Department of Economics
Working Paper Series

The Empirical Significance of Econometric Models

Thomas Mayer
University of California, Davis

May 16, 2006

Paper # 06-20

This essay discusses some, but by no means all the important problems that arise in econometric testing of econometric models. Specifically, it disusses the reliabilty of the underlying data and their processing, the problem of relating theories and data, ceteris paribus conditions and testability, data mining and the misuse of significance test.

UCDAVIS

Department of Economics
One Shields Avenue
Davis, CA 95616
(530)752-0741

http://www.econ.ucdavis.edu/working_search.cfm

THE EMPIRICAL SIGNIFICANCE OF ECONOMETRIC MODELS

Thomas Mayer

University of California, Davis

Abstract

This essay discusses some, but by no means all the important problems that arise in econometric testing of economic models. Specifically, it discusses the reliability of the underlying data and of their processing, the problem of relating theories and data, ceteris paribus conditions and testability, data mining and the misuse of significance test.

The Empirical Significance of Econometric Models

Thomas Mayer

Most of the papers in this volume analyze in detail a specified problem of economic measurement. This paper takes a more general approach and surveys a number of problems that limit the empirical evaluations of economic models. It takes as given that economic models should have direct empirical relevance, so that they need to be empirically tested. It therefore excludes what might be called "housekeeping models", that is models intended to clarify theory by, for example, unifying lower-level models. It also excludes what Allen Gibbard and Hal Varian (1978) have called "caricature models", that is models that investigate what happens if a particular variable is allowed to assume extreme values. Such a model can be extremely useful in providing understanding, what Fritz Machlup (1950) has called a sense of "Ahaness"[1]. But it is usually not subject to econometric testing.

Instead, I focus on some specific problems in testing the type of models that usually are subjected to econometric tests. Such problems being numerous and space being limited, I take up just a few, concentrating on those that are at least to some extent remediable by available methods, and ignore others, such as the problem of inferring causality, the Lucas critique and also the fundamental problems discussed in the last chapter of Aris Spanos' (1986) textbook. Though they are hardly trivial I also do not discuss the problems created by ideological commitment, excessive loyalty to one's school of thought, the reluctance to admit error, and the frenzied following of fashion.[2] However, I have not been reluctant to discuss problems that – in

principle - are well known, but ignored in practice. But before looking at difficulties of econometric testing let us ask whether they really matter, that is whether economists take empirical tests seriously.

I. Do Economists Take Empirical Evidence Seriously?

Some pessimists have argued that in economics empirical evidence is not taken seriously when it conflicts with an appealing theoretical model. For example McCloskey (1985, p. 182) wrote: "no proposition about economic behavior has yet been overturned by econometrics, at any rate not to the standard that the hypothetico-deductive model of science would demand.", and Aris Spanos (1986, p. 660) stated " ... to my knowledge no economic theory was ever abandoned because it was rejected by some empirical econometric test, nor was a clear-cut decision between competing theories " made on the basis of such a test. Such statements are hard to evaluate because they are vague. Do they include only major, generally accepted propositions or also claims made in a specific paper that has not been widely cited? Moreover, what does "overturned" mean? Suppose a paper makes a claim that is then rejected by an econometric test in another paper, and neither paper is cited thereafter. Does that count? Or suppose a model that implies that raw material prices rise over time at a rate equal to the interest rate, is rejected not only by econometric tests, but also by informal observation. Does that count? If it does it is an obvious counter- example. Besides, well-entrenched propositions - in the physical sciences as well as in economics - seldom fall as a result of a single piece of evidence. And that is more a sign of common sense and good judgment than of faulty methodology.

Let's therefore bet more specific and ask whether economists have stuck with irrational stubbornness to their model of rational, maximizing behavior, despite extensive empirical evidence to the contrary? (Cf. Hausman, 1992.) In an important way they have, but with two substantial qualifications. First, even within mainstream economics this theory is now being challenged by behavioral economics. Second, much of economic theory, for example, Keynesian theory, monetarism, and comparative- cost theory, require only weak versions of rational maximizing behavior, versions that are much less challenged by the empirical

evidence cited against the stronger versions. Empirical evidence has much more influence on what economists actually do when dealing with practical problems than it has on what is emblazoned on their banner. Hard-core versions of new classical theory, Ricardian equivalence, efficient markets theory, etc., that require a rigorous maximizing model cater primarily to niche markets. Perhaps economists should be blamed, not for sticking to disconfirmed hypotheses, but for fooling others (in particular philosophers) by proclaiming what they do not really believe or practice. A reason they do that is that the weakened version of the rational-maximizing principle that economists use in practice is hard to formulate, particularly since it depends on the specific issue being addressed.

This is no to deny that economists sometimes do show excessive fondness for empirically weak theories. Some of their beliefs seem invulnerable to empirical evidence. Macro- economists widely accept rational-expectations theory despite the empirical evidence against it. (See Goldfarb and Stekler, 2000.) I conjecture that no economist ever accepted this theory because he or she found its empirical evidence convincing. Rather, it is widely accepted because of its elegance, and because it seems to be a necessary implication of rational behavior, and must therefore be defended to the death. But even here ongoing research on how agents learn, and the introduction of such learning models into macro models - a move that does not bring the assumption of rational behavior into question - is now taming rational-expectations theory. Eventually reason wins out - even in academia.

A more specific look at individual papers, also tells a mixed story. It is common for papers to test the hypothesis or model that they present. As discussed below, due to data mining successes on these tests are not all that impressive. Yet, in the process of data mining the hypothesis is modified, so that the data do have some influence.

Hence, neither the simplistic dogma that economics is a "science - which among other things ignores that we lack an adequate definition of science – nor the cynical view that economic theories are invulnerable

to contradictory data is plausible. It is therefore useful to look at some problems that inhibit the empirical testing of economic models.

II. Reliability of the Data and of their Processing

As several economists have pointed out (see for instance, Leontief 1971) most economists show little concern about the quality of their data.[3] To be sure, they make allowance for sampling error, but that's about it. The standard justifications for this unconcern are first that the obvious need to quantify and test our hypotheses forces us to use whatever data we can find, and as long as they are the best available data, well, that's all we can be expected to do. Second, previous researchers have already decided what the best data sets are, so we can just use these.

