Synthese
DOI 10.1007/s11229-015-0744-y

Error statistical modeling and inference: Where methodology meets ontology

Aris Spanos[1] · **Deborah G. Mayo**[2]

Received: 27 June 2014 / Accepted: 6 April 2015

Abstract In empirical modeling, an important desiderata for deeming theoretical entities and processes as real is that they can be reproducible in a statistical sense. Current day crises regarding replicability in science intertwines with the question of how statistical methods link data to statistical and substantive theories and models. Different answers to this question have important methodological consequences for inference, which are intertwined with a contrast between the ontological commitments of the two types of models. The key to untangling them is the realization that behind every substantive model there is a statistical model that pertains exclusively to the probabilistic assumptions imposed on the data. It is not that the methodology determines whether to be a realist about entities and processes in a substantive field. It is rather that the substantive and statistical models refer to different entities and processes, and therefore call for different criteria of adequacy.

Keywords Error statistics · Statistical vs. substantive models · Statistical ontology · Misspecification testing · Replicability of inference · Statistical adequacy

1 Introduction

Constructing and assessing scientific theories, the repositories of our scientific ontology, revolve around our methodology of data generation, modeling and inference—much of which is probabilistic and statistical in nature. An important desiderata for

Thanks are due to two anonymous reviewers for many contructive and valuable comments/suggestions.

✉ Aris Spanos
aris@vt.edu

[1] Department of Economics, Virginia Tech, Blacksburg, VA 24061, USA

[2] Department of Philosophy, Virginia Tech, Blacksburg, VA 24061, USA

deeming theoretical entities and processes ‘real’ is that they can be reproducible in a statistical sense. Current day crises regarding replicability in science intertwine with unclarity as to how statistical methods link data to statistical and substantive theories and models. If one is unclear about the linkages, how can we expect to critically evaluate inferences about warranted statistical processes and effects?

Statistical science involves methods for collecting, modeling, and drawing inferences from data in contexts where there is a threat of potential errors from different sources. The data are inexact and noisy, the phenomenon of interest is at best roughly captured in our models. In empirical modeling and inference, one cannot avoid all errors and unreliabilities but there are methods that have been reasonably successful in controlling and assessing these errors, as well as securing the statistical reliability of inferences. Error refers to any erroneous interpretation of the data, whether statistical or theoretical. Probability arises in these methods in order to assess the frequencies of erroneous inferences in terms of the relevant error probabilities associated with different inferential procedures. The methods that use probability to control and assess error probabilities we call error statistical; see Mayo (1996), Mayo and Spanos (2011). Examples would be confidence intervals, prediction intervals, significance tests, Neyman-Pearson (N-P) tests and model validation or Mis-Specification (M-S) tests.

What allows these methods to work in linking statistical and substantive models is often unclear, and that is what we will be addressing. One’s conception of the distinction between substantive and statistical models has important methodological consequences for inference, but these are also entwined with a contrast between the ontological commitments of the two types of models. It is not that the methodology impinges on whether to be a realist about entities and processes in a substantive field of inquiry—one is free to adopt any ontology on the substantive level—it is rather that there are different entities and processes being talked about in the two models, and different criteria for their adequacy. Only by getting clear on this can the different ways they are treated methodologically be understood, and that is the focus of this paper.

The traditional way of viewing empirical modeling is as a curve-fitting problem where a substantive model is foisted on a particular data set. Accordingly, a statistical probe of a substantive model is traditionally viewed as little more than adding an error term to a theory model, which is estimable with the data in question. In this conception, the statistical model and the substantive model are referring to the same entities and processes. We think this is a mistake. Instead, we view the substantive inquiry as consisting of posing questions in the context of a highly idealized statistical model. Such a model is statistically adequate when it accounts for the chance regularity patterns existing in the data and representing statistical systematic information. That is, the adequacy of the statistical model means that the data could have been generated by a stochastic process as described in the model. By contrast, adequacy of the substantive model would mean it adequately describes the portion of the world giving rise to the particular data. Inadequacy at the substantive level means that the theory model differs systematically from the actual data generating mechanism that gave rise to the phenomenon of interest; this can arise from false causal claims, missing variables, confounding factors, etc. Inadequacy at the statistical level means that one or more of the probabilistic assumptions imposed on the data are invalid.

We come to learn about the stochastic mechanism that gave rise to the data by something rather concrete: the error probabilities of our inferential methods. By regarding these as quantifying an inferential procedure's capacities for error detection, they are considered as real properties of the procedures used to probe substantive questions of interest in the context of the statistical model. As such, what statistical adequacy achieves is to ensure the error-reliability of inferences stemming from posing questions about those aspects of the world. Error-reliability refers to the nominal (assumed) error probabilities being close enough to the actual ones. These ontological/methodological connections lead to a number of requirements:

(1) It is necessary to identify an intermediate model that links the substantive questions to the data. That link comes in the form of a statistical model which pertains to the probabilistic assumptions imposed (often implicitly) on the data. Typical linkage methods involve estimation and testing procedures connecting parameters of the substantive model to those of the statistical model.

(2) It is vital to ensure the validity of the probabilistic assumptions underlying the statistical model on which the reliable assessment of the inferential tools' capacities depends. Inadequate statistical models will undermine the error-reliability of inference, which is vital to determine the capabilities of our statistical methods. We take seriously R.A. Fisher's (1935) conception that one has a real phenomenon when one knows how to conduct an experiment (or simulation) that will rarely fail to give a statistically significant result.

Brief outline: Section 2 brings out certain crucial weaknesses of the traditional approach to theory appraisal in empirical modeling in an attempt to show how the error statistical approach can address the problems raised by (1)–(2). Section 3 revisits the issues of reliability and replicability in empirical modeling. Section 4 discusses statistical model validation.

2 Theory appraisal in empirical modeling

In this section, we compare and contrast the traditional approach to theory testing with the error statistical approach. We argue that weaknesses in the traditional approach grow out of ontological commitments concerning the connection between a substantive model and the data.

2.1 The traditional approach to theory testing

In fields like economics, where the available data are usually observational, the traditional approach to empirical modeling attributes to theory a *pre-eminent* role and assigns to the data the *subordinate* role of 'quantifying the substantive model $\mathcal{M}_{\varphi}(\mathbf{z})$ presumed valid'. This perspective is usually implemented by adopting a methodology that views the quantification of $\mathcal{M}_{\varphi}(\mathbf{z})$ as a curve-fitting problem using data $\mathbf{Z}_0$ to estimate the unknown structural parameters φ. This traditional approach to econometric modeling introduces the probabilistic structure needed for the quantification of $\mathcal{M}_{\varphi}(\mathbf{z})$ via stochastic (usually white-noise) error terms attached to the theory model.

The implicit ontological stance is that the substantive model provides an adequate enough approximation to the 'true' data generating mechanism that gave rise to $\mathbf{Z}_0$ so as to render the error term statistically non-systematic (e.g. white-noise). That is, the substantive model is assumed to account for all the statistical systematic information in the data apart from some non-systematic noise; see Spanos (2010a). This might not be an unreasonable ontological stance in certain circumstances in disciplines like physics (Mayo 2010a) and astronomy (Spanos 2007), even when modeling with observational data. However, in fields like economics, where one is modeling economy-wide phenomena that involve the incessant interaction of millions of different economic agents, such a stance in conjunction with a methodology focusing exclusively on quantifying $\mathcal{M}_{\varphi}(\mathbf{z})$ using data $\mathbf{Z}_0$ often leads to untrustworthy inferences.

To illustrate how this traditional approach can easily give rise to spurious inference results, let us consider an example from financial economics.

The *Capital Asset Pricing Model* (CAPM) represents a highly idealized model of a rational investor's behavior that builds on *Markowitz's portfolio theory* and focuses on *risk premiums*, the difference between the expected return from a portfolio of risky assets and *a risk-free* rate of return, and the market excess return:

> The capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965) marks the birth of asset pricing theory (resulting in a Nobel Prize for Sharpe [and Markowitz] in 1990). Four decades later, the CAPM is still widely used in applications, such as estimating the cost of capital for firms and evaluating the performance of managed portfolios. ... The attraction of the CAPM is that it offers powerful and intuitively pleasing predictions about how to measure risk and the relation between expected return and risk. (Fama and French 2004, p. 25).

