Mis-Specification Testing in Retrospect and Prospect

Aris Spanos
Department of Economics,
Virginia Tech, Blacksburg,
VA 24061, USA

October 2016

Abstract

The primary objective of this paper is threefold. First, to take a retrospective view of Mis-Specification (M-S) testing, going back to Karl Pearson (1900). Second, to critically discuss several arguments questioning the value and role of M-S testing. Instead, they favor relying on weak, but non-testable probabilistic assumptions, combined with invoking asymptotic results and vague robustness claims. Third, to make a case for extending/modifyng Fisher’s statistical framework by clearly separating the modeling from the inference facet. The proposed framework separates, specification, estimation, M-S and respecification with a view to secure a statistically adequate model, from inference proper. The proposed framework helps to elucidate various foundational issues associated with M-S testing. A case is made for employing joint M-S tests based on auxiliary regressions to enhance the effectiveness and reliability of probing for potential statistical misspecifications.

Key words: Misspecification testing; statistical model; specification; respecification; statistical vs. substantive adequacy; Neyman-Pearson testing; error probabilities; reliability of inference; weak probabilistic assumptions; non-testable assumptions; asymptotic inference.
1 Introduction

The problem of misspecification arises when any of the assumptions invoked by a particular statistical inference procedure is invalid for the data in question. Departures from the invoked assumptions distort the sampling distribution of a statistic (estimator, test, predictor). As a result of such distortions the reliability of an inference procedure is often undermined, in the sense that it could give rise to inconsistent estimates or/and sizeable discrepancies between the actual type I and II error probabilities and the nominal ones – the ones derived by invoking these assumptions. Applying a .05 significance level test, when the actual type I error is closer to .9 will lead an inference astray.

Mis-Specification (M-S) testing aims to assess the validity of the assumptions invoked by a statistical model upon which the inference procedure is based. The usefulness of M-S tests is twofold:

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

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 (a) weak but often non-testable probabilistic assumptions, combined with (b) asymptotic inference results, and (c) vague ‘robustness’ claims, as a substitute for establishing the validity of the relevant probabilistic assumptions using comprehensive M-S testing.

The current neglect of statistical adequacy is mainly due to the fact that the overall statistical modeling framework has been beclouded with 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) insightful categorization into: Specification (the choice of the statistical model), Estimation and Distribution (inference), with M-S testing being included in Distribution. This grouping of M-S testing with other forms of inference, however, left a lot of unanswered questions concerning its role and nature. 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 is the inadvertent blending of statistical and substantive premises which arises because in most scientific fields the statistical model is implicitly specified via probabilistic assumptions attached to the error term(s) of a substantive (structural) model. As a result, the
econometric literature conflates two very different forms of misspecification: statistical and substantive. The most serious form of misspecification, as viewed by econometrics textbooks is that of ‘omitted variables’ (see Greene, 2011), which has nothing to do with statistical misspecification; see Spanos (2006a). To make matters worse, M-S testing is often conflated with Neyman-Pearson’s (N-P) hypothesis testing because their key differences have not been sufficiently elucidated.

The primary objective of this paper is to attempt a retrospective view M-S testing with a view to shed light on the issues and problems raised. The first modern M-S test was the celebrated Karl Pearson (1900) goodness-of-fit test. Pearson’s test is rarely viewed as a M-S test, however, because the line between the latter and other types of testing are blurred to this day. In an attempt to elucidate the blurred lines and the 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. Section 3, discusses the importance of M-S testing in establishing the statistical adequacy of a statistical model – the validity of its probabilistic assumption vis-a-vis the data in question. Section 4, revisits the ambivalent attitude toward M-S testing and questions the pertinence of the various arguments behind the current neglect of the problem of testing the validity of model assumptions. Section 5 discusses briefly several foundational issues pertaining to M-S testing.

2 M-S testing: a brief historical overview

2.1 Karl Pearson

Viewed retrospectively, the modern era on M-S testing was initiated by Karl Pearson’s (1900) paper on goodness-of-fit; see Cox (2002). There is an element of anachronism in this view, however, because Pearson’s approach to statistics was very different from the current approach which is primarily due to Fisher (1922).

In its simplest form, Pearson’s approach would begin with a set of data $x_0:= (x_1, ..., x_n)$ in search of a descriptive model in the form of a frequency curve. This would be attained by summarizing $x_0$ in the form of a histogram, and then fitting a frequency curve $f_0(x)$ from the Pearson family generated by:

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

(1)

to describe it as closely as possible. The choice of $f_0(x)$ would be based on the values
of the estimated parameters $\hat{\theta} := (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)$ 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:

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

(2)

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

$$\mathbb{P}(\eta(X) > \eta(x_0)) = \rho(x_0).$$

(3)

His rationale was that the bigger the value of $\eta(x_0)$, 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 $x_0$; 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 nature of the finite sampling distribution of:

$$\sqrt{n}(X_n - \mu) \sim \mathcal{St}(n-1), \text{ for any } n > 1,$$

(4)

where $X_n = \frac{1}{n} \sum_{t=1}^{n} X_t$, $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2$, $\mathcal{St}(n-1)$ 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.

A formal proof of the result was provided by Fisher (1915) who brought out explicitly the probabilistic assumptions that gave rise to the result, in the form of the simple Normal model:

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

(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); see Savage (1976). Instead of starting with data $x_0$ in search of a descriptive model, like Pearson, he proposed to interpret the data as a random sample from a pre-specified 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 re-arrangement 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. From Fisher’s perspective data $x_0$ are interpreted as a typical realization from $M_\theta(x)$.

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 derivates, 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 \textit{a posteriori}.” (p. 314)

Indeed, 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 an objective justification for statistical induction stemming from being able to objectively test the adequacy of its premises, i.e. the probabilistic assumptions imposed on data $x_0$:

“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 $\mu$ in (4) is replaced by the known value $\mu_0$ yields a test statistic, which, when evaluated under $H_0$, yields the same distributional result:
\[
\tau(X) = \frac{1}{\sqrt{n}} \frac{X_n - \mu_0}{s} \overset{H_0}{\sim} \text{St}(n-1).
\]  

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:

\[
P(\tau(X) > \tau(x_0); H_0) = p(x_0),
\]

interpreting it as an indicator of ‘discordance’ between data \(x_0\) and \(H_0\), using a threshold value \(c_0\) [e.g., .01, .025, .05]; \(p(x_0) < c_0\) data \(x_0\) indicate that \(H_0\) is discordant with data \(x_0\). 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 re-frame Pearson’s chi-square test and ground it on firmer foundations by:

(a) unveiling the implicit probabilistic assumptions of IID underlying (2),
(b) explaining that (2) constitutes an asymptotic approximation (as \(n \to \infty\)),
(c) (2) stems from evaluating the test statistic under the null \((H_0): \mu = \mu_0\),
(d) making explicit the hypothesis of interest:

\[H_0: f_0(x) = f^*(x) \in \text{Pearson}(\theta),\]

where \(f_0(x)\) denotes the assumed density and \(f^*(x)\) the ‘true’ density,
(e) correcting Pearson’s degrees of freedom, and
(f) formalizing Pearson’s p-value into: \(P(\eta(X) > \eta(x_0); H_0) = p(x_0)\).

Note that, the expression ‘\(f^*(x)\)’ is the ‘true’ density’ is a shorthand for saying that ‘data \(x_0\) constitute a typical realization of the sample \(X\) with distribution \(f^*(x)\)’. 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 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:

(a) Narrowed down the scope of Fisher testing by focusing exclusively on testing within the boundaries of the prespecified statistical model \(M_\theta(x)\).
(b) Framed the hypotheses of interest in terms of the parameters \(\theta \in \Theta\).
(c) Extended Fisher’s null hypothesis \((H_0)\) from a single value to a subset of \(\Theta\), say \(H_0: \theta \in \Theta_0\).
(d) 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:

\[
\mathbb{R}_X^n = \left\{ \begin{array}{c}
\frac{C_0}{C_1} \leftrightarrow \Theta_0 \\
\frac{C_1}{C_0} \leftrightarrow \Theta_1 
\end{array} \right\} = \Theta
\]

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 a M-S test but not a N-P test since that latter:

(i) is always framed in terms of the parameters of the statistical model \(M_\theta(x)\) in question, and

(ii) probes within the boundaries of the prespecified statistical model \(M_\theta(x)\).

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 which 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 1920’s 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 \(\tau(X)\) assuming some skewed distribution (possibly log-Normal). We know this from Gosset’s reply:

"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." (quoted by Lehmann, 1999).

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 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 Egon Pearson (1930). Gosset realized that simulation alone was not enough, and called upon Fisher again calling for his help very bluntly:

“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’s 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.

(a) 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.

