JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12200

MIS-SPECIFICATION TESTING IN RETROSPECT

Aris Spanos*

Department of Economics
Virginia Tech
Blacksburg VA

Abstract. The primary objective of this paper is threefold. First, to undertake a retrospective view of Mis-Specification (M-S) testing, going back to the early 20th century, with a view to (i) place it in the broader context of modeling and inference and (ii) bring out some of its special features. Second, to call into question several widely used arguments undermining the importance of M-S testing in favor of relying on weak probabilistic assumptions in conjunction with generic robustness claims and asymptotic inference. Third, to bring out the crucial role of M-S testing in securing trustworthy inference results. This is achieved by extending/modifying Fisher's statistical framework with a view to draw a clear line between the modeling and the inference facets of statistical induction. The proposed framework untangles the statistical from the substantive (structural) model and focuses on how to secure the adequacy of the statistical model before probing for substantive adequacy. A case is made for using joint M-S tests based on custom-built auxiliary regressions with a view to enhance the effectiveness and reliability of probing for potential statistical misspecifications.

Keywords. Error probabilities; Misspecification testing; Neyman–Pearson testing; Nontestable nssumptions; Reliability of inference; Respecification; Robustness; Specification; Statistical model; Statistical vs. substantive adequacy; Weak probabilistic assumptions

1. Introduction

The problem of misspecification arises when certain assumptions invoked by a statistical inference procedure are invalid. Departures from the invoked assumptions distort the sampling distribution of a statistic (estimator, test, and predictor), and as a result, the reliability of an inference procedure is often undermined. For instance, invalid assumptions could give rise to inconsistent estimators or/and sizeable discrepancies between the actual types I and II error probabilities and the nominal ones – the ones derived by invoking these assumptions. Applying a 0.05 significance level test, when the actual type I error is closer to 0.9, will lead an inference astray.

Mis-Specification (M-S) testing aims to assess the validity of the assumptions comprising a statistical model. Its usefulness is twofold:

(i) it can alert a modeler to potential problems with unreliable inferences,
(ii) it can shed light on the nature of departures from the model assumptions.

*Corresponding author contact email: aris@vt.edu; Tel: +540 231 7707.

Since its introduction at the beginning of the 20th century, M-S testing has been one of the most confused and confusing facets of statistical modeling. As a result, its role and importance in securing the reliability and precision of inference has been seriously undervalued by the statistics and econometric literatures. The current conventional wisdom relies primarily on (i) weak, but often nontestable, probabilistic assumptions, combined with (ii) asymptotic inference results, and (iii) vague robustness claims, as a substitute for using comprehensive M-S testing to establish statistical adequacy: the validity of the probabilistic assumptions comprising the model.

The current neglect of statistical adequacy is mainly due to the fact that the current statistical modeling is beclouded by conceptual unclarities stemming from the absence of a coherent empirical modeling framework that delineates the different facets of modeling and inference beyond Fisher's (1922) initial categories: *Specification* (the choice of the statistical model), *Estimation* and *Distribution* (inferences based on sampling distributions). Fisher's grouping of M-S testing under sistribution, however, left a lot of unanswered questions concerning its nature and role. How does M-S testing differ from other forms of inference? How would one go about establishing the validity of the model assumptions. What would one do when certain assumptions are found wanting? A pioneer of 20th century statistics attests to that by acknowledging the absence of a systematic way to validate statistical models:

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

The question of specifying and validating a statistical model has received comparatively little attention for a number of reasons. The most crucial stems from viewing empirical modeling as curve fitting that revolves around the quantification of the substantive (structural) model presumed "true." In such cases, the theory information is treated as "established knowledge" instead of tentative conjectures to be tested against the data. From this perspective, the statistical model – comprising the probabilistic assumptions imposed on the data – is implicitly specified via the error term(s) attached to the structural model. As a result, the statistical premises are inadvertently blended with the substantive premises of inference, and the empirical literature conflates two very different forms of misspecification: statistical and substantive. The end result is that a typical estimated model in applied econometrics is usually both statistically and substantively misspecified, but one has no way of separating the two and apportioning blame. This raises a thorny issue that plagues inductive inference more broadly, known as Duhem's problem; see Mayo (1996). Hence, it should come as no surprise that econometrics textbooks consider 'omitted variables' as the most serious form of statistical misspecification (omitted variables bias and inconsistency (Greene, 2011), when in fact it has nothing to do with the statistical assumptions; it is an issue relating to substantive adequacy (Spanos, 2006).

The primary objective of this paper is to undertake a retrospective view of M-S testing with a view to shed light on the issues and problems raised in establishing statistical adequacy. The first M-S test of the modern era was Pearson's (1900) goodness-of-fit test. It is rarely viewed as such, however, because the line between the latter and other types of testing is blurred to this day. In an attempt to shed some light on this and other foundational issues raised by M-S testing, Section 2 provides a brief historical review of statistical testing in general, and M-S testing in particular, in order that the key differences between the two are foregrounded. Section 3 discusses the importance of M-S testing in establishing the adequacy of a statistical model. Section 4 revisits the ambivalent attitude toward M-S testing and calls into question various arguments often used to justify its neglect. Section 5 discusses an extension/elaboration of Fisher's (1922) statistical modeling framework intending to address the issues and problems brought up in Sections 2– 4, as well as provide a more coherent and effective approach to M-S testing with a view to establish statistical adequacy.

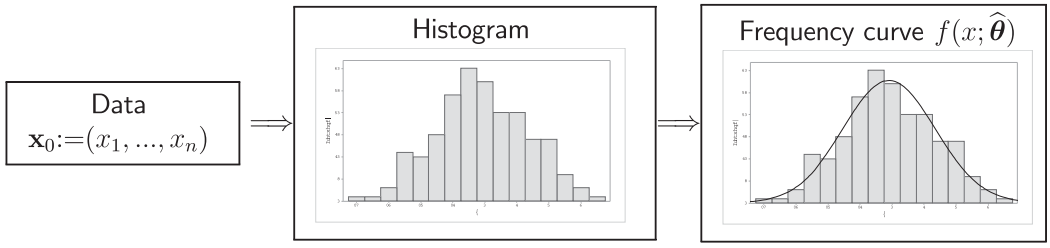

Figure 1. Karl Pearson's Approach to Statistics.

2. M-S Testing: A Brief Historical Overview

2.1 *Karl Pearson*

Viewed retrospectively, the modern era on M-S testing was initiated by Pearson's (1900) paper on goodness-of-fit. There is an element of anachronism in this view, however, because Pearson's approach to statistics was very different from the current approach that is primarily due to Fisher (1922).

In its simplest form, Pearson's approach would begin with a set of data $\mathbf{x}_0 := (x_1, x_2, \ldots, x_n)$ in search of a descriptive model in the form of a frequency curve; see figure 1. This would be attained by summarizing $\mathbf{x}_0$ in the form of a histogram, and then fitting a frequency curve $f_0(x)$ from the Pearson family generated by:

$$\frac{d\ln f(x)}{dx} = \frac{(x-\theta_1)}{\theta_2+\theta_3 x+\theta_4 x^2},\ \boldsymbol{\theta} := (\theta_1,\theta_2,\theta_3,\theta_4) \in \Theta \subset \mathbb{R}^4,\ x \in \mathbb{R}_X \tag{1}$$

to describe it as closely as possible. The choice of $f_0(x)$ would be based on the values of the estimated parameters $\widehat{\boldsymbol{\theta}} := (\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \widehat{\theta}_4)$ using Pearson's method of moments.

To test the appropriateness of the choice of the frequency curve $f_0(x)$, Pearson (1900) proposed a goodness-of-fit test based on the standardized *distance function*:

$$\eta(\mathbf{X}) = \sum_{i=1}^{m} \frac{(\widehat{f}_i - f_i)^2}{f_i} = n\sum_{i=1}^{m} \frac{[(\widehat{f}_i/n) - (f_i/n)]^2}{(f_i/n)} \sim \chi^2(m) \tag{2}$$

where $(\widehat{f}_i,\ i=1,2,\ldots,m)$ and $(f_i,\ i=1,2,\ldots,m)$ denote the empirical $(f(x;\widehat{\boldsymbol{\theta}}))$ and assumed $(f_0(x))$ frequencies. For inferring whether the choice of $f_0(x)$ was appropriate or not, Pearson introduced a primitive version of the p-value:

$$\mathbb{P}(\eta(\mathbf{X}) > \eta(\mathbf{x}_0)) = p(\mathbf{x}_0) \tag{3}$$

His rationale was that the bigger the value of $\eta(\mathbf{x}_0)$, the worse the goodness-of-fit, and the smaller the tail area probability.

It is important to note that the choice of $f_0(x)$ is not the only probabilistic assumption imposed on the data. For the histogram to be a statistically meaningful summary of the data, the independence and identically distributed (IID) assumptions are also implicitly imposed on the data $\mathbf{x}_0$; see Spanos (1999).

2.2 *R.A. Fisher*

Fisher's (1922) approach to statistical inference was partly inspired by the "Student" (1908) path breaking paper, written by William Gosset. His main result was to deduce the *finite* sampling distribution of the pivotal quantity:

$$\frac{\sqrt{n}(\overline{X}_n-\mu)}{s}\sim \mathsf{St}(n-1),\quad \text{for any } n>1 \tag{4}$$

where $\overline{X}_n=\frac{1}{n}\sum_{t=1}^{n}X_t$, $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{\mu}_n)^2$, and $\mathsf{St}(n-1)$ denotes a Student's t distribution with $(n-1)$ degrees of freedom. Gosset's demonstration was based on Karl Pearson's curve-fitting over histograms using simulated data.

A formal proof of (4) was given by Fisher (1915), who brought out explicitly the invoked probabilistic assumptions in the form of the *simple Normal model*:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}): X_t\sim \text{NIID}(\mu,\sigma^2),\quad t=1,2,\ldots,n \tag{5}$$

where "NIID" stands for normal, independent, and identically distributed.

Fisher's (1922) most remarkable achievement was to initiate the recasting of *statistical induction* into its modern variant by introducing explicitly the notion of a statistical model such as (5). Instead of starting with data $\mathbf{x}_0$ in search of a descriptive model, like Pearson, he proposed to interpret the data as a *random sample* from a prespecified statistical model ("hypothetical infinite population"), specified in terms of a few unknown parameters:

> "... the object of statistical methods is the reduction of data ... This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters ..." (p. 311)
>
> "The postulate of randomness thus resolves itself into the question, "Of what population is this a random sample" which must frequently be asked by every practical statistician. It will be seen from the above examples that the process of the reduction of data is, even in the simplest cases, performed by interpreting the available observations as a sample from a hypothetical infinite population ..." (p. 312)

This is not a trivial rearrangement of Pearson's procedure, but a complete recasting of the problem of statistical induction, with the notion of a parametric statistical model delimiting its premises. A key element of Fisher's perspective comes in the form of viewing data $\mathbf{x}_0$ as a typical realization from $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

Fisher (1922, p. 313) went on to define the different stages of statistical modeling:

> "The problems which arise in reduction of data may be conveniently divided into three types:
>
> (1) Problems of Specification. These arise in the choice of the mathematical form of the population.
> (2) Problems of Estimation. These involve the choice of methods of calculating from a sample statistical derivatives, or as we shall call them statistics, which are designed to estimate the values of the parameters of the hypothetical population.
> (3) Problems of Distribution. These include discussions of the distribution of statistics derived from samples, or in general any functions of quantities whose distribution is known."

Focusing on specification, he called for testing the adequacy of the model:

> "As regards problems of specification, ... the adequacy of our choice may be tested *a posteriori* ." (p. 314)

Despite the well-documented animosity between the two pioneers of modern statistics (see Stigler, 2005), Fisher went on to offer a very rare praise for Karl Pearson. In addition to the introduction of the Pearson family of distributions, Fisher praised Pearson for being a pioneer in M-S testing:

> "Nor is the introduction of the Pearsonian system of frequency curves the only contribution which their author has made to the solution of problems of specification: of even greater importance is the introduction of an objective criterion of goodness of fit. For empirical as the specification of the hypothetical population 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)

He clearly appreciated the importance of M-S testing for providing a justification for statistical induction that stems from being able to objectively test the adequacy of its premises, i.e., the probabilistic assumptions imposed on data $\mathbf{x}_0$:

> "The possibility of developing complete and self-contained tests of goodness of fit deserves very careful consideration, since therein lies our justification for the free use which is made of empirical frequency formulae." (p. 314)

Fisher used the result (4) to formulate his significance testing by introducing the notion of a "null" hypothesis framed in terms of the unknown parameter(s) of the prespecified model (5), say the mean: $H_0 : \mu = \mu_0$.

He realized that when the unknown μ in (4) is replaced by the hypothesized null value μ_0 yields a test statistic, whose evaluation under H_0 gives rise the same distributional result:

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{s} \overset{H_0}{\sim} \mathsf{St}(n-1) \tag{6}$$

Fisher was explicit about the nature of *reasoning* underlying significance testing:

> "In general, tests of significance are based on *hypothetical* probabilities calculated from their null hypotheses." (Fisher, 1956, p. 47)

He then used (6) to formalize Pearson's p-value:

$$\mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0);\ H_0) = p(\mathbf{x}_0) \tag{7}$$

interpreting it as an indicator of "discordance" between data $\mathbf{x}_0$ and H_0, using a threshold value c_0 [e.g., 0.01, 0.025, 0.05]; $p(\mathbf{x}_0) < c_0$ data $\mathbf{x}_0$ indicate that H_0 is discordant with data $\mathbf{x}_0$. That is, Fisher (1922) recast Pearson's test into his significance testing based on p-values by emphasizing the use of finite sampling distribution (p. 314).

His recasting of statistical induction enabled Fisher to reframe Pearson's chi-square test and ground it on firmer foundations by:

(i) unveiling the implicit probabilistic assumptions of IID underlying (2),
(ii) explaining that (2) constitutes an *asymptotic* approximation (as $n \to \infty$),
(iii) (2) stems from evaluating the test statistic *under the null* (H_0: $\mu = \mu_0$),
(iv) making explicit the hypothesis of interest:

$$H_0 : f_0(x) = f^*(x) \in \mathsf{Pearson}(\boldsymbol{\theta})$$

where $f_0(x)$ denotes the assumed density and $f^*(x)$ is the "true" density,
(v) correcting Pearson's degrees of freedom, and
(vi) formalizing Pearson's p-value into: $\mathbb{P}(\eta(\mathbf{X}) > \eta(\mathbf{x}_0); H_0) = p(\mathbf{x}_0)$.

Note that the expression "$f^*(x)$ that is the 'true' density" is a shorthand for saying that "data $\mathbf{x}_0$ constitute a typical realization of the sample $\mathbf{X}$ with distribution $f^*(\mathbf{x})$." It is interesting to note that chapters 3 and 4 of Fisher's (1925) book discuss several M-S tests for departures from Normality, Independence, and Homogeneity (ID). In chapters 6–8, he discusses significance tests pertaining to unknown parameters, such as means, variances, covariances, correlation coefficients, and regression coefficients. What is missing from Fisher's discussion is an attempt to bring out any differences between M-S testing and the traditional significance tests.

2.3 *Neyman and Pearson*

A retrospective view of Neyman–Pearson's (N-P) extension/modification of Fisher's testing framework reveals the following changes that enabled them to put forward an optimal theory of testing. The N-P approach:

1. Narrowed down the scope of Fisher testing by focusing exclusively on testing *within* the boundaries of the prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.
2. Framed the hypotheses of interest in terms of the parameters $\theta \in \Theta$.
3. Extended Fisher's null hypothesis (H_0) from a single value to a subset of Θ, say H_0: $\theta \in \Theta_0$.
4. Supplemented H_0 with an alternative hypothesis (H_1), defined to be its complement relative to the parameter space, say H_1: $\theta \in \Theta_1 := \Theta - \Theta_0$. This defines a partition of the parameter space that corresponds to a partition of the sample space into an acceptance (C_0) and a rejection (C_1) region:

$$\mathbb{R}^n_X = \left\{ \begin{array}{c} \boxed{C_0} \leftrightarrow \boxed{\Theta_0} \\ \underline{\boxed{C_1}} \leftrightarrow \underline{\boxed{\Theta_1}} \end{array} \right\} = \Theta$$

These modifications/changes to Fisher's framing enabled Neyman and Pearson (1933) to define optimality in terms of uniformly most powerful (UMP) tests.

Returning to Pearson's goodness-of-fit test, the only generic way to add an alternative hypothesis is in terms of negating the null:

$$H_0 : f^*(x) = f_0(x) \text{ vs. } H_1 : f^*(x) \neq f_0(x), \text{ for all } x \in \mathbb{R}_X$$

This renders it an M-S test but *not* an N-P test since that latter:

(i) is always framed in terms of the parameters of the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in question and
(ii) probes *within* the boundaries of the prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

Does that imply that all goodness-of-fit tests are M-S tests? The answer is *no* because there goodness-of-fit tests in the context of other statistical models that are proper N-P tests since they satisfy (i)–(ii). The quintessential example is the F-test framed in terms of the R^2 in the context of the linear regression model; see Greene (2011).

A key question since the 1930s has been: How does M-S testing differ from traditional hypothesis testing? One can argue that no clear and persuasive answers have been provided by the subsequent statistical literature.