Sounds compelling - but isn't. Yes, empirical testing is important, but in some cases even the best available data may not be reliable enough to test the model, and then you should either develop a better data set on our own, or else admit that your model cannot, at least at present, be adequately tested. Or if the available data are neither wholly reliable nor totally inadequate you still might use them, but only if you inform the reader about the problem, and perhaps do some robustness testing. That others have used a data set is not an adequate justification for your using it, not only because of uncertainty about whether the previous use was valid, but also because, while for some purposes crude estimates suffice, for others they do not. Don't assume that the sophistication of your econometrics can necessarily compensate for the inadequacy of your data.(Cf. Chatfield, 1991) Time spent on cleaning the data, or looking for a data set that provides a better measure of your model's variables, and getting intimately familiar with the data (see Kennedy, 2002; Magnus, 1999), may not impress a referee, but it may improve the results more than the same time spent learning the latest technique. As Daniel Hamermesh (2000, p. 365) has remarked: "data may be dirty, but in many cases the dirt is more like mud than Original Sin."

In more concrete terms suppose the data seem to disconfirm the hypothesis because the t value of the critical coefficient is low, or because other regression diagnostics look poor. Both of these may be due to data errors and not an error in the hypothesis. To illustrate with an extreme case, albeit one involving an identity rather than a hypothesis, few would deny that total exports equal total imports, even though the data show them not to. Conversely, data errors may sometimes favor the hypotheses. For example, because of a lack of better data the compilers of a series may have estimated an important component by a simple trend. If the model contains a regressor dominated by a similar trend this data error could provide spurious support for the model.

Because of the reluctance of economists to get involved in the messy details of how their data were derived certain standard conventions are used without question. To illustrate the type of problem frequently swept under the rug consider the savings ratio. How many economists who build models to explain this ratio discuss whether they should use the savings estimates given in the National Income and Product Accounts (NIPA). or else the very different savings estimates that can be derived from the flow-of-funds accounts? The former are generally used even though they derive saving by subtracting consumption from income, and are therefore at least potentially subject to large percentage errors.[4] (The flow-of-funds estimates also have their problems.) Moreover, as Marshall Reinsdorf (2005) has pointed out, there are some specific problems with the interpretation of the NIPA savings data, and that the particular measure one should use depend on the purpose at hand. One is that the personal income data include income received on behalf of households by pension funds and nonprofit organizations that serve households, that is income that households may not be aware of and take into consideration when deciding on their consumption. Data on the difference between the NIPA personal savings rate and the savings rate of households that exclude these receipts are available since 1992, and while the difference is trivial in 1992-1994, it amounts to 0.7 percentage points - that is about 30 percent of the savings ratio - in 1999 and 2000. Another problem is that the NIPA data treat as interest

income nominal instead of real interest receipts. (And a similar thing applies to interest payments on consumer debt.) Using real instead of nominal interest payments reduces the personal savings rate by 1.5 to 2.4 percentage points during 1980-92, but only by 0.5 to 1.2 percentage points in 1993-2000

Another problem is the treatment of capital gains and losses. The NIPA data exclude capital gains from income, and hence from saving, but they deduct the taxes paid on realized capital gains from disposable personal income, and thus indirectly from personal saving. Using an alternative measure that includes in disposable personal income federal taxes on capital gains changes the recorded savings rate by only 0.5 percentage points in 1991-92 but by 1.65 percentage points in the unusual year, 2000. And then there is the important question whether at least some of the unrealized capital gains and losses shouldn't be counted as saving, since over the long run capital gains are a major component of the yield on equities.

Other data sets have other problems. For instance the difficulties of measuring the inflation rate are well known, and since real GDP is derived by deflating nominal GDP, errors in estimating the inflation rate generate corresponding errors with the opposite sign in estimated real GDP. Moreover, real GDP estimates are downward biased because of an underground economy that might account for 10 percent or more of total output. Furthermore, GDP revisions are by no means trivial, which raises the question of how reliable the final estimates are. Balance of payments statistics, too, are notoriously bad. The difficult of defining money operationally has led to the quip that the demand for money is stable; it is just the definition of money that keeps changing. And even if one agrees on the appropriate concept of money, real time estimates of the quarterly growth rates of money are unreliable. The problems besetting survey data, such as misunderstood questions and biased answers are also large. Moreover, in using survey data it has become a convention in economics not to worry about a possible bias due to non-response, even when the non-response rate is, say 65 percent.

My point here is not that the available data are too poor to test our models. That I believe would be an overstatement. It is also not that economists use wrong data sets, but rather that they tend to select their data sets in a mechanical way without considering alternatives, or asking whether the data are sufficiently accurate for the purpose at hand. And they need to pay attention to just what the question is that they are trying to answer (see Magnus, 1999). Otherwise there is the danger of confusing the question for which we have data with the question we claim to answer.

There is also a serious danger of errors in data entry, in calculations, and in the transcription of regression results. Dewald, Thursby and Anderson (1986), show that such errors were frequent and substantial. Perhaps as a result of this paper they are now much less common, but perhaps not.[5] Downloading data from a standard database is not a complete safeguard against errors. Without even looking for them I have twice found a substantial error in a widely used database.

Moreover, since various popular software packages can yield sharply different results, regression programs, too, can generate substantial errors. (See Lovell and Selover, 1994, McCullough and Vinod, 1999, McCullough, 2000.) In particular, McCullough and Vinod speak of:

> the failure of many statistical packages to pass even rudimentary benchmarks for numerical accuracy. ... [E]ven simple linear procedures, such as calculation of the correlation coefficient can be horrendously inaccurate. ... While all [three popular] packages tested did well on linear regression benchmarks - gross errors were uncovered in analyses of variance routines. ... [There are] many procedures for which we were unable to find a benchmark and for which we found discrepancies between packages: linear estimation with AR(1) errors, estimation of an ARMA model, Kalman filtering, ... and so on." (pp. 633, 635, 650, 655)

Because this paper appeared in the Journal of Economic Literature many economists were surely aware of it. One might therefore have expected them to have recalculated computations in their previously published papers using alternative software packages, and the journals to be full of errata notices. This did not happen. (Mea culpa.) In checking Google for references to the McCullough and Vinod paper for such corrections I did not find a single one.[6]

Even allowing for the natural reluctance to retract one's results, and a tendency for herding (and hence for thinking that if nobody else worries, why should I?) this nonchalant attitude is not easy to reconcile with the claim that economics is a "science", or even that it is a serious discipline. And yet this "who cares?" attitude should not be surprising to someone who takes our portrayal of "economic man" seriously, because there is only a small chance that an error will be caught. But while we therefore need a system of routinely checking at least some published results (say 5 percent) to discourage both carelessness and occasionally even fraud, we are not likely to get one.