(b) 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 insight. For instance, the proper comparison is not between the error probabilities of the t-test under a non-Normal distribution with those of the alternative test. This comparison ignores the fact that the t-test is no longer the ‘optimal’ test under non-Normality. Hence, the proper comparison is between the alternative robust test and the optimal test under a different distributional assumption. For instance, the nonparametric literature in the late 1940s early 1950s established that the t-test is robust to non-Normal distributions which 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 $\bar{X}$, neither of which exists! The case
of the Uniform is also problematic. Assuming that the Uniform is the appropriate
distribution:

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

the robustness of the t-test to symmetric departures is questionable, because the
optimal test for the hypothesis: $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$, is no longer the t-test,
but the test (Neyman and Pearson, 1928):

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

where $(X_{[1]}, X_{[n]})$ denote the smallest and the largest elements of the sample, and $F(2, 2(n-1))$ the F distribution with 2 and $2(n-1)$ degrees of freedom. One can argue
that the relevant error probabilities for comparison purposes are no longer the ones
associated with the t-test under the Uniform distribution, but the ones associated
with the test based on (8), which is the 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 $(\hat{\alpha}_3, \hat{\alpha}_4)$ under Normality, and suggested that these
results can be used to construct tests for 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), 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), Wilcoxon (1945).

(iii) Testing the assumption of Normality; see Pearson (1930; 1931; 1935), Geary
derivation of the sampling distributions of the sample third ($\hat{\alpha}_3$) and fourth ($\hat{\alpha}_4$)
central moments led the skewness-kurtosis test based on:

$$SK(X) = \frac{m_3}{\mu_3^2} + \frac{m_4}{\mu_4} (\alpha_4 - 3)^2 \frac{H_0}{\alpha} \chi^2(2), \quad \mathbb{P}(SK(X) > SK(x_0); H_0) = p(x_0).$$

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 transformed the original continuous random variable $X$ with a prespecified cdf $F_0(.)$ using the probability integral transformation: $Z = F_0(X) \sim U(0, 1)$.

Second, he replaced the generic alternative hypothesis $H_1$: $f^*(z) \neq f_0(z)$, with a particular family:

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

The proposed family of distributions takes the form:

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

where $\varphi = (\varphi_1, \varphi_2, ..., \varphi_p)$, $c(\varphi)$ is a normalizing constant to ensure that:

(a) $\int_{z \in [0, 1]} f_p(z; \varphi) dz = 1$, and (b) $h_i(z), i = 1, ..., p$, denote Legendre polynomials 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 $\varphi$:

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

(9)

giving rise to the test defined by:

$$\psi_p^2(Z) = \sum_{i=1}^{p} u_i^2 = \sum_{i=1}^{p} \left( \sum_{k=1}^{n} \frac{h_k(z_i)}{\sqrt{n}} \right)^2 \sim \chi^2(\delta; p), \quad \delta = \sum_{i=1}^{p} \varphi_i^2,$$

$$C_1(\alpha) = \{Z: \psi_p^2(Z) > c_\alpha\}, \quad \int_{c_\alpha}^{\infty} \chi^2(p)dz.\quad (10)$$

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

$$u_i^2 = (\sum_{k=1}^{n} \frac{h_k(z_i)}{\sqrt{n}})^2, \quad i = 1, 2, ..., p,$$

(11)

provide additional information as to the direction of departure from the prespecified $F_0(.)$. Neyman (1937) was able to show that this test is locally UMP Unbiased and symmetric of size $\alpha$. He recommended $p = 5$ will be sufficient for detecting most forms of departures. This can be very convenient because the $u_i^2$ can be related directly to the sample raw moments of $Z$, e.g.

$$u_1^2 = 3(2\mu'_1 - 1)^2, \quad u_2^2 = 5(6\mu'_2 - 6\mu'_1 + 1)^2, \quad u_3^2 = 7(20\mu'_3 - 30\mu'_2 + 12\mu'_1 - 1)^2, ...$$
The importance of this modification stems from the fact that when Pearson’s chi-square test is viewed in the context of the N-P framework it becomes a special case of Neyman’s test with a limited implicit alternative. In particular, if we assume that \( N_k \) is the observed number of \( X_i \)'s in the interval \( k \), then \( N_k \) has a Binomial distribution with parameters \((p_k, n)\), \( k=1, 2, ..., p \); assuming that the assumption relating to the form of the original distribution \( f(x) \) is valid. Hence, \( E(N_k)=np_k \), giving rise to the statistic:

\[
\eta(X)=\sum_{k=1}^{p} \frac{(N_k-np_k)^2}{np_k} \frac{H_0}{\alpha} \chi^2(p-1),
\]

(12)

with an implicit local alternative of the form: \( \pi_k=p_k+(\delta_k/\sqrt{n}) \), where the \( \delta_k \)'s 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:

\[
H_0: \varphi =0, \quad \text{vs.} \quad H_1: \varphi = (\delta/\sqrt{n}),
\]

(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 statistical models. The only constraint is that this family includes the model being tested as a special case. Constructing M-S tests by postulating a broader model that nests parametrically the statistical model in question was a

\[\text{fl tricked by the next influential test discussed next.}\]

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:

\[
u(X)=[\sum_{k=2}^{n} (\hat{u}_k-\hat{u}_{k-1})^2]/[\sum_{k=1}^{n} \hat{u}_k^2] \approx 2(1-\bar{\rho}), \quad \hat{u}_k=X_k-\bar{\mu}, \quad \bar{\mu}=\frac{1}{n} \sum_{k=1}^{n} X_k.
\]

Anderson (1942) derived the distribution of:

\[
\bar{\rho}=[\sum_{k=2}^{n} (X_k-\hat{\mu})(X_{k-1}-\hat{\mu})]/[\sum_{k=2}^{n} (X_k-\hat{\mu})^2].
\]

Durbin and Watson (1950) extended these results to the Linear Regression (LR) model:

<table>
<thead>
<tr>
<th>Table 1: Linear Regression (LR) model</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y = X\beta + u )</td>
</tr>
<tr>
<td>[i] ( E(u</td>
</tr>
</tbody>
</table>

By focusing on testing the non-correlation assumption in [ii], 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 described in the following steps.

Step 1. They particularized the form of departure from the independence assumption:

\[
E(\varepsilon_t\varepsilon_s|X_t=x_t)\neq 0, \quad t > s, \quad t, s=1, 2, ..., n,
\]

(14)

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

Step 2. They nested the original model into an encompassing model (table 2).
Table 2: Autocorrelation-Corrected (A-C) LR model

\[ y = X\beta + \varepsilon, \quad \varepsilon = \rho\varepsilon_{t-1} + u, \quad |\rho| < 1, \]

[i] \( E(u|X) = 0, \) [ii] \( \text{Cov}(u|X) = \sigma_u^2 V_n, \) [iii] \( \text{Rank}(X) = k < n. \)

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

\[ H_0: \rho = 0, \quad \text{vs.} \quad H_1: \rho \neq 0, \]  

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

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

where \( \hat{\varepsilon}_t = y_t - x_t^T \hat{\beta}, \) denotes the OLS residuals, and \( \hat{\beta} = (X^T X)^{-1} X^T y \) the OLS estimator. When the observed test statistic \( D-W(y_0) \) 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. Textbook econometrics found their answer to this respecification question in Cochrane and Orcutt (1949). The answer was to adopt the alternative specification in table 2 invoked by the D-W test. When assumption [ii]* is true and \( V_n \) is known, the strategy recommends replacing \( \hat{\beta}, \) which is inefficient, with Aitken’s (1935) GLS estimator, \( \hat{\beta} = (X^T V_n^{-1} X)^{-1} X^T V_n^{-1} y, \) that is more efficient. Cochrane and Orcutt (1949) proposed a way to estimate \( \beta \) to derive a Feasible GLS (FGLS) estimator \( \hat{\beta} = (X^T \hat{V}_n^{-1} X)^{-1} X^T \hat{V}_n^{-1} y; \) see Greene (2011).

2.7 Box and Jenkins

The important breakthrough in time series analysis came with Box and Jenkins (1970) who re-invented and popularized differencing of the original data \( y_t^* \) as a way to deal with the non-stationarity of the transformed series:

\[ y_t = \Delta^d y_t^*, \]  

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

in order to achieve mean-stationarity. They proposed the ARMA(p,q) model:

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

They did not just propose a statistical model, but a modeling strategy which was viewed as an iterative process that involves several stages, identification, estimation, diagnostic checking, and prediction.

Stage 1: Identification: the choice of \( (p, d, q) \) using graphical Exploratory Data Analysis techniques, the sample autocorrelations (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: to assess the validity of the assumptions underlying the error term using the residuals.

Stage 4: A forecasting procedure based on the estimated model is constructed.

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-Jenkins 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, Stationary.

(ii) M-S Testing. Simple diagnostic checking was not effective enough to probe the validity of the model assumptions.