2.4 *Misspecification: Gosset, Fisher, and Egon Pearson*

During the early 1920s, there were few misgivings about the appropriateness of the IID assumptions because the modeling was primarily based on experimental data, but doubts were raised about the

appropriateness of the Normality assumption. The first to raise these doubts in a letter to Fisher dated 1923 was Gosset:

> "What I should like you to do is to find a solution for some other population than a normal one." (see Lehmann, 1999)

He went on to explain that he tried the rectangular (uniform) distribution but made no progress, and he was seeking Fisher's help in tackling the "robustness" problem. In his reply that was unfortunately lost, Fisher must have derived the sampling distribution of $\tau(\mathbf{X})$ assuming some skewed distribution (possibly log-Normal). We know this from Gosset's reply (quoted by Lehmann, 1999):

> "I like the result for z [t-test] in the case of that horrible curve you are so fond of. I take it that in skew curves the distribution of$z$ is skew in the opposite direction."

After this exchange, Fisher showed little interest in following up on Gosset's requests to address the problem of working out the implications of non-Normality for the reliability of inference associated with the simple Normal model; t, chi-square, and F tests. Indeed, in one of his letters to Gosset, Fisher washed his hands by saying that "it was none of his business," to derive the implications of departures from Normality.

Egon Pearson, however, shared Gosset's concerns on the robustness of Normal-based tests that are under non-Normality, and tried to address the issue in a series of papers in the late 1920s and early 1930s using simulation; see Pearson (1930). Gosset realized that simulation alone was not enough, and asked Fisher for help:

> "How much does it [non-Normality] matter? And in fact that is your business: none of the rest of us have the slightest chance of solving the problem: we can play about with samples [i.e. perform simulation studies], I am not belittling E.S.P.'s work, but it is up to you to get us a proper solution." (quoted by Lehmann, 1999).

In this passage, one can discern the high-esteem Gosset held Fisher for his conceptual and technical abilities. Fisher's reply was equally blunt:

> "I do not think what you are doing with nonnormal distributions is at all my business, and I doubt if it is the right approach. What I think is my business is the detailed examination of the data, and of the methods of collection, to determine what information they are fit to give, and how they should be improved to give more or other information. In this job it has never been my experience to make the variation more normal; I believe even in extreme cases a change of variate [i.e. a transformation] will do all that is wanted. ... Where I differ from you, I suppose, is in regarding normality as only a part of the difficulty of getting data [inference]; viewed in this collection of difficulties I think you will see that it [non-normality] is one of the least important." (quoted by Lehmann, 1999).

It is clear from this that Fisher understood the problem of departures from the statistical model assumptions very differently from Gosset. His answer alludes to three issues that were, unfortunately, not well understood at the time or since.

1. Departures from IID are considerably more serious for the reliability of Normal-based inference than Normality. The truth of the matter is that there are no robustness results for generic departures from the IID assumptions. To make matters worse, no easy "corrections" are available either.
2. Deriving the consequences of non-Normality on the reliability of Normal-based inference is not the right approach in addressing the problem of departures from the assumptions because this ignores the "optimality" issue.

Unfortunately, the subsequent statistical literature on "robustness" has largely ignored Fisher's insights, viewing the problem of robustness as a choice between Normality or bust. What about optimal tests under different non-Normal distributions? This literature has focused exclusively on comparisons between the error probabilities of the t-test with those of a robust test under non-Normality, and ignores the fact that the t-test is no longer the optimal test. For instance, the nonparametric literature in the late 1940s and early 1950s established that the t-test is robust to non-Normal distributions that are *symmetric*; see Geary (1947). It has been shown in this literature that the relative *asymptotic efficiency* of the Wilcoxon–Mann–Witney (W-M-W) test relative to the t-test is 1.0 when the distribution is uniform instead of Normal and ∞ when the distribution is Cauchy; see Hettmansperger (1984). These comparisons, however, are questionable. In the case of the Cauchy, it is meaningless to compare the t-test with that of W-M-W test because the t-test is based on the mean and variance of $\overline{X}$, neither of which exists! The case of the uniform distribution:

$$X_k \sim \mathsf{U}(a-\mu, a+\mu),\ f(x) = \frac{1}{2\mu},\ (a-\mu) \leq x \leq (a+\mu),\ \mu > 0$$

the optimal test for the hypotheses H_0: $\mu = \mu_0$ vs. H_1: $\mu \neq \mu_0$, is no longer the t-test, but the test (Neyman and Pearson, 1928):

$$w(\mathbf{X}) = \frac{(n-1)\left(\left(\frac{1}{2}\right)\left[X_{[1]} + X_{[n]}\right] - \mu_0\right)}{\left(\frac{1}{2}\right)\left[X_{[n]} - X_{[1]}\right]} \overset{H_0}{\sim} \mathsf{F}(2, 2(n-1)),\ C_1 := \{\mathbf{x} : |w(\mathbf{x})| > c_\alpha\} \tag{8}$$

where $(X_{[1]}, X_{[n]})$ denote the smallest and the largest elements of the sample, and $\mathsf{F}(2, 2(n-1))$ is the F distribution with 2 and $2(n-1)$ degrees of freedom. Hence, the relevant comparison is not between the t-test under the uniform distribution and the robust test, but the latter with the test based on (8), an optimal N-P test.

One might object to this line of reasoning as impractical because the point of the robustness literature is to "buy" insurance against non-Normality. This is a very weak argument because one can easily distinguish between a realization of an IID process from a Normal, Uniform, Student's t, and other symmetric distributions using a simple t-plot. Difficulties in making a clear diagnosis arise when the underlying process is not IID, but there are ways to overcome such difficulties; see Spanos (1999, ch. 5).

(c) Fisher's alluding to ".. the detailed examination of the data . . . " in the above quotation can be interpreted as arguing that a more effective way to address the problem of non-Normality in empirical modeling is:
  (i) to *test* the Normality assumption, and if any departures are detected,
  (ii) proceed to *respecify* the statistical model by selecting another distribution.

Indeed, Fisher (1930) went on to derive the sampling distributions of the sample skewness and kurtosis coefficients $(\widehat{\alpha}_3, \widehat{\alpha}_4)$ under Normality, and suggested their use to test Normality.

In summary, Gosset's initial concerns relating to departures from the assumption of Normality and Fisher's reaction gave rise to several interrelated lines of research that unfolded over the next four decades.

(i) The implications of non-Normality for the reliability of inference procedures associated with the simple Normal model, such as the t-test and F-test; see Pitman (1937), Geary (1947), Box (1953), and Box and Watson (1962).
(ii) The use of distribution free models to derive nonparametric tests that rely on indirect distributional assumptions; see Kolmogorov (1941), Lilliefors (1967), Mood (1940), Wald and Wolfowitz (1943), Wolfowitz (1944), and Wilcoxon (1945).
(iii) Testing the assumption of Normality; see Pearson (1930, 1931, 1935), Geary (1935), Shapiro and Wilks (1965), and D'Agostino and Pearson (1973). Fisher's (1930) derivation of the sampling

distributions of the sample third ($\widehat{\alpha}_3$) and fourth ($\widehat{\alpha}_4$) central moments led the *skewness-kurtosis test* based on:

$$SK(\mathbf{X}) = \tfrac{n}{6}\widehat{\alpha}_3^2 + \tfrac{n}{24}(\widehat{\alpha}_4 - 3)^2 \overset{H_0}{\underset{\alpha}{\sim}} \chi^2(2), \quad \mathbb{P}(SK(\mathbf{X}) > SK(\mathbf{x}_0); H_0) = p(\mathbf{x}_0)$$

D'Agostino and Pearson (1973) proposed a modification of this test to improve its finite sample properties.

(iv) Respecifying the statistical model to account for any detected departures from its probabilistic assumptions. This was the most neglected aspect of Fisher's perspective on statistical misspecification.

2.5 *Jerzy Neyman*

In an attempt to enhance the power of Pearson's chi-square test, Neyman (1937) modified the problem in two important respects. First, he used the probability integral transformation: $Z = F_0(X) \sim \mathrm{U}(0, 1)$, where X is a continuous random variable with cumulative distribution function $F_0(.)$

Second, he particularized the generic alternative hypothesis H_1: $f^*(z) \neq f_0(z)$ into:

$$H_1 : f^*(z) \in f_p(z; \boldsymbol{\varphi}), \quad \text{for all } z \in [0, 1]$$

The proposed family of distributions takes the form:

$$f_p(z; \boldsymbol{\varphi}) = c(\boldsymbol{\varphi}) \exp\left\{\sum_{i=1}^{p} \varphi_i h_i(z)\right\}, \quad \text{for } z \in [0, 1], \ p < n$$

where $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p)$, $c(\boldsymbol{\varphi})$ is a normalizing constant such that: (i) $\int_{z \in [0,1]} f_p(z; \boldsymbol{\varphi}) dz = 1$, and (ii) $h_i(z)$, $i = 1, \ldots, p$, denote Legendre polynomials defined over [0, 1]:

$$h_0(z) = 1, \ h_1(z) = \sqrt{3}(2z - 1), \ h_2(z) = \sqrt{5}(6z^2 - 6z + 1)$$

$$h_3(z) = \sqrt{7}(20z^3 - 30z^2 + 12z - 1), \text{ etc.}$$

This enabled Neyman to parameterize H_0 and H_1 in terms of $\boldsymbol{\varphi}$:

$$H_0 : \boldsymbol{\varphi} = 0, \text{ vs. } H_1 : \boldsymbol{\varphi} \neq 0 \tag{9}$$

giving rise to the test defined by:

$$\begin{aligned} &\psi_p^2(\mathbf{Z}) = \sum_{i=1}^{p} u_i^2 = \sum_{i=1}^{p} \left[\sum_{k=1}^{n} \tfrac{h_i(z_k)}{\sqrt{n}}\right]^2 \underset{a}{\sim} \chi^2(\delta; p), \ \delta = \sum_{i=1}^{p} \varphi_i^2 \\ &C_1(\alpha) = \left\{\mathbf{z} : \psi_p^2(\mathbf{z})) > c_\alpha\right\}, \qquad \int_{c_\alpha}^{\infty} \chi^2(p) dz \end{aligned} \tag{10}$$

The idea behind Neyman's smooth test is that the individual components:

$$u_i^2 = \left(\sum_{k=1}^{n} \tfrac{h_i(z_k)}{\sqrt{n}}\right)^2, \quad i = 1, 2, \ldots, p \tag{11}$$

provides information as to the direction of departure from $F_0(.)$. Neyman (1937) was able to show that this test is locally UMP unbiased and symmetric of size α. He recommended that $p = 5$ will be sufficient for detecting most forms of departures.

The importance this reformulation stems from the fact that Neyman's test includes the Pearson chi-square test as a special case with a particular implicit alternative. That is, assuming that N_k is the observed number of X_is in the interval k, then $N_k \sim \mathrm{Bin}(p_k, n)$, $k = 1, 2, \ldots, p$, under $f_0(x)$ is valid. This yields the test statistic:

$$\eta(\mathbf{X}) = \sum_{k=1}^{p} \tfrac{(N_k - np_k)^2}{np_k} \overset{H_0}{\underset{\alpha}{\sim}} \chi^2(p - 1) \tag{12}$$

Table 1. Linear Regression (LR) Model

$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u},$
[i] $E(\mathbf{u}\mid\mathbf{X})=\mathbf{0}$, [ii] $Cov(\mathbf{u}\mid\mathbf{X})=\sigma_0^2\mathbf{I}_n$, [iii] Rank$(\mathbf{X})=k<n$.

Table 2. Autocorrelation-Corrected (A-C) LR Model

$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}=\rho\boldsymbol{\varepsilon}_{-1}+\mathbf{u},\ \lvert\rho\rvert<1,$
[i] $E(\mathbf{u}\mid\mathbf{X})=\mathbf{0}$, [ii]* Cov$(\mathbf{u}\mid\mathbf{X})=\sigma_0^2\mathbf{V}_n$, [iii] Rank$(\mathbf{X})=k<n$.

with an implicit local alternative of the form: $\pi_k=p_k+(\delta_k/\sqrt{n})$,
where the δ_ks sum to zero, i.e., $\sum_{k=1}^{p}\delta_k=0$. In relation to (11), this form of a *local alternative* can be interpreted as including only the first term:

$$H_0:\boldsymbol{\varphi}=0,\ \text{ vs. }\ H_1\colon\boldsymbol{\varphi}=\left(\boldsymbol{\delta}/\sqrt{n}\right) \tag{13}$$

which brings out its limited probing in M-S testing.

Neyman's (1937) "smooth" test demonstrated that the alternative in a M-S test does not have to be a generic negation of the null, but can take the form of an explicit family of distributions that nests parametrically the original distribution. This method of constructing M-S tests was popularized by the next influential test.

2.6 *Durbin and Watson*

The first misspecification test for serial correlation in the context of the simple Normal model (Table 1) was given by Von Neumann (1941) based on:

$$v(\mathbf{X})=\left[\sum_{k=2}^{n}(\widehat{u}_k-\widehat{u}_{k-1})^2\right]\Big/\left[\sum_{k=1}^{n}\widehat{u}_k^2\right]\simeq 2(1-\widehat{\rho}),\ \widehat{u}_k=X_k-\widehat{\mu},\ \widehat{\mu}=\frac{1}{n}\sum_{k=1}^{n}X_k$$

Anderson (1942) derived the distribution of: $\widehat{\rho}=[\sum_{k=2}^{n}(X_k-\widehat{\mu})(X_{k-1}-\widehat{\mu})]/[\sum_{k=2}^{n}(X_{k-1}-\widehat{\mu})^2]$.

Durbin and Watson (1950) extended these results to the linear regression (LR) model (Table 1) by relaxing the *noncorrelation* assumption in [ii] (Table 1).

Durbin–Watson (1950) proposed the first M-S test that became part of the practitioner's tool kit almost immediately. Their multifaceted contribution can be summarized in the following steps.

Step 1. They particularized the generic departure from [ii]:

$$E(\varepsilon_t\varepsilon_s\mid\mathbf{X}_t=\mathbf{x}_t)\neq 0,\ t\neq s,\ t,s=1,2,\ldots,n \tag{14}$$

by postulating the AR(1) model for the error term: $\varepsilon_t=\rho\varepsilon_{t-1}+u_t,\ |\rho|<1$.

Step 2. They nested the original model into an encompassing model in Table 2.

Step 3. In the context of the AC-LR model, the temporal independence assumption could be tested in terms of the hypotheses:

$$H_0:\rho=0,\qquad \text{vs. }\ H_1:\rho\neq 0 \tag{15}$$

using the D-W test based on a statistic and a rejection region:

$$\text{D-W}(\mathbf{y}) = \left[\sum_{t=1}^{n}\widehat{\varepsilon}_t^2\right]^{-1}\sum_{t=2}^{n}(\widehat{\varepsilon}_t - \widehat{\varepsilon}_{t-1})^2, \quad C_1 = \{\mathbf{y} : d_U(\alpha) < \text{D-W}(\mathbf{y}) < d_L(\alpha)\} \tag{16}$$

where $\widehat{\varepsilon}_t = y_t - \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}$ denotes the OLS residuals, and $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is the OLS estimator. When the observed test statistic D-W($\mathbf{y}_0$) is smaller (bigger) than the lower (upper) bound $d_L(\alpha)$ ($d_U(\alpha)$), H_0 is rejected.

What is especially noteworthy is that Durbin and Watson refused to commit themselves on what one should do next, by declaring that:

> "We shall not be concerned in either paper with the question of what should be done if the test gives an unfavorable result." (p. 409)

Step 4. Cochrane and Orcutt (1949) proposed an answer to the *respecification* question: adopt the particularized alternative of the D-W test (Table 2). When assumption [ii]* is true and $\mathbf{V}_n$ is known, this strategy recommends replacing $\widehat{\boldsymbol{\beta}}$ with the more efficient Aitken's (1935) GLS estimator: $\breve{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{V}_n^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}_n^{-1}\mathbf{y}$.

When $\mathbf{V}_n$ is unknown, Cochrane and Orcutt proposed a way to estimate it to derive a *Feasible GLS* estimator $\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^\top\widetilde{\mathbf{V}}_n^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\widetilde{\mathbf{V}}_n^{-1}\mathbf{y}$. Their proposals were adopted wholeheartedly by textbook econometrics; see Greene (2011).

2.7 *Box and Jenkins*

The important breakthrough in time series analysis came with Box and Jenkins (1970) who reinvented and popularized *differencing* of the original data y_t^*:

$$y_t = \Delta^d y_t^*, \quad \text{where } \Delta^d = (1-L)^d, \quad d > 0 \text{ integer}, t \in \mathbb{N}$$

as a way to achieve stationarity. They proposed the ARMA(p,q) model:

$$y_t = \alpha_0 + \sum_{k=1}^{p}\alpha_k y_{t-k} + \sum_{k=1}^{q}\gamma_k\varepsilon_{t-k} + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{NIID}(0,\sigma^2),\ t \in \mathbb{N}$$

together with a modeling strategy viewed as an iterative process with several stages.

Stage 1: *Identification*: selecting particular values for (p, d, q) using graphical techniques, such as the correlogram and the sample partial autocorrelations.
Stage 2: *Estimation*: estimate the ARIMA model after the choice (p, d, q) is made.
Stage 3: *Diagnostic checking*: validate the error term assumptions using the residuals.
Stage 4: *Forecasting*: use the estimated model to forecast beyond n.

A remarkable feature of their diagnostic checking was to purposefully choose a more *general specification* than the one suggested by the identification stage in order to "put the model in jeopardy" (Box and Jenkins, 1970, p. 286).

Weaknesses of Box–Jenking modeling. Initially, the Box–Jenkins (1970) approach to time series modeling was considered a major success, but gradually several weaknesses emerged.

