enough n. Therefore, the straight rule, when viewed in the context of model-based inference, is just a fallacious assertion. This unwarranted claim undermines the appropriateness of estimation-based effects sizes that are widely viewed as a replacement for p-values in the current discussions on the replicability and trustworthiness of evidence; see Spanos (2020).

Second, in the context of model-based induction Hacking's (1965) "Other propositions needed to complete their vindication of induction" include (i) the validity of the inductive premises (IID) for data $\mathbf{x}_0$ that ensures the trustworthiness of evidence, as well as (ii) the optimality of the particular estimator $\widehat{\theta}(\mathbf{X})$, that secures the effectiveness of the inference; both issues lie at the core of *inductive (statistical) inference*: how we learn from data about phenomena of interest.

Third, the justification of model-based induction does not invoke *a priori* stipulations such as the 'uniformity' of *nature* and the 'representativeness' of the *sample*, as in the case of enumerative induction and Karl Pearson's curve-fitting, but relies on establishing the validity of model assumptions using comprehensive Mis-Specification (M-S) testing. As Fisher (1922) argued: "For empirical as the specification of the hypothetical population [statistical model] may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts." (p. 314).

3.2 Model-based frequentist statistics: foundational issues

Model-based frequentist statistics, as cast by Fisher (1922, 1925) and extended by Neyman and Pearson (1933), and Neyman (1937), has been plagued by several foundational problems that have bedeviled its proper implementation since the 1930s, including the following two.

Foundational issue 1. How one could secure statistical adequacy: the validity of the probabilistic assumptions comprising the chosen $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$.

The statistics and econometric literature paid little attention to the systematic testing of the validity of the model assumptions (M-S testing), and what would one do when any of the assumptions are found wanting (respecification); see Spanos (1986).

Foundational issue 2. When data $\mathbf{x}_0$ provides good evidence for or against a hypothesis or an inferential claim? (Mayo, 1996). Fisher's p-value and Neyman-Pearson's accept/reject H_0 results did not provide a coherent evidential interpretation that could address this question.

Error statistics refines the Fisher recasting of frequentist inference by embracing the distinction between *the modeling* and the *inference facet*, to address issue 1; see Mayo and Spanos (2004). In an attempt to address issue 2, error statistics *extends* the F-N-P approach by distinguishing between *pre-data* and *post-data* phases of frequentist testing to supplement the original framing with a post-data severity evaluation of testing results. This provides a sound *evidential account* that can be used to address several misconceptions and problems raised about F-N-P testing, including the large n problem; see Mayo and Spanos (2006).

Foundational issue 3. What is the nature of the reasoning underlying frequentist inference? Spanos (2012) made a case for two types of reasoning. *Factual reasoning* (used in estimation and prediction), under $\theta = \theta^*$, whatever the value θ^* happens to be in Θ, which underlines the evaluation of the sampling distributions of estimators, pivotal functions and predictors. *Hypothetical reasoning* (used in testing), underlying the evaluation of sampling distributions of test statistics under the scenarios, (null) $H_0: \theta \in \Theta_0$, and (alternative) $H_1: \theta \in \Theta_1$. Parenthetically, these forms of reasoning underlying frequentist inference are at odds with the *universal* reasoning, for all $\theta \in \Theta$, underlying Bayesian inference; see Spanos (2017).

Example 1 (contained). For the simple Normal model in (4), assuming σ^2 is <u>known</u> for simplicity, (6) implies that:

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \backsim \text{N}(0,1). \tag{9}$$

What is not so obvious is how to interpret (9), since $d(\mathbf{X}; \mu) = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ involves the unknown parameter μ, and why $E(d(\mathbf{X}; \mu)) = 0$ is not apparent. A simple answer is that since $\overline{X}_n$ is an unbiased estimator of μ, i.e. $E(\overline{X}_n) = \mu^*$, and thus $E(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}) = 0$. For that to be the case, however, (9) must be evaluated under $\mu = \mu^*$, which is known as *factual reasoning*. Hence, a more informatory way to specify (9) is:

$$d(\mathbf{X};\mu^*)=\frac{\sqrt{n}(\overline{X}_n-\mu^*)}{\sigma}\overset{\mu=\mu^*}{\backsim}\mathrm{N}(0,1). \tag{10}$$

For estimation and prediction the underlying reasoning is factual. For hypothesis testing, however, the reasoning is hypothetical and take the form:

$$d(\mathbf{X};\mu_0)=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\overset{\mu=\mu_0}{\backsim}\mathrm{N}(0,1),\quad d(\mathbf{X};\mu_1)=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\overset{\mu=\mu_1}{\backsim}\mathrm{N}(0,\delta_1), \tag{11}$$

where $\delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$, for $\mu_1\neq\mu_0,\ \mu_i\in\mathbb{R},\ i=0{,}1$; see Spanos (2019).

3.2.1 Estimation (point and interval)

Example 1 (continued). For the simple Normal model in (4) with σ^2 known, the Maximum Likelihood (ML) estimator of μ is $\widehat{\theta}_{ML}(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n}X_i$. Its optimality revolves around its sampling distribution evaluated using *factual reasoning* ($\boldsymbol{\theta}=\boldsymbol{\theta}^*$):

$$\widehat{\theta}_{ML}(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n}X_i\overset{\mu=\mu^*}{\backsim}\mathrm{N}(\mu^*,\frac{\sigma^2}{n}). \tag{12}$$

It can be shown that (12) implies that $\widehat{\theta}_{ML}(\mathbf{X})$ is unbiased, sufficient, fully efficient, and strongly consistent; see Lehmann and Romano (2005).

Confidence Intervals (CIs), $[L(\mathbf{X}),\ U(\mathbf{X})]$ are evaluated in terms of their capacity measured by the coverage probability $(1-\alpha)$ to overlay θ^* between the lower and upper bounds (Neyman, 1952):

$$\mathbb{P}(L(\mathbf{X})\leq\theta<U(\mathbf{X});\ \mu=\mu^*)=1-\alpha.$$

Example 1 (continued). For (4) with σ^2 known, the $(1-\alpha)$ CI takes the form:

$$\mathbb{P}(\overline{X}_n-c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})\leq\mu<\overline{X}_n+c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}});\ \mu=\mu^*)=1-\alpha, \tag{13}$$

stemming from the distribution of the pivot:

$$d(\mathbf{X};\mu)=\frac{\sqrt{n}(\overline{X}_n-\mu^*)}{\sigma}\overset{\mu=\mu^*}{\backsim}\mathrm{N}(0,1). \tag{14}$$

3.2.2 Neyman-Pearson (N-P) testing

Example 1 (continued). Consider testing the hypotheses:

$$H_0\colon\mu\leq\mu_0\text{vs.}\,H_1\colon\mu>\mu_0, \tag{15}$$

in the context of the simple Normal model in (4) with σ^2 known. It is important to emphasize that the framing of H_0 and H_1 should constitute a partition of $\mathbb{R}$, because for N-P testing the whole range of values of μ is relevant for statistical inference purposes, irrespective of whether only a few values are of substantive interest.

A α-significance level Uniformly Most Powerful (UMP) test is defined by (Lehmann and Romano, 2005):

$$T_\alpha := \{d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma},\ C_1(\alpha) = \{\mathbf{x}: d(\mathbf{x}) > c_\alpha\}\}, \tag{16}$$

where c_α is the α-significance level threshold based on:

$$d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{s} \overset{\mu=\mu_0}{\backsim} \mathrm{N}(0{,}1). \tag{17}$$

The type I error probability and the p-value are evaluated using (17):

$$\mathbb{P}(d(\mathbf{X}) > c_\alpha;\ \mu = \mu_0) = \alpha, \quad \mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0);\ \mu = \mu_0) = (\mathbf{x}_0).$$

The power of T_α, defined by:

$$\mathcal{P}(\mu_1) = \mathbb{P}(d(\mathbf{X}) > c_\alpha;\ \mu = \mu_1),\ \mathrm{forall}\mu_1 > \mu_0, \tag{18}$$

is based on the distribution:

$$d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma} \overset{\mu=\mu_1}{\backsim} \mathrm{N}(\delta_1, 1),\ \delta_1 = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}, \mathrm{forall}\mu_1 > \mu_0, \tag{19}$$

where δ_1 is the non-zero mean parameter. It is important to emphasize that the power of a test provides a measure of its *generic* (for any sample value $\mathbf{x} \in \mathbb{R}_X^n$) *capacity* to detect discrepancies from H_0. Also, none of the above error probabilities (type I, II, power) are conditional on values of μ since it is neither an event nor a random variable; $d(\mathbf{X})$ is evaluated under *hypothetical* values of θ; see Spanos (2019).