(iii) Respecification. There was no obvious alternative family of models one could use if the assumptions of linearity and homoskedasticity of the ARIMA model turned out to be invalid.

(iv) Differencing. \( \Delta^d y_t \) is not a universal method to achieve stationarity for a stochastic process \( \{y^*_t, \ t \in \mathbb{N}\} \). It is appropriate when the AR(p) model has d unit roots:
\[
y_t = \alpha_0 + \sum_{k=1}^{p} \alpha_k y_{t-k} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \ t \in \mathbb{N}.
\]

### 2.8 Attempts to frame/systematize M-S testing

In the statistics literature several attempts have been made to extend M-S testing for the Linear Regression model beyond autocorrelation to departures from Normality, homoskedasticity and linearity using the residuals; see Anscombe (1961) and Anscombe and Tukey (1963). Building on that, Ramsey (1969) proposed a number of M-S tests for “omitted variables, incorrect functional form, simultaneous equation problems and heteroskedasticity.” (p. 350). The paper was influential in generating further interest in the econometric literature with a view to formalize M-S testing using general testing procedures such as the Likelihood ratio, Lagrange Multiplier and Wald type tests. These attempts included the important contributions by White (1980; 1982), Domowitz and White (1982), Newey (1985) and Tauchen (1985) inter alia; see Godfrey (1988) for an comprehensive overview of these developments.

Arguably, the most concerted effort to systematize M-S testing was initiated by the LSE tradition founded by Sargan (1964) and enhanced by Jim Durbin’s contributions in this area. The LSE perspective favored a thorough probing of the model assumptions to account for all statistical information in the data and respecify if the model is misspecified; see Hendry (1980, 2003).

A strategy for M-S testing was initially formalized by Mizon (1977) and applied more broadly by other members of the LSE tradition, especially Hendry and his co-authors; see Davidson et al (1978). This also encouraged practitioners to use graphical techniques that bring out the chance regularities in the data with a view to render statistical model specification more effective; see Spanos (1999).
3 Statistical adequacy and its role in inference

Extending Fisher’s notion of a statistical model to accommodate non-random samples, takes the generic form:

\[ M_\theta(z) = \{ f(z; \theta), \ \theta \in \Theta \}, \ z \in \mathbb{R}^n_z, \]  

(17)

where \( f(z; \theta), \ z \in \mathbb{R}^n_z \) denotes the (joint) distribution of the sample \( Z := (Z_1, ..., Z_n) \), and \( \Theta, \mathbb{R}^n_z \) 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 \( z_0 := (z_1, z_2, ..., z_n) \) a typical realization thereof’. The ‘typicality’ of \( z_0 \) can – and should – be assessed using trenchant M-S testing that pertains to appraising the adequacy of the probabilistic assumptions representing the statistical model \( M_\theta(z) \) vis-a-vis data \( z_0 \); see Spanos (2006b).

In the context of statistical modeling and inference, Mis-Specification (M-S) testing refers to the formal testing procedures used to evaluate the validity of the pre-specified statistical model (17). The statistical model \( M_\theta(z) \), comprises the totality of the probabilistic assumptions imposed (directly or indirectly) on the data in question \( z_0 := (z_1, z_2, ..., z_n) \) when one estimates a model using statistical techniques. In parametric inference one can derive the sampling distribution (cdf) of any statistic \( \Phi = \gamma(Z) \) via:

\[ P(\Phi \leq \varphi) = \int \int \cdots \int_{\{z: \gamma(z) \leq \varphi\}} f(z; \theta) \, dz. \]  

(18)

A statistically misspecified \( M_\theta(z) \) would vitiate any procedure relying on \( f(z; \theta) \), (or the likelihood \( L(\theta; z_0) \)), and render all inductive inferences unreliable. Misspecification also undermines Bayesian inference via the posterior, \( \pi(\theta|z_0) \propto \pi(\theta) \cdot L(\theta; z_0) \). Similarly, it affects the reliability of nonparametric inference since it invariably relies on strong dependence and heterogeneity assumptions, as well as ‘indirect’ distributional assumptions, in the form of the existence of certain moments or/and assumptions concerning the smoothness of the unknown distribution.

Statistical inference is often viewed as the quintessential form of inductive inference: learning form a particular set of data \( z_0 \) 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:

\[ \text{if } M_\theta(z), \text{ then } Q(z), \]  

(19)

where \( M_\theta(z) \) specifies the premise of inference and \( Q(z) \) denotes the deduced inference propositions (optimal properties of estimators, tests and predictors). In deductive logic one is not particularly concerned about the soundness of \( M_\theta(z) \), but in statistical induction the soundness (validity) of the premises vis-a-vis data \( z_0 \) is of paramount importance, that 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 \( z_0 \) pertinent to the stochastic phenomenon of interest; it is the cornerstone of inductive reasoning. When any of the probabilistic assumptions defining \( M_\theta(z) \) is invalid for data \( z_0 \), \( Q(z) \) is called into question. This is primarily 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 used to calibrate the inference procedures are likely to be very different from the actual ones, rendering the inference results unreliable. Applying a .05 significance level test when the actual type I error probability is closer to .6 will give rise to untrustworthy results. Such 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 non-testable probabilistic assumptions and vague robustness results. The practitioner needs to apply thorough M-S testing. $M \sim (z)$ is said to be statistically adequate when all its probabilistic assumptions are valid for data $z_0$. 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

<table>
<thead>
<tr>
<th>Table 3 - The simple Normal model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical GM:</td>
</tr>
<tr>
<td>$X_i = \mu + u_i, \ t \in \mathbb{N}$,</td>
</tr>
<tr>
<td>[1] Normal:</td>
</tr>
<tr>
<td>$X_i \sim \mathcal{N}(\mu, \sigma)$,</td>
</tr>
<tr>
<td>[2] Constant mean:</td>
</tr>
<tr>
<td>$E(X_i) = \mu$, for all $t \in \mathbb{N}$,</td>
</tr>
<tr>
<td>[3] Constant variance:</td>
</tr>
<tr>
<td>$Var(X_i) = \sigma^2$, for all $t \in \mathbb{N}$,</td>
</tr>
<tr>
<td>[4] Independence:</td>
</tr>
<tr>
<td>${X_t, \ t \in \mathbb{N}}$ - independent process.</td>
</tr>
</tbody>
</table>

Let us focus on the simple Normal model in table 1, where we for simplicity of exposition assume that $\sigma^2$ is known. It can be shown that for testing the hypotheses:

$$H_0: \mu \leq \mu_0 \text{ vs. } H_1: \mu > \mu_0,$$

there is an $\alpha$-level UMP defined by: $T_\alpha := \{d(\mathbf{x}), C_1(\alpha)\}$:

$$d(\mathbf{x}) = \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{\sigma}, \ C_1(\alpha) = \{\mathbf{x}: d(\mathbf{x}) > c_\alpha\}. \quad (21)$$

What is often insufficiently emphasized is that the associated nominal error probabilities, i.e. the significance $\alpha$, as well as the power of test $T_\alpha$, are likely to be different from the actual error probabilities when any of the assumptions [1]-[4] are invalid for data $\mathbf{x}_0$.

To illustrate that, let us consider the case where the independence assumption [4] is false for the underlying process $\{X_t, \ t \in \mathbb{N}\}$, and instead:

$$\text{Corr}(X_i, X_j) = \rho, \ 0 < \rho < 1, \text{ for all } i \neq j, \ i, j = 1, \ldots n. \quad (22)$$
How does such a misspecification affect the reliability of test $T_\alpha$? Historically, the first to consider this question was Gosset who attempted to trace the effects of correlated samples on the distribution of the sample mean; see Student (1909).