(i) Model selection within a prespecified family. The ARIMA(p,d,q) family essentially assumes that $\{y_t = \Delta^d y_t^*,\ t \in \mathbb{N}\}$ is Normal, Markov, and Stationary.
(ii) M-S Testing. Simple diagnostic checking is not effective enough to establish the validity of the model assumptions.

(iii) Respecification. There is very limited scope in respecifying an ARIMA model when found wanting, because it has built-in linearity and homoskedasticity.

(iv) Differencing. $\Delta^d y_t$ is not a universal method to achieve stationarity for a process $\{y_t^*,\ t \in \mathbb{N}\}$. It is appropriate when the AR(p) model for y_t^* has d unit roots:

$$y_t^* = \alpha_0 + \sum_{k=1}^{p} \alpha_k y_{t-k}^* + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{NIID}(0, \sigma^2),\ t \in \mathbb{N}$$

2.8 *Attempts to Frame/Systematize M-S Testing*

Several attempts have been made in the statistics literature to extend M-S testing for the LR model beyond autocorrelation, to include departures from Normality, homoskedasticity, and linearity using the residuals; see Anscombe (1961) and Anscombe and Tukey (1963). Building on that, Ramsey (1969, p. 350) proposed a number of M-S tests for omitted variables, incorrect functional form, simultaneity problems, and heteroskedasticity. The paper was influential in generating further interest in the econometric literature in formalizing M-S testing using general testing procedures such as the likelihood ratio, Lagrange multiplier, and Wald-type tests. These attempts included important contributions by Breusch and Pagan (1979, 1980), White (1980, 1982), Domowitz and White (1982), Newey (1985), and Tauchen (1985) inter alia; see Godfrey (1988) for a comprehensive overview.

An early attempt to implement M-S testing was initiated by the LSE tradition founded in the mid-1960s by Denis Sargan and Jim Durbin; see Hendry (2003) and Phillips (1988). The LSE perspective put emphasis on testing model assumptions and respecifying when the model is found to be misspecified; see Hendry (1980). A strategy for M-S testing was initially framed by Mizon (1977) and applied more broadly by other members of the LSE tradition, especially Hendry and his coauthors; see Davidson *et al.* (1978).

3. Statistical Adequacy and Its Role in Inference

Extending Fisher's notion of a statistical model to accommodate nonrandom samples takes the generic form:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}) = \{f(\mathbf{z};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}, \quad \mathbf{z} \in \mathbb{R}_Z^n \tag{17}$$

where $f(\mathbf{z};\boldsymbol{\theta}), \mathbf{z} \in \mathbb{R}_Z^n$ denotes the (joint) distribution of the sample $\mathbf{Z}:=(Z_1, \ldots, Z_n)$, and $\Theta, \mathbb{R}_Z^n$ denoting the parameter and sample spaces, respectively. His question "of what population is this a random sample" is extended to "what statistical model would render the observed data $\mathbf{z}_0:=(z_1, z_2, \ldots, z_n)$ a typical realization thereof." The "typicality" of $\mathbf{z}_0$ can – and should – be assessed using trenchant M-S testing that pertains to appraising the adequacy of the probabilistic assumptions comprising the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ vis-a-vis data $\mathbf{z}_0$; see Spanos (1999).

In the context of statistical modeling and inference, M-S testing refers to the formal testing procedures used to evaluate the validity of the prespecified statistical model (17). The statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, comprises the totality of the probabilistic assumptions imposed (directly or indirectly) on the data in question $\mathbf{z}_0:=(z_1, z_2, \ldots, z_n)$ when one estimates a model using statistical techniques. In parametric inference, one can derive the sampling distribution (cdf) of any *statistic* $Y_n=g(\mathbf{Z})$ via:

$$\mathbb{P}(Y_n \leq y) = F(y) = \int \int \cdots \int_{\{g(\mathbf{z}) \leq y\}} f(\mathbf{z};\boldsymbol{\theta}) d\mathbf{z} \tag{18}$$

A statistically misspecified $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ would vitiate any procedure relying on $f(\mathbf{z};\boldsymbol{\theta})$ (or the likelihood $L(\boldsymbol{\theta};\mathbf{z}_0)$), and render all inductive inferences unreliable. This includes Bayesian inference since the

Table 3. The Simple Normal Model.

Statistical GM:	$X_t = \mu + u_t, t \in \mathbb{N}$,
[1] Normal:	$X_t \sim \mathsf{N}(.,.)$,
[2] Constant mean:	$E(X_t) = \mu$, for all $t \in \mathbb{N}$,
[3] Constant variance:	$Var(X_t) = \sigma^2$ (known),
[4] Independence:	$\{X_t, t \in \mathbb{N}\}$ - independent process.

posterior is $\pi(\boldsymbol{\theta}|\mathbf{z}_0) \propto \pi(\boldsymbol{\theta}) \cdot L(\boldsymbol{\theta}; \mathbf{z}_0)$. It also affects the reliability of nonparametric inference since it invokes dependence and heterogeneity assumptions, as well as "indirect" distributional assumptions.

Statistical inference is often viewed as the quintessential form of *inductive inference*: learning form a particular set of data $\mathbf{z}_0$ about the stochastic phenomenon that gave rise to the data. However, it is often insufficiently recognized that this inductive procedure is embedded in a *deductive argument*:

$$\text{if } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}), \quad \text{then} \quad \mathbb{Q}(\mathbf{z})$$

where $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ specifies the premise of inference and $\mathbb{Q}(\mathbf{z})$ denotes the deduced inference propositions (optimal properties of estimators, tests, and predictors). The inbuilt deduction transmits the validity of the premises to the reliability of inference. In deductive logic, one is not particularly concerned about the soundness of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, but in statistical induction, the soundness (validity) of the premises vis-a-vis data $\mathbf{z}_0$ is of paramount importance, which was clearly recognized by Fisher (1922); see Section 2.2. Indeed, the ampliative dimension (going beyond the premises) of statistical induction relies on statistical adequacy to render the specific information in the form of data $\mathbf{z}_0$ pertinent to the stochastic phenomenon of interest; it is the cornerstone of inductive reasoning. When any of the probabilistic assumptions defining $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is invalid for data $\mathbf{z}_0$, $\mathbb{Q}(\mathbf{z})$ is called into question because the sampling distributions of the statistics used as a basis for inference will be different from those derived via (18). In particular, the *nominal* error probabilities are likely to be very different from the *actual* ones, rendering the inference results unreliable. Large discrepancies can easily arise in practice even in cases of "minor" departures from the model assumptions; see Spanos and McGuirk (2001).

How can one guard against such discrepancies? Not by invoking weak but nontestable probabilistic assumptions or vague robustness results. The practitioner needs to apply thorough M-S testing. $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is said to be *statistically adequate* when all its probabilistic assumptions are valid for data $\mathbf{z}_0$. Statistical adequacy will ensure the reliability of inference.

3.1 *Misspecification and the Unreliability of Inference*

As a prelude to M-S testing, it is worth illustrating how particular departures from the model assumptions can affect the reliability of inference by inducing discrepancies between the nominal and actual error probabilities. The example below reinforces Fisher's discerning reply to Gosset that compared to departures from IID, non-Normality "is one of the least important."

3.2 *Simple Normal Model and Misspecification*

Consider the simple Normal model in Table 3. For testing the *hypotheses*:

$$H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0 \tag{19}$$

Table 4. Type I Error of T_α When $\text{Corr}(X_i, X_j) = \rho$

ρ	0.0	0.05	0.1	0.2	0.3	0.5	0.75	0.8	0.9
α^*	0.05	0.249	0.309	0.359	0.383	0.408	0.425	0.427	0.431

there is an α-level UMP defined by: T_α:=$\{d(\mathbf{X}), C_1(\alpha)\}$ (Lehmann, 1986):

$$d(\mathbf{X}) = \tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma},\ C_1(\alpha) = \{\mathbf{x}\text{: } d(\mathbf{x}) > c_\alpha\} \tag{20}$$

What is often insufficiently emphasized is that the associated nominal error probabilities (I and II) are likely to be different from the actual ones when any of the assumptions [1]–[4] are invalid for data $\mathbf{x}_0$.

To illustrate that, consider the case where assumption [4] is false, and instead:

$$\text{Corr}(X_i, X_j) = \rho, 0 < \rho < 1, \text{ for all } i \neq j,\ i, j = 1, \ldots n \tag{21}$$

How does such a misspecification affect the reliability of test T_α? Starting with the sampling distribution: $\overline{X}_n \sim \mathsf{N}(\mu, \frac{\sigma^2}{n})$, the Normality and unbiasedness of $\overline{X}_n$ will *not* be affected, but its variance needs to include the covariances:

$$\text{Var}(\overline{X}_n) = \tfrac{1}{n^2}\left(\sum_{i=1}^{n}\text{Var}(X_i) + 2\sum_{j=i+1}^{n}\text{Cov}(X_i, X_j)\right) = \frac{\sigma^2 d_n(\rho)}{n} > \frac{\sigma^2}{n}$$

since $d_n(\rho) = (1+(n-1)\rho) > 1,\ 0 < \rho < 1,\ n > 1$. This suggests that the *actual* sampling distribution of $\overline{X}_n$ (assuming (21)) is:

$$\overline{X}_n \sim \mathsf{N}\left(\mu, \frac{d_n(\rho)\sigma^2}{n}\right) \tag{22}$$

As a result of (22), the *actual* distribution of $d(\mathbf{X})$ under H_0 is:

$$d^*(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma\sqrt{d_n(\rho)}} \overset{\mu_0}{\sim} \mathsf{N}(0, 1) \tag{23}$$

Hence, the *actual* type I error probability will be higher than the *nominal*. Let $\alpha = 0.05$ ($c_\alpha = 1.645$), $\sigma = 1$, and $n = 100$. To find the *actual type I error probability*, we need to evaluate the tail area of the distribution of $d^*(\mathbf{X})$ beyond $c_\alpha = 1.645$. In view of (23), we can deduce that: $\alpha^* = \mathbb{P}(d(\mathbf{X}) > c_\alpha; H_0) = \mathbb{P}(Z > \frac{c_\alpha}{\sqrt{d_n(\rho)}}; \mu = \mu_0)$, for $Z \sim \mathsf{N}(0, 1)$.

The results in Table 4 indicate that test T_α has now become "unreliable" since its nominal type I error probability (α) is different from the actual one (α^*).

In view of (22), the actual *type II error probability* is based on:

$$d^*(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma\sqrt{d_n(\rho)}} \overset{\mu=\mu_1}{\sim} \mathsf{N}\left(\frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma\sqrt{d_n(\rho)}}, 1\right) \tag{24}$$

Hence, the *actual power* shown in Table 5 is evaluated via:

$$\pi^*(\mu_1) = \mathbb{P}(d(\mathbf{X}) > c_\alpha; H_1(\mu_1)) = \mathbb{P}(Z > (1/\sqrt{d_n(\rho)}))\left[c_\alpha - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}\right]; \mu = \mu_1)$$

The main conclusion from Table 5 is that (i) for small values of μ_1 (0.01, 0.02, 0.05.1), the power *increases*, but for larger values of μ_1 (0.2, 0.3, 0.4), the power *decreases* as $\rho \to 1$. This undermines the "probativeness" of a test, rendering it like a *defective* smoke alarm that has the tendency to go off when burning toast, but it will not be triggered by real smoke until the house is fully ablaze; see Mayo (1996).

Table 5. Power $\pi^*(\mu_1)$ **of** T_α When $\text{Corr}(X_i, X_j) = \rho$

ρ	$\pi^*(0.01)$	$\pi^*(0.02)$	$\pi^*(0.05)$	$\pi^*(0.1)$	$\pi^*(0.2)$	$\pi^*(0.3)$	$\pi^*(0.4)$
0.0	0.061	0.074	0.121	0.258	0.637	0.911	0.991
0.05	0.262	0.276	0.318	0.395	0.557	0.710	0.832
0.1	0.319	0.330	0.364	0.422	0.542	0.659	0.762
0.3	0.390	0.397	0.418	0.453	0.525	0.596	0.664
0.5	0.414	0.419	0.436	0.464	0.520	0.575	0.630
0.8	0.431	0.436	0.449	0.471	0.515	0.560	0.603
0.9	0.435	0.439	0.452	0.473	0.514	0.556	0.598

4. Traditional Perspective on Misspecification

Statistical adequacy is crucially important for inference because "no trustworthy evidence for or against a substantive theory (or claim) can be secured on the basis of a statistically misspecified model." In light of that, "why are applied econometricians so reluctant to address the problem of statistical misspecification?"

4.1 *On the Reluctance to Test Model Assumptions*

The conventional wisdom for the role of M-S testing in econometrics is aptly summed up by Hansen's (1999, p. 195) advice in lieu of validating the LR model:

> "... omit the tests of normality and conditional heteroskedasticity, and replace all conventional standard errors and covariance matrices with heteroskedasticity-robust versions."

Hence, it is not surprising that very few applied papers in econometric journals provide sufficient evidence for the statistical adequacy of their estimated models. There are several reasons for this neglect, including the following.

1. The applied econometrics literature appears to seriously underestimate the potentially devastating effects of statistical misspecification on the reliability of inference. This misplaced confidence in the reliability of inference stems from a number of different questionable arguments and claims often used in the traditional literature.
   (i) The first is based on invoking *generic robustness results* whose generality and applicability is often greatly overvalued.
   (i) The second is that *asymptotic sampling distributions* render one's inferences less vulnerable to statistical misspecification.
   (iii) The third is that using *weaker* probabilistic assumptions would render an inference less vulnerable to statistical misspecification.
   (iv) The fourth claims that M-S testing is futile because as the sample size (n) increases, the p-values decrease toward zero, rendering every model misspecified.

2. In econometrics, the statistical premises are misleadingly blended with the substantive premises of inference. For instance, in the case of LR model, the assumption that no relevant (irrelevant) explanatory variables have been excluded (included) is considered an integral part of the statistical premises. As a result of confusing the two premises, the "omitted variables" problem, a substantive misspecification, has been the prime specification error in textbook econometrics; see Greene (2011).

3. Practitioners rarely have a complete list of probabilistic assumptions defining statistical models. Even in cases where some of the probabilistic assumptions are made explicit, e.g., the LR model,

the list is often incomplete and usually specified in terms of the unobservable error term. This undermines the effectiveness of any form of M-S testing rendering it ad hoc and partial at best.

4. M-S testing is often confused with N-P testing primarily because the same test procedures, such as likelihood ratio, Lagrange multiplier, and Wald, are employed for both types of testing. This has led to a number of misleading claims and charges against M-S testing such as calling into question the legitimacy and value of the latter, including "vulnerability to multiple testing," "illegitimate double use of data," "pretest bias," "infinite regress," etc..

4.2 *Evaluating Claims Discouraging Model Validation*

4.2.1 *Revisiting Generic Robustness Claims*

Box (1953) defined robustness to refer to the sensitivity of inference procedures (estimators, tests, and predictors) to departures from the model assumptions. According to Box, a procedure is said to be robust against certain departure(s) from the model assumptions when the inference is not *very sensitive* to the presence of *modest departures* from the premises; some assumptions "do not hold, to a greater or lesser extent." Since the premises of inference are never exactly "true," it seems only reasonable that one should evaluate the sensitivity of the inference method to modest departures.

When reading the above passage, one is struck by the vagueness of the various qualifications concerning "modest departures," "degrees of insensitivity," and assumptions holding "to a greater or lesser extent." The problem is that establishing the degree of "insensitivity" that renders the reliability of an inference procedure "tolerable" under specific departures from the model assumptions is a very difficult task. As argued above, a natural way one can render those claims less vague in the case of hypothesis testing is to evaluate the difference between the nominal and actual error probabilities under different departures from model assumptions.

What is often insufficiently appreciated in practice is that departures from model assumptions can take an infinite number of forms, but there are no robustness results for generic departures, such as:

$$\text{Corr}(X_i, X_j) \neq 0, \text{ for all } i \neq j, \ i, j = 1, \ldots n. \tag{25}$$

This is because one cannot evaluate the discrepancy between nominal and actual error probabilities under generic departures such as (25). Worse, the discrepancy between the actual and nominal error probabilities of the *t*-test based on (6) will be very different, depending, not only on the particular form of dependence, say:

$$\begin{aligned} &\text{(i) Exchangeable :} && \text{Corr}(X_i, X_j) = \rho, 0 < \rho < 1, \text{ for all } i \neq j, \ i, j = 1, \ldots n, \\ &\text{(ii) Markov :} && \text{Corr}(X_i, X_j) = \rho^{|i-j|}, \ -1 < \rho < 1, \ i \neq j, \ i, j = 1, \ldots n, \end{aligned}$$

but also on the magnitude of ρ. This implies that before one can provide a reasonable evaluation of this discrepancy, one needs to establish the appropriateness of the specific departures for the particular data to demonstrate their potential relevance. The problem is that if one needs to establish the particular form of dependence appropriate for one's data, it becomes pointless to return to the original (misspecified) model if one were able to reach a (statistically adequate) model after respecification.