Two crucial features of N-P testing are often flouted by statistical textbooks and practitioners alike giving rise to several confusions and misinterpretations. These features can be found in the classic paper by Neyman and Pearson (1933) who proposed two crucial preconditions for the effectiveness of N-P testing in learning from data which relate to the framing of hypotheses: (i) H_0 and H_1 should constitute a partition of Θ, in a way that renders (ii) the type I error probability as the most serious; see also Neyman (1952). The partition of Θ is crucial in light of the primary objective of frequentist inference since the 'true' value θ^* might lie outside the union of H_0 and H_1, turning an N-P test into a wild goose chase.

Example 1 (continued). For the simple Normal model in (4), Berger and Wolpert (1988) invoke the N-P lemma to frame the hypotheses as:

$$H_0: \mu = 1 \ \mathrm{vs.} H_1: \mu = -1, \tag{20}$$

Unfortunately, the N-P lemma assumes a partition $\Theta := \{\theta_0, \theta_1\}$; see Spanos (2011). Condition (ii) suggests that when no reliable information about the potential range of values of θ^* is available, the N-P test is likely to be more effective by using:

$$H_0: \theta = \theta_0 \mathrm{vs.} H_1: \theta \neq \theta_0. \tag{21}$$

When such reliable information is available, however, the N-P test will be more effective by using a directional framing for H_0 and H_1, as in (15), ensuring that H_1 includes the potential range of values of θ^* as departures from the null value, say θ_0. This is because the power of the test – its capacity to detect discrepancies from θ_0– should be defined over the range most called for, the potential range of values of θ^*.

3.3 Statistical adequacy and Mis-Specification (M-S) testing

The current state of affairs on model validation is insightfully described by Freedman (2010), p. 16: "Bayesians and frequentists disagree on the meaning of probability and other foundational issues, but both schools face the problem of model validation. Statistical models have been used successfully in the physical and life sciences. However, they have not advanced the study of social phenomena. How do models connect to reality? When are they likely to deepen understanding? When are they likely to be sterile and misleading? ... I believe model validation to be a central issue. Of course many of my colleagues will be found to disagree. For them, fitting models to data, computing standard errors, and performing significance test is "informative," even though the basic statistical assumptions (linearity, independence of errors, etc.) cannot be validated. This position seems indefensible, nor are the consequences trivial. Perhaps it is time to reconsider."

Establishing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ calls for testing the validity of its probabilistic assumptions vis-a-vis data $\mathbf{x}_0$, such as NIID in the case of (4). The most effective way to secure statistical adequacy is to separate the *modeling*, which includes (a) *specification* – the initial choice of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ –, (b) *M-S testing* and (c) *respecification* when any of its assumptions are found wanting, from the *inference* facet because (i) the latter presumes the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and (ii) they pose very different questions to the data; see Spanos (2018). The modeling facet aims to secure the validity of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and the inference facet ensures the optimality of inference procedures with a view to secure the reliability and precision of inferential results. Treating the two as a single combined inference problem is akin to conflating the construction of a boat to given specifications (modeling) with sailing it in a competitive race (inference). The two are clearly related since the better the construction the more competitive the boat, but imagine trying to build a boat from a pile of plywood in the middle of the ocean while racing it.

Since inference presupposes the validity of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, statistical adequacy needs to be secured before optimal inference procedures can be reliably employed. Neyman-Pearson (N-P) constitutes *testing within* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ aiming to *narrow down* Θ to a much smaller subset, presupposing its validity. In contrast, M-S testing poses the question whether the particular $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ could have give rise to data $\mathbf{x}_0$, for any value of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and constitutes *testing outside* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ since the default null is $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is valid vs. its negation $\neg\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) := [\mathcal{P}(\mathbf{x}) - \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$, i.e. some other statistical model in $[\mathcal{P}(\mathbf{x}) - \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$, where $\mathcal{P}(\mathbf{x})$ is the set of all possible statistical models that could have given rise to $\mathbf{x}_0$. The problem is practice is how to operationalize $[\mathcal{P}(\mathbf{x}) - \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ to render possible comprehensive M-S testing; see Spanos (2018).

4 Empirical modeling in econometrics

4.1 Traditional curve-fitting and respecification

Empirical modeling across different disciplines involves an intricate blending of *substantive* subject matter and *statistical information*. The substantive information stems form a theory or theories about the phenomenon of interest, and could range from simple tentative conjectures to intricate *substantive* (structural) models, say $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, framed in terms of mathematical equations formulating the theory that are estimable in light of the available data $\mathbf{Z}_0 := (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n)$. The substantive information has an important and multifaceted role to play by demarcating the crucial aspects of the phenomenon of interest (suggesting the relevant variables and data), as well as enhancing the learning from data when it does not belie the statistical information in $\mathbf{Z}_0$ that stems from the *chance regularity patterns* exhibited by data $\mathbf{Z}_0$. Scientific knowledge often begins with substantive conjectures based on subject matter information, but it becomes knowledge when its veracity is established by being tested thoroughly against actual data generated by the phenomenon of interest.

The Pre-Eminence of Theory (PET) perspective, which has dominated empirical modeling in economics since the early 19th century, amounts to theory-driven curve-fitting guided by probabilistic assumptions assigned to the error term, and evaluated by goodness-of-fit measures; see Reiss (2008), Spanos (2009). The assignment of probabilistic assumptions to error term terms stems from a standpoint that relationships among the variables are mathematical (deterministic) in nature, but these are subject to stochastic disturbances due to simplification, approximation, and measurement errors. The theory-driven curve(s) are framed in terms of a structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ and the aim is to use the data $\mathbf{Z}_0$ to quantify it by estimating $\boldsymbol{\varphi} \in \boldsymbol{\Phi}$. In this sense, the data $\mathbf{Z}_0$ play only a subordinate role in availing the quantification by attaching random error term(s) to transform the curves into a stochastic model amenable to statistical analysis. The traditional textbook approach to empirical modeling in economics is summed up by Pagan (1984) as follows: "Four steps almost completely describe it: a model is postulated, data gathered, a regression run, some t-statistics or simulation performance provided and another empirical regularity was forged." (p. 103)

Although his description is meant to be a witty caricature of textbook econometrics, like all perceptive parodies, it contains more than one home truth.

The first home truth is that the phenomenon of interest is rarely explicitly described so that one can evaluate the empirical findings in relation to what has been learned from data about that phenomenon.

The second home truth is that the modeling begins with a prespecified substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ – an estimable form of that theory in light of the available data – meant to provide a description/explanation of the phenomenon of interest.

The third home truth is that $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ is treated as established knowledge, and not as tentative conjectures to be tested against the data because the primary aim is to quantify $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ by estimating the unknown structural parameters $\boldsymbol{\varphi} \in \boldsymbol{\Phi} \subset \mathbb{R}^p$.

The fourth truth is that the selection of data is often ad hoc, in the sense that the theory variables are assumed to coincide with the particular data $\mathbf{Z}_0$ chosen. No attempt is made to (i) compare the theoretical variables, often defined in terms of intentions of individual economic agents stemming out of an optimization problem, but the data refer to observed quantities and prices generated by the market, and (ii) provide a cogent bridging of the gap between them; see Spanos (2015).

