# Econometrics in Retrospect and Prospect*

**Aris Spanos**†
Department of Economics,
Virginia Tech, Blacksburg,
VA 24061, USA
<aris@vt.edu>

March 2005, 6th draft

## Contents



*Forthcoming, chapter 1 of the **Palgrave Handbook of Econometrics, vo. 1: Theoretical Econometrics**

†I'm most grateful to Deborah Mayo for some invaluable suggestions and comments on an earlier draft of the chapter. Thanks are also due to Karim Abadir, Bernt Stigum, Terence Mills and an associate editor for several useful comments and suggestions.

# 1 Introduction

The vision of *econometrics*, articulated in the early 20th century, was that it would supply economics with genuine *empirical foundations* by endowing it with an *inductive component*, based on *statistics*. The primary purpose of this chapter is to put forward a retrospective view of econometrics, by revisiting this vision. My objective is to assess the extent to which this vision has been fulfilled, take stock of what has been accomplished so far, and emphasize what remains to be done. The intention is to elucidate the problems that have posed obstacles to fulfilling this vision, as well as offer concrete suggestions for making progress in overcoming them.

The viewing angle of this retrospective is one of an academic econometrician who has actively studied, taught, practiced, and grappled with the broader issues of empirical modeling for more than a quarter century. I undertake this task cognizant of the fact that it exposes me to "**two serious charges: that of tedium and that of presumption**", as well as offers "... **the greatest opportunity for internecine strife**." (Harrod, 1938, p. 383). Having said that, it is important to emphasize at the outset that the critical stance of this chapter is *not* aimed at individual authors and their contributions, but solely at the current state of econometric modeling and its underlying methodological framework.

## 1.1 Econometrics: the vision of early 20th century pioneers

Viewing **econometrics** broadly as the utilization of statistics to provide *empirical foundations* to **economics**, its roots can be traced back to the late 19th and early 20th century. During the 19th century, there was general consensus that economic theorizing begins with certain initial postulates comprising the *premises*, proceeds to derive *deductively* certain 'economic laws', and then their *appropriateness* in enhancing our understanding of economic phenomena is assessed. In a certain sense this constitutes a *primitive* version of what became known as the *hypothetico-deductive method*. The disagreements during the 19th century centered around the *nature* and the *method of assessment* of the initial postulates and the derived economic laws. Despite the widespread misuse of the terms 'deduction' and 'induction' in methodological discussions during the 19th century (see Redman, 1997), the consensus at the end of that century was that '*deductive*' and '*inductive*' methods, broadly defined, were considered complementary, and, *statistics* viewed as *quantitative induction*, could help to provide pertinent empirical foundations to economics; see Keynes (1891).

By the late 19th century the *deductive component* was considered to be largely in place (Marshall, 1890, chs. 3-4), and the leading economists of that generation (Jevons, Menger, Edgeworth, Walras and Pareto) sought ways to provide *empirical foundations* to economics. The **vision** of these early pioneers was articulated by Moore (1908), pp. 1-2, in the following way:

> Economics will become an empirical science when its deductive component is supplemented with an adequate *inductive component* based on statistics.

Jevons (1871) expressed the same vision even earlier:

"**The deductive science of Economics must be verified and rendered useful by the purely empirical science of statistics.**" (ibid., p. 12)

A similar vision provided the cornerstone upon which the Econometric Society was founded in 1930 (see Frisch, 1933).

In the next sub-section we will summarize the current state of econometrics vis-a-vis this vision, in an attempt to provide the vantage point for the retrospective view reasoned in the sequel.

## 1.2 A summary of the current state of affairs in econometrics

At the dawn of the 21st century, econometrics has developed from the humble beginnings of 'curve fitting' by least squares in the early 20th century, into a powerful array of statistical tools for modeling all types of data, from the traditional *time series*, to *cross-section* as well as *panel data*. In view of this, the question that naturally arises is the extent to which the *inductive vision* of econometrics has been fulfilled.

The **main thesis** of this chapter is that a century later this vision remains *largely unrealized*. The impressive developments in econometrics during the 20th century concern primarily the sophistication and rigor of techniques and methods for 'quantifying theory models', but these do not amount to a comprehensive methodology for 'learning from data about observable economic phenomena'.

In order to provide the background for the current state of affairs in econometrics, it is enlightening to compare Leontief's appraisal of the development of econometrics in 1948 and then in 1971. Leontief (1948), after tracing the development of econometrics from I. Fisher and Moore, to Cobb and Douglas, Schultz, Roos and Tinbergen, he appeared to be very optimistic about its prospects. The reason for his optimism was primarily the new methodological framework, introduced by Haavelmo (1944), based on 'the modern theory of statistical inference':

"**Considerable progress has been achieved in recent years toward the understanding of proper and improper application of statistical procedures to economic analysis.**" (Leontief, 1948, p. 393)

He embraced Haavelmo's view that the new methodological framework was:

"**... essentially a systematic attempt to develop a method to bridge the commonly recognized *gap between abstract theory and the actually observed facts* which it is supposed to explain.**" (ibid., p. 394)

He considered the period 1933-1948 as one dominated by methodological concerns (he called it a 'reflective stage'), as opposed to actual empirical modeling. Secondary reasons for his optimism were: (i) the formalization of *simultaneity* and the associated *structural models*, and (ii) the move away from *errors-in-variables* toward *errors in equations* (ibid. p. 402). His concluding message called for a *constructive dialogue between theorists and econometricians* (ibid., p. 393).

20 years later his optimism had vaporized; Leontief (1971) blazoned out:

"**The weak and all too slowly growing empirical foundation clearly cannot support the proliferating superstructure of pure,**

**or should I say, speculative economic theory."** (ibid., p. 1)

He blamed the 'indifferent performance' on the **unreliability of empirical evidence** arising from non-testable probabilistic assumptions:

> "**... the validity of these statistical tools depends itself on the acceptance of certain convenient assumptions pertaining to stochastic properties of the phenomena which the particular models are intended to explain; assumptions that can be seldom verified."** (p. 3)

His main conclusion concerning the empirical foundations for economics was:

> "**In no other field of empirical inquiry has so massive and sophisticated a statistical machinery been used with such indifferent results. Nevertheless, theorists continue to turn out model after model and mathematical statisticians to devise complicated procedures one after another."** (p. 3)

*More than 30 years later, has the reliability of empirical evidence improved?* The situation is arguably worse today. The rapid accumulation of new economic data, together with the widespread use of statistical software on personal computers, have lowered the cost of producing empirical evidence, but has not improved their reliability. Indeed, one can make a case that, at the dawn of the 21st century, the applied econometric literature is filled with a disorderly assemblage of 'study-specific', 'period-specific', and largely *unreliable evidence*, which collectively provide a completely inadequate empirical foundation for economics. The experience of this author with published empirical evidence has been that very few, if any, empirical studies can even survive a thorough probing of their *statistical premises*, regardless of any empirical merit the underlying theories might have. It is argued that the unreliability of evidence is symptomatic of an inadequate methodological framework, and thus the critical discourse of this chapter is not directed toward individual authors and their work, but at the methodological framework itself.

As things currently stand, the **economic theorist** feels no obligation to take account of such 'empirical evidence' (and rightly so!), and the development of theoretical models is driven by a variety of motivating factors, such as mathematical sophistication and rigor, fecundity, generality and simplicity, to the exclusion of *empirical adequacy*. Similarly, the (theoretical) **econometrician** is content to continue devising sophisticated statistical techniques, unconcerned with the appropriateness of these methods for economic data or the unreliability and imprecision of the ensuing empirical results; 'goodness of fit' reigns supreme as the primary criterion for assessing the appropriateness of a new technique. In journal publishing a premium is placed on asymptotically 'optimal' procedures based on the *weakest* set of probabilistic assumptions, such as the Generalized Method of Moments (GMM), proposed by Hansen (1982), as well as nonparametric methods.

Caught in the middle, the **applied econometrician** stares with esteem at the mathematical dexterity of the other two, but finds himself modeling data from observable economic phenomena which are usually not the result of the 'ideal circumstances' envisaged by the theory, but of an ongoing complex data generation process which

shows no respect for *ceteris paribus* clauses, and tramples over individual agents' intentions with no regard for *rationality*.

At the start of the 21st century *econometric modeling* has made meager progress towards its primary objective of furnishing apposite empirical foundations to economics. Empirical evidence in economics has been undermined as a way to test economic theories; see Summers (1991). The primary reason for this is that the current textbook approach to empirical modeling has given rise to mountains of *unreliable evidence* that amount to nothing more than heaps of statistically meaningless 'non-regularities', which, unfortunately, are given 'theoretical meaning' (using unwarranted statistical inferences), under the guise of *identification*. Worse still, these 'non-regularities' are often used as the basis of *empirical support* for theories, as well as for policy analysis and predictions. What is conspicuously missing from current econometric modeling are genuinely reliable methods and procedures that enable one to discriminate between the numerous models and theories that could fit the same data equally well or better.

The **primary problem of current econometric modeling** is: *unreliable evidence* built upon *unwarranted inductive inferences*. One of the **main contributing factors** is that the premises of induction, the probabilistic assumptions comprising the underlying statistical model, are often incompletely specified and they are rarely probed thoroughly for departures. This is symptomatic of the current methodological framework where the emphasis on the 'quantification of theoretical relationships' gives rise to a *theory-dominated* approach to empirical modeling, which invariably fails to take into account: (i) the huge gap between theory and data, (ii) the probabilistic structure of the data, and (iii) the different ways an inference could be in error. By making assumptions about *errors,* the emphasis is placed on the least restrictive assumptions, irrespective of their verifiability, that would 'justify' the quantification method. The idea is that the less restrictive the assumptions, the less susceptible to misspecification the results are likely to be. As argued in section 8, this is clearly a deductively motivated argument that forestalls both the *reliability* and the *precision* of inference; weak assumptions give rise to imprecise inference, and non-verifiable assumptions render the substantiation of statistical adequacy impossible; see Spanos (2001b).

## 1.3 A call to arms

The above critical summary of the current state of affairs in econometrics is offered in the spirit of constructive self-criticism, and not as an indictment of a field that the author shares in the responsibility for its current plight. It is offered as a 'call to arms' to ameliorate the current predicament of econometrics with a view to fulfill the vision of the early pioneers in providing empirical foundations to economics.

In order to meet this challenge, one needs to take a retrospective of the development of econometrics during the 20th century with a view to tracing the source of particular weaknesses in current practice. In the next sections a selective summary

of some of these developments is given in light of three important influences on the development of econometrics.

*First*, how the **developments in probability and statistics** during the 20th century influenced the evolution of econometrics. The basic idea is that, even though some of the problems in econometric modeling can be traced to 'inveterate' problems in statistics, econometrics has not made judicious use of certain important developments in probability and statistics.

*Second*, how the **underlying methodological framework** of empirical modeling in economics has been affected by both its own internal dynamics as well as the broader 'currents' in philosophy of science. The basic idea is to trace the development of econometric methodology from the broad problem of bridging the gap between theory and data, in the early 20th century, to the current approach of focusing narrowly on the 'quantification of theoretical relationships'.

*Third*, how the various 20th century developments in econometrics have been **distilled into the current textbook approach** to econometrics. The demarcation of econometrics in the early 1960s by two popular textbooks has distilled the earlier developments through the Gauss-Markov 'curve fitting' perspective to the detriment of empirical modeling in economics.

For a more balanced discussion of the history of econometrics see Stigler (1965), Christ (1985), Epstein (1987), Fox (1989), Morgan (1990), Heckman (1992), Quin (1993), Hendry and Morgan (1995).

# 2 The 'unreliability' of empirical evidence

The question that naturally arises is 'why *econometric modeling* has made such meager progress toward its primary objective of furnishing apposite empirical foundations to economics?'

Making the critical arguments in the constructive way I would like is complicated by the fact that the notions of 'deduction' and 'induction' have been greatly misused in economics since the early 19th century (see Redman, 1997), and the idea of 'empirical foundations' has shifted greatly since then. Any adequate retrospective must recognize that:

(1) There is unclarity and debate as to what is required of an 'inductive component' in order that it supply genuine empirical foundations, and along side this,

(2) there are the disagreements and controversies regarding the nature and role of statistics in science in general, and in economics in particular.

It is well known that economists employ data to 'quantify theoretical relationships', the results of which are often taken as providing 'support' for their theories, but it is apparent, given the heaps of unreliable and non-incisive published evidence, that this does not suffice for pertinent empirical foundations. What is *not* manifest is the source of the unreliability of this evidence.

A textbook approach econometrician begins with a theory which he uses to derive

a theory-model in the form of functional relationships among variables of interest (exclusively determined by the theory in question). The object of the empirical modeling is to 'quantify' this theoretical relationship(s) and/or verify the theory in question. The quantification/verification is guided by both theoretical (sign, size of estimated parameters), as well as statistical considerations, such as $R^2$ 'goodness of fit', t-ratios and F-tests. Typically, a textbook applied econometrician begins with a theory model, more or less precisely specified, and proceeds to transform it into a statistical model in the context of which the *quantification* will take place. The primary objective of the quantification is to use data to provide empirical evidence *for* the theory in question. This is achieved by viewing the theory model as furnishing the systematic *error* component and a *white noise error* as appending the non-systematic component. The *quantification* is driven by a search for an 'optimal' estimator (OLS, GLS, FGLS, IV, 2SLS, 3SLS, k-class, LIML, FIML, etc.) for each different set of error assumptions. It is invariably assumed that the data chosen measure the concepts in terms of which the theory in question is articulated.

In a typical scenario, after an unsuccessful first attempt at 'quantification' using the **Classical Linear Regression** (CLR) model, an applied econometrician finds himself faced with *non-white noise residuals*, indicating that the OLS estimator $\widehat{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is no longer optimal. The recommendation from the textbook approach is to retain the original systematic component $\mathbf{X}\boldsymbol{\beta}$, but allow for *non-white errors,* by modifying the assumptions of the error term $\mathbf{u}$. The underlying rationale is that these 'error-fixing' *corrections* are used to get *better* estimators as well as *valid* standard errors for the theoretical model under quantification. For instance, if the Durbin-Watson (D-W) indicates a departure from the no-autocorrelation assumption: $E(\mathbf{uu}^\top)=\mathbf{I}_T$, 'correcting' for autocorrelation amounts to adopting the GLS estimator $\widetilde{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{V}_T^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}_T^{-1}\mathbf{y}$, where $\mathbf{V}_T=E(\mathbf{uu}^\top)$. The formal justification for this move is that $\widetilde{\boldsymbol{\beta}}$ is more optimal because $Cov(\widehat{\boldsymbol{\beta}})\geq Cov(\widetilde{\boldsymbol{\beta}})$. The success of this inference is assessed in terms of the value of the new D-W test statistic, as well as the theoretical validity of the sign and magnitude of the estimated coefficients. As argued in the sequel, the PRIMARY PROBLEM with the textbook approach is that it invariably leads to *unreliable inferences* because the inference procedures used to choose the 'appropriate' estimator *have very limited capacity to uncover the different ways such an inference could be false.*

*First*, by focusing exclusively on the error term the textbook perspective overlooks the ways in which the systematic component may be misspecified, and sometimes fails to acknowledge other implicit assumptions. Moreover, the 'error-fixing' strategy ignores the fact that by definition $\mathbf{u}=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, and thus 'fixing' the error involves modeling the systematic component indirectly; this sometimes leads to internally inconsistent models; see Spanos (1995b).

*Second*, the 'error-fixing' strategy is designed to 'save the theory', because by retaining the systematic component, it ignores alternative theories which might fit the same data equally well or even better. That is, the 'error-fixing' strategy misuses

data in ways that 'appear' to provide empirical (inductive) *support* for the theory in question, when in fact the inferences are unwarranted.

*Third*, the 'error-fixing' strategies are littered with flawed reasoning (see Spanos, 1986, 2000). For instance, when autocorrelated *residuals* are interpreted as autocorrelated *errors,* the inference to the new estimator $\widetilde{\boldsymbol{\beta}}$ (or equivalently, the new model based on $E(\mathbf{uu}^{\top}) = \mathbf{V}_T$) is unwarranted because the method used to choose it had no chance to uncover the various other forms of departures that could have been responsible for the presence of residual autocorrelation; see Spanos (1986), McGuirk and Spanos (2003). The general reasoning flaw in this *respecification* strategy is that adopting the alternative hypothesis in a misspecification test often amounts to committing the **fallacy of rejection**: in non-exhaustive cases, evidence *against* the null is erroneously interpreted as evidence *for* the alternative. Hence, after such 'error-fixing' takes place, by choosing the 'optimal' estimator associated with the new set of error assumptions, one often ends up with a *misspecified model* because these new assumptions have not been tested. This is an instance of a classic *inductive fallacious* move: infer a claim, or save a model from anomaly, by adjusting a feature to 'account for' the data, when in fact the data underdetermines this particular save (see Mayo and Spanos, 2004).

Admittedly, *misspecification testing* (assessing the validity of the probabilistic assumptions) and *respecification* (choosing a more appropriate model), raise some fundamental issues concerning *inference* and *evidence* in general, such as double use of data, which are neither trivial nor obvious in the case of empirical modeling; see Spanos (1999, 2000, 2001a). Indeed, misspecification testing is often considered as unwarranted data mining; see Kennedy (2003). This might help to explain why these problems persisted for so long in econometrics. Nevertheless, addressing these issues is necessary for improving the reliability of inference in econometric modeling.

## 2.1 A methodology of error inquiry

A philosophy of science that will help to organize the issues I wish to highlight is the **error statistical account** articulated by Mayo (1991, 1996). Unlike other philosophical accounts of evidence, whether data $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_n)$ provide *evidence for* a hypothesis $H$, is not a mere matter of *logical* or *probabilistic* relationships between $\mathbf{Z}$ and $H$, but is an *empirical* issue that requires considering how $\mathbf{Z}$ were generated and how $H$ was selected, in order to determine the overall reliability of the inference. Moreover, the error statistical account deals explicitly with the *frequentist statistical methodology* which provides the foundation for empirical modeling in economics.

What is missing from economics, when compared to other more successful sciences, is **a constructive dialogue** between theory and data, as a result of which *learning can* take place. As argued by Mayo, when a theory in a scientific discipline is found wanting after being confronted with data, the rejection should give rise to some form of *reliable inquiry* into its causes. Inquiries which should lead to *some form of learning*, however rudimentary: perhaps the data were inappropriate or inaccurate,

or the theory needs to be revised. Indeed, she considers 'learning from an anomaly' a demarcation criterion for scientific inquiry.

It is argued that what is needed for modern econometrics is a **methodology of error inquiry** that encourages detecting and identifying the different ways an inductive inference could be in error by applying efficacious methods and procedures which would detect such errors when present with very high probability. **Learning from error**, according to Mayo (1996), amounts to deliberate and reliable argument from error based on **severe testing**: "... a testing procedure with an overwhelmingly good chance of revealing the presence of specific error, if it exists – but not otherwise." (p. 7). Mere fit is insufficient for $\mathbf{Z}$ to pass $H$ severely, such a good fit must be something very difficult to achieve, and so very improbable, were $H$ to be in error. Nevertheless, the fact that $\mathbf{Z}$ were used to arrive at a good fit with $H(\mathbf{Z})$ does not preclude counting $\mathbf{Z}$ as good evidence for $H(\mathbf{Z})$ – it all depends on whether the procedure for arriving at $H(\mathbf{Z})$ would find evidence erroneously with very low probability.

Mayo (1996) applies the severe test reasoning to the Neyman-Pearson (N-P) hypothesis testing framework in order to supplement it with a post-data evaluation procedure that addresses most of the criticisms of frequentist testing; see also Mayo and Spanos (2003). In particular, it provides the reasoning to address the question "When do data $\mathbf{Z}$ provide evidence for a hypothesis or a claim $H$?" In this evidential intepretation of tests one wants to avoid tests which are either too sensitive or not sensitive enough for the inference in question.

Mayo's way of implementing the severe testing reasoning, is to localize the error probing by viewing it in the context of compartmented, but highly interconnected pieces (models), comprising the overall theory testing. These interconnected pieces come in the form of a *hierarchy of models* (primary, experimental and data) devised to link the primary hypothesis to the data; see Mayo (1996), ch. 5. The idea behind the hierarchy of models is to localize errors and apply severe testing in a piece-meal way at a level which enables one to pose questions concerning errors one at a time in an exhaustive fashion and then pieced together to provide an overall testing procedure.