In addition, certain arguments for using "robust" inference procedures are akin to being encouraged to buy insurance for *unforeseen hazards*. An example of that can be found in the statistics literature relating to Wilcoxon-type tests vs. the *t*-test in the context of simple statistical models (IID) under various departures from Normality; see Hettmansperger (1984). What is often insufficiently appreciated by this literature is that the various comparisons of "asymptotic efficiency" between robust tests and the *t*-test under several non-Normal distributions are often dubious in both value and substance. Unlike natural hazards, one has additional information in the form of the data that can be used to evaluate potential

departures from these assumptions. In the case of non-Normality, a simple t-plot of IID data realizations can easily help one to distinguish between different families of distributions. In cases where the data exhibit t-heterogeneity and/or t-dependence, one could subtract such features using auxiliary regressions based on trends and lags to be able to assess the underlying the distributional assumption. Of course, such initial choices are subject to validation using formal M-S testing. The same argument can be extended to detecting departures from IID when the data exhibit trends and cycles; see Spanos (1999, ch. 5).

4.2.2 *Weak vs. Strong Probabilistic Assumptions*

The traditional econometric textbook perspective appears to promote the inferential strategy of keeping the probabilistic assumptions as weak as possible and relying on asymptotic theory. This is often justified on the basis of an instinctive impression that "weaker assumptions are less vulnerable to misspecification" and the latter becomes less pernicious as $n\to\infty$. That is, the traditional perspective encourages practitioners to adopt a combination of generic robustness and asymptotic procedures. For instance, the most widely invoked robustness claim relates to dealing with departures from the homoskedasticity and no-autocorrelation assumptions (see Table 1), by using OLS estimators in conjunction with robust standard errors (SEs), known as heteroskedasticity/autocorrelation consistent (HAC) SE; see Hansen (1999) and Greene (2011). It turns out, however, that the use of such robust SEs often gives rise to unreliable inferences because the relevant actual error (types I and II) probabilities are very different from the nominal ones for a given sample size n and the discrepancy does not improve as n increases; see Spanos and McGuirk (2001).

The counterargument is that weak assumptions can be as fallible as strong ones. What guards an inference from misspecification is the testability of the assumptions. As argued by Fisher (1922):

> "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).

Weaker probabilistic assumptions, such as replacing Normality with the existence of the first two moments, usually imply broader but partly nontestable inductive premises, forsaking the possibility to secure the reliability of inference. The claim that broader inductive premises are less vulnerable to misspecification is analogous to the claim that a wider fishing net will always produce a larger catch. This, however, ignores the fact that the size of the catch also depends on the pertinency of the casting location. In the case of a fully parametric model with testable assumptions, M-S testing can guide the modeler toward a more appropriate "location"; this adeptness is missing from a weak but nontestable set of probabilistic assumptions.

Example. This very point is exemplified in Bahadur and Savage (1956) who replaced the Normality assumption in the simple Normal model (Table 1) with a broader family of distributions F defined by the existence of its first two moments:

$$\mathcal{M}_F(\mathbf{x}): X_k \sim \text{IID}(\mu, \sigma^2),\ x_k \in \mathbb{R},\ k = 1, 2, \ldots, n \tag{26}$$

Then, they posed the question: whether there is a reasonably reliable test within the family F for testing the hypotheses, H_0: $\mu = 0$, vs. H_1: $\mu \neq 0$, analogous to the t-test; see Lehmann (1986). The surprising answer is that no such test exists. Any t-type test based on F will be *biased* and *inconsistent*:

> "It is shown that there is neither an effective test of the hypothesis that $\mu = 0$, nor an effective confidence interval for μ , nor an effective point estimate of μ . These conclusions concerning μ flow from the fact that μ is sensitive to the tails of the population distribution." (Bahadur and Savage, 1956, p. 1115)

Table 6. AR(1) Model (Phillips, 1987).

	$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t, \quad t \in \mathbb{N} := (1, , 2, \ldots)$
(i)	$E(u_t) = 0$, for all $t \in \mathbb{N}$,
(ii)	$\sup_t E\|u_t\|^{\delta+\varepsilon} < 0$ for $\delta > 2, \varepsilon > 0$,
(iii)	$\lim_{n\to\infty} E(\frac{1}{n}\left(\sum_{t=1}^{n} u_t\right)^2 = \sigma_\infty^2 > 0$,
(iv)	$\{u_t, \ t \in \mathbb{N}\}$ is *strongly mixing* with mixing coefficient $\alpha_m \underset{m\to\infty}{\to} 0$ such that $\sum_{m=1}^{\infty} \alpha_m^{1-\delta/2} < \infty$.

Table 7. Normal, AutoRegressive (AR(1)) Model

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

$$\left.\begin{array}{lll} [1] & \text{Normality:} & (y_t, y_{t-1}) \backsim \mathsf{N}(.,.), \\ [2] & \text{Linearity:} & E\left(y_t|\sigma(y_{t-1})\right) = \alpha_0 + \alpha_1 y_{t-1}, \\ [3] & \text{Homoskedasticity:} & \text{Var}\left(y_t|\sigma(y_{t-1})\right) = \sigma_0^2, \\ [4] & \text{Markov:} & \{y_t, \ t\in\mathbb{N}\} \text{ is a Markov process}, \\ [5] & \text{t-invariance:} & \left(\alpha_0, \alpha_1, \sigma_0^2\right) \text{ are } \textit{not} \text{ changing with } t, \end{array}\right\} t\in\mathbb{N}.$$

$$\alpha_0 = E(y_t) - \alpha_1 E(y_{t-1}) \in \mathbb{R}, \ \alpha_1 = \frac{\text{Cov}\,(y_t, y_{t-1})}{\text{Var}\,(y_{t-1})} \in (-1, 1), \ \sigma_0^2 = \text{Var}(y_t) - \frac{\text{Cov}\,(y_t, y_{t-1})^2}{\text{Var}\,(y_{t-1})} \in \mathbb{R}_+$$

Note: that $\sigma(y_{t-1})$ denotes the sigma-field generated by y_{t-1}.

The intuitive reason for this result is that the family of distributions F defined by the existence of its first two (or even all) moments is too broad to enable one to tame the tails sufficiently to be able to evaluate types I and II error probabilities.

Example. A typical example of a traditional weak set of probabilistic assumptions for the AR(1) model that provides the basis for unit root testing is given in a path breaking paper by Phillips (1987); see Table 6. How would a practitioner decide that the probabilistic assumptions (i)–(iv) are appropriate for a data set $\mathbf{y}_0$? In practice, modelers will take such assumptions at face value since they are not testable.

In contrast, Table 7 gives a complete set of testable probabilistic assumptions for the same model. Andreou and Spanos (2003) used the assumptions [1]–[5] to illustrate how statistical misspecifications undermined the inference results of numerous published papers in the case of testing "trend versus difference stationarity."

Another widely used, but highly misleading, claim is that misspecification becomes less pernicious as n increases, and thus the asymptotic sampling distributions are more "robust" than the finite sample ones. First, certain forms of misspecification, such as the presence of heterogeneity, become more and more pernicious as n increases. Second, limit theorems invoked by consistent and asymptotically normal (CAN) estimators and associated tests also rely on probabilistic assumptions, such as (i)–(iv) above, which are usually nontestable, rendering the reliability of the resulting inferences dubious. Finally, the truth of the matter is that all inference results will rely exclusively on the n available data points $\mathbf{x}_0$ and nothing more. 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$n$."

Asymptotic theory based on "$n \to \infty$" relates to the "capacity" of inference procedures to pinpoint μ^*, the "true" μ, as data information accrues $\{x_k\}_{k=1}^{\infty} := (x_1, x_2, \ldots, x_n, \ldots)$ approaching the limit at ∞. In

that sense, asymptotic properties are useful for their value in excluding potentially unreliable estimators and tests, but they do not guarantee the reliability of inference procedures for a given data $\mathbf{x}_0$. For instance, an inconsistent estimator is likely to give rise to unreliable inference, but a consistent one does not guarantee the trustworthiness of the inference results. Such asymptotic results tell us nothing about the trustworthiness of the inference results based on data $\mathbf{x}_0 := (x_1, \ldots, x_n)$. The latter is inextricably bound up with the particular $\mathbf{x}_0$ and n.

One can just imagine an econometrician asking (tongue-in-cheek) why Bahadur and Savage (1956) did not use:

$$\tau(\mathbf{X}) \overset{\mu=\mu_0}{\underset{n\to\infty}{\sim}} \mathsf{N}(0, 1) \tag{27}$$

oblivious to the fact that this conjuring is tantamount to imposing approximate Normality. To be more specific, when the log-likelihood can be approximated by a quadratic function of $\boldsymbol{\theta} := (\mu, \sigma^2)$ (like the Normal), then (27) is likely to be accurate enough; see Geyer (2013). Believing that the validity of (27) stems from invoking the heuristic "as $n \to \infty$" is just an illusion. The trustworthiness of any inference results invoking (27) stems solely from the approximate validity of the probabilistic assumptions imposed on $\mathbf{x}_0$ for the specific n.

The above comments are particularly relevant when comparing *parametric* with *nonparametric models*. The only real difference between the two is that a parametric model is based on a direct and testable distributional assumption, but the nonparametric relies instead on *indirect* distributional assumptions that are often *nontestable*. Such indirect distributional assumptions include: (i) the existence of certain moments up to order p, as well as (ii) smoothness restrictions on the unknown density function $f(z)$, $z \in \mathbb{R}_Z$ (continuity, symmetry, differentiability, unimodality, boundedness, and continuity of derivatives of $f(z)$ up to order $m > 1$); see Wasserman (2006). It is important to emphasize that both types of models impose direct dependence and heterogeneity probabilistic assumptions. How do nonparametric models address the nonexistence of optimal inference procedures raised by the Bahadur–Savage (1956) result? By imposing sufficient smoothness restrictions on $f(z)$, presented as harmless mathematical restrictions imposed for convenience. As argued next, imposing nontestable assumptions on the data is the surest way to undermine the reliability of inductive inference.

At a more subtle level, the conventional wisdom favoring weaker assumptions is burdened with a fundamental confusion between *mathematical deduction* and *statistical induction*. In mathematical deduction, there is a premium for results based on the weakest (minimal) set of assumptions comprising the deductive premises. A deductive argument is logically "valid" if it is impossible for its premises to be true, while its conclusion is false. The logical validity of a deductive argument does not depend on the "soundness" of the premises. In statistical induction, the empirical validity (soundess) of the premises is of paramount importance because the deductive argument (the inference procedure) will preserve that validity to secure the reliability of the inference procedures derived on the basis of the premises. Hence, for statistical induction, the premium is attached to the strongest (maximal) set of *testable* assumptions comprising the inductive premises. Such a maximal set, when validated vis-a-vis the data, provides the most effective way to learn from data because the inference procedures stemming from it are both reliable and optimal (precise). In this sense, the weakest link when using CAN estimators and related inference procedures in practice is the conjuration of limit theorems that rely on mathematically convenient but empirically nontestable assumptions. Despite their value as purely deductive propositions, such limit theorems are insidious for modeling purposes when their assumptions turn out to be invalid for one's data.

In several papers in the 1940s and 1950s, Fisher railed against "mathematicians" who view statistics as a purely deductive field, by ignoring its inductive roots stemming from the link between the probabilistic assumptions and the actual data; see Box (1978, pp. 435–438). He was also highly critical of valuating technical dexterity and rigorous derivations at the expense of active thinking. As stated by Mahalanobis (1938):

"The explicit statement of a rigorous argument interested him [Fisher], but only on the important condition that such explicit demonstration of rigor was needed. Mechanical drill in the technique of rigorous statement was abhorrent to him, partly for its pedantry, and partly as an inhibition to the active use of the mind. He felt it was more important to think actively, even at the expense of occasional errors from which an alert intelligence would soon recover, than to proceed with perfect safety at a snail's pace along well-known paths with the aid of the most perfectly designed mechanical crutches."

A case can be made that current textbook econometrics is often taught as a purely deductive field with the emphasis placed on rigorous arguments grounded on "perfectly designed mechanical crutches," instead of active thinking. Their empirical illustrations are of questionable value as exemplars of empirical modeling that gives rise to learning from data about phenomena of interest because they are often based on nonvalidated inductive premises.

4.2.3 *A Large Enough Sample Size (n) and Misspecification*

It is often argued that with a large enough sample size, all statistical models will be found misspecified. This claim would have implied that a small enough n would render all models statistically adequate, which is a peculiar conclusion. It is certainly true that as the sample size n increases, the p-value decreases toward zero, and thus the presence of any discrepancy from the null (H_0), however tiny, will lead to rejecting H_0 with a large enough n. There is nothing paradoxical about this result since this happens for all *consistent* tests; their power goes to one for any discrepancy from the null, however, small. The above claim is an instance of a classic fallacy (Mayo and Spanos, 2006):

1. The fallacy of rejection: Evidence *against* H_0 is misinterpreted as evidence *for* a particular alternative H_1. This fallacy arises when the test in question has high power (e.g., large n) to detect substantively minor discrepancies. That is, a small p-value or a rejection of H_0 does not automatically provide evidence that H_1 is valid.

Example. A quintessential example of this fallacy is the case where the D-W test, based on the A-C LR model in Table 2, rejects H_0 and textbook econometrics advises practitioners to "fix" the problem by replacing the OLS with the GLS estimator, adopting the A-C LR model. A rejection of H_0 provides evidence *against* H_0 and *for* the presence of generic temporal dependence:

$$E(\varepsilon_t \varepsilon_s | \mathbf{X}_t = \mathbf{x}_t) \neq 0 \tag{28}$$

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

$$H_1 : E(\varepsilon_t \varepsilon_s | \mathbf{X}_t = \mathbf{x}_t) = \left(\frac{\rho^{|t-s|}}{1-\rho^2}\right)\sigma_u^2, \; t > s, \quad t, s = 1, 2, \ldots, n \tag{29}$$

How one can secure evidence for (29) is discussed in Section 5.3.

2. The fallacy of acceptance: *no* evidence against H_0 is misinterpreted as evidence *for* H_0. This fallacy can easily arise in cases where the test in question has low power (e.g., small n) to detect discrepancies of interest.

The fallacious claims stem from ignoring the power of the test: its capacity to detect different discrepancies from H_0. An oversensitive test (e.g., very large n) is likely to pick up minor discrepancies from H_0, and since the power of a consistent test increases with n, one needs to take that into account. Analogously, an undersensitive test (e.g., small n) is likely to leave sizeable discrepancies undetected. In this sense, it is not true that a large enough n would render all models misspecified. What it implies is that the significance level α should be adjusted as n changes to avoid situations where the power for a particular discrepancy is either very low of very high. Indeed, practitioners in statistics are advised

to decrease the significance level as n increases using certain rules of thumb; see Lehmann (1986). For instance, Good (1988) suggested standardizing the p-value $p(\mathbf{x}_0)$ to a fixed sample size $n = 100$ using the rule of thumb: $p_{100}(\mathbf{x}_0) = \min(0.5, [p(\mathbf{x}_0)\cdot\sqrt{n/100}]),\ n > 10$.

For example, $p(\mathbf{x}_0) = 0.04$ for $n = 1000$ corresponds to $p_{100}(\mathbf{x}_0) = 0.126$.

A more formal way to addresses both of the above fallacies is to use the postdata severity evaluation that determines the warranted discrepancy from H_0. This takes into consideration the power of the test in going from the p-value or accept/reject results to an evidential account for a particular data $\mathbf{z}_0$; see Mayo and Spanos (2006).

4.3 *Statistical vs. Substantive Premises of Inference*

What is often insufficiently appreciated in econometrics is that behind every structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, there is always a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ that pertains solely to the probabilistic assumptions imposed (often implicitly on the data $\mathbf{Z}_0$. The crucial reason for delineating the two premises is that one needs to secure statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ first before one can reliable probe substantive adequacy of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$. In a sense, statistical adequacy is the price a modeler needs to pay for using reliable statistical procedures to pose the substantive questions of interest. Adding a variable to an LR model and using the significance of its coefficient as evidence for its relevance can lead to spurious inference results when the original regression is statistically misspecified; one has no reason to trust the t-statistic, unless the probabilistic assumptions invoked by the t-test have been validated vis-a-vis data $\mathbf{Z}_0$.

One way to untangle the two premises is to ground the statistical and structural models on different sources of information, the chance regularities and the theoretical information, respectively. The structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ stems exclusively from substantive (theory) information. Its only connection to the data is that the original theory model, which might include latent variables, needs to be modified to render it estimable with the particular data $\mathbf{Z}_0$. The statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ stems solely from the statistical information in the data (chance regularity patterns), and that takes the form of a probabilistic structure assigned to the stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{N}\}$ underlying data $\mathbf{Z}_0$ with a view to account for its chance regularities. The choice of the statistical model is made with a twofold objective in mind:

(i) to account for the chance regularities in data $\mathbf{Z}_0$ by choosing a probabilistic structure for the stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{N}\}$ so as to render $\mathbf{Z}_0$ a "truly typical realization" thereof, and
(ii) to parameterize $\{\mathbf{Z}_t,\ t \in \mathbb{N}\}$ in an attempt to specify $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ so as to embed (parametrically) $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ in its context via certain restrictions, say $\mathbf{G}(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \mathbf{0},\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\varphi} \in \Phi$, relating the statistical and structural parameters.