The fifth truth is that the estimation of $\boldsymbol{\varphi}$ amounts to foisting $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ onto the data by viewing it as curve-fitting guided by probabilistic assumptions assigned to the error term, to be evaluated using goodness-of-fit measures.

The sixth truth is that the estimated $\boldsymbol{\varphi}$ of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ and the associated statistics, such as tratios and goodness-of-fit measures, are usually taken at face value without any attempt to secure their reliability. Indeed, the validity of the probabilistic assumptions ascribed to the error term is treated as an afterthought that determines the estimation method for the curve-fitting, and if any departures from these assumptions are indicated by the computer program output, such as a Durbin-Watson (D-W) statistic close to zero, all one has to do is to modify the original estimation method to 'account' for the departure designated by the alternative hypothesis of the D-W test. To be more specific, for the LR model in (8) the DW test operationalizes $[\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$ by embedding (8) into:

$$Y_t=\beta_0+\beta_1 x_t+u_t,\ u_t=\rho u_{t-1}+\varepsilon_t\ \ (\varepsilon_t|X_t=x_t)\backsim \text{NIID}(0,\sigma_\varepsilon^2),\ t\in\mathbb{N}, \qquad (22)$$

and testing the hypotheses: $H_0:\rho=0$ vs. $H_1:\rho\neq 0$. When H_0 is rejected, the traditional respecification is to accept H_1, i.e. adopt (22) as the respecified model. This is fallacious because rejecting H_0 entitles one to infer that $E(u_t u_s|X_t=x_t)\neq 0$ for $t>s$, but not that $E(u_t u_s|X_t=x_t)=(\rho^{|t-s|}/(1-\rho^2))\sigma_\varepsilon^2,\ t,s=1,2,\ldots,n.$ Such a claim will require one to estimate (22) and test all its probabilistic assumptions to ensure statistical adequacy; see McGuirk and Spanos (2009). Indeed, this traditional respecification strategy constitutes a quintessential example of the *fallacy of rejection*: (mis)interpreting reject H_0 [evidence against H_0] as evidence *for* a particular H_1; see Spanos (2019).

A related fallacy is that *of acceptance*: (mis)interpreting a large p-value or accept H_0 [no evidence against H_0] as evidence for H_0 – this can arise when a test has very low power (e.g. small n).

$Table1\colon AR(1)model\colon traditionalspecification$

$$y_t=\alpha_0+\alpha_1 y_{t-1}+u_t,\quad t\in\mathbb{N}.1$$

$$(i)E(u_t)=0,\qquad (ii)\sup_t E|u_t|^{\delta+\varepsilon}<0\,for\ \delta>2,\qquad \varepsilon>0,$$

$$(iii)\ \lim_{n\to\infty}E(\frac{1}{n}(\sum_{t=1}^{n}u_t)^2=\sigma_\infty^2>0,$$

$$(iv)\{u_t,\qquad t\in\mathbb{N}\}\ is\ strongly\ mixing\ with\ \ \alpha_m\underset{m\to\infty}{\to}0\ such\ that\sum\nolimits_{m=1}^{\infty}\alpha_m^{1-\delta/2}<\infty.$$

A case in point is the literature of unit root testing in the context of the AR(1) model traditionally specified as in table 1; see Choi (2015), p. 21. As Phillips and Xiao (1998) show, the unit root test based on $H_0\colon\alpha_1=1$ vs. $H_1\colon\alpha_1<1$, has very low power ($<.33$) for $n\leq 100$, and thus, the null is often erroneously accepted. Worse still, none of the invoked probabilistic assumptions (i)-(iv) (table 1) are testable with data $\mathbf{y}_0$, and thus the literature on unit root testing largely ignores the statistical misspecification problem; see Andreou and Spanos (2003).

$Table2\colon Normal, AutoRegressive\big(AR(1)\big)ModelllStatisticalGM\colon$

$$\begin{aligned} y_t = \alpha_0 + \alpha_1 y_{t-1} & \\ & + u_t, \quad t[1]Normality\colon (y_t, y_{t-1}) \backsim \mathrm{N}(.,.),2]Linearity\colon E(y_t|\sigma(y_{t-1})) \\ & = \alpha_0 + \alpha_1 y_{t-1}, 3]Homoskedasticity\colon Var(y_t|\sigma(y_{t-1})) = \sigma_0^2, 4]Markov\colon \{y_t, \\ & t \\ & \in \mathbb{N}\}isaMarkovprocess, 5]t \\ & - invariance\colon (\alpha_0, \alpha_1, \sigma_0^2)arenotchangingwitht, \}t \in \mathbb{N}. \alpha_0 \\ & = E(y_t) - \alpha_1 E(y_{t-1}) \in \mathbb{R}, \qquad \alpha_1 = \frac{Cov(y_t, y_{t-1})}{Var(y_{t-1})} \in (-1,1), \\ & \sigma_0^2 = Var(y_t) - \frac{Cov(y_t, y_{t-1})^2}{Var(y_{t-1})} \\ & \in \mathbb{R}_+ Notethat\sigma(y_{t-1})denotesthesigma - fieldgeneratedbyy_{t-1}. \end{aligned}$$

The specification in table 2 brings out the inappropriateness of relying on non-testable probabilistic assumptions relating to 'as $n \to \infty$'. As argued by Le Cam (1986), p. xiv: "... limit theorems "as n tends to infinity" are logically devoid of content about what happens at any particular n. All they can do is suggest certain approaches whose performance must then be checked on the case at hand. Unfortunately the approximation bounds we could get are too often too crude and cumbersome to be of any practical use." In fact, the assumptions whose validity for $\mathbf{y}_0$ will secure the reliability of any test based on AR(1) are given in table 2. It is worth noting that when the AR(1) model is properly specified (table 2) using the probabilistic reduction $f(y_1, y_2, \ldots, y_n; \boldsymbol{\psi}) = f_1(y_1; \boldsymbol{\psi}_1) \prod_{t=2}^{n} f(y_t|y_{t-1}; \boldsymbol{\theta})$, stemming from the probabilistic assumptions Normality, Markovness and stationarity, the coefficient $\alpha_1 \in (-1,1)$ and thus $\alpha_1 = Corr(y_t, y) = 1$ lies outside its parameter space. This is not unrelated to the low power mentioned above; see Spanos (2011).

The seventh home truth is that the 'empirical regularity forged' is usually another set of statistically 'spurious' numbers added to the ever-accumulating mountain of untrustworthy evidence gracing prestigious journals, which stems primarily from the inadequate criteria used to determine success in publishing in these journals:

[i] **statistical:** goodness-of-fit/prediction, statistical significance,

[ii] **substantive:** theoretical meaningfulness, explanatory capacity,

[iii] **pragmatic:** simplicity, generality, elegance.

The problem is that the criteria [i]–[iii] do not secure the reliability of inference and the trustworthiness of the ensuing evidence. As shown in Spanos (2007a), excellent fit is neither necessary nor sufficient for statistical adequacy, because the former seeks 'small' residuals, but the latter relies on non-systematic (white-noise) residuals. The criteria [i]–[iii] are not even sufficient for the evaluation of the cogency of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ in shedding adequate light on the phenomenon of interest. The combination of [i]-[iii] neglects a fundamental criterion:

[iv] **epistemic**: empirical adequacy, that relates to both *statistical adequacy* – validating the implicit statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ vis-a-vis data $\mathbf{Z}_0$, as well as *substantive adequacy* – probing the cogency of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ vis-a-vis the phenomenon of interest. Let us unpack this claim.

