Learning from Data: the Role of Error in Statistical Modeling and Inference*

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

October 2012

Abstract

A widely held view in economics is that the current untrustworthiness of empirical evidence and the inability to forecast economic phenomena is largely due to the fact that the economy is too complicated and the resulting data are too heterogeneous to be amenable to statistical modeling. The primary objective of this paper is to argue that whether economic data are amenable to statistical modeling is largely an empirical issue and the current untrustworthiness is mainly due to an inadequate implementation of statistical modeling and inference. The latter stems from two main sources: (i) commingling the statistical and substantive premises of inference, and (ii) neglecting the validation of the latter the secures the reliability of inference. It is argued that error statistics, a refined/extended frequentist perspective, attains learning from data about phenomena of interest by employing reliable and effective procedures to establish trustworthy evidence.

†I would like to thank Deborah Mayo for numerous creative discussions on several issues discussed in this paper.
1 Introduction

‘Learning from data’ has a long and questionable history in economics going back to the 17th century. Despite the fact that Statistics and Economics share common roots going back to ‘Political Arithmetic’ in the mid-17th century, the relationship between them has been one of mutual distrust and uneasy collaboration. They developed into separate disciplines after the demise of Political Arithmetic in the early 18th century, due mainly to:

(a) the inability of political arithmeticians to distinguish between real regularities and artifacts in observed data, combined with
(b) the irresistible proclivity to misuse data analysis in an attempt to make a case for one’s favorite policies.

These indulgences led political arithmeticians to exorbitant speculations and claims based on abusing the information in the data to such an extent that eventually discredited these data-based methods beyond salvation; see Spanos (2008).

Political economy (Economics) was separated from Political Arithmetic, and established itself as the first social science at the end of the 18th century (Blaug, 1997), partly as a result of attributing a subordinate role to the data and emphasizing the role of theory, causes and explanations. Political economists made their case by contrasting their primarily deductive methods to the discredited data-based (inductive) methods utilized by political arithmeticians.

In the early 19th century Statistics emerged as a ‘cleansed’ version of Political Arithmetic by shaking off the discredited data-based inferential component and promising to focus exclusively on the collection and tabulation of data. The idea was that statistics could not be a separate discipline, but had to rely on other disciplines for any inferences drawn from data. The founding document the Statistical Society of London in 1834 separated statistics from political economy as follows:

“The Science of Statistics differs from Political Economy because although it has the same end in view, it does not discuss causes, nor reason upon probable effects; it seeks only to collect, arrange, and compare, that class of facts which alone can form the basis of correct conclusions with respect to social and political government.” (JSSL, 1834, p. 2).

Mill’s thesis. Mill (1844) was the first philosopher to articulate the emphasis on deductively derived propositions and attributing a subordinate role to data, by proposing a methodological framework to accommodate Ricardo’s (1817) clear deductive perspective. He argued that causal mechanisms underlying economic phenomena are too complicated – they involve too many contributing factors – to be disentangled using the available observational data. This is in contrast to physical phenomena whose underlying causal mechanisms are not as complicated – they involve only a few dominating factors – and the use of experimental data can help to untangle them by ‘controlling’ the ‘disturbing’ factors. Hence, economic theories could not establish precise enough implications of theories. At best, one can demonstrate general tendencies whose validity can be assessed using observational data. These tendencies are framed in terms of the primary causal contributing factors with the rest of the
numerous (possible) disturbing factors relegated to *ceteris paribus* clauses whose appropriateness cannot, in general, be assessed using observational data. This means that any empirical evidence that contradict the implications of a theory can always be explained away as due to counteracting disturbing factors. As a result, Mill (1844) attributed to statistics and data analysis the auxiliary role of investigating the *ceteris paribus* clauses in order to shed light on the unaccounted by the theory disturbing factors which prevent the establishment of the tendencies predicted by the theory in question; see Hausman (1992) for a modern interpretation of Mill’s thesis.

**Keynes’s thesis.** More recently, the view that economic phenomena are too complicated and heterogeneous to be amenable to traditional statistical methods using observational data was re-iterated by Keynes (1939) in his criticisms of Tinbergen’s (1939) work on the business cycle. Among other less important issues and a number confusions, Keynes raised several foundational problems associated with the use of linear regression in learning from data (Boumans and Davis, 2010; Hendry and Morgan, 1995; Morgan, 1990), including:

(i) the need to account for all the relevant contributing factors (including non-measurable ones), at the outset,

(ii) the conflict between observational data and the *ceteris paribus* clauses underlying economic theories,

(iii) the spatial and temporal heterogeneity of economic phenomena stemming from complex interactions among numerous economic agents,

(iv) the validity of the assumed functional forms of economic relations, and

(v) the ad hoc specification of the lags and trends in economic relations.

In this paper it is to argue that both standpoints articulated by Mill and Keynes stem from inadequate understanding of the nature and applicability of statistical induction as it relates to learning from data about phenomena of interest.

Keynes’s criticism reminds one of an earlier charge by Robbins (1935) — expressing a widely-held view among economists — who dismissed the use of statistics for theory appraisal in economics. Their argument was that:

(vi) statistical techniques are only applicable to data which can be considered as ‘random samples’ from a particular population, i.e. Independent and Identically Distributed (IID) random variables; see Frisch (1934), p. 6. In light of that, statistical analysis of economic data had no role to play in theory assessment.

The primary objective of this paper is to argue that the claims (i)-(vi) are largely misinformed for two main reasons.

First, these claims largely ignore two important developments in the 1930s:

(a) the new frequentist statistical approach pioneered by Fisher, Neyman and Pearson, and

(b) the theory of stochastic processes that greatly extended the intended scope of statistical modeling to include data that exhibit both heterogeneity and dependence.

Second, these claims conflate two different dimensions of empirical modeling: the *statistical* and the *substantive premises.*
In particular, issues (i)-(ii) pertain to the substantive adequacy of the estimated model, but issues (iii)-(vi) pertain to its statistical adequacy.

Section 2 provides a summary of the frequentist approach to inference with a view to shed light on its objective and underlying reasoning in learning from data, as well as bring out its applicability to non-IID data. Section 3 discusses the Pre-Eminence of Theory perspective that has dominated empirical modeling in economics since Ricardo (1817), in an attempt to highlight its key weaknesses: [i] conflating the statistical with the substantive premises of inference, and [ii] neglecting the validation of the latter would secure the reliability of inference (Mayo and Spanos, 2004), in an attempt to motivate the need for a broader methodological framework that allows for the data to play a more prevalent role. The error statistical framework (Mayo, 1996) is presented in section 4 as a refinement/extension of the Fisher-Neyman-Pearson approach, that sheds light on the issues raised in (i)-(vi) by addressing the key weaknesses [i]-[ii].

2 The frequentist approach to statistical inference

Fisher (1922) pioneered modern frequentist statistics as a model-based approach to statistical induction anchored on the notion of a statistical model:

\[ M_\theta(z) = \left\{ f(z; \theta), \theta \in \Theta \right\}, \ z \in \mathbb{R}^n_z, \ \Theta \subset \mathbb{R}^m, \ m < n, \]

where \( f(z; \theta) \) is the joint distribution of the sample \( Z := (Z_1, ..., Z_n) \). The mathematical apparatus of frequentist statistical inference was largely in place by the late 1930s. Fisher (1922; 1925;1934) largely created the current theory of ‘optimal’ point estimation and formalized significance testing based on the p-value reasoning. Neyman and Pearson (1933) proposed an ‘optimal’ theory for hypothesis testing, by modifying/extending Fisher’s significance testing. Neyman (1937) proposed an ‘optimal’ theory for interval estimation analogous to N-P testing.

At this stage it important to note that the transition from the early 20th century descriptive statistics paradigm, relying on vague allusions to large samples, to model-based statistical inference, primarily in the hands of Fisher and Neyman and Pearson, relying on a given finite \( n > 1 \), was neither apparent nor without its polemists. It’s no accident that several statistics textbooks that were influential in economics, like Bowley (1937), Mills (1938) and Allen (1949), missed the change of paradigms altogether. In particular, the overwhelming majority of their chapters revolved around descriptive techniques based on a vague frequentist interpretation of probability, but without a clear probabilistic foundation, akin to the Karl Pearson and not the F-N-P perspective. For instance, correlation and regression are viewed from a curve-fitting perspective instead of being based on a well-defined statistical model (regression) with clearly specified probabilistic assumptions.

2.1 Broadening the scope of statistical modeling

Specifying statistical models in terms the joint distributions \( f(z; \theta) \) of a stochastic process \( \{ Z_t, t \in \mathbb{N} \} \) enables one to account for different forms of dependence and
heterogeneity in data $Z_0$, by viewing the latter as a ‘typical’ realization of the stochastic process $\{Z_t, \ t \in \mathbb{N}\}$ underlying $\mathcal{M}_\theta(z)$; see Kolmogorov (1933). Whether economic data exhibit sufficient invariance over $t \in \mathbb{N}$ to be amenable to statistical modeling and inference is an empirical issue that pertains to the validity of the statistical premises $\mathcal{M}_\theta(z)$. That is, the only restriction on the heterogeneity/dependence in the data in question is that they exhibit a sufficient $t$-invariance that encapsulates the recurring features of the phenomenon being modeled in the form of constant parameters $\theta$.