Mayo (1996) proposed a methodology of error inquiry concerned with 'learning from experiments' to provide coherent philosophical foundations for **new experimentalism** (see Chalmers, 1999, 13). Her notion of 'experiment', however, is broad enough: "Any planned inquiry in which there is a deliberate and reliable argument from error." (p. 7), to render it appropriate for certain *non-experimental* situations. The idea is that in cases where no literal control or manipulation over the phenomena being modeled is possible, one can still ascertain deliberate and reliable argument from error by probing thoroughly the different ways one's claim can be wrong. This renders the error-statistical methodology eminently relevant for empirical modeling in economics. The main difficulty in implementing the severe testing reasoning arises from the fact that in certain situations one cannot accomplish the same level of cogency as in cases of carefully controlled experiments. This, however, could serve as a challenge to improve the current procedures and methods used to probe for error,

as well as to delineate the limits of empirical modeling using non-experimental data. Indeed, the hope of this author is that severe testing reasoning will help to furnish criteria that alleviate some of the discomforts associated with several methodological problems in econometric modeling such as pre-test bias, post-designation, data snooping, and other forms of exploratory data analysis. Moreover, the value of ascertaining that one is *not* able to infer that $\mathbf{Z}$ provides good evidence for $H$, is that it provides an important source of guidance as to what to try next.

In direct analogy to Mayo's hierarchy of models, the methodological framework for econometric modeling discussed in section 8 consists of a sequence of models ranging from the *theory model* to the *estimable model* that links the theory (substantive information) to the *data* via by the *statistical model* (statistical information), and the synthesis of substantive and statistical information gives rise to the *empirical model*; see fig. 1. The reliability of evidence is assessed at all levels of different models by using severe testing reasoning, quantitative as well as qualitative, to probe the different types of errors that arise in that context.

In the context of a **statistical model** the primary sources of error are:

**(I) Statistical misspecification**: some of the probabilistic assumptions comprising the statistical model (premises of induction) are invalid for data $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_n)$.

A misspecified statistical model gives rise to unreliable inferences because the *actual* error probabilities are very different from the *nominal* ones – the ones assumed to hold if the premises are true. Applying a .05 significance level t-test, when the actual type I error is .95, renders the test highly unreliable; see Mayo (1996).

**(II) Inaccurate data**: data $\mathbf{Z}$ are marred by *systematic errors* imbued by the collection/compilation process.

The inaccuracy and inadequacies of economic data as of the late 1950s has been documented by Morgenstern (1950/1963) and Kuznets (1950). Since then the accuracy of economic data has been improving steadily, but not enough attention has been paid to issues concerning how the collection, processing and aggregation of data in economics might imbue *systematic errors* that can undermine their statistical analysis; see Abadir and Talmain, (2002) for an illuminating discussion.

In the context of an **estimable model** the relevant source of error is:

**(III) Incongruous measurement**: data $\mathbf{Z}$ do not measure the concepts $\boldsymbol{\xi}$ envisioned by the theory model (e.g. intentions vs. realizations).

This is a fundamental issue because typically theory models are built on static intentions but the available data measure realizations of ongoing convoluted processes. Moreover, the overwhelming majority of economic data are collected by government agencies and private institutions, not the modelers themselves. Measuring economic theory concepts constitutes a form of 'experimentation' in the sense used by Mayo in the above quotation that needs to be recognized as such before the incongruous measurement problem can be adequately addressed.

In the context of a **theory model** the primary source of error is:

**(IV) External invalidity**: the circumstances envisaged by the theory differ

'systematically' from the *actual* Data Generating Process (DGP); *ceteris paribus* clauses, missing confounding factors, causal claims (see Hoover, 2001), etc.

This is a most fundamental issue that has been inadequately addressed in econometrics, because external invalidity effects and is affected by all the other sources of error. The typology of different models is designed to delineate questions concerning different errors and render their probing much more effective than lumping them together in one overall error term, as the textbook approach practices!

It is unfortunate that the recent methodology of economics literature (see Backhouse, 1994; Maki, 2002, inter alia) has largely ignored the methodology of econometrics in their discussions of the various philosophy of science perspectives in economics; see Stigum (2003) for a notable exception.

In summary, the current textbook approach to econometrics does not provide a methodological framework for a reliable/effective error probing inquiry that leads to learning from data. What renders the overwhelming majority of published evidence in econometrics *unreliable* is that they were produced by methods and procedures which *had very limited ability to detect errors* if, in fact, they were present.

For the attempt to trace the roots of the current state of econometrics, it will be convenient to divide the 20th century developments in econometrics into four broad periods: 1900-1930: the promising beginnings; 1930-1943: the period of zymosis;

1944-1962: the Cowles Commission; 1963-present: the textbook approach.

# 3 Econometrics 1900-1930: Promising beginnings

## 3.1 Economic methodology at the end of the 19th century

The methodological discussions during the 19th century concerning the proper method of economics focused primarily on whether the method of the physical sciences, as initially envisaged by Bacon and articulated by Newton, could or should be applied to the social sciences. There was wide-held consensus that a *primitive hypothetico-deductive method* was appropriate for economic theorizing; this might explain the appeal of Popper's falsifiability in economics (Blaug, 1992). The disagreements during the 19th century centered around the *nature* and the *method of assessment* of the initial postulates (premises) and the (deductively derived) economic laws. In very broad terms, economists like Hume, Smith, McCulloch, and Say advocated that the grounding and method of assessment for both the initial postulates and the economic laws should be *empirical* (anchored in the observable world), and they are often labelled 'inductivists'. At the other extreme, economists like Ricardo, Senior, Torrens and Cairnes opposed this and, instead, emphasized the deductive aspects of theorizing from plausible premises, arguing that the plausibility of the premises and the appositeness of the deductions did not necessitate empirical grounding or testing; these economists are often labelled 'deductivists'. There was consensus that economics differed from physics in some important respects, such as the inaccessibility of the experimental method, and the presence of innumerable factors in economic

phenomena, but there was no agreement as to how the *primitive hypothetico-deductive method* should be modified to accommodate these differences. Mill (1874, 1884) straddled both camps by arguing for inductively established initial postulates, but put the emphasis on *deductively derived* (as opposed to *experimentally* determined) economic laws establishing 'tendencies' instead of the exact predictions churned out by physical laws. He construed the empirical analysis of economic laws as establishing 'empirical uniformities' that provide a way to shed light on the question of delineating the 'pertubations' (the less important factors) influencing a particular phenomenon; see Mill (1874), p. 154. Mill's methodology influenced both Marshall (1890) and Keynes (1891) in so far as they both adopted (i) the empirical grounding of initial postulates, (ii) the importance of deductively derived economic laws, as well as (iii) the view that economic laws only establish 'tendencies'. Indeed, Marshall (1890, pp. 31-32) adopts, almost verbatim, Mill's thesis that economics is an *inexact science* (the inherent presence of innumerous pertubations) more like the field of tidology (concerned with sea tides) than the *exact science* of astronomy (Mill, 1884, pp. 587-8).

Jevons (1871, 1874) redressed the balance between induction and deduction in Mill's discussion by re-stating the hypothetico-deductive nature of the proper method for economics, introducing probability into induction, and reaffirming the role of empirical testing for deductively derived laws.

Keynes (1891) played the role of the master synthesizer of the methodological discussions during the earlier 19th century, by couching the consensus view in a way that emphasized economics as a *positive science,* and not as a *normative art*, and ascribing to **statistics** a much greater role than hitherto (pp. 342-346):

"**The functions of statistics in economic enquiries are: ... descriptive, ... to suggest empirical laws, which may or may not be capable of subsequent deductive explanation, ... to supplement deductive reasoning by checking its results,... enabling the deductive economist to test and, where necessary, modify his premises, ... measure the force exerted by disturbing agencies.**"

## 3.2 Early 20th century statistics

The most important development in statistics during the 20th century is undoubtedly the recasting of *statistical induction* into its modern form by R. A. Fisher (1922). His *modus operandi* was the notion of **a statistical model** in the form of a pre-specified 'hypothetical infinite population', with data $\mathbf{x} := (x_1, x_2, ..., x_n)$ interpreted as a *representative sample* from that 'population'.

Before Fisher, the notion of a statistical model was only implicit, and its role was primarily confined to the *description* of the distributional features of *the data in hand* using the histogram and the first few sample moments. The problem was that statisticians would use descriptive summaries of the data to claim generality beyond the data in hand. The conventional wisdom at the time is summarized by Mills (1924) as 'statistical description' vs. 'statistical induction', where the former is always valid and "**may be used to perfect confidence, as accurate descriptions of the given**

**characteristics"** (p. 549), but the validity of the latter depends on the inherent assumptions of (a) 'uniformity' for the *population* and (b) the 'representativeness' of the *sample* (pp. 550-2).

The fine line between *statistical description* and *statistical induction* was nebulous until the 1920s, for a number of reasons. *First*: "**No distinction was drawn between a sample and the population, and what was calculated from the sample was attributed to the population."** (Rao, 1992, p. 35). *Second*, it was thought that the 'inherent assumptions' for the validity of statistical induction are *not* empirically verifiable; see Mills p. 551. *Third*, there was (and, unfortunately, still is) a widespread belief, exemplified in the above quotation from Mills, that statistical description *does not* require any *assumptions* because 'it's just a summary of the data'. The reality is that there are *appropriate* and *inappropriate* summaries of the data. For instance, the arithmetic average is an inappropriate summary of a trending time series because it measures no feature of that data; the same applies to the sample variance and higher central moments over the arithmetic mean; see Spanos (2001b). It's an appropriate summary in the case where the data constitute a realization of an *Independent and Identically Distributed (IID)* process, because it represents a reliable estimate of the mean of the process.

Karl Pearson elevated descriptive statistics to a higher level of sophistication by proposing the 'graduation (smoothing) of histograms' into 'frequency curves', and introducing a whole new family of such curves; see Pearson (1895). The problem of *statistical induction* was understood by Pearson (1920) in terms of being able to ensure the 'stability' of empirical results in subsequent samples, by invoking 'uniformity' and 'representativeness' assumptions. This is a form of *induction by enumeration,* which attempts to generalize observed *events,* like '80% of A's in this data are B's', beyond the data in hand; see Salmon (1967).

## 3.3 The prospect of early 20th century econometrics

In the early 20th century, pioneers, like Moore, aspiring to provide empirical foundations to economics, had several things going for them. Substantive progress was made in the collection and systematization of economic data; statistical offices, index numbers, etc. The Neoclassical Economic theory was in the process of complete mathematization, providing them with economic models which were amenable to empirical inquiry. At the end of the 19th century there were several developments in statistical methods, like periodogram analysis, correlation and multiple correlation (regression), which seemed appropriate for analyzing economic data.

At the same time, however, these pioneers were facing an insuperable obstacle in so far as the inductive component of statistics was lacking. Statistics amounted to a toolkit of descriptive methods for summarizing data, accompanied by ill-defined allusions to inductive inference using probabilistic terms. These descriptive methods, emanating from the works of Graunt, Petty, Quetelet, Galton and Karl Pearson were largely *disjoined* from the theory of *probability* as developed by Bernoulli, Gauss,

Poisson, Cournot, Venn, Peirce and Edgeworth; see Stigler (1986). To make matters more confusing, the language of probability, with references to the 'law of error', 'probable error' and 'correlation', abounded in the statistics literature of that time. Proper integration of probability theory with statistical inference did not begin until the 1920s with R. A. Fisher (1922).

## 3.4 Henry L. Moore

In the selective reappraisal of the development of econometrics, Moore is chosen as the quintessential pioneer during the early 20th century. His empirical studies were instrumental in generating discussions on how the newly developed statistical tools by Galton, Karl Pearson and Yule can be utilized to render economics an empirical science. This early period is important because some of the crucial weaknesses of the current textbook approach can be traced back to Moore (1911, 1914).

The inductive procedure was seen by Moore at the outset as comprising two intertwining branches. One proceeds *from data* (and statistical laws established via correlation and regression) *to theory,* and the other *begins with theory* and *relates that to data* either as theory-quantification or as theory-testing. The basic objective of empirical modeling was seen to involve both *explanation* as well as *prediction.* Moore used the *data-to-theory* inductive process to study business cycles by first establishing 'statistical regularities' (using periodogram analysis, correlation and regression), and then relating them to theory. Fitting periodograms to rainfall series (1839-1910) from the Ohio Valley, he detected two cycles of 8 and 33 years, and went on to correlate them with cycles of the yield of agricultural products, such as corn, hay, oats and potatoes. He (mis-)interpreted the observed correlation as evidence for a 'causal connection' between rainfall and yield of crops, and proceeded to complete the inductive intertwine using the *theory-to-data* inductive process by evaluating the 'law of demand' for these products.

Moore fitted demand curves for a number of agricultural products using the least squares method. For illustrative purposes, let us use Moore's (1914) *demand for corn*, based on annual observations for the period 1866-1911, as a typical example. The reported 'interpolated curve' was:

$$y_t = 7.79 - 0.886x_t, \tag{1}$$

where $x_t = \frac{(100)\Delta p_t}{p_t}$ and $y_t = \frac{(100)\Delta q_t}{q_t}$, $p_t-$ average price per bushel, $q_t-$ production in bushels. His criteria for considering (1) a 'statistical law of demand' are:

> "**... *simplicity* of the formula, its *fecundity*, its *closeness of fit*, its *ease of calculation*, its *a priori validity.***" (Moore, 1908, p. 21)

A priori validity is justified in terms of the sign and magnitude of the estimated coefficients as they relate to a theoretical demand function. Taking the estimates $(7.79, -0.886)$ at face value, without any accompanying measures of uncertainty (e.g. standard errors), Moore's analysis constitutes descriptive statistics with all its limitations. However, that did not stop Moore from drawing inductive inferences of the

form 'the demand elasticity for corn is -1.129' (p. 84), or predicting the price of corn for 1912 using (1) (p. 78). 'What provides the justification for such inferences?'

### 3.4.1 Moore's empirical 'non-regularities'

Taking an anachronistic look at Moore's analysis, using today's vantage point, we can say that its primary limitation is that his criteria are inadequate to provide a sound underpinning for his inductive inferences. For Moore, 'excellence of fit' is both necessary and sufficient for establishing 'statistical laws' (see Moore, 1914, p. 17). What's missing is any discussion of the different ways the inferences might be in error. As argued in section 2, the most rudimentary way such an inference might be false is *statistical misspecification*.

From today's vantage point, the implicit statistical model for Moore's curve fitting is the **Gauss Linear model** (see Spanos, 1986, ch.18), whose assumptions are:

[1] $u_t \backsim \mathsf{N}(.,.)$, [2] $E(u_t)=0$, [3] $Var(u_t)=\sigma^2$, [4] $E(u_t u_s)=0,\ t\neq s$.

Re-estimating (1) using Moore's data yields:

$$y_t = \underset{(2.175)}{7.219} - \underset{(.083)}{0.699}x_t + \underset{(14.447)}{\widehat{u}_t}\ ,\ R^2=.622,\ s=14.447,\ n=45; \tag{2}$$

where the standard errors are reported in brackets. Before we rush into pronouncements that Moore was right after all, by quoting t-ratios, F-tests and the $R^2$, it is important to remind ourselves that the reliability of any inductive inference based on (2) depends on whether assumptions [1]-[4] are valid for the corn data. Some basic misspecification tests (see appendix) are reported in table 1, with the associated p-values (in square brackets).

| **Table 1 - Misspecification tests** | |
|---|---|
| **Non-Normality**: | $D'AP = 3.252[.197]$ |
| **Non-linearity**: | $F(1,42) = 18.560[.000095]^*$ |
| **Heteroskedasticity**: | $F(2,40) = 14.902[.000015]^*$ |
| **Autocorrelation**: | $F(2,42) = 18.375[.000011]^*$ |

The tiny p-values indicate significant departures from assumptions [2]-[4]. Hence, Moore's inferences concerning the sign and the magnitude of the coefficients $(\beta_0, \beta_1)$ are *unwarranted*; (2) constitutes *unreliable evidence*. This is not intended to be a criticism of Moore's empirical work, but to highlight the fact that the overwhelming majority of published applied papers almost 100 years later are unlikely to pass this same *statistical adequacy test*.

### 3.4.2 Moore's 'upward sloping demand curve'

'Where does this leave Moore's revolutionary vision of econometrics as the inductive component of economics?' The statistical unreliability in his empirical work did not help his mission. Without a way to distinguish between genuine *empirical regularities* and **non-regularities**, both forms of inductive inference, from data-to-theory and vice versa, were inadvertently exposed to disrepute. This came sooner than later in

the form of a 'positively sloping statistical demand curve' for *pig iron* (Moore, 1914, pp. 110-116). 'Goodness of fit' was enough for Moore to pronounce an 'interpolated line' between the percentage change of the production of pig iron $y_t$ and the average price $x_t$, 'a new type of demand curve'. Re-estimation of his equation yields:

$$y_t = \underset{(2.800)}{-4.575} + \underset{(.129)}{0.521}x_t + \underset{(16.540)}{\widehat{u}_t}\ ,\ R^2 = .288,\ s = 16.540,\ n = 42; \qquad (3)$$

where the estimates of $(\beta_0, \beta_1, \sigma^2)$ are almost identical to Moore's original estimates. Before one can establish (3) as a statistical regularity, never mind 'a new type of demand curve', one needs to ensure its statistical adequacy. It turns out that (3) *is misspecified* because most of the assumptions [1]-[4] are invalid.

Unfortunately, Moore's (1914) publication of the 'new demand curve' gave rise to fomented discussions which, despite some positive lessons being learned, painted prematurely a rather distorted picture of the issues involved in bridging the gap between theory and data; a picture that continues to befuddle econometric modeling to this day. The ensuing discussion by Lenoir, Wright, Working and Ezekiel (see Stigler (1962), Christ (1985), Fox (1989), Morgan (1990), Hendry and Morgan (1995), inter alia), focused almost exclusively on (a) the wrong sign and (b) Moore's handling of the *ceteris paribus* clause. The ensuing econometric literature on what was wrong with Moore's upward sloping 'statistical demand curve' concluded that Moore committed a blunder, because he actually estimated a *supply curve*; see Morgan (1990). In addition, Moore's attempt to address the issues raised by the ceteris paribus clause, using a combination of (i) 'data adjustment', such as 'trend ratios' (detrended data) and 'link relatives' $(x_t/x_{t-1})$, and (ii) 'model adjustment', by including other potentially relevant factors, was deemed inadequate.

Viewing this discussion in light of the above classification of the different ways an inference might be in error (I)-(IV), we can say that, given that the estimated curve (3) is statistically misspecified, there is no 'statistical regularity' to confer theoretical meaning upon. In addition, Moore's discussion of the data (p. 113) provides enough hints for the potential presence of certain serious systematic errors; the distorted dynamics brought out by the misspecification tests attest to that. To make matters worse, external validity is seriously undermined because Moore's data did not measure the *theoretical concept 'demand',* defined as representing the 'aggregate intentions to buy'; see Spanos (1995a) for an extensive discussion.

Unfortunately, as a result of the discussions in the late 1910s and early 1920s, economic theory was inadvertently granted the authority to bestow theoretical meaning upon statistical 'non-regularities' in the name of **identification.** In their attempt to flesh out the ceteris paribus clause in the case of a demand function, the econometric literature of the 1920s focused on how certain omitted factors could have caused shifts in the supply curve, but kept constant the demand curve, so as to identify the latter from data on 'quantities transacted' and the corresponding 'prices'. The end result was that the all-encompassing problem of *bridging the gap between theory and data* was, in effect, transformed into 'story telling' to identify a theory relationship in a way which ignores the probabilistic structure of the *data*.

### 3.5 The 'deductive' method re-instated?

During the first quarter of the 20th century, economic theory was going through a process of mathematization that emphasized the *deductive component* of the primitive hypothetico-deductive method, and moved away from the empirical grounding of the initial postulates and the testing of the deduced laws using statistics. Robbins (1932/1937) re-instated a *radical* form of the *deductive method*, which viewed economic theory as a system of deductions from a set of initial postulates derived from introspection, and not amenable to empirical verification. He poked fun at statistical methods applied to economics because of their dependence on the notion of a *random sample*, and argued that economic data could never qualify as such (ibid. p. 102). Hutchison (1938) attempted a rebuttal by arguing in favor of importing into economics the 'empiricism' of the newly founded 'Logical Positivism' in philosophy of science, known as the Vienna Circle; see Chalmers (1999). His case for confining economic inquiry to empirically testable propositions, however, was undermined by his emphasis on the *testability of the initial postulates* rather than the *theory's predictions*. As a result, his form of 'empiricism' did not have any discernible impact on the development of econometrics, but generated some inflamed discussions in economic methodology concerning the *nature* (realisticness) and *testability* of the initial postulates in economic theory; see Knight (1940), Friedman (1953), Machlup (1955).

The gallant efforts by Mitchell (1927), and his co-workers at the National Bureau of Economic Research, to restore faith in empirical regularities associated with economic times series, such as trends and cycles, was gravely handicapped by the inadequacy of their probabilistic framework to 'model' these regularities in a systematic way. The proper probabilistic framework for modeling time series data, in the form of stochastic processes, was not erected until the mid-1930s by Kolmogorov and Khinchine; see Spanos (1999), ch. 8.