To simplify the discussion, we focus on the CAPM substantive (structural) model with one asset that is specified in terms of $y_t := \left(r_t - r_{ft}\right)$, which denotes the excess returns of a particular asset and $x_t := (r_{Mt} - r_{ft})$, the excess returns of the market [r_t—returns of a particular asset, r_{Mt}—market returns (e.g., S&P 500), r_{ft} – returns of risk free asset (e.g., 3-month treasury bill rate)]. For instance, when the returns on a particular asset is 3.5 % and that of the risk free asset is 1.5 %, the excess returns is 2 %. The substantive CAPM for a particular asset takes the simple form:

$$\boxed{\mathcal{M}_{\varphi}(\mathbf{z}):\ y_t = \beta x_t + \varepsilon_t,\ \ \varepsilon_t \backsim \mathsf{NIID}\,(0,\omega),\, t{=}1, ..., n,} \qquad (1)$$

where '$\mathsf{NIID}(0,\omega)$' stands for 'Normal, Independent and Identically Distributed with mean 0 and variance ω'.

What renders (1) a substantive model is the theoretical meaning bestowed on the substantive parameters $\varphi := (\beta,\ \omega)$ and the interpretation of the error term ε_t. Intuitively, β represents the sensitivity of the particular asset to the market excess return: $\beta > 1$ indicates individual asset returns better than the market returns, $\beta{=}1$ the same as the market, and $\beta < 1$ worse than the market. Hence, when the investor expects the market returns r_{Mt} to trend upward over the next several periods (bull market), investing in assets with $\beta > 1$ is the rational strategy, and the reverse when r_{Mt} is expected to trend downward (bear market). Similarly, the variance $\omega{=}\sigma^2{-}\beta^2\sigma_M^2$ decomposes into the *market systematic risk* $\sigma_M^2 = Var(r_{Mt})$ (non-diversifiable) and the *individual* asset risk σ^2 (diversifiable). The latter is diversifiable because it can be reduced by including additional assets in one's portfolio.

In the traditional literature the CAPM is tested by being embedded into a statistical model as follows:

> The Sharpe-Lintner CAPM says that the expected value of an asset's excess return (the asset's return minus the risk-free interest rate, $(r_t - \mu_{ft})$) is completely explained by its expected CAPM risk premium (its beta times the expected value of $(r_{Mt} - \mu_{ft})$). This implies that 'Jensen's alpha,' the intercept term in the time-series regression
>
> $$y_t = \alpha + \beta x_t + \varepsilon_t, \quad t{=}1, ..., n, \tag{2}$$
>
> is zero for each asset. (Fama and French 2004, p. 32)

That is, the structural model is parametrically embedded into a broader linear regression model (2) and its validity is assessed by testing the following restriction:

$$H_0 : \alpha = 0 \text{ versus } H_1 : \alpha \neq 0. \tag{3}$$

The unknown parameter α is referred to as "Jensen's (1968) alpha", with $\alpha{=}0$ interpreted (substantively) as indicating that the investment has earned a return adequate for the assumed risk; $\alpha < 0$ indicates an inadequate return for the risk, and $\alpha > 0$ as indicating an excess return for the risk.

Empirical illustration of the model (Lai and Xing 2008, pp. 72–81). The data are *monthly observations* for the period Aug. 2000 to Oct. 2005 ($n{=}64$), where $y_t{:=}\left(r_{kt} - \mu_{ft}\right)$ is excess (log) returns of Exxon-Mobil Corp., $x_t{:=}\left(r_{Mt} - \mu_{ft}\right)$ is the market excess (log) returns, where r_{Mt} is the returns based on the S&P 500 index; the risk free returns (μ_{ft}) is based on the 3-month Treasury bill rate. Estimation of the statistical model (2) yields:

$$y_t {=} \underset{(.004)}{.003} + \underset{(.101)}{.693} x_t + \underset{(.021)}{\widehat{u}_t}, \quad R^2{=}.433, \; n{=}64, \tag{4}$$

where the standard errors are given in parentheses below the point estimates. The authors proceed to test (3) using a t-test yielding: $\tau_\alpha(\mathbf{z}_0){=}\frac{.003}{.004}{=}.84[.402]$—the p-value, in square brackets, is not small enough to reject H_0. They conclude that failing to reject H_0, in conjunction with the significance of β $[\tau_\beta(\mathbf{z}_0){=}\frac{.693}{.101}{=}6.86[.000]]$ and the goodness-of-fit $R^2{=}.433$, provide *evidence for* the CAPM. That is, from the curve-fitting perspective the good fit and the statistically significant coefficients 'appear' to indicate that the data provide good evidence validating the substantive model.

This is an example of how a deficient methodology stemming from the presupposition that the structural model constitutes a good approximation of the actual mechanism that generated the data, can mislead modelers as to what they think they are finding out about the world. The traditional perspective on the pre-eminence of theory takes for granted that the substantive model is the only mediator between the data and the

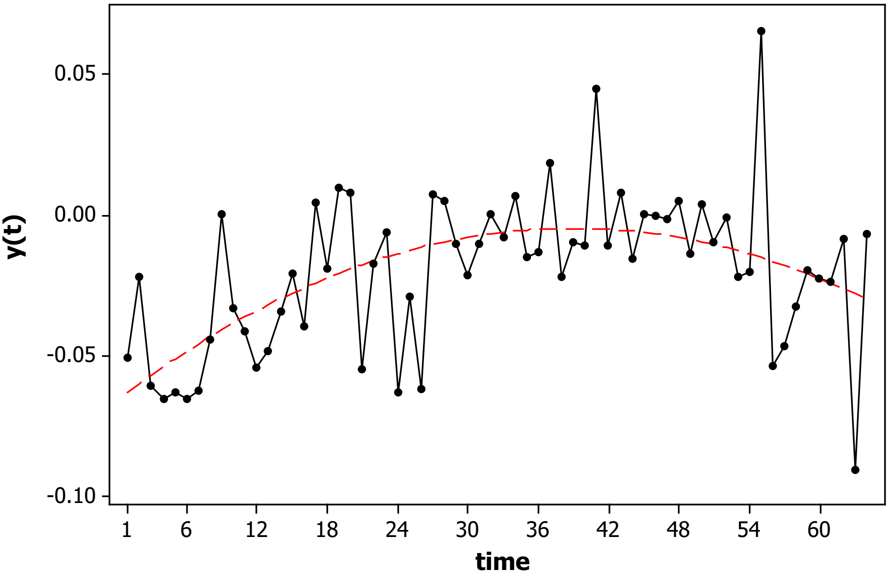

Fig. 1 Exxon-Mobil excess returns

phenomenon of interest. The traditional perspective downplays or ignores the fact that behind every substantive model $\mathcal{M}_{\varphi}(\mathbf{z})$, there is a statistical model $M_{\theta}(\mathbf{z})$ that pertains exclusively to the probabilistic assumptions (statistical inductive premises) imposed on the data (often implicitly) via the error term ε_t. The validity of these statistical premises vis-a-vis the data is what underwrites the reliability of all inferences relating to $\mathcal{M}_{\varphi}(\mathbf{z})$.

As shown below, the authors' findings are seriously called into question because (4) is *statistically misspecified*. That is, some of the probabilistic assumptions invoked by the above quoted t-statistics and the R^2 are invalid and thus the resulting inferences are statistically spurious: the *nominal* error probabilities are very different from the *actual* ones. Large discrepancies between the actual and nominal error probabilities can easily arise with what some modelers might consider as 'minor' misspecifications.

For example, the t-plots of the data (Figs. 1, 2) used in estimating (4) indicate that the sample mean (average) of the data over time (t=1, 2, ..., n) changes in a systematic way exhibiting mean-heterogeneity (trending) which could be approximated by a quadratic polynomial in t, say $\mu(t)=\delta_0+\delta_1 t+\delta_2 t^2$, indicated by a dashed line. As shown in Sect. 4.1, ignoring the presence of an even lower degree trend ($\mu(t)=\delta_0 + \delta_1 t$) in one's data (see Fig. 3) constitutes a serious statistical misspecification. A trend that induces a huge discrepancy between the nominal type I error for the primary hypothesis (3) of .05, and the actual one of .973 (for n=100). Applying a .05 significance level test when the actual type I error is greater than .97 will lead an inference astray; see Spanos and McGuirk (2001).

