Newsletter of the Cliometrics Society
Vol. 2, no