## 4 Modern statistical inference

### 4.1 R. A. Fisher

One of Fisher's most remarkable, but least recognized, achievement was to recast *statistical induction*. Instead of starting with data $\mathbf{x} := (x_1, x_2, ..., x_n)$ in search of a descriptive model, he would interpret the data as *a representative sample* from a pre-specified 'hypothetical infinite population'; Fisher (1922, 1925). This might seem like a trivial re-arrangement of Karl Pearson's procedure, but in fact constitutes a complete reformulation of statistical induction from generalizing observed 'events' related with the data, to modeling the underlying 'process' that gave rise to the data.

The modeling process begins with a *prespecified parametric statistical model* (envisioned in the form of 'a hypothetical infinite population'), chosen so as to ensure that the observed data $\mathbf{x}_0$ can be viewed as a *random sample* from that 'population', i.e. a truly 'typical realization' of the sample $\mathbf{X}$:

"**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."** (Fisher, 1922, p. 313).

A *key concept* in Fisher's approach to inference is the **likelihood function**:

$$L(\boldsymbol{\theta};\mathbf{x}) = \ell(\mathbf{x})\cdot f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta,$$

where $\ell(\mathbf{x})$ denotes a proportionality constant, and $\{f(\mathbf{x};\boldsymbol{\theta}),\mathbf{x}\in\mathbb{R}^n_X\}$ the *distribution of the sample* - the latter 'encapsulates' the information in the statistical model. Given that the choice of a statistical model is appropriate when it renders the observed data $\mathbf{x}_0$ a 'truly typical' realization, Fisher (1922) defined the 'optimal' estimator of $\boldsymbol{\theta}$ to be the value $\widehat{\theta}(\mathbf{x}_0)\in\Theta$ which attributes to $\mathbf{x}_0$ the highest chance of occurring; calling $\widehat{\theta}(\mathbf{X})$ the *maximum likelihood estimator*. In Karl Pearson's perspective, $\widehat{\theta}(\mathbf{X})$ does not exist, and the distinction between $\theta$ and $\widehat{\theta}(\mathbf{x}_0)$ is blurred by an implicit asymptotic argument, ascribing to both the same logical status. In Fisher's perspective the *estimator* $\widehat{\theta}(\mathbf{X})$ constitutes an inference procedure purporting to zero in on the true value $\theta^*$ of $\theta$, an unknown parameter, and the *estimate* $\widehat{\theta}(\mathbf{x}_0)$ constitutes an instantiation of a $\widehat{\theta}(\mathbf{X})$ corresponding to a particular data $\mathbf{x}_0$. The reliability of the inference procedure is assessed in terms of the associated *error probabilities*.

### 4.1.1 Fisher's recasting of statistical induction

Fisher's most enduring contribution is his devising a general way to 'operationalize' errors by *embedding* the *material experiment* into a **statistical model**, and taming errors via probabilification, i.e. to define *frequentist error probabilities* in the context of a statistical model. These error probabilities are (a) *deductively* derived from the statistical model, and (b) provide a measure of the 'trustworthiness' of the inference procedure: how often a certain method will give rise to reliable inferences concerning the underlying Data Generating Mechanism (DGM). The form of induction envisaged by Fisher and Peirce (see Mayo (1996), ch. 12) is one where the reliability of the inference was emanating from the 'trustworthiness' of the procedure used to arrive at the inference. As argued by (Fisher, 1935, p. 14):

> **"In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure."**

In the parlance of inductive logic, the inference is reached by an inductive procedure which, with high probability, will reach true conclusions (estimation, testing, prediction) from true (or approximately true) premises (statistical model). This is in contrast to *induction by enumeration* where the focus is on observed 'events' and not on the 'process' generating the data.

Fisher's crucial contributions to the built-in deductive component, known as 'sampling theory', in the form of deriving the finite sampling distributions of several estimators and test statistics, recast statistical induction in terms of 'reliable procedures' based on 'ascertainable error probabilities'.

Fisher recognized that the *trustworthiness* of an inference procedure depends crucially on the *adequacy* of the assumed statistical model vis-a-vis data $\mathbf{x}_0$. Viewing

$\mathbf{x}_0$ as 'truly representative' realization of a stochastic process $\{\mathbf{X}_t,\ t \in \mathbb{T}\}$ underlying the prespecified *statistical model*, Fisher was able to recast statistical induction and render its *premises explicit* as well as *testable*; see Spanos (1999). The *soundness* of the premises is assessed using **misspecification tests**, and in turn, *statistical adequacy* ensures the reliability of inference. It is interesting to note that most of the tests proposed by Fisher (1925) were *misspecification tests* concerned with testing the Normality and IID assumptions.

Fisher's view concerning the relevance of statistics to the social sciences, and economics in particular, was clearly expressed in Fisher (1925, p. 2):

"**Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of science.**"

Fisher went on to criticize the perspective on 'statistics as a branch of economics' expounded by Bowley (1920), calling it an 'unfortunate misapprehension'.

To do justice to Fisher's numerous contributions to modern statistics will require a major digression; see Bartlett (1965), Rao (1992), Spanos (2005a).

### 4.1.2 Neyman and observational data

The question that naturally arises at this stage is 'to what extent is Fisher's modeling strategy relevant for non-experimental data?' The contention is that the insights and experience gained after three centuries of interaction between experimentation and statistics can be used to shed light on learning from observational data. Fisher focused primarily on modeling data from agricultural experiments, but Neyman had substantial experience in modeling observational data from a variety of fields including astronomy, biology, epidemiology and economics. His empirical research (see Neyman, 1950, 1952) has effectively extended Fisher's view of the modeling process by viewing statistical models broadly as 'chance mechanisms', emphasizing the gap between the *phenomenon of interest* and the *statistical model* chosen to 'represent' it, and distinguishing between *structural* and *statistical models* (Neyman, 1976, Lehmann, 1990). These are important innovations which are of paramount interest to econometrics.

## 4.2 Outstanding issues in modern statistics

The probabilistic foundations of statistics, as we understand them today, were in place by the 1950s, but the field was still struggling with the proper form of its own *inductive reasoning*; see Mayo (2004). Fisher was arguing for 'inductive inference' spearheaded by his significance testing (see Fisher, 1955, 1956), and Neyman was arguing for 'inductive behavior' based on Neyman-Pearson testing (see Neyman, 1956); see Lehmann (1993). Neither account of inductive reasoning, however, provided an adequate account of how to address the question 'when do data $\mathbf{Z}$ provide *evidence* for (or against) a hypothesis or a claim $H$?' The *pre-data* error–probabilistic account of inference seemed inadequate for a *post-data* evaluation of the inference reached; see Hacking (1965), pp. 99-101.

Pre-data *error probabilities* in estimation (see Fisher, 1922,1925, 1935), are used to determine the 'optimality' of an estimator $\widehat{\boldsymbol{\theta}}$ (stemming from its capacity to zero

in on the *true value* $\boldsymbol{\theta}^*$), using its sampling distribution evaluated under the 'true state of nature' (*factual reasoning*), say $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. This factual reasoning, however, renders these error probabilities impossible to operationalize *post-data*, because of its dependence on the unknowable $\boldsymbol{\theta}^*$. Symptomatic of this difficulty is the confusion surrounding the use of error probabilities in conjunction with confidence intervals.

In contrast to estimation, *error probabilities of type I and II* in hypothesis testing are determined by evaluating the sampling distribution of a test statistic using *counterfactual reasoning*: under hypothetic values of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (the parameter space) relating to the *null* as well as *alternative hypotheses*. Fisher's *significance testing* utilizes *counterfactual reasoning* under the null, but Neyman and Pearson (1933) extend that to scenarios under alternative hypotheses. N-P *pre-data error probabilities* provided a theory for 'optimal' tests; based on their capacity to discriminate between true and false hypotheses. The counterfactual reasoning in testing *poses sharper questions and elicits more informative answers* whose 'trustworthiness' can be assessed using error probabilies, without invoking the true $\boldsymbol{\theta}^*$. It turns out that the *severe testing reasoning* (section 2.1) can be used to supplement Neyman-Pearson testing with a post-data evaluation component; see Mayo (1996). This supplement can be utilized to address the problem raised by Hacking (1965), as well as several criticisms leveled against frequentist testing; see Mayo and Spanos (2003).

In addition to the problems of (a) the proper form of its own *inductive reasoning*, (b) structural vs. statistical models, (c) post-data interpretation of Neyman-Pearson testing, (d) the appropriate error probabilities in multiple testing situations, (e) double use of data, (f) pre-designation vs. post-designation, one needs to address the chronic problems of (g) *statistical model selection (specification)* and (h) *validation*. According to Rao (2004):

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

In section 8, it is argued that some of these inveterate problems, including statistical model specification and validation, can be addressed in the context of a Fisher-Neyman modeling framework when judiciously adapted/extended using the error-statistical account based on severe testing reasoning (section 2.1).

## 4.3 Time Series Modeling

Although time series provided the main type of data for statistical modeling from its early beginnings in the 17th century, modeling using descriptive statistics techniques was problematic because it ignored the inherent (a) heterogeneity and (b) temporal dependence exhibited by such data. Initially, time series data were treated as if they were realizations of IID processes, but by the late 19th century periodogram and correlation analysis was available to capture these inherent features. However, the search for periodicities using periodograms, and the use of temporal correlations, conflated the two features of time series data (a)-(b), and invariably led to unreliable inferences; false periodicities and nonsense correlations (see Morgan, 1990). The

first statistical models for time series data were proposed by Yule (1927) and Slutsky (1927) in the form of the Autoregressive (AR(p)) and Moving Average (MA(q)) formulations, respectively. These models, however, were viewed in the context of the curve fitting tradition because the necessary probabilistic perspective was yet to be developed; Kolmogorov and Khinchine introduced the notions of stationarity, ergodicity and Markov dependence in the mid 1930s. Using these new concepts, Wold (1938) provided a probabilistic perspective that united the two formulations to form the ARMA(p,q) model; see Spanos (2001a).

## 4.4 Bayesian Statistics

The Bayesian approach to the problem of inductive inference is framed in terms of 'revising' one's prior beliefs $\pi(\boldsymbol{\theta})$ concerning the parameters of interest $\boldsymbol{\theta}$, in view of the data $\mathbf{x}_0$, by evaluating the *posterior* distribution:

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}_0) \propto \pi(\boldsymbol{\theta}) \cdot L(\boldsymbol{\theta}; \mathbf{x}_0), \ \boldsymbol{\theta} \in \Theta,$$

where $L(\boldsymbol{\theta}; \mathbf{x}_0)$ denotes the likelihood function; $\pi(\boldsymbol{\theta} \mid \mathbf{x}_0)$ provides the basis of any inference concerning $\boldsymbol{\theta} \in \Theta$. The Bayesian inductive inference account frames induction in terms of the change in the prior probability for a claim, say $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, in view of data $\mathbf{x}_0$, to yield the posterior $\pi(\boldsymbol{\theta}_0 \mid \mathbf{x}_0)$. In particular, $\pi(\boldsymbol{\theta}_0 \mid \mathbf{x}_0) > \pi(\boldsymbol{\theta}_0)$ is interpreted as data $\mathbf{x}_0$ providing *evidence for* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

The Bayesian approach to statistical inference was built on a subjective interpretation of probability by Bayes and Laplace, and was criticized by Venn, Boole (see Hald, 1998), Peirce (1878), as well as Fisher (1922). Keynes (1921) proposed a subjective 'rational degree of belief' interpretation of probability as a 'logical property' of propositions conditional on data, which had an important impact on the philosophy of confirmation (Carnap, 1962), but less impact on statistical inference by influencing Jeffreys with regard to the weaknesses of the frequency interpretation of probability. Jeffreys (1939) proposed a notion of 'objective' *non-informative priors*, which often lead to Bayesian inferences analogous to the frequentist approach, but, in addition, enable one to attach probabilities to parameters and thus to hypotheses. Further revival of the Bayesian approach was fervently promoted by Savage (1954) through reviving the work of Ramsey and de Finetti based on a purely *personal/subjective* interpretation of probability. Zellner (1971) was instrumental in re-introducing Jeffreys (1939) into both econometrics and statistics.

## 4.5 Nonparametric Statistics

In an attempt to reduce the reliance of statistical inference propositions (optimal estimators and test statistics) on the Normal distribution, the statistics literature in the early 1940s initiated an approach to inference designed to replace Normality with more general forms of distributional assumptions, such as continuity and unimodality of the density function; these methods became known as *nonparametric* or *distribution free*; see Scheffe (1943) for an early survey. These methods relied mostly on *order*

*and rank statistics*, leading to inference propositions which are 'robust' to departures from Normality; see Lehmann (1975).

Another important line of development in this literature was initiated by Rosenblatt (1956) who proposed a nonparametric estimator of the density function based on 'smoothing' the histogram; this gave rise to *kernel smoothing* methods, which have been extended to the estimation of multivariate and conditional densities as well as regression and skedastic functions; see Simonoff (1996).

# 5 Econometrics 1930-1942: a period of zymosis

The period of zymosis, 1930-1942, is largely dominated by the efforts of three eminent pioneers: Ragnar Frisch, Jan Tinbergen and Tjalling Koopmans. The prevailing view in the 1920s was that the application of statistical inference tools in economics was questionable because: (i) it is impossible to apply the 'experimental method' to the study of economic phenomena, (ii) there is always an unlimited number of potential factors influencing economic phenomena - hence the invocation of *ceteris paribus* clauses-, (iii) economic phenomena are intrinsically heterogeneous (spatial and temporal variability), and (iv) economic data are vitiated with errors of measurement.

## 5.1 Ragnar Frisch

Frisch advanced the 'curve fitting' perspective of the early statistical efforts on estimating demand/supply functions and analyzing business cycles to a higher level of sophistication, but at the same time he contributed significantly to forestalling the introduction of the probabilistic perspective into econometric modeling.

In relation to estimating demand/supply functions, Frisch (1934) treated all observable random variables, including prices and quantities, $(x_{kt},\ k = 1, 2, ..., m)$, as a priori symmetrical and comprising two orthogonal components: a *systematic* (non-random) unobservable $(\mu_{kt})$, and an *erratic* (random) unobservable $(\varepsilon_{kt})$:

$$x_{kt} = \mu_{kt} + \varepsilon_{kt},\ k = 1, 2, ..., m, \tag{4}$$

He then considered 'curve fitting' (which he called *confluence analysis*) as a problem in vector space geometry where the number $r$ of 'independent' linear relationships among the systematic components:

$$a_{0j}\mu_{0t} + a_{1j}\mu_{1t} + \cdots + a_{mj}\mu_{mt} = 0,\ j = 1, 2, ..., r, \tag{5}$$

is determined by the singularity of the matrix $[\mu_{kt}]_{k=1,..,m}^{t=1,..,T}$. Since this matrix is not observable, these linear relationships can only be determined using the data matrix $[x_{kt}]_{k=1,..,m}^{t=1,..,T}$, which includes the errors.

Frisch decided that Fisher's methods were inappropriate for non-experimental data, and consciously disparaged the probabilistic perspective as only applicable to cases where experimental control was possible. His answer to the non-experimental nature of economic data came in the form of a **hybrid specification** of a *deterministic theory* (5) (referred to as *structural*) combined with *erratic errors* carrying the probabilistic understructure.

It is interesting to note that in a reply to a question by E. B. Wilson in 1940 concerning Frisch's scheme, Fisher charged economists of perpetuating a major *confusion* between 'statistical' regression coefficients and "... coefficients in abstract economic laws"; see Bennett (1990), p. 305.

## 5.2 Jan Tinbergen

Tinbergen's most influential contribution came in the form of the two monographs (1939) on "Statistical Testing of Business Cycle Theories", where he proposed the first *macroeconometric models* for the US economy in the form of a system of dynamic equations; see Morgan, 1990, for an extensive discussion. Despite the title these monographs had very little to do with *statistical testing* as such. The emphasis was placed almost exclusively on estimation, with only occasional references to tests of significance for regression coefficients. It was an attempt to combine Fisher's theory of estimation relating to regression with Frisch's confluence analysis by utilizing Koopmans (1937) hybrid formulation.

For our present purposes, Tinbergen's empirical work is interesting because it represents the first version of the methodological framework that was to dominate empirical modeling in econometrics for the rest of the century. Tinbergen extended Moore's combination of statistical 'goodness of fit' criteria and 'economic validity' as the basis of 'empirical reliability' by utilizing more modern statistical techniques to allow some leeway in determining the lags and trends empirically (ibid. p. 26). Like Moore, Tinbergen (1939) raised the issue of 'static theory' vs. 'time series' (dynamic) data, and argued in favor of utilizing 'sequence analysis' instead of the long run equilibrium in developing theory models. He went beyond Moore by reporting standard errors of estimators and making extensive use of graphical techniques which included a plot of the residuals.

In reviewing Tinbergen (1939), Keynes (1939) raised several problems associated with the use of regression in econometrics. Viewed from a current perspective, the issues raised by that debate comprise a mixture of substantive and statistical problems: (i) the inclusion of all relevant factors, (ii) the measurability of certain factors, (iii) the interdependence of these factors, (iv) the use of ceteris paribus clauses, (v) the spatial and temporal heterogeneity of economic phenomena, (vi) the functional form of estimated relationships, and (vii) the specification of the dynamics, including lags and trends; see Morgan (1990), Epstein (1987). The perceived outcome of the debate, which gave Keynes the moral high ground and Tinbergen the success on pragmatic grounds, did very little to restore the reliability of empirical evidence in economics.

## 5.3 Tjalling Koopmans

The tension between the descriptive 'curve fitting', as extended by Frisch, and R. A. Fisher's probabilistic perspective is best demonstrated in Koopmans (1937). His stated intention was to integrate Frisch's confluence analysis with Fisher's statistical inference (sampling) framework. More than any other econometrician of his

time, Koopmans appreciated Fisher's statistical inference framework, and made an impressive effort in his 1937 book to bring into econometrics some of its most important features, including the distinction between unknown 'parameters' and their 'estimators', the problems of 'specification, estimation and distribution', as well as the method of Maximum Likelihood.

Despite his impressive efforts, he could not integrate Frisch's 'curve fitting' and Fisher's probabilistic perspectives because there was a fundamental incompatibility between the two perspectives. In light of Wold (1938), one can argue that Frisch's scheme (4) was probabilistically inappropriate for time series data, because his orthogonal decomposition $\mathbf{X}_t = \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t$, would render $\boldsymbol{\varepsilon}_t = (\mathbf{X}_t - \boldsymbol{\mu}_t)$ *non-systematic*, only if the systematic component is defined by $\boldsymbol{\mu}_t = E(\mathbf{X}_t \mid \mathcal{D}_t)$, and the conditioning information set $\mathcal{D}_t$ includes the past of the process, e.g. $\mathcal{D}_t = \sigma(\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, ...\mathbf{X}_0)$, where $\sigma(\mathbf{Z})$ denotes the sigma-field generated by $\mathbf{Z}$; see Stigum (1990). In such as case, $\boldsymbol{\mu}_t$ could not be realistically viewed as a deterministic function of $t$.

# 6 Econometrics 1943-1962: Haavelmo and the Cowles Commission

## 6.1 Haavelmo 1944

Haavelmo was a student and an assistant to Frisch in Oslo, but spent the late 1930s and early 1940s in the USA, where he interacted with two early pioneers in statistics, Neyman and Wald. The single most significant publication in econometrics during the 20th century is arguably Haavelmo (1944). This appraisal is based not only on its actual influence on the development of econometrics, via the Cowles Commission, but also on its potential impact, which never materialized.

The primary **influences of the monograph** include: (i) introducing the *probabilistic foundations* as well as the *methods of modern statistical inference* (maximum likelihood and Neyman-Pearson testing) into econometrics, (ii) formalizing the notion of *interdependence* (**simultaneity**) in economic phenomena, (iii) emphasizing the importance of *autonomous* (**structural**) relationships in empirical modeling (see Christ (1985), Epstein (1987), Morgan (1990)), and (iv) introducing different types of **models** as the primary tools for empirical modeling. Haavelmo argued convincingly that the probabilistic perspective of stochastic processes provides the proper framework for: (a) modeling time series data which exhibit both dependence as well as heterogeneity, and (b) embedding theory models into statistical (probabilistic) models for the purposes of inference:

"**... For no tool developed in the theory of statistics has any meaning ... without being referred to some stochastic scheme.**" (ibid., p. iii)

"**... economists might get more useful and reliable information (and also fewer spurious results) out of their data by adopting more clearly formulated probabilistic models; and that such formulation might help in suggesting what data to look for and how to collect them.**" (p. 114)

He emphasized how the *joint distribution of the observables* can be used, not only for inference, but also as a basis for *statistical model specification*:

"**The class of $n$-dimensional probability laws can, therefore, be considered as a rational classification of all a priori conceivable mechanisms that could rule the behavior of the $n$ observable variables considered.**" (p. 49)

It's no exaggeration to say that the last quotation reads like an informal stating of **Kolmogorov's theorem**: Under some general regularity conditions, the probability structure of the stochastic process $\{X_t,\ t \in \mathbb{T}\}$ is *completely specified* if one is given the joint probability distribution $F_n(x_{t_1}, x_{t_2}, \ldots x_{t_n})$, for all $n \geq 1$ and all $(t_1, ..., t_n) \in \mathbb{T}$; see Kolmogorov (1933), Doob (1953). This was formalized and extended to give rise to the Probabilistic Reduction approach (see section 8.1).