Starting with the sampling distribution: $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$, the Normality and unbiasedness of $\overline{X}_n$ will not be affected, but its variance will be different due to the covariances ($\text{Cov}(X_i, X_j) = \rho \sigma^2$, $i \neq j$, $i, j = 1, \ldots, n$). In particular:

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

$$
d_n(\rho) = (1 + (n-1)\rho) > 1, \quad 0 < \rho < 1, \quad n > 1.
$$

This suggests that the actual sampling distribution of $\overline{X}_n$ (assuming (22)) is:

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>0</th>
<th>$0 &lt; \rho &lt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$,</td>
<td>$\overline{X}_n \sim N\left( \mu, \frac{\text{d}_n(\rho)\sigma^2}{n} \right)$.</td>
<td></td>
</tr>
</tbody>
</table>

As a result of (23), the actual distribution of $d(X)$ under $H_0$ is:

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

(24)

Hence, the actual type I error probability will be higher than the nominal since:

$$
\text{Var}(\overline{X}_n) = \frac{\text{d}_n(\rho)\sigma^2}{n} > \frac{\sigma^2}{n}, \quad \text{since} \quad \text{d}_n(\rho) > 1.
$$

Let $\alpha = .05$ ($\alpha = 1.645$), $\sigma = 1$ and $n = 100$. To find the actual type I error probability we need to evaluate the tail area of the distribution of $d^*(X)$ beyond $c_\alpha = 1.645$. In view of (24) we can deduce that:

$$
\alpha^* = P\left( d(X) > c_\alpha; H_0 \right) = P\left( Z > \frac{c_\alpha}{\sqrt{d_n(\rho)}}; \mu = \mu_0, \right),
$$

where $Z \sim N(0, 1)$. The results in Table 4 indicate that test $T_\alpha$ has now become ‘unreliable’ because its assumed type I error is different from the actual one, giving rise to unreliable inferences.

| Table 4 - Type I error of $T_\alpha$ when $\text{Corr}(X_i, X_j) = \rho$ |
|--------|-----|-----|-----|-----|-----|-----|-----|-----|
| $\rho$ | .0  | .05 | .1  | .2  | .3  | .5  | .75 | .8  | .9  |
| $\alpha^*$ | .05 | .249 | .309 | .359 | .383 | .408 | .425 | .427 | .431 |

Turning to the type II error probability, in view of (23), the actual distribution of $d(X)$ under $H_1$ is:

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

(25)

Hence, the actual power is evaluated via:

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

giving rise to the results in Table 5. The main conclusion from table 5 is that (i) for small values of $\mu_1$ (.01, .02, .05.1), the power increases as $\rho \to 1$, but for larger
values of \( \mu_1 \) (2, 3, 4), the power decreases. This undermines the ‘probativeness’ of a test. It has become like a defective smoke alarm which 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 \( T_\alpha \) when Corr(\( X_i, X_j \))=\( \rho \) |
|---|---|---|---|---|---|---|---|
| \( \rho \) | \( \pi^*(.01) \) | \( \pi^*(.02) \) | \( \pi^*(.05) \) | \( \pi^*(.1) \) | \( \pi^*(.2) \) | \( \pi^*(.3) \) | \( \pi^*(.4) \) |
| .0 | .061 | .074 | .121 | .258 | .637 | .911 | .991 |
| .05 | .262 | .276 | .318 | .395 | .557 | .710 | .832 |
| .1 | .319 | .330 | .364 | .422 | .542 | .659 | .762 |
| .3 | .390 | .397 | .418 | .453 | .525 | .596 | .664 |
| .5 | .414 | .419 | .436 | .464 | .520 | .575 | .630 |
| .8 | .431 | .436 | .449 | .471 | .515 | .560 | .603 |
| .9 | .435 | .439 | .452 | .473 | .514 | .556 | .598 |

Spanos and McGuirk (2001) use simulation to show how the reliability of inference in the context of the Linear Regression model will be completely undermined when one ignores a trending mean (non-ID), because all the MLE estimators will be inconsistent.

4 Traditional perspective on misspecification

The key reason why securing statistical adequacy of an estimated model is extremely important stems from the fact that ‘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 concerning the validation of the LR model in econometrics is aptly summarized by Hansen (1999) in the form of the following recommendation to practitioners:

‘... omit the tests of normality and conditional heteroskedasticity, and replace all conventional standard errors and covariance matrices with heteroskedasticity-robust versions.” (p. 195)

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.
(a) The first is based on invoking *generic robustness results* whose generality and applicability is often greatly overvalued.

(b) The second is that *asymptotic sampling distributions* render one’s inferences less vulnerable to statistical misspecification.

(c) The third is that using *weaker* probabilistic assumptions would render an inference less vulnerable to statistical misspecification.

(d) 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 the prespecified statistical model in question. Even in cases where some of the probabilistic assumptions are made explicit, e.g. the Linear Regression model, the list of assumptions 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 Neyman-Pearson (N-P) testing primarily because the same test procedures, 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’, ‘pre-test bias’, ‘infinite regress’, etc..

4.2 Evaluating claims discouraging model validation

Let us evaluate the pertinency of the arguments in 1. and 2. above.

4.2.1 Revisiting generic robustness claims

Box (1953) defined robustness to refer to the sensitivity of inference procedures (estimators, tests, 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:

$$Corr(X_i, X_j) \neq 0, \text{ for all } i \neq j, \ i, j = 1, ... n. \quad (26)$$

This is because one cannot evaluate the discrepancy between nominal and actual error probabilities under generic departures such as (26). 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:

(i) Exchangeable: $$Corr(X_i, X_j) = \rho, \ 0 < \rho < 1, \text{ for all } i \neq j, \ i, j = 1, ... n,$$

(ii) Markov: $$Corr(X_i, X_j) = \rho^{i-j}, \ -1 < \rho < 1, \ i \neq j, \ i, j = 1, ... n,$$

but also on the magnitude of \( \rho \); see Spanos (2009). 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, the whole robustness line of reasoning is undermined. This is because 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. For instance, comparing a Wilcoxon-type test with the t-test in the case where the true distribution is Cauchy is completely misplaced since the t-test is based on the sample mean and variance which do not exist in the case of the Cauchy distribution. More importantly, unlike natural hazards, in statistical modeling 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, one can easily use the data to distinguish between realizations of IID processes from different families of distributions with confidence using simple t-plots. 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

A strong case can be made that the traditional econometric textbook perspective encourages "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 combine generic robustness with asymptotic results based on \( n \to \infty \). For instance, the most widely invoked robustness claim in econometrics relates to dealing with departures from the homoskedasticity and no-autocorrelation assumptions (see table 1) using OLS estimators in conjunction with robust standard errors (SE), known as Heteroskedasticity/Autocorrelation Consistent (HAC) SE; see Hansen (1999), Greene (2011). It turns out, however, that the use of such robust SEs gives rise to unreliable inferences because the relevant actual error (type I and II) probabilities are very different from the nominal ones for the particular sample size \( n \) and the discrepancy does not improve as \( n \) increases; see Spanos and McGuirk (2001).

Weak assumptions can be as fallible as strong ones. What guards an inference from misspecification is the testability of the assumptions. Weaker probabilistic assumptions, such as replacing Normality with the existence of the first two moments, usually imply broader but partly non-testable 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 bigger 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 contrast, a fully parametric model with testable assumptions can guide the modeler toward a more appropriate ‘location’ using M-S testing. That is, a weak but non-testable set of probabilistic assumptions renders learning from data tantamount to a blind ‘hit or miss’ strategy. As argued by Fisher (1935):

"In inductive logic, ..., an erroneous assumption of ignorance is not innocuous; it often leads to manifest absurdities." (p. 49)

**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:

\[
\mathcal{M}_f(x) : X_k \sim \text{IID}(\mu, \sigma^2), \quad x_k \in \mathbb{R}, \quad k = 1, 2, ..., n. \tag{27}
\]

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; 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 \( \mu \), nor an effective point estimate of \( \mu \). These conclusions concerning \( \mu \) flow from the fact that \( \mu \) is sensitive to the tails of the population distribution." (Bahadur and Savage, 1956, p. 1115)
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 type 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 de-
cide that the probabilistic assumptions (i)-(iv) are appropriate for a data set $y_0$? 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.

### Table 6: AR(1) model: Phillips (1987)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>$E(u_t) = 0$, for all $t \in \mathbb{N}$,</td>
</tr>
<tr>
<td>(ii)</td>
<td>$\sup_t E[</td>
</tr>
<tr>
<td>(iii)</td>
<td>$\lim_{n \to \infty} E\left(\frac{1}{n} \sum_{t=1}^{n} u_t^2\right) = \sigma_\infty^2 &gt; 0$,</td>
</tr>
<tr>
<td>(iv)</td>
<td>${u_t, t \in \mathbb{N}}$ is strongly mixing with mixing coefficient $\alpha_m \to 0$ such that $\sum_{m=1}^{\infty} \alpha_m^{-\delta/2} &lt; \infty$.</td>
</tr>
</tbody>
</table>

### Table 7: Normal, AutoRegressive (AR(1)) Model

<table>
<thead>
<tr>
<th>Statistical GM:</th>
<th>$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t$, $t \in \mathbb{N}$,</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1] Normality:</td>
<td>$(y_t, y_{t-1}) \sim \mathcal{N}(\ldots)$,</td>
</tr>
<tr>
<td>[2] Linearity:</td>
<td>$E(y_t</td>
</tr>
<tr>
<td>[3] Homoskedasticity:</td>
<td>$\text{Var}(y_t</td>
</tr>
<tr>
<td>[4] Markov:</td>
<td>${y_t, t \in \mathbb{N}}$ is a Markov process,</td>
</tr>
<tr>
<td>[5] t-invariance:</td>
<td>$(\alpha_0, \alpha_1, \sigma_0^2)$ are not changing with $t$,</td>
</tr>
</tbody>
</table>

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

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

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 be-
comes less pernicious as $n$ increases, and thus the asymptotic sampling distributions
are more ‘robust’ than the finite sample ones. First, certain forms of misspecifi-
cation, such as the presence of heterogeneity, become more and more pernicious as
$n$ increases; see Spanos and McGuirk (2001). Second, limit theorems invoked by Consis-
tent and Asymptotically Normal (CAN) estimators and associated tests, also rely
on probabilistic assumptions, such as (i)-(iv) above, which are usually non-testable,
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
$x_0$ 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 \( n \to \infty \) relate to the 'capacity' of inference procedures to pinpoint $\mu^*$, the 'true' $\mu$, as data information accrues \( \{x_k\}_{k=1}^\infty := (x_1, x_2, \ldots, x_n, \ldots) \) approaching the limit at $\infty$. 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 $x_0$. 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 $x_0 := (x_1, \ldots, x_n)$. The latter is inextricably bound up with the particular $x_0$ and $n$.