The 'apparent' evidence in favor of the CAPM is built on statistical artifacts (invalid probabilistic assumptions) that deceive one into thinking that (4) accounts for the phenomenon of interest. Such erroneous inferences can be traced to a certain ontological

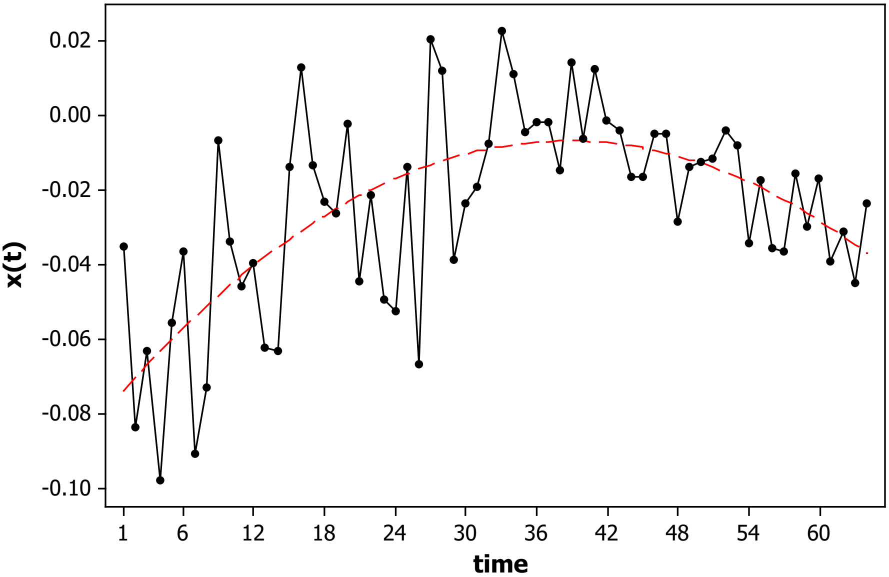

Fig. 2 Market excess returns

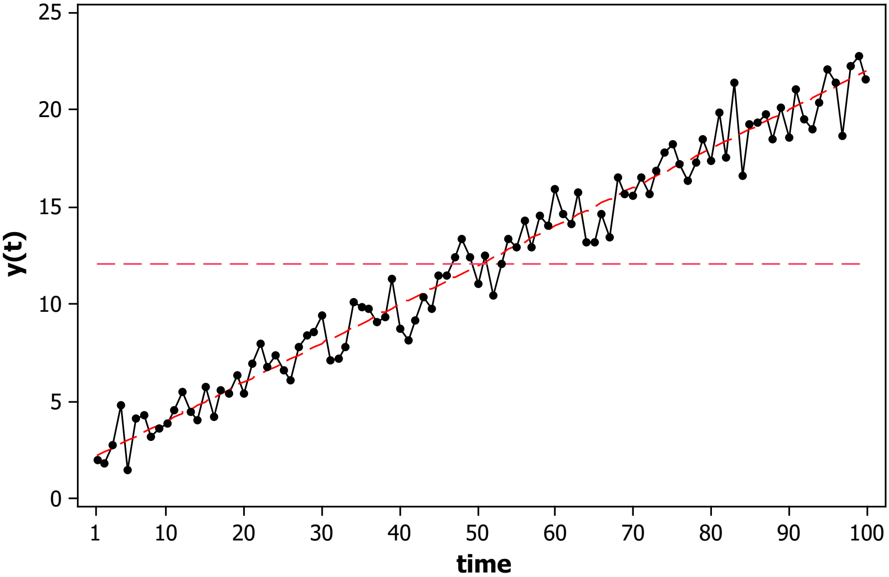

Fig. 3 Typical realization of a NI($2+.2t$, 1) process

stance underlying curve-fitting: the modeler assumes that the structural model constitutes a good approximation of the actual mechanism that has generated the data. In principle, such a stance can be justified by providing pertinent evidence on how good the approximation is—in a statistical sense—using the data. This will require one to demonstrate that (4) does account for all the statistical systematic information in the data apart from some non-systematic noise, i.e. it is statistically adequate. The problem arises because the strategy of foisting the structural model on the data, and appraising the fitted model (4) using goodness-of-fit measures, leaves little room for securing statistical adequacy. As a result, any departures from the probabilistic assumptions (indirectly) imposed on the data by the estimation and testing procedures will undermine the reliability of any inference based on (4), including goodness-of-fit/prediction measures; see Spanos (2010a).

The next sub-section proposes a way to circumvent this unreliability of inference by separating the substantive from the statistical premises of inference, *ab initio*, and securing the adequacy of the latter before probing the adequacy of the former.

2.2 Substantive versus statistical models

What are the **potential errors** that could invalidate the authors' claim that the data provide evidence *for* the CAPM? Intuitively, the move from the substantive (1) to the statistical model (2) by adding a constant α seems *ad hoc* and unjustified. More formally, bridging the gap between $\mathcal{M}_{\varphi}(\mathbf{z})$ and $\mathbf{Z}_0$ requires proper justification to ensure the estimated model *does* account for the chance regularities in data $\mathbf{Z}_0$.

This can be achieved using the notion of a statistical model $M_{\theta}(\mathbf{z})$ as a mediator between $\mathcal{M}_{\varphi}(\mathbf{z})$ and $\mathbf{Z}_0$. This mediating enables the modeler to distinguish between two primary sources of errors that can render inferences unreliable:

[a] *Statistical inadequacy*: one or more of the probabilistic assumptions (implicitly) imposed on the data $\mathbf{Z}_0$ are invalid.

[b] *Substantive inadequacy*: the circumstances envisaged by the theory in question differ 'systematically' from the *actual* data generating mechanism that gave rise to the phenomenon of interest. Substantive inadequacy can arise from flawed *ceteris paribus* clauses, missing confounding factors, false causal claims, etc.

The traditional perspective downplays the fact that the justification of the inference rests on the validity of the statistical premises represented by the probabilistic assumptions imposed (indirectly) on the data via the error assumptions in (1). In practice, the statistical premises need to be brought out explicitly and specified in terms of the observable process $\{\mathbf{Z}_t := (Y_t, X_t),\ t\in\mathbb{N}\}$ underlying data $\mathbf{Z}_0 := (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$. This is achieved by viewing the data as a realization of the process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ and the statistical model $M_{\theta}(\mathbf{z})$ as a particular parameterization of $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$. This affords $M_{\theta}(\mathbf{z})$ 'a life of its own' stemming from the probabilistic structure that represents adequately the chance regularities in data $\mathbf{Z}_0$, not from the substantive model; Spanos (2006).

A complete set of testable assumptions for the linear regression model in terms of the observable process $\{(Y_t|X_t{=}x_t),\ t\in\mathbb{N}\}$, is given in Table 1. These assumptions are related to those of the error terms in (1) but the two do not coincide; Mayo and Spanos (2004). The statistical Generating Mechanism (GM) and assumptions [1]-[5] define the statistical model underlying (2) in purely probabilistic terms without invoking any substantive information relating to the unknown parameters $\theta := \left(\beta_0, \beta_1, \sigma^2\right)$. As shown in the last line of Table 1, the statistical parameters θ are defined in terms of the moments (means, variances and covariances) of the process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$.

From the error-statistical perspective, the statistical model $M_{\theta}(\mathbf{z})$ in Table 1 is viewed as a generating mechanism that could have given rise to data $\mathbf{Z}_0$ only when it

Table 1 Normal, Linear Regression Model

$Statistical\ GM:$	$Y_t=\beta_0+\beta_1 x_t+u_t,\quad t\in\mathbb{N}.$	
[1] Normality:	$(Y_t\vert X_t{=}x_t)\backsim \mathsf{N}(.,.),$	$t\in\mathbb{N}.$
[2] Linearity:	$E\,(Y_t\vert X_t{=}x_t)=\beta_0+\beta_1 x_t,$	
[3] Homoskedasticity:	$Var\,(Y_t\vert X_t{=}x_t)=\sigma^2,$	
[4] Independence:	$\{(Y_t\vert X_t{=}x_t),\ t\in\mathbb{N}\}$ independent process ,	
[5] t-invariance:	$\theta:=\left(\beta_0,\beta_1,\sigma^2\right)$ are *not* changing with t,	
$\beta_0=E(Y_t){-}\beta_1 E(X_t),\ \beta_1=\frac{Cov(Y_t,X_t)}{Var(X_t)},\ \sigma^2=Var(Y_t)-\frac{[Cov(Y_t,X_t)]^2}{Var(X_t)}$		

accounts for all the chance regularities in data $\mathbf{Z}_0$, i.e. $M_\theta(\mathbf{z})$ is *statistically adequate*: assumptions [1]–[5] are valid for $\mathbf{Z}_0$. The statistical model $M_\theta(\mathbf{z})$ is no longer viewed as derived from the substantive model, but as a related but separate entity specified with a view to: (i) account for all the chance regularity patterns contained in data $\mathbf{Z}_0$; and (ii) nest parametrically the substantive model $\mathcal{M}_\varphi(\mathbf{z})$.