In the natural sciences, too, mechanical checking of other people's results is rare. (See Mirowski and Skilivas, 1991.) But instead of checking the mechanics, such as the correctness of calculations, natural scientists try to replicate the results, that is they look for similar results in similar circumstances. (See Backhouse, 1992.) For example, they may repeat an experiment at a different temperature. If they get similar results then that confirms the original findings, and if they do not, that can be read either as a limitation of the domain of the model or as casting doubt on it. If many replications fail to confirm the original findings these are then treated as, at best, a special case. Such replication is not common in economics.

III. Three Basic Problems in Testing Economic Theories

Suppose that the data set contains no errors and provides unequivocal information, that there are no mechanical errors, and that the regression program is accurate. That still leaves several serious problems.

1. Relating Theories and Data

The traditional procedure is to select as the regressors the major variables implied by the model, run the regression, and then, if necessary add, or perhaps eliminate, some regressors until the diagnostics look good. An alternative procedure coming from LSE econometricians is to use a large number of regressors, some of which may not be closely tied to the hypothesis being tested, and then narrow the analysis by dropping those with insignificant coefficients. Such a search for the data generating process (DGP) usually puts more stress

on meeting the assumptions of the underlying statistical model, emphasizes misspecification tests, and rejects quick fixes, such as adding an AR term, than does the traditional procedure, though it does not reject the criteria used in the traditional approach. Thus Spanos (1986, pp. 669-70) cites the following criteria: "theory consistency, goodness of fit, predictive ability, robustness (including nearly orthogonal explanatory variables), encompassing [the results of previous work and] parsimony."

What is at stake here is a more fundamental disagreement than merely a preference for either starting with a simple model and then adding additional variables until the fit becomes satisfactory. or else starting with a general model and then dropping regressors that are not statistically significant. Nobody can start with a truly general model and if the reduction does not provide a satisfactory solution a LSE econometrician, too, is likely to add additional regressors at that stage (see Keuzenkamp and McAleer, 1995).

The more fundamental disagreement can be viewed in two ways. The first is as emphasizing economic theory versus emphasizing statistical theory. In the former case one may approach a data set with strong priors based on the theory's previous performance on other tests or it's a priori plausibility. One then sees whether the new data set is also consistent with that theory rather than asking which hypothesized DGP gives the most satisfactory diagnostics, Suppose that the quantity theory gives a good fit for the inflation rates of twenty countries . However, for each of these countries one can estimate a DGP that gives a better fit, but contains an extra variable that differs from country to country. One may then still prefer the quantity theory. LSE econometricians would probably agree, but in practice their method tends to stress econometric criteria rather than the other criteria relevant to theory choice. This issue is well stated by Friedman and Schwartz, 1991, pp. 39, 49) who wrote in their debate with Hendry and Ericsson (1991) that one should:

> [E]xamine a wide variety of evidence quantitative and nonquantitative ... ; test results from one body of evidence on the other bodies, using econometric techniques as one tool in this process, and build up a collection of simple hypotheses. ... [R]egression analysis is a good tool for deriving hypotheses. But any hypothesis must be tested with data or nonquantitative evidence other than that used in deriving the regression, or available when the regression

> was derived. Low standard errors of estimate, high t values and the like are often tributes to the ingenuity and tenacity of the statistician rather than reliable evidence. ...

A good illustration of this approach is a paper in which Friedman (2005) tried to confirm the quantity theory by comparing changes in the growth rate of money and subsequent recessions in the U.S. in the 1920s and 1990s and in Japan in the 1980s. He has only three observations, so obviously he used no econometrics. Nonetheless, I found it persuasive - not as conclusive evidence, but as circumstantial evidence, because his findings fit in with much other evidence. By contrast, an adherent of the LSE approach would presumably find it unconvincing.

The second, and deeper way of viewing the disagreement is to treat it as dispute about the criterion to be applied to economics. Should one require of economics rigor close to that of mathematics and of physics, with the latter's (alleged) reliance on crucial experiments, and hence be hard-nosed about meeting econometric criteria, or should one consider this as a generally unattainable goal, and settle for more amorphous but extensive circumstantial evidence? Thus in an unjustly neglected book Benjamin Ward (1972) argued that economics should model itself more on law, with its emphasis on circumstantial evidence, than on physics. There are problems with both extremes. The rigorous approach requires us to abandon suggestive evidence even when nothing better is available. The other may degenerate into journalism.

A related issue in the interaction of theory and data is whether data are to be used only to test models, or also to inspire them. Thus Arnold Zellner (1992) advocates searching for "ugly facts", that is puzzling phenomena that cry out for explanation. This fits in with Friedman and Schwartz's just-cited suggestion of looking at many different types of observations rather than analyzing just one particular data set. And it seems to have played an important role in Friedman's own work on the permanent income theory, thus demonstrating its fruitfulness.

A third issue is the choice between simple, or more precisely what Zellner (1992) calls "sophisticatedly simple", models and complex models. Although economists in evaluating their own and their colleague's work seem to adhere to a labor theory of value, several econometricians have warned against automatically assuming that a complex model is more useful and predicts better than a simple model. (See Keuzenkamp and McAleer, 1995, Kennedy, 2002, Makridakis and Hibon, 2000, Zellner, 1995.)

2. Ceteris Paribus Conditions and Testability

Another serious problem both in testing and in applying a model is that the ceteris paribus conditions that define its domain are often insufficiently specified. If we are not told what they are, and the extent to which they can be relaxed without significant damage to the model's applicability, then data cannot be said to refute it, but only to constrain its domain. In day-to-day work this shows up as the question of what variables have to be included among the auxiliary regressors. A dramatic illustration is Edward Leamer's (1978) tabulation of the results obtained when one includes various plausible auxiliary regressors in equations intended to measure the effect of capital punishment on the homicide rate. The results are all over the map. And the same is true in a recent follow-up study (Donohue and Wolfers, 2006). Similarly, as Thomas Cooley and Stephen LeRoy (1981) have shown, in demand functions for money, the negative interest elasticity predicted by theory does not emerge clearly from the data, but depends on what other regressors are used.

The ideal solution would be to specify the ceteris paribus conditions of the theoretical model so precisely that it would not leave any choice about what auxiliary regressors to include. But we cannot list <u>all</u> the ceteris paribus conditions. New classical theorists claim to have a solution: the selection of auxiliary regressors must be founded on rational-choice theory. But that is unpersuasive. In their empirical work the new classicals substitute for utility either income, or both income and leisure variables, plus perhaps a risk-aversion variable. But behavioral and experimental economics, as well as neuroscience, provide much evidence that there is more to utility than that. And the well documented bounds on rationality open the door

to all sorts of additional variables that are not in the new classical's utility function. Similarly, market imperfections complicate a firms' decisions.