The **potential impact of the monograph** derives from Haavelmo's insightful diagnosis of the methodological problems bedeviling empirical modeling in economics in the 1930s. He proposed a broad methodological framework that mapped out clearly *the gap between theory and data*, and proposed several systematic ways to contemplate how to bridge this gap:

"**When we set up a system of theoretical relationships and use economic names for the otherwise purely theoretical variables involved, we have in mind some *actual experiment*, or some *design* of an experiment, which we could at least imagine arranging, in order to measure those quantities in real economic life that we think might obey the laws imposed on their namesakes.**" (p. 6)

If we return to the different ways (I)-(IV) an inference can be wrong (see section 2.3), we can see that Haavelmo made a serious attempt to map out the issues (III)-(IV) associated with *substantive inadequacies*. His suggestion for dealing with **incongruous measurement** was to distinguish between 'true', 'theoretical' and 'observational' variables (p. 7): " '**true' variables represent our ideal as to accurate measurements of reality "as it is in fact"**," 'theoretical' variables "**are the true measurements that we should make if reality were actually in accordance with our theoretical model"**, and 'observational' variables are those quantified by the available data; see pp. 5-7. He argued that, given a theoretical model and its associated design, as well as a set of observations", two questions need to be posed:

**(1) Have we actually observed what we meant to observe, i.e. can the given set of observations be considered as a result obtained by following our design of "ideal" experiments?**

**(2) Do the "true" variables actually have the properties of the theoretical variables?"** (p. 7)

"**... one should study very carefully the actual series considered and the conditions under which they were produced, before identifying them with the variables of a particular theoretical model.**" (p. 7)

His suggestion to bring out the **external invalidity** issue was to compare the conditions envisaged by the theory model with those of the 'actual DGM', and bridge

the gap by modifying the former to 'resemble' the latter:

"**We try to choose a theory and a design of experiments to go with it, in such a way that the resulting data would be those which we get by passive observation of reality.**" (p. 14)

Unfortunately for econometrics, Haavelmo's suggestions concerning the gap between theory and data were not pursued further by the subsequent literature, perhaps because their implementation is non-trivial; see Spanos (1989).

## 6.2 The Cowles Commission: 1932-1954

The research of the *Cowles Commission for Research in Economics* during the period 1932-9, at Colorado Springs, was dominated by the *descriptive statistics*, 'curve fitting' perspective as exemplified by the work of Davis (1941), Roos (1934) and Tintner (1940). After the Cowles Commission was moved to Chicago (1939-1954), its main research agenda changed drastically and focused on formalizing and extending the work of Haavelmo (1943, 1944), with a view to improving the macroeconometric modeling exemplified by Tinbergen (1939); see Morgan (1990). The focus of the Cowles Commission group in Chicago narrowed the intended scope of Haavelmo's monograph down to the problem of inference in the context of the Simultaneous Equations Model (SEM), utilizing the modern statistical inference methods developed by Fisher and Neyman-Pearson. Let us summarize their main results.

**The Structural Form (SF)** of the model often comes in the form of a system of simultaneous linear equations:

$$\mathbf{\Gamma}^{\top}\mathbf{y}_t = \mathbf{\Delta}^{\top}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t \backsim \mathsf{N}(\mathbf{0}, \mathbf{\Omega}),\ \boldsymbol{\alpha} = \mathbf{h}\left(\mathbf{\Gamma}, \mathbf{\Delta}, \mathbf{\Omega}\right) \in \Phi,\ t \in \mathbb{T}, \tag{6}$$

where $\mathbf{y}_t : m \times 1$ denotes a vector of *endogeneous* variables and $\mathbf{x}_t : k \times 1$ a vector of *exogenous* variables; let $\boldsymbol{\alpha}$ denote a $n_1 \times 1$ vector of the (structural) unknown parameters in $(\mathbf{\Gamma}, \mathbf{\Delta}, \mathbf{\Omega})$. The corresponding **Reduced Form (RF)** is:

$$\mathbf{y}_t = \mathbf{B}^{\top}\mathbf{x}_t + \mathbf{u}_t, \quad \mathbf{u}_t \backsim \mathsf{N}(\mathbf{0}, \mathbf{\Sigma}),\ \boldsymbol{\theta} = \mathbf{g}\left(\mathbf{B}, \mathbf{\Sigma}\right) \in \Theta, \quad t \in \mathbb{T}, \tag{7}$$

where $\boldsymbol{\theta}$ denotes a $n_2 \times 1$ $(n_2 \geq n_1)$ vector of the unknown parameters in $(\mathbf{B}, \mathbf{\Sigma})$. The two forms are interrelated parametrically via the system of equations:

$$\text{(i) } \mathbf{B}(\boldsymbol{\theta})\mathbf{\Gamma}(\boldsymbol{\alpha}) = \mathbf{\Delta}(\boldsymbol{\alpha}), \quad \text{(ii) } \mathbf{\Omega}(\boldsymbol{\alpha}) = \left(\mathbf{\Gamma}^{\top}(\boldsymbol{\alpha})\mathbf{\Sigma}(\boldsymbol{\theta})\mathbf{\Gamma}(\boldsymbol{\alpha})\right).$$

**Identification** of the structural parameters $\boldsymbol{\alpha}$ takes the parameters $\boldsymbol{\theta}$ as given, and poses the question: "can one 'solve' *uniquely* for $\boldsymbol{\alpha} = \mathbf{H}(\boldsymbol{\theta})$ the implicit system of equations (i)-(ii)?"

When the SF (6) is **just identified**, the mapping: $\boldsymbol{\alpha} = \mathbf{H}(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \mathbf{\Theta}$, is bijective, defining a *reparameterization* with $\boldsymbol{\theta} = \mathbf{H}^{-1}(\boldsymbol{\alpha}),\ \boldsymbol{\alpha} \in \mathbf{\Phi}$. When the structural model is **overidentified** the mapping is surjective, defining a *reparameterization/restriction,* because $\mathbf{H}^{-}(.)$, the *pre-image* of the mapping $\mathbf{H}(.)$, imposes restrictions on the statistical parameters:

$$\boldsymbol{\theta}^{*} = \mathbf{H}^{-}(\boldsymbol{\alpha}),\ \boldsymbol{\alpha} \in \mathbf{\Phi} \Rightarrow \boldsymbol{\theta}^{*} \in \mathbf{\Theta}_1 \subset \mathbf{\Theta}.$$

The test of overidentifying restrictions is based on the implicit hypotheses:
$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^*$ vs. $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^*$; see Spanos (1986), ch. 25.

Using modern statistical inference methods, the Cowles Commission group was commendably successful in addressing the technical problems of *identification*, *estimation*, *testing* and *prediction* in the context of the SEM; see Koopmans, 1950, Hood and Koopmans, 1953. Their crucial influences on the development of econometrics during the second half of the 20th century come in the form of (a) fortifying the use of modern statistical methods, and (b) focusing attention on estimating 'autonomous structure' for empirical modeling purposes in econometrics. A retrospective view of the Cowles Commission's influence reveals that the research agenda of this group was, in some respects, a lot more influential in shaping the research in *theoretical econometrics* until the 1970s, than in influencing *empirical modeling* in econometrics. The mainstream theoretical research effort of the 1960s and 1970s was primarily channeled into simplifying and extending the work of the Cowles Commission group, especially in estimation (LIML, FIML, 2SLS, 3SLS, IV, k-class, etc.) and identification of the SEM; see Hendry (1976) for a unifying survey. The narrowing of Haavelmo's blueprint by the Cowles Commission down to the problem of *simultaneity bias* of the OLS estimator, and the additional rigidity introduced into his methodological framework, contributed to its early perceived failure as a general approach to empirical modeling; indeed, most of the protagonists turned away from econometrics in the mid 1950s; see Epstein (1987). The rigidity, as well as the restrictiveness of the probabilistic assumptions on the errors, were perceptively criticized by Orcutt (1951). Moreover, the empirical work on macroeconometric modeling that followed the Cowles Commission era by Klein and Goldberger (1955) brought out the inappropriateness of the proposed framework.

As a result of the Cowles Commission methodology, the theory-dominated methodological framework of the 1920s and 1930s was greatly fortified against any predilections towards empirical regularities and the data-to-theory inductive process. Their *theory-dominated* perspective commences with a fully specified structural model which, in turn, determines the reduced form in conjunction with the probabilistic structure of the error terms. A way to explain this perspective is to see it as an attempt to imitate the 'perceived' form of empirical modeling in fields like physics, where the 'laws' are clearly known a priori and observation can only help to quantify them; discrepancies between the laws and the data can only arise from measurement and/or sampling errors. Needless to say, this perceived form of empirical modeling in physics is misleading; see Cartwright (1983).

As argued in Spanos (1986, 1990), the reduced form is the (implicit) *statistical model* in the context of which the structural model is embedded. Hence, the statistical model, not only 'has *no* life of its own', but it is wholly dominated by the theory model in both form and probabilistic structure via the error term assumptions. The fact of the matter is that if any of the assumptions [1]-[5] (see table 2, where $\mathbf{B}^\top = \left(\boldsymbol{\beta}_0^\top, \mathbf{B}_1^\top\right)$, $\mathbf{x}_t := (1, \mathbf{x}_{1t})$) turn out to be invalid for data $\mathbf{Z} := (\mathbf{X}, \mathbf{y})$, any inference

based on the estimated structural model (6) is likely to be unreliable. It is interesting to note that the limited use of the probabilistic perspective by the Cowles Commission was raised by Vining (1949), p. 85.

**Table 2 - The Multivariate Linear Regression (MLR) Model**

$$\mathbf{y}_t = \boldsymbol{\beta}_0 + \mathbf{B}_1^\top \mathbf{x}_{1t} + \mathbf{u}_t,\ t \in \mathbb{T},$$

| | | |
|---|---|---|
| [1] | **Normality:** | $D(\mathbf{y}_t \mid \mathbf{x}_{1t}; \boldsymbol{\psi})$ is Normal, |
| [2] | **Linearity:** | $E(\mathbf{y}_t \mid \mathbf{X}_{1t} = \mathbf{x}_{1t}) = \boldsymbol{\beta}_0 + \mathbf{B}_1^\top \mathbf{x}_{1t}$, linear in $\mathbf{x}_t$, |
| [3] | **Homoskedasticity:** | $Cov(\mathbf{y}_t \mid \mathbf{X}_{1t} = \mathbf{x}_{1t}) = \boldsymbol{\Sigma}$, free of $\mathbf{x}_{1t}$, |
| [4] | **Independence:** | $\{(\mathbf{y}_t \mid \mathbf{X}_{1t} = \mathbf{x}_{1t}),\ t \in \mathbb{T}\}$ - independ. process, |
| [5] | **t-homogeneity:** | $(\boldsymbol{\beta}_0, \mathbf{B}_1^\top, \boldsymbol{\Sigma})$ are not functions of $t \in \mathbb{T}$, |

where $\mathbf{B}_1 := Cov(\mathbf{X}_{1t})^{-1} Cov(\mathbf{X}_{1t}, \mathbf{y}_t)$, $\boldsymbol{\beta}_0 := E(\mathbf{y}_t) - \mathbf{B}_1^\top E(\mathbf{X}_{1t})$,

$$\boldsymbol{\Sigma} := Cov(\mathbf{y}_t) - Cov(\mathbf{y}_t, \mathbf{X}_{1t}) Cov(\mathbf{X}_{1t})^{-1} Cov(\mathbf{X}_{1t}, \mathbf{y}_t).$$

In conclusion, the Cowles Commission group provided a crucial impetus to the development of econometrics by presenting econometricians with technical problems arising from the SEM, but contributed little to the advancement of the methodology of empirical modeling. Their methodological framework constituted a retrogression from that of Haavelmo (1944), especially as it relates to the error probing inquiries concerning the ways (I)-(IV) that an inductive inference might be false. Theirs was a theory-dominated approach to empirical modeling that allowed no real role for the systematic statistical information in the data.

# 7 Econometrics 1963 - present: the textbook approach

## 7.1 Formulating the textbook approach

The textbook approach to econometrics was shaped in the early 1960s by Johnston (1963) and Goldberger (1964), by demarcating its intended scope to be the 'quantification of theoretical relationships in economics', and formalizing the Gauss-Markov (G-M) perspective; both elements have dominated econometrics to this day. This is both a tribute to the appeal of the edifice built by these two influential pioneers, and an indictment to the uncritical attitude of the subsequent literature. It is argued that the G-M perspective constitutes a return to the earlier 'curve fitting' paradigm, where the theory dominates the specification of the statistical model and, in conjunction with 'white noise' **error terms**, decrees the probabilistic structure the available data is alleged to have.

### 7.1.1 The Gauss-Markov blueprint

The key statistical model of the Gauss-Markov perspective is the so-called **Classical Linear Regression (CLR)**, specified in the form of:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \text{ where } \mathbf{y} : T \times 1,\ \mathbf{X} : T \times k, (T > k)$$

together with the **Gauss-Markov (G-M) assumptions**:

$$(\mathbf{1}) \quad E(\mathbf{u}) = \mathbf{0}, \qquad (\mathbf{2}) \quad E(\mathbf{uu}') = \sigma^2 \mathbf{I}_T,$$
$$(\mathbf{3}) \quad \text{Rank}(\mathbf{X}) = k, \qquad (\mathbf{4}) \quad \mathbf{X} \text{ is fixed in repeated samples.}$$

The error term is thought to include (i) approximation errors, (ii) omitted factors, (iii) variability in human behavior, (iv) aggregation errors, and (v) errors of measurement; see Valavanis (1959). A glance at this list of potential errors suggests that one is facing an impossible task in trying to disentagle, not to mention probe thoroughly, all these possible sources of error in the context of one error term.

A caricature of the methodological framework of the textbook approach begins with a theory which one uses to derive a theory-model in the form of functional relationships among variables of interest (exclusively determined by the theory in question). The object of the empirical modeling is to 'quantify' this theoretical relationship(s) and/or verify a theory. The quantification/verification is guided by both theoretical (sign, size of estimated parameters), as well as statistical considerations, such as $R^2$ 'goodness of fit', t-ratios and F-tests. The textbook perspective shares with the Cowles Commission the predominance of the theory in the specification of the statistical model, but it allows for more flexibility in so far as statistical considerations are allowed to influence the final choice of the model. In addition, the error term, that was primarily viewed as relating to autonomous shocks in the Cowles Commission tradition, becomes a 'catch-all' component carrying the probabilistic structure of the underlying statistical model. Indeed, the error term in the G-M perspective takes center stage in determining the sampling properties of inference procedures, and is endowed with 'a random life of its own'; any departures from assumptions **(1)-(2)** are addressed by 'modeling' the error term! 'How did this change from the Cowles Commission perspective come about?'

A case can be made that Richard Stone played an important role in influencing the forging of the textbook approach to econometrics; see Gilbert (1991). The methodological framework is clearly discernible in Stone (1954a), and the CLR model with assumptions (1)-(4) is clearly specified in that form by Stone (1954b, ch. 19). The roots of the G-M assumptions can be traced back to Aitken (1934) and David and Neyman (1938), and the extension of assumption (2) to the case where $E(\mathbf{uu}') = \mathbf{\Omega}_T \neq \sigma^2 \mathbf{I}_T$, can be traced back to the Cambridge tradition of Yule (1926), Cohrane and Orcutt (1949) and Durbin and Watson (D-W) (1950). In particular, the last two papers had a lasting influence in the shaping of the textbook approach, both in terms of testing for the presence, as well as the modeling, of error-autocorrelation. Indeed, the D-W test was the first misspecification test to be widely applied.

This textbook approach blueprint begins with the CLR model with assumptions (1)-(4), and views the other statistical models of interest in econometrics as variations/extensions of this model. The central axis around which the textbook modeling strategy unfolds, however, is not the choice of an appropriate statistical model *in view of the data*, but the choice of an 'optimal' estimator in view of the theory model. Textbook econometrics revolves around the Gauss-Markov theorem as justifying the use of

the OLS estimator under assumptions **(1)-(4)**. Whenever any one of these assumptions is violated, the OLS estimator loses its optimality and a better estimator is sought: GLS, FGLS, Ridge, Instrumental Variables (IVs), 2SLS, 3SLS, LIML, FIML, k-class, double-k class etc.; see Kennedy (2003) for a clear exposition. A bird's eye view of this blue-print is as follows.

**Chapter 1**. The *Gauss-Markov theorem.* Under assumptions **(1)-(4)**, the *OLS estimator* $\widehat{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, is the *Best, Linear, Unbiased Estimator* (BLUE) of $\boldsymbol{\beta}$. That is, $\widehat{\boldsymbol{\beta}}$ is a relatively efficient estimator (minimum variance) within the class of unbiased estimators, which are also restricted to be linear functions of $\mathbf{y}$. Every subsequent chapter deals with a particular violation of assumptions **(1)-(4)** and discusses its consequences for $\widehat{\boldsymbol{\beta}}$, as well as the remedies that give rise to another 'optimal' estimator. **Chapter 2**. **Non-zero mean:** $E(\mathbf{u}) \neq \mathbf{0}$. **Chapter 3**. **Autocorrelation**: $E(\mathbf{uu}') = \boldsymbol{\Omega}_T \neq \sigma^2\mathbf{I}_T$. **Chapter 4**. **Heteroskedaticity**: $E(\mathbf{uu}') = \boldsymbol{\Lambda}_T = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_T^2)$. **Chapter 5**. **Multicollinearity**: Rank$(\mathbf{X}) < k$. **Chapter 6**. **Stochastic regressors**: $E(\mathbf{X}^\top\mathbf{u}) \neq \mathbf{0}$, **(a) Errors-in-variables, (b) Simultaneous Equations.**

As we can see, the textbook approach replaced the Simultaneous Equations Model (SEM) of the Cowles Commission from center stage with the Classical Linear Regression (CLR) model, viewing the SEM as a variant of the CLR. A comparison between Johnston (1963) and Greene (2002), perhaps the most successful recent graduate textbook in econometrics, reveals that, besides raising the level of mathematical sophistication of statistical inference, the main difference between them is that the latter includes several new statistical models of interest in econometrics, together with their associated statistical inference propositions - primarily estimation. Additional models, include **discrete dependent, limited dependent and duration models for cross-section data**, **models for time series data (ARMA(p,q), ARIMA(p,d,q), VAR(p), ARCH), and models for panel data**, which were introduced into textbook econometrics during the 1980s and 1990s. There is no doubt that the impressive developments in econometrics came in the form of more and better *statistical models* and their *associated estimation procedures.* These additions, however, have the hallmark of the Gauss-Markov 'curve fitting' perspective, which has a *theory-dominated* view of empirical modeling, with the emphasis placed on 'the quantification of theory models'.

## 7.2 The critics of the textbook approach

In an attempt to remedy the perceived failings of the textbook approach, in this subsection we will briefly mention a number of alternative approaches to empirical modeling in econometrics from the point of view of how they suggest modifying the textbook approach in order to address some of its problems discussed above; for more extensive discussions of some of these approaches see Pagan (1987), Spanos (1988), Hendry et al (1989), Granger (1990). All these critics agree on one thing: *the textbook approach often gives rise to unreliable inferences and delivers poor predictive*

*performance, but they disagree on the sources of unreliability and on how to address these problems.*

**A. The Box-Jenkins** ARIMA(p,d,q) time series modeling approach is based on:

$$y_t^* = \alpha_0 + \sum_{k=1}^{p} \alpha_k y_{t-k}^* + \sum_{\ell=1}^{q} \beta_\ell \varepsilon_{t-\ell} + \varepsilon_t, \quad \varepsilon_t \backsim \mathsf{NIID}(0, \sigma^2),\ t \in \mathbb{T}, \tag{8}$$

where $y_t^* := \Delta^d y_t$, brought out the statistical inadequacy problem as it relates to the temporal structure and the heterogeneity exhibited by economic time series data. The use of 'differencing', in order to render the data stationary, was questioned in the late 1970s and that led to the 'unit root revolution' in econometric time series modeling; see Dickey and Fuller (1979), Phillips (1987).