It is important to emphasize the distinction between the chance regularity patterns exhibited by data $\mathbf{Z}_0$, and the mathematical concepts from probability theory needed to frame and model (account for) these regularities. The statistical model $M_\theta(\mathbf{z})$ aims to provide a mathematical formulation that adequately summarizes (models, captures) all of these regularities in terms of the *probabilistic structure* (assumptions) pertaining to the underlying process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$. That is, the data $\mathbf{Z}_0$ are viewed as a typical realization of $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ and the 'typicality' is appraised by testing the assumptions defining $M_\theta(\mathbf{z})$ vis-a-vis data $\mathbf{Z}_0$ using M-S testing. In turn, an adequate $M_\theta(\mathbf{z})$ enables one to pose the substantive questions of interest to data $\mathbf{Z}_0$, including the appraising of the substantive adequacy of $\mathcal{M}_\varphi(\mathbf{z})$.

Statistical adequacy plays a crucial role in all aspects of inference because:

(i) It secures the **error-reliability** of any inference procedures by ensuring that the relevant:

$$\boxed{\textit{actual}\text{ error probabilities }\simeq\textit{ nominal}\text{ (assumed) error probabilities}}$$

Unreliability of inference manifests itself in terms of significant discrepancies between the actual and the nominal error probabilities; see Sects. 3.2, 4.1.

(ii) It provides a **sound link** between the substantive model $\mathcal{M}_\varphi(\mathbf{z})$ and the phenomenon of interest by adequately accounting for the chance regularities in $\mathbf{Z}_0$.

(iii) It offers a **reliable basis** for testing the empirical validity of substantive models like the CAPM in (1), as well as probing their substantive adequacy.

(iv) It ensures that the statistical model in Table 1 can be used to **replicate** data $\mathbf{Z}_0$ by generating simulated data that exhibit (approximately) the same chance regularities as data $\mathbf{Z}_0$. This stems from viewing $M_\theta(\mathbf{z})$ as a parameterization of the stochastic process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ whose probabilistic structure is chosen so that $\mathbf{Z}_0$ constitutes a typical realization thereof. The *chance regularity patterns* that exist in $\mathbf{Z}_0$ render such data not only amenable to statistical modeling and inference, but when captured adequately by $M_\theta(\mathbf{z})$, the latter can be used to replicate them.

Table 2 M-S testing results for (4)

[1] Normality:	$\chi^2(2) = 12.153[.002]$
[2] Linearity:	$F(1, 61) = 1.98[.165]$
[3] Homoskedasticity:	$F(1, 60) = 1.15[.288]$
[4] Independence:	$F(2, 60) = 3.61[.048]$
[5] t-homogeneity:	$F(2, 60) = 14.2[.000]$

The ability to use $M_\theta(\mathbf{z})$ to generate at will numerous sample realizations allows one to construct the empirical counterpart to the relevant sampling distributions and the associated error probabilities using computer simulation, as demonstrated in Sect. 4.1. This gives operational meaning to Fisher's view of replicability:

> In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (Fisher 1935, p. 14)

An important insight that can be traced back to Fisher (1922) is that one can learn from data about the world by *embedding* the material experiment into a statistical model, in a way that renders measurable the capacity of the relevant inferential procedures using error probabilities. In a certain sense, the error probabilities play an analogous role to the physical controls associated with calibrating the capacity of instruments to distinguish between signal and noise in physical experiments. Error probabilities quantify the capacity of the inferential procedures to distinguish between systematic and non-systematic effects. Moreover, the *experimental knowledge* gained from well-designed physical experiments corresponds to the *empirical regularities* accounted for by a statistically adequate model. The ability to simulate a statistical model whose adequacy has been secured enables one to demonstrate the reality of an effect using the relevant empirical error probabilities associated with the particular inference procedure.

When any of the assumptions [1]–[5] are invalid for data $\mathbf{Z}_0$, none of the results in (i)–(iv) are assured. In particular, what renders resampling methods, including the simple bootstrap, effective tools for learning about the underlying generating mechanism, is the validity of the statistical assumptions.

Example. Using the Mis-Specification (M-S) tests described in the Appendix, it is shown in Table 2 that the estimated model (4) is statistically misspecified; the results indicate departures from assumptions [1], [4] and [5]—p-values are given in square brackets. That is, the estimated regression in (4) suffers from several misspecifications, in addition to t-heterogeneity indicated by Figs. 1 and 2.

In light of the serious consequences of statistical misspecification, 'why does the traditional approach neglect M-S testing?' There are several reasons, the most important ones stem from the perspective of the pre-eminence of theory.

[i] Rather than view the probabilistic assumptions as separate from the validity of the substantive model, the substantive error ε_t is viewed as pertaining to the structural model $\mathcal{M}_\varphi(\mathbf{z})$ vis-a-vis the phenomenon of interest. These could include errors of measurement, errors of approximation, omitted effects, external shocks etc. That perspective precludes distinguishing between the statistical and the substantive premises of inference.

[ii] Due to the fact that probabilistic assumptions are made in terms of the error ε_t term, it is often unclear how these assumptions pertain to $\mathbf{Z}_0$ and how one can validate them.

In summary, the key problem with the traditional approach to econometric modeling is that foisting the substantive model on the data can result in an estimated model like (4) which is both statistically and substantively misspecified, but one has no principled way to distinguish between these two sources of misspecification. The key to circumventing this *Duhemian ambiguity* (Mayo 1997) is to untangle the statistical from the substantive premises.

2.3 An ontology of a statistical model?

In general, behind every structural model, generically specified by:

$$\mathcal{M}_\varphi(\mathbf{z})=\{f_S(\mathbf{z};\varphi),\ \varphi\in\Phi\subset\mathbb{R}^p\},\ \mathbf{z}\in\mathbb{R}^n_Z,\ p<n, \tag{5}$$

there exists (often implicit) a statistical model, taking the generic form:

$$M_\theta(\mathbf{z})=\{f(\mathbf{z};\theta),\ \theta\in\Theta\subset\mathbb{R}^m\},\ \mathbf{z}\in\mathbb{R}^n_Z,\ m\geq p, \tag{6}$$

that can be viewed as a parameterization of the observable stochastic process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{Z}_0$ and summarized by the joint distribution of the sample $\mathbf{Z}:=(\mathbf{Z}_1,...,\mathbf{Z}_n),\ f(\mathbf{z};\theta),\ \mathbf{z}\in\mathbb{R}^n_Z$. That means that any form of statistical misspecification will distort these sampling distributions in different ways, and thus induce discrepancies between the actual and nominal probabilities associated with different inferential procedures.

The connection between the two models is crucial because:

(a) the statistical model $M_\theta(\mathbf{z})$ can be viewed as a *parameterization* of the observable stochastic process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{Z}_0$,

(b) the structural model $\mathcal{M}_\varphi(\mathbf{z})$ can be viewed as a *reparameterization/restriction* of $\mathcal{M}_\theta(\mathbf{z})$ via: $\mathbf{G}(\theta,\varphi)=\mathbf{0},\ \theta\in\Theta,\ \varphi\in\Phi,$ where θ and φ denote the statistical and substantive parameters of interest, respectively. That is, φ is determined by imposing restrictions on θ. In the above example, these restrictions come down to α=0, but in other cases they can be very complicated; Spanos (1990).

(c) The adequacy of $M_\theta(\mathbf{z})$ underwrites the error reliability of inferences based on $\mathcal{M}_\varphi(\mathbf{z})$. The statistical adequacy of $M_\theta(\mathbf{z})$ is what enables the actual error probabilities to approximate closely the nominal (assumed) ones, and thereby reflect accurately how well-tested the primary hypotheses are. Hence, the substantive model $\mathcal{M}_{\varphi}(\mathbf{z})$ is *empirically valid* when: (i) the implicit statistical model $M_{\theta}(\mathbf{z})$ is adequate, and (ii) the restrictions $\mathbf{G}(\varphi,\theta)=\mathbf{0}$ are *data-acceptable*.

Untangling $M_{\theta}(\mathbf{z})$ from $\mathcal{M}_{\varphi}(\mathbf{z})$ delineates two very different questions:
[a] statistical adequacy: does $M_{\theta}(\mathbf{z})$ adequately account for the chance regularity patterns in $\mathbf{Z}_0$?
[b] substantive adequacy: does the model $\mathcal{M}_{\varphi}(\mathbf{z})$ shed adequate light on (describe, explain, predict) the phenomenon of interest?

