American Economic Review

May 1985 (vol. 75, #2)

179.

The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests

By DONALD N. MCCLOSKEY*

Roughly three-quarters of the contributors to the *American Economic Review* misuse the test of significance. They use it to persuade themselves that a variable is important. But the test can only affirm a likelihood of excessive scepticism in the face of errors arising from too small a sample. The test does *not* tell the economist whether a fitted coefficient is large or small in an economically significant sense.

The criticism is distinct from the criticism that in a world of publication-counting deans, there is an incentive to mine the data, giggling uncomfortably when caught. Economists, being professional cycnics, are much amused by data mining and significance fishing (Gordon Tullock, 1959; Edward Ames and Stanley Reiter, 1961; Edgar Feige, 1975; Edward Leamer, 1978; Thomas Mayer, 1980; Michael Lovell, 1983; Frank Denton, 1985). But the point here is that even under classical conditions the t-test is irrelevant much of the time.

Neither criticism is controversial or arcane. Statisticians, psychometricians, sociometricians, econometricians, and other metrical folk have understood them both for 60 years (see Kenneth Arrow, 1960; Zvi Griliches, 1976). Both should be parts of the statistical education of an economist, yet almost none of the texts in econometrics mention them.

I. An Example: Tests of Purchasing Power Parity

The usual test of purchasing power parity (see J. R. Zecher and myself, 1984) fits prices at home (P) to prices abroad (P^*), allowing for the exchange rate (e): $P = \alpha + \beta(eP^*) +$ error term. The equation can be in levels or rates of change. If the coefficient b is statistically significantly different from 1.0, the hypothesis of purchasing power parity is rejected; if not, not. The test seems to tell about substantive significance without any tiresome inquiry into how true a hypothesis must be in order to be true. The table of t will tell.

But a number is large or small relative only to some standard. Forty degrees of frost is paralyzing cold by the standard of Virginia, a normal day by the standard of Saskatoon in January, and a heat wave by the standard of most interstellar gas. A *New Yorker* magazine cartoon shows water faucets labeled "Hot (A Relative Concept)" and "Cold (A Relative Concept)." Nothing is large-in-itself. It is large (or yellow, rich, cold, stable, well-integrated, selfish, free, rising, monopolistic) relative to something with which it can be interestingly compared. The remark "But how large is large?" is one of those seminar standbys, applying to any paper, like "Have you considered simultaneity bias?" or "Are there unexploited opportunities for entry?" It's usually a good question, inheriting some of its excellence from its father in thought, the mind-stunning "So What?" (and its Jewish mother: "So What Else is New?"). You say the coefficient is 0.85 with a standard error of 0.07? So?

The literature does not discuss how near the slope has to be to 1.0 to be able to say that purchasing power parity succeeds or fails. It does not answer how large is large. The only standard offered is statistical significance, that is, how surprising it would be to get the observed sample if the hypothesis of $\beta = 1.0$ were in fact exactly true.

But "exactly" true is not relevant for most economic purposes. What is relevant is merely that β is in the neighborhood of 1.0,

*Departments of Economics and of History, University of Iowa, Iowa City, IA 52242. I thank Leanne Swenson for research assistance, and seminar participants at Iowa, Nebraska, Yale, Chicago, and McMaster for comments. The work was supported by the John Simon Guggenheim Foundation, the Institute for Advanced Study, and the Program in Humanities, Science and Technology of the National Endowment for the Humanities.

where "the neighborhood" is defined by *why* it is relevant—for policy, for academic reputation, for the progress of knowledge. The question requires thought about the loss function. One begins to think that the neighborhood of small loss might be large. And even outside it, one begins to think that $\beta = .10$, say, would still be economically significant, were the fit tight enough to constrain prices at home; or that even a coefficient of -7854.86 would belie closed economy models of inflation.

The usual test does not discuss standards. It gives them up in favor of irrelevant talk about the probability of a type I error in view of the logic of random sampling. Most economists appear to have forgotten how narrow is the question that a statistical test of significance answers. It tells the intrepid investigator how likely it is that, *because of the small size of the sample he has*, he will make a mistake of excessive skepticism in rejecting a true hypothesis (in this case, $\beta = 1.0$). Though not to be scorned, it isn't much. It warns him about a certain narrow kind of foolishness.

The elementary but neglected point is that statistical tests of significance are merely about one sort of unbiased errors in *sampling*. The standard error, after all, is $(s^2/N)^{1/2}$. Except in the limiting case of literally zero correlation, if the sample were large enough all the coefficients would be significantly different from everything. The inverse of the square root of a extremely large number is very small. Any social scientist with large samples has had such logic impressed on him by events. A psychologist, Paul Meehl, for instance, reports a sample of 55,000 Minnesota highschool seniors which "reveal statistically significant relationships in 91 percent of pairwise associations among a congeries of 45 miscellaneous variables such as sex, birth order, religious preference,..., dancing, interest in woodworking.... The majority of variables exhibited significant relationships with *all but three of the others*, often at a very high confidence level" (1967, p. 259).