4.2 Traditional econometric techniques

The dominance of the Pre-eminence of Theory (PET) perspective in applied economics and econometrics seem to have largely ignored Fisher's (1922) paradigm shift of recasting Karl Pearson's descriptive statistic since it shares with the latter the curve-fitting perspective evaluated by goodness-of-fit measures. The difference between curve-fitting a frequency curve vs. a structural model, $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, is not as important when the trustworthiness of evidence is a primary objective. A case in point is the anachronistic attribution of the Method of Moments to Pearson by the econometrics literature (Greene, 2018), oblivious to the fact that Pearson's method was designed for a very different paradigm, where one would begin with the data $\mathbf{z}_0$ in search of a descriptive model $f(x;\widehat{\boldsymbol{\psi}}) \in \mathcal{F}_P(x;\boldsymbol{\psi}),\ x \in \mathbb{R}$, and not lead off with a prespecified model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, assumed to have given rise to data $\mathbf{z}_0$; see Spanos (2019).

The emphasis in current applied econometrics is placed on the recipe-like mechanical implementation of inference procedures, such as Instrumental Variables (IV), Generalized Method of Moments (GMM), Vector Autoregresion (VAR), structural VAR, calibration, and matching moments. The probabilistic assumptions are assigned to error terms and treated as an afterthought when deriving Consistent and Asympotically Normal (CAN) estimators, and then forgotten at the inference facet. Indeed, the notion of probabilistic assumptions imposed on one's data could be invalid is hardly mentioned in the recent textbooks on macroeconometrics; see Canova (2007). As one would expect, Canova (2007) points out the major advancements in the mathematical, statistical and computational tools over the last 20 years in econometrics. The problem is that 'these improvements in tools' are inversely related to the trustworthiness of the empirical evidence. The empirical examples used in most recent textbooks in applied econometrics are *not* exemplars of how to do empirical modeling that give rise to learning from data, but illustrations on how to apply recipe-like procedures, ignoring the problem of securing the *trustworthiness* of the empirical evidence when employing the proposed tools.

In an attempt to justify the neglect of statistical model validation, traditional textbooks often invoke misleading robustness results. Popular examples in textbook econometrics are the Heteroskedasticity-Consistent (HC) and Autocorrelation-Consistent (AC) Standard Errors (SEs), as well as HAC SEs; see Wooldridge (2010). HAC SEs are used to justify ignoring any departures from homoskedasticity and no-autocorrelation assumptions such as [3] and [4] in table 4. Unfortunately, the claim that such SEs based on asymptotic arguments can circumvent the unreliability of inference problem is unfounded; see Spanos and McGuirk (2001). As shown in Spanos and Reade (2015), HC and HA SEs do nothing to ensure that the actual error probabilities approximate closely the nominal one. As argued above, the idea that a consistent estimator of the SE of an estimator could save an inference from unreliability stems from another misapprehension of asymptotic properties.

A strong case can be made that the published literature in prestigious journals in econometrics, pays little to no attention to statistical adequacy: validating the estimated models. There are several reasons for that, including the fact that the PET perspective dominates current practice (Spanos, 2018):

(i) Views empirical modeling as theory-driven curve-fitting guided by error-term probabilistic assumptions and evaluated using goodness-of-fit measures.

(ii) Conflates the modeling with the inference facet, ignoring the fact that they pose very different questions to the data. It's similar to conflating the construction of a boat to given specifications with sailing it in a competitive race!

(iii) Blends the statistical with the substantive information/model and neglects both statistical and substantive adequacy. Under-appreciates the potentially devastating effects of statistical misspecification on the reliability of inference for both the substantive questions of interest as well as probing for substantive adequacy.

(iv) M-S testing (probing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$) is often conflated with N-P testing (probing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$), and as a result, M-S testing is often criticized for being vulnerable to pre-test bias, double use of data, data-mining, etc.; see Spanos (2010b).

(v) Statistical respecification is viewed as 'error-fixing' based on modifying the probabilistic assumptions assigned to the error term $\{\varepsilon_t,\ t \in \mathbb{N}\}$ so that one can get 'good' estimators for the curve-fitting.

(vi) Current practice in econometrics relies unduly on asymptotics, in particular CAN estimators, and practitioners seem unaware that does not suffice to secure the reliability of inferences. Since limit theorems, such as the LLN and CLT, only tells what happens at the limit (∞), asymptotic properties are useful for their value in excluding totally unreliable estimators and tests, but they do not guarantee the reliability of inference procedures for a given data $\mathbf{Z}_0$ and n. For instance, an inconsistent estimator will give rise to unreliable inferences, but a consistent one does not guarantee their reliability.

(vii) There is a huge divide between a theoretical econometrician and a practitioner. An important contributor to the uninformed implementation of statistical procedures, such as IV, GMM and VAR, that continues unabated to give rise to untrustworthy evidence, is a subtle disconnect between the theoretician (theoretical econometrician), that leaves the practitioner hopelessly unable to assess the appropriateness of different methods for his/her particular data. The theoretician develops the statistical techniques associated with different statistical models for different types of data (time-series, cross-section, panel), and the practitioner implements them using data, often observational. As observed by Rust (2016): "It is far easier to publish theoretical econometrics, an increasingly arid subject that meets the burden of mathematical proof. But the overabundance of econometric theory has not paid off in terms of empirical knowledge, and may paradoxically hinder empirical work by obligating empirical researchers to employ the latest methods that are often difficult to understand and use and fail to address the problems that researchers actually confront." Each will do a much better job at their respective tasks if only they understood sufficiently well the work of the other. The theoretician will be more cognizant of the difficulties for the proper implementation of these tools, and make a conscious effort to elucidate their scope, applicability, and limitations. Such knowledge will enable the practitioner to produce trustworthy evidence by applying such tools only when appropriate. For instance, in proving that an estimator is CAN, the theoretician could invoke *testable* assumptions comprising the relevant $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. This will give the practitioner a chance to appraise the appropriateness of different methods and do a much better job in producing trustworthy evidence by testing the validity of the invoked assumptions; see Spanos (2018).

Unfortunately, empirical modeling in economics is currently dominated by a serious disconnect between these two since the theoretician is practicing *mathematical deduction* and the practitioner uses recipe-like *statistical induction* by transforming formulae into numbers misusing the data. The theoretician has no real motivation to render the invoked $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ testable. If anything, the motivation stemming from the perceived esteem level reflecting his/her technical dexterity is to make $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ even less testable and obtuse by invoking the misleading claim that weaker assumptions are less vulnerable to misspecification. Also, the practitioner has no real motivation to do the hard work of establishing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, given that no journal editor asks for that, and is happy to give credit/blame to the theoretician.

4.3 Traditional modeling and the trustworthiness of evidence

Despite bold assertions in book titles, such as "Mostly Harmless Econometrics" by Angrist and Pischke (2008), ignoring the probabilistic assumptions one imposes on a particular data $\mathbf{W}_0$, is anything but 'harmless', when trustworthy evidence and learning from data are important objectives in the empirical modeling. Moreover, "better research designs, either by virtue of outright experimentation or through the well-founded and careful implementation of quasi-experimental methods" (p. 26), as claimed by Angrist and Pischke (2010), will not take the 'con' out of econometrics, since the untrustworthiness of evidence stemming from imposing (implicitly or explicitly) invalid probabilistic assumptions on one's data plagues modeling with experimental data as well; see Rust (2016). A real-life example of statistical misspecification due to ignoring heterogeneity in cross-section experimental data, is the case of the sleep aid Ambien. After going through the rigorous procedures and protocols a new medical treatment has to follow before approval, and several years on the market, as well as millions of prescriptions, it was discovered (retrospectively) that female patients are more susceptible to the risk of 'next day impairment' because their body metabolizes Ambien more slowly than male patients; see Spanos (2020). If the rigorous process based on the 'gold standard' for evidence, the Randomized Controlled Trials (RCTs) for a new treatment, could not safeguard the trustworthiness of evidence from statistical misspecification, one wonders how any impromptu "better research designs" and "quasi-experimental methods" would do better; see Deaton (2010), Heckman (1997) and Reiss (2015) for further discussion.