The theory of stochastic processes, developed in the 1930s and 1940s, has extended considerably the intended scope of statistical modeling. This demarcation comes in the form of limit theorems [Law of Large Numbers and Central Limit] specifying sufficient probabilistic restrictions (distribution, dependence and heterogeneity) on $\{Z_t, \ t \in \mathbb{N}\}$ under which ‘learning from data’ is potentially possible. The earlier limit theorems holding when $\{Z_t, \ t \in \mathbb{N}\}$ is IID, have been greatly extended to hold in the case of a Martingale Difference (MD) process, which covers numerous forms of non-IID assumptions.

To illustrate the generality of a MD process consider a wild process $\{\phi_\tau, \ \tau \in \mathbb{N}\}$ with no direct distributional, and very general dependence and heterogeneity restrictions:

$$E(\phi_\tau) = \mu(t) < \infty, \ Var(\phi_i) = \sigma^2(t) < \infty, \ Cov(\phi_i, \phi_j) = c(t, s) < \infty, \ i, j = 1, 2, \ldots$$

It can be shown that the transformed process $\{Y_k, \ k \in \mathbb{N}\}$:

$$Y_k = Z_k - E(Z_k|D_{k-1})$$

where $D_{k-1} = \sigma(Z_{k-1}, Z_{k-2}, \ldots, Z_1), \ k = 2, 3, \ldots,$

is ‘well-behaved’ because it is an MD process with properties:

(i) mean homogeneous: $E(Y_k|D_{k-1}) \Rightarrow E(Y_k) = E\{E(Y_k|D_{k-1})\} = 0$,

(ii) orthogonal: $E(Y_i Y_j) = E\{E(Y_i Y_j|D_{i-1})\} = 0$, for $i > j, \ i, j = 1, 2, \ldots$

Hence, assuming that $E(Y_k^2) < c < \infty$, one can deduce that:

$$Var\left(\frac{1}{n} \sum_{k=1}^{n} Y_k\right) = \frac{1}{n^2} \sum_{k=1}^{n} E(Y_k^2) \leq \frac{nc}{n^2} \rightarrow 0.$$  (2)

This condition ensures that the various limit theorems, including the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT), hold for the process $\{Y_k, \ k \in \mathbb{N}\}$; see Doob (1953). This result is very important because limit theorems demarcate the scope of empirical modeling by specifying conditions under which ‘learning from data’ is potentially possible as $n \rightarrow \infty$.

### 2.2 Frequentist estimation

The key features of the frequentist approach to inference are:

[a] The interpretation of probability is grounded on the Strong Law of Large Numbers providing a sound link between relative frequencies and mathematical probability under certain restrictions on the probabilistic structure of the process $\{Z_t, \ t \in \mathbb{N}\}$ underlying data $Z_0$; see Spanos (2012).

[b] The chance regularities exhibited by data $Z_0$ constitute the only relevant statistical information for the specification of the probabilistic structure of the generic
stochastic process \( \{Z_t, t \in \mathbb{N}\} \) underlying the data. The statistical model \( M_\theta(z) \) parameterizes \( \{Z_t, t \in \mathbb{N}\} \) with a view to pose the substantive questions of interest. 

Substantive information comes in the form of restrictions on statistical parameters, but should not be imposed at the outset; it needs to be tested before imposed.

[c] The primary aim of the frequentist approach is to learn from data about the ‘true’ statistical data-generating mechanism \( M^*(z) = \{f(z; \theta^*)\}, z \in \mathbb{R}_Z^p \).

In frequentist estimation learning from data using an estimator \( \hat{\theta} \) of \( \theta \in \Theta \) is evaluated in terms of the sampling distribution \( f_n(\hat{\theta}; \theta^*) \) where \( \theta^* \) denotes the true value of \( \theta \). The finite sample properties [any \( n > 1 \)] are defined directly in terms of this distribution, and they include unbiasedness, full efficiency and sufficiency.

**Unbiasedness.** An estimator \( \hat{\theta} \) of \( \theta \) is said to be unbiased if: \( E(\hat{\theta})=\theta^* \), where \( \theta^* \) denotes the true value of \( \theta \) in \( \Theta \), whatever that happens to be.

<table>
<thead>
<tr>
<th>Table 1 - The simple Bernoulli model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical GM: ( X_t = \theta + u_t, t \in \mathbb{N}, )</td>
</tr>
<tr>
<td>[1] Bernoulli: ( X_t \sim \text{Ber}(\theta) ),</td>
</tr>
<tr>
<td>[2] Constant mean: ( E(X_t)=\theta, 0 \leq \theta \leq 1 ), for all ( t \in \mathbb{N} ),</td>
</tr>
<tr>
<td>[3] Constant variance: ( \text{Var}(X_t)=\theta(1-\theta) ), for all ( t \in \mathbb{N} ),</td>
</tr>
<tr>
<td>[4] Independence: ( {X_t, t \in \mathbb{N}} - \text{independent process} ).</td>
</tr>
</tbody>
</table>

**Example.** In the case of the simple Bernoulli model (table 1), the Maximum Likelihood Estimator (MLE) \( \hat{\theta}_n(X):=\overline{X}_n=\frac{1}{n}\sum_{t=1}^{n} X_t \) is an unbiased estimator of \( \theta \). Evaluating its sampling distribution under the TSN \( (\theta=\theta^*) \) one can deduce that:

\[
\hat{\theta}_n(X) \xrightarrow{\theta=\theta^*} \text{Bin}(\theta^*, \frac{\theta^*(1-\theta^*)}{n}).
\] (3)

Hence, factual reasoning establishes that \( \hat{\theta}_n(X) \) has a mean equal to the true \( \theta^* \), whatever that value happens to be. The same sampling distribution can be used to show that \( \hat{\theta}_n(X) \) is also fully efficient, as well as a sufficient estimator of \( \theta \); see Cox and Hinkley (1974). Taken together these properties ensure that \( f(\hat{\theta}_n(X); \theta^*) \) is not only located at \( \theta^* \), but its dispersion around that value is as small as possible.

The asymptotic properties of an estimator \( \hat{\theta} \) of \( \theta \) are defined in terms of the asymptotic sampling distribution \( f_\infty(\hat{\theta}; \theta^*) \) aiming to approximate \( f_n(\hat{\theta}; \theta^*) \) as \( n \to \infty \), which in the above case takes the form:

\[
\hat{\theta}_n(X) \xrightarrow{\alpha} \text{N}(\theta^*, \frac{\theta^*(1-\theta^*)}{n}).
\] (4)

Of particular interest is the asymptotic property of consistency.

**Consistency (strong).** \( \hat{\theta}_n \) zeroes-in on \( \theta^* \) as \( n \to \infty \), denoted by \( \hat{\theta}_n \xrightarrow{a.s.} \theta^* \):

\[
\mathbb{P}(\lim_{n \to \infty} \hat{\theta}_n(X)=\theta^*)=1.
\]

This property is illustrated in figures 1 and 2, where for a large enough \( n \) the estimate \( \hat{\theta}_n(x_0) \) ‘points out’ the true \( \theta \) with probability one!
Consistency (weak or strong) is an extension of the Law of Large Numbers (LLN), and a consistent estimator of $\theta$ indicates potential (as $n \to \infty$) learning from data about the true but unknown parameter $\theta^*$. Its importance stems from the fact that consistency is a minimal property of an estimator; necessary but not sufficient.

In this sense consistency demarcate the limits of statistical modeling in the sense that they specify necessary and/or sufficient restrictions on the probabilistic structure of a stochastic process $\{X_t, t \in \mathbb{N}\}$ under which a potentially operational statistical model can be specified. However, to render the potential learning designated by consistency into actual learning one needs to supplement consistency with certain finite sample properties like full efficiency and sufficiency to ensure that learning can take place with the particular data $x_0 := (x_1, x_2, ..., x_n)$ of sample size $n$.

**Empirical example.** Let the simple Bernoulli model (table 1) be selected for as the appropriate $\mathcal{M}_\theta(x)$ for data $x_0$, $n=9869$, boys=5152, girls=4717. Assuming the model is valid, how do we learn from data about $\mathcal{M}^*(x)$, $x \in \mathbb{R}^n_\mathcal{X}$? By using the data $x_0$ in conjunction with effective inference procedures to reduce the infinite parameter space $\Theta:=[0,1]$ to a smaller subset, ideally down to the point $\theta=\theta^*$, i.e. use data $x_0$ as a guide to zero-in on the true $\mathcal{M}^*(x)$, $x \in \mathbb{R}^n_\mathcal{X}$.

Despite the fact that the sample size is very large, $n=9869$, one should not infer that this estimate is ‘very close’ to $\theta^*$, since in inductive inference one needs to calibrate the uncertainty associated with ‘how close’.

It turns out that point estimation is often considered *inadequate* for the purposes of scientific inquiry because a ‘good’ point estimator $\hat{\theta}_n(X):=\overline{x}_n$, by itself, does not provide any measure of the reliability and precision associated with the estimate $\theta_n(x_0)$. This is the reason why $\hat{\theta}_n(x_0)$ is often accompanied by its standard error.