The large-sample case makes clear the irrelevance of statistical significance to the main question: so what? In the usual test of purchasing power parity, a sample size of a million yielding a very tight estimate that $\beta = 0.999$, "significantly" different from 1.0, could be produced under the usual procedures as evidence that the theory had "failed." Common sense, presumably, would rescue the investigator from asserting that if $\beta = 0.999$, with a standard error of .00000001, we should abandon purchasing power parity, or run our models of the American economy without the world price level. Similar common sense should be applied to findings that $\beta = .80$ or 1.30 with sample sizes of 30. It is not.

The point can be put most sharply by supposing that we *knew* the coefficient to be, say, 0.85. Suppose God told us. God does not play dice with the universe, and His is no mere probabilistic assurance. Would the scientific task be finished? No, it would not. We would still need to decide, by some criterion of why it matters (a human, not a divine, concern), whether 0.85 is high enough to affirm the theory. No mechanical procedure can relieve us of this responsibility. Nor is it a decision that should be made privately, as a matter of "mere opinion." It is the most important scientific decision, and it should be made out in the open. The test of significance doesn't make it.

II. A History of Consciousness

The overuse of statistical significance arises largely from its name. Surely, it insinuates, we serious scientists should be interested first of all in "significant" coefficients: the wise and good would not wish to waste time on trivialities. The appeal is part of the rhetoric of statistics (compare my book, 1985, ch. 2). The British inventors of statistics, as recipients of classical educations, were skillful in naming their ideas. As William Kruskal, a statistician of note, has argued:

> Suppose that Sir R. A. Fisher—a master of public relations—had not taken over from ordinary English such evocative words as "sufficient," "efficient," and "consistent" and made them into precisely defined terms of statistical theory. He might, after all, have used

> utterly dull terms for those properties of estimators, calling them characteristics A, B, and C.... Would his work have had the same smashing influence that it did? I think not, or at least not as rapidly. [1978, p. 98]

As the words spread to less sophisticated research workers, the task of undoing the rhetorical damage commenced. The earliest paper making the point of the present one was written in 1919 by, alarmingly, Edwin Boring. Attacks on the mechanical use of significance became early a commonplace in statistical education. By 1939, for example, a *Statistical Dictionary of Terms and Symbols* of no great intellectual pretensions was putting the point utterly plainly: "A significant difference is not necessarily large, since, in large samples, even a small difference may prove to be a significant difference. Further, the existence of a significant difference may or may not be of practical significance" (A. K. Kurtz and H. A. Edgerton, 1939, article "Significant Difference"). M. G. Kendall and A. Stuart's *Advanced Theory of Statistics* explicitly recognized the mischief in the rhetoric, recommending the phrase "size of the test" in preference to "significance level" (1951, p. 163n); the sociometricians Denton Morrison and Ramon Henkel (whose book *The Significance Test Controversy*, 1970, is the best reading on the subject) suggest that "significance test" be replaced by the less portentous "sample error decision procedure" (p. 198).

In the 1930's, Jerzy Neyman and E. S. Pearson, and then more explicitly Abraham Wald, argued that actual statistical decisions should depend on substantive, not merely statistical, significance. As Wald wrote in 1939:

> The question as to how the form of the weight (i.e., loss) function $W(\theta, \omega)$ should be determined, is not a mathematical or statistical one. The statistician who wants to test certain hypotheses must first determine the relative importance of all possible errors, which will entirely depend on the special purposes of his investigation. [p. 302]

Economists have largely ignored Wald's economical logic, with the result that few textbooks in econometrics mention that the goodness or badness of a hypothesis cannot be decided on merely statistical grounds.

III. The Practice of Economists

It is not easy, then, to justify the use of probabilistic models to answer nonprobabilistic questions. One might retort that good economists do not make such mistakes. But they do, as may be seen from their best practice, in this *Review*. From the fifty full-length papers using regression analysis in the four regular issues of 1981, 1982, and 1983, I took a sample of ten for close scrutiny. Since the purpose is to criticize a socially accepted practice, not to embarrass individual writers, I shall withhold the names here (a larger version of the paper contains them, and a still larger one will examine all fifty).

Of the ten papers, only two do not admit experimenting with the regressions, sometimes with hundreds of different specifications. None propose to alter their levels of significance. Only two of the ten do not use a sign test in conjunction with a significance test: the variable has a statistically significant coefficient *and the right (or expected) sign*. Little statistical theory seems to lie behind the practice, although it seems sensible enough—a beginning, indeed, of looking beyond statistical significance to the size of the coefficient. One of the papers uses a sample of convenience so convenient that it looks like a universe, about which sampling theory can tell nothing: all counties in Alabama, Mississippi, North Carolina, and South Carolina. Four of the ten use true samples, such as the opinions of 6,000 Swedes on the current and expected rate of inflation. The only doubt here is the disproportion of effort in dealing with sampling errors when others are probably more serious. At $N = 6000$ we can surely dismiss Student and attend to bias. As Leamer remarked recently, "when the sampling uncertainty... gets small

compared to the misspecification uncertainty..., it is time to look for other forms of evidence, experiments, or nonexperiments" (1983, p. 33). The other five papers use time-series. One can only ask quietly and pass on: from what universe is a time-series a random sample, and if there is such a universe, is it one we wish to know about?