If theory cannot constrain sufficiently the variables that have to be held paribus by the inclusion of regressors for them one possible solution could be to open the floodgates, allow all sorts of plausible variables in, and call the model confirmed only if it works regardless of which auxiliary regressors are included. In this spirit Edward Leamer (1978, 1983) has advocated "extreme bounds analysis", that is, deciding what regressors are plausible, running regressions with various combinations of them, and then treating as confirmed only those hypotheses that survive all of these tests. This procedure has been criticized on technical grounds (see McAleer et al, 1983; Hoover and Perez, 2000). It also has the practical disadvantage that it allows very few hypotheses to survive. Since if economists refuse to answer policy questions they leave more space for the answers of those who know even less, it is doubtful that they should become the Trappist monks that extreme bounds analysis would require of them. However, it may be possible to ameliorate this problem by adopting a, say 15 percent significance level instead of the 5 percent level.

Full-scale extreme bounds analysis has found few adherents. Instead, economists now often employ an informal and limited version by reporting as robustness tests, in addition to their preferred regressions, also the results of several alternative regressions that use different auxiliary regressors or empirical definitions of the theoretical variables.[7] This can be interpreted along Duhem-Quinian lines as showing that the validity of the maintained hypothesis does not depend on the validity of certain specific auxiliary hypotheses. While this is a great improvement over reporting just the results of the favored regression it is not clear that economists test - and report on - a sufficient number of regressors and definitions. Indeed, that is not likely because data mining creates an incentives- incompatibility problem between authors (agents) and readers (principals).

3. Data Mining

Usually, by the time she runs her regressions a researcher has already spent much effort on the project. Hence, if her initial regressions fail to confirm her hypothesis she has a strong incentive to try other regressions, perhaps with differently defined variables, different functional forms, different sample periods, different auxiliary variables, or different techniques, and to do so until she obtains favorable results. Such pre-testing makes the t values of the final regression – as traditionally calculated - worthless.[8] Just as bad, if not worse, such biased data mining also means that the final results "confirm" the hypothesis only in the sense of showing that it is not necessarily inconsistent with the data, that there are some decisions about auxiliary regressors, etc., that could save the hypothesis. Suppose a researcher has run, say ten alternative regressions, three of which support his hypothesis and seven that do not. He will be tempted to present one of his successful regressions as his main one and mention the other two successful ones as robustness tests, while ignoring the seven regressions that did not support his hypothesis.[9]

Data mining can occur not only in conventional econometric tests, but also in calibrations, where there may be many diverse microeconomic estimates among which the calibrator can pick and choose. (Cf. Hansen and Heckman, 1996). To be convincing a calibration test requires making a compelling case for the particular estimates of the coefficients that has been picked out of the often quite diverse ones in the literature, not just giving a reference to the coefficient found in a particular paper.

Though much practiced (see Backhouse and Morgan, 2000) data mining, is widely deplored (see for instance Leamer 1983; Cooley and LeRoy 1981). But it has its defenders. Thus Adrian Pagan and Michael Veall (2000) argue that since economists seem willing to accept the output of data miners they cannot be all that concerned about it. But what choice do they have? They do no know what papers have been hyped by biased data mining, and being academic economists they have to read the journals and refer to them. Pagan and Veall also argue that data mining does little damage because if a paper's results seem important but are

not robust, it will be replicated and its fragility will be exposed. But while path-breaking papers are likely to be replicated, by no means all unreplicated papers are unimportant; much scientific progress results from normal science. And even when papers are replicated time passes until the erroneous ones are spotted, and in the meantime they shunt researchers onto the wrong track.

A much more persuasive defense of data mining is the need to obtain as much information as possible from the data, so that the learning that results from trying many regressions needs to coexist with the task of testing the model. (See Greene, 2000; Spanos, 2000). Thus Hoover and Peres (2000), who focus on generating accurate values for the coefficients of a hypothesis rather than on testing it, argue (mainly in the context of general-to-specific modeling) that we need to try many specifications to find the best one, while Keuzenkamp and McAleer (1995, p. 20) write: "specification freedom is a nuisance to purists, but is an indispensable aid to practical econometricians. (See also Backhouse and Morgan, 2000

Testing, Hoover and Perez argue, should be done in some other way, thus separating the task of exploring a data set from the task of drawing inferences from it. That would be the ideal solution – as Magnus (1999) points out in traditional mathematical statistics books estimation and testing are treated as different topics. But in macroeconomics such multiple independent data sets are generally not available, or if they are they relate to different countries which may complicate research . In much microeconomic work with sample surveys or experimental data, it is, in principle, possible to divide the sample into two, and to use one to formulate and the other to test the hypothesis. But in practice, funds are often too limited for that. Suppose, for example, that your budget allows you to draw a sample of a 1000 responses. Would you feel comfortable using only 500 responses to estimate the coefficients when another 500 are sitting on your desk? Moreover, a researcher who has two samples can mine surreptitious by peeking at the second sample when estimating the coefficients from the first sample.[10] (See Inoue and Killian, 2002.)

The other polar position on data mining - one usually not stated so starkly but implicit in much criticism of data mining - is to limit each researcher to testing only a single variant of her model. But that is a bad rule, not only for the reasons just mentioned, but also because it leaves too much to luck. A researcher might just happen on the first try to pick the only variant of twenty equally plausible ones that provides a good fit (See Bronfenbrenner, 1972.) Moreover, even if all data mining by individual researchers were eliminated, it would not put a stop to the harmful effects of data mining because of a publication bias. Only those papers that come up with acceptable t values and other regression diagnostics tend to be published, so that, at least in the short run, there would still be a bias in favor of the hypothesis.[11] Moreover, it is hard to imagine such a rule of one regression per researcher being effectively enforced.

A more feasible solution that avoids both extremes is to permit data mining, but only as long as it is done transparently. A basic idea underlying the organization of research is the division of labor; instead of having every scientist investigate a particular problem, one scientist does so, and her discoveries become known to the others. This does not work well if she withholds information that detracts from the validity of her work, for example, that her results require the assumption that the lag is six months rather than three, nine or twelve months. Hence, a data miner should let readers know if plausible assumptions other than the ones she used yield results that are meaningfully different.

Though I think this is the best of all available alternatives, it, too, has its problems. One is the difficulty (impossibility?) of ensuring that researchers mention all their alternative regressions that significantly reduce the credibility of her maintained hypothesis. Your conscience may urge you to do so, but fear that your rivals do not, urges you to override your conscience. A second is that a researcher is likely to run some regressions that she does not take seriously, just to see what would happen if... . Do they have to be reported? And if not, where does one draw the line? Another problem is that a researcher who intends to run,

say twelve variants of the maintained hypothesis, and happens to get a good result in say the first two, has a strong incentive to quit while she is ahead, so that potential knowledge is lost.