Although this parametric nesting of the structural $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ within the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is generally applicable to all structural models of interest in econometrics, it has been used in traditional econometrics in the context of the simultaneous equations model (SEM), without the emphasis on the statistical vs. structural distinction. The *structural model* in the SEM is specified using the generic formulation:

$$\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z}) : \boldsymbol{\Gamma}^{\top}(\boldsymbol{\varphi})\mathbf{y}_t + \boldsymbol{\Delta}^{\top}(\boldsymbol{\varphi})\mathbf{x}_t = \boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t \sim \mathsf{N}(\mathbf{0}, \Omega(\boldsymbol{\varphi})) \tag{30}$$

where $\mathbf{y}_t$ denotes the endogenous variables, $\mathbf{x}_t$ the exogenous variables, and $\boldsymbol{\varphi} \in \boldsymbol{\Phi}$ denotes the unknown structural parameters in the coefficient matrices $\boldsymbol{\Gamma}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\Omega}$. Corresponding to this, there is a *reduced form*:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}) : \mathbf{y}_t = \mathbf{B}^{\top}(\boldsymbol{\theta})\mathbf{x}_t + \mathbf{u}_t,\ \mathbf{u}_t \sim \mathsf{N}(\mathbf{0}, \Sigma(\boldsymbol{\theta})) \tag{31}$$

What has not been sufficiently appreciated is that the reduced form in (31) is the implicit statistical model behind (30). Instead of being viewed as a derived formulation, (31) has a life of its own as a multivariate

version of the LR model in Table 9, specified in terms of the statistical parameters: $\boldsymbol{\theta} \in \boldsymbol{\theta}$. The two sets of parameters are related via the identifying restrictions:

$$\mathbf{G}(\boldsymbol{\varphi}, \boldsymbol{\theta}) = \mathbf{0} : \mathbf{B}(\boldsymbol{\theta})\boldsymbol{\Gamma}(\boldsymbol{\varphi}) + \boldsymbol{\Delta}(\boldsymbol{\varphi}) = \mathbf{0}, \ \boldsymbol{\Omega}(\boldsymbol{\varphi}) = \boldsymbol{\Gamma}^{\top}(\boldsymbol{\varphi})\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{\Gamma}(\boldsymbol{\varphi}), \text{ for } \boldsymbol{\varphi} \in \boldsymbol{\Phi}, \boldsymbol{\theta} \in \boldsymbol{\theta} \tag{32}$$

Moreover, the statistical adequacy of the implicit statistical model in (31) underwrites the reliability of any inferences based on the estimated structural parameters $\widehat{\boldsymbol{\varphi}}$. That is, when reduced form in the form of the LR model is statistically misspecified, any inferences based on $\widehat{\boldsymbol{\varphi}}$ are likely to be unreliable; see Spanos (1990). This is because the sampling distribution (finite and asymptotic) of the estimator $\widehat{\boldsymbol{\varphi}}$, – instrumental variables (IV) or maximum likelihood (ML) – depends crucially on assumptions [1]–[5]; see Phillips (1983).

Untangling $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ from $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ delineates two different types of inadequacy:

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

[b] Substantive inadequacy: the conditions envisaged by the theory in question differ "systematically" from the *actual* data generating mechanism (GM) that gave rise to the phenomenon of interest. Substantive inadequacy arises from highly unrealistic structural models, flawed *ceteris paribus* clauses, missing confounding factors, etc. This includes the testing of the overidentifying restrictions stemming from $\mathbf{G}(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \mathbf{0}$.

The above perspective should be contrasted with the traditional textbook approach where $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ is estimated directly, without validating the (implicit) $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. The inevitable result is that the estimated $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ is misspecified, both statistically and substantively, but one has no way to delineate the two errors and apportion blame with a view to address the errors; see Spanos (2010b).

4.4 *All Models Are Wrong But Some Are Useful*

The confusion between statistical and substantive premises of inference can be used to explain why applied econometricians are particularly receptive to the catchphrase: "All models are wrong, but some are useful." This is often invoked as an alibi to ignore M-S testing because the slogan creates the erroneous impression that statistical "misspecification" is inevitable. The catchphrase, however, is only half the insight as expressed by Box (1979, p. 202). The other half pertains to viewing empirical modeling as an iterative process driven by diagnostic checking based on the residuals (p. 204):

> "How can we avoid the possibility that the parsimonious model we build by such an iteration might be misleading? There are two answers: a) Knowing the scientific context of an investigation we can allow in advance for more important contingencies. b) Suitable analysis of residuals can lead to our fixing up the model in other needed directions."

A closer look at the slogan suggests that with "all models are wrong," Box alludes to the fact that models are always approximations of the real world:

> "Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model." (p. 202)

which clearly pertains to *substantive* adequacy. In light of the fact that a structural model cannot be an exact description of the phenomenon of interest because it invariably involves abstraction, simplification, and approximation, one can make a strong case that substantive inadequacy is more or less inevitable. For instance, establishing that no potentially relevant variables have been omitted from one's model is an impossible task. Statistical adequacy, however, can be established using M-S testing.

Surely, in b) above, Box is advising practitioners to guard against misspecification.

Table 8. Stages of Empirical Modeling

1. Specification
2. Estimation
3. Mis-Specification (M-S) testing
4. Respecification

■ Statistically adequate model

5. Inference

4.5 *M-S Testing: Foundational Issues*

Despite the plethora of M-S tests in the econometric literature, no systematic way to apply these tests has emerged; see Godfrey (1988). The literature left practitioners perplexed since numerous issues about M-S testing remained unanswered. These include: (i) the choice among different M-S tests, (ii) the use of omnibus (nonparametric) vs. parametric tests, (iii) the difference between M-S tests, N-P tests, Fisher's significance tests and Akaike-type model selection procedures, (iv) what to do when any of the model assumptions are invalid, as well as (v) the potential vulnerability of M-S testing to charges such as data mining, (b) multiple testing and (c) pre-test bias., and (v) what to do when any of the model assumptions are invalid.

It is argued that these issues can be addressed by having a coherent framework where different aspects of modeling and inference can be delineated. To do that one needs to return to Fisher's early writings to pick up and formalize some of his suggestions and ideas, by extending/modifying the original framework. In particular, we need to separate the modeling from the inference facets lumped together by Fisher under "problems of distribution." The modeling facets include *M-S testing* and *Respecification* (with a view to achieve statistical adequacy) that are given in Table 8.

The above perspective suggests that when the underlying statistical model is statistically misspecified, it needs to be respecified with a view to account for the chance regularities in data $\mathbf{z}_0$, but retain the parametric nesting of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ whenever possible. It is important to emphasize that a lot of confusions in empirical modeling stem from blending M-S testing and respecification with inference proper. Examples of that include the "model averaging" (Claeskens and Hjort, 2008), the "pretest bias" problem (Saleh, 2006), and the model selection literatures. The problem is that the inference facet, e.g., N-P testing, presupposes that the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is valid for data $\mathbf{z}_0$, i.e., $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ could have generated data $\mathbf{z}_0$.

Hence, averaging two models where one is statistically adequate and the other is misspecified will give rise to a misspecified mixture of models. Model averaging makes sense only when one begins with several statistically adequate models based on different sets of explanatory variables, and the averaging is used to enhance the substantive adequacy of the aggregated model.

Similarly, the pretest bias problem arises because M-S testing is blended with the inference facet using a decision-theoretic framing in a misguided attempt that institutionalizes the fallacies of acceptance and rejection; see Spanos (2010a).

By the same token, selecting a statistical model on statistical adequacy grounds (thorough M-S testing) is very different from model selection based on Akaike-type procedures. These procedures begin with a broad family of models $\mathsf{F}_z = \{\mathcal{M}_{\boldsymbol{\varphi}_i}(\mathbf{z}),\ i = 1, \ldots m\}$, and then select a best model $\mathcal{M}_{\boldsymbol{\varphi}_k}(\mathbf{z})$ within this family by trading goodness-of-fit against parsimony; see Konishi and Kitagawa (2008). However, goodness-of-fit is neither necessary nor sufficient for statistical adequacy; when F_z is misspecified one is using the wrong log-likelihood as a goodness-of-fit measure. Akaike-type model selection procedures invariably give rise to unreliable inferences because: (i) they ignore the preliminary step of validating F_z (Lehmann, 1990), and (ii) their ranking of the different models in F_z is equivalent to pairwise N-P testing comparisons without keeping track of the relevant error probabilities; see Spanos (2010a).

5. The Probabilistic Reduction (PR) Approach

This PR perspective provides a purely probabilistic construal of a statistical model $\mathcal{M}_{\theta}(\mathbf{z})$ and places statistical modeling in a broader framework that allows the *ab initio* separation of the statistical (chance regularities) and substantive information in the form of a structural model $\mathcal{M}_{\varphi}(\mathbf{z})$, and their subsequent fusing with a view to end up with an empirical model that is both statistically adequate and substantively data-acceptable. The two types of information are viewed as complementary, and their fusing is achieved without compromising the integrity of either. This perspective can be viewed as an extension/modification of Fisher's (1922) framework aiming to separate the modeling from the inference facet, as well as shed light on several modeling problems, including (i)–(v) mentioned above.

The basic idea is that a modeler begins with the substantive subject matter information usually framed into a structural model $\mathcal{M}_{\varphi}(\mathbf{z})$. This model demarcates the crucial aspects of the phenomenon of interest by choosing the relevant variables behind data $\mathbf{Z}_0$. The traditional perspective specifies the associated statistical model indirectly via the probabilistic structure of the error process, along the lines of the SEM discussed above. By disentangling the two models, the PR perspective proposes that the specification of the statistical model $\mathcal{M}_{\theta}(\mathbf{z})$ be based directly on the probabilistic structure of the stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{N}\}$ underlying the data $\mathbf{Z}_0$, by being viewed as a particular parameterization of this process. That is, the choice of the data $\mathbf{Z}_0$ affords the statistical model a life of its own, separate from the structural model. As argued in Section 4.3, the two models are related in so far as the specification (selection) of $\mathcal{M}_{\theta}(\mathbf{z})$ is based on two interrelated aims: (i) to account for the systematic statistical information in data $\mathbf{Z}_0$, and (ii) to select a parameterization for $\mathcal{M}_{\theta}(\mathbf{z})$ that parametrically nests $\mathcal{M}_{\varphi}(\mathbf{z})$, so that one can pose the substantive questions of interest. The PR perspective on specification has several advantages, including: (i) enabling the modeler to distinguish between the statistical and the substantive premises of inference, (ii) providing a complete and internally consistent set of testable probabilistic assumptions, (iii) providing a well-defined parameterization for the model parameters, (iv) bringing out the interrelationship among the probabilistic assumptions, and (v) unifying the specification of statistical models for different types data: time series, cross section, and panel. At the practical level, the problem of specification is viewed in the context of the set of all possible models that could have generated data $\mathbf{z}_0$, say $\mathcal{P}(\mathbf{z})$, where $\mathcal{M}_{\theta}(\mathbf{z})$ is specified by partitioning $\mathcal{P}(\mathbf{z})$ using probabilistic assumptions from three broad categories: distribution, dependence, and heterogeneity.

The PR perspective on specification operationalizes Fisher's discerning reply (Section 2.4) to Gosset's (1923): "What I think is my business is the detailed examination of the data, ... to determine what information they are fit to give ...." Graphical techniques, such as simple t-plots and scatter plots, are often sufficient to detect a variety of chance regularity patterns that can be accounted for using appropriate probabilistic assumptions for $\{\mathbf{Z}_t,\ t \in \mathbb{N}\}$. The selection aims to render data $\mathbf{Z}_0$ a typical realization of $\{\mathbf{Z}_t,\ t \in \mathbb{N}\}$; see Spanos (1999, chapters 5 and 6). Formally, the appropriateness of the selected probabilistic assumptions is appraised by M-S testing.

Example 1. Consider the specification procedure for the AR(1) model (Table 7). The t-plot of the data in Figure 2 exhibits distinct irregular cycles, indicating positive dependence. Subtracting this dependence, using an auxiliary regression between y_t and y_{t-1}, yields the dememorized data in Figure 3, which exhibit bell-shape symmetry associated with a Normal distribution. No indications of departures from IID suggest that data $\mathbf{y}_0$ could be viewed as a typical realization of a Normal (N), Markov (M), Stationary (S) process $\{y_t,\ t \in \mathbb{N}\}$. The parameterization for the AR(1), specified in terms of $f(y_t|y_{t-1};\boldsymbol{\theta})$, results from a reduction of the joint distribution of the process $\{y_t, t \in \mathbb{N}\}$ using sequential conditioning in conjunction with those assumptions:

$$f(y_1,\ldots,y_n;\boldsymbol{\phi}) \overset{\text{M}}{=} f(y_1;\boldsymbol{\varphi}_1)\prod_{t=2}^{n} f_t(y_t|y_{t-1};\boldsymbol{\varphi}_t) \overset{\text{M\&S}}{=} f(y_1;\boldsymbol{\varphi}_1)\prod_{t=2}^{n} f(y_t|y_{t-1};\boldsymbol{\varphi})\ \forall \mathbf{y} \in \mathbb{R}^n \qquad (33)$$

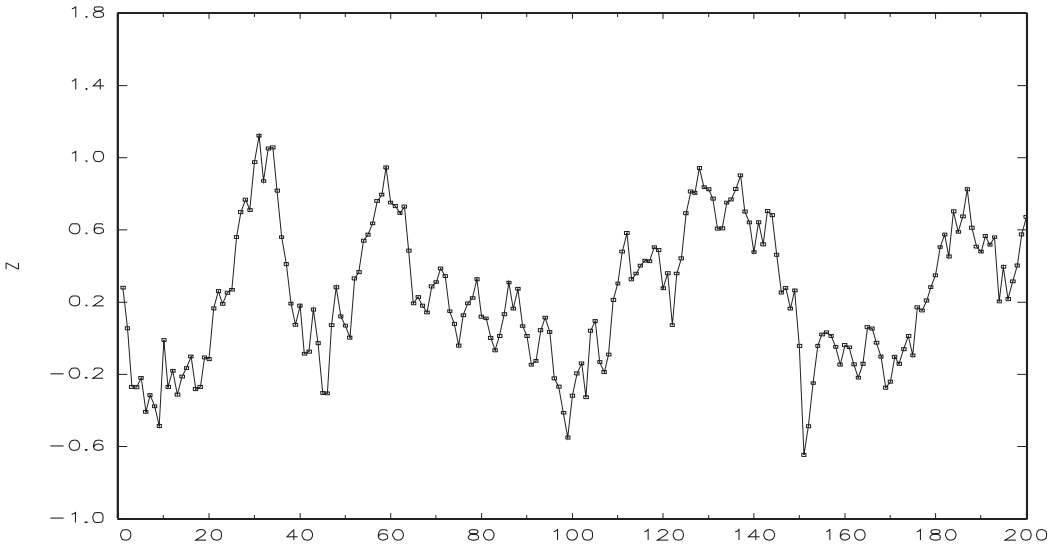

Figure 2. t-Plot of z_t.

Table 9. Linear Regression Model

Statistical GM:	$y_t=\beta_0+\beta_1^\top \mathbf{x}_t+u_t,\ t\in\mathbb{N}.$		
[1]	Normality:	$(y_t\vert\mathbf{X}_t=\mathbf{x}_t)\backsim \mathsf{N}(.,.),$	$t\in\mathbb{N}.$
[2]	Linearity:	$E(y_t\vert\mathbf{X}_t=\mathbf{x}_t)=\beta_0+\beta_1^\top \mathbf{x}_t,$	
[3]	Homosk/city:	$\text{Var}(y_t\vert\mathbf{X}_t=\mathbf{x}_t)=\sigma^2,$	
[4]	Independence:	$\{(y_t\vert\mathbf{X}_t=\mathbf{x}_t),\ t\in\mathbb{N}\}$ indep. process,	
[5]	t-invariance:	$(\beta_0,\beta_1,\sigma^2)$ are *not* changing with t,	
	$\beta_0=E(y_t)-\beta_1^\top E(\mathbf{X}_t),\ \beta_1=[\text{Cov}(\mathbf{X}_t)]^{-1}\text{Cov}(\mathbf{X}_t,y_t),$		
	$\sigma^2=\text{Var}(y_t)-\text{Cov}(\mathbf{X}_t,y_t)^\top[\text{Cov}(\mathbf{X}_t)]^{-1}\text{Cov}(\mathbf{X}_t,y_t)$		

This reduction yields the AR(1) in Table 7, with the parameterization of $(\alpha_0,\alpha_1,\sigma_0^2)$.

Example 2. The same procedure can be followed to specify the LR model assuming that $\{\mathbf{Z}_t:=(y_t,\mathbf{X}_t),\ t\in\mathbb{N}\}$ is Normal, IID, giving rise to the LR model in Table 9:

$$D(\mathbf{Z}_1,\ldots,\mathbf{Z}_n;\boldsymbol{\phi})\overset{\text{IID}}{=}\prod_{t=1}^{n}D(\mathbf{Z}_t;\boldsymbol{\varphi})=\prod_{t=1}^{n}D(y_t\vert\mathbf{X}_t;\boldsymbol{\psi}_1)\cdot D(\mathbf{X}_t;\boldsymbol{\psi}_2),\ \forall\mathbf{Z}_t\in\mathbb{R}^{m\times n} \tag{34}$$

The LR model is specified exclusively in terms of the conditional distribution $D(y_t\vert\mathbf{X}_t;\boldsymbol{\psi}_1)$, due to weak exogeneity; see Engle *et al.* (1983). To be more specific, the LR model comprises the statistical GM in conjunction with the probabilistic assumptions [1]–[5] (Table 9).