**Interval estimation** can be used to rectify this weakness of point estimation by providing the relevant error probabilities associated with inferences pertaining to ‘covering’ the true value $\theta^*$ of $\theta$. In the above example provides the relevant error probabilities associated with the $(1-\alpha)$ Confidence Interval (CI):

$$\mathbb{P}\left(\overline{x}_n - c_{\frac{\alpha}{2}} s_x \leq \theta \leq \overline{x}_n + c_{\frac{\alpha}{2}} s_x\right) = 1-\alpha, \quad (5)$$
where \( s^2 = \frac{(\bar{X}_n - \mu_0)}{n} \). The inferential claim is that this random interval will cover (overlay) \( \theta^* \), *whatever that happens to be*, with probability \((1-\alpha)\).

For the above numerical example, \( 1.96s_\bar{x} = .009855 \), thus the observed .95 CI is:

\[
[.51215, .53186].
\]  

Does this imply that the true \( \theta^* \) lies within those two bounds with probability .95? Not necessarily! The observed CI (6) either includes or excludes \( \theta^* \), but we do not know. This weakness can be alleviated using hypothesis testing.

### 2.3 Frequentist testing

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

Consider the simple Normal model (table 2), and let the hypotheses of interest be:

\[
H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0. \tag{7}
\]

The test \( T_\alpha:=\{d(\mathbf{X}), \ C_1(\alpha; \mu_0)\} \), where for \( \bar{X}_n = \frac{1}{n} \sum^n_{i=1} X_i \) and \( s^2 = \frac{1}{n-1} \sum^n_{i=1} (X_i - \bar{X}_n)^2 \):

\[
d(\mathbf{X}) = \sqrt{\frac{n-1}{s}} \frac{\bar{X}_n - \mu_0}{H_0} \text{St}(n-1), \quad C_1(\alpha):=\{x : d(\mathbf{x}) > c_\alpha\}, \tag{8}
\]

where \( \text{St}(n-1) \) denotes the Student’s \( t \) distribution with \((n-1)\) degrees of freedom, is Uniformly Most Powerful (UMP); see Lehmann (1986). The relevant significance level and \( p\)-value are defined by:

\[
\mathbb{P}(d(\mathbf{X}) > c_\alpha; H_0) = \alpha, \quad \mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0); H_0) = p(\mathbf{x}_0). \tag{9}
\]

The *power* of this test is defined by:

\[
\pi(\mu_1) = \mathbb{P}(d(\mathbf{X}) > c_\alpha; \mu_1) \quad \text{for} \quad \mu_1 = \mu_0 + \gamma_1, \quad \gamma_1 \geq 0, \tag{10}
\]

stemming from the sampling distribution:

\[
d(\mathbf{X}) = \sqrt{\frac{n-1}{s}} \frac{\bar{X}_n - \mu_0}{\mu_1 - \mu_0} \text{St}(n-1; \delta), \quad \delta = \frac{\sqrt{n}}{\sigma} (\sqrt{\frac{n-1}{s}} \frac{\bar{X}_n - \mu_0}{\mu_1 - \mu_0} \mu_1 - \mu_0). \tag{11}
\]

UMP means that this test has the highest capacity to detect any discrepancy of the form \( \mu_1 > \mu_0 \) than any other \( \alpha \)–significance level N-P test.

At this stage it is important to bring out two difference between factual and hypothetical reasoning underlying (7)-(10).

*First*, practitioners often confuse the following two distinct sampling distributions:

\[
\text{(a)} \quad d(\mathbf{X}) = \sqrt{\frac{n-1}{s}} \frac{\bar{X}_n - \mu_0}{\mu_0} \text{St}(n-1), \quad \text{(b)} \quad d(\mathbf{X}; \mu) = \sqrt{\frac{n-1}{s}} \frac{\bar{X}_n - \mu^*}{\mu^*} \text{St}(n-1), \tag{12}
\]
where \(d(X)\) is a test statistic evaluated under the null, and \(d(X;\mu)\) is a pivot evaluated under the TSN, overlooking the fact that \(\mu_0\) in (a) is a predesignated value of \(\mu\), but \(\mu^*\) in (b) denotes the true (but unknown) value of \(\mu\).

Second, in factual reasoning that there is only one scenario, but in hypothetical reasoning there is an infinity of possible scenarios. This has two important implications.

(a) Due to the multiplicity of hypothetical scenarios, testing poses sharper questions and often elicits more precise answers.

(b) The error probabilities associated with hypothetical reasoning are also definable post-data, using (11). This explains why post data CIs have degenerate coverage error; the TSN scenario has played out and the relevant error probabilities stemming from (??) are either one or zero.

2.4 Statistical misspecification and unreliable inference

A strong case can be made that the single most crucial contributor to the accumulation of untrustworthy evidence in applied econometrics is statistical misspecification. The overwhelming majority of statistical evidence presented in prestigious applied econometric journals are totally untrustworthy because the implicit probabilistic assumptions invoked by the inference procedures used by the modeler are invalid for the data used; see Spanos (2006a).

What goes wrong when \(M_\theta(x)\) is statistically misspecified? The assumed likelihood function is invalid \(L(\theta;x_0)\) and that gives rise to erroneous inferences. For Bayesian inference, it induces an erroneous posterior \(p(\theta|x_0)=\pi(\theta)L(\theta;x_0)\). For frequentist inference this gives rise to unreliable fit/prediction measures, as well as false error probabilities. A false \(f(x;\theta)\Rightarrow\) a false sampling distribution \(F_n(t)\) for any statistic \(T_n=g(X)\):

\[
F(t;\theta):=\mathbb{P}(T_n \leq t;\theta) = \int \int \cdots \int f(x;\theta) \ dx.
\]

That is, statistical misspecification induces a discrepancy between the actual and nominal (assumed) error probabilities. The surest way to draw an invalid inference is to apply a .05 significance level test when its actual – due to misspecification – type I error probability is closer to .99; see Spanos and McGuirk (2001).

Example. To illustrate the potentially devastating effects of statistical misspecification let us consider the simple Normal model (table 2), and evaluate the nominal and actual error probabilities in the case when [4] is false.

Instead of [4], we assume that \(\text{Corr}(X_i, X_j)=\rho, \ 0 < \rho < 1, \ \text{for all} \ i\neq j, \ i, j=1, \ldots n\), focusing on the case where: \(n=100, \ \alpha=.05 \ (c_\alpha=2.01)\).

What are the effects of such a misspecification on the relevant error probabilities of the t-test in (8) for the hypotheses in (7)? For \(\rho=.1\), the nominal \(\alpha=.05\) becomes an actual type I error \(\rho=.1\) of \(\alpha^*=.317\), and the discrepancy \(\alpha^*−\alpha=.267\) increases as \(\rho\rightarrow1\).
Similarly, the distortions in power are sizeable, ranging from $\pi^*(.01) - \pi(.01) = .266$ to $\pi^*(.3) - \pi(.3) = -.291$.

Again, the discrepancy between nominal and actual power increases as $\rho \to 1$; for more details see Spanos (2009).

3 The Pre-Eminence of Theory (PET) perspective

The single most important contributor to the untrustworthiness of empirical evidence in economics is the methodological framework, known as the Pre-Eminence of Theory (PET) perspective, that has dominated empirical modeling in economics since Ricardo (1817). This framework asserts that empirical modeling takes the form of constructing simple idealized models which capture certain key aspects of the phenomenon of interest, with a view to use such models to shed light or even explain economic phenomena, as well as gain insight concerning potential alternative policies. From the PET perspective empirical modeling is strictly theory-driven with the data playing only a subordinate role in quantifying theory-models (assumed to be true); see Spanos (2010a).

The focus in current textbook econometrics is primarily on addressing the ‘quantification’ of theory models, notwithstanding the gap between the variables envisaged by theory and what the available data actually measure. What is esteemed is the mathematical sophistication of the inferential techniques (estimators and tests) called for in quantifying alternative theoretical models using different types of data (time series, cross-section and panel). Since the early 1950s, prestigious econometric journals are overflowing with a bewildering plethora of estimation techniques, and their associated asymptotic theory based on invoking mathematically convenient assumptions that are often non-testable. Even in the case where some of these assumptions are testable they are rarely checked against the data.

3.1 Textbook econometrics and untrustworthy evidence

As a result, the incessant accumulation of asymptotic inferential results leave the practitioner none the wiser as to ‘how’ and ‘when’ to apply them to practical modeling problems with a view to learn from data about economic phenomena of interest. All the practitioners can do is apply these techniques to a variety of data hoping that occasionally some of the computer outputs will enable them to ‘tell a story’ and thus publish to survive academia. The trouble is that most of the empirical ‘evidence’ upon which these stories rely upon is untrustworthy, shedding no real light on economic phenomena of interest, primarily because the link between the theory and data is precarious at best. How did we reach this state of affairs?

A crucial implication of the textbook econometric PET perspective is that by making probabilistic assumptions about error (shock) terms the emphasis is placed on the least restrictive assumptions that would justify (asymptotically) the ‘quantification’ technique, irrespective of whether they are testable or not. The idea is that weaker assumptions are less vulnerable to misspecification; hence the current
popularity of the Generalized Method of Moments (GMM) as well as Nonparametric methods. This is a flawed argument because weaker, but indirect, assumptions about the stochastic process \{Z_t, t \in \mathbb{N}\} underlying the data \(Z_0\), are not necessarily less vulnerable to statistical misspecification, when they largely ignore the probabilistic structure of the data. Indeed, weaker non-tested/testable assumptions are likely to undermine both the reliability and precision of inference; see Spanos (2007).