In macroeconomics another way of ameliorating the effects of data mining would be to require an author to publish, perhaps three to five years after the appearance of his paper, a follow-up note on how well his model fits the subsequent data. (See Greene, 2000). This is preferable to asking him to hold out the last few year's data when fitting his model, because of the danger that he may be influenced, either consciously or unconsciously, by what he knows happened in these last few years. Moreover, it would provide only a small sample. Besides, policymakers may want to know how well the model performed during the last five years. All in all, there is no perfect solution to the problem of biased data mining, but requiring transparency seems a reasonable compromise.

V. Significance Tests

Most economists seem to view significance tests as a standard accoutrement of a "scientific" paper. They might be surprised that in psychology their use has come in for much criticism, and that there was even an unsuccessful attempt to ban them in journals published by the American Psychological Association.[12]

(i)

Within economics D. McCloskey (1985, Chapter 9) has argued that significance tests are useless because what matters is the magnitude of a coefficient, its "policy oomph" as she calls it, and not its t value, which depends on sample size. Given a large enough sample even a substantively trivial coefficient can be statistically significant without being substantively significant. McCloskey is right in stressing that one should usually look at the size of a coefficient. But that does not mean that significance tests are unimportant. A researcher usually has to clear two hurdles. She must show that her results are substantial enough to be interesting, and that they are unlikely to be just due to sampling error. And in some special cases even a statistically significant but substantively trivial coefficient may be highly relevant if we are choosing between

two theories that have sharply different and tight implications on this point; for example the slight bending of light that supports relativity theory against Newtonian theory. (Elliot and Granger, 2004; Horowitz, 2004)[13]

Although the distinction between substantive and statistical significance seems obvious Steven Ziliak and Deidre McCloskey (2004) claim that it is widely ignored. But while it is confused in many cases (see Elliot and Granger, 2004, and Thorbecke 2004), their claim that this is the common practice is questionable.[14] Kevin Hoover and Mark Siegler (unpublished) have re-examined Ziliak and McCloskey's data, and conclude that this confusion is not widespread. But even if it occurs only some of the time, that is too much. A good heuristic is to think of a significance test, not as a test of the hypothesis, but as a test of the adequacy of the sample. A problem may also result from the interaction of statistical and substantive significance. An economist may first check for statistical significance, and having reassured himself about that, check for substantive significance, and make a confident statement that the coefficient - by which he means its point estimate - is substantively as well as statistically significant. But the confidence intervals should also be checked for substantive significance. If a test of the law of one price finds that the difference between two prices is both statistically significant and substantively large, there is still not a strong case against the law of one price if the lower confidence interval, though it does not include zero, does include a substantively insignificant value.

(ii)

I now turn to a problem that is less frequently acknowledged (but see Darnell 1997) and therefore needs more discussion. This is the confusion of "not confirmed" with "disconfirmed", a confusion that sometimes shows up in econometric practice, even though the distinction is well known in the abstract.[15] Imagine first an ideal world in which the dependent variable is explained entirely by a few independent variables, all data are measured without error, and the sample encompasses the universe, so that if the correct model is used the standard error of the regression is zero. In this world if a hypothesis implies that the regression coefficient of

x is zero, and it is not so in the data, we can say that the hypothesis has been disconfirmed. But what happens in a stochastic model? Suppose the estimated coefficient is 1.0 with a standard error of 0.25, so that its t value is 4 and the hypothesis is rejected. So far no problem. But now suppose that the standard error is greater, so that the hypothesis that the true value of the coefficient is zero is significant only at the 20 percent level. Then, the usual procedure is to say that the hypothesis has not been disconfirmed. And while this may be stated cautiously as "the data do not reject the hypothesis," often the clear implication is that the test confirmed the hypothesis in the following way: the data were given a chance to reject it, but did not do so. And the more often a hypothesis survives a potentially disconfirming test, the more credible it is. But in the case just described does this make sense? In repeated sampling in only one fifth of the runs would random errors generate that large a discrepancy between the actual and predicted values. And that should count as potential evidence against, not for, the hypothesis.[16]

A related problem arises if a hypothesis is tested more than once. Suppose that on the first test an estimated coefficient that according to the hypothesis should be zero is positive with a t value of 1.7. Suppose further that on a second test using an independent data set it is again positive with a t value of 1.6, and on a third test it is positive with a t value of 1.5. If failure to reject at the 5 percent level is interpreted as confirmation, then the second and third tests must be treated as strengthening the plausibility of the hypothesis, since three tests have now failed to reject it. But the correct message of the second and third tests is just the opposite. The probability of three successive sampling errors that large and with the same sign is so low that the hypothesis should be rejected.

These problems arise from our unsurprising eagerness to have significance tests do more than they are capable of. We want them to classify hypotheses as either confirmed or disconfirmed. But all they can do is tell us the probability that the observed result is just due to sampling error or other noise in the data. We then add the rule of thumb that when the probability is less than 5 percent that the observed error is just a

sampling or noise error, we refuse to accept the hypothesis. But there is a wide gap between refusing to accept the hypothesis, and accepting the proposition that it is false. In many cases the correct conclusion is neither to accept nor to reject it, but to suspend judgment. Yet this point is sometimes missed. For example, if the cross-equation restrictions of a model cannot be rejected at the 5 percent level, we act as though they have been satisfied, even if they can be rejected at, at say the 12 percent level. The 5 percent criterion was intended to be a tough taskmaster, but all too often has become a progressive educator.

This raises a difficult problem. Suppose we subject a hypothesis to a tough test, tough in the sense that it tests an implication that is rigorously derived from the hypothesis, and as far as we can tell cannot also be deduced from some other reasonable hypothesis, (See Kim, de Marchi and Morgan, 1995). Suppose that on this test the t value of the difference between the predicted and the estimated coefficient is less than, say 0.1. Since it seems unlikely that we got such a small t just by chance, it is reasonable to say that the data support the hypothesis. On the other hand, if the t value is 1.5, then the probability that the difference is due to sampling error is low. If we do not have a null hypothesis that tells us what t value to expect if the hypothesis is false, we cannot confidentially say whether to treat the 1.5 t value as enhancing or as reducing the credibility of the hypothesis. We have to rely on our subjective judgment - precisely the situation that we, though not the originators of significance tests (See Gigerenzer, 2004), sought to avoid. (Cf. Darnell, 1007.)