It is important to emphasize that, for statistical adequacy purposes, only the validity of assumptions [1]–[5] is relevant because they constitute the statistical premises of inference. That is, if assumptions

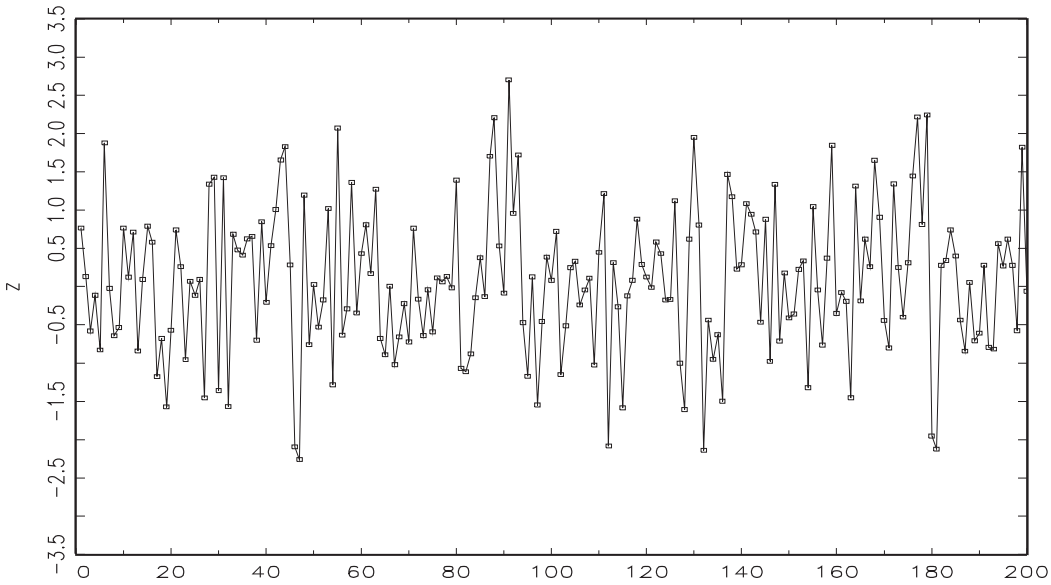

Figure 3. A t-Plot of Dememorized z_t.

[1]–[5] are valid for data $\mathbf{Z}_0 := (\mathbf{y}, \mathbf{X})$, the statistical procedures associated with the LR model have the optimality properties attributed to them; note that assumptions [1]–[4] are directly related to those in Table 1.

Of particular interest in the specification of the above statistical models is the Statistical GM, which has a dual role to play. The first is to provide a generic way that artificial data that have the same probabilistic structure as data $\mathbf{Z}_0$ can be simulated on a computer, and the second to provide the link between the statistical and substantive information. From a purely probabilistic perspective, the statistical GM stems from an orthogonal decomposition of y_t, defined on a probability space $(S, \mathcal{F}, \mathbb{P}(.))$, into a systematic $\mu_t = E(y_t|\mathcal{D}_t)$ and nonsystematic $u_t = y_t - E(y_t|\mathcal{D}_t)$ component with respect to the conditioning information set $\mathcal{D}_t \subset \mathcal{F}$, to define:

$$y_t = E(y_t|\mathcal{D}_t) + u_t, \ t \in \mathbb{N} \tag{35}$$

where $\mathcal{D}_t$ is selected with a view to capture all the systematic information in $\mathbf{Z}_0$ rendering the "educed" $u_t = y_t - E(y_t|\mathcal{D}_t)$ nonsystematic. More formally, u_t constitutes a second-order Martingale Difference (MD) process, i.e.,

$$\text{(i) } E(u_t|\mathcal{D}_t) = 0, \quad \text{(ii) } E(\mu_t u_t|\mathcal{D}_t) = 0, \quad \text{(iii) } E\left(u_t^2|\mathcal{D}_t\right) = \text{Var}(y_t|\mathcal{D}_t), \ t \in \mathbb{N} \tag{36}$$

In practice, $\mathcal{D}_t$ is commonly generated by observable random variables defined on $(S, \mathcal{F}, \mathbb{P}(.))$. As shown in White (1999, p. 59–60), $\mathcal{D}_t = \sigma(\mathbf{X}_t, \mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, \ldots, \mathbf{Z}_1)$ is the generic choice of the relevant $\mathcal{D}_t$ when no dependence and heterogeneity assumptions (restrictions) are imposed on the observable process $\{\mathbf{Z}_t := (y_t, \mathbf{X}_t), \ t \in \mathbb{N}\}$ that would render $\{(u_t|\mathcal{D}_t), \ t \in \mathbb{N}\}$ an MD process.

In the case of the LR model, the independence and Normality assumptions narrow down the appropriate choice to $\mathcal{D}_t = (\mathbf{X}_t = \mathbf{x}_t)$. In this sense, the statistical error term u_t is viewed as: [i]*derived*, denoting the nonsystematic information in $\mathbf{Z}_0$ relative to μ_t, and [ii]*local* , in the sense that it pertains to the statistical

model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ vis-a-vis the data $\mathbf{Z}_0$. That is, the appraisal of concern is how adequately $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ accounts for the chance regularities in $\mathbf{Z}_0$.

The above interpretation of the statistical error term enables one to draw a clear distinction between the potentially relevant:

1. *Statistical information* in data $\mathbf{Z}_0$, in terms of which $\mathcal{D}_t$ is defined, such as temporal dependence, generically represented with lags $\mathbf{Z}_{t-i} := (y_{t-i}, \mathbf{X}_{t-i})$, $i = 1, \ldots, p$ and heterogeneity, generically represented by dummies and trends (based on the ordering $t = 1, \ldots, n$).
2. *Substantive information* beyond the data $\mathbf{Z}_0$, such as omitted but relevant variables, errors of measurement, simultaneity problems, etc.
   This distinction indicates clearly that $\mathbf{Z}_{t-i}$ is not an "omitted" set of variables, in the sense of explanatory variables $\mathbf{W}_t$ *not* in $\mathbf{X}_t$, but a "discounted" one since $\mathbf{Z}_{t-i}$ and trend terms t, t^2 are already in the current statistical information set $(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n)$.

In contrast to the statistical, the structural error term $\boldsymbol{\varepsilon}_t$ in (30) is viewed as: [i]* *autonomous,* potentially representing errors of approximation, omitted effects, errors of measurement, shocks, etc., and [ii]* *global,* in the sense that the relevant appraisal of concern is how adequately $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ sheds light on the phenomenon of interest. This probing of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ represents a very different and a lot more open-ended level of comparison than that of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$.

The purely probabilistic construal of a statistical model, as shown in Table 9, enables ones to distinguish these from statistical misspecifications since $E(\mu_t u_t|\mathcal{D}_t) = 0$ by definition. The PR in (34) yields the relevant model parameterization $(\boldsymbol{\theta})$ in terms of the primary parameters $\boldsymbol{\varphi} := (\mu_1, \boldsymbol{\mu}_2, \sigma_{11}, \boldsymbol{\sigma}_{21}, \boldsymbol{\sigma}_{22})$ (Table 9). Hence, it is no accident that numerous substantive misspecifications, as opposed to statistical, are associated with $E(\mathbf{X}_t^\top \boldsymbol{\varepsilon}_t) \neq \mathbf{0}$.

5.1 *M-S Testing*

The primary objective of M-S testing is to probe within $\mathcal{P}(\mathbf{z})$-all possible statistical models that could have given rise to $\mathbf{Z}_0$, for potential departures from the assumptions defining $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}) = \{f(\mathbf{z};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\},\ \mathbf{z} \in \mathbb{R}_Z^n$. The generic null and alternative hypotheses take the form:

$$H_0 : f^*(\mathbf{z}) \in \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}) \text{ vs. } \overline{H}_0 : f^*(\mathbf{z}) \in [\mathcal{P}(\mathbf{z}) - \mathcal{M}_{\theta}(\mathbf{z})]$$

where $f^*(\mathbf{z})$ denotes the "true" distribution of the sample. A key feature of M-S testing is that it probes *outside* the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. In contrast, for the archetypal N-P hypotheses:

$$H_0 : \boldsymbol{\theta}_0 \in \Theta_0 \text{ vs. } H_1 : \boldsymbol{\theta}_0 \in \Theta_1 \tag{37}$$

where (Θ_1, Θ_0) constitute a partition of Θ, the probing is *within* the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, since (37) can be equivalently expressed in the form:

$$H_0 : f^*(\mathbf{x}) \in \mathcal{M}_0(\mathbf{x}) = \{f(\mathbf{x};\boldsymbol{\theta}_0),\ \boldsymbol{\theta}_0 \in \Theta_0\} \text{ vs. } H_1 : f^*(\mathbf{x}) \in \mathcal{M}_1(\mathbf{x}) = \{f(\mathbf{x};\boldsymbol{\theta}_0),\ \boldsymbol{\theta}_0 \in \Theta_1\}$$

Another crucial difference is that for M-S testing, one needs to operationalize $[\mathcal{P}(\mathbf{x}) - \mathcal{M}_{\theta}(\mathbf{x})]$ by replacing $\overline{H}_0$ with a more specific feasible alternative H_1. This renders M-S testing highly vulnerable to the fallacy of rejection.

This operationalization can take a number of different forms, including parametric and nonparametric tests, as well as directions of departure. *Nonparametric* (omnibus) tests, such as the runs test, the Pearson, and the Kolmogorov tests, are usually *nondirectional* in the sense that the alternative hypothesis is defined as the negation of the null, rendering them particularly useful for M-S testing because that implies a broader local scope. Their low local power is a blessing in M-S testing because their indicating departures from certain assumptions provides better evidence for such departures than parametric (directional) tests

with very high power. The most serious weakness of nonparametric M-S tests is that they rarely provide information about the source of departure. For that information, a practitioner needs to use parametric (directional) tests. Parametric M-S tests often take the form of encompassing $\mathcal{M}_{\theta}(\mathbf{z})$ into a broader model $\mathcal{M}_{\psi}(\mathbf{z})$ and testing the nesting restrictions, as in the case of the Durbin–Watson test. This has the appearance of an N-P test, but such an interpretation is unwarranted because there is no reason to believe that $\mathcal{M}_{\psi}(\mathbf{z})$ is statistically adequate; a crucial presupposition for N-P testing. Note also that, in contrast to N-P testing, the type II (not I) error is the most crucial for M-S testing; see Spanos (2010a).

5.1.1 *Joint M-S Testing Using Auxiliary Regressions*

A strong case can be made that the best strategy to avoid "erroneous" diagnoses, minimize the number of maintained assumptions, and enhance the scope of the tests is to use *joint M-S testing*. As shown above, the model assumptions are usually interrelated, and thus testing them individually can give rise to misleading diagnoses. As shown next, in the case of the LR model, there are natural groupings of the assumptions according to how their potential departures might change/modify the regression and skedastic functions of the original model.

The joint M-S testing based on auxiliary regressions has several distinct advantages over other procedures based on individual test statistics, such as Lagrange multiplier tests for homoskedasticity, the Durbin–Watson, and the Box–Pierce tests for no-autocorrelation, the Ramsey RESET test; see Godfrey (1988). In addition to minimizing the error of misdiagnoses, the explicit estimation of the auxiliary regressions enables the modeler to view the statistical significance of each individual term. For instance, a practitioner can easily conceal the presence of first-order autocorrelation in the residuals by using a Box–Pierce test with a high order p of lags.

To simplify the discussion, we focus on the LR regression model in Table 9, but the approach can be easily extended to all statistical models of interest in econometrics.

Conditional expectation orthogonality. Consider a set of random variables defined on the probability space $(S, \mathcal{F}, \mathbb{P}(.))$ with bounded variance, including $\mathbf{Z} := (y, \mathbf{X})$ (an $m \times 1$ vector) such that $E(|\mathbf{Z}|^2) < \infty$. For any $D \subset \mathcal{F}$ and any random variable ξ relative to D (Williams, 1991):

$$E(y-\xi)^2 = E\,(y - E(y|D))^2 + E\,[E(y|D) - \xi]^2 \tag{38}$$

This implies that $E(y-\xi)^2$ is minimized when $\xi^* = E(y|D)$. In the case where $D = \sigma(\mathbf{X})$-the σ-field generated by $\mathbf{X}$, i.e., $\xi^* = E(y|\sigma(\mathbf{X}))$. The minimization in (38) follows from the orthogonality between $u = y - E(y|\sigma(\mathbf{X}))$ and any random variable with respect to $\sigma(\mathbf{X})$ that can take the form of any Borel function of $\mathbf{X}$ (Doob, 1953):

$$E\,([y - E(y|\sigma(\mathbf{X}))] \cdot h(\mathbf{X})) = 0, \text{ for any Borel-function } h(\mathbf{X}) \tag{39}$$

This result can be extended to regression functions in the sense that the orthogonality:

$$E\,([y_t - g(\mathbf{X}_t))] \cdot h(\mathbf{X}_t)) = 0, \text{ for all Borel-functions } h(\mathbf{X}_t),\ t \in \mathbb{N} \tag{40}$$

holds if and only if: $g(\mathbf{X}_t) = E(y_t|\sigma(\mathbf{X}_t)),\ t \in \mathbb{N}$. For $u_t = y_t - E(y_t|\sigma(\mathbf{X}_t))$, the orthogonality takes the form:

$$E\,(u_t \cdot h(\mathbf{X}_t)) = 0, \quad t \in \mathbb{N} \tag{41}$$

In light of the fact that $u_t^r, r = 2, 3, \ldots$, define random variables whose mean exists, one can extend the above orthogonality to higher conditional moment functions. Of particular interest is the second, where $E(u_t^2|\sigma(\mathbf{X}_t))$:

$$E\left(\left[u_t^2 - g_2(\mathbf{X}_t)\right] h(\mathbf{X}_t)\right) = 0, \quad \text{for all Borel-functions } h(\mathbf{X}_t),\ t \in \mathbb{N} \tag{42}$$

if and only if $g_2(\mathbf{X}_t) = E(u_t^2|\sigma(\mathbf{X}_t)),\ t \in \mathbb{N}$.

In the case of the LR model, the construction of the M-S tests will be based on seeking legitimate $D_t \subset \mathcal{F}$ for which the orthogonalities below might *not* hold:

$$\text{(i) } E\left([y_t - E(y_t|D_t)]\cdot h_1(D_t)\right) = 0,\ \text{(ii) } E\left(\left[u_t^2 - E\left(u_t^2|D_t\right)\right]\cdot h_2(D_t)\right) = 0,\ t \in \mathbb{N}$$

D_t is operationally legitimate for M-S testing purposes if D_t is a proper subset of the statistical universe of discourse $\mathcal{F}_Z := \sigma(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n)$. Of particular interest is the choice $\mathcal{D}_t = \sigma(\mathbf{X}_t, \mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, \ldots, \mathbf{Z}_1)$, which would render $\{(u_t|\mathcal{D}_t),\ t \in \mathbb{N}\}$ a second-order MD process without imposing independence.

Such potential nonorthogonalities can be framed in terms of using auxiliary regressions of the form:

$$u_t = \delta_1 + \boldsymbol{\Gamma}_1^\top \mathbf{h}_1(D_t) + v_{1t},\ u_t^2 = \delta_2 + \boldsymbol{\Gamma}_2^\top \mathbf{h}_2(D_t) + v_{2t}, \quad t = 1, 2, \ldots, n \tag{43}$$

where $\mathbf{h}_r(D_t)$, $r = 1, 2$, denote the vectors of different Borel functions relating to D_t chosen with a view to pick up different potential departures from the model assumptions. In a certain sense, M-S testing based on (43) amounts to probing for departures from the process $\{(u_t|D_t),\ t \in \mathbb{N}\}$ being a Normal, MD process.

To illustrate the above auxiliary regressions and simplify the discussion, we focus on the LR regression model in Table 9, but the approach can be easily extended to other statistical models.

Example. Possible Borel functions of the original statistical information set $(\mathbf{Z}_t := (y_t, \mathbf{X}_t),\ t = 1, 2, \ldots, n)$ that can be used to define potential statistical information not accounted for by the original model, say $D_t = (\boldsymbol{\psi}_t, \mathbf{z}_{t-1}, \mathbf{t})$:

$$\boldsymbol{\psi}_t := (x_{it} \cdot x_{jt})_{i,j},\ i \geq j = 2, \ldots, k,\ \mathbf{z}_{t-1} := (y_{t-1}, \mathbf{x}_{t-1}), \quad \mathbf{t} := (t, t^2, .., t^p)$$

Conditioning on D_t will give rise to an alternative regression function:

$$E(y_t|D_t) = \alpha_0 + \alpha_1^\top \mathbf{x}_t + \alpha_2^\top \boldsymbol{\psi}_t + \boldsymbol{\alpha}_3^\top \mathbf{z}_{t-1} + \boldsymbol{\Delta}^\top \mathbf{t} \tag{44}$$

As mentioned above, $(\boldsymbol{\psi}_t, \mathbf{z}_{t-1}, \mathbf{t})$ are not "omitted" but "discounted" variables because they are already in the statistical information set: $(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n)$. Comparing the original $E(y_t|\mathbf{X}_t = \mathbf{x}_t) = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t$ with the respecified model, one can construct an auxiliary regression for testing departures from the assumptions [2], [4], and [5]:

$$\text{(GM1): } y_t = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t + u_t,\ \text{(GM2): } y_t = \alpha_0 + \boldsymbol{\alpha}_1^\top \mathbf{x}_t + \alpha_2^\top \boldsymbol{\psi}_t + \boldsymbol{\alpha}_3^\top \mathbf{z}_{t-1} + \boldsymbol{\Delta}^\top \mathbf{t} + v_{1t} \tag{45}$$

Subtracting the estimated GM1, $y_t = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1^\top \mathbf{x}_t + \widehat{u}_t$, from GM2 yields an auxiliary regression in terms of the residuals of the original regression:

$$\widehat{u}_t = (\alpha_0 - \widehat{\beta}_0) + (\boldsymbol{\alpha}_1 - \widehat{\boldsymbol{\beta}}_1)^\top \mathbf{x}_t + \boldsymbol{\alpha}_2^\top \boldsymbol{\psi}_t + \boldsymbol{\alpha}_3^\top \mathbf{z}_{t-1} + \boldsymbol{\delta}^\top \mathbf{t} + v_{1t} \tag{46}$$

This auxiliary regression can be easily extended/modified to include higher order powers of $\mathbf{x}_t$, as well as higher order lags and/or using orthogonal polynomials in t. In cases where the sample size is not large enough, one can use the fitted values $\widehat{y}_t = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1^\top \mathbf{x}_t$ and the residuals $\widehat{u}_{t-i}$, in place of higher order functions of $\mathbf{x}_t$ and lags of $\mathbf{Z}_t$, respectively. The F-type tests for the joint hypotheses:

$$H_0: \overbrace{\boldsymbol{\alpha}_2 = \mathbf{0}}^{[2]},\ \overbrace{\boldsymbol{\alpha}_3 = \mathbf{0}}^{[4]},\ \overbrace{\boldsymbol{\Delta} = \mathbf{0}}^{[5]}, \text{ vs. } H_1 : \boldsymbol{\alpha}_2 \neq \mathbf{0} \text{ or } \boldsymbol{\alpha}_3 \neq \mathbf{0} \text{ or } \boldsymbol{\delta} \neq 0 \tag{47}$$

provides an M-S test for [2], [4], and [5], as they affect the regression function.