The textbook ‘quantification of theory’ perspective has influenced, not only the way statistical models are specified (attaching error terms to theory models), but also the M-S testing and respecification facets of modeling.

The Linear Regression (LR) model was made the cornerstone of textbook econometrics:

<table>
<thead>
<tr>
<th>Table 3: Traditional Linear Regression (LR) model</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y_t = \beta_0 + \beta_1 x_t + u_t, \quad t \in \mathbb{N} := (1, 2, ..., n, ...))</td>
</tr>
<tr>
<td>(i) ((u_t</td>
</tr>
<tr>
<td>(iii) (E(u_t^2</td>
</tr>
</tbody>
</table>

with all other models being viewed as modifications/extensions of this model; see Greene (2011). A key issue with table 3 is that the error assumptions [i]-[iv] provide a misleading picture of the statistical premises since:

(a) Their validity cannot be directly evaluated vis-a-vis \(Z_0 = (z_1, ..., z_n)\).

(b) The list of assumptions is incomplete; there are missing assumptions.

(c) The parameters \((\beta_0, \beta_1, \sigma^2)\) are not given a proper statistical parameterization.

(d) What ultimately matters is not the set of assumptions pertaining to the error term, but what they imply for the probabilistic structure of the observable process \(\{Z_t, t \in \mathbb{N}\}, Z_t := (y_t, X_t)\) in terms of which the distribution of the sample \(D(z_1, z_2, ..., z_n; \phi)\) and the likelihood function \(L(\phi; Z_0)\) are defined.

(e) The substantive and statistical information are inseparably intermingled by viewing the theory model as the systematic component and attaching probabilistic error terms to define the non-systematic component.

The last of these weaknesses has contributed significantly to providing a misleading perspective on both Mis-Specification (M-S) testing and respecification. To illustrate that consider the first M-S test to probe for departures from assumption [iv], proposed by Durbin and Watson (1950). Their proposed test was based on assuming a particular form of autocorrelation for the error term:

\[u_t = \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1,\]  

and framing the null and alternative hypotheses in terms of the additional parameter:

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

Ignoring the fact that \(E(u_t u_s | X_t = x_t) \neq 0\) can take an infinite number of forms, rejection of \(H_0\) is interpreted by textbook econometrics as providing evidence for \(H_1\). Worse, the form of respecification recommended is to adopt \(H_1\), by making the Autocorrelation-Corrected (A-C) linear regression the new maintained model; see Greene (2011) inter alia. This form of respecification constitutes a classic example of the fallacy of rejection: evidence against \(H_0\) (mis)interpreted as evidence for \(H_1\); see section 4.
3.2 An empirical example: the CAPM

Consider the following example from financial econometrics. The structural model is known as the Capital Asset Pricing Model (CAPM) (Lai and Xing, 2008):

\[(r_{kt} - r_{ft}) = \beta_k (r_{Mt} - r_{ft}) + \varepsilon_{kt}, \quad k=1,2,...,m, \quad t=1,...,n,\]  

(17)

where \(r_{kt}\) – returns of asset \(k\); \(r_{Mt}\) – market returns; \(r_{ft}\) – returns of a risk free asset. To simplify the discussion, let us focus on just one of the \(m\) equations (one for each asset). To test its validity vis-a-vis the data, the CAPM (17) is usually embedded into the linear regression model (table 3):

\[y_t = \beta_0 + \beta_1 x_t + u_t, \quad t=1,...,n,\]  

(18)

where \(y_t := (r_t - r_{ft})\), \(x_t := (r_{Mt} - r_{ft})\). In light of the fact that (18) nests (17) via the parametric restriction \(\beta_0 = 0\), the validity of the CAPM is often tested using the hypotheses:

\[H_0: \beta_0 = 0, \quad vs. \quad H_1: \beta_0 \neq 0.\]  

(19)

**Data.** The relevant data \(Z_0\) come in the form of *monthly log-returns* of Citigroup \((r_t = \Delta \ln P_t)\) for the period Aug. 2000 to Oct. 2005, with the market portfolio returns \((r_{Mt})\) measured by the SP500 index and the risk free asset by the 3-month Treasury bill (3-Tb) rate return \((r_{ft})\); see Lai and Xing (2008), pp. 72-81.

Estimating the statistical model (18) for Citigroup yields:

\[y_t = .0053 + 1.137 x_t + \hat{u}_t, \quad R^2 = .725, \quad s = .0188, \quad n = 64.\]  

(20)

On the basis of (20) the PET assessment will be as follows (Lai and Xing, 2008, pp. 72-81):

(a) the signs and magnitudes of the estimated coefficients are in accordance with the CAPM (\(\beta_0 = 0\) and \(\beta_1 > 0\)),

(b) the beta coefficient \(\beta_1\) is statistically significant, on the basis of the t-test:

\[\tau(Z_0; \beta_1)=\frac{1.137}{.089}=12.775[.000],\]

where the p-value is given in square brackets,

(c) the CAPM restriction \(\beta_0 = 0\) is not rejected by the data, even at a .1 significance level, using the t-test:

\[\tau(Z_0; \beta_0)=\frac{.0053}{.0188}=2.806[.108],\]

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

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

Such confirming evidence, however, are questionable unless one can show that the invoked inductive premises are valid for the above data. As shown in section 4, the above estimated model is seriously misspecified statistically, rendering the above inferences unreliable.
4 Error Statistics: a modeling framework

In the context of the F-N-P perspective, probability plays two interrelated roles. First, \( f(z; \theta) \), \( z \in \mathbb{R}^n \) attributes probabilities to all legitimate events relating to the sample \( Z \). Second, it provides all relevant error probabilities associated with any statistic (estimator, test or predictor) \( T_n=g(Z) \) via (13). Arguably, the most enduring contribution of the F-N-P perspective comes in the form of rendering the errors for statistical induction ascertainable by embedding the ‘material experiment’ into a statistical model in terms of which the relevant error probabilities are defined, thus providing a measure of the reliability of the inference procedure; see Spanos (2006a).

4.1 Outstanding foundational problems and issues

Despite the wide-spread influence of the Fisher-Neyman-Pearson (F-N-P) framework in numerous applied fields, several key foundational issues and problems, bedeviling this approach since 1930s, remained unresolved. These problems include, not only issues pertaining to inference, extensively discussed in the statistics literature, but also problems pertaining to modeling arising in fields where substantive information plays an important role. A partial list of these key problems is given below.

A. Foundational issues pertaining to Modeling

[i] Statistical model specification: choosing a statistical model \( M_\theta(z) \) as it relates to:
(a) statistical information contained in data \( Z_0 \), and
(b) substantive (subject matter) information,
[ii] statistical adequacy: establishing the validity of a statistical premises \( M_\theta(z) \) vis-a-vis data \( Z_0 \),
[iii] statistical model re-specification: how to respecify a model \( M_\theta(z) \) when found misspecified, and
[iv] substantive adequacy: how to assess the adequacy of a substantive model \( M_\varphi(z) \) in shedding light (describe, explain, predict) on the phenomenon of interest.

B. Foundational issues pertaining to Inference

[v] the role of pre-data (type I and II) vs. post-data (p-value) error probabilities,
[vi] safeguarding frequentist inference against the following fallacies:
(a) the fallacy of acceptance: (mis)interpreting accept \( H_0 \) [no evidence against \( H_0 \)] as evidence for \( H_0 \),
(b) the fallacy of rejection: (mis)interpreting reject \( H_0 \) [evidence against \( H_0 \)] as evidence for a particular \( H_1 \); e.g. statistical vs. substantive significance.

As argued next, these issues and problems can be adequately addressed in the context of the ‘error statistical approach’, which can be viewed as a refinement/extension of the Fisher-Neyman-Pearson (F-N-P); Mayo and Spanos (2011). In particular, error statistics aims to:

[A] Refine the F-N-P approach by proposing a broader framework that distinguishes the statistical from the substantive premises, with a view to address modeling issues like [i]-iv], and
Extend the F-N-P approach by supplementing it with a post-data severity assessment with a view to address a variety of inference problems, including [v]-[vi] (Mayo and Spanos, 2006).

The most important modeling foundational problem that has frustrated every applied field that relies on observational data relates to the role of theory and data in empirical modeling and pertains to the issues [i]-[iv] raised above. The inevitable result of foisting one’s favorite theory on the data is often an estimated model \( M_{\theta}(z) \) which is both statistically and substantively misspecified, but one has no principled way to distinguish between the two sources of misspecification and apportion blame:

\[
\text{is the substantive subject matter information false? or are the statistical premises of inference misspecified?} \tag{21}\]

The key to circumventing this Duhemian ambiguity (Mayo, 1997) is to find a way to disentangle the statistical \( M_{\theta}(z) \) from the substantive premises \( M_\phi(z) \), without compromising the integrity of either source of information.