(iii)

Another criticism of significance tests is that despite their prevalence they have had little influence. Hugo Keuzenkamp and Jan Magnus (1995) have offered a prize to anyone finding an example of a significance test that changed economist's thinking about some proposition. So far at least, this prize has not been successful claimed. However, the many papers that have sunk some propositions, such as the total interest inelasticity of the demand for money, would probably not have been taken seriously, or even been published, if the relevant coefficients had not been statistically significant. But since statistical significance was just a supportive point

in their argument they do not qualify as examples with which to claim the Keuzenkamp-Magnus prize. Moreover, the requirements for the prize are also hard to meet because one of them is that "the particular test has been persuasive to others" (Keuzenkamp and Magnus, 1955, p. 21). But while we can observe changes in the opinions of our colleagues, it is much harder to determine why they changed their minds. Moreover, important propositions are often sunk not on by a single hit, but by unrelenting bombardment from many tests. (Cf. Hoover and Siegler, 2005.)

V. In Conclusion: All is Not Bleak

This discussion may seem to have struck an unrelieved pessimistic note. But all attempts to advance knowledge, not just economic measurements, face obstacles. For example, economic theory has its unrealistic assumption (and implication) of rational income maximization. All the same, it has greatly advanced our understanding. Moreover, the large volume of economic modeling over the last few decades has improved our understanding of the economy and our predictive ability, think, for example, of asymmetric information theory, modern finance theory and behavioral economics.

And other fields have their problems too. In medicine a study found that:

> 16 percent of the top cited clinical research articles on postulated effective medical interventions that have been published within the last 15 years have been contradicted by subsequent clinical studies, and another 16 percent have been found to have initially stronger effects than subsequent research." (Ioannidis, 2005, p. 223)

Other studies of medical statistics, see for instance, James Mills (1993) and An-Wen Chen et al (2004), have complained about data mining and biased reporting of results. And yet medical knowledge has increased. And so has our knowledge of natural science, even though Emili Garcia-Berthou (2004) found that in 11.6 percent of a sample of papers published in *Nature* the reported test statistics, degrees of freedom and p values were inconsistent.

The preceding tale of woe is therefore not a plea for giving up, but instead an argument for modesty in the claims we make. Our papers seem to suggest that there is at least a 95 percent probability that our conclusions are correct. Such a claim is both indefensible and unneeded. If an economist takes an important proposition for which the previous evidence suggested a 50:50 probability and shows that it has a 55:45 probability of being right, she has done a useful job. It is also a plea to improve our work by paying more attention to such mundane matters as the quality and meaning of our data, potential computing errors, and the need to at least mention unfavorable as well as favorable results of robustness tests. To be sure, that would still leave some very serious problems, such as the transition from correlation to causation and the limited availability of reliable data, but that there is some opportunity for improvement is a hopeful message. Moreover, that some problems we face are insoluble should make economists feel good about themselves, since it suggests that their failure to match the achievements of most natural sciences is not an indication of intellectual inferiority.[17]

ENDNOTES

1. Caricature model can, however, be dangerous if used for prediction. The classic example is David Ricardo's prediction that rents would rise enough to keep wages at the subsistence level, which lead Joseph Schumpeter to warn against the "Ricardian Vice." In our own time while early new classical theory with its assumption of complete price flexibility may have provided insight, it would have been a disastrous basis for policy.

2. Elsewhere (Mayer 2001b), I have argued that ideological differences do not explain very much of the disagreement among economists. (For a contrary conclusion see Fuchs, Kruger and Poterba, 1998). In (Mayer, 1998) I have presented a case study of how adherence to schools of thought and similar obstacles have inhibited the debate about a fixed monetary growth rate rule.

3. Previously Oskar Morgenstern (1963) had provided a long list of errors that resulted from economists' not knowing enough about their data, and Andrew Kamarck (1963) has presented more recent examples. The appearance of downloadable databases probably exacerbated this problem. In the old days when economists had to take the data from the original sources they were more likely to read the accompanying description of the data. Another exacerbating factor is the much greater use of research assistants. A researcher who has to work with the data herself is more likely to notice anomalies in the data than are assistants who tend to follow instructions rather than "waste" time by thinkiFng about the data.

4. More precisely, "personal outlays for personal consumption expenditures, for interest payments on consumer debt, and for current transfer payments are subtracted from disposable personal income." (Reinsdorf, 2004, p. 18.) The extent to which errors in estimating either income or consumption affect estimates of the savings ratio depends not only on the size of these errors, but also on their covariance. Suppose income is actually 100, but is estimated to be 101, while consumption is estimated correctly at 95. Then, saving is estimated to be 6 rather than 5, a 20 percent error. But if income has been overestimated by 1 because consumption was overestimated by 1, then these errors lower the estimated savings ratio only by 0.05 percent of income, that is by 1 percent of its actual value.

5. Over many years of working first with desk calculators and then with a PC I have found that even if one checks the data carefully, in any large project mechanical errors do creep in. Calculation errors may be as common, or even more common, now than they were in the days of desk calculators. One is more likely to be dividing when one should be multiplying, if one can do so with a single key stroke, than in the old days when in the tedious hours of using a desk calculator one had plenty of time to think about what one was doing.

6. I did the search on November 4, 2005 using the Google "scholar" option. It is, of course, possible, though unlikely, that some errata were published that did not cite the McCullough-Vinod paper. It is also possible that when working on subsequent papers some economists did check whether other programs gave results similar to the one they used, though I do not recall ever seeing any indication of this. Also, some economists may have tried several programs and abandoned their project when they found that these programs gave substantially different results. It would be interesting to know whether economists in government or business, whose errors could result in large losses, recalculated some of their regressions using different programs.

7. I have the impression that this has become much more common in recent years.

8. For the correct procedure of calculating t values in the presence of pre-testing see Jan Magnus and James Durbin (1999) and Dimitry Danilov and Jan Magnus (2004a and 2004b).

9. It is often far from obvious whether the results of additional regressions confirm or disconfirm the maintained hypothesis. Suppose, that this hypothesis implies that the coefficient of x is positive. Suppose further that it is positive and significant in the main regression. But in additional regressions that include certain other auxiliary regressors, though again positive, it is significant only at the 20 percent level. Although taken in isolation these additional regressions would usually be read as failures to confirm, they should perhaps be read as enhancing the credibility of the maintained hypothesis, because they suggest that even if the auxiliary hypothesis that these regressors do not belong in the regression is invalid, there is still only a relatively small likelihood that the observed results are due merely to sampling error. Good theory choice takes more than attention to t values.