In light of **[a]**, the ontological commitments in specifying $M_{\theta}(\mathbf{z})$ concern: [i] the existence of perceptible chance regularity patterns in data $\mathbf{Z}_0$, and [ii] the availability of a rich enough probability theory to represent these regularities mathematically in terms of concepts from three broad categories (Spanos 1999): *Distribution* (e.g. Normal, Poisson), *Dependence* (e.g. Independence, Markov), *Heterogeneity* (e.g. ID, stationarity). That is, what renders data $\mathbf{Z}_0$ amenable to statistical modeling and inference is *not* the existence of a certain *population* (hypothetical or otherwise) from which different sample realizations can be chosen repeatedly, but the existence of the chance regularity patterns in the particular data. The 'population' metaphor is inappropriate in general because it is overly influenced by the IID assumptions to be pertinent for non-IID data. In turn, statistical adequacy ensures the existence of a true value $\theta^{*}\in\Theta$ such that $M_{\theta^{*}}(\mathbf{z})=\{f(\mathbf{z};\theta^{*})\},\ \mathbf{z}\in\mathbb{R}^{n}_{Z}$, could have generated data $\mathbf{Z}_0$. Viewing $M_{\theta}(\mathbf{z})$ as such a stochastic generating mechanism renders repeatability possible both in principle and in practice. Any simulated data generated by a statistically adequate $M_{\theta}(\mathbf{z})$ will exhibit the same chance regularity patterns as $\mathbf{Z}_0$.

In light of **[b]**, the ontological commitments associated with the substantive model $\mathcal{M}_{\varphi}(\mathbf{z})$ are very different from those associated with $M_{\theta}(\mathbf{z})$, irrespective of whether one adopts a realist, an instrumentalist or a constructive empiricist view of theories.

That statistically adequate models suffice to critically evaluate the substantive weaknesses in models and theories is the key to delimiting what may be extracted from data at a given time, and how to develop better tests, models and instruments. In this sense, the error-statistical perspective on statistical models as mediators between $\mathcal{M}_{\varphi}(\mathbf{z})$ and data $\mathbf{Z}_0$ does not entail any specific ontological stance on the nature of theories and theory models. It aims to provide a broad enough framework to enable one to learn from data about phenomena of interest by bringing out the potential errors that can undermine this goal.

In relation to the nature of theory models, the above conception of a statistical model is not restricted to the less well developed theories of economics as opposed to say physics. This error statistical perspective was used to understand why Kepler's 1609 first law—the motion of the planets around the sun is elliptical—was originally just an empirical regularity because it was actually based on a statistically adequate model when retrospectively estimated using the original Brahe data; Spanos (2007). It took almost 60 years before Newton could provide a substantive interpretation to that empirical regularity and show that the statistical model underlying Kepler's law was describing a real world mechanism. Moreover, developing parametric rivals within statistical models is what enabled substantive theories to be developed in experimental general relativity, long before their theoretical interpretation was available; Mayo (2010b).

This error statistical conception of a statistical model is appropriate whether the data are observational or experimental. Its appropriateness in the case of experimental data—resulting from controlled chemical processes—in the discovery of 'argon' in the 1890's is illustrated in Spanos (2010d). In that case there was no actual generating mechanism that one can model using the particular data, but the data do exhibit chance regularities, rendering them amenable to statistical modeling and inference. In the case of experimental data it is sometimes thought that one need not worry about the probabilistic assumptions because of the precautions taken or the experimental designs applied in generating the data. The truth is that one needs to ensure that the experimental controls and designs applied in generating the data actually had the intended effect. That can only be established by testing the validity of the statistical premises vis-a-vis the data.

Finally, the proposed statistical vs. substantive model distinction can be used to shed light on the slogan widely used in statistics and elsewhere that "all models are wrong but some are useful!" attributed to Box (1976). This catchphrase, however, is often used in an unintended manner; one which confuses the statistical and substantive adequacy questions **[a]–[b]** above. It is one thing to claim that the structural model $\mathcal{M}_{\varphi}(\mathbf{z})$ is wrong in the sense that it is not an exact picture of reality in a substantive sense, and quite another to claim that the implicit statistical model $M_{\theta}(\mathbf{z})$ could *not* have generated data $\mathbf{Z}_0$, because its probabilistic assumptions are invalid. Indeed, the usefulness of models for inference purposes, one can argue, stems from their statistical adequacy, irrespective of any substantive inadequacies.

3 Reliability and replicability in empirical modeling

In this section, we revisit two of the key notions raised above that pertain to the reliability of inference and the replicability of inference results.

Current discussions of the replication crisis seem to assume that if an effect is established by significance testing is genuine, then it will be replicated by most, if not all, studies of the same phenomenon. The scarcity of published studies that replicate previous empirical results is generally thought to be primarily due to biases and improper interpretations of inference results of authors and journal editors. While this gets to one part of the problem, such assessments misdiagnose the real sources of the observed non-replicability of certain inference results. When Baggerly and Coombes (2009) found themselves unable to replicate the results of Potti et al. (2006), the problem had largely to do with the inadequacy of the models being used to predict whether cancer would be sensitive to a given chemotherapy regimen. Even though Potti's results were statistically spurious, he claimed to be able to replicate them. Sure enough, statistically spurious results can be easily replicable when the replicators use defective procedures that commit similar errors.

For example, evidence in favor of the CAPM is found by dozens of MBA students every week by following the traditional approach to theory testing which invariably ignores the fact that most of the probabilistic assumptions underwriting their inference results are invalid for their data. That only confirms dubious (unreal) effects discovered by a shared defective methodology. What is important about replicability is that there is an infinity of spurious results one can 'discover' when using different defective methodologies. There are numerous ways a statistical model can be misspecified, with each form of misspecification giving rise to different spurious results. What secures the uniqueness of inference results is statistical adequacy. This means that the question of replicability is inextricably bound up with the reliability of inference and cannot be properly discussed separately.

3.1 Revisiting reliability: the error-statistical perspective

In error statistical inference there is a well-defined notion of *reliability* pertaining to whether the actual error probabilities associated with an inference procedure approximate closely the nominal (assumed) ones. This error reliability can be undermined in a number of different ways.

The most well-known source of error unreliability in the literature is associated with techniques of significance seeking, cherry picking, multiple testing and the like, which increase the chance of erroneously outputting significance, thereby violating the control supposedly afforded by size and power. For example, if a published result of a clinical trial alleges statistical significance (benefit from a drug), at a small significance level .01, but ignores 19 other non-significant trials, it makes it easy to find a positive result on one factor or other; see Cox and Mayo (2010).

The discrepancy between actual and nominal error probabilities stems from evaluating the nominal error probabilities using the wrong sampling distribution. The wrong sampling distribution is one that does not reflect the particular procedure's actual reliability in answering the question of interest. It is crucial to take account of the ways such selection effects alter the capabilities of tests and lead to erroneous error probabilities (Mayo 1996; Mayo and Cox 2010). However adjustments for multiple testing and cherry picking are dependent on first having a statistically adequate model. Finding model violations can cut short the process of checking for bias due to selection effects. The error statistical notion of reliability affords a standpoint to block what is well-known: 'If you torture the data long enough, they will confess to anything'. If we insist on ensuring that the actual error probabilities approximate closely the nominal ones, the data would not confess to claims that are not warranted by reliable evidence.

3.2 Revisiting replicability: the error-statistical perspective

What does replicability mean in the context of frequentist testing? To begin with, it does not mean that every data set $\mathbf{Z}_0$ will give rise to identical inferential results. Indeed, such a requirement will be unattainable even in the case where the different data sets $\mathbf{z}^{(k)} := (z_1, ..., z_n)^\top$, $k = 1, 2, ..., N$ have the same probabilistic structure and sample size n. To explain why, let us consider a simple example where the relevant sampling distributions can be simulated on a computer.

Table 3 Adequate Linear Regression model

Replications N=10, 000	True/Estim: $Y_t = 1.5 + 0.5x_t + u_t$			
	n=50		n=100	
True values	Estimates mean	Estimates Std	Estimates mean	Estimates Std
β_0=1.5	1.502	.122	1.500	.087
β_1=.5	0.499	.015	0.500	.008
σ^2=.75	0.751	.021	0.750	.010
$\mathcal{R}^2$=.25	0.253	.090	0.251	.065
t-statistic	Type I error probability		Type I error probability	
H_0: β_0=0	Nominal	Actual	Nominal	Actual
τ_{β_0}	.05	.049	.05	.05
t-statistic	Type I error probability		Type I error probability	
H_0: β_1=0	Nominal	Actual	Nominal	Actual
τ_{β_1}	.05	.047	.05	.049

Example. Consider the Linear Regression model (Table 1) with statistical GM:

$$Y_t = \beta_0 + \beta_1 x_t + \sigma\epsilon_t,\ \ \epsilon_t \backsim \mathsf{N}(0,1),\ \ t{=}1, 2, ..., n, \tag{7}$$

where $\epsilon_k \backsim \mathsf{N}(0,1)$ denotes *pseudo-random* numbers from $\mathsf{N}(0,1)$.