The most important question is whether the economists in the sample mix up statistical and substantive significance. Even on purely statistical grounds the news is not good: none of the papers mention the word "power," though all mention "significance." Statisticians routinely advise examining the power function, but economists do not follow the advice. Some follow its spirit, avoiding the excessive gullibility of the type II error by treating the machinery of hypothesis testing with a certain reserve. Most do not. Only three of the ten do not jump with abandon from statistical to substantive significance. The very language, though mostly formulaic, sometimes exposes the underlying attitude. One paper slipped into using the phrase "statistically important."

Seven of the papers, then, let statistical significance do the work of substantive significance. Usually this is accomplished by a fallacy of equivocation. The result that is on page 10 (statistically) significant turns up as (economically) significant on p. 20. In the worst cases there is no attempt to show how large the effects are, or whether the statistical tests of the their largeness are powerful, or what standard of largeness one should use. In four of the seven papers with significant errors in the use of significance there is some discussion of how large a coefficient would need to be to be large, but even these let statistical significance do most of the work. And even in the three papers that recognize the distinction and apply it consistently, there is flirtation with intellectual disaster. The siren song of "significance" is a hazard to navigation.

IV. What is to be Done?

If we do not wish to leave science to chance, we must rethink the use of statistical significance in economics. Econometrics courses should teach the relevant decision theory, as judging from results they appear not now to do. It would help if the standard statistical programs did not generate *t*-statistics in such profusion. The programs might be written to ask "Do you really have a probability sample?," "Have you considered power?", and, above all, "By what standard would you judge a fitted coefficient large or small?" Or perhaps they could merely say, printed in bold capitals beside each equation, "So What Else is New?"

REFERENCES

Ames, Edward and Reiter, Stanley, "Distributions of Correlation Coefficients in Economic Time Series," *Journal of the American Statistical Association*, September 1961, *56*, 627–56.

Arrow, Kenneth, "Decision Theory and the Choice of a Level of Significance for the *t*-Test," in Ingram Olkin et al., eds., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford: Stanford University Press, 1960, 70–78.

Boring, Edwin G., "Mathematical Versus Scientific Significance," *Psychological Bulletin*, 1919, *15*, 335–38.

Denton, Frank T., "Data Mining as an Industry," *Review of Economics and Statistics*, forthcoming 1985.

Feige, Edgar, "The Consequences of Journal Editorial Policies and a Suggestion for Revision," *Journal of Political Economy*, December 1975, *83*, 1291–96.

Griliches, Zvi, "Automobile Prices Revisited: Extensions of the Hedonic Hypothesis," in N. E. Terleckyj, ed., *Household Production and Consumption*, NBER, Studies in Income and Wealth, Vol. 40, 1976, 325–90.

Kendall, M. G. and Stuart, A., *Advanced Theory of Statistics*, 3rd ed., Vol. II, London: Griffin, 1951.

Kruskal, William H., "Formulas, Numbers, Words: Statistics in Prose," *The American Scholar*, 1978; reprinted in D. Fiske, ed., *New Directions for Methodology in Social and Behavioral Sciences*, San Francisco: Jossey-Bass, 1981.

Kurtz, A. K. and Edgerton, H. A., *Statistical Dictionary of Terms and Symbols*, New York: Wiley & Sons, 1939.

Leamer, Edward, *Specification Searches: Ad Hoc Inferences with Nonexperimental Data*, New York: Wiley & Sons, 1978.

_____, "Let's Take the Con Out of Econometrics," *American Economics Review* March 1983, *73*, 31–43.

Lovell, Michael C., "Data Mining," *Review of Economics and Statistics*, February 1983, *45*, 1–12.

McCloskey, Donald N., *The Rhetoric of Economics*, Madison: University of Wisconsin Press, 1985.

Mayer, Thomas, "Economics as a Hard Science: Realistic Goal or Wishful Thinking?," *Economic Inquiry*, April 1980, *18*, 165–78.

Meehl, Paul E., "Theory Testing in Psychology and Physics: A Methodological Paradox," *Philosophy of Science*, June 1967, *34*, 103–15, reprinted in Morrison and Henkel, 1970.

Morrison, Denton E. and Henkel, Ramon E. "Significance Tests Reconsidered," *American Sociologist*, May 1969, *4*, 131–40, reprinted in Morrison and Henkel, 1970.

_____ **and** _____, *The Significance Test Controversy — A Reader*, Chicago: Aldine, 1970.

Tullock, Gordon, "Publication Decisions and Tests of Significance: A Comment," *Journal of the American Statistical Association*, September 1959, *54*, 593.

Wald, Abraham, "Contributions to the Theory of Statistical Estimation and Testing Hypotheses," *Annals of Mathematical Statistics*, December 1939, *10*, 299–326.

Zecher, J. R. and McCloskey D. N., "The Success of Purchasing Power Parity," in M. D. Bordo and A. J. Schwartz, eds., *Retrospective on the Classical Gold Standard*, Chicago: University of Chicago Press, 1984.