10. This is not necessarily dishonest. If a macroeconomist sets a few year's data aside as a second sample, she usually knows something about what the data are likely to show simply by having lived through this period, And someone working with survey data may have inadvertently learned something about the second sample in the process of splitting the data or in talking to her research assistant.

Bronfenbrenner, Martin (1972) "Sensitivity Analysis for Econometricians," *Nebraska Journal of Economics* 2, 57-66.

Chan, An-Wen, Hrøbjartsson, Asbjørn, Haar, Mette, Gøtzsche, Peter, Altman, Douglas (2004) "Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials," *Journal of the American Medical Association*, 291, 2457-65.

Chatfield, Christopher (1991) "Avoiding Statistical Pitfalls," *Statistical Science,* 6, August, 240-52.
Chow, Sui (1996) *Statistical Significance*, London, Sage Publishing.

Cooley, Thomas and LeRoy Stephen (1981) "Identification and Estimation of Money Demand," *American Economic Review*, 71, 825-43.

Danilov, Dimitry and Magnus, Jan (2004a) "Forecast Accuracy after Pretesting with an Application to the Stock Market," *Journal of Forecasting*, 23, 251-74.

Danilov, Dimitry and Magnus Jan (2004b) "On the Harm that Pretesting can Cause," *Journal of Econometrics*, 122, 27-46

Darnell, Adrian (1997) "Imprecise tests and Imprecise Hypotheses", *Scottish Journal of Political Economy,* 44, 247-68.

Dewald, William, Thursby, Jerry and Anderson, Richard (1986) "Replication in Economics: *The Journal of Money, Credit and Banking Project", American Economic Review*, (76) 587-603.
Donohue, John, III and Wolfers, Justin (2006) "Uses and Abuses of Empirical Evidence in the Death Penalty Debate," NBER Working Paper 11982.

Elliot, Graham and Granger, Clive (2004) "Evaluating Significance: Comment or 'Size Matters'," *Journal of Socio-Economics*, 33, 547-50

Fuchs, Victor, Kruger, Alan and Poterba, Joseph (1998) "Economists' Views about Parameters, Values and Policies: Survey Results in Labor and Public Economics," *Journal of Economic Literature,* 36, 1387-1425,

Friedman, Milton (2005) "A Natural Experiment in Monetary Policy Covering Three Periods of Growth and Decline in the Economy and the Stock Market," *Journal of Economic Perspectives,* 19, 145-50.

Friedman, Milton and Schwartz, Anna (1991) "Alternative Approaches to Analyzing Economic Data," *American Economic Review*, 81 39-49.

Garcia-Berthou, Emili and Alcarez, Charles (2004) "Incongruence between Test Statistics and p Values in Medical Papers," *Medical Research Methodology,* 4, http://www.biomedcentral.com/147-2289-4-13.

11. But only in the short run; as Robert Goldfarb (1995) has shown, once a significant hypothesis is widely accepted only those tests that disconfirm it tend to be published, because only they provide new information.
12. See for instance, Bruce Thompson (2004); Open Peer Comment (1996). Sui Chow (1996 p. 11) who, even though he defended the use of significance tests, wrote: "the overall assessment of the ... [null-hypotheses significance test procedure] in psychology is not encouraging. The puzzle is why so many social scientists persist in using the process." He argued persuasively that these criticisms of significance tests are largely due to researchers trying to read too much into them. And Magnus (1999) points out that what matters in constructing a model in not whether a variable that is of no interest per se is significant or not, but whether its inclusion improves the estimation of the variables that are of interest.

13. McCloskey (1985) also argues that in many cases the sample is in effect the whole universe, so that tests for sampling error are meaningless. Hoover and Perez's (2005) response is that the hypothesis being tested is intended to be general and thus cover actual or potential observations outside the sample period. However, in economic history, some hypotheses do relate to only a limited period.
14. The confusion of statistical and substantive significance has also been a problem in biology (see Pfannkuch and Wild, 2000).

15. For some specific instances see Robertson (1999); Viscusi and Hamilton, (1999); Loeb and Page (2000); McConnell and Perez-Quiros (2000); Papell el al (2000); Wei (2000). For a further discussion of this problem see Mayer (2001b).

16. All of this is entirely consistent with the proposition in philosophy of science that failure to be disconfirmed on a hard test raises the credibility of a hypothesis, because the term "not disconfirmed" is used in two different senses. In the context of significance testing it means that we cannot be certain that the hypothesis is false because there is an at least 5 percent probability that the divergence between its prediction and our data is due merely to sampling error. In the context of philosophy-of- science failure to be disconfirmed means that the probability that the proposition is false is less than 50 percent.

17. Alexander Rosenberg, a philosopher of science who specializes in the philosophies of economics and biology, describes economics as "a subject on which at least as much sheer genius has been lavished as on most natural sciences." (Rosenberg, 1978, p. 685, italics in original.) That is flattering, but not entirely convincing.

ACKNOWLEDGEMENTS

I am indebted for helpful comments to Jan Magnus and Marshall .
Reinsdorf.

REFERENCES

Backhouse, Roger (1992) "The Significance of Replication in Econometrics," Discussion Paper 92-25, Economics Department, University of Birmingham.

Backhouse, Roger and Morgan, Mary (2000) "Introduction: Is Data Mining a Methodological Problem?", *Journal of Economic Methodology*, 7, 173-82.

Gibbard, Allen and Varian, Hal (1978) "Economic Models," *Journal of Philosophy*, 1975, 665-77.

Gigerenzer, Gerd (2004) "Mindless Statistics," *Journal of Socio-Economics,* 33, 587-606.

Goldfarb, Robert (1995) "The Economist-as-Audience Needs a Plausible Model of Inference," *Journal of Economic Methodology*, 2, 201-22.

Goldfarb, Robert and Stekler H. O. (2000) "Why Do Empirical Results Change? Forecasts as Tests of Rational Expectations," *History of Political Economy, Annual Supplement*, 95-116.

Greene, Clinton (2000). "I am not, nor have I have been a Member of the Data-Mining Discipline," *Journal of Economic Methodology*, 7, 217-239

Hamermesh (2000) "The Craft of Labrometrics," *Industrial and Labor Relations Review*, 53, 363-80.

Hansen, Lars P. and Heckman, James (1996) "The Empirical Foundations of Calibration," *Journal of Economic Perspectives.* 10, 87-104.

Hausman, Daniel (1992) *The Inexact and Separate Science of Economics,* Cambridge, Cambridge university Press.

Hendry, David and Ericsson, Neil (1991) "An Econometric Ana*lysis of U.K. Money Demand in Monetary Trends in the United States and the United Kingdom* by Milton Friedman and Anna J, Schwartz," *American Economic Review*, 81, 8-39.