What is often insufficiently realized is that behind every inference based on observed data \( Z_0 \) in conjunction with a substantive (scientific) model \( M_\phi(z) \), there is a distinct statistical model \( M_{\theta}(z) \) (often implicit) providing the inductive premises of inference. Separating the two enables one to establish statistical adequacy first in order to secure the reliability of inference procedures, and then proceed to assess the substantive hypotheses of interest reliably. In addition, sound probabilistic foundations for the notion of a statistical model deal effectively with the crucial problem of distinguishing between real regularities and artifacts in data; a key problem that led to the demise of Political Arithmetic.

A moment’s reflection suggests that from this perspective the above Keynesian thesis amounts to claiming that to learn anything from data one must know everything at the outset, i.e. construct a substantive model \( M_\phi(z) \) and select an appropriate data set \( Z_0 \) that would render \( M_\phi(z) \) both statistically and substantively adequate. Although widely held, this is a highly misleading thesis!

As argued in the sequel, despite the fact that the substantive (structural) model \( M_\phi(z) \) may always come up short in fully capturing or explaining a highly complex phenomenon of interest, a statistical model \( M_{\theta}(z) \) may be entirely adequate to reliably test and evaluate the appropriateness or completeness of \( M_\phi(z) \). Indeed, the combination of the complex nature of economic phenomena and the observational nature of economic data call for more not less reliance on statistical methods.

In fields that rely on observational data, establishing statistical adequacy is of paramount importance because it can help to guide (reliably) the challenging search for adequate substantive explanations of complex phenomena, by posing substantive questions like ‘is \( W_t \) a confounding factor?’ in the context of a statistically adequate model. Without securing statistical adequacy, any answer based on statistical procedures cannot be relied upon due to the potential discrepancy between actual and nominal error probabilities stemming from statistical misspecification! A statistically
adequate model can also provide a guiding constrain for new theories aspiring to explain complex phenomena in so far as they should be required, at the very least, to account for the statistical information in the data.

### 4.2 Bridging the gap between theory and data

Error statistics proposes a systematic account that can be used to address the methodological problems in [i]–[vi], as well as the foundational problems bedeviling frequentist inference; see Mayo and Spanos (2011). In this sense error statistics can be viewed as a refinement/extension of the Fisher-Neyman-Pearson approach to frequentist statistics, that proposes ways to deal adequately with several key foundational issues in the theory-data confrontation including the following responses to the problems [i]–[vi] above.

(A) From theory $T$ to testable hypotheses $h$: fashioning an abstract and idealized theory $T$ into into a structural (substantive) model $M_\varphi(z)$ that is estimable in light of data $Z_0$, with $h$ representing the substantive hypotheses of interest framed in the context of $M_\varphi(z)$.

![Diagram 1: Sequence of interlinked models in empirical modeling](image)

To address this problem error statistics proposes a sequence of interconnected models (see diagram 1) aiming to bridge the gap between theory and data; see Spanos (1986, 1989), Mayo (1996).

From the theory side of the bridge one constructs a theory model (a mathematical formulation), say $M_\varphi(z; \xi)$ which might often include latent variables $\xi$. In connect-
ing $M_{\varphi}(z; \xi)$ to the available data $Z_0$ one needs to transform the theory model into an estimable (in light of data $Z_0$) form, the substantive (structural) model $M_{\varphi}(z)$. The construction of a structural model is particularly taxing in economics because one needs to reflect on the huge gap between the circumstances envisaged by the theory and its the concepts (intentions, plans), and the actual phenomenon of interest giving rise to the available data; see Haavelmo (1944).

From the data side, the statistical information (chance regularity patterns exhibited by data) is distilled by a statistical model $M_{\theta}(z)$ with a view to meet two interrelated aims:

(a) to account for the chance regularities in data $Z_0$ by choosing a probabilistic structure for the generic stochastic process $\{Z_t, t \in \mathbb{N}\}$ so as to render $Z_0$ a ‘truly typical realization’ thereof, and

(b) to parameterize ($\theta \in \Theta$) this probabilistic structure in an attempt to specify $M_{\theta}(z)$ in such a way so as to embed (parametrically) $M_{\varphi}(z)$ in its context via restrictions of the form $G(\theta, \varphi)=0$, $\theta \in \Theta$, $\varphi \in \Phi$, relating the statistical and substantive parameters.

In light of these identification restrictions (see diagram 1) one might interested in specific substantive hypotheses of interest $h$ that can now be framed in terms of the statistical parameters $\theta$, say:

$$H_0 : h(\theta)=0, \text{ vs. } H_1 : h(\theta) \neq 0, \theta \in \Theta. \quad (22)$$

(B) From raw data to reliable evidence: transforming a finite and incomplete set of raw data $Z_0$ – containing uncertainties, impurities and noise – into reliable ‘evidence’ summarized by a statistically adequate model $M_{\theta}(z)$ pertinent for appraising the substantive hypotheses of interest $h$.

This takes the form of establishing the statistical adequacy of $M_{\theta}(z)$ using thorough Mis-Specification (M-S) testing (validation), guided by discerning graphical techniques and effective probing strategies, and respecification in cases where the original model is found wanting. A statistically adequate model $M_{\theta}(z)$ secures the reliability of statistical inferences based on it by ensuring that the actual error probabilities approximate closely enough the assumed (nominal) ones. It is important to note that goodness-of-fit/prediction is neither necessary nor sufficient for statistical adequacy; Spanos (2007).

(C) Using reliable evidence to appraise substantive hypotheses: assessing whether $Z_0$ provides evidence for or against $h(\theta)=0$.

A statistically adequate model $M_{\theta}(z)$ summarizes the statistical information in the data to provide the reliable evidence against which the theory in question can be confronted. Error statistics calls for probing the substantive hypotheses of interest in (22) only in the context of a statistically adequate model $M_{\theta}(z)$ [to ensure the reliability of the test in question]. Moreover, it proposes extending the N-P accept/reject result to providing evidence for or against $h(\theta)=0$ using the severity reasoning. The latter establishes the warranted discrepancy from the null and enables one to address the crucial fallacies of acceptance and rejection; see Mayo and Spanos (2006).
Similarly, one can initiate the process of assessing the substantive adequacy of $M_\varphi(z)$ by testing the over-identifying restrictions in the form of the hypotheses:

$$H_0 : G(\theta, \varphi) = 0, \text{ vs. } H_1 : G(\theta, \varphi) \neq 0, \text{ for } \theta \in \Theta, \varphi \in \Phi.$$  

When empirically valid, the learning from data is passed from $\theta$ to $\varphi$, rendering $M_\varphi(z)$ both statistically and substantively meaningful; see Spanos (1990). $M_\varphi(z)$ can now provide the basis for shedding light on (describe, explain, predict) the phenomenon of interest.

4.3 Probing for and eliminating/controlling errors

One of the key advantages of bridging the gap between theory and data in terms of a sequence of interrelated models (diagram 1) is that it makes the framing of potential errors during the modeling process more transparent.

(I) Errors in theory model construction. The mathematical framing of an abstract theory $T$ into a theory model $M_\psi(z; \xi)$ might turn out to be inappropriate or incomplete.

(II) Errors in eliminating latent variables by transforming $M_\psi(z; \xi)$ into a structural model $M_\varphi(z)$ to render it estimable with data $Z_0$.

(III) Inaccurate data errors: data $Z_0$ are marred by systematic errors imbibed by the collection/compilation process. Such systematic errors are likely to distort the statistical regularities and give rise to misleading inferences; see Morgenstern (1963).

(IV) Incongruous measurement errors: data $Z_0$ do not adequately quantify the concepts envisioned by the theory. This, more than the other substantive sources of error, is likely to be the most serious one in ruining the trustworthiness of empirical evidence; see Spanos (1995).

(V) Statistical misspecification errors: one or more of the probabilistic assumptions of the statistical model $M_\psi(z)$ are invalid for data $Z_0$.

(VI) Substantive inadequacy errors: the circumstances envisaged by the theory in question differ ‘systematically’ from the actual data generating mechanism underlying the phenomenon of interest. This inadequacy can easily arise from impractical ceteris paribus clauses, missing confounding factors, false causal claims, etc.; see Guala (2005).

4.4 Substantive vs. statistical premises of inference

The confusion between statistical and substantive assumptions permeates the whole econometric literature. A glance at the complete specification of the linear regression model in econometric textbooks reveals that along the probabilistic assumptions pertaining to the observable processes [made via the error term] underlying the data in question, there are substantive assumptions concerning omitted variables (Spanos, 2006c), measurement errors and non-simultaneity; see Greene (2011) and Kennedy (2008), Wooldridge (2009) inter alia.

The key question is: How can one disentangle the two types of assumptions?
Error statistics views empirical modeling as a **piecemeal process** that relies on distinguishing between the **statistical** $M_\theta(z)$ and **substantive** $M_\varphi(z)$, clearly delineating the following two questions:

**a) statistical adequacy:** does $M_\theta(z)$ account for the chance regularities in $Z_0$?

**b) substantive adequacy:** is model $M_\varphi(z)$ adequate as an explanation (causal or otherwise) of the phenomenon of interest?

As sketched above, the substantive premises stem from the theory or theories under consideration. The more difficult question is Where do the statistical premises come from?