5 Recasting curve-fitting into model-based inference

How does one secure reliability of inference and the trustworthiness of evidence when the modeling begins with a substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$? By a recasting the curve-fitting approach into a model-based induction with a view to accommodate the substantive information encapsulated by $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, but distinguishing between $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ and the statistical$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, and ensure that its probabilistic assumptions are specified in terms of the observable process $\{\mathbf{Z}_t, t \in \mathbb{N}\}$ and not the error term. This is needed to establish the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, which, in turn, will ensure the reliability of the statistical procedures used to congruously coalesce the two models into an empirical model, which is both statistically and substantive adequate.

5.1 Statistical vs. substantive models

A closer look at Fisher's (1922, 1925) recasting of statistics reveals that in his framing there is always a 'material experiment', often specified in terms of alternative experimental designs – a simple $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ – that is embedded into a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. It turns out that behind every substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ there is an implicit statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ that comprises the probabilistic assumptions imposed on data $\mathbf{Z}_0$, but one needs to bring it out explicitly and test the validity of these assumptions. This renders the current debate between structural vs. reduce form models (Low and Meghir, 2017) a false dilemma, since the reduce form of any structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ comprises the probabilistic assumptions (implicitly or explicitly) imposed on data $\mathbf{z}_0$, i.e. the built-in statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, whose statistical adequacy determines the reliability of inference of the estimated $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$.

In direct analogy to $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ the substantive model is generically specified by:

$$\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z}) = \{f(\mathbf{z}; \boldsymbol{\varphi}),\ \boldsymbol{\varphi} \in \Phi \subset \mathbb{R}^p\},\ \mathbf{z} \in \mathbb{R}^n_Z,\ p \leq m. \tag{23}$$

A congruous blending of the two models is based on relating their parameterizations $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ by ensuring that $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ is parametrically nested in $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$.

The first step in that direction is to 'transfer' the probabilistic assumptions from the error term to the observable process $\{\mathbf{Z}_t,\ t \in \mathbb{N}\}$ underlying $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, and separate the statistical from the substantive assumptions by distinguishing between statistical and substantive adequacy:

[a] **Statistical adequacy**: $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ adequately accounts for the chance regularities in $\mathbf{z}_0$, or equivalently, the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ are valid for data $\mathbf{z}_0$. It is 'local' because it relates to the particular data and their chance regularities.

[b] **Substantive adequacy**: the extent to which $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ sheds adequate light (describe, explain, predict) on the phenomenon of interest. Hence, any assumptions relating to ceteris paribus clauses, omitted variables, causality, etc., are substantive since they encode 'tentative information' about 'how the world really works'. In this sense, substantive adequacy is phenomenon-oriented because it relates to the relationship between $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ and the phenomenon of interest. Indeed, the traditional criteria [ii] substantive and [iii] pragmatic relate to the substantive adequacy. The problem is that without securing the statistical adequacy first, none of these criteria can be properly implemented in practice.

It is important to emphasize at this point that the widely invoked slogan '*All models are wrong, but some are useful*' attributed to George Box (1979), is invariably misinterpreted as suggesting that statistical misspecification is inevitable. The 'wrongness' Box refers to, however, is not statistical but substantive: "Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model." (p. 202). Box, goes on to emphasize empirical modeling as an iterative process of selecting a model, testing its probabilistic assumptions using the residuals, and respecifying it when any of them are invalid!

5.2 The tale of two linear regressions

To illustrate the difference between a statistical and a substantive perspective let us compare and contrast the traditional textbook specification of the Linear Regression (LR) model (table 3) with the model-based specification (table 4).

$Table3: LinearRegressionmodel: traditionalspecificationY_t = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t + \varepsilon_t,\quad t \in \mathbb{N},1l\{1\}(\varepsilon_t|\mathbf{X}_t = \mathbf{x}_t) \backsim \mathrm{N}(.,.),\ \{2\}E(\varepsilon_t|\mathbf{X}_t = \mathbf{x}_t) = 0,1l\{3\}\ Var(\varepsilon_t|\mathbf{X}_t = \mathbf{x}_t) = \sigma_\varepsilon^2,\ \{4\}\ Cov(\varepsilon_t\varepsilon_s|\mathbf{X}_t = \mathbf{x}_t) = 0,\ t > s,\ t,s \in \mathbb{N}.$

$Table4\text{:}\,Normal, LinearRegressionmodel\ \ 1lllStatisticalGM\text{:}\,Y_t=\beta_0+\beta_1x_t+u_t,\quad t\in\mathbb{N}\ \ ,lll[1]Normal$
$x_t)\backsim \mathrm{N}(.,.),2]Linearity\text{:}\,E(Y_t|X_t=x_t)=\beta_0+\beta_1x_t,3]Homoskedasticity\text{:}\,Var(Y_t|X_t=$
$x_t)=\sigma^2,4]Independence\text{:}\,\{(Y_t|X_t=x_t),\ t\in\mathbb{N}\}indep.process,5]t-invariance\text{:}\,\boldsymbol{\theta}\text{:}=$
$(\beta_0,\beta_1,\sigma^2)arenotchangingwitht,\ \}t\in$
$\mathbb{N}.1l\beta_0=E(y_t)-\beta_1E(X_t),\ \beta_1=(\frac{Cov(X_t,y_t)}{Var(X_t)}),\ \sigma^2=Var(y_t)-\beta_1Cov(X_t,y_t).$

In terms of their assumptions, the two specifications differ in several respects.
First, table 3 is usually supplemented by additional assumptions that include:
{5} $\mathbf{X}_t$ is fixed at $\mathbf{x}_t$ in repeated samples,
{6} All relevant variables have been included in $\mathbf{X}_t$,
{7} No collinearity: rank$(\mathbf{X}^\top\mathbf{X})=m+1,\ \mathbf{X}=(\mathbf{1},\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_m)$.

Second, the Generating Mechanism (GM) in table 3 is (implicitly) *substantive*: how the phenomenon of interest generated data $\mathbf{Z}_0\text{:}=(\mathbf{y}_0,\mathbf{X}_0)$, but the GM in table 4 is *statistical*: how the stochastic mechanism underlying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ could have generated data $\mathbf{Z}_0$. Equivalently, this could represent how one could generate data Y_t given $X_t=x_t$ on a computer using pseudo-random numbers for u_t.

Third, the error terms ε_t and u_t, associated with the two specifications in tables 3 and 4, are interpreted very differently because they represent different types of errors. For a statistical model, such as in (3) (table 4), the error term u_t is assumed to represent the *non-systematic* statistical information in data $\mathbf{Z}_0\text{:}=(\mathbf{y}_0,\mathbf{X}_0)$, neglected by the systematic component $m(t)=E(Y_t|X_t=x_t)$; more formally, $\{(u_t|\mathcal{D}_t),\ t\in\mathbb{N}\}$ is a Martingale Difference process relative to the information $\mathcal{D}_t\subset\mathfrak{J}$ of the probability space $(S,\mathfrak{J},\mathbb{P}(.))$ underlying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Hence, the statistical error term u_t is: [i] *Derived* in the sense that $u_t=Y_t-E(Y_t|X_t=x_t)$ represents the non-systematic component of the orthogonal decomposition of Y_t defining the statistical GM:

$$Y_t=E(Y_t|X_t=x_t)+u_t,$$

where by 'design' $E(u_t|X_t=x_t)=0$ and $E(m(t)\cdot u_t|X_t=x_t)=0$. Hence, the probabilistic structure of $\{(u_t|X_t=x_t),\ t\in\mathbb{N}\}$ is completely determined by that of $\{(Y_t|X_t=x_t),\ t\in\mathbb{N}\}$ ([1]-[5], table 4). This implies that when any of the assumptions [1]-[5] are invalid, u_t will include the systematic statistical information in $\mathbf{Z}_0$ unaccounted for by $m(t)$. [ii] *Data-oriented,* in the sense that its validity/invalidity (departures from assumptions [1]-[5]) revolves solely around the statistical systematic information in $\mathbf{Z}_0$.