Analogous reasoning can be used to derive an auxiliary regression corresponding to the skedastic function $E(u_t^{2\sim}|\widehat{\ }\mathbf{X}_t = \mathbf{x}_t) = \sigma^2$:

$$\widehat{u}_t^2 = \gamma_0 + \boldsymbol{\alpha}_2^\top \boldsymbol{\psi}_t + \boldsymbol{\alpha}_3^\top \mathbf{z}_{t-1}^2 + \boldsymbol{\Delta}^\top \mathbf{t} + v_{2t} \tag{48}$$

where $\mathbf{z}_{t-1}^2$ denotes quadratic functions of $\mathbf{z}_{t-1}$, and the joint hypotheses are:

$$H_0: \overbrace{\boldsymbol{\alpha}_2 = \mathbf{0}}^{[3]},\ \overbrace{\boldsymbol{\alpha}_3 = \mathbf{0}}^{[4]},\ \overbrace{\boldsymbol{\Delta} = \mathbf{0}}^{[5]}, \text{ vs. } H_1 : \boldsymbol{\alpha}_2 \neq \mathbf{0} \text{ or } \boldsymbol{\alpha}_3 \neq \mathbf{0} \text{ or } \boldsymbol{\delta} \neq 0 \tag{49}$$

It is important to emphasize that (48) distinguishes between testing for heteroskedasticity $\text{Var}(y_t|\mathbf{X}_t = \mathbf{x}_t) = g(\mathbf{x}_t),\ \mathbf{x}_t \in \mathbb{R}_X^k$, and for conditional variance heterogeneity $\text{Var}(y_t|\mathbf{X}_t = \mathbf{x}_t) = g(t),\ t \in \mathbb{N},\ \mathbf{x}_t \in \mathbb{R}_X^k$; the latter can be trends or shifts in $Var(y_t|\mathbf{X}_t = \mathbf{x}_t)$. This reveals how misleading the traditional strategy of using HCSE (Hansen, 1999) and the HAC SEs can be, when employed as a panacea for all potential forms of departure from assumptions [2]–[5], affecting the regression and skedastic functions; see Greene (2011).

The above auxiliary regressions are only indicative of factors that might be used in practice; several variations/extensions one might consider, include powers of $\widehat{y}_t$ and $\widehat{u}_{t-i}$, so long as the extra terms are functions of the original statistical information.

It is also interesting to note that the validity of assumptions [1]–[5] ensures that the error $\{(u_t||\mathbf{X}_t = \mathbf{x}_t),\ t \in \mathbb{N}\}$ is a Normal, MD process.

[1] Normality is the only assumption that the auxiliary regressions ((46)–(48)) do not test for. This is because the available M-S tests for [1] assume that the other assumptions are valid, rendering their results questionable when any of the assumptions [2]–[5] are invalid. Hence, for a reliable test of Normality one should secure the validity of [2]–[5] beforehand.

The above auxiliary regressions can be modified/extended to other statistical models of interest in econometrics, including models for cross section and panel data.

5.2 *Empirical Example*

Lai and Xing (2008, pp. 72–81) illustrate the capital asset pricing model (CAPM) using *monthly data* for the period August 2000–October 2005 ($n = 64$). For simplicity, let us focus on one of their equations where: y_t is excess (log) returns of Intel, x_t is the market excess (log) returns based on the SP500 index; the risk free returns is based on the 3-month Treasury bill rate. Estimation of the statistical (LR) model that nests the CAPM when the constant is zero yields:

$$y_t = \underset{(0.009)}{0.02} + \underset{(0.237)}{1.996} x_t + \widehat{u}_t,\ R^2 = 0.536,\ s = 0.0498,\ n = 64 \tag{50}$$

where the SEs are given in parentheses. The authors proceed to test the significance of the regression coefficients (β_0, β_1) using t-tests, and conclude that for $\alpha = 0.025$, the coefficient β_0 is statistically insignificant but β_1 is significant, providing evidence for the CAPM.

The above inference results will be trustworthy when the LR probabilistic assumptions [1]–[5] (Table 9) are valid for the particular data; otherwise, their trustworthiness will be questionable. A glance at the t-plots of the data $\{(y_t, x_t),\ t = 1, 2, \ldots, n\}$ (Figures 4 and 5) suggests that assumptions [4] and [5] are likely to be invalid because the data exhibit very distinct time cycles and trends in the mean, and a shift in the variance after observation $t = 30$; see also the residuals in Figure 6.

These misspecifications are confirmed by the following auxiliary regressions:

$$\widehat{u}_t = \underset{(0.011)}{0.02} + \underset{(0.311)}{0.370} x_t + \underset{(0.079)}{0.091} t + \underset{(0.094)}{0.253} t^2 + \underset{(0.075)}{0.172} t^3 - \underset{(0.122)}{0.343}\, \widehat{u}_{t-1} + \widehat{v}_{1t}$$
$$R^2 = 0.197,\ s = 0.0463,\ n = 63 \tag{51}$$

$$\widehat{u}_t^2 = \underset{(0.001)}{0.002} - \underset{(0.0005)}{0.002} t + \underset{(0.086)}{0.141}\, \widehat{y}_t^2 + \widehat{v}_{2t},\ \ R^2 = 0.2,\ s = 0.0037,\ n = 64 \tag{52}$$

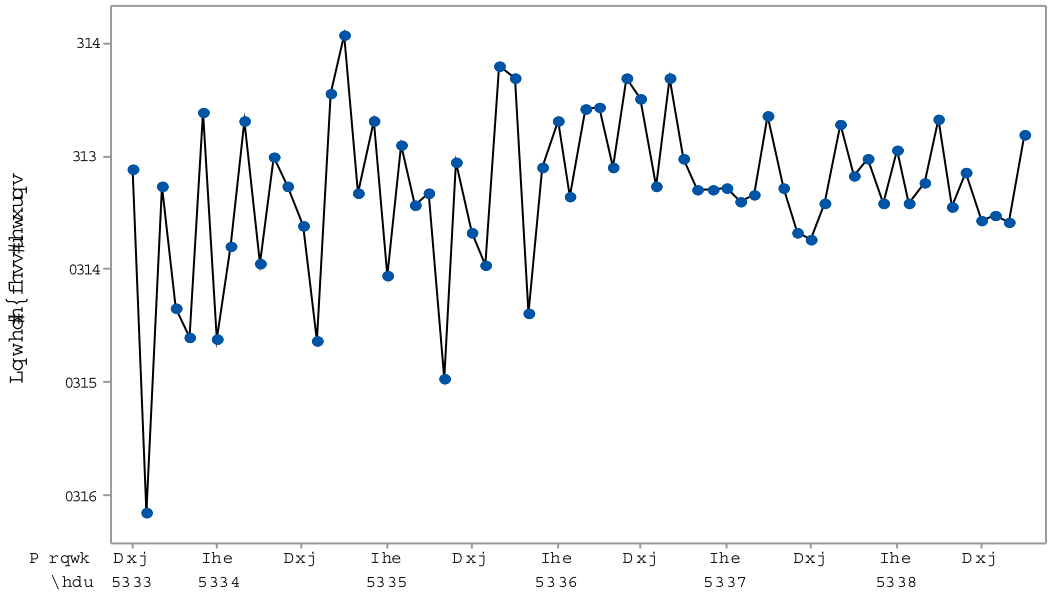

Figure 4. Intel Corp. Excess Returns. [Colour figure can be viewed at wileyonlinelibrary.com]

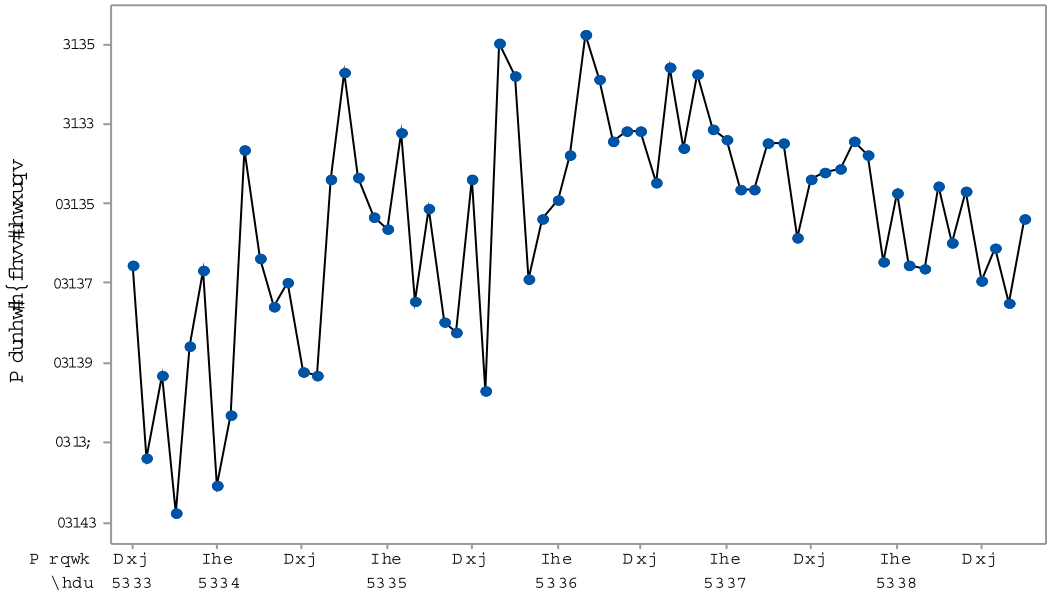

Figure 5. Market Excess Returns. [Colour figure can be viewed at wileyonlinelibrary.com]

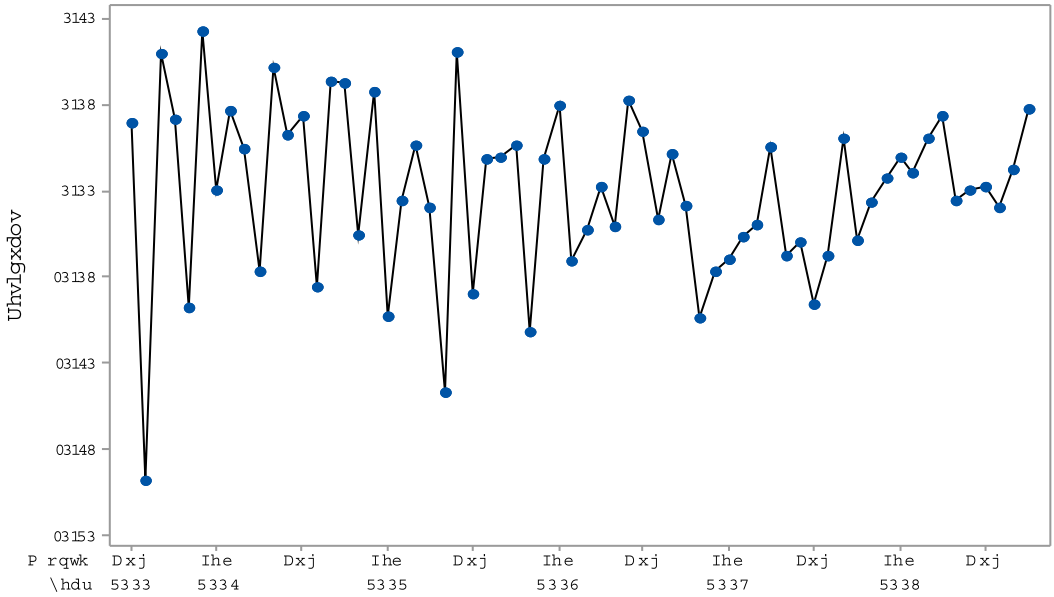

Figure 6. t-Plot of the Residuals from (50). [Colour figure can be viewed at wileyonlinelibrary.com]

The above M-S testing results indicate clearly that no reliable inferences can be drawn on the basis of the estimated model in (50) since assumptions [3]–[5] are invalid. Note that the use of HAC SEs cannot address these misspecifiations; a proper respecification would lead to a Student's t DLR model.

Substantive adequacy. To illustrate the perils of a misspecified model, consider posing the question: is z_{t-1} last period's excess returns of General Motors (see Figure 7) an omitted variable in (50)? Adding z_{t-1} gives rise to:

$$y_t = \underset{(0.008)}{0.013} + \underset{(0.233)}{2.086}x_t - \underset{(0.129)}{0.296}z_{t-1} + \widehat{\epsilon}_t,\ R^2 = 0.577,\ s = 0.0483,\ n = 64 \tag{53}$$

which, when taken at face value, suggests that z_{t-1} is a relevant omitted variable. The truth is that any variable that picks up the unmodeled trend will misleadingly appear to be statistically significant. Indeed, a simple respecification of the original model, such as adding trends and lags to account for the indicated departures:

$$y_t = \underset{(0.015)}{0.035} + \underset{(0.315)}{2.319}x_t + \underset{(0.081)}{0.086}t + \underset{(0.101)}{0.190}t^2 + \underset{(0.075)}{0.158}t^3 - \underset{(0.122)}{0.316}y_{t-1} + \underset{(0.498)}{0.524}x_{t-1} - \underset{(0.177)}{0.164}z_{t-1} + \widehat{\varepsilon}_t$$
$$R^2 = 0.645,\ s = 0.0462,\ n = 63$$

renders z_{t-1} insignificant. For a more detailed example, see Spanos (2010b).

5.3 *Respecification*

The PR perspective views respecification as a repeat of the specification facet with a view to select a new statistical model, using the tripartite partitioning of $\mathcal{P}(\mathbf{z})$. The aim is to select probabilistic assumptions that account for the chance regularities not accounted for by the original model. A discerning interpretation of a comprehensive M-S testing results and not one departure at a time could guide the respecification by repartitioning $\mathcal{P}(\mathbf{z})$ with a view to specify a new statistical model that accounts for the statistical

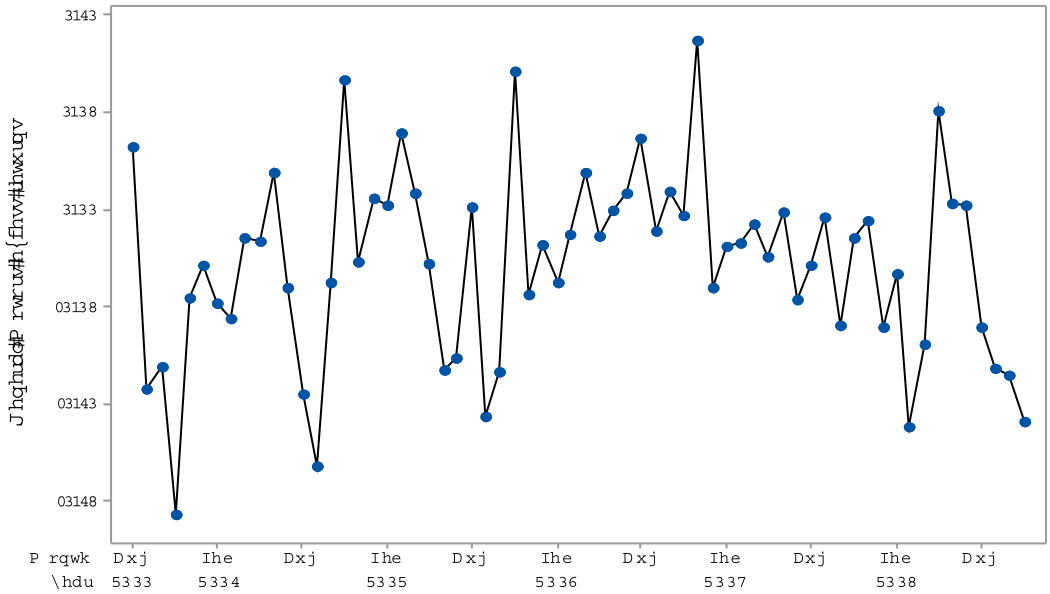

Figure 7. General Motors Excess Returns. [Colour figure can be viewed at wileyonlinelibrary.com]

information the original model did not. A tentative new model is estimated and its own assumptions are tested thoroughly.