The fashioning of the statistical model (premises) $M_\theta(z)$ begins with a given data $Z_0$, irrespective of the theory or theories that led to the choice of $Z_0$. Once selected, data $Z_0$ take on ‘a life of its own’ as a particular realization of an underlying generic stochastic process $\{Z_t, \ t\in\mathbb{N}\}$. The link between data $Z_0$ and the process $\{Z_t, \ t\in\mathbb{N}\}$ is provided by the key question:

‘what probabilistic structure pertaining to the process $\{Z_t, \ t\in\mathbb{N}\}$ would render data $Z_0$ a truly typical realization thereof?’

Adopting a statistical perspective, one views data $Z_0$ as a realization of a generic (vector) stochastic process $\{Z_t, \ t\in\mathbb{N}\}$, regardless of what the variables $Z_t$ measure substantively, thus separating the ‘statistical’ from the ‘substantive’ information.

**Example.** What does a ‘typical realization’ of Normal, Independent and Identically Distributed (NIID) process $\{Z_t, \ t\in\mathbb{N}\}$ look like? Like fig. 1, but not 2; see Spanos (1999).

A pertinent answer to the ‘typicality’ question provides the relevant probabilistic structure for $\{Z_t, \ t\in\mathbb{N}\}$. The next step is to choose a parameterization $\theta \in \Theta$, based on this probabilistic structure, to define the relevant statistical model $M_\theta(z)$.

**Example 1.** For the data in fig. 1, an appropriate model is the simple Normal:

$$M_\theta(z): Z_t \sim \text{NIID}(\mu, \sigma^2), \ \theta := (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+, \ z_t \in \mathbb{R}, \ t\in\mathbb{N}$$

**Example 2.** For fig. 2, an appropriate model is the Normal, mean-heterogeneous:

$$M_\theta(z): Z_t \sim \text{NI}((\mu_0 + \mu_1 t, \sigma^2), \ \theta := (\mu_0, \mu_1, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_+, \ z_t \in \mathbb{R}, \ t\in\mathbb{N}$$
The particular parameterization \( \theta \in \Theta \) giving rise to \( \mathcal{M}_\theta(z) \) is chosen with a view to embed (nest) the substantive model \( \mathcal{M}_\varphi(z) \) in its context, e.g. via certain restrictions \( G(\theta, \varphi)=0, \varphi \in \Phi \); see Spanos (1990).

A statistical model of particular interest in economics is the Normal/Linear Regression (LR) model, which from this perspective can be viewed as a parameterization of an observable NIID process \( \{Z_t:=(y_t, X_t), \ t \in \mathbb{N}\} \), stemming from the reduction:

\[
D(Z_1, Z_2, \ldots, Z_n; \phi) = \prod_{t=1}^n D(Z_t; \varphi(t)) \overset{\text{IID}}{=} \prod_{t=1}^n D(Z_t; \varphi) = \prod_{t=1}^n D(y_t|X_t; \varphi_1)D(X_t; \varphi_2)
\]

where Normality of \( D(Z_t; \varphi) \) ensures that the Normal LR can be specified exclusively in terms of the conditional distribution \( D(y_t|X_t; \varphi_1) \), as given in table 4. Note that the multivariate Normal model, the Principal Components and Factor Analysis models can be viewed as alternative parameterizations of the same NIID process \( \{Z_t:=(y_t, X_t), \ t \in \mathbb{N}\} \).

<table>
<thead>
<tr>
<th>Table 4 - Normal/Linear Regression model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical GM: ( y_t = \beta_0 + \beta_1 x_t + u_t, \ t \in \mathbb{N} ),</td>
</tr>
<tr>
<td>[1] Normality: ( (y_t</td>
</tr>
<tr>
<td>[2] Linearity: ( E(y_t</td>
</tr>
<tr>
<td>[3] Homoskedasticity: ( Var(y_t</td>
</tr>
<tr>
<td>[4] Independence: ( {E(y_t</td>
</tr>
<tr>
<td>[5] t-invariance: ( \theta := (\beta_0, \beta_1, \sigma^2) ) do not change with ( t ).</td>
</tr>
</tbody>
</table>

\[
\beta_0 = E(y_t) - \beta_1 E(x_t), \quad \beta_1 = \frac{Cov(x_t, y_t)}{Var(x_t)} \sigma^2 = Var(y_t) - \beta_1 Cov(y_t, x_t)
\]

Comparing the two specifications (table 3 vs. 4), it is clear that there is an equivalence between assumptions [i]-[iii] and [1]-[3], [iv] is strengthened to independence, and assumption [5] is missing from table 3. In contrast to table 3, table 4 provides a complete, internally consistent and testable set of probabilistic assumptions, without which model validation will be hopeless.

4.5 M-S testing: the error-statistical perspective

Adequate understanding of M-S testing requires distinguishing it from Neyman-Pearson (N-P) testing. The key difference is that N-P testing constitutes probing within the boundaries of \( \mathcal{M}_\theta(z)=\{f(z; \theta), \ \theta \in \Theta\} \), by framing the hypotheses in terms of a partition of the parameter space \( \Theta \):

\[
H_0: \ \theta \in \Theta_0 \text{ vs. } H_1: \ \theta \in \Theta_1, \text{ where } \Theta_0 \cup \Theta_1 = \Theta \text{ and } \Theta_0 \cap \Theta_1 = \varnothing.
\]

There are only two types of errors because it is assumed that the statistical adequacy of \( \mathcal{M}_\theta(z) \) has been secured. In contrast, M-S testing constitutes probing outside the boundaries of \( \mathcal{M}_\theta(z) \), and thus no such adequacy can be secured; see Spanos (2000).

The generic hypotheses for M-S testing take the form:

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

19
where \( f^*(z) \) is the ‘true’ distribution of the sample, and \( \mathcal{P}(z) \) denotes the set of all possible statistical models that could have given rise to \( z_0 \). This difference raises a number of conceptual and technical issues, including the following.

1. **How to particularize** \([\mathcal{P}(z) - \mathcal{M}_\theta(z)]\) to construct M-S tests.

A most inefficient way to do this is to attempt to probe \([\mathcal{P}(z) - \mathcal{M}_\theta(z)]\) one model at a time \( M_i(z) \), \( i = 1, 2, \ldots \); a hopeless task. A much more efficient way that enables the elimination of an *infinite* number of alternatives at a time, is to construct M-S tests by modifying the original tripartite [Distribution, Dependence, Heterogeneity] partitioning of \( \mathcal{P}(z) \) that gave rise to \( \mathcal{M}_\theta(z) \) in particular directions of departures gleaned from Exploratory Data Analysis to construct alternative models \( \mathcal{M}_1(z) \) or broad subsets of \([\mathcal{P}(z) - \mathcal{M}_\theta(z)]\); see Spanos (2006b).

2. **Type II error is the more serious.** In contrast to N-P testing the more serious error in M-S testing is the type II error [accepting \( H_0 \) when false]. This is an instance of the *fallacy of acceptance*: (mis)interpreting accept \( H_0 \) [no evidence against \( H_0 \)] as evidence for it. Falsely rejecting \( \mathcal{M}_\theta(z) \) as misspecified can be remedied at the re-specification stage; no such recourse exists for falsely accepting \( \mathcal{M}_\theta(z) \) as statistically adequate. Hence, a good strategy is to apply multiple M-S tests to each assumption (or combination), ensuring that some of these tests have high enough power.

3. **Securing the effectiveness/reliability of the diagnosis.** In M-S testing the objective is to probe as *broadly* away from the null as possible, and thus omnibus tests with low power but broad (local) probing capacity have an important role to play when combined with high power directional tests. The reliability of the diagnosis is enhanced by:

   (i) **Astute ordering** of M-S tests so as to exploit the interrelationship among the model assumptions with a view to ‘correct’ each other’s diagnosis, and

   (ii) **Joint M-S tests** (testing several assumptions simultaneously) designed to minimize the maintained model assumptions. This probing strategy aims to distinguish between the different sources of misspecification. For the LR model (table 4), one can construct joint M-S tests by framing potential departures in terms of how they might alter the first two conditional moments of \( \mathcal{M}_\theta(z) \). This yields *auxiliary regressions* based on the studentized residuals: \( \tilde{v}_t = \sqrt{n}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t)/s \):

\[
\begin{align*}
\tilde{v}_t &= \gamma_{10} + \gamma_{11} x_t + \gamma_{12} t + \gamma_{13} x_t^2 + \gamma_{14} x_{t-1} + \gamma_{15} y_{t-1} + \epsilon_{1t}, \\
\tilde{v}_t^2 &= \gamma_{20} + \gamma_{21} x_t + \gamma_{22} t + \gamma_{23} x_t^2 + \gamma_{24} x_{t-1}^2 + \gamma_{25} y_{t-1}^2 + \epsilon_{2t},
\end{align*}
\]

If the residuals are **non-systematic** – the assumptions of \( \mathcal{M}_\theta(z) \) are valid for data \( z_0 \) – the added terms will be statistically insignificant. Otherwise departures from model assumptions are indicated.

4. **Increased vulnerability to the fallacy of rejection:** (mis)interpreting reject \( H_0 \) [evidence against \( H_0 \)] as evidence for \( H_1 \). M-S testing is more akin to Fisher’s significance
testing where the alternative, although implicitly present, one should *never* accept it without further testing.