When table 3 is viewed in the context of curve-fitting ε_t is a *structural error* term, assumed to represent the *non-systematic* substantive information unaccounted for by $Y_t=\beta_0+\beta_1x_t$. In this sense, ε_t is: [i]* *Autonomous* in the sense that its probabilistic structure also depends on other relevant substantive information that $Y_t=\beta_0+\beta_1x_t$ might have overlooked, including omitted variables, unobserved confounding factors, external shocks, and systematic errors of measurement/approximation. [ii]* *phenomenon-oriented,* in the sense that the validity of the probabilistic structure of ε_t revolves around how adequately $Y_t=\beta_0+\beta_1x_t$ accounts for the phenomenon of interest. Hence, when probing for substantive adequacy one needs to consider the different ways $Y_t=\beta_0+\beta_1x_t$ might depart from the actual data generating mechanism giving rise to the phenomenon of interest; not just the part that generated $\mathbf{Z}_0$.

Fourth, when the assumptions {1}-{4} (table 3) are viewed from a purely *probabilistic* perspective one can see that they relate directly to assumptions [1]-[4] (table 4); see Spanos (2019). In particular:

$$\{2\} E(\varepsilon_t | X_t = x_t) = 0 \Leftrightarrow [2] E(Y_t | X_t = x_t) = \beta_0 + \beta_1 x_t. \tag{24}$$

On the other hand, assumption {2} (table 3) in textbook econometrics is referred to as an *exogeneity* assumption (Greene, 2018, p. 55), which reveals that {2} is viewed from a *substantive* (curve-fitting) perspective, where a potential departure can arise when ε_t includes an omitted but relevant variable, say W_t, such that $Cov(X_t, W_t) \neq 0$, implying that ¬{2}: $E(\varepsilon_t | \mathbf{X}_t = \mathbf{x}_t) \neq 0$. This argument makes no sense when ε_t is viewed as a statistical error term (see (24)) since it has nothing to do with data $\mathbf{Z}_0$, but it does make sense when ε_t is viewed as an autonomous substantive error term $\varepsilon_t = Y_t - \beta_0 - \beta_1 x_t$ that includes any systematic substantive information neglected by $Y_t = \beta_0 + \beta_1 x_t$. This issue is particularly important in econometrics because $E(\varepsilon_t | \mathbf{X}_t = \mathbf{x}_t) \neq 0$ is used to motivate one of the most widely used (and abused – Spanos, 2007b) methods of estimation, known as the Instrumental Variables (IVs) method; see Wooldridge (2010). Another variation on the substantive departure (¬{2}) gives rise to the so-called omitted variable bias, which is erroneously viewed as a form of statistical misspecification in the econometric literature; see Spanos (2006c).

Fifth, assumptions {5}-{7} are not probabilistic assumptions that make sense in the context of a statistical model, since {5} is superfluous when X_t is viewed as a conditioning variable, {6} is a substantive assumption (Spanos, 2010c), and {7} is a condition that relates to the particular data $\mathbf{Z}_0$, and not the generating mechanism; see Spanos (2019).

Sixth, when viewed from a purely probabilistic perspective, there are two clear differences between tables 3 and 4. The first is that all assumptions [1]-[5] relate to the observable process $\{(Y_t | X_t = x_t),\ t \in \mathbb{N}\}$ and are directly testable vis-a-vis data $\mathbf{Z}_0$, with [5] missing from table 3. The second difference is the implicit statistical parameterization in table 4, indicating what 'statistical' (as opposed to substantive) parameters the unknown $\boldsymbol{\theta}$ represents. This is crucial because the statistical GM in conjunction with this parameterization separates the statistical from the substantive perspective, indicating that one does not need to invoke a substantive model to estimate the statistical model in table 4. This clear separation of the statistical and substantive models, ab initio, stems from viewing the former as a particular parameterization of the stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{N}\}$ underlying the data $\mathbf{Z}_0$. To derive the particular parameterization one can invoke the Kolmogorov extension theorem that enables one to fully describe the stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{N}\}$ using its joint distribution $D(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n; \boldsymbol{\phi})$; see Billingsley (1995). Note that the probabilistic reduction that relates

$D(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n; \boldsymbol{\phi}) \overset{\text{IID}}{=} \prod_{t=1}^{n} D(Y_t | \mathbf{x}_t; \boldsymbol{\varphi}_1) D(\mathbf{X}_t; \boldsymbol{\varphi}_2),\ \forall \mathbf{z}_t \in \mathbb{R}^{nm}$ to the distribution $D(Y_t | \mathbf{x}_t; \boldsymbol{\varphi}_1)$,

underlying the LR model in table 4, also ensures the internal consistency of assumptions [1]-[5]; see Spanos (2019).

The parameterization of $\boldsymbol{\theta}$ provides the first link between the statistical and substantive models because $\boldsymbol{\theta}$ is chosen in such a way to parametrically nest the substantive model parameters $\boldsymbol{\varphi}$. This relationship can be expressed in the generic form:

$$\mathbf{G}(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \mathbf{0},\ \boldsymbol{\varphi} \in \mathbb{R}^p,\ \boldsymbol{\theta} \in \mathbb{R}^m,\ p \leq m. \tag{25}$$

Diagram 1 can be easily extended to accommodate a substantive $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ model in addition to the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, as articulated above; see Spanos (2020).

5.3 From statistical and substantive to empirical models

As emphasized above, what renders the estimated LR model (table 4) and the associated statistical inference a statistical regularity is the validity of [1]-[5] and nothing else. It becomes an empirical regularity when a worthy substantive model explains the phenomenon of interest without belying the statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$.

Kepler's first law. Spanos (2007a) illustrates this using Kepler's 1609 statistical regularity for the motion of the planets ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$) and the substantive model ($\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$) provided by Newton almost 80 years later. In particular, Kepler's first law states that 'a planet moves around the sun in an elliptical motion with one focus at the sun'. The loci of the elliptical motion based on r-distance of the planet from the sun, and ϑ- angle between the line joining the sun and the planet and the principal axis of the ellipse. Using the polar coordinates transformations $y{:}=(1/r)$ and $x{:}=\cos\vartheta$, Kepler's first law becomes $y_t=\alpha_0+\alpha_1 x_t$, which can be estimated as a LR model in (??). Estimating (??) using the original Brahe data for Mars ($n=28$) yields:

$$y_t = \underset{(.000002)}{.662062} + \underset{(.000003)}{.061333}x_t + \hat{u}_t,\ n=28,\ R^2=.9999,\ s=.00001115, \tag{26}$$

which can be shown to be statistically adequate; see Spanos (2007a).

The substantive interpretation of Kepler's first law had to wait for Newton's (1687) *Law of Universal Gravitation* (LUG): $F=[G(m\cdot M)]/r^2$, where F is the force of attraction between two bodies of *mass* m (planet) and M (sun), G is a constant of gravitational attraction, and r is the *distance* between the two bodies. LUG attributed a clear structural interpretation to β_0 and β_1: $\beta_0=[MG/4\kappa^2]$, where κ denotes the Kepler constant, $\beta_1=([(1/d)-\beta_0]$, where d is the shortest distance between the planet and the sun; see Hahn (1998). Also, the error term ε_t enjoys a substantive interpretation in the form of 'departures' from the elliptical motion due to potential *measurement errors* and *unmodeled effects*. Hence, the assumptions {1}-{4} (table 1) could be inappropriate in cases where: (i) the data suffer from 'systematic' observation errors, (ii) the third body problem effect is significant, (iii) the general relativity terms (Lawden, 2002) are significant.