**B.** The **Sims VAR(p) methodology** extended the ARIMA models to:

$$\mathbf{Z}_t = \mathbf{a}_0 + \sum_{k=1}^{p} \mathbf{A}_k \mathbf{Z}_{t-k} + \mathbf{E}_t, \quad \mathbf{E}_t \backsim \mathsf{NIID}(\mathbf{0}, \mathbf{\Omega}),\ t \in \mathbb{T}, \tag{9}$$

in order to model both the contemporaneous as well as the temporal structure of economic time series. Sims (1980, 1982) diagnosed the unreliability of the textbook approach to structural models as emanating from two sources: the neglect of the temporal structure of these series and the 'implausibility' of the substantive information forced upon the data.

**C.** The **Sargan-Hendry LSE tradition** (see Sargan, 1964, Hendry, 2003) was also motivated by the neglect of the temporal structure of the data, but tried to preserve a link between the statistical and substantive information. Hence, the LSE tradition proposed the **Autoregressive Distributed Lag** (AD(p,q)) model:

$$y_t = \alpha_0 + \sum_{k=1}^{p} \alpha_k y_{t-k} + \sum_{\ell=1}^{q} \boldsymbol{\beta}_\ell^\top \mathbf{x}_{t-\ell} + \varepsilon_k, \quad \mathbf{x}_t : k \times 1,\ \varepsilon_k \backsim \mathsf{NIID}(0, \sigma^2),\ t \in \mathbb{T}, \tag{10}$$

as a data-based formulation capturing the short run dynamics, with equilibrium economic theory being relevant in determining the long-run solution of this stochastic difference equation. To deal with the 'unreliability of evidence' problem, the LSE tradition suggests allowing the data to determine the choice of optimal $(p, q)$ by following a 'General to Specific' procedure from large initial values $(p, q)$ and testing downwards for a more parsimonious representation (see Hendry, 1995, Mizon, 1995). To avoid statistically unreliable inferences, the LSE approach recommends *diagnostic tests* in the spirit of the Box-Jenkins approach.

A particular case of the AD(p,q) model, known as the *error correction model*, which, in the case $p = q = 1$, takes the form:

$$\Delta y_t = \gamma_0 + \gamma_1 [y_{t-1} - \boldsymbol{\gamma}_2^\top \mathbf{x}_{t-1}] + \boldsymbol{\gamma}_3^\top \Delta \mathbf{x}_{t-1} + v_t,\ v_k \backsim \mathsf{NIID}(0, \sigma_v^2),\ t \in \mathbb{T}, \tag{11}$$

proved to be empirically a very successful specification; see Hendry (1993). It turned out that the empirical success of the error-correction model can be explained in terms of the cointegrating properties of certain non-stationary time series; see Engle and Granger (1987). This provided a connection between the AD and the Sims VAR approach, extending the error-correction model (11) to systems of equations; see Johansen (1991), Boswijk (1995).

**D.** The **Lucas-Sargent tradition** (Lucas, 1976, Lucas and Sargent, 1981) attempts to address the problem of unreliable empirical evidence associated with the textbook approach in two interrelated ways. First, by improving economic theory to account for the dynamic structure of economic time series, and, second, by basing forecasts and policy evaluations not on empirical regularities but on estimated structural models which are invariant to policy interventions.

**E. Leamer's Bayesian-oriented criticism** of the practised variant of textbook econometrics was motivated by the great disparity between the formal textbook approach and its practised variant, which he called 'cookbook' econometrics; see Leamer (1978). His suggestion is to make honest modelers out of cookbook econometricians by formalizing their ad hoc procedures using informal Bayesian procedures such as extreme bounds analysis (see Leamer and Leonard, 1983). This aims to expose the possible *fragility* of estimated relationships by testing their robustness to changes in prior information.

## 7.3 Whither textbook econometrics?

Although the various critics of the textbook approach pinpoint to crucial weaknesses of the approach, and make constructive suggestions on how to address them, none of them constitute a comprehensive methodology of econometric modeling that can replace what they are berating; see Pagan (1987). The question that naturally arises is the extent to which one can systematize some of the suggestions by the critics to put forward such a more comprehensive methodology that can potentially replace the textbook methodology.

Taking stock of what the critics **A-C** bequeathed to empirical modeling, one can argue that the Box-Jenkins (B-J) approach offered four primary innovations:

**(i) Predesignated family of models**. The modeling of time series data within a *predesignated family of models* (ARIMA(p,d,q)) that was thought to adequately capture their temporal dependence and heterogeneity (including seasonality).

**(ii) Modeling as an interative process**. Empirical modeling is not a one stage activity, but an *iterative process* that involves several stages, *identification*, *estimation* and *diagnostic checking*, before any *prediction* is made.

**(iii) Model assessment by diagnostic checks**. *Diagnostic checks,* based on the residuals from the fitted model, provide a way to detect model inadequacies with a view to improve the original model.

**(iv) Warranted Exploratory Data Analysis (EDA)**. *Exploratory data analysis* is not intrinsically sinful or deceptive, but can be legitimately used to select (identify) a model within the predesignated family.

**(v) Model overfitting**. For assessing the adequacy of a selected (identified) model, it's advisable to embed it into a more *general specification* in order to put the model 'in jeopardy' (see Box and Jenkins, 1970, p. 286).

The B-J approach constituted a major departure from the rigid textbook approach, where the model is assumed to be devised in a single step and in advance

of any data. The almost algorithmic nature of the B-J iterative modeling process added to its appeal, but it was its predictive success in the early 1970s (see Cooper, 1972, inter alia) that convinced econometricians that they could ignore the temporal dependence and heterogeneity of times series data at the detriment of the predictive performance of their models (see Granger and Newbold, 1986).

*Sims* (1980), in an attempt to broaden the intended scope of the B-J approach to include *prediction* and *description,* extended the predesignated family of models to the VAR(p) (9) in order to capture the *joint* temporal dependence and heterogeneity of related (via some form of theory) times series data.

The LSE tradition embraced (i)-(v) and broadened the predesignated family of models even further to allow for *prediction*, *description* and *explanation.* The main motivation for introducing the AD(p,q) family (10), which also incorporates the *contemporaneous dependence* (captured by $\boldsymbol{\beta}_0^\top \mathbf{x}_t$) among related times series, was to leave the door open for linking such data-oriented models to economic theory via their long-run solution. In addition, the LSE approach formalized (v) to:

**(vi) General to Specific (G-to-S) procedure**. To secure adequacy, one should commence from a general dynamic model, such as (10), and test downwards to a more specific but *congruent* empirical model (Hendry, 1995, p. 365).

The modeling conceptions (i)-(vi) were never integrated into the textbook approach because they cannot be accommodated into its methodological framework without a major overhaul. As a result, the critics **A-C** have generated their own modeling cultures within applied macroeconomics focusing on time series data. The countercharge by Textbook Econometricians (TE) is that the critics **A-C** simply ignore, or pay lip service to, economic theory; a return to the 'measurement without theory'. Models (8)-(10) are clearly not theory motivated, and thus *ad hoc* and arbitrary when viewed in the context of the textbook approach. By the same token, the Lucas-Sargent approach did have some influence on the textbook approach because it called for no changes in its basic methodological framework. The primary lessons from that approach are that:

(vii) The ultimate objective of empirical modeling should be to 'unveil' structural models with high degree of invariance to policy changes.

(viii) Static equilibrium economic models need to be extended to dynamic models to be able to account for the temporal structure exhibited by economic data.

The thesis defended in this paper is that critics **A-C** brought out certain crucial weaknesses of the textbook approach that invariably lead to unreliable inferences, but their proposals did not go far enough to addresses the unreliability of inference problem; (a) they raised but did *not* address adequately the statistical unreliability problem, and (b) they did not provide an adequate link between these predesignated family of data-oriented models and economic theory, neglecting the substantive unreliability problem. A TE would ask 'where do models (8)-(10) come from?' and 'how does one justify them if not on theory grounds?' The only reply is that these models are justified on *pragmatic* grounds, because they 'capture' the systematic informa-

tion in some time series data much better than the various structural models in the literature. The TE would reply that even on pragmatic grounds, not all time series data can be adequately modeled using this Linear/Normal/Homoskedastic families of models, and they are clearly inappropriate for cross-section and panel data. What can one do when models (8)-(10) are found to be wanting? Moreover, even when these models turn out to be appropriate, they are overly data-specific and could not possibly capture the invariant (structural) features of the phenomena being modeled because they largely ignore the available substantive subject matter information.

To be able to address these questions and the associated methodological concerns, one needs to envision a methodological framework where the issue of *statistical and/or substantive* unreliability can be adequately addressed without jeopardizing either form of information.

# 8 The prospect of 21st century econometrics

Summarizing the argument so far, the persistent unreliability of empirical evidence produced by the textbook approach can be traced to three primary sources.

The *first* source of unreliability stems from the fact that data are viewed 'in light of a theory', but not 'in light of a stochastic process that would render the data a truly typical realization thereof'. The perceived outcomes of the Keynes vs. Tinbergen and the Koopmans vs. Vining debates re-enforced this *theory-dominated* view of inductive inference.

The *second* source of unreliability stems from the way the gap between theory and data is often being glossed over by implicitly assuming that: (i) the available data provide congruous measurements to the concepts envisaged by economic theory, (ii) the circumstances envisaged by theory coincide with the actual DGM, and thus, (iii) the substantive information 'encompasses' the statistical, apart from white noise error(s), whatever the data!

The *third* source of unreliability arises from the fact that inductive inferences are often not probed adequately in the different ways they can be in error. One can make a case that little progress has been made in addressing the problems of (I) Statistical misspecification, (II) Inaccurate data, and (III) Incongruous measurement, (IV) External invalidity, since Moore (1914).

These concerns were instrumental in motivating a methodological framework, called the Probabilistic Reduction (PR) approach, proposed in Spanos (1986).

## 8.1 The Probabilistic Reduction (P-R) approach

The Probabilistic Reduction (PR) approach is discussed in this sub-section as a methodological framework whose primary objective is to foster a self-reflective meeting ground of the disparate approaches; a framework wherein the strengths and weaknesses of the various approaches can be discussed and compared. It offers a methodological forum in the context of which legitimate methodological concerns, such as

the mistrust of data-oriented models and preliminary data analysis, the imperativeness of predesignation, the abuse of substantive information, as well as the neglect of statistical information, can be illuminated and hopefully addressed.

### 8.1.1 Substantive vs. Statistical Information

The starting point of the PR approach is that empirical models constitute a blending of *substantive* and *statistical* information, and their primary objective is to enable us to *learn* about observable phenomena of interest using data. The substantive information derives from the subject matter theory, and the statistical information is reflected in the 'chance regularity (recurring) patterns' exhibited by the data. In an attempt to address the problems raised by the textbook approach, and explain when the choice of models such as (8)-(10) makes sense, the two kinds of information are encapsulated initially by two different models, the **theory** and **statistical models**; see fig. 1. The former is stipulated in terms of theoretical variables, some of which might be unobservable, but the latter is specified exclusively in terms of the observable random variables underlying the data $\mathbf{Z} := (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T)$. The idea is that the Lucas-Sargent call for structural modeling can be accommodated as theory models, but, at the same time, the data-oriented models, suggested by critics **A-C,** can be justified as statistical models. The problem is to find ways to link the two without compromising the integrity of either the substantive or the statistical information.

### 8.1.2 Statistical model specification

Let us focus on statistical models first. A statistical model presupposes the choice of the relevant data $\mathbf{Z}$, (chosen by a theory or theories) and is specified by utilizing the Fisher-Neyman probabilistic perspective which views $\mathbf{Z}$ as a 'realization' of a (vector) stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$ whose probabilistic structure is such that would render $\mathbf{Z}$ 'truly typical'. This probabilistic structure, according to *Kolmogorov's theorem*, can be fully described, under certain mild regularity conditions, in terms of the joint distribution $D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_T; \boldsymbol{\phi})$; see Doob (1953). A statistical model constitutes a *parameterization* of the assumed probabilistic structure of $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$ and can be viewed as a reduction from this joint distribution.

To illustrate how the same $D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_T; \boldsymbol{\phi})$ can give rise to different (but related) statistical models, consider the case where $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$ is assumed to be a Markov of order p (M(p)), and Stationary (S) process. On the basis of these assumptions one can deduce:

$$D(\mathbf{Z}_1, ..., \mathbf{Z}_T; \boldsymbol{\phi}) \overset{\mathsf{M}(p)}{=} D(\mathbf{Z}_1; \boldsymbol{\varphi}_1) \prod_{t=2}^{T} D_t(\mathbf{Z}_t|\ \mathbf{Z}_{t-1}^p; \boldsymbol{\varphi}_t) \overset{\mathsf{M}(p)\&\mathsf{S}}{=} D(\mathbf{Z}_1; \boldsymbol{\varphi}_1) \prod_{t=2}^{T} D(\mathbf{Z}_t|\ \mathbf{Z}_{t-1}^p; \boldsymbol{\varphi}), \tag{12}$$

where $\mathbf{Z}_{t-1}^p : (\mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, ..., \mathbf{Z}_{t-p})$, i.e., the joint distribution $D(\mathbf{Z}_1, ..., \mathbf{Z}_T; \boldsymbol{\phi})$ is reduced to a product of conditional distributions $D(\mathbf{Z}_t|\ \mathbf{Z}_{t-1}^p; \boldsymbol{\varphi})$. Assuming, in addition that $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$ is Normal (N), $D(\mathbf{Z}_t|\ \mathbf{Z}_{t-1}^p; \boldsymbol{\varphi})$ gives rise to the VAR(p) model (9). That is, the VAR(p) model provides a particular representation of a (N, M(p), S)

process $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$. However, in view of the fact that:

$$D(\mathbf{Z}_t|\ \mathbf{Z}_{t-1}^p;\boldsymbol{\varphi})=D(y_t|\ \mathbf{X}_t,\mathbf{Z}_{t-1}^p;\boldsymbol{\varphi})\cdot D(\mathbf{X}_t|\ \mathbf{Z}_{t-1}^p;\boldsymbol{\varphi}),$$

it can be easily shown that the AD(p,p) (10), based on $D(y_t|\ \mathbf{X}_t,\mathbf{Z}_{t-1}^p;\boldsymbol{\varphi})$, provides another parameterization of the same stochastic process $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$. How does one decide which statistical model to choose? The choice depends entirely on how a particular parameterization would enable one to embed the structural model in its context. This, of course, presupposes that there is a structural model providing a bridge between the theory and the statistical model. How the two sources of information are sythesized will be discussed after a more detailed discussion of how the PR approach addresses some of the statistical issues left open by critics **A-C**.

This PR reduction provides a justification for the VAR(p) and AD(p,p) as *statistical models*, but also brings out their most crucial weakness; they do not constitute a panacea for modeling time series data. When any of the probabilistic assumptions (N, M(p), S) are inappropriate, these models are likely to be misspecified. Moreover, the critics **A-C** did not adequately address the problems of (a) **specification** (how one initially chooses a statistical model), (b) **MisSpecification (M-S) testing** (how one proceeds to probe thoroughly the different ways a statistical model can be false), and (c) **respecification** (how one proceeds when the initial statistical model if found wanting). Indeed, one can argue that there is a lot in common between the critics **A-C** and the modeling strategy of model selection based on *Akaike type information criteria*. All these procedures narrow the *problem of statistical model specification* down to *model selection* within predesignated family of models. As perceptively stated by Lehmann (1990): "... this view of model selection ignores a preliminary step: the specification of the class of models from which the selection is to be made." (p. 162). If the predesignated family of models turns out to be statistically misspecified, the selection process is likely to lead to misleading choices; the *actual* error probabilities are likely to be different from the assumed ones. Moreover, once such a family is selected on statistical adequacy grounds, no further model selection is needed, because in integral part of establishing statistical adequacy is the choice of maximum lags/trends needed to capture the systematic (recurring) information in the data.

**Table 3: The VAR(1) model**

$$\boxed{\mathbf{Z}_t=\mathbf{a}_0+\mathbf{A}_1^\top\mathbf{Z}_{t-1}+\mathbf{E}_t,\ \ t\in\mathbb{T}.}$$

| | | |
|---|---|---|
| [1] | Normality: | $D(\mathbf{Z}_t\mid\mathbf{Z}_{t-1}^0;\boldsymbol{\psi})$ is Normal, |
| [2] | Linearity: | $E(\mathbf{Z}_t\mid\sigma(\mathbf{Z}_{t-1}^0))=\mathbf{a}_0+\mathbf{A}_1^\top\mathbf{Z}_{t-1}$, linear in $\mathbf{Z}_{t-1}$, |
| [3] | Homoskedasticity: | $Cov(\mathbf{Z}_t\mid\sigma(\mathbf{Z}_{t-1}^0))=\boldsymbol{\Omega}$, free of $\mathbf{Z}_{t-1}^0$, |
| [4] | Markov dependence: | $\{\mathbf{Z}_t,\ t\in\mathbb{T}\}$ is a vector Markov process, |
| [5] | t-homogeneity: | $(\mathbf{a}_0,\mathbf{A}_1,\boldsymbol{\Omega})$ are not functions of $t\in\mathbb{T}$. |

where $\mathbf{A}_1{:=}Cov(\mathbf{Z}_{t-1})^{-1}Cov(\mathbf{Z}_{t-1},\mathbf{Z}_t)$, $\mathbf{a}_0{:=}E(\mathbf{Z}_t)-\mathbf{A}_1^\top E(\mathbf{Z}_{t-1})$,
$\boldsymbol{\Omega}:=Cov(\mathbf{Z}_t)-Cov(\mathbf{Z}_t,\mathbf{Z}_{t-1})Cov(\mathbf{Z}_{t-1})^{-1}Cov(\mathbf{Z}_{t-1},\mathbf{Z}_t)$.

As argued in Spanos (1986, 2001a), the specification of statistical models in terms of *error assumptions,* such as those in (8)-(10), is inadequate because: (i) such specifications are often incomplete (they invariably involve hidden assumptions), and (ii) some of the error assumptions seem non-testable, or innocuous until they are translated in terms of the observable r.v's. In the case of the VAR(1) model (to simplify notation) the complete set of probabilistic assumptions in terms of $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$ are given in table 3. If we compare assumptions [1]-[5] with NIID for the error we can see that assumption [5] is not explicitly stated in the error specification. Indeed in the textbook specification the t-invariance of $\mathbf{\Omega}$ is often conflated with homoskedasticity [3]; the distinction becomes clear when the VAR(1) model is seen in the context of the reduction (12).

It goes without saying that an *incomplete specification* will be a major stumbling block for ensuring *statistical adequacy.* A key element of the PR approach is a *complete* specification of the probabilistic assumptions in terms of the observable processes involved. The complete set of statistical model assumptions reveals itself when the distribution underlying it, say $D(\mathbf{Z}_t|\ \mathbf{Z}_{t-1}^p;\boldsymbol{\varphi})$, is related to the joint distribution $D(\mathbf{Z}_1,...,\mathbf{Z}_T;\boldsymbol{\phi})$, via the reduction (12) itself. At the same time, this reduction demarcates the model in question relative to all other possible models ($\mathcal{P}$) that could, potentially, have given rise to data $\mathbf{Z} := (\mathbf{z}_1,\mathbf{z}_2,...,\mathbf{z}_T)$, because all the potential models in $\mathcal{P}$ can be perceived as reductions from the same joint distribution based on different *reduction assumptions* from three broad categories: (D) **Distribution,** (M) **Dependence,** (H) **Heterogeneity.** For instance, if one were to replace the Normality assumption with that of a *Student's t distribution*, a whole new family of models corresponding to (9)-(11) would arise from the reduction in (12); see Spanos (1986, 1992). Viewing the model in question, say $M_0$, in the context of all potential models $\mathcal{P}$, provides a broad and flexible enough framework to address the issues of M-S testing and respecification left open by the critics **A-C**.

### 8.1.3 Statistical vs. Structural error terms

An important dimension of the PR approach is that one is able to probe for different types of errors at different levels in the spirit of the error statistical account briefly discussed in section 2.1. Although it is important to distinguish clearly between the different *types of errors* (I)-(IV) and the *error terms* associated with different models, the two notions are related in so far as it is important to delineate the types of errors to probe for in assessing the non-systematic nature of error terms.

In the context of a statistical model, such as the VAR(1):

$$\mathbf{Z}_t = \mathbf{a}_0 + \mathbf{A}_1\mathbf{Z}_{t-1} + \mathbf{E}_t,$$

the error term $\mathbf{E}_t$ represents a particular case of the *martingale difference* process $\mathbf{E}_t^* = \mathbf{Z}_t - E(\mathbf{Z}_t \mid \sigma(\mathbf{Z}_{t-1}^0))$, when $E(\mathbf{Z}_t \mid \sigma(\mathbf{Z}_{t-1}^0)) = \mathbf{a}_0 + \mathbf{A}_1\mathbf{Z}_{t-1}$. What is crucially important in probing for misspecification errors in the context of a statistical model is to realize that the *universe of discourse* for statistical purposes is confined to $\mathcal{F} := \sigma(\mathbf{Z}_1,\mathbf{Z}_2,\dots\mathbf{Z}_T)$, denoting the sigma-field generated by $(\mathbf{Z}_1,\mathbf{Z}_2,\dots\mathbf{Z}_T)$. That

is, one needs to probe the different ways assumptions [1]-[5] could be false *relative* to the information set $\mathcal{F}$. In particular, missing factors can only be relevant for *statistical* (as opposed to substantive) misspecification purposes if they belong to $\mathcal{F}$. This renders the assumptions of the statistical error term *empirically testable* using the data $(\mathbf{z}_1, \mathbf{z}_2, \dots \mathbf{z}_T)$; see Spanos (2005b) for further details. In assessing the non-systematic nature of this error term one needs to focus primarily on probing for errors associated with (I) statistical misspecification, and (II) inaccurate data.