Let us return to the Durbin-Watson (D-W) test for assessing assumption [4] (table 1) whose hypotheses in (16) can be viewed as stemming from the fact that for:

\[
\mathcal{M}_0(z) : \quad y_t = \beta_0 + \beta_1 x_t + u_t, \quad u_t \sim \text{NiID}(0, \sigma_u^2)
\]

\[
\mathcal{M}_1(z) : \quad y_t = \beta_0 + \beta_1 x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{NiID}(0, \sigma_\varepsilon^2)
\]

\[
\mathcal{M}_1(z)|_{\rho=0} = \mathcal{M}_0(z).
\]

In the traditional textbook approach when the null is rejected the recommendation is to adopt \(\mathcal{M}_1(z)\), ignoring the fact that the D-W test can provide good evidence against \(\mathcal{M}_0(z)\), but no evidence for \(\mathcal{M}_1(z)\); the rejection could have been due to any one of an infinite number of alternatives in \([\mathcal{P}(z) - \mathcal{M}_0(z)]\). Securing evidence for \(\mathcal{M}_1(z)\) requires one to test its own assumptions.

Despite the blatancy of the fallacy of rejection, the econometric literature ennobled it into the *pre-test bias* problem (Leeb and Pötscher, 2005), adding another misguided problem to textbook econometric modeling; see Spanos (2010b).

### 4.6 Validating the statistical premises of the CAPM

Thorough M-S testing reveals (see table 5) that the estimated model (20) is statistically misspecified; assumptions [1], [3]-[5] are invalid.

<table>
<thead>
<tr>
<th>Table 5- M-S testing results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Normality:</strong> (D'AP = 1.169[.557]^*)</td>
</tr>
<tr>
<td><strong>Homoskedasticity:</strong> (F(2, 59) = 4.950[.010]^*)</td>
</tr>
<tr>
<td><strong>t-invariance:</strong> (F_{\beta}(2, 60) = 4.611[.014]^*)</td>
</tr>
</tbody>
</table>

Indeed, one can easily see why the above statistical model is so seriously misspecified by just looking at the t-plots of the data, which are clearly not exhibiting realizations of NiID processes (fig. 1). Both t-plots exhibit mean and variance heterogeneity as well as temporal dependence in the form of cycles.

The misspecifications indicated by the M-S testing results in table 5 call into question, not only the above quoted t-test results, but also other criteria used by
the Pre-Eminence of Theory (PET) perspective to evaluate the appropriateness of an estimated model:

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

[ii] substantive: theoretical meaningfulness, explanatory capacity,

[iii] pragmatic: simplicity, generality, elegance,

because, directly or indirectly, they invoke the same misspecified inductive procedures. This implies that the estimated model cannot be used as a basis of reliable inferences to assess the validity of the substantive questions of interest.

To illustrate the serious errors such a strategy can give rise to, let us consider the question of confounding in the context of (20) by posing the question whether last period’s excess returns of General Motors is a relevant variable in explaining \((r_t - \mu_{ft})\)-excess returns of Citigroup. The augmented model yields:

\[
(r_t - \mu_{ft}) = 0.055 + 1.131(r_{Mt} - \mu_{ft}) - 0.144(r_{t-1} - \mu_{ft-1}) + \hat{\nu}_t, \quad R^2 = 0.747, \quad s = 0.0182,
\]

Hence, the t-test result \((\tau(Z_0) = 2.571)\), when taken at face value, indicates that the answer is yes! However, this is false because any data series that exhibits certain regularity patterns, including generic trends and lags, is likely to appear significant when added to the original misspecified model:

\[
(r_t - \mu_{ft}) = 0.0296 + 1.344(r_{Mt} - \mu_{ft}) - 0.134t + 0.168t^2 + \hat{\nu}_t, \quad R^2 = 0.745, \quad s = 0.0184,
\]

\[
(r_t - \mu_{ft}) = 0.0034 + 1.24(r_{Mt} - \mu_{ft}) - 0.175(r_{t-1} - \mu_{ft-1}) + \hat{\nu}_t, \quad R^2 = 0.75, \quad s = 0.0181.
\]

Indeed, this can be used to explain the ‘apparent success’ of the Fama and French (1996) multifactor model:

\[
M_\psi(z) : (r_{it} - r_{ft}) = \beta_0 + \beta_1(r_{Mt} - r_{ft}) + \sum_{i=2}^{p} \beta_{ii} f_{it} + \varepsilon_{it}, \quad i = 1, 2, ..., m, \quad t = 1, ..., n,
\]

where \(f_{it}, i = 1, 2, ..., p\) are omitted relevant factors, as nothing more than misguided inferences stemming from a misspecified statistical model.

The question that naturally arises at this stage, is what is one supposed to do next? Three things are clear.

First, no evidence for or against a substantive claim (theory) can be secured on the basis of a statistically misspecified model. Second, the so-called refinements/extensions of the CAPM (Levy, 2011) are misguided attempts driven by unreliable inference results! This is also the case in various other applied subfields in economics. Third, before any reliable testing of the CAPM can be applied one needs to secure the statistical adequacy; the validity of model assumptions vis-a-vis data \(Z_0\).

### 4.7 Statistical vs. substantive adequacy

The distinction between substantive adequacy can be used to shed light on several confusions that permeate the current textbook approach to econometric modeling, including the following.

1. **Statistical adequacy** is the price for using reliable statistical procedures to learn from data; it ensures the reliability of inference by securing the closeness of
the actual and nominal error probabilities. Learning from data depends crucially on establishing a sound link between the process generating the data $Z_0$ and the assumed $M_\theta(z)$, by securing statistical adequacy.

2. Statistical adequacy is necessary but not sufficient for substantive adequacy. For substantive adequacy one needs to probe for additional potential errors vis-a-vis the phenomenon of interest, impractical ceteris paribus clauses, intentions vs. realizations, omitted effects, false causal claims, etc.

Example. The textbook discussion of the ‘omitted variables’ bias and inconsistency problem is mired in confusion. A closer look at the argument reveals that it has nothing to do with statistical misspecification. Despite the allusions to biases and inconsistencies of estimators, it is an issue pertaining to the substantive adequacy of a structural model $M_\varphi(z)$; Spanos (2006c).

3. It is often claimed that a ‘good’ model is one that ‘fits the data well’. This is misleading because good fit/prediction is neither necessary nor sufficient for the statistical adequacy of $M_\theta(z)$; see Spanos (2007). Having said that, such criteria become relevant for substantive adequacy in the case where the overidentifying restrictions $G(\theta, \varphi) = 0$ of $M_\varphi(z)$ are valid vis-a-vis a statistically adequate model $M_\theta(x)$. That is, for a data-pertinent structural model $M_\varphi(z)$ goodness-of-fit can provide a measure of its comprehensiveness vis-a-vis the phenomenon of interest.

4. The slogan: "all models are false, but some are useful", attributed to George Box, is based on confusing substantive and statistical inadequacy. A substantive model $M_\varphi(z)$ may always come up short in fully capturing or explaining a phenomenon of interest [e.g. not realistic enough], but a statistical model $M_\theta(z)$ may be entirely satisfactory [when statistically adequate] to reliably test substantive questions of interest, including assessing the appropriateness of $M_\varphi(z)$.

5. Statistical inadequacy can be used to account for the kind of "spurious results" as stemming from statistical misspecification. Time series data that exhibit both temporal dependence and heterogeneity will give rise to a statistically misspecified Linear Regression (LR) model because invariably assumptions [2]-[5] (table 4) are likely to be invalid; see Spanos and McGuirk (2001).

6. Statistical adequacy depends only on $M_\theta(z)$ and $Z_0$ and can be established independently by different practitioners using thorough M-S testing; it is well-defined and objective. Hence, to paraphrase Keynes (1940):

"If 70 well-trained econometricians were shut up in 70 separate rooms with the same $M_\theta(z)$ and $Z_0$, when they emerge they should have the same misspecification diagnosis."

Such a claim cannot be made in relation to establishing substantive adequacy because the search is more open-ended in the sense that there is an infinity of scenarios one could potentially probe for.

7. In the social sciences, where the actual DGM is highly complex and the available data are (often) observational, a statistically adequate $M_\theta(z)$ could play a more crucial role in guiding the search for substantively adequate explanations (theories) by delineating ‘what there is to explain’. Kepler’s ‘law’ for the elliptical motion
of the planets was an empirical regularity that guided Newton toward a substantively adequate model, that of universal gravitation, 60 years after the empirical regularity was first established; Spanos (2007).

8. The distinction between substantive vs. statistical adequacy elucidates the crucial differences between statistical model specification and model selection based on Akaike-type criteria. These model selection procedures often give rise to unreliable inferences primarily because (Spanos, 2010b):
(a) they assume away the problem of statistical model specification and validation,
(b) their ranking of models is equivalent to pairwise comparison between them using Neyman-Pearson testing, but without ‘controlling’ the relevant error probabilities.

4.8 Post-data severity evaluation

Error statistics provides an important extension of the Fisher-Neyman-Pearson (F-N-P) approach to frequentist inference comes in the form of a post-data evaluation of the N-P accept (reject) results using severe testing reasoning to weave an evidential interpretation in the form of the smallest (largest) discrepancy γ from the null warranted with data \( x_0 \) (Mayo and Spanos, 2006).