Empirical sampling distributions. The simulation takes the form of generating N realizations of sample size n=50 and n=100, using (7) and for each realization one estimates the unknown parameters $\left(\beta_0, \beta_1, \sigma^2, \mathcal{R}^2\right)$, as well as the t-tests for the hypotheses: H_0: $\beta_i{=}\beta_i^*$ vs. H_1: $\beta_i{\neq}\beta_i^*$, i=0, 1, where β_i^*, i=0, 1 denote the true values. For a large enough N, say N=10, 000, one can construct the empirical counterparts to the theoretical sampling distributions (Cox and Hinkley 1974):

$$\widehat{\beta}_0 \backsim \mathsf{N}(\beta_0, Var(\widehat{\beta}_0)),\ \ \widehat{\beta}_1 \backsim \mathsf{N}(\beta_1, Var(\widehat{\beta}_1)),\ \ \tfrac{(n-2)s^2}{\sigma^2} \backsim \chi^2(n-2).$$

Table 3 summarizes these empirical sampling distributions using descriptive statistics (sample mean and variance) to bring out the reliability and precision of the inference in question. The true values of the parameters are reported in column 1 (Table 3), and the sample mean (average) and standard error (std) of the sampling distributions of the point estimates in columns 2-3, for n=50. Columns 4-5 report the same statistics for a larger sample size n=100 to see how the precision and reliability of inference changes with more data information. The sample mean of the $N = 10, 000$

point estimates for n=50 of the point estimates, (1.502, .499, .751, .253), are extremely close to the true parameters, $\left(\beta_0=1.5, \beta_1=.5, \sigma^2=.75, \mathcal{R}^2=.25\right)$. Moreover, as expected from sampling theory, the accuracy of the sample mean of the point estimates becomes even more accurate for n=100.

The t-tests in Table 3 report the nominal (assumed) type I error probability (α=.05), as well as the actual one based on the relative frequency of rejections in N=10, 000 sample realizations. For n=50 the actual type I error probabilities associated with τ_{β_0} and τ_{β_1} are .049 and .047, respectively, which are very close to the nominal α=.05. For n=100 the actual and nominal error probabilities are even closer, as expected. That is, the simulation results in Table 3 show us empirically what error statistical sampling theory is intended to achieve. It is worth noting that one can easily extend the above simulations to evaluate the empirical power of the t-tests by selecting different discrepancies from (β_0, β_1) .

Replicability. A key feature of replicability stemming from the simulation results in Table 3 is that for any particular sample realization $\mathbf{z}^{(k)}$, the point estimates and the observed test statistics take different values over a certain range. This is summarized by the *sample mean* and the associated standard error (Std) of each of the N different realizations. Hence, even in the best case scenario where for different studies:

(i) the sample size n is the same,
(ii) the data constitute realizations of the same generating mechanism,
(iii) the estimated model in each study is statistically adequate,

inference results for each study, such as point and interval estimates, accept/reject testing results, including p-values, will *not* coincide. In particular, since the p-value $p(\mathbf{z}_0)$, when viewed as a statistic $p(\mathbf{Z})$ for different realizations of $\mathbf{Z}$, is Uniformly distributed over the range [0,1] (Cox and Hinkley 1974), it follows that for a significant result, $p(\mathbf{z}_0) \leq \alpha$, the probability of getting a p-value equal or less than $p(\mathbf{z}_0)$ in many replications is equal to $p(\mathbf{z}_0)$, i.e. $\mathbb{P}(p(\mathbf{Z}) \leq p(\mathbf{z}_0); H_0) \leq p(\mathbf{z}_0)$.

It is expected that the p-values will be very different for different sample sizes n because they are vulnerable to the large n problem; see Senn (2001).

What is replicable then? The error statistical perspective sheds light on the replicability issue using the post-data severity evaluation of individual inference results. Given that published studies rarely use the same or similar data, or even the same sample size n data, what is replicable in practice is not the inference result as such, but the *discrepancy* from the null that is warranted with high severity. For instance, despite the significant differences in the inferential results of numerous studies modeling the ratio of male to female newborns for different localities, different time periods and different sample sizes—the reported p-values vary considerably from .000001 to .4—the warranted discrepancies from H_0: θ=.5 are very similar and take value between .01 and .018. More importantly, the warranted discrepancies approximate closely the substantive discrepancy grounded in human biology, even in cases where H_0 is accepted ($p(\mathbf{z}_0)$=.394); see Spanos (2010e). This is because, in contrast to the p-value, the discrepancy outputted by the severity evaluation takes into account the sample size n as it affects the power of the test; Mayo and Spanos (2006).

4 Statistical model validation

All approaches to statistical inference, including the frequentist, the Bayesian and the nonparametric, rest on a notion of a statistical model. A statistically misspecified $\mathcal{M}_\theta(\mathbf{z})$, by definition assumes an erroneous distribution of the sample $f(\mathbf{z};\theta)$, $\mathbf{z}\in\mathbb{R}_Z^n$, giving rise to a wrong likelihood via: $L(\theta;\mathbf{z}_0)\propto f(\mathbf{z}_0;\theta)$, $\theta\in\Theta$. That in turn will give rise to an erroneous posterior. What differentiates alternative approaches to inference is how well they step up to the plate to ensure the reliability of inference by checking the adequacy of the assumed statistical model, or by securing learning from data despite certain violations of statistical assumptions (robustness).

4.1 The error-statistical perspective on model validation

An error statistician pursues statistical model validation by thoroughly testing the model assumptions, such as [1]–[5] in Table 1, using several different types of tools: informal graphical analyses of data, non-parametric and parametric M-S tests, and simulation-based methods, including resampling. To illustrate the serious effects of statistical misspecification on the reliability of inference, consider a variant of the simulation experiment reported in Table 3, where assumption [5] is rendered invalid: the data exhibit a linear trend in the mean as in Fig. 3.

The statistical misspecification due to the presence of mean-heterogeneity in the data—by omitting the trend (t)—is shown to devastate both the accuracy of the point estimates as well as the error reliability of the t-tests. Table 4 shows that for $n{=}50$ there are huge differences between the true values of the parameters $\beta_0{=}1.5$, $\beta_1{=}.50$, $\sigma^2{=}.75$, $\mathcal{R}^2{=}.25$, and the sample mean of the point estimates $\overline{\beta}_0{=}.48$, $\overline{\beta}_1{=}1.97$, $\overline{\sigma}^2{=}2.96$, $\overline{R}^2{=}.983$. Moreover, the accuracy of the sample mean of the point estimates worsens as n increases to 100 in direct contrast to the results in Table 3. Similarly, the t-tests in Table 4 show that for $n{=}50$, there are huge discrepancies between the nominal type I error probabilities (.05, .05) and the actual ones (.782, 1.0) associated with the t-tests based on $\left(\tau_{\beta_0},\tau_{\beta_1}\right)$. That is, one will reject 100% of the time the true null hypothesis H_0: $\beta_1{=}\beta_1^*$. For $n{=}100$, the actual type I probability for τ_{β_0} increases to .973.

The intuitive explanation of what goes wrong in Table 4 is that one ignores the fact that the data exhibit a mean trend as in Fig. 3 (dashed line), but instead assumes a constant mean (solid line), i.e. it is assumed that the data look like the data shown in Fig. 4. This invalid assumption gives rise to *inconsistent estimators* of the means of (x_t,y_t), in the form of $\left(\overline{x}{=}\frac{1}{n}\sum_{i=1}^n x_i,\ \overline{y}{=}\frac{1}{n}\sum_{i=1}^n y_i\right)$, when the actual means are $(E(X_t){=}1{+}.1t,\ E(Y_t){=}2{+}.2t)$. This move, in turn, yields inconsistent (*spurious*) estimators of the variances, covariances and correlation coefficients defining the unknown parameters $\theta{:=}\left(\beta_0,\beta_1,\sigma^2\right)$ (Table 1).