Duhem's thesis. The distinction between the statistical $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ and substantive $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ model can be used to address *Duhem's* (1914) thesis that 'no hypothesis can be tested separately from the set of auxiliary hypotheses' needed for such empirical tests. The statistical assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ are the only 'auxiliary hypotheses' needed, and their validity can be established independently of the substantive hypotheses in (25). Indeed, the adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is needed for testing the validity of such substantive hypotheses. For instance, the statistically adequate model in (26), can provide the basis for testing the Copernicus hypothesis that the motion of a planet around the sun is circular ($y_t=\alpha_0$) using the hypotheses: H_0: $\beta_1=0$ vs. H_1: $\beta_1\neq 0$, using $\tau(\mathbf{z}_0;\beta_1)=\frac{.061333}{.000003}=20444[.000000]$, which strongly rejects H_0.

The propensity interpretation of probability. The distinction between $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ and $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ models can also be used to address a conundrum associated with the propensity interpretation of probability, attributed to Popper and Peirce, as it relates to Humphreys (1985) paradox: the propensity interpretation has a built-in *causal connection* between different events, say A and B, which renders reversing conditional probabilities such as $\mathbb{P}(A|B)$ to $\mathbb{P}(B|A)$ meaningless when A is the effect and B is the cause. The paradox goes away by noting that the propensity interpretation is associated with real-world stochastic mechanisms, such as a radioactive atom has a 'propensity to decay' that gives rise to stable relative frequencies. This suggests that the mechanism is viewed as a substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ that carries with it substantive information, including causal. Thus, even though the statistical information encapsulated in $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ satisfies all the rules of conditional probability, in the context of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ the substantive causal information imposes additional restrictions (including causal) on $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ which are often testable via (25); see Spanos (2019). Hence, there is no conflict between the frequentist and propensity interpretations of probability, as the former is germane to the statistical $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and the latter to the substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$.

5.4 Revisiting the Koopmans vs. Vining debate

Koopmans (1949), in his exchange with Vining (1949), used the historical episode of Kepler's statistical regularities concerning planetary motion to criticize the primitive state of development of empirical business cycle modeling represented by Burns and Mitchell (1946), as opposed to that of theory-driven curve-fitting modeling of the Cowles Commission; see Morgan (1990). He called the former the 'Kepler stage' of empirical modeling, in contrast to the 'Newton stage', where these statistical regularities were given a substantive interpretation using Newton's LUG.

Arguably, Koopmans did not draw the right lessons from this episode, in the sense that the inductive process best describing it is that of *data-to-theory,* because the statistical regularity of the elliptical motion of Mars around the sun was established based on (a) meager substantive information, but (b) reliable statistical information using Brahe's data, and (c) was instrumental in inspiring Newton to devise the LUG; Newton called the elliptical motion Kepler's *first law*.

The right lesson to be learned from this episode is that a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ provides the starting point of the statistical regularities in data $\mathbf{z}_0$ a worthy theory aiming to explain the particular phenomenon of interest needs to account for, and that's how Newton understood Kepler's statistical regularity.

6 Summary and conclusions

Using a philosophy of science perspective, the above discussion provides a critical view of current econometric modeling and inference with a view to provide a deeper understanding of what econometricians are engaged in and what they are trying to accomplish in empirical modeling.

The primary aim of the above discussion is to place econometrics in the broader statistical context of model-based statistical induction and focus on issues that call for conceptual clarification and coherence, detect gaps in traditional econometric arguments and frame alternative conceptual perspectives. The success of current econometric methodology has been evaluated with respect to its effectiveness in giving rise to 'learning from data' about economic phenomena of interest.

The overall assessment is that current econometric methodology has so far failed to shed sufficient light on economic phenomena, for a several reasons. The most important is that viewing empirical modeling as curve-fitting guided by impromptu stochastic error terms, and evaluated by goodness-of-fit will not give rise to learning from data. In hard sciences (physics, chemistry, geology, astronomy) curve-fitting is more successful due to several special features: (a) laws of nature are usually *invariant* with respect to the time and location. Their experimental investigation is: (b) guided by *reliable substantive knowledge* pertaining to the phenomenon of interest, (c) framed in terms of tried and trusted *procedural protocols*, and (d) empirical knowledge has a high degree of *cumulativeness*. In contrast, empirical modeling in *social sciences* pertains to (a)* fickle human behavior that is not invariant to time or location. The empirical modeling in the soft sciences (including economics) is: (b)* guided by tentative conjectures that are often misconstrued as *established knowledge*, (c)* by foisting a substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ on the data without validating the implicit $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. (d)* The end result is invariably an estimated $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ that is *statistically and substantively misspecified*; Spanos (2007a).

To meliorate the untrustworthiness of the evidence problem arising from curve-fitting, the traditional approach needs to be modified in ways that allow the systematic statistical information in data (chance regularities) to play a more crucial role than the subordinate one of 'quantifying substantive models presumed true'. Hence, the need for a much broader and more coherent modeling framework based on several nuanced distinctions, including (i) statistical vs. substantive information/model/adequacy, (ii) statistical modeling vs. inference, (iii) factual vs. hypothetical reasoning in frequentist inference, (iv) Neyman-Pearson testing (within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$) vs. M-S testing (outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$), (v) pre-data vs. post-data error probabilities, and (vi) untestable vs. testable probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. The cornerstone of this framework is the concept of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ and its adequacy. This is crucial because the combination of observational data and the absence of reliable substantive knowledge pertaining to the phenomenon of interest, a statistically adequate model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ can provide the basic benchmark for what a worthy substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ needs to explain to begin with.

The above discussion calls for certain changes in the current paradigm of econometric modeling and inference, including the overall conceptual framework, the research methods, the objectives, the professional and educational subject system, as well as the standards for what constitutes a real contribution to trustworthy evidence in applied economics. The proposed framework offers suggestions for journal editors and referees on several ways to ameliorate the untrustworthiness of published empirical evidence. First, decline forthwith papers that ignore establishing the adequacy of the invoked statistical model(s) by their inferences. Second, call out authors for uninformed implementation of inference procedures and unwarranted interpretations of their results. Third, demand that authors probe adequately for any potential substantive misspecifications, after the adequacy of the underlying statistical model has been secured. Fourth, demand from theoreticians to ensure that the probabilistic assumptions underlying their proposed tools are testable. As argued by Rust (2016): "... journals should increase the burden on econometric theory by requiring more of them to show how the new methods they propose are likely to be used and be useful for generating new empirical knowledge."

References

Andreou, E. and A. Spanos (2003) “Statistical adequacy and the testing of trend versus difference stationarity”, *Econometric Reviews*, **3**: 217-237.

Angrist, J.D. and J.S. Pischke (2008) *Mostly harmless econometrics: An empiricist’s companion*, Princeton University Press, Princeton, NJ.

Angrist, J.D. and J.S. Pischke (2010) “The credibility revolution in empirical economics: How better research design is taking the con out of econometrics”, *Journal of economic perspectives*, 24(2): 3-30.

Berger, J.O. and R.W. Wolpert (1988), *The Likelihood Principle*, Institute of Mathematical Statistics, Lecture Notes - Monograph series, 2nd edition, vol. 6, Hayward, CA.

Billingsley, P. (1995) *Probability and Measure*, 4th ed., Wiley, NY.

Box, G.E.P. (1979) “Robustness in the Strategy of Scientific Model Building,” in *Robustness in Statistics*, ed. by Launer, R. L. and G. N. Wilkinson, Academic Press, NY.

Burks, A.W. (1958), editor, *Collected Papers of Charles Saunderss Peirce*, volumes I-VIII, Harvard University Press, Cambridge, MA.

Burns, A.F. and W.C. Mitchell (1946) *Measuring Business Cycles*, NBER, NY.

Canova, F. (2007) *Methods of Macroeconometric Research*, Princeton University Press, Princeton, NJ.

Choi, I. (2015) *Almost all about unit roots: Foundations, developments, and applications*, Cambridge University Press, Cambridge.

Deaton, A. (2010) Instruments, randomization, and learning about development”, *Journal of economic literature*, 48(2): 424-55.

Dickhaus, T. (2018) *Theory of nonparametric tests*, Springer, NY.

Doob, J.L. (1953) *Stochastic Processes*, Wiley, NY.