A **structural model** of the form:

$$\mathbf{y} = h(\mathbf{X}; \boldsymbol{\phi}) + \boldsymbol{\varepsilon}(\mathbf{X}, \mathbf{U}), \tag{13}$$

where $\mathbf{y} := (y_1, \dots, y_m)$, $\mathbf{X} := (X_1, \dots, X_k)$, with $\boldsymbol{\phi}$ denoting the unknown *structural parameters*, $\boldsymbol{\varepsilon}(\mathbf{X}, \mathbf{U}) = \mathbf{y} - h(\mathbf{X})$ represents the *structural error term*, which is regarded as a function of both $\mathbf{X}$ and $\mathbf{U}$: a set of potentially relevant (observable or unobservable) factors. The structural error term is 'autonomous' and represents all unmodeled influences**;** see Spanos (2005b) for further details. Hence, in assessing the non-systematic nature of this error term one needs to focus primarily on probing for errors associated with (III) incongruent measurement, and (IV) external invalidity; in particular, probing for missing factors to ensure substantive reliability.

### 8.1.4 MisSpecification (M-S) Testing/Respecification

In the context of the PR approach, the question posed by **M-S testing** is in the form of the hypotheses:

$$H_0 : f_0(\mathbf{z}) \in M_0, \text{ against } H_1 : f_0(\mathbf{z}) \in [\mathcal{P} - M_0], \tag{14}$$

where $f_0(\mathbf{z})$ denotes the 'true' distribution of $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$. The real problem is how can one probe $[\mathcal{P} - M_0]$ adequately, given that it invariably entails an infinity of models. The answer is provided by the reduction itself because the specification of the particular model, say $M_0$, amounts to imposing probabilistic assumptions which partition $\mathcal{P}$. Re-partitioning provides an effective strategy to probe $[\mathcal{P} - M_0]$ the different ways $M_0$ could be in error, when it is based on information concerning potential departures from $M_0$. The use of graphical techniques, justified as a form of qualitative severe testing reasoning, can be utilized to render re-partitioning deliberatively effective. It should be noted that M-S tests differ from Neyman-Pearson (N-P) tests in so far as the probing in the case of the latter takes place *within* the boundary of a prespecified model $M_0$, but the former constitutes probing $[\mathcal{P} - M_0]$, *without* $M_0$, rendering M-S tests a type of Fisherian *significance tests*; see Spanos (1999). An effective M-S testing strategy should take account of the assumptions underlying the tests themselves, as well as avoid the problem of infinite regress. With that in mind, the use of a combination of *parametric* and *non-parametric* M-S tests is often the best strategy to probe $[\mathcal{P} - M_0]$ more exhaustively; see Mayo and Spanos (2004).

The above described procedure constitutes a more effective M-S testing strategy than the limited scope of error diagnostic checks based on the residuals from the

fitted model proposed by the Box-Jenkins approach. Moreover, this view of M-S testing is different from that of the LSE tradition, as explained by Hendry and Richard (1982), in so far as the latter views diagnostic tests as 'design criteria' and not as genuine 'independent checks' on the validity of the model. It is also important to note that *statistical adequacy* is not the same as *model congruency* propounded by the LSE tradition. Model congruency specifies a set of desirable properties of empirical models in general, which constitute a mixture of statistical and substantive criteria; see Hendry (1995), p. 365. In contrast, *statistical adequacy* is confined to statistical information and is always defined *relative* to the probabilistic assumptions comprising the particular statistical model in question. Hence, unlike a congruent model, a statistically adequate model does not have to be homoskedastic, etc. (see Spanos, 1994, 1995). More importantly, it does *not* have to be consistent with a particular theory; it only needs to provide an embedding framework for the theory or theories under consideration.

The same re-partitioning of $\mathcal{P}$, using reduction assumptions from the three broad categories, can be used to *respecify* the original statistical model in view of the detected departures from $M_0$. The details of statistical model **respecification,** which involve the relationship between *reduction* and *model assumptions*, the creative use of M-S test results as a whole, as well as the utilization of graphical techniques, are discussed in Spanos (1999, 2000). The important point to make is that each respecified model is assessed for statistical adequacy by testing *its* own assumptions and the process ends when such a statistically adequate model is found. This provides a respecification strategy with the emphasis placed on constructing a new statistical model within $\mathcal{P}$ that adequately captures the statistical systematic information in the data.

### 8.1.5 Bridging the gap between theory and data

The PR approach highlights the gap between theory and data and proposes generic ways to bridge it using a *sequence of interlinked models*, emanating from the phenomenon of interest and furcating into models based on the *statistical* information in the data (statistical model), and models based on the *substantive* information (theory and estimable models). The two families of models are blended at the level of the **empirical model**; see fig.1. Often economic theory comes in the form of static equilibrium relationships like a demand/supply model, but the available data refer to quantities transacted and the corresponding prices. In such cases, there is a need to construct **estimable models,** specified in terms of the observables, which could serve as a link between the theory and statistical models; Spanos (1989). The term *structural model* stands for a theory or an estimable model when the latter is called for. The insistence of the textbook and the Lucas-Sargent approaches on estimating *structural models* is distinctly desirable, but often its proper implementation requires one to probe theory models for potential errors that can arise from incongruous measurements, as well as external invalidity. Dynamic decision functions might not go far

enough to bridge the gap between intentions and the available data, and the devising of estimable models, in the form of quantity and price adjustment equations, is called for; see Spanos (1995a)

A statistically adequate model provides a convenient summary of the *statistical information* in the data. It might turn out that one can find no such statistical model for a particular data $\mathbf{Z}$, in which case the error-probing should give rise to a re-consideration of either the structural model or the choice of the data, or both. The idea is that an adequate statistical model might not have any *explanatory power*. The structural model, when its restrictions on the statistically adequate model are data-acceptable, contains the substantive information that will supplement the statistical information to bestow upon the statistical model additional explanatory power, giving rise to an *empirical model*. Hence, the empirical model is likely to be less data-specific and thus more informative than the statistical model. Hopefully, the empirical model will 'reflect' the invariance structure of the phenomenon being modeled, and it will be of use for prediction, explanation as well as policy analysis purposes. We refer to the data-acceptability of a the structural model within the context of an embedding adequate statistical model **identification**. We note that this use of the term is different from that of the Box-Jenkins approach, which refers to the choice of the optimal values of (p,d,q) in an ARIMA(p,d,q) model, or that of the textbook approach which adopted the Cowles Commission use of the term as explained in section 6.2.

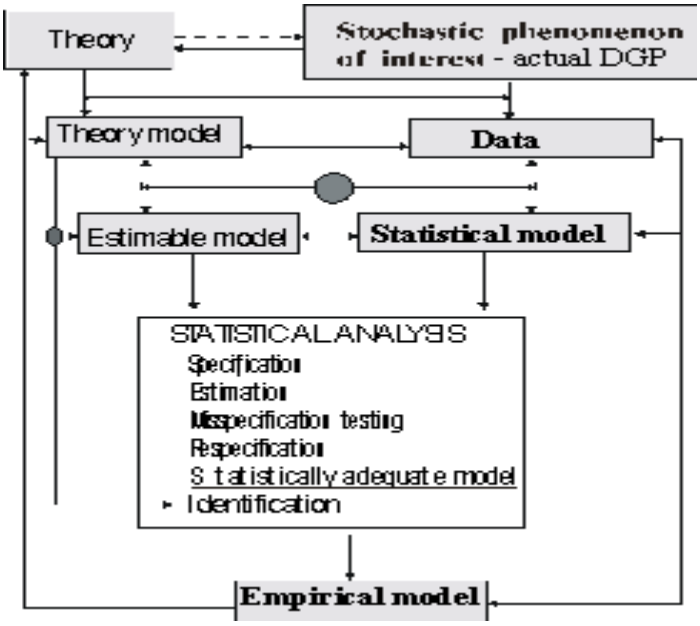


Fig. 1: The Probabilistic Reduction approach

It goes without saying that the PR approach *does not* offer a systematic method and a logic on how to 'discover' appropriate structural models; it offers a framework

to facilitate this process by bringing out the distinctions that help to ask the right questions, even though the answers might be very difficult at times; they often depend crucially on detailed substantive subject matter information as well as on in-depth knowledge of the substantive aspects of the data in question. At the same time, however, the PR approach suggests that the process of structural model choice does not rely exclusively on imagination and inspiration; it interacts in important ways with the statistical information and that interaction can be systematized in ways which can often render the imagination more creative and effective.

### 8.1.6 Substantive Adequacy

It is important to emphasize that a statistically adequate is necessary for the assessment of *substantive adequacy*. Once this is established one can proceed to pose questions concerning substantive information, including external validity. A missing confounding factor that does not necessarily lead to invalid statistical inferences, but it might lead to invalid substantive inferences. The only way to assess that, however, is in the context of a statistically adequate model which 'includes' the confounding factor. Sometimes, statistical misspecification which cannot be rectified by respecifying the statistical model within the same $\mathcal{P}$ might be an indication of a missing factor whose systematic effect is left in the residuals of the estimated statistical model. Again, this can only be established in the context of a statistically adequate model that includes that factor. Moreover, strange results in M-S testing often provide good indications for *inaccurate data*; Moore's (1914) data for pig iron provide an example of distorted dynamics due to imbued systematic errors.

### 8.1.7 The PR approach and recent developments in statistics

In addition to adopting the Fisher-Neyman probabilistic perspective, the Probabilistic Reduction (PR) approach integrates a number of different 20th century developments in statistics to give rise to a model-based statistical methodology with very specific rules for model specification as well model validation, thus addressing the issues raised by Rao (2004). *Firstly*, the PR approach is built squarely on the frequentist foundations laid down by Fisher and extended by Neyman and Pearson. It avoids some of the problems with frequentist testing by using post-data evaluation based on severity; see Mayo (1996), Mayo and Spanos (2003). *Secondly*, it integrates the various developments associated with *nonparametric techniques* (Lehmann, 1975, Simonoff, 1996) and *Exploratory Data Analysis* (see Tukey, 1977, Mosteller and Tukey, 1977), into the PR framework as tools to enhance the effectiveness of the process from statistical model specification to a statistically adequate model. Specification, misspecification testing and respecification rely heavily on the use of graphical techniques, as well as nonparametric methods grounded on both rank statistics and kernel smoothing techniques; see Spanos (1999, 2001b), Mayo and Spanos (2004). However, the statistically adequate model in the context of which the structural model will be embedded is invariably a *parametric statistical model*, in order to ensure both the embedding as well the precision of inference. *Thirdly*, the PR approach brings out the important role

that bootstrapping (Efron and Tibshirani, 1993) and other resampling techniques (see Politis at al, 1999) can play in evaluating the relevant error probabilities associated with particular inductive inferences.

The severe testing reasoning ( section 2.1), is used to address a number of issues such as (i) model selection vs. model specification, (ii) misspecification testing vs. Neyman-Pearson testing, (iii) respecification vs. error-fixing, (iv) statistical vs. substantive significance, (v) pre-designation vs. post-designation, (vi) what constitutes *data mining*, *omitted variable bias*, *pre-test bias*, *data snooping*; see Spanos (2000).

## 8.2 Addressing the unreliability of evidence conundrum

As argued above, the textbook approach to econometrics focuses on 'saving the theory' by fashioning confirming empirical evidence, with little concern for assessing the different ways such inferences can be false. The source of the problem is the mistrust for 'statistical regularities', accompanied by an undue confidence in theory's ability to weed out such untrustworthy 'regularities'; an attitude that can be traced back to the 1920s. The mistrust for 'statistical regularities' is based on the belief that they are too easy to contrive, and if one is willing to try hard enough, the data can be molded to 'fit' any theory or hypothesis. There is also the lingering mistrust that if the choice of data $\mathbf{Z}$ is not theory-motivated, it will lead to a plethora of meaningless 'data-based regularities'. This was certainly true at the time of Moore, but Fisher's approach to statistical inference changed that drastically to make the fallacious nature of such an argument obvious.

The **Fisher-Neyman probabilistic perspective** asserts that:

(a) Every statistical (inductive) inference is based on certain *premises,* in the form of a *statistical model* parameterizing an underlying stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$ that would have produced data $\mathbf{Z} := (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T)$ as 'a matter of course'. (b) Statistical adequacy is necessary for establishing the reliability of 'statistical regularities'. Unfortunately, the Fisher-Neyman probabilistic perspective has not as yet permeated empirical modeling in economics because deep-rooted erroneous impressions, such as there is no such thing as statistical information separate from substantive information, are often hard to eliminate.

An estimated regression model with a high $R^2$ does not constitute a 'statistical regularity', unless the probabilistic assumptions [1]-[5] of the statistical model (table 2) have been vigorously probed, using potent misspecification tests and none detected, i.e. establish *statistical adequacy.* In this author's experience, statistically adequate empirical regularities are rare because assumptions such as [4]-[5] are not easily met in practice. Hence, even though statistical 'non-regularities' based on goodness of fit criteria abound, severely-probed statistically adequate regularities are very rare indeed. Once statistical adequacy is established, the reliability of the ensuing inductive inference assures the trustworthiness of the assessment of the substantive information (in the form of a structural model). Statistically *inadequate* premises give rise to *non-regularities*, and 'theoretical meaningfulness' cannot transform it into a regularity of

any sort.

The difficulty in establishing statistical adequacy suggests that mindless assemblages of data are unlikely to give rise to any such regularities; but even if they sometimes do, probing for *incongruous measurement* and *external validity* assessments, will weed them out. Hence, the attitude that the process of establishing statistical regularities cannot even begin unless one has a complete and detailed structural form of the type envisaged by the Cowles Commission group is misplaced in fields like economics. The process might begin with low level theory or theories (conjectures), and use statistical regularities to guide the elaboration/testing of such theories. Indeed, the fleshing out of ceteris paribus clauses can be reliably done using statistically adequate models which ensure the trustworthiness of the inferences based upon them. That is, statistically adequate models provide the cornerstone for the process to probe for both incongruous measurement and the external invalidity of theories. Often, the inability to establish statistical adequacy within the chosen information set is a strong signal that important factors effecting the phenomenon of interest might be missing. This suggests that statistically adequate regularities can be of great help in disciplines like economics, where substantive information is not as precise or reliable as in some other fields like physics. Instead of subordinating observation to theory, it might pay to allow observation to play some role in the 'discovery' process as well. Once a statistically adequate model is reached, one can assess the extent to which the theory in question accounts for these regularities, as well as contemplate alternative ways inductive inferences based on these regularities might be substantively false due to incongruous measurement and/or external invalidity.

### 8.2.1 Revisiting Moore's 'demand' for corn

Let us return to Moore's 'demand for corn' and consider the issue of respecifying his 'interpolated curve' in order to ensure statistical adequacy. It turns out that using his transformed data $(x_t, y_t,\ t = 1, ..., n)$, no statistically adequate model is possible, suggesting that Moore's 'data adjustment' is likely to have introduced systematic errors. However, if one returns to the original data, $p_t-$ average price per bushel, $q_t-$ production in bushels, the following dynamic linear regression model:

$$\ln q_t = \underset{(.794)}{.401} - \underset{(.078)}{0.679} \ln p_t + \underset{(.082)}{0.734} \ln p_{t-1} + \underset{(.082)}{0.960} \ln q_{t-1} + \underset{(.109)}{\widehat{u}_t}\ , \qquad (15)$$
$$R^2 = .909,\ s = .109,\ n = 44,$$

turns out to be statistically adequate, using similar misspecification tests as in the appendix. (15) can now provide the basis for a dialogue between theory and data to probe for other errors, including *theoretical validity*, *inaccurate data*, *incongruous measurements* and *external invalidity*. For instance, it's clear from Moore's own discussion of how these data were compiled that some systematic errors are likely to have crept in. It is also obvious from his discussion of the observed data that they do not measure 'demand' as understood in economic theory; *intentions to buy* corresponding to hypothetical prices. In view of the fact that the data measure actual production and average observed prices, the estimated equation could not be a

demand schedule, but it could be a *quantity adjustment equation*; see Spanos (1995a). Can economic theory shed further light by providing a rationalization for (15) as an adjustment equation? To what extent is such an adjustment equation appropriate for the US market? These are questions that need to be addressed by the modeler.

## 8.3 The Gauss-Markov perspective and reliability/precision

One of the important elements of the PR approach summarized above is that, the specification of statistical models, based on the probabilistic structure of the vector stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$, should be as specific and complete as possible. The rationale is that the narrower the premises, the more precise the inferences - assuming that the postulated structure is valid for the data in question - hence the important role ascribed to *statistical adequacy*. The *weaker* (less specific) the postulated probabilistic structure, the *less precise and incisive* the inferences. Moreover, weaker probabilistic assumptions do not necessarily render the inference more reliable, but they inevitably contribute to the imprecision of inference; see Spanos (2000).

This goes against the textbook conventional wisdom which emphasizes the weakest probabilistic structure that would 'justify' a method yielding 'consistent' estimators of the parameters of interest. In particular, he Gauss-Markov (G-M) theorem, as well as analogous theorems concerning the 'optimality' of IV, GMM and non-parametric methods, distance themselves from strong probabilistic assumptions, such as Normality, in an attempt to claim greater generality and less susceptibility to misspecification. Indeed, these methods are often motivated by claims of weak probabilistic assumptions as a way to overcome unreliability; Matyas (1999), p. 1.

The rationale underlying this argument is that the reliance on weaker probabilistic assumptions will render OLS, IV and GMM-based inferences less prone to statistical misspecifications and thus more reliable. The cornerstone of this rationale is the G-M theorem. Minimum variance within the class of unbiased estimators of $\boldsymbol{\beta}$, which are also linear in $\mathbf{y}$, of course, amounts to relative efficiency within a very narrow and no so interesting class of estimators. For Gauss in 1809 linearity might have been an attractive property, but now it has no value for inference purposes beyond its historical context. This is a mathematically interesting result, but it does not serve well the **reliability** and **precision** of inference. To see this let us consider how an econometrician using this theorem can proceed to draw inferences on the basis of $\widehat{\boldsymbol{\beta}}$.

**A. Finite sample inference.** The sampling distribution $\mathsf{D}(.)$ of $\widehat{\boldsymbol{\beta}}$ is:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \ \backsim \ \overset{?}{\mathsf{D}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$

When testing hypotheses, such as $H_0 : \boldsymbol{\beta} = \overline{\boldsymbol{\beta}}$, vs. $H_1 : \boldsymbol{\beta} \neq \overline{\boldsymbol{\beta}}$, the type I and II error probabilities cannot be evaluated directly because $\mathsf{D}(.)$ is *unknown*. Hence, the only way finite sample inference is possible is to use its first two moments, in conjunction with inequalities such as Chebyshev's or Liapunov's, to provide upper/lower bounds for such error probabilities. It is well known, however, that these upper/lower bounds can be very crude in practice (see Spanos (1999), ch. 10), leading to very imprecise inferences.

**B. Asymptotic inference.** One hopes that as $T \to \infty$ the crudeness of finite sample inferences will be ameliorated via the asymptotic sampling distribution: $\sqrt{T}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \underset{\alpha}{\backsim} \mathsf{N}(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{Q}_X^{-1})$,where $\mathbf{Q}_X = \lim_{T\to\infty}\left(\frac{\mathbf{X}^\top\mathbf{X}}{T}\right)$. The main problem with all asymptotic inferences is that its approximation error cannot be evaluated for a given $T$. One can, however, contemplate a number of scenarios where the approximation might be good, bad or uncertain.

**(a) Good**. When assumptions [1]-[5] (see table 2) are valid, the approximation:

$$\widehat{\boldsymbol{\beta}} \backsim \mathsf{N}(\boldsymbol{\beta},\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}), \tag{16}$$

will be excellent because it's true for any $T$.

**(b) Bad**. When any of the assumptions [2]-[5] (see table 2) are invalid, the approximation (16) is likely to be bad. Even small deviations from these assumptions can have a sizeable distorting affect on the error probabilities. It is important to note that the distorting affect of certain forms of misspecification, such as trending data, does not decrease with the sample size, it *increases*; see Spanos and McGuirk (2001).