4.2 An ontology of Mis-Specification (M-S) testing?

The inferences revolving around the unknown θ of a statistical model $\mathcal{M}_\theta(\mathbf{z})$ depend on the model assumptions such as [1]–[5] (Table 1), but the M-S tests for [1]–[5]

Table 4 Misspecified Linear Regression model

Replications N=10, 000	True: $Y_t = 1.5+.15t+.5x_t+u_t$ Estimated: $Y_t=\beta_0+\beta_1 x_t+u_t$			
	$n = 50$		$n = 100$	
True values	Estimates mean	Estimates Std	Estimates mean	Estimates Std
β_0=1.5	0.481	.472	0.223	.327
β_1=.5	1.971	.053	1.993	.021
σ^2=.75	2.964	.392	2.997	.281
$\mathcal{R}^2$=.25	0.983	.004	0.998	.001
t-statistic	Type I error probability		Type I error probability	
H_0: β_0=0	Nominal	Actual	Nominal	Actual
τ_{β_0}	.05	.782	.05	.973
t-statistic	Type I error probability		Type I error probability	
H_0: β_1=0	Nominal	Actual	Nominal	Actual
τ_{β_1}	.05	1.000	.05	1.000

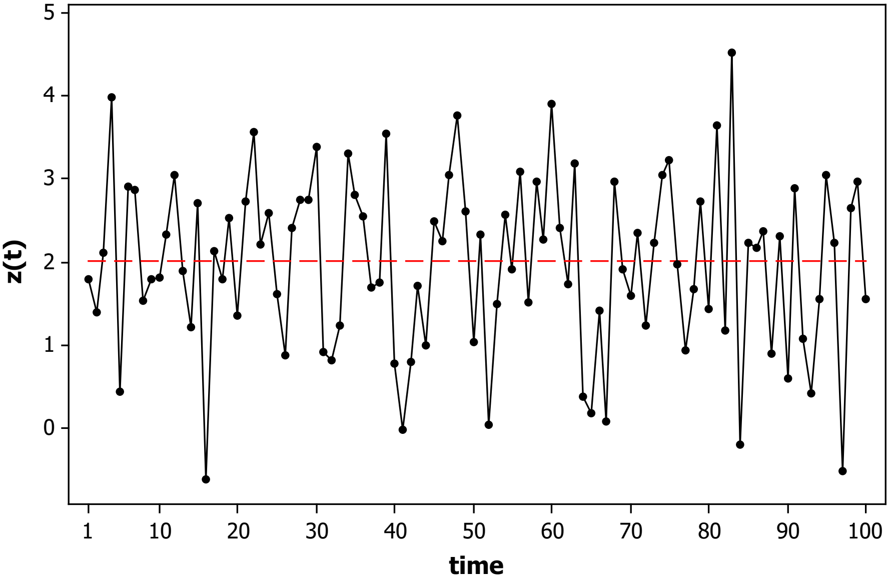

Fig. 4 Typical realization of a NIID(2, 1) process

should not depend on the true values of θ. The logic of M-S tests is this: a test statistic $d(\mathbf{Z})$ is constructed to measure the distance between what is observed $\mathbf{z}_0$ and what is expected assuming the null hypothesis H_0 holds, so as to derive the distribution of $d(\mathbf{Z})$ under H_0. The generic form of the hypothesis of interest in M-S tests is:

$$H_0: \text{ the assumption(s) of statistical model } \mathcal{M}_\theta(\mathbf{z}) \text{ hold for data } \mathbf{z}_0. \tag{8}$$

The relevant p-value would be $p(\mathbf{z}_0){=}P(d(\mathbf{Z}){>}d(\mathbf{z}_0);\,H_0)$, and if it is very small, then $d(\mathbf{z}_0)$ indicates violations of the assumption(s) of $\mathcal{M}_\theta(\mathbf{z})$ being tested. The idea underlying model validation is to construct M-S tests using a $d(\mathbf{Z})$ whose distribution under the null ($\mathcal{M}_\theta(\mathbf{z})$ is valid) is known, and at the same time they have power against potential departures from the model assumptions. M-S tests can be regarded as posing 'secondary' questions to the data as opposed to the primary ones. Whereas primary statistical inferences take place within a specified (or assumed) model $\mathcal{M}_\theta(\mathbf{z})$, the secondary inference has to put its assumptions to the test; so to test $\mathcal{M}_\theta(\mathbf{z})$'s assumptions, one stands outside $\mathcal{M}_\theta(\mathbf{z})$, as it were.

As we have underscored, the *ontological commitments* in selecting a statistical model $\mathcal{M}_\theta(\mathbf{z})$ concern the existence of perceptible chance regularities in data $\mathbf{Z}_0$ that $\mathcal{M}_\theta(\mathbf{z})$ could account for, and the objective in M-S testing is to appraise that; see Spanos (2013). The statistical error term $u_t{:=}Y_t{-}E\,(Y_t|X_t{=}x_t)\,{=}(Y_t{-}\beta_0{-}\beta_1 x_t)$ represents $(Y_t|X_t{=}x_t)$ in terms of which [1]–[5] are specified. Hence, it should come as no surprise that M-S testing probes the estimated errors (*residuals*) from (4):

$$\{\widehat{u}_t{=}(Y_t{-}\widehat{\beta}_0{-}\widehat{\beta}_1 x_t),\ t{=}1,2,...n\}, \tag{9}$$

for possible departures from assumptions [1]–[5] (Table 1) indicated by lingering statistical systematic information. There are other good reasons why M-S tests use the residuals, including the fact that often they constitute a maximal ancillary statistic, which is independent of the minimal sufficient statistic upon which primary inferences are based (Cox and Mayo 2010). Given that they pose very different questions to data $\mathbf{Z}_0$, this independence offers a valuable and powerful tool in separating testing assumptions from testing primary hypotheses; see Spanos (2010b). Should M-S testing of $\mathcal{M}_\theta(\mathbf{z})$ detect significant departures from the model assumptions [1]–[5], a return to the drawing board is called for. The original NIID assumptions for $\{\mathbf{Z}_t,\ t{\in}\mathbb{N}\}$ need reexamination with a view to account for the overlooked statistical information; see Spanos (2006).

By contrast, the traditional approach attempts to address the respecification problem by modifying the error term! Fiddling with the error term, they may accommodate any perceived departures from its original assumptions. We may call this "error-fixing" strategies.

4.3 'Error-fixing' stratagems in model respecification

The tendency in traditional econometrics to seek to 'correct' the error term $\{\varepsilon_t,\ t{\in}\mathbb{N}\}$ assumptions upon discovering that the original model is misspecified is closely tied to the common conception to blur the ontological commitments of the substantive and statistical models. That is, the curve-fitting perspective encourages retaining the relationship $y_t = \alpha + \beta x_t$ given by the theory and 'correcting' the error term probabilistic assumptions.

In contrast, when the statistical model $M_\theta(\mathbf{z})$ is viewed as a parameterization of the process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{Z}_0$, the respecification takes the form of changing the original probabilistic assumptions of $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ in order to account for statistical misspecifications.

Example. Assumption [4] (Independence) of the linear regression model (Table 1), is often tested in the traditional approach by replacing the original error with an AR(1) model for the error term to specify the encompassing model:

$$Y_t{=}\beta_0 + \beta_1 x_t + u_t,\ u_t{=}\rho u_{t-1}{+}\epsilon_t,\ |\rho| < 1. \tag{10}$$

This renders the original model a special case under the null when testing:

$$H_0 : \rho = 0 \text{ vs. } H_1 : \rho \neq 0.$$

When H_0 is rejected, the traditional approach recommends adopting the alternative model $Y_t{=}\beta_0{+}\beta_1 x_t{+}u_t,\ u_t{=}\rho u_{t-1}{+}\epsilon_t,\ \epsilon_t \backsim \text{NIID}(0, \sigma_\epsilon^2)$, as a way to respecify the original model. This, however, is a classic example of *the fallacy of rejection*: misinterpreting evidence *against* H_0 as evidence *for* the particular H_1. Rejecting H_0 provides evidence *for* the generic departure:

$$E(u_t u_s | X_t{=}x_t) \neq 0, t > s,\ t, s{\in}\mathbb{N}, \tag{11}$$

but it does *not* provide evidence *for* the particular form entailed by H_1:

$$H_1 \rightarrow E(u_t u_s | X_t{=}x_t){=}(\tfrac{\rho^{|t-s|}}{1-\rho^2})\sigma_\epsilon^2,\ t > s,\ t, s{\in}\mathbb{N}. \tag{12}$$

The fact that H_1 involves additional assumptions that have not been validated precludes their passing with severity; see Mayo and Spanos (2004). In practice, the alternative hypothesis that could potentially explain the non-independence in data $\mathbf{Z}_0$ can take numerous different forms, including the process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ is Markov dependent instead of independent. Assuming that $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ is Normal, Markov and stationary gives rise to a Dynamic Linear Regression model with statistical GM:

$$Y_t = \alpha_0 + \alpha_1 x_t + \alpha_2 Y_{t-1} + \alpha_3 x_{t-1} + u_t,\ t{\in}\mathbb{N}. \tag{13}$$

The respecified model in (13) includes the encompassing model in (10) as a special case when $\alpha_3{=}-\alpha_1\alpha_2$, which is a gratuitous restriction that is often false in practice; see McGuirk and Spanos (2008). More importantly, (13) brings out the fact that is often overlooked: validating the new model demands that its own assumptions be severely tested against the data.

5 Conclusions

Statistical methodology can mislead us as to what we think we are finding out about the world, not because of inherent idealizations in our models, nor inherent weakness in methods, but to flawed conceptions of the statistical ontology/methodology linking intermediate models to data and substantive phenomena.