In an attempt to simplify the discussion, consider testing the following hypotheses:

\[ H_0 : \theta = \theta_0 \] vs. \[ H_1 : \theta > \theta_0 \], where \( \theta_0 = \frac{18}{35} \approx .5143 \), in the context of the simple Bernoulli model (table 1) for \( \alpha = .01 \Rightarrow c_\alpha = 2.326 \). It can be shown that the test defined by \( T_\alpha^\gamma := \{ d(X), C_1^\gamma (\alpha) \} \):

\[
d(X) = \left( \frac{n}{\theta_0 (1-\theta_0)} \right)^{1/2} \frac{H_0}{\theta_0 (1-\theta_0)} \text{ Bin}(0, 1; n), \quad C_1^\gamma (\alpha) = \{ x : d(x) > c_\alpha \},
\]

is an \( \alpha \)-level, Uniformly Most Powerful (\( \alpha \)-UMP) test; see Lehmann (1986). Applying this test to the above data \( x_0 \), where \( n=9869 \), boys=5152, girls=4717 yields:

\[
d(x_0) = \left( \frac{9869}{5152} \cdot \frac{18}{35} \right) = 1.541, \quad p_\gamma (x_0) = P(d(X) > 1.541; H_0) = .062.
\]

Hence, the null hypothesis is accepted, and the p-value indicates that \( H_0 \) wouldn’t have been rejected even if \( \alpha = .05 \) were to be chosen.

Does the acceptance mean that there is evidence for \( \theta_0 = \frac{18}{35} \)? Nor necessarily! As argued above, acceptance (rejection) of \( H_0 \) is vulnerable to the fallacy of acceptance (rejection).

An effective way to circumvent such fallacious inferences is to use the post-data severity evaluation (Mayo and Spanos, 2006). A hypothesis \( H \) passes a severe test \( T_\alpha \) with data \( x_0 \) if:

(S-1) \( x_0 \) accords with \( H \), and

(S-2) with very high probability, test \( T_\alpha \) would have produced a result that accords less well with \( H \) than \( x_0 \) does, if \( H \) were false.

In light of the fact that for the above example \( d(x_0) = 1.541 > 0 \), the relevant inferential claim is of the form \( \theta > \theta_1 = \theta_0 + \gamma \), and the ‘discordance’ condition (S-2) of severity gives rise to:

\[
SEV(T_\alpha^\gamma; \theta > \theta_1) = P(x : d(x) \leq d(x_0); \theta_0 = \theta_1),
\]

24
whose evaluation is based on the sampling distribution:

$$d(X) = \sqrt{\frac{\hat{\theta} - \theta_0}{\theta_0(1-\theta_0)}} \sim \text{Bin} \left( \delta(\theta_1), V(\theta_1); n \right), \text{ for } \theta_1 > \theta_0, \quad (25)$$

where $\delta(\theta_1) = \sqrt{\frac{\hat{\theta} - \theta_0}{\theta_0(1-\theta_0)}} \geq 0$, $V(\theta_1) = \frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)}$, $0 < V(\theta_1) \leq 1$.

Table 4: Accept $H_0$: $\theta = \frac{15}{35}$ vs. $H_1$: $\theta > \frac{15}{35}$

<table>
<thead>
<tr>
<th>Relevant claim</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>$\theta &gt; \theta_1 = [0.5143 + \gamma]$, $SEV(T_{\alpha}^2; x_0; \theta &gt; \theta_1)$</td>
</tr>
<tr>
<td>-0.0043</td>
<td>$\theta &gt; 0.510$, 0.992</td>
</tr>
<tr>
<td>-0.0033</td>
<td>$\theta &gt; 0.511$, 0.986</td>
</tr>
<tr>
<td>-0.0021</td>
<td>$\theta &gt; 0.5122$, 0.975</td>
</tr>
<tr>
<td>-0.0013</td>
<td>$\theta &gt; 0.513$, 0.964</td>
</tr>
<tr>
<td>-0.00053</td>
<td>$\theta &gt; 0.51376$, 0.950</td>
</tr>
<tr>
<td>-0.00003</td>
<td>$\theta &gt; 0.514$, 0.945</td>
</tr>
<tr>
<td>0.00000</td>
<td>$\theta &gt; 0.5143$, 0.938</td>
</tr>
<tr>
<td>0.0007</td>
<td>$\theta &gt; 0.515$, 0.919</td>
</tr>
<tr>
<td>0.0013</td>
<td>$\theta &gt; 0.5156$, 0.900</td>
</tr>
<tr>
<td>0.0017</td>
<td>$\theta &gt; 0.516$, 0.885</td>
</tr>
<tr>
<td>0.0027</td>
<td>$\theta &gt; 0.517$, 0.842</td>
</tr>
<tr>
<td>0.0037</td>
<td>$\theta &gt; 0.518$, 0.789</td>
</tr>
<tr>
<td>0.0057</td>
<td>$\theta &gt; 0.520$, 0.657</td>
</tr>
</tbody>
</table>

Table 4 reports the evaluation of the severity for several key discrepancies, including $\theta_0 \approx 0.5143$ (first conjectured by Nicholas Bernoulli) and the substantively determined value: $\theta^* = 0.5122$. In human biology (Hardy, 2002) it is commonly accepted that the sex ratio at birth is approximately 105 boys to 100 girls, yielding $\theta^* = \frac{105}{205} = 0.5122$. The severity evaluates in table 4 indicate that for a high enough threshold probability, say .90, the smallest warranted discrepancy associated with the claim:

$$\theta > \theta_1 = \theta_0 + \gamma, \text{ is } \gamma \leq 0.5156.$$

Severity and learning from data. The inferential claims associated with the post-data severity evaluation enable one to get a lot more accurate information pertaining to what the data say about $\theta$. Taking an overly simplistic perspective, one can argue that these evaluations provide very good evidence for a tiny range of values:

$$\left[0 \leq \theta \leq 1 \right]^{SEV(T_{\alpha}^2; x_0; \theta > \theta_1)} \frac{.51 \leq \theta \leq .5156}, \quad (26)$$

with the severity evaluation gradating them. It is important to notice that the point estimate $\hat{\theta}(x_0) = 0.522$ lies outside (26), and the .95 observed CI $[0.51215, 0.53186]$ hardly overlaps with this range of values! Moreover, the above interval should not be interpreted as attaching probabilities to different values of $\theta$, but as post-data $(x_0)$ error probabilities, based on the test $T_{\alpha}^2$, attached to particular inferential claims $(\theta > \theta_1)$.
5 Summary and conclusions

A widely held view in economics is that the current untrustworthiness of empirical evidence and the inability to forecast economic phenomena is largely due to the fact that the economy is too complicated and the resulting data are too heterogeneous to be amenable to statistical modeling.

The paper has argued that whether economic data are amenable to statistical modeling is largely an empirical issue and the current untrustworthiness is mainly due to an inadequate implementation of statistical modeling and inference. The latter stems from two main sources: (i) commingling the statistical and substantive premises of inference that leads to numerous confusions, and (ii) neglecting the validation of the latter gives rise to unreliable inferences.

A case was made that a more appropriate implementation of frequentist statistics can be achieved using the error statistical perspective, a refinement/extension of the Fisher-Neyman-Pearson approach, that attains learning from data about phenomena of interest by addressing (i)-(ii) in a way that gives rise to reliable and effective procedures to establish trustworthy evidence.

The key to disentangling the statistical from the substantive premises is to adopt a purely probabilistic construal of a statistical model \( \mathcal{M}_\theta(z) \) so that its specification relies solely on the statistical information contained in the data, and at the same time it allows for the embedding of the substantive model \( \mathcal{M}_\varphi(x) \) in its context. Viewing \( \mathcal{M}_\theta(z) \) as a parametrization of probabilistic structure of observable stochastic process \( \{Z_t, t \in \mathbb{N}\} \) underlying the data \( Z_0 \), enables one to secure its statistical adequacy vis-a-vis \( Z_0 \) using thorough M-S testing, before posing substantive questions of interest.

The substantive and statistical information are reconciled by embedding the structural model \( \mathcal{M}_\varphi(x) \) into a statistically adequate \( \mathcal{M}_\theta(z) \) and assessing the parameter restrictions \( G(\theta, \varphi) = 0 \).

Learning from data is achieved in three steps. First, the statistical \( \mathcal{M}_\theta(z) \) and substantive \( \mathcal{M}_\varphi(x) \) premises are disentangled. Second, the adequacy of the statistical \( \mathcal{M}_\theta(z) \) premises is secured using thorough M-S testing and respecification. A statistically adequate \( \mathcal{M}_\theta(z) \) provides a sound basis for learning from data because it secures the error reliability of the inferences procedures pertaining to the statistical parameters \( \theta \). Third, the empirical validity of \( \mathcal{M}_\varphi(x) \) is assessed by testing the restrictions \( G(\theta, \varphi) = 0 \). When the validity of these restrictions is established, \( \mathcal{M}_\varphi(x) \) provides a sound basis for assessing the substantive adequacy of the underlying theory and learning from data about the phenomenon of interest.

References


[38] 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.