**(c) Uncertain**. When only assumption [1] is invalid, the approximation (16) might or might not be bad. For instance, if $D(\mathbf{y}_t \mid \mathbf{X}_t;\boldsymbol{\psi})$ is highly non-Normal, say highly skewed, the approximation is not very good even for moderate sample sizes; see Spanos and McGuirk (2001). Moreover, we know that when $D(\mathbf{y}_t \mid \mathbf{X}_t;\boldsymbol{\psi})$ is skewed, the reliability of inference concerning $H_0$ is also adversely affected by the 'non-Normality' of the $\{\mathbf{X}_t,\ t\in\mathbb{T}\}$ process; see Ali and Sharma (1996).

The real problem in practice is that a G-M modeler has *no idea* which of the above scenarios applies to a particular situation *unless* thorough *misspecification testing* is applied to assess the validity of [1]-[5] (see table 2) with data $\mathbf{Z} := (\mathbf{y},\mathbf{X})$. Unfortunately, the G-M theorem discourages probing for departures from these assumptions (especially Normality), which could shed light on how good the asymptotic approximation might be for a given $T$. Hence, the modeler would be considerably more informed about the potential reliability of any inferences based on the CLR model if assumptions [1]-[5] (see table 2) were probatively assessed, even in the case where some of them are rejected by the data, rather than invoke the G-M theorem and be oblivious of the appropriateness of assumptions [1]-[5] for data $\mathbf{Z}$.

In addition to the reliability problem, the textbook modeler has to face the *imprecision* of inference issue because weaker assumptions give rise to less precise inference. Hence, the textbook strategy to address the problem of potential unreliability emanating from statistical misspecification leads inevitably to imprecision of inference, rendering empirical evidence non-decisive! The apparent *trade-off* between reliability and precision of inference, created by the G-M perspective, can be effectively addressed by thorough misspecification testing before any inference is made.

## 8.4 Statistical reliability and structural models

The statistical unreliability problem becomes much worse when the structural model is 'framing' the inference with the statistical model *not* even specified *explicitly*. This

can be best illustrated by the textbook discussion of Instrumental Variables (IVs) as providing a way to tackle the problems of *bias* and *inconsistency* for the OLS estimator $\widehat{\boldsymbol{\alpha}}=\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{y}$, arising from the dependence between the stochastic regressors $\mathbf{X}_t$ and the error term $\varepsilon_t$ :

$$y_t=\boldsymbol{\alpha}^{\top}\mathbf{X}_t+\varepsilon_t,\quad \text{where } \mathbf{X}_t:m\times 1,\ E(\mathbf{X}_t\varepsilon_t)\neq\mathbf{0},\ t=1,...,T. \tag{17}$$

How does one decide that condition (a) $E(\mathbf{X}_t\varepsilon_t)\neq\mathbf{0}$ holds? One is supposed to tell an economic theory 'story' on how some of the omitted variables (that one can think of) included in $\varepsilon_t$, are likely to be correlated with $\mathbf{X}_t$. Is the story sufficient to render (a) operational? Clearly not *statistically operational*, because $\varepsilon_t$ is never observable, and for every story one can tell, somebody else can tell another claiming the opposite.

How does the textbook narrative solve the bias and inconsistency problems? By finding another vector of observable variables $\mathbf{Z}_t:p\times 1,\ p\geq m$ such that (b) $E(\mathbf{Z}_t\varepsilon_t)=\mathbf{0}$. How is this an operational solution? One is supposed to tell another economic theory 'story' on how the omitted variables included in $\varepsilon_t$ (that one can think of) are unlikely to be correlated with $\mathbf{Z}_t$, but they correlated with $(y_t,\mathbf{X}_t)$, i.e. (c) $E(\mathbf{Z}_t y_t)=\boldsymbol{\sigma}_{31}\neq\mathbf{0}$, (d) $E(\mathbf{Z}_t\mathbf{X}_t^{\top})=\boldsymbol{\Sigma}_{32}\neq\mathbf{0}$, (e) $E(\mathbf{Z}_t\mathbf{Z}_t^{\top})=\boldsymbol{\Sigma}_{33}>\mathbf{0}$. What about the omitted factors in $\varepsilon_t$ that one *cannot* think of, some of which are not even measurable? Condition (b) is clearly non-operational in the same sense that (a) is not, and the 'solution' boils down to verifying (c)-(e) by calculating their sample analogues, say (c)' $\frac{1}{T}\left(\mathbf{Z}^{\top}\mathbf{y}\right)\neq\mathbf{0}$, (d)' $\frac{1}{T}\left(\mathbf{Z}^{\top}\mathbf{X}\right)\neq\mathbf{0}$, (e)' $\frac{1}{T}\left(\mathbf{Z}^{\top}\mathbf{Z}\right)>\mathbf{0}$. This, of course, is pitiably inadequate from the statistical viewpoint because there will be thousands of instruments whose sample second moments would seem to satisfy (c)'-(e)', and statistical reliability cannot be based on who can tell the 'best story'; it's like playing tennis with the net down! The reliability of inference always depends on the adequacy of the underlying statistical model. However, one will be hard pressed to find the implicit statistical model mentioned in the traditional textbook discussion.

Viewing this argument in light of the discussion in section 8.4, (17) is a structural model whose 'embedding' statistical model is the Multivariate Linear Regression (MLR) in table 2, with $\mathbf{y}_t:=(y_t,\mathbf{X}_t)$ and $\mathbf{x}_{1t}:=\mathbf{z}_t$. As shown in Spanos (1986), (17) constitutes a reparameterization/restriction of the MLR model, and the reliability of any inference based on the IV estimator $\widetilde{\boldsymbol{\alpha}}=\left(\mathbf{X}^{\top}\mathbf{P}_Z\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{P}_Z\mathbf{y}$ where $\mathbf{P}_Z=\mathbf{Z}\left(\mathbf{Z}^{\top}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}$, depends crucially on (i) testing thoroughly assumptions [1]-[5] (table 2), detecting no departures, and then (ii) testing and accepting the overidentifying restrictions. This author is unaware of any empirical studies using IVs which have passed these tests. In conclusion, if the substantive information concerning the potentially relevant factors is not assessed in relation to a statistically adequate (implicit) statistical model, IV-based inference is likely to be unreliable.

## 8.5 Recent developments[1]

The above criticisms of the Gauss-Markov (G-M) perspective apply equally well to its more recent extensions such as the *Generalized Method of Moments* (see Matyas, 1999, Hayashi, 2000) as well as certain *nonparametric* (see Pagan and Ullah, 1999) and *semi-parametric* methods (see Horowitz, 1998), when used as a basis for substantive inference. The specification of statistical models relying exclusively on substantive information, the absence of specific distributional assumptions, and the presence of non-testable assumptions, as well as the use of statistical models with 'broad' premises, are not conducive to reliable/precise inferences. As argued above, nonparametric procedures can be of great value, not for imbedding structural models and/or as a basis of substantive inferences, but for exploratory data analysis purposes. Such models can contribute decisively to the specification, misspecification testing and respecification facets of modeling, as expounded in the context of the PR approach; see Spanos (1999).

The dominating influence of the G-M perspective is clearly apparent in the textbook discussion of the statistical models introduced into econometrics since the 1960s, including *discrete and limited dependent* and *duration models* for **cross-section data** (see Wooldridge, 2002) and models for **panel data** (see Baltagi, 2001, Arellano, 2003). The discussion of these statistical models: (i) views the probabilistic structure of statistical models almost exclusively in terms of assumptions concerning *error terms*, (ii) these specifications are often incomplete, and invariably involve non-testable assumptions, (iii) the statistical analysis focuses primarily on *estimation*, and (iv) respecification is often confined to 'error-fixing'. As argued above, when framing the structural model the error terms plays a crucial role, but thinking in terms of an autonomous error term carrying the probabilistic structure of a statistical model can lead to numerous confusions.

Over the last 20 years or so, certain significant developments have taken place in time series econometrics, beginning with Granger and Newbold (1974) revisiting the spurious regression problem raised initially by Yule (1926). By simulating two uncorrelated Normal random walks ($x_t = x_{t-1} + \varepsilon_{1t},\ y_t = y_{t-1} + \varepsilon_{2t},\ E(\varepsilon_{1t}\varepsilon_{2t}) = 0$), they showed that inferences based the regression model $y_t = \beta_0 + \beta_1 x_t + u_t$ are seriously unreliable; the *actual* error probabilities are very different from the *nominal* ones. Phillips (1986) explained these results using non-standard inference propositions associated with unit roots in AR(p) models; see Dickey and Fuller (1979). Unfortunately, the literature missed the crucial lesson that *misspecification* is the real source of spurious regression, and instead focused on how the traditional inference propositions (sampling distributions of estimators and test statistics) need to be modified when modeling unit root processes; Phillips (1987). As a result, the empirical evidence accumulated by the unit root testing literature is largely unreliable,

[1]It is very important to re-iterate that the criticisms in this section are not directed toward the authors of the textbooks referenced, whom I regard with esteem, but towards the perspective that dominates current econometrics.

and in need to be re-considered in light of statistically adequate models; see Andreou and Spanos (2003). Moreover, the inadequate attention paid to the probabilistic prespective of unit root (UR) processes has misled this literature into the false premise that one can test for the presence of a unit root in the context of an AR(p) model, when in fact the AR and UR processes are *non-nested*; see Spanos and McGuirk (2002). On the positive side, the focus on unit roots has led to important theoretical developments in time series econometrics, such as *cointegration* and *error-correction* models; Engle and Granger (1987), Johansen (1991), Boswijk (1995).

Despite these important developments, the dominating influence of the G-M perspective is still apparent in the textbook discussion of statistical models for **times series data** (see Hamilton, 1994, Greene, 2003), but there are some encouraging signs that this literature is moving toward the probabilistic perspective propounded above. The first sign is that the Sims VAR and the Sargan-Hendry AD models seemed to have influenced the literature enough to pay more attention to the probabilistic structure of the data. Related to this, there has been growing interest in model evaluation and diagnostic checking, not only in the research literature (Newey, 1985; Tauchen, 1985, inter alia), but also at the textbook level; see Hendry (1995), Mills (1993) and Patterson (2000). Despite these encouraging signs, the current focus of statistical model specification is still on probabilistic assumptions for the error term, leading often to incomplete and sometimes internally inconsistent specifications; see Spanos (1995b). This renders misspecification testing and respecification a much less systematic activity than the PR approach envisages. In this sense the probabilistic perspective proposed in Spanos (1986) had very little impact on traditional econometric textbooks. Having said that, the discussion concerning 'model evaluation and diagnostic testing' in the last edition of the Johnston textbook (see Johnson and DiNardo, 1997) is very encouraging.

One can make a case that, what is currently needed, is to extend the Fisher-Neyman probabilistic perspective, as utilized in the case of time series models (see Spanos, 2001a), to the other statistical models for cross-section and panel data, as well as develop elaborated methods and procedures for specification, misspecification testing and respecification for these models. Such a probabilistic perspective will provide a complete and internally consistent set of testable probabilistic assumptions (in terms of observables) for these models, which is a necessary first step for ascertaining statistical adequacy. For instance, such a specification for the **Logit/Probit** 'regression-like' models (see Wooldridge, 2002, ch. 13) will look like table 4, where $F(z_k) = \frac{\exp(z_k)}{1+\exp(z_k)}$ and $F(z_k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_k} \exp(-\frac{u^2}{2})du$, $z_k := \boldsymbol{\alpha}^\top \mathbf{x}_k$, for the Logit and Probit models, respectively; the Poisson and duration 'regression-like' models are specified similarly. This makes it clear that the choice between the logit and probit models should be based on statistical adequacy grounds (testing thoroughly [2]-[5]),

and not on some 'goodness of fit' and/or 'theoretical meaningfulness' criteria.

**Table 4: The Logit/Probit model**

$$\boxed{Y_k = F(\boldsymbol{\alpha}^\top \mathbf{x}_k) + u_k,\ k \in \mathbb{N},}$$

| | | |
|---|---|---|
| [1] | Bernoulli: | $D(y_k \mid \mathbf{x}_k; \theta_k)$ is Bernoulli distributed, |
| [2] | Logit/Probit: | $E(Y_k \mid \mathbf{X}_k = \mathbf{x}_k) = F(\boldsymbol{\alpha}^\top \mathbf{x}_k)$, |
| [3] | Heteroskedasticity: | $Var(Y_k \mid \mathbf{X}_k = \mathbf{x}_k) = F(\boldsymbol{\alpha}^\top \mathbf{x}_k)[1 - F(\boldsymbol{\alpha}^\top \mathbf{x}_k)]$, |
| [4] | Independence: | $\{(Y_t \mid \mathbf{X}_k = \mathbf{x}_k), k \in \mathbb{N}\}$ - independent process, |
| [5] | $k$-homogeneity: | the parameters $\boldsymbol{\alpha}$ are not functions of $k \in \mathbb{N}$. |

Another example of how the probabilistic perspective can shed light on the structure of the **panel data models** concerns the standard question of 'fixed vs. random effects'. This question is posed as a choice between the following models:

$$\begin{array}{lll} \textbf{(a) fixed effects}: & y_{it} = \alpha_i + \mathbf{x}_{it}^\top \boldsymbol{\beta} + \varepsilon_{it}, & i \in \mathbb{N},\ \ t \in \mathbb{T}, \\ \textbf{(b) random effects}: & y_{it} = \alpha + \mathbf{x}_{it}^\top \boldsymbol{\beta} + (u_i + \varepsilon_{it}), & i \in \mathbb{N},\ \ t \in \mathbb{T}, \end{array}$$

accompanied by several assumptions (most of them non-testable) concerning the 'white-noise' error terms $(u_i, \varepsilon_{it})$, the most crucial being:

$$\begin{array}{lll} E(\varepsilon_{it} \mid \mathbf{X}) = 0, & E(u_i \mid \mathbf{X}) = 0, & E(\varepsilon_{it} u_j \mid \mathbf{X}) = 0, \text{ for all } i, j \text{ and } t, \\ E(\varepsilon_{it}^2 \mid \mathbf{X}) = \sigma_\varepsilon^2, & E(u_i^2 \mid \mathbf{X}) = \sigma_u^2, & E(\varepsilon_{it}\varepsilon_{jt} \mid \mathbf{X}) = 0,\ E(u_i u_j \mid \mathbf{X}) = 0,\ i \neq j; \end{array}$$

see Greene (2003), ch. 13. Transforming these assumptions in terms of the observables via $E(y_{it} \mid \mathbf{X}_{it} = \mathbf{x}_{it})$ and $Var(y_{it} \mid \mathbf{X}_{it} = \mathbf{x}_{it})$, reveals that the fixed effects term captures first moment heterogeneity, since in (a) $\alpha_i = E(y_{it}) - E(\mathbf{X}_{it})^\top \boldsymbol{\beta}$. By contrast, the random effects model imposes first order homogeneity (in (b) $\alpha = E(y_{it}) - E(\mathbf{X}_{it})^\top \boldsymbol{\beta}$), and $u_i$ can only capture second moment 'ephemeral' heterogeneity since it relates to the conditional variance $Var(y_{it} \mid \mathbf{X}_{it} = \mathbf{x}_{it}) = \sigma_u^2 + \sigma_\varepsilon^2$. Hence, the view that these two models are substitutes is misleading. Additional insight will be gained when all the probabilistic assumptions concerning the error terms are transformed into assumptions relating to the conditional distribution $D(y_{it} \mid \mathbf{X}_{it}; \boldsymbol{\theta})$, and the parameters $(\alpha_i, \alpha, \boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_\varepsilon^2)$ are given explicit statistical parameterizations stemming from the joint distribution $D(y_{it}, \mathbf{X}_{it}; \boldsymbol{\phi})$, analogous to those of the parameters in tables 2-3.

Although a detailed discussion of Bayesian methods in econometrics (see Zellner, 1971, Bauwens et al, 1999) is beyond the scope of this chapter, it is important to emphasize that the G-M perspective also permeates these methods, and they are equally susceptible to the unreliability and imprecision of inference problems discussed above. While a more explicit specification of the probabilistic assumptions of the model helps, the use of priors cannot address the issues raised by a misspecified statistical model. Equally ineffective for addressing statistical misspecification is Leamer's call for a Bayesian formalization of 'cookbook econometrics', because the formalization of fallacious 'error-fixing' strategies would create 'honest' but misguided modelers; it would not deal with the reliability and precision of inference problems.

# 9 Conclusions

The above assessment of 20th century developments in econometrics suggests that, despite the impressive developments in econometric methods and techniques, the vision to furnish apposite empirical foundations to economics remains largely unfulfilled. The current methodological framework, based on the *Gauss-Markov 'curve fitting' perspective,* has given rise to a *theory-dominated* approach to empirical modeling that invariably leads to unreliable empirical evidence. The unreliability of evidence stems primarily from applying statistical techniques for quantification *without* proper justification, i.e. they are produced by methods and procedures which *have very limited ability to detect errors* if, in fact, they were present. Inductive inferences which concern the sign and magnitude of the coefficients in a regression can only be justified in terms of the *validity of the premises*; the probabilistic assumptions underlying the statistical model are approximately true for the data in question. That requires thorough misspecification testing, not 'goodness of fit'.

It is argued that what is needed is a **methodological framework** which: (a) brings out the *gap between theory and data,* (b) encourages *error probing inquiries* at all stages of modeling, (c) affords the data 'a life of their own' by adopting the Fisher-Neyman probabilistic perspective, (d) specifies statistical models in terms of the observable processes, and (e) separates the statistical and structural model specification; the former is built upon the *statistical information* contained in the data, and the latter on the *substantive information* emanating from the theory. The probabilistic perspective views the *statistical model* in the context of all possible such models that could have given rise to data $\mathbf{Z}$, and provides the foundation and overarching framework for establishing statistical adequacy: specification, misspecification, respecification. Reliable theory testing takes place only when the structural model is confronted with a statistically adequate description of the systematic statistical information in the data. In addition to *statistical misspecification*, this **methodological framework** should emphasize *probing for errors* concerning the *accuracy of the data*, whether they measure what they are supposed to (*congruous measurement),* as well as whether the inductive inferences reached can be extended to the phenomenon of interest (*external validity).*

This methodological framework will create the pre-conditions for a **constructive dialogue between theory and data,** revolving around the question 'when do data $\mathbf{Z}$ provided evidence for a theory or a claim $H$? The form of inductive reasoning that best addresses this question is based on the notion of *severe testing* (see Mayo, 1996), which strongly encourages the probing of the different ways an inference might be in error. The severe testing reasoning can also shed light on several important methodological issues which concern the nature, interpretation, and justification of methods and models that are relied upon to learn from observational data. Only then, can 'learning from data' contribute significantly towards establishing economics as an empirical science.

# 10 Appendix - Basic Misspecification Tests

**Normality.** D'Agostino, R. B. and E. S. Pearson (1973) test.

**Linearity:** $H_0 : \gamma_2 = \gamma_3 = 0$, using the artificial regression:

$$\hat{u}_t = \gamma_0 + \gamma_1 x_t + \gamma_2 x_t^2 + \gamma_3 x_t^3 + v_t. \tag{18}$$

**Homoskedasticity:** $H_0 : \delta_1 = \delta_2 = 0$, using the artificial regression:

$$\hat{u}_t^2 = \delta_0 + \delta_1 x_t^2 + \delta_2 x_t^3 + v_t. \tag{19}$$

**Independence:** $H_0 : \alpha_2 = \alpha_3 = 0$, using the artificial regression:

$$\hat{u}_t = \alpha_0 + \alpha_1 x_t + \alpha_2 y_{t-1} + \alpha_3 x_{t-1} + v_t. \tag{20}$$

For the details, see Spanos (1986, 1999).

# References


[1] Abadir, K. and G. Talmain, (2002), "Aggregation, Persistence and Volatility in a Macro Model," *Review of Economic Studies*, **69**, 749-779.

[2] Aitken, A. C. (1934), "On least squares and liner combinations of observations," *Proceedings of the Royal Society of Edinburgh*, **55**, 42-48.

[3] Ali, M. M. and S. C. Sharma (1996), "Robustness to nonnormality of regression F-tests," *Journal of Econometrics*, **71**, 175-205.

[4] Andreou, E. and A. Spanos (2003), "Statistical adequacy and the testing of trend versus difference stationarity", *Econometric Reviews*, **22**, 217-252 (with discussion).

[5] Arellano, M. (2003), *Panel Data Econometrics*, Oxford University Press, Oxford.

[6] Backhouse, R. E. (1994), *New Directions in Economic Methodology*, Routledge, London.

[7] Baltagi, B. H. (2001), *Econometric Analysis of Panel Data*, 2nd ed., Wiley, NY.

[8] Bartlett, M. S. (1965), "R. A. Fisher and the Last Fifty Years of Statistical Methodology," *Journal of the American Statistical Association*, **60**, 395-409.

[9] Bennett, J. H. (1990), ed., *Statistical Inference and Analysis: Selected correspondence of R. A. Fisher*, Clarendon Press, Oxford.