One can just imagine an econometrician sneering at the Bahadur and Savage (1956) result by asking (tongue-in-cheek) why it did not occurred to them to use the asymptotic distribution:

$$
\tau(X) \overset{\mu=\mu_0}{\sim} N(0, 1)
$$

oblivious to the fact that such conjuration is tantamount to imposing approximate Normality. To be more specific, when the log-likelihood can be approximated by a quadratic function of $\theta:=(\mu, \sigma^2)$ [like the Normal], then (28) is likely to be accurate enough; see Geyer (2013). Supposing that the validity of (28) stems from invoking the heuristic 'as $n \to \infty$' is just an illusion. Such an invocation will not render the inference results based on $n=50$ any more trustworthy as the Bahadur and Savage result attests. The latter stems solely from the approximate validity of the probabilistic assumptions imposed on $x_0$ for a given $n$.

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 non-testable. Such indirect distributional assumptions include: (a) the existence of certain moments up to order $p$, as well as (b) smoothness restrictions on the unknown density function $f(z)$, $z \in \mathbb{R}_2$ (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 non-existence 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 non-testable 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. This
provides the cornerstone of mathematics where the ‘soundness’ of the premises is not of any particular interest. Contrary to the conventional wisdom, for statistical induction the opposite is true. 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 precise. Hence, the weakest link when using CAN estimators and related inference procedures in practice is the conjuration of limit theorems which rely on mathematically convenient but empirically non-testable 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. 435-8. Current textbook econometrics is sometimes taught as a purely deductive field. Their use of empirical illustrations aims primarily to demonstrate how to carry out the calculations. Their value as exemplars of empirical modeling that gives rise to learning from data about phenomena of interest is questionable because they are invariably based on non-validated 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 any null hypothesis (H0) will be rejected 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):

(a) The fallacy of rejection: evidence against H0 is misinterpreted as evidence for a particular alternative H1. 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 H0 does not automatically provide evidence that H1 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 H0 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 H0 provides evidence against H0 and for the presence of generic temporal dependence:

\[ E(\varepsilon_t \varepsilon_s | X_t = x_t) \neq 0, \]

but it does not provide evidence for the particular form assumed by H1:

\[ H_1: E(\varepsilon_t \varepsilon_s | X_t = x_t) = \left( \frac{t-s}{t-p} \right) \sigma_u, \quad t > s, \quad t, s = 1, 2, ..., n. \]

How one can secure evidence for (30) is discussed in section 5.3.
(b) 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 over-sensitive 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 $\alpha$ should be adjusted as $n$ changes to avoid situations where the power for a particular discrepancy is either very low or very high. Indeed, practitioners in statistics are advised to decrease the significance level as $n$ increases using certain rules of thumb; see Lehmann (1986), Good (1992). A more formal way to addresses both of the above fallacies is to use the post-data 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 $z_0$; 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 $M_\phi(z)$ there is always a statistical model $M_\theta(z)$ which pertains solely to the probabilistic assumptions imposed (often implicitly on the data $Z_0$. The crucial reason for delineating the two premises is that one needs to secure statistical adequacy of $M_\theta(z)$ first before one can reliably probe substantive adequacy of $M_\phi(z)$. 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 a 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 $Z_0$.

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 $M_\phi(z)$ 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 $Z_0$. The statistical model $M_\theta(z)$ 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 $\{Z_t, t \in \mathbb{N}\}$ underlying data $Z_0$ with a view to account for its chance regularities. The choice of the statistical model is made with a twofold objective in mind:

(a) to account for the chance regularities in data $Z_0$ by choosing a probabilistic structure for the stochastic process $\{Z_t, t \in \mathbb{N}\}$ so as to render $Z_0$ a ‘truly typical
realization’ thereof, and
(b) to parameterize \{Z_t, t \in \mathbb{N}\} in an attempt to specify \mathcal{M}_\theta(z) so as to embed (parametrically) \mathcal{M}_\varphi(z) in its context via certain restrictions, say \textbf{G}(\theta, \varphi) = 0, \theta \in \Theta, \varphi \in \Phi, relating the statistical and structural parameters.

Although this parametric nesting of the structural \mathcal{M}_\varphi(z) within the statistical model \mathcal{M}_\theta(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:

\begin{equation}
\mathcal{M}_\varphi(z) : \ 
\Gamma^\top(\varphi)y_t + \Delta^\top(\varphi)x_t = \epsilon_t, \ 
\epsilon_t \sim \text{N}(0, \Omega(\varphi)),
\end{equation}

where \(y_t\) denotes the endogenous variables, \(x_t\) the exogenous variables, and \(\varphi \in \Phi\) denotes the unknown structural parameters in the coefficient matrices \(\Gamma, \Delta\) and \(\Omega\). Corresponding to this, there is a reduced form:

\begin{equation}
\mathcal{M}_\theta(z) : \ y_t = B^\top(\theta)x_t + u_t, \ 
u_t \sim \text{N}(0, \Sigma(\theta)).
\end{equation}

What has not been sufficiently appreciated is that the reduced form in (32) is the implicit statistical model behind (31). Instead of being viewed as a derived formulation, (32) has a life of its own as a multivariate version of the LR model in table 9, specified in terms of the statistical parameters: \(\theta \in \Theta\). The two sets of parameters are related via the identifying restrictions:

\begin{equation}
\text{G}(\varphi, \theta) = 0: \ 
B(\theta)\Gamma(\varphi) + \Delta(\varphi) = 0, \ 
\Omega(\varphi) = \Gamma^\top(\varphi)\Sigma(\theta)\Gamma(\varphi), \ 
\text{for all } \varphi \in \Phi, \theta \in \Theta.
\end{equation}

Moreover, the statistical adequacy of the implicit statistical model in (32) underwrites the reliability of any inferences based on the estimated structural parameters \(\hat{\varphi}\). That is, when reduced form in the form of the LR model is statistically misspecified, any inferences based on \(\hat{\varphi}\) are likely to be unreliable; see Spanos (1990). This is because the sampling distribution (finite and asymptotic) of the estimator \(\hat{\varphi}\) – Instrumental Variables (IV) or Maximum Likelihood (ML) – depends crucially on assumptions [1]-[5]; see Phillips (1983).

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

[a] Statistical inadequacy: one or more of the probabilistic assumptions (implicitly) imposed on the data \(Z_0\) are invalid.

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

The above perspective should be contrasted with the traditional textbook approach where \(\mathcal{M}_\varphi(z)\) is estimated directly, without validating the (implicit) \(\mathcal{M}_\theta(z)\). The inevitable result is that the estimated \(\mathcal{M}_\varphi(z)\) 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 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. Indeed, in b) above, Box is advising practitioners to guard against misspecification.

The confusion between substantive and statistical misspecification is particularly pernicious when a structural model is estimated directly and the estimation is viewed as a form of curve-fitting: quantifying theoretical relationships presumed ‘true’. In such cases the theory information is treated as ‘established knowledge’ instead of tentative conjectures to be tested against the data. The end result is that the estimated model 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 Spanos (2012).

4.5 M-S testing: foundational issues

Despite the plethora of M-S tests proposed in the econometric literature testing for different individual and combinations of assumptions, 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 the plethora of such 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) the vulnerability of M-S testing to charges such as data mining, (b) multiple testing and (c) pre-test bias, and (v) how to proceed when any of the assumptions being tested are invalid.
The main thesis of this paper is that these issues can be adequately 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); see table 8.

<table>
<thead>
<tr>
<th>Table 8: Stages of empirical modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Specification</td>
</tr>
<tr>
<td>2. Estimation</td>
</tr>
<tr>
<td>3. Mis-Specification (M-S) Testing</td>
</tr>
<tr>
<td>4. Respecification</td>
</tr>
<tr>
<td>• Statistically adequate model</td>
</tr>
<tr>
<td>5. Inference</td>
</tr>
</tbody>
</table>

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 $z_0$, but retain the parametric nesting of $M_{\varphi}(z)$ 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 ‘pre-test 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 $M_\theta(z)$ is valid for data $z_0$, i.e. $M_\theta(z)$ could have generated data $z_0$.