Duhem, P. (1914) *The Aim and Structure of Physical Theory*, English translation published by Princeton University Press, Princeton.

Fisher, R. A. (1922) "On the mathematical foundations of theoretical statistics", *Philosophical Transactions of the Royal Society A*, **222**: 309-368.

Fisher, R. A. (1925) "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*, **22**: 700-725.

Freedman, D.A. (2010) *Statistical Models and Causal Inference,* Cambridge University Press, Cambridge.

Greene, W. H. (2018) *Econometric Analysis*, 8th ed., Prentice Hall, NJ.

Hacking, I. (1965) "Salmon's Vindication of Induction", *The Journal of Philosophy*, 62(10): 260-266.

Hacking, I. (1980) "The Theory of Probable Inference: Neyman, Peirce and Braithwaite", pp. 141–60 in Mellor, D. (ed.), *Science, Belief and Behavior: Essays in Honour of Richard B. Braithwaite*, Cambridge University Press, Cambridge.

Hahn, A.J. (1998) *Basic Calculus: From Archimedes to Newton to its Role in Science*, Springer, New York.

Hajek, A. (2007) "Interpretations of Probability", in the *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/probability-interpret/.

Heckman, J.J. (1997) "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations", *Journal of Human Resources*, 32(3): 441–62.

Henderson, L. (2020) "The Problem of Induction", The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/induction-problem/>.

Hoover, K.D. (2006) “The Methodology of Econometrics,” pp. 61-87 in Mills, T.C. and K. Patterson (2006).

Hume, D. (1748) *An Enquiry Concerning Human Understanding*, Oxford University Press, Oxford.

Humphreys, P. (1985) “Why propensities cannot be probabilities”, *The Philosophical Review*, **94**, 557–570.

Kolmogorov, A.N. (1933) *Foundations of the theory of Probability*, 2nd English edition, Chelsea Publishing Co. NY.

Koopmans, T.C. (1947) “Measurement Without Theory," *Review of Economics and Statistics*, **17**: 161-172.

Lawden, D.F. (2002) *Introduction to Tensor Calculus, Relativity and Cosmology*, Dover, New York.

Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*, Springer.

Lehmann, E.L. and J.P. Romano (2005) *Testing Statistical Hypotheses*, Springer, NY.

Low, H and C. Meghir (2017) “The Use of Structural Models in Econometrics”, J*ournal of Economic Perspectives*, 31(2): 33–58.

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

Mayo, D.G. and A. Spanos (2004) “Methodology in Practice: Statistical Misspecification Testing”, *Philosophy of Science*, **71**: 1007-1025.

Mayo, D.G. and A. Spanos. (2006) “Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction,” *The British Journal for the Philosophy of Science,* **57**: 323-357.

Mayo, D. G. and A. Spanos (eds.) (2010) *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, Cambridge University Press, Cambridge.

McGuirk, A. and A. Spanos (2009) “Revisiting Error Autocorrelation Correction: Common Factor Restrictions and Granger Non-Causality,” *Oxford Bulletin of Economics and Statistics,* **71**: 273-294.

Mills, F. C. (1924) *Statistical Methods*, Henry Holt and Co., NY.

Morgan, M.S. (1990) *The History of Econometric Ideas*, Cambridge University Press, Cambridge.

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.

Neyman, Jerzy (1952) *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd ed., Washington: U.S. Department of Agriculture.

Neyman, J. 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.

Pagan, A.R. (1987) “Three econometric methodologies: a critical appraisal”, *Journal of Economic Surveys,* **1**, 3-24. Reprinted in C. W. J. Granger (1990).

Phillips, P.C.B. and Z. Xiao (1998), “A Primer on Unit Root Testing”, *Journal of Economic Surveys*, **12**: 423-470.

Qin, D. (1993) *The formation of econometrics: A historical perspective*, Clarendon Press, Oxford.

Reiss, J. (2008) *Error in economics: towards a more evidence–based methodology*, Routledge, NY.

Reiss, J. (2013) *Philosophy of Economics: A contemporary introduction*, Routledge, NY.

Reiss, J. (2015) *Causation, Evidence, and Inference*. Routledge, NY.

Rust, J. (2016) "Mostly useless econometrics? Assessing the causal effect of econometric theory", pp. 23-34 in *Causal Inferences in Capital Markets Research*, edited by Iván Marinovic, NOW, Hanover, MA.

Salmon, W. (1967) *The Foundations of Scientific Inference*, University of Pittsburgh Press, Pittsburgh, PA.

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

Spanos, A. (2006a) "Econometrics in Retrospect and Prospect," pp. 3-58 in *New Palgrave Handbook of Econometrics*, vol. 1, (ed.) Mills, T.C. and K. Patterson, MacMillan, London.

Spanos, A. (2006b) "Where Do Statistical Models Come From? Revisiting the Problem of Specification," pp. 98-119 in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo, Lecture Notes-Monograph Series, vol. 49, Institute of Mathematical Statistics.

Spanos, A. (2006c) "Revisiting the omitted variables argument: substantive vs. statistical adequacy," *Journal of Economic Methodology*, **13**: 179–218.

Spanos, A. (2007a) "Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach," *Philosophy of Science*, **74**: 1046–1066.

Spanos, A. (2007b) “The Instrumental Variables Method revisited: On the Nature and Choice of Optimal Instruments,” pp. 34-59 in *Refinement of Econometric Estimation and Test Procedures*, ed. by G. D. A. Phillips and E. Tzavalis, Cambridge University Press, Cambridge.

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

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

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

Spanos, A. (2010c) “Theory Testing in Economics and the Error Statistical Perspective”, pp. 202-246 in *Error and Inference*, edited by Mayo, D. G. and A. Spanos (2010).

Spanos, A. (2011) “Revisiting Unit Root Testing in the context of an AR(1) model with Variance Heterogeneity”, Virginia Tech working paper.

Spanos, A. (2012) “Philosophy of Econometrics”, pp. 329-393 in *Philosophy of Economics*, U. Maki (editor), in the series *Handbook of Philosophy of Science*, Elsevier (editors) D. Gabbay, P. Thagard, and J. Woods.

Spanos, A. (2013) “A Frequentist Interpretation of Probability for Model-Based Inductive Inference”, *Synthese*, **190**: 1555–1585.

Spanos, A. (2015) “Revisiting Haavelmo’s Structural Econometrics: Bridging the Gap between Theory and Data”, *Journal of Economic Methodology*, **22**: 171-196.

Spanos, A. (2017) “Why the Decision-Theoretic Perspective Misrepresents Frequentist Inference”, chapter 1, pp. 3-28, *Advances in Statistical Methodologies and Their Applications to Real Problems*, ISBN 978-953-51-4962-0.

Spanos, A. (2018) “Mis-Specification Testing in Retrospect”, *Journal of Economic Surveys*, **32**(2): 541–577.

Spanos, A. (2019) *Introduction to Probability Theory and Statistical Inference: Empirical Modeling with Observational Data*, 2nd edition, Cambridge University Press, Cambridge.

Spanos, A. (2020) “Yule–Simpson’s paradox: the probabilistic versus the empirical conundrum”, *Statistical Methods & Applications*, https://doi.org/10.1007/s10260-020-00536-4.

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

Spanos, A. and J.J. Reade (2015) “Heteroskedasticity/Autocorrelation Consistent Standard Errors and the Reliability of Inference”, VT working paper.

Vining, R. and Koopmans, T.C. (1949) “Methodological Issues in Quantitative Economics, " *Review of Economics and Statistics*, **31**: 77-94.

Von Mises, R. (1928/1957) *Probability, Statistics and Truth*, Dover, NY.

Wasserman, L. (2006) *All of Nonparametric Statistics*, Springer, NY.

Wooldridge, J.M. (2010) *Econometric analysis of cross section and panel data*, MIT press., 2nd ed., Cambridge, MA.

Yule, G.U. (1916) *An Introduction to the Theory of Statistics*, 3rd ed., Griffin, London.