[10] Blaug, M. (1992), The Methodology of Economics, Cambridge University Press, Cambridge.

[11] Boswijk, P. H. (1995), "Efficient inference on cointegration parameters in structural error correction models," *Journal of Econometrics*, **69**, 133-158.

[12] Bowley, A. L. (1920), *Elements of Statistics*, 4th ed., P.S. King and Son, London.

[13] Box, G. E. P. and G. M. Jenkins (1970/1976) *Time series analysis: forecasting and control*, (revised edition) Holden-Day, San Francisco.

[14] Burns, A. F. and W. C. Mitchell (1946), *Measuring Business Cycles*, National Bureau of Economic Research, New York.

[15] Cartwright, N. (1983), *How the Laws of Physics Lie*, Clarendon Press, Oxford.

[16] Chalmers, A. F. (1999), *What is this thing called Science?*, 3rd ed., Hackett, Indianapolis.

[17] Christ, C. F. (1985), "Early Progress in Estimating Quantitative Economic Relationships in America," *American Economic Review*, **75**, 39-52.

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

[19] Cooper, R. L. (1972), "The predictive performance of quarterly econometric models of the United States," in *Econometric Models of Cyclical Behavior*, ed. by Hickman, B. G., Columbia University Press, NY.

[20] D'Agostino, R. B. and E. S. Pearson (1973), "Tests for departure from normality. Empirical results for the distributions of $b_2$ and $\sqrt{b_1}$. *Biometrika*, **60**, 613-622.

[21] David, F. N. and J. Neyman (1938), "Extension of the Markoff theorem on least squares," *Statistical Research Memoirs*, **II**, 105-116.

[22] Davis, H. T. (1941), *The Theory of Econometrics*, Principia Press Inc., Indiana.

[23] Dickey, D. A. and W. A. Fuller (1979), "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, **74**, 427-431.

[24] Doob, J. L. (1953), *Stochastic Processes*, Wiley, New York.

[25] Durbin, J. and G. S. Watson (1950), "Testing for serial correlation in least squares regression I", *Biometrika*, **37**, 409-428.

[26] Efron, B. and R. Tibshirani (1993), *An Introduction to Bootstrap*, Chapman and Hall, London.

[27] Engle, R. F. and C. W. J. Granger (1987), "Co-integration and Error-correction: representation, estimation and testing," *Econometrica*, **55**, 251-276.

[28] Epstein, R. J. (1987), *A History of Econometrics*, North Holland, Amsterdam.

[29] Fisher, I. (1911), *The purchasing power of money*, MacMillan, New York.

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

[31] Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh.

[32] Fisher, R. A. (1935), *The Design of Experiments*, Oliver and Boyd, Edinburgh.

[33] Fisher, R. A. (1955) "Statistical methods and scientific induction," *Journal of the Royal Statistical Society*, **B**, **17**, 69-78.

[34] Fox, K. A. (1989), "Agricultural Economists in the Econometric Revolution: Institutional Background, Literature and Leading Figures," *Oxford Economic Papers*, **41**, 53-70.

[35] Friedman, M. (1953), *Essays in Positive Economics*, University of Chicago Press, Chicago.

[36] Frisch, R. (1933), "Editorial," *Econometrica*, **1**, 1-4.

[37] Frisch, R. (1934), *Statistical Confluence Analysis by Means of Complete Regression Systems*, Univeritetets Okonomiske Institutt, Oslo.

[38] Gilbert, C. L. (1991), "Richard Stone, demand theory and the emergence of modern econometrics, *Economic Journal*, **101**, 288-302.

[39] Goldberger, A. S. (1964), *Econometric Theory*, John Wiley & Sons, New York.

[40] Granger, C. W. J. (1990), (ed.) *Modelling Economic Series*, Clarendon Press, Oxford.

[41] Granger, C. W. J. and P. Newbold (1974), "Spurious regressions in econometrics," *Journal of Econometrics*, **2**, 111-120.

[42] Granger, C. W. J. and P. Newbold (1986), *Forecasting Economic Time Series*, 2nd, ed., Academic Press, London.

[43] Greene, W. H. (2002), *Econometric Analysis*, 5th ed., Prentice Hall, NJ.

[44] Haavelmo (1943), "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, **11**, 1-12.

[45] Haavelmo, T. (1944), "The probability approach to econometrics", *Econometrica*, **12**, suppl., 1-115.

[46] Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge University Press, Cambridge.

[47] Hald, A. (1998), *A History of Mathematical Statistics From 1750 to 1930*, Wiley, NY.

[48] Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press, NJ.

[49] Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators, " *Econometrica*, **97**, 93-115.

[50] Harrod, R. F. (1938), "Scope and Method of Economics," *Economic Journal*, **48**, 383-412.

[51] Hayashi, F. (2000), *Econometrics*, Princeton University Press, Princeton, NJ.

[52] Heckman, J. J. (1992), "Haavelmo and the Birth of Modern Econometrics: A Review of The History of Econometric Ideas by Mary Morgan," *Journal of Economic Literature*, **XXX**, 876-886.

[53] Hendry, D. F. (1976), "The Structure of Simultaneous Equations Estimators," *Journal of Econometrics*, **4**, 51-88.

[54] Hendry, D. F. (1993), *Econometrics: Alchemy or Science?*, Blackwell, Oxford.

[55] Hendry, D. F. (1995), *Dynamic Econometrics*, Oxford University Press, Oxford.

[56] Hendry, D. F. and J.-F. Richard (1982), "On the Formulation of Empirical Models in Dynamic Econometrics," *Journal of Econometrics*, **20**, 3-33.

[57] Hendry, D. F., E. E. Leamer and D. J. Poirier (1990), "The ET dialogue: a conversation on econometric methodology," *Econometric Theory*, **6**, 171-261.

[58] Hendry, D. F., A. Spanos and N. Ericsson (1989) "Trygve Haavelmo's contributions to econometrics", *Socialokonomen*, **11**, 12-17.

[59] Hendry, D. F. and M. S. Morgan (1995), (ed.) *The foundations of econometric analysis*, Cambridge University Press, Cambridge.

[60] Hood, W. C. and Koopmans, T. C. (1953), (eds.), *Studies in Econometric Method*, Cowles Commission Monograph, No. 14, John Wiley & Sons, New York.

[61] Hoover, K. D. (2001), *Causality in Macroeconomics*, Cambridge University Press, Cambridge.

[62] Horowitz, J. L. (1998), *Semiparametric Methods in Econometrics*, Springer-Verlag, N.Y.

[63] Hutchison, T. W. (1938), *The Significance and Basic Postulates of Economic Theory*, Augustus M. Kelly reprints, 1965, New York.

[64] Jeffreys, H. (1939), *Theory of Probability*, Oxford University Press, Oxford.

[65] Jevons, W. S. (1871), *The Theory of Political Economy*, MacMillan, London.

[66] Jevons, W. S. (1874), *The Principles of Science*, MacMillan, London.

[67] Johansen, S. (1991), "Estimation and hypothesis testing of cointegrating vector of Gaussian vector autoregressive models," *Econometrica,* **59**, 1551-1580.

[68] Johnston, J. (1963), *Econometric Methods*, 4th ed., McGraw-Hill Book Co., New York.

[69] Johnston, J. and J. DiNardo (1997), *Econometric Methods*, McGraw-Hill Book Co., New York.

[70] Judge, G. G., C. R. Hill, W. E. Griffiths, H. Lutkepohl and T-C. Lee (1988) *Introduction to the theory and practice of econometrics*, Wiley, New York.

[71] Kennedy, P. (2003), *A Guide to Econometrics*, 5th edition, MIT Press, Cambridge.

[72] Keynes, J. M. (1921), *A Treatise on Probability*, McMillan, London.

[73] Keynes, J. M. and Tinbergen, J. (1939-40), "Professor's Tinbergen's Method," *Economic Journal*, **49**, 558-568, "A Reply," by J. Tinbergen, and "Comment," by Keynes, **50**, 141-156.

[74] Keynes, J. N. (1891), *The Scope and Method of Political Economy*, McMillan, London.

[75] Klein, L. R. and A. S. Goldberger (1955), *An Econometric Model of the United States, 1929-1952*, North-Holland, Amsterdam.

[76] Knight, F. H. (1940), ""What is Truth" in Economics?," *Journal of Political Economy*, **48**, 1-32.

[77] Kolmogorov, A. N. (1933), *Foundations of the theory of Probability*, 2nd English edition, Chelsea Publishing Co. New York.

[78] Koopmans, T. C. (1939), *Linear Regression Analysis of Economic Time Series*, Netherlands Economic Institute, Publication No. 20, Haarlem, F. Bohn.

[79] Koopmans, T. C. (1945), "Statistical Estimation of Simultaneous Economic Relations," *Journal of the American Statistical Association*, **40**, 448-466.

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

[81] Koopmans, T. C. (1950), (ed.), *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph, No. 10, John Wiley & Sons, New York.

[82] Kuznets, S. (1950), book review of *On the accuracy of economic observations*, by O. Morgenstern (1950), *Journal of the American Statistical Association*, **45**, 576-579.

[83] Leamer, E. E. (1978), *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley, New York.

[84] Leamer, E. E. and H. B. Leonard (1983), "Reporting the fragility of regression estimates, *Review of Economics and Statistics*, 65, 306-317.

[85] Lehmann, E. L. (1990), "Model specification: the views of Fisher and Neyman, and later developments", *Statistical Science*, **5**, 160-168.

[86] Lehmann, E. L. (1993), "The Fisher and Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association*, **88**, 1242-9.

[87] Leontief, W. W. (1948), "Econometrics," in *A Survey of Contemporary Economics*, ed. H. S. Ellis, R. D. Irwin, Homewood, Illinois.

[88] Leontief, W. W. (1971), "Theoretical Assumptions and Nonobserved Facts," *American Economic Review*, 1-7.

[89] Lucas, R. E. (1976), "Econometric Policy Evaluation: a Critique," pp. 19-46. of *The Phillips Curve and Labour Markets*, ed. by K. Brunner and A. M. Metzer, Carnegie-Rochester Conference on Public Policy, I. North-Holland, Amsterdam.

[90] Lucas, R. E. and T. J. Sargent (1981), *Rational Expectations and Econometric Practice*, George Allen & Unwin, London.

[91] Maki, U. (2002), *Fact and Fiction in Economics*, Cambridge University Press, Cambridge.

[92] Marshall, A. (1890), *Principles of Economics*, McMillan, London.

[93] Matchlup, F. (1955), "The Problem of Verification in Economics," *Southern Economic Journal*, **22**,1-21.

[94] Matyas, L. (1999), (editor), *Generalized Method of Moments Estimation*, Cambridge University Press, Cambridge.

[95] Mayo, D. G. (1991), "Novel Evidence and Severe Tests", *Philosophy of Science,* **58**, 523-552.

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

[97] Mayo, D. G. (2004), "Philosophy of Statistics," forthcoming in S. Sarkar (ed.) *Routledge Encyclopedia of the Philosophy of Science.*

[98] Mayo, D. G. and A. Spanos (2003), "Severe Testing as a Basic Concept in the Neyman-Pearson Philosophy of Induction," Virginia Tech working paper.

[99] Mayo, D. G. and A. Spanos (2004), "Methodology in Practice: Statistical Misspecification Testing," forthcoming, *Philosophy of Science.*

[100] McGuirk, A. and A. Spanos (2003) "Revisiting Error Autocorrelation Correction: Common Factor Restrictions and Granger Non-Causality," Virginia Tech working paper.

[101] Mill, J. S. (1874), *Essays on some unsettled questions of political economy*, 2nd ed., Longmans, London.

[102] Mill, J. S. (1884), *A System of Logic*, 8th ed., Harper and Brothers, New York.

[103] Mills, F. C. (1924/1938), *Statistical Methods*, Henry Holt and Co., New York.

[104] Mills, T. C. (1993), *The Econometric Modelling of Financial Time Series*, Cambridge University Press, Cambridge.

[105] Mitchell, W. C. (1927), *Business Cycles: The Problems and its Setting,* National Bureau of Economic Research, New York.

[106] Mizon, G. E. (1995), "Progressive Modelling of Macroeconomic Time Series: The LSE Methodology," in *Macroeconometrics: Developments, Tensions and Prospects*, Hoover, K. D. (ed.), Kluwer, Dordrecht.

[107] Moore, H. L. (1908), "The Statistical Complement of Pure Economics," *Quarterly Journal of Economics*, **23**, 1-33.

[108] Moore, H. L. (1911), *The Law of Wages*, McMillan, New York.

[109] Moore, H. L. (1914), *Economic Cycles - Their Laws and Cause*, McMillan, New York.

[110] Morgan, M. S. (1990), *The history of econometric ideas*, Cambridge University Press, Cambridge.

[111] Morgenstern, O. (1950/1963) *On the accuracy of economic observations*, 2nd edition, Princeton University Press, New Jersey.

[112] Mosteller, F. and J. W. Tukey (1977), *Data Analysis and Regression*, Addison-Wesley.

[113] Newey, W. K. (1985), "Maximum Likelihood specification testing and conditional moment tests," *Econometrica*, **53**, 1047-1070.

[114] Neyman, J. (1950), *First Course in Probability and Statistics*, Henry Holt, New York.

[115] Neyman, J. (1952), *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd ed. U.S. Department of Agriculture, Washington.

[116] Neyman, J. (1956), "Note on an Article by Sir Ronald Fisher," *Journal of the Royal Statistical Society*, B, **18**, 288-294.

[117] Neyman, J. (1976), "Tests of Statistical Hypotheses and their use in Studies of Natural Phenomena," *Communications in Statistics – Theory and Methods*, **5**, 737-751.

[118] 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.

[119] Orcutt G. H. (1952), Book review of *Statistical Inference in Dynamic Economic Models*, ed. by Koopmans, T. C. (1950), *The American Economic Review*, **42**, 165-169.

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

[121] Pagan, A.R. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press, Cambridge.

[122] Patterson, K. (2000), *Introduction to Applied Econometrics: A Time Series Approach*, MacMillan, London.

[123] Pearson, K. (1895), "Contributions to the mathematical theory of evolution II. Skew variation in homogeneous material", *Philosophical Transactions of the Royal Society of London, series A*, **186**, 343-414.

[124] Pearson, K. (1920), "The Fundamental Problem of Practical Statistics," *Biometrika*, **XIII**, 1-16.

[125] Peirce, C. S. (1878), "The Probability of Induction," *Popular Science Monthly*, **12**, 705-718.

[126] Phillips, P. C. B. (1986), "Understanding spurious regression in econometrics," *Journal of Econometrics*, **33**, 311-40.

[127] Phillips, P. C. B. (1987), "Time series regressions with a unit root," *Econometrica*, **55**, 227-301.

[128] Politis, D. N. and J. P. Romano, M. Wolf (1999), *Subsampling*, Springer, New York.

[129] Quin, D. (1993), *The Formation of Econometrics: a Historical Perspective*, Clarendon Press, Oxford.

[130] Rao, C. R. (1992), "R. A. Fisher: The Founder of Modern Statistics," *Statistical Science*, **7**, 34-48.

[131] Rao, C. R. (2004), "Statistics: Reflections on the Past and Visions for the Future," *Amstat News*, **327**, 2-3.

[132] Redman, D. A. (1997), *The Rise of Political Economy as a Science*, The MIT Press, Cambridge.

[133] Robbins, L. (1932/1937), *An Essay on the Nature and Significance of Economic Science*, McMillan, London.

[134] Rosenblatt, M. (1956) "Remarks on some nonparametric estimates of a density function", *Annals of Mathematical Statistics*, **27**, 832-35.

[135] Salmon, W. (1966), *The Foundations of Scientific Inference*, University of Pittsburgh Press, Pittsburgh.

[136] Savage, L. J. (1954) *The Foundations of Statistics*, Wiley, New York.

[137] Scheffe, H. (1943), "Statistical Inference in the Non-parametric case," *Annals of Mathematical Statistics*, 14, 305-332.

[138] Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, Springer-Verlag, New York.

[139] Sims, C. A. (1980), "Macroeconomics and Reality," *Econometrica*, **48**, 1-48.

[140] Sims, C. A. (1982), "Policy Analysis with Econometric Models," *Brookings Papers on Economic Activity*, 107-164.

[141] Slutsky, E. (1927), "The summation of random causes as the source of cyclic processes" (in Russian); English translation in *Econometrica*, **5**, 1937.

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

[143] Spanos, A. (1988), "Towards a Unifying Methodological Framework for Econometric Modelling", *Economic Notes*, 107-34.

[144] Spanos, A. (1989), "On re-reading Haavelmo: a retrospective view of econometric modeling", *Econometric Theory*, **5**, 405-429.

[145] Spanos, A. (1990), "The Simultaneous Equations Model revisited: statistical adequacy and identification", *Journal of Econometrics*, **44**, 87-108.

[146] Spanos, A. (1995a), "On theory testing in Econometrics: modeling with non-experimental data", *Journal of Econometrics,* 67:189-226.

[147] Spanos, A. (1995b), "On Normality and the Linear Regression model", *Econometric Reviews*, **14**, 195-203.

[148] Spanos, A. (1999), *Probability Theory and Statistical Inference: econometric modeling with observational data*, Cambridge University Press, Cambridge.

[149] Spanos, A. (2000), "Revisiting Data Mining: 'hunting' with or without a license," *The Journal of Economic Methodology*, **7**, 231-264.

[150] Spanos, A. (2001a), "Time series and dynamic models," ch. 28, pp. 585-609, *A Companion to Theoretical Econometrics*, edited by B. Baltagi, Blackwell Publishers, Oxford.

[151] Spanos, A. (2001b), "Parametric versus Non-parametric Inference: Statistical Models and Simplicity," pp. 181-206 in *Simplicity, Inference and Modelling*, edited by A. Zellner, H. A. Keuzenkamp and M. McAleer, Cambridge University Press.

[152] Spanos, A. (2004), "Structural Equation Modeling, Causal Inference and Statistical Adequacy," forthcoming in the proceeding of the *12th International Congress of Logic, Methodology and Philosophy of Science*.

[153] Spanos, A. (2005a), "Where Do Statistical Models Come From? Revisiting the Problem of Specification," forthcoming in the 2nd Erich Lehmann symposium volume.

[154] Spanos, A. (2005b), "Structural vs. Statistical Models in Econometric Modeling," Virginia Tech working paper.

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

[156] Spanos, A. and A. McGuirk (2002), "Revisiting the foundations of unit root testing: statistical parameterizations and implicit restrictions," Virginia Tech working paper.

[157] Stigler, G.J. (1954), "The early history of the empirical studies of consumer behavior", *The Journal of Political Economy*, **62**, 95-113. Reprinted in Stigler, G. J. (1962).

[158] Stigler, G. J. (1962), "Henry L. Moore and Statistical Economics", *Econometrica*, **30**, 1-21.

[159] Stigler, S. M. (1986), *The history of statistics: the measurement of uncertainty before 1900,* Harvard University Press, Cambridge, Massachusetts.

[160] Stigum, B. P. (1990), *Toward a Formal Science of Economics*, MIT Press, Cambridge.

[161] Stigum, B. P. (2003), *Econometrics and the Philosophy of Economics*, Princeton University Press, Princeton.

[162] Stone, J. R. N. (1954a), "Linear Expenditure systems and demand analysis: an application to the pattern of British demand," *Economic Journal,* **64,** 511-527.

[163] Stone, J. R. N. (1954b), *The Measurement of Consumers' Expenditure and Behaviour in the United Kingdon, 1920-1938*, Cambridge University Press, Cambridge.

[164] Summers, L. (1991), "The Scientific Illusion in Empirical Macroeconomics," *Scandinavian Journal of Economics*, **93**, 129-143.

[165] Tauchen, G. E. (1985), "Diagnostic testing and evaluation of maximum likelihood models," *Journal of Econometrics*, **30**, 415-443.

[166] Tinbergen, J. (1939), *Statistical Testing of Business Cycle Research*, 2 vols., League of Nations, Geneva.

[167] Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley.

[168] Valavanis, S. (1959), *Econometrics*, McGraw-Hill, New York.

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

[170] Wold, H. O. (1938), *A Study in the Analysis of Stationary Time Series*, Almquist and Wicksell, Uppsala.

[171] Wooldridge, J. M. (2002), *Econometric Analysis of Cross-Section and Panel Data*, The MIT Press, Cambridge, Massachusetts.

[172] Yule, G.U. (1926), "Why do we sometimes get nonsense correlations between time series-a study in sampling and the nature of time series ", *Journal of the Royal Statistical Society*, **89**, 1-64.

[173] Yule, G.U. (1927), "On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers ", *Philosophical Transactions of the Royal Society,A*, **226**, 267-298.

[174] Zellner, A. (1971), *Introduction to Bayesian Inference in Econometrics*, Wiley, New York.