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 pre-test 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 $F=\{\mathcal{M}_{\varphi_i}(z), i=1, \ldots, m\}$, and then select a best model $\mathcal{M}_{\varphi_i}(z)$ 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 $F$ 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 $F$ (Lehmann, 1990), and (ii) their ranking
of the different models in $F$ is equivalent to N-P testing comparisons among them 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 $M_\theta(z)$ and places statistical modeling in a broader framework which allows the \textit{ab initio} separation of the statistical (chance regularities) and substantive information in the form of a structural model $M_\varphi(z)$, 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 problem of specification is viewed in the context of the set of all possible models that could have generated data $z_0$, say $P(z)$. The statistical model $M_\theta(z)$ is specified by partitioning $P(z)$ using probabilistic assumptions from three broad categories: Distribution, Dependence and Heterogeneity. Viewing $M_\theta(z)$ as a particular parameterization of the observable vector stochastic process $\{Z_t, t\in\mathbb{N}\}$ underlying the data $Z_0$, has several advantages, including (i) providing a complete and internally consistent set of testable probabilistic assumptions, (ii) providing a well-defined parameterization for the model parameters, (iii) bringing out the interrelationship among the probabilistic assumptions, and (iv) enabling the modeler to distinguish between the statistical and the substantive premises of inference.

This 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 ....” Simple t-plots and scatter plots are often sufficient to enable a practitioner to detect a variety of chance regularity patterns that can be accounted for by selecting the appropriate probabilistic assumptions from the above three categories for $\{Z_t, t\in\mathbb{N}\}$. The selection aims to render data $Z_0$ a typical realization of $\{Z_t, t\in\mathbb{N}\}$; see Spanos (1999), chs 5-6. Naturally, the appropriateness of the selected probabilistic assumptions will be formally appraised at the M-S testing stage of modeling.

\textbf{Example 1.} Consider the specification procedure for the AR(1) model (table 7). The t-plot of the data in fig. 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 fig. 3, which exhibit bell-shape symmetry associated with a Normal distribution. No indications of departures from IID, suggest that data $y_0$ could be viewed as a typical realization of $\{Z_t, t\in\mathbb{N}\}$; see Spanos (1999), chs 5-6. Naturally, the appropriateness of the selected probabilistic assumptions will be formally appraised at the M-S testing stage of modeling.
\[
M = f(y_1; \varphi_1) \prod_{t=2}^{n} f(y_t|y_{t-1}; \varphi_t), \quad \forall y \in \mathbb{R}^n \tag{34}
\]

This reduction gives rise to the AR(1) in table 7, together with the parameterization of \((\alpha_0, \alpha_1, \sigma_0^2)\) in the last row.

**Example 2.** The same procedure can be followed to specify the Linear Regression (LR) model assuming that \(\{Z_t:=(y_t, X_t), \ t \in \mathbb{N}\}\) is Normal, IID, giving rise to the LR model in table 9:

\[
D(Z_1, ..., Z_n; \varphi) \overset{iid}{=} \prod_{t=1}^{n} D(y_t; X_t; \psi_1) = \prod_{t=1}^{n} D(y_t|X_t; \psi_1), \quad \forall Z_t \in \mathbb{R}^{m \times n}. \tag{35}
\]

The LR model is specified exclusively in terms of the conditional distribution \(D(y_t|X_t; \psi_1)\), due to weak exogeneity; see Engle et al (1983). To be more specific, the LR model comprises the statistical Generating Mechanism (GM) in conjunction with the probabilistic assumptions [1]-[5] (table 9). It is important to emphasize for statistical adequacy purposes only the validity of assumptions [1]-[5] is relevant. That is, if assumptions [1]-[5] are valid for data \(Z_0:=(y, X)\), the statistical procedures associated with the LR model have the optimality properties attributed to them. These assumptions, with the exception of [5], are directly related to those based on the error term in table 1, and they constitute the statistical premises of inference.

**Table 9: Linear Regression Model**

<table>
<thead>
<tr>
<th>Statistical GM: ( y_t = \beta_0 + \beta_1^\top x_t + u_t, \ t \in \mathbb{N} ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1] Normality: ( (y_t</td>
</tr>
<tr>
<td>[2] Linearity: ( E(y_t</td>
</tr>
<tr>
<td>[3] Homoskedasticity: ( \text{Var}(y_t</td>
</tr>
<tr>
<td>[4] Independence: ( {y_t</td>
</tr>
<tr>
<td>[5] t-invariance: ( (\beta_0, \beta_1, \sigma^2) ) are not changing with ( t ),</td>
</tr>
<tr>
<td>( \beta_0 = E(y_t) - \beta_1^\top E(X_t), \ \beta_1 = \text{Cov}(X_t)</td>
</tr>
<tr>
<td>\sigma^2 = \text{Var}(y_t) - \text{Cov}(X_t, y_t)^\top \text{Cov}(X_t)</td>
</tr>
<tr>
<td>( t \in \mathbb{N} ).</td>
</tr>
</tbody>
</table>
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 \( Z_0 \) 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 \( \mu_t = E(y_t|D_t) \) and non-systematic \( u_t = y_t - E(y_t|D_t) \) component with respect to the conditioning information set \( D_t \subset \mathcal{F} \), to define:

\[
y_t = E(y_t|D_t) + u_t, \quad t \in \mathbb{N}. \tag{36}
\]

\( D_t \) is selected with a view to capture all the systematic information in \( Z_0 \) rendering the ‘duced’ \( u_t = y_t - E(y_t|D_t) \) non-systematic. More formally \( u_t \) constitutes a 2nd order Martingale Difference (MD) process, i.e.

\[
\begin{align*}
(1) \quad & E(u_t|D_t) = 0, \\
(2) \quad & E(\mu_t u_t|D_t) = 0, \\
(3) \quad & E(u_t^2|D_t) = \text{Var}(y_t|D_t), \quad t \in \mathbb{N}. \tag{37}
\end{align*}
\]

In practice, \( D_t \) is commonly generated by observable random variables defined on \( (S, \mathcal{F}, \mathbb{P}(.)) \). As shown in White (1999), p. 59-60, \( D_t = \sigma(X_t, Z_{t-1}, Z_{t-2}, ..., Z_t) \) is the generic choice of the relevant \( D_t \) when no dependence and heterogeneity assumptions (restrictions) are imposed on the observable process \( \{Z_t := (y_t, X_t), \quad t \in \mathbb{N}\} \) that would render \( \{(u_t|D_t), \quad t \in \mathbb{N}\} \) a MD process.

In the case of the LR model, the independence and Normality assumptions narrow down the appropriate choice to \( D_t = (X_t = x_t) \). In this sense, the statistical error term \( u_t \) is viewed as:

[i] derived, denoting the non-systematic statistical information in \( Z_0 \) relative to \( \mu_t \),

[ii] local, in the sense that it pertains to the statistical model \( M_0(z) \) vis-a-vis the data \( Z_0 \). That is, the appraisal of concern is how adequately \( M_0(z) \) accounts for the chance regularities in \( Z_0 \).

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

(a) statistical information in data \( Z_0 \), in terms of which \( D_t \) is defined, such as temporal dependence, generically represented with lags \( Z_{t-i} := (y_{t-i}, X_{t-i}), \quad i = 1, ..., p \) and heterogeneity, generically represented by dummies and trends (based on the ordering \( t=1, ..., n \)), and

(b) substantive information beyond the data \( Z_0 \), such as omitted but relevant variables, errors of measurement, simultaneity problems, etc.

This distinction indicates clearly that \( Z_{t-i} \) is not an omitted set of variables in the same sense that additional explanatory variables not in \( X_t \) are, since terms such as lags \( Z_{t-i} \), and trend terms \( t, t^2 \) are already in the current statistical universe of discourse \( \sigma(Z_1, Z_2, ..., Z_n) \).