The distinction between substantive and statistical models can give rise to fallacious inferences—to those who identify them too readily—but it can also enable building a statistical platform for finding out the strengths and weaknesses of substantive models. If there are any ontological consequences of error statistical modeling and inference, they will not be in terms of the kinds of theories we may learn about. They do not limit us to repeatable economies. It suffices that hypothetical repetitions of sample realizations afford ways to connect the chance regularities in the data, stemming from a phenomenon of interest, to the probabilistic structure of a statistical model accounting for these regularities. The error-statistical framework puts the spotlight on testing if the models are adequate—not necessarily as real-world representations—but as statistically error-reliable platforms for posing questions about the phenomena of interest. The foundation of an adequate statistical model enables the lineup between purely statistical parameters and substantive ones. The fact that the key components of inquiry may essentially be translated into statistical questions explains how we may attain reliable inferences, even if they are directed to highly idealized models.

6 Appendix: M-S testing and auxiliary regressions

In light of the fact that the Linear Regression model (Table 1) is specified in terms of the conditional mean and variance:

$$E\left(Y_t \mid X_t{=}x_t\right){=}\beta_0{+}\beta_1 x_t,\ Var\left(Y_t \mid X_t{=}x_t\right){=}\sigma^2,\quad t{\in}\mathbb{N}, \tag{14}$$

one can test for any departures from the linear regression assumptions: [1] Normality, [2] linearity, [3] homoskedasticity, [4] independence, and [5] t-invariance, by expanding the orthogonal decompositions stemming from (14) (Spanos 1999):

$$u_t{=}\overbrace{E\left(u_t \mid X_t{=}x_t\right)}^{0}{+}v_{1t},\ u_t^2{=}\overbrace{E\left(u_t^2 \mid X_t{=}x_t\right)}^{\sigma^2}{+}v_{2t},\quad t{\in}\mathbb{N}, \tag{15}$$

to include additional terms representing potential violations from these assumptions. Whereas the adequacy of the model assumes that $E\left(u_t \mid X_t{=}x_t\right){=}0$, the true error might be non-zero when any of the assumptions [2]-[5] are invalid; similarly for $E\left(u_t^2 \mid X_t{=}x_t\right){=}\sigma^2$. A particular example of such auxiliary regressions whose terms are only indicative of the kind of terms one could use to seek out any remaining systematic information in the residuals, is:

$$\begin{aligned}\widehat{u}_t &= \overbrace{\gamma_{10}+\gamma_{11}x_t}^{[1],[2],[4],[5]} + \overbrace{\gamma_{12}t{+}\gamma_{13}t^2}^{\overline{[5]}}{+}\overbrace{\gamma_{14}x_t^2}^{\overline{[2]}} + \overbrace{\gamma_{15}x_{t-1}+\gamma_{16}y_{t-1}}^{\overline{[4]}} + v_{1t},\\ &H_0: \gamma_{11}=\gamma_{12}=\gamma_{13}=\gamma_{14}=\gamma_{15}{=}\gamma_{16}{=}0\end{aligned} \tag{16}$$

$$\widehat{u}_t^2 = \overbrace{\gamma_{20}}^{[1],[3],[5]} + \overbrace{\gamma_{21}x_t}^{[3]} + \overbrace{\gamma_{22}t+\gamma_{23}t^2}^{[5]} + \overbrace{\gamma_{24}x_t^2}^{[3]} + \overbrace{\gamma_{25}x_{t-1}^2 + \gamma_{26}y_{t-1}^2}^{[4]} + v_{2t},$$
$$H_0 : \gamma_{21} = \gamma_{22} = \gamma_{23} = \gamma_{24} = \gamma_{25} = \gamma_{26}=0 \qquad (17)$$

In each case the null hypotheses H_0 assert that the model assumptions hold, taking us back to (15). The terms beyond $\gamma_{10} + \gamma_{11}x_t$ in (16) and beyond γ_{20} in (17) represent different types of statistical systematic information that the original model might have overlooked. The interesting upshot of this is that the additional terms represent potential violations, which are expressed in generic terms that represent systematic statistical information already in **Z** and do not directly refer to any specific substantive factors. Their statistical significance, however, raises questions about how generic terms such as t and t^2—which represent substantive ignorance—can be replaced by relevant explanatory variables for substantive adequacy purposes; see Spanos (2010c).

One has reduced the problem of probing for model violations to testing the statistical significance of these additional terms, individually or in groups, using simple t-tests and F-tests (Spanos 1999). A rejection of a null hypothesis indicates departures from the underlying model assumption(s).

References

Baggerly, K. A., & Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*, *3*, 1309–1334.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical*, *71*, 791–799.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Cox, D. R., & Mayo, D. G. (2010). Objectivity and conditionality in frequentist inference. In D. G. Mayo & A. Spanos (Eds.), *Error and inference* (pp. 276–304). Cambridge: Cambridge University Press.

Fama, E. F., & French, K. R. (2004). The capital asset pricing model: Theory and evidence. *The Journal of Economic Perspectives*, *18*, 25–46.

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

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945–1964. *Journal of Finance*, *23*, 389–416.

Lai, T. L., & Xing, H. (2008). *Statistical models and methods for financial markets*. NY: Springer.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, *47*, 13–37.

McGuirk, A., & Spanos, A. (2008). Revisiting error autocorrelation correction: Common factor restrictions and granger non-causality. *Oxford Bulletin of Economics and Statistics*, *71*, 273–294.

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

Mayo, D. G. (1997). Duhem's problem, the Bayesian way, and error statistics, or "What's belief got to do with It?". *Philosophy of Science*, *64*, 222–244.

Mayo, D. G. (2010a). Learning from error, severe testing, and the growth of theoretical knowledge. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 28–57). Cambridge: Cambridge University Press.

Mayo, D.G. (2010b). Learning from error: The theoretical significance of experimental knowledge, *The modern schoolman*. Guest editor, Kent Staley. Volume 87, Issue 3/4, March/May 2010, *Experimental and theoretical knowledge*, The 9th Henle conference in the history of philosophy, 191–217.

Mayo, D. G., & Cox, D. R. (2010). Frequentist statistics as a theory of inductive inference. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability,* *and the objectivity and rationality of science* (Vol. 7, pp. 247–275). Cambridge: Cambridge University Press.

Mayo, D. G., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science*, *71*, 1007–1025.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science*, *57*, 323–57.

Mayo, D. G., & Spanos, A. (2010). *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*. Cambridge: Cambridge University Press.

Mayo, D. G., & Spanos, A. (2011). Error statistics. In D. Gabbay, P. Thagard, & J. Woods (Eds.), *Philosophy of statistics, handbook of philosophy of science*. Amsterdam: Elsevier.

Potti, A., Dressman, H. K., Bild, A., Riedel, R. F., Chan, G., Sayer, R., et al. (2006). Genomic signatures to guide the use of chemotherapeutics. *National Medicine*, *12*, 1294–1300.

Senn, S. J. (2001). Two cheers for P-values. *Journal of Epidemiology and Biostatistics*, *6*(2), 193–204.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, *19*, 425–442.

Spanos, A. (1990). The simultaneous equations model revisited: Statistical adequacy and identification. *Journal of Econometrics*, *44*, 87–108.

Spanos, A. (1999). *Probability theory and statistical inference: Econometric modeling with observational data*. Cambridge: Cambridge University Press.

Spanos, A. (2006). Where do statistical models come from? Revisiting the problem of specification, pp. 98–119 In *Optimality: The second Erich L. Lehmann Symposium*, Rojo, J. (Ed.) Lecture notes-monograph series, vol. 49, Institute of Mathematical Statistics.

Spanos, A. (2007). Curve-fitting, the reliability of inductive inference and the error-statistical approach. *Philosophy of Science*, *74*(5), 1046–1066.

Spanos, A. (2010a). Theory testing in economics and the error statistical perspective. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 202–246). Cambridge: Cambridge University Press.

Spanos, A. (2010b). Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification. *Journal of Econometrics*, *158*, 204–220.

Spanos, A. (2010c). Statistical adequacy and the trustworthiness of empirical evidence: Statistical vs. substantive information. *Economic Modelling*, *27*, 1436–1452.

Spanos, A. (2010d). The discovery of argon: A case for learning from data? *Philosophy of Science*, *77*(3), 359–380.

Spanos, A. (2010e). Is frequentist testing vulnerable to the base-rate fallacy? *Philosophy of Science*, *77*, 565–583.

Spanos, A. (2013). A frequentist interpretation of probability for model-based inductive inference. *Synthese*, *190*, 1555–1585.

Spanos, A., & McGuirk, A. (2001). The model specification problem from a probabilistic reduction perspective. *Journal of the American Agricultural Association*, *83*, 1168–1176.