What is important to reiterate is that in M-S testing, the null is always H_0: $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is valid, but the alternative $\overline{H}_0$: $\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is nonoperational. Hence, the modeler needs to select particularized alternatives or directions of departure that could never span $\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. This implies that when H_0 is rejected, the specific alternative H_1, specifying a particularized subset of $\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, is never an option for respecification purposes without further testing. In particular, the M-S testing based on auxiliary regressions, such as (46) and (48), could only provide information about departures from the original model assumptions. Significant coefficients indicate particular directions of departure from these assumptions. The auxiliary regressions do not provide a clear answer as to what the respecified model should look like. That is decided by the statistical adequacy of the respecified model; its assumptions are tested anew and shown to be valid. For instance, the significance of $\boldsymbol{\psi}_t$ in (46) and (48) provides only an indication that the assumptions [2] and [3] are invalid. The relevant nonlinear and heteroskedastic functional forms, however, are unlikely to coincide with the polynomials used in (46)–(48). Their functional forms are determined by the joint distribution of $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$, and can be validated by M-S testing.

The PR perspective provides a broader and more coherent vantage point from that stemming from the error process $\{(u_t|\mathbf{X}_t=\mathbf{x}_t),\ t\in\mathbb{N}\}$. For instance, it views the LR model as specified in terms of the regression and skedastic functions of $D(y_t|\mathbf{X}_t;\boldsymbol{\theta})$:

$$E(y_t|\mathbf{X}_t=\mathbf{x}_t)=h(\mathbf{x}_t),\ \ \mathrm{Var}(y_t|\mathbf{X}_t=\mathbf{x}_t)=g(\mathbf{x}_t),\ \ \mathbf{x}_t\in\mathbb{R}^k_X$$

where the functional forms $h(.)$ and $g(.)$ and the relevant parameterization $\boldsymbol{\theta}$ stem from the joint distribution $D(y,\mathbf{X}_t;\boldsymbol{\varphi})$. From this perspective, departures from particular assumptions might relate to both functions. For instance, the move of retaining the Linearity and Normality assumptions, but adopting an arbitrary form of heteroskedasticity (Greene, 2011), can easily give rise to an internally inconsistent set of probabilistic assumptions.

In contrast, the traditional respecification usually takes the form of ad hoc "error-fixing" moves, based on adopting the particularized alternative H_1 of the M-S test applied; a classic case of the fallacy of rejection. This arises because the results of the M-S testing for the original model do not justify the move to the new model; only the statistical adequacy of the latter accomplishes that. The source of the problem for the traditional respecification is that $\{(u_t|\mathbf{X}_t=\mathbf{x}_t),\ t\in\mathbb{N}\}$ provides a much narrower perspective on M-S testing and respecification than $\{\mathbf{Z}_t:=(y_t,\mathbf{X}_t),\ t\in\mathbb{N}\}$. "Correcting" an error assumption by modeling $u_t=y_t-\beta_0-\boldsymbol{\beta}_1^\top\mathbf{x}_t$ retains (unchanged) the original systematic component $E(y_t|\mathbf{X}_t=\mathbf{x}_t)=\beta_0+\boldsymbol{\beta}_1^\top\mathbf{x}_t$ (Table 9). As a result, this gives rise to respecified models with unnecessary and often implausible restrictions.

Example. Sargan (1964) showed that the move of replacing the OLS estimators of $(\boldsymbol{\beta},\sigma^2)$ with the GLS estimators, or equivalently, adopting the AC-LR model $(y_t=\beta_0+\boldsymbol{\beta}_1^\top\mathbf{x}_t+u_t,\ u_t=\rho u_{t-1}+\varepsilon_t)$ when the no-autocorrelation assumption of the LR model $(y_t=\beta_0+\boldsymbol{\beta}_1^\top\mathbf{x}_t+u_t)$ is rejected, amounts to adopting the respecified model:

$$y_t=\beta_0(1-\rho)+\boldsymbol{\beta}_1^\top\mathbf{x}_t+\rho y_{t-1}-\rho\boldsymbol{\beta}_1^\top\mathbf{x}_{t-1}+\varepsilon_t,\ (\varepsilon_t|\mathbf{X}_t=\mathbf{x}_t)\sim\text{NIID}\ \left(0,\sigma_\varepsilon^2\right)$$

This is a special case of the dynamic LR (DLR) model:

$$y_t=\alpha_0+\boldsymbol{\alpha}_1^\top\mathbf{x}_t+\alpha_2y_{t-1}+\boldsymbol{\alpha}_3^\top\mathbf{x}_{t-1}+v_t,\ (v_t|D_t^0)\sim\text{NMD}\left(0,\sigma_0^2\right)\tag{54}$$

subject to the *common factor* (CF) restrictions: $\boldsymbol{\alpha}_3+\boldsymbol{\alpha}_1\alpha_2=\mathbf{0}$. Note that $D_t^0:=[\mathbf{X}_t=\mathbf{x}_t,\mathbf{X}_{t-1}=\mathbf{x}_{t-1},\sigma(y_{t-1})]$ is the respecified conditioning information set and "$(v_t|D_t^0)\sim\text{NMD}(0,\sigma_0^2)$" denotes a "Normal, Martingale Difference" process.

Viewed from the PR respecification, (54) arises naturally (without the CF restrictions) when the no-autocorrelation assumption is replaced with Markov dependence for $\{\mathbf{Z}_t:=(y_t,\mathbf{X}_t),\ t\in\mathbb{N}\}$ and $D_t:=(\mathbf{X}_t=\mathbf{x}_t)$ with D_t^0, giving rise to the new statistical GM: $y_t=E(y_t|D_t^0)+v_t$. From the PR perspective, the move to adopt the AC-LR model is justified only when: (i) the DLR model in (54) is statistically adequate, and (ii) the CF restrictions are valid. (i) holds when all its five assumptions, which constitute a hybrid of the LR model (Table 9) and the AR(1) model (Table 8), are valid for data $\mathbf{Z}_0$. Note that even when (i) holds, but (ii) does not, the OLS estimators of $(\boldsymbol{\beta},\sigma^2)$ are both biased and inconsistent; see Spanos (1986).

6. Summary and Conclusions

The primary objective of empirical modeling is "to learn from data $\mathbf{Z}_0$" about observable phenomena of interest using a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ as the link between theory and data. Substantive subject matter information, codified in the form of a structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, plays an important role in demarcating and enhancing this learning from data when it does not belie the statistical information in $\mathbf{Z}_0$. Behind every structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, there is a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ that comprises the probabilistic assumptions imposed on one's data, and nests $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ via generic restrictions of the form $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$.

The paper proposes a broadening of Fisher's (1922) framework with a view to separate the modeling from the inference facets in order to bring out the different potential errors and use strategies to safeguard against them. The statistical misspecification of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is a crucial error because it undermines the reliability of the inference procedures based on it. Relying on weak assumptions, combined with vague "robustness" claims and heuristics invoking $n\to\infty$, will not circumvent this error in practice. The brief historical overview of M-S testing brings out several foundational issues that can be addressed in the proposed PR framework. Establishing the adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ calls for a thorough M-S testing, combined with a coherent respecification strategy that relies on changing the assumptions imposed on $\{\mathbf{Z}_t,t\in\mathbb{N}\}$. Joint M-S tests based on auxiliary regressions provide a most effective tool for detecting departures from

the model assumptions. The traditional respecification of adopting the particular alternative H_1 used by the M-S test is fallacious.

Distinguishing between statistical and substantive inadequacy is crucial because it is one thing to argue that a structural model $\mathcal{M}_{\varphi}(\mathbf{z})$ is only an approximation of the reality it aims to explain, and entirely another to claim that the probabilistic assumptions imposed on $\mathcal{M}_{\theta}(\mathbf{z})$ are invalid for data $\mathbf{Z}_0$. $\mathcal{M}_{\varphi}(\mathbf{z})$ may always come up short in securing substantive adequacy vis-a-vis the phenomenon of interest, but $\mathcal{M}_{\theta}(\mathbf{z})$ may be perfectly adequate for answering substantive questions of interest. This could happen when $\mathcal{M}_{\varphi}(\mathbf{z})$ does not belie the data, i.e., when (i) the (implicit) $\mathcal{M}_{\theta}(\mathbf{z})$ is statistically adequate, and (ii) the overidentifying restrictions $\mathbf{G}(\varphi, \theta) = \mathbf{0}$ are data-acceptable. Further probing is needed to establish that $\mathcal{M}_{\varphi}(\mathbf{z})$ adequately captures (describes, explains, and predicts) the phenomenon of interest.

Acknowledgements

Author thanks two anonymous referees whose constructive comments helped to improve the original paper.

References

Aitken, A.C. (1935) On least-squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh* 55: 42–48.

Anderson, R.L. (1942) Distribution of the serial correlation coefficient. *The Annals of Mathematical Statistics* 13: 1–13.

Andreou, E. and Spanos, A. (2003) Statistical adequacy and the testing of trend versus difference stationarity. *Econometric Reviews* 22: 217–237.

Anscombe, F.J. (1961) Examination of residuals. In J. Neyman (ed.), *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1 (pp. 1–36), University of California Press, Berkeley, CA.

Anscombe, F.J. and Tukey, J.W. (1963) The examination and analysis of residuals. *Technometrics* 5: 141–160.

Bahadur, R.R. and Savage, L.J. (1956) The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics* 27: 1115–1122.

Balakrishanan, N. and Meshbah, M. (eds.), (2002) *Goodness-of-Fit Tests and Model Validity*. Boston: Birkhauser.

Box, G.E.P. (1953) Non-normality and tests on variances. *Biometrika* 40: 318–335.

Box, G.E.P. (1979) Robustness in the strategy of scientific model building. In R.L. Launer and G.N. Wilkinson (eds.), *Robustness in Statistics* (pp. 201–236). London: Academic Press.

Box, G.E.P. and Jenkins, G.M. (1970) *Time Series Analysis: Forecasting and Control*, 1976 revised edn. San Francisco, CA: Holden-Day.

Box, G.E.P. and Watson, G.S. (1962) Robustness to non-normality of regression tests. *Biometrika* 49: 93–106.

Box, J.F. (1978) *R. A. Fisher: The Life of a Scientist*. NY: Wiley.

Breusch, T.S. and Pagan, A.R. (1979) A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47: 1287–1294.

Breusch, T.S. and Pagan, A.R. (1980) The Lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies* 47: 239–253.

Cochrane, D. and Orcutt, G.H. (1949) Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association* 44: 32–61.

Davidson, J.E.H., Hendry, D.F., Srba, F. and Yeo, S.J. (1978) Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal* 88: 661–692.

D'Agostino, R.B. and Pearson, E.S. (1973) Tests for departures from normality: empirical results for the distributions of β_2 and $\sqrt{\beta_1}$. *Biometrika* 62: 243–250.

Domowitz, I. and White, H. (1982) Misspecified models with dependent observations. *Journal of Econometrics* 20: 35–58.

Durbin, J. and Watson, G.S. (1950) Testing for serial correlation in least squares regression, I. *Biometrika* 37: 409–428.
Engle, R.F., Hendry, D.F. and Richard, J.F. (1983) Exogeneity. *Econometrica* 51: 277–304.
Fisher, R.A. (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10: 507–521.
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) *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.
Fisher, R.A. (1930) The moments of the distribution for normal samples of measures of departure from normality. *Proceedings of the Royal Society of London* 130: 16–28.
Fisher, R.A. (1956) *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.
Geary, R.C. (1935) The ratio of the mean deviation to the standard deviation as a test of normality. *Biometrika* 27: 310–332.
Geary, R.C. (1947) Testing for normality. *Biometrika* 34: 209–242.
Geyer, C.J. (2013) Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. In G. Jones, and X. Shen (eds.), *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton* (pp. 1–24). Hayward, CA: Institute of Mathematical Statistics.
Godfrey, L.G. (1988) *Misspecification Tests in Econometrics*. Cambridge: Cambridge University Press.
Good, I.J. (1988) The interface between statistics and philosophy of science. *Statistical Science* 3(4): 386–397.
Greene, W.H. (2011), *Econometric Analysis*, 7th edn. NJ: Prentice Hall.
Hansen, B.E. (1999) Discussion of 'Data mining reconsidered'. *The Econometrics Journal* 2: 192–201.
Hendry, D.F. (1980) Econometrics-alchemy or science?' *Economica* 47: 387–406.
Hendry, D.F. (2003) J. Denis Sargan and the origins of LSE econometric methodology. *Econometric Theory* 19: 457–480.
Hettmansperger, T.P. (1984) *Statistical Inference Based on Ranks*. NY: Wiley.
Kolmogorov, A.N. (1941) Confidence limits for an unknown distribution function. *Annals of Mathematical Statistics* 12: 461–463.
Lai, T.L. and Xing, H. (2008) *Statistical Models and Methods for Financial Markets*. NY: Springer.
CamLe, L. (1986) *Asymptotic Methods in Statistical Decision Theory*. NY: Springer-Verlag.
Lehmann, E.L. (1986) *Testing Statistical Hypotheses*, 2nd edn. NY: Wiley.
Lehmann, E.L. (1990) Model specification: the views of Fisher and Neyman, and later developments. *Statistical Science* 5: 160–168.
Lehmann, E.L. (1999) "Student" and small-sample theory. *Statistical Science* 14: 418–426.
Lilliefors, H.W. (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* 62: 399–402.
Mahalanobis, P.C. (1938) Prof. R. A. Fisher. *Sankhya* 4: 241–244.
Mayo, D.G. (1996) *Error and the Growth of Experimental Knowledge*. Chicago: The University of Chicago Press.
Mayo, D.G. and Spanos, A. (2006) Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *The British Journal of the Philosophy of Science* 57: 321–356.
Mizon, G.E. (1977) Model selection procedures. In M.J. Artis and R.A. Nobay (eds.), *Studies in Modern Economic Analysis* (pp. 97–120). Oxford: Blackwell.
Mood, A.M. (1940) The distribution theory of runs. *Annals of Mathematical Statistics* 11: 367–392.
Newey, W.K. (1985) Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53: 1047–1070.
Neyman, J. (1937) 'Smooth test' for goodness of fit. *Scandinavian Actuarial Journal* 3–4: 149–199.
Neyman, J. and Pearson, E.S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* 20A: 175–240.
Neyman, J. and Pearson, E.S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, A* 231: 289–337.
Pearson, E.S. (1930) A further development of tests for normality. *Biometrika* 22: 239–249.
Pearson, E.S. (1931) Analysis of variance in cases of non-normal variation. *Biometrika* 23: 114–133.
Pearson, E.S. (1935) A comparison of β_2 and Mr Geary's w_n criteria. *Biometrika* 27: 333–352.

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* 5: 157–175.
Phillips, P.C.B. (1983) Exact small sample theory in the simultaneous equations model. In M.D. Intriligator and Z. Griliches (eds.), *Handbook of Econometrics*, Vol. 1 (pp. 449–516). Amsterdam: North-Holland.
Phillips, P.C.B. (1987), "Time Series Regression with a Unit Root", *Econometrica*, 55: 277–301.
Phillips, P.C.B. (1988) The ET interview: Professor James Durbin. *Econometric Theory* 4: 125–157.
Pitman, E.J.G. (1937) Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika* 29: 322–335.
Rao, C.R. (2004) Statistics: reflections on the past and visions for the future. *Amstat News* 327: 2–3.
Ramsey, J.B. (1969) Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society Series B*, 31: 350–371.
Sargan, D.J. (1964) Wages and prices in the U.K.: a study in econometric methodology. In P. Hart, G. Mills and J.K. Whitaker (eds.), *Econometric Analysis for National Economic Planning*, Vol. 16 of Colston Papers (pp. 25–54). London: Butterworths.
Shapiro, S.S. and Wilk, M.B. (1965) An analysis of variance test for normality. *Biometrika* 52: 591–611.
Spanos, A. (1986) *Statistical Foundations of Econometric Modelling*. Cambridge: Cambridge University Press.
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) Revisiting the omitted variables argument: substantive vs. statistical adequacy. *Journal of Economic Methodology* 13: 179–218.
Spanos, A. (2010a) Akaike-type criteria and the reliability of inference: model selection vs. statistical model specification. *Journal of Econometrics* 158: 204–220.
Spanos, A. (2010b) Statistical adequacy and the trustworthiness of empirical evidence: statistical vs. substantive information. *Economic Modelling* 27: 1436–1452.
Spanos, A. and McGuirk, A. (2001) The model specification problem from a probabilistic reduction perspective. *Journal of the American Agricultural Association* 83: 1168–1176.
Stigler, S. (2005) Fisher in 1921. *Statistical Science* 20: 32-49.
Student (1908) The probable error of the mean. *Biometrika* 6: 1–25.
Tauchen, G. (1985) Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics* 30: 415–443.
von Neumann, J. (1941) Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics* 12: 367–395.
Wald, A. and Wolfowitz, J. (1943) An exact test for randomness in the nonparametric case based on serial correlation. *Annals of Mathematical Statistics* 14: 378–388.
Wasserman, L. (2006) *All of Nonparametric Statistics*. NY: Springer.
White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and direct test for heteroskedasticity. *Econometrica* 48: 817–838.
White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1–25.
White, H. (1999) *Asymptotic Theory for Econometricians*, revised edn. London: Academic Press.
Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics* 1: 80–83.
Williams, D. (1991) *Probability with Martingales*. Cambridge: Cambridge University Press.
Wolfowitz, J. (1944) Asymptotic distribution of runs up and down. *Annals of Mathematical Statistics* 15: 163–172.