In contrast to the statistical, the structural error term \( \varepsilon_t \) in (31) is viewed as:

[i]* autonomous, potentially representing errors of approximation, omitted effects, errors of measurement, shocks etc., and

30
5.1 M-S testing

The primary objective of M-S testing is to probe within \( \mathcal{P}(z) \)-all possible statistical models that could have given rise to \( Z_0 \), for potential departures from the assumptions defining \( \mathcal{M}_\theta(z) = \{ f(z; \theta), \theta \in \Theta \}, \ z \in \mathbb{R}^n \). The generic null and alternative hypotheses take the form:

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

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

\[
H_0: \theta_0 \in \Theta_0 \quad \text{vs.} \quad H_1: \theta_0 \in \Theta_1,
\]

where \( (\Theta_1, \Theta_0) \) constitute a partition of \( \Theta \), the probing is within the boundaries of \( \mathcal{M}_\theta(z) \), since (38) can be equivalently expressed in the form:

\[
H_0: f^*(x) \in \mathcal{M}_0(x) = \{ f(x; \theta_0), \ \theta_0 \in \Theta_0 \} \quad \text{vs.} \quad H_1: f^*(x) \in \mathcal{M}_1(x) = \{ f(x; \theta_0), \ \theta_0 \in \Theta_1 \}.
\]

Another crucial difference is that for M-S testing one needs to operationalize \([\mathcal{P}(x) - \mathcal{M}_\theta(x)]\) 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. Nonparametric (omnibus) tests, such as the runs test, the Pearson and the Kolmogorov tests, are usually non-directional 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 \( \mathcal{M}_\theta(z) \) into a broader model \( \mathcal{M}_\psi(z) \) and testing the nesting restrictions, as in the case of the Durbin-Watson test. This has the appearance of a N-P test, but such an interpretation is unwarranted because there is no reason to believe that \( \mathcal{M}_\psi(z) \) 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.
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 \((S, \mathcal{F}, \mathbb{P}(\cdot))\) with bounded variance, including \(Z:= (y, X)\) (a \(m \times 1\) vector) such that \(E\left(\mid Z \mid^2\right) < \infty\). For any \(D \subset \mathcal{F}\) and any random variable \(\xi\) relative to \(D\) (Williams, 1991):

\[
E(y-\xi)^2 = E(y-E(y|D))^2 + E[E(y|D)-\xi]^2.
\]

(39)

This implies that \(E(y-\xi)^2\) is minimized when \(\xi^* = E(y|D)\). In the case where \(D = \sigma(X)\)-the \(\sigma\)-field generated by \(X\), i.e. \(\xi^* = E(y|\sigma(X))\). The minimization in (39) follows from the orthogonality between \(u = y - E(y|\sigma(X))\) and any random variable with respect to \(\sigma(X)\) which can take the form of any Borel function of \(X\) (Doob, 1953):

\[
E\left(\mid y - E(y|\sigma(X)) \mid h(X)\right) = 0, \text{ for any Borel-function } h(X).
\]

(40)

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

\[
E\left(\mid y - E(y|\sigma(X)) \mid h(X)\right) = 0, \text{ for all Borel-functions } h(X), \ t \in \mathbb{N},
\]

(41)

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

\[
E(u_t \cdot h(X_t)) = 0, \ t \in \mathbb{N}.
\]

(42)

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

\[
E\left(\mid u_t^2 - g_2(X_t)\mid h(X_t)\right) = 0, \text{ for all Borel-functions } h(X), \ t \in \mathbb{N},
\]

(43)

if and only if \(g_2(X_t) = E(u_t^2|\sigma(X_t))\), \(t \in \mathbb{N}\).
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:

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

\( D_t \) is operationally legitimate for M-S testing purposes if \( D_t \) is a proper subset of the statistical universe of discourse \( \sigma (Z_1, Z_2, ..., Z_n) \). Of particular interest is the choice \( D_t = \sigma (X_t, Z_{t-1}, Z_{t-2}, ..., Z_1) \), that would render \( \{ u_t | D_t \}, \ t \in \mathbb{N} \) a 2nd order MD process without imposing independence.

Such potential non-orthogonalities can be framed in terms of using auxiliary regressions of the form:

\[
u_t = \delta_1 + \gamma_1^\top h_1(D_t) + v_{1t}, \quad u_t^2 = \delta_2 + \gamma_2^\top h_2(D_t) + v_{2t}, \quad t = 1, 2, ..., n,
\]

where \( h_r(D_t), \ r = 1, 2, \) denote 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 (44) amounts to probing for departures from the process \( \{ u_t | D_t \}, \ t \in \mathbb{N} \) being a Normal, Martingale Difference 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 \( (Z_t := (y_t, X_t), \ t = 1, 2, ..., n) \) that can be used to define potential statistical information not accounted for by the original model, say \( D_t = (\psi_t, z_{t-1}, t) \):

\[
\psi_t := (x_{it} - x_{jt})_{i, j}, \ i, j = 2, ..., k, \quad z_{t-1} := (y_{t-1}, x_{t-1}), \quad t := (t, t^2, ..., t^p).
\]

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

\[
E(y_t | D_t) = \alpha_0 + \alpha_1^\top x_t + \alpha_2^\top \psi_t + \alpha_3^\top z_{t-1} + \delta^\top t.
\]

As mentioned above, \((\psi_t, z_{t-1}, t)\) are not ‘omitted’ explanatory variables in the traditional sense used in textbook econometrics (Greene, 2011) because they are already in the current statistical information set: \( (Z_1, Z_2, ..., Z_n) \). This respecified regression function can be compared with the original \( E(y_t | X_t = x_t) = \beta_0 + \beta_1^\top x_t \) to construct an auxiliary regression for testing departures from the assumptions [2], [4] and [5]:

\[
(GM1): \ y_t = \beta_0 + \beta_1^\top x_t + u_t, \quad (GM2): \ y_t = \alpha_0 + \alpha_1^\top x_t + \alpha_2^\top \psi_t + \alpha_3^\top z_{t-1} + \delta^\top t + v_{1t},
\]

Subtracting the two gives rise to the auxiliary regression:

\[
u_t = (\alpha_0 - \beta_0) + (\alpha_1 - \beta_1)^\top x_t + \alpha_2^\top \psi_t + \alpha_3^\top z_{t-1} + \delta^\top t + v_{1t}.
\]

This can be rendered operational by replacing the errors with the residuals to yield:

\[
\tilde{u}_t = (\alpha_0 - \beta_0) + (\alpha_1 - \beta_1)^\top x_t + \alpha_2^\top \psi_t + \alpha_3^\top z_{t-1} + \delta^\top t + v_{1t}.
\]

It is important to emphasize that this auxiliary regression can be easily extended/modified to include higher order powers of \( x_t \), 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 \( \hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t \) and the residuals \( \hat{u}_{t-i} \), in place of higher order functions of \( x_t \) and lags of \( Z_t \), respectively. The F-type tests for the joint hypotheses:

\[
H_0: \alpha_2 = 0, \quad \alpha_3 = 0, \quad \delta = 0, \quad \text{vs.} \quad H_1: \alpha_2 \neq 0 \quad \text{or} \quad \alpha_3 \neq 0 \quad \text{or} \quad \delta \neq 0,
\]

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

\[
\hat{u}_t^2 = \gamma_0 + \alpha_2 + \alpha_3 z_{t-1}^2 + \delta^T t + v_{2t},
\]

where \( z_{t-1}^2 \) denotes quadratic functions of \( z_{t-1} \), and the joint hypotheses are:

\[
H_0: \alpha_2 = 0, \quad \alpha_3 = 0, \quad \delta = 0, \quad \text{vs.} \quad H_1: \alpha_2 \neq 0 \quad \text{or} \quad \alpha_3 \neq 0 \quad \text{or} \quad \delta \neq 0.
\]

It is important to emphasize that (50) distinguishes between testing for heteroskedasticity \( \text{Var}(y_t \mid X_t=x_t) = \sigma^2 \), \( x_t \in \mathbb{R}^k \), and for conditional variance heterogeneity \( \text{Var}(y_t \mid X_t=x_t) = g(t), \ t \in \mathbb{N}, \ x_t \in \mathbb{R}^k \); the latter can be trends or shifts in \( \text{Var}(y_t \mid X_t=x_t) \). This reveals how misleading the traditional strategy of using HCSE (Hansen, 1999) and the Heteroskedasticity and Autocorrelation Consistent (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, but in practice there are many variations/extensions one might consider, including powers of the fitted values and lags of the residuals, 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 \( \{ (u_t \mid X_t=x_t), \ t \in \mathbb{N} \} \) is a Normal, MD process.

The only assumption that these auxiliary regressions (48) and (50) do not test for [1] Normality, and there are good reasons for that. The available M-S tests for [1] assume that the other assumptions are valid, rendering the results questionable when any of the other assumptions [2]-[5] is 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 apply to other statistical models of interest in econometrics, including models for cross-section and panel data; see Spanos (2012).

### 5.2 Empirical example

Lai and Xing (2008), pp. 72-81, illustrate the Capital Asset Pricing Model (CAPM) using monthly data for the period Aug. 2000 to Oct. 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:

\[
y_t = .02 + 1.996 x_t + \hat{u}_t, \quad R^2 = .536, \quad s = .0498, \quad n = 64,
\]

34
where the standard errors are given in parentheses. The authors proceed to test the significance of the regression coefficients ($\beta_0, \beta_1$) using t-tests, and conclude that for $\alpha=.025$ the coefficient $\beta_0$ is statistically insignificant but $\beta_1$ is significant, providing evidence for the CAPM.