Hoover. Kevin and Perez, Stephen (2000) "Three Attitudes towards Data Mining, *Journal of Economic Methodology*, 7, 195-210.

Hoover, Kevin and Siegler, Mark (unpublished) "Sound and Fury: McCloskey and Significance Testing in Economics."

Horowitz, Joel (2004) "Comment on 'Size Matters," *Journal of Socio- Economics*, 33, 571-75.
Inoue, Atsushi and Killian, Lutz (2002) "In-sample or Out-of-sample Tests of Predictability: Which should we Use?", European Central Bank, Working Paper No. 195.

Ioannidis, John (2005) "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *Journal of the American Medical Association*, 294, 219-27.

Kamarck, Andrew, (1983) *Economics and the Real World*, Oxford, Blackwell.

Kennedy, Peter (2002) "Sinning in the Basement: What are the Rules? The Ten Commandments of Applied Econometrics," *Journal of Economic Surveys*, 16, # 4 569-85.

Keuzenkamp, Hugo and Magnus, Jan (1995) "On Tests and Significance in Econometrics," *Journal of Econometrics*, 67, 5-24.

Keuzenkamp, Hugo and McAleer, Michael (1995) "Simplicity, Scientific Inference and Econometric Modelling," *Economic Journal*, 1-21.

Kim, Jimbang, de Marchi, Neil and Morgan, Mary (1995), "Empirical Model Peculiarities and Belief in the Natural Rate Hypothesis," *Journal of Econometrics,* 67, 81-102.

Leamer, Edward (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, New York, John Wiley.

Leamer, Edward (1983) "Let's take the Con out of Econometrics," *American Economic Review,* 73, 31-43.

Leontief, Wassily (1971) "Theoretical Assumptions and Nonobervable Facts," *American Economic Review* 61,1-7.

Loeb, Suzanna and Page, Marianne (2000) "Examining of the Link between Teacher Wages and Student Outcome," (2000) *Review of Economics and Statistics,* 82, 393-408

Lovell, Michael (1994) and Selover, David, "Software Reviews", *Economic Journal.* 104, 713-26.
Machlup, Fritz (1950) *Methodology of Economics and Other Social Sciences*, New York, Academic Press.

Magnus, Jan (1999) "The Success of Econometrics," *De Economist*, 147, 55-71.

Magnus, Jan and Durbin, James (1999) "Estimation of Regression Coefficients of Interest when Other Regression Coefficients are of no Interest," *Econometrica,* 67, 639-43.

Mayer, Thomas (1998) "Monetarists versus Keynesians on Central Banking," in R. Backhouse, Roger, Hausman, Daniel, Mäki, Uskali and Salanti, Andrea (eds.) *Economics and Methodology*, London. MacMillan.

Mayer, Thomas (2001a) "Misinterpreting a Failure to Disconfirm as a Confirmation" www.econ.ucdavis.edu/

Mayer, Thomas (2001b) "The Role of Ideology in Disagreements among Economists: A Quantities Analysis," *Journal of Economic Methodology*, 8, 253-74.

McAleer, Michael, Pagan, Adrian and Volker, Paul (1983) "What will take the Con out of Econometrics?", *American Economic Review,* 73, 293-307

McCloskey, D. (1985) *The Rhetoric of Economics*, Madison, University of Wisconsin Press.

McConnell Margaret and Perez-Quiros, Gabriel (2000) "Output Fluctuations in the United States: What has Changed since the early 1980s," *American Economic Review*, 90,1464-76.
McCullough, B. D. (2000) "Is it Safe to Assume that Software is Accurate? *International Journal of Forecasting* 16, 349-57.

McCullough, B.D. and Vinod, H.D. (1999) "The Numerical Reliability of Econometric Software," *Journal of Economic Literature*, 37, 633-65.

Mills, James (1993) "Data Torturing," *New England Journal of Medicine*, 329, 1196-99.

Mirowski, Philip and Skilivas, Steven (1991) "Why Econometricians don't Replicate although they do Reproduce". *Review of Political Economy*, 3, 2, 146-63.

Morgenstern, Oskar (1963) *On the Accuracy of Economic Observations*, Princeton, Princeton University Press.

"Open Peer Comments" (1996) *Brain and Behavioral Research,* 19, 188-228

Pagan, Adrian, and Veall, Michael (2000) "Data Mining and the Econometrics Industry: Comments on the Papers by Mayer and Hoover and Perez", *Journal of Economic Methodology,* 7, 211-16.

Papell, David, Murray, Christian and Ghiblawi, Hala (2000) "The Structure of Unemployment,"*Review of Economics and Statistics,* 82, 309-15.

Pfannkuch, Maxine and Wild, Chris (2000) "Statistical Thinking and Statistical Practice: Themes Gleaned from Professional Statisticians", *Statistical Science*, 15, # 2, 132-52,

Reinsdorf, Marshall (2004) "Alternative Measures of Personal Saving," *Survey of Current Business*, v. 84, 17-24

Robertson, Raymond (2000) "Wage Shocks and North American Labor- Market Integration," *American Economic Review*, 9, 742-64.

Rosenberg, Alexander (1978) "The Puzzle of Economic Modeling," *Journal of Philosophy*, 1975, November, 679-83.

Spanos, Aris (1986) *Statistical Foundations of Econometric Modeling*, Cambridge, Cambridge University Press.

Spanos, Aris (2000) "Revisiting Data Mining: 'Hunting' with or without a License,", *Journal of Economic Methodology*, 7, 231-64.

Thompson, Bruce (2004) "The 'Significance' Crisis in Psychology and Education." *Journal of Socio-Economics*, 33, 607-13.

Thorbecke, Erik (2004) "Economic and Statistical Significance," Comments on 'Size Matters'." *Journal of Socio-Economics*, 33, 571-75.

Viscusi. W. K. and Hamilton J. K. (1999) "Are Risk Regulators Rational? Evidence from Hazardous Waste Cleanup Decisions," *American Economic Review*, 89, 210-27.

Ward, Benjamin (1972) *What's Wrong with Economics*, New York, Basic Books,

Wei, Sang-Jin (2000) "How Taxing is Corruption on International Investment?" *Review of Economics and Statistics*, 82, 1-11.

Zellner, Arnold (1992) "Statistics, Science and Public Policy," *Journal of the American Statistical Association,* 87, 1-6

Ziliak Steven and McCloskey, Deidre (2004) "Size Matters: The Standard Error of Regressions in the *American Economic Review*", *Journal of Socio-Economics*, 33, 527-46