The above inference results will be trustworthy when the Linear Regression (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 $\{(y_t, x_t), \ t=1, 2, \ldots, n\}$ (fig. 4-5), suggests that assumptions [4]-[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 fig. 6.

These misspecifications are confirmed by the following auxiliary regressions:

$$\hat{u}_t = .02 + .370 xt + .091 t + .253 t^2 + .172 t^3 - .343 \hat{u}_{t-1} + \hat{v}_t,$$

$$R^2=.197, \ s=.0463, \ n=63,$$

$$\hat{u}_t^2 = .002 - .002 t + .141 \hat{y}_t^2 + \hat{v}_t^2,$$

$$R^2=.2, \ s=.0037, \ n=64$$

The above M-S testing results indicate clearly that no reliable inferences can be drawn on the basis of the estimated model in (52) since assumptions [3]-[5] are invalid. Note that the use of HAC SEs cannot address these misspecifications; a proper respecification would lead to a Student’s t DLR model; see Heracleous and Spanos (2006).

Fig. 4: t-plot of Intel excess returns

Fig. 5: t-plot of market excess returns

These specifications are confirmed by the following auxiliary regressions:

$$\hat{u}_t = .02 + .370 xt + .091 t + .253 t^2 + .172 t^3 - .343 \hat{u}_{t-1} + \hat{v}_t,$$

$$R^2=.197, \ s=.0463, \ n=63,$$

$$\hat{u}_t^2 = .002 - .002 t + .141 \hat{y}_t^2 + \hat{v}_t^2,$$

$$R^2=.2, \ s=.0037, \ n=64$$

The above M-S testing results indicate clearly that no reliable inferences can be drawn on the basis of the estimated model in (52) since assumptions [3]-[5] are invalid. Note that the use of HAC SEs cannot address these misspecifications; a proper respecification would lead to a Student’s t DLR model; see Heracleous and Spanos (2006).

Fig. 6: t-plot of the residuals from (52)

Fig. 7: General Motors excess returns
To illustrate the perils of posing substantive questions of interest on a misspecified model, consider an omitted variables question: is last period’s excess returns of General Motors \((z_{t-1})\) (see fig.7) an omitted variable in (52)? Adding \(z_{t-1}\) gives rise to:

\[
y_t = .013 + 2.086x_t - .206z_{t-1} + \tilde{\epsilon}_t, \quad R^2 = .577, \quad s = .0483, \quad n = 64,
\]

which suggests that \(z_{t-1}\) is a relevant omitted variable. Of course, any variable which picks up the unmodeled information, will misleadingly appear to be statistically significant. Indeed, a simple respecification of the original model, such as adding trends and a lag to account for the indicated departures:

\[
y_t = .035 + 2.319x_t + .086t + .190t^2 + 1.58t^3 - .316y_{t-1} + .524x_{t-1} - .164z_{t-1} + \tilde{\epsilon}_t,
\]

renders \(z_{t-1}\) statistically insignificant. For a more detailed empirical example on how the PR modeling strategy is applied in practice, 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 \(P(z)\). 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, could guide the respecification by re-partitioning \(P(z)\) with a view to specify a new statistical model that accounts for the statistical information the original model did not.

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

\[
E(y_t | X_t=x_t) = h(x_t), \quad Var(y_t | X_t=x_t) = g(x_t), \quad x_t \in \mathbb{R}^k_x,
\]

where the functional forms \(h(.)\) and \(g(.)\), and the relevant parameterization \(\theta\) stem from the joint distribution \(D(y, X_t; \varphi)\). 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; see Spanos (1995).

In contrast, textbook econometrics respecification usually takes the form of ad hoc ‘error-fixing’ strategies that often involve adopting the particular alternative \(H_1\) used in the M-S test applied; a classic case of the fallacy of rejection. For instance, the error process \(\{(u_t | X_t=x_t), \ t \in \mathbb{N}\}\) provides a much narrower perspective on M-S testing because, by definition \(u_t = y_t - \beta_0 - \beta_1^\top x_t\), and thus the error process retains the original systematic component \(E(y_t | X_t=x_t) = \beta_0 + \beta_1^\top x\) (table 1). As shown by Sargan (1964), the respecification the LR model (table 1) by adopting the AC-LR model (table 2), associated with the Durbin-Watson test:

\[
y_t = \beta_0 + \beta_1^\top x + u_t, \quad u_t = \rho u_{t-1} + \tilde{\epsilon}_t \quad \rightarrow \quad y_t = \beta_0 (1-\rho) + \beta_1^\top x_t + \rho y_{t-1} - \rho \beta_1^\top x_t + \tilde{\epsilon}_t,
\]
gives rise to a restricted Dynamic Linear Regression (DLR) model:

\[ y_t = E(y_t \mid X_t, Z_{t-1}) + \alpha_0 + \alpha_1 x_t + \alpha_2 y_{t-1} + \alpha_3 x_{t-1} + v_t. \]  

(56)

The DLR model would naturally arise from the PR respecification by replacing Independence with Markov dependence for the observable process \( \{Z_t := (y_t, X_t), t \in \mathbb{N}\} \). The AC-LR model is nested within (56) by imposing \( k \) common factor restrictions: \( \alpha_3 + \alpha_1 \alpha_2 = 0 \). These restrictions are not as innocuous as it appears at first sight because they impose highly unappetizing restrictions on the temporal structure of \( \{Z_t := (y_t, X_t), t \in \mathbb{N}\} \) (McGuirk and Spanos, 2009). Worse, the OLS estimators \((\hat{\beta}, s^2)\) will be both biased and inconsistent unless the common factor restrictions are valid; see Spanos (1986).

6 Summary and conclusions

The primary objective of empirical modeling is ‘to learn from data \( Z_0 \)’ about observable phenomena of interest using a statistical model \( M_{\theta}(z) \) as the link between theory and data. Substantive subject matter information, codified in the form of a structural model \( M_{\varphi}(z) \), plays an important role in demarcating and enhancing this learning from data when it does not belie the statistical information in \( Z_0 \). Behind every structural model \( M_{\varphi}(z) \) there is a statistical model \( M_{\theta}(z) \) which comprises the probabilistic assumptions imposed on one’s data, directly or indirectly. An effective way to unveil \( M_{\theta}(z) \) is to view it as providing a convenient a summary of the statistical information in data \( Z_0 \), with convenience referring to selecting a parameterization that nests \( M_{\varphi}(z) \) via generic restrictions of the form \( G(\theta, \varphi) = 0 \).

The paper makes a case that what is needed for improving empirical modeling is a broadening of Fisher’s (1922) framework that separates the modeling from the inference facets with a view to bring out the different potential errors and use strategies to safeguard against them. The statistical misspecification of \( M_{\theta}(z) \) 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 arguments invoking \( n \to \infty \), will not circumvent this error in practice. The brief historical overview of M-S testing, going back to Pearson (1900), brings out several foundational issues that can be addressed in the proposed PR framework. Establishing the adequacy of \( M_{\theta}(z) \) calls for a thorough M-S testing, combined with a coherent respecification strategy that relies on assumptions imposed on \( \{Z_t, t \in \mathbb{N}\} \). Joint M-S tests based on auxiliary regressions provide a most effective tool for detecting departures from the model assumptions. When any assumptions are found wanting, the original \( M_{\theta}(z) \) needs to be respecified with a view to account for all the systematic information in the data. The traditional respecification of adopting the particular alternative \( H_1 \) used for the M-S test constitutes an example of the fallacy of rejection.

Distinguishing between statistical and substantive inadequacy is crucial because it is one thing to argue that a structural model \( M_{\varphi}(z) \) is only an approximation of the reality it aims to explain, and entirely another to claim that the assumed
statistical model \( M_\theta(z) \) could not have generated data \( Z_0 \). Hence, \( M_\varphi(z) \) may always come up short in securing substantive adequacy vis-a-vis the phenomenon of interest, but \( M_\theta(z) \) may be perfectly adequate for answering substantive questions of interest. \( M_\varphi(z) \) does not belie the data when: (a) the (implicit) \( M_\theta(z) \) is statistically adequate, and (b) the overidentifying restrictions \( G(\varphi, \theta) = 0 \) are data-acceptable. Although not contradicting the data is necessary for rendering \( M_\varphi(z) \) substantively adequate, it is not sufficient. Further investigating is needed to establish that \( M_\varphi(z) \) adequately captures (describes, explains, predicts) the phenomenon of interest. For that probing, the statistical adequacy of \( M_\theta(z) \) is of paramount importance to ensure the reliability of the inference procedures used.

**References**


[50] Neyman, J. and E.S. Pearson (1933) “On the problem of the most efficient tests of statistical hypotheses”, *Phil. Trans. of the Royal Society, A*, 231, 289-337.


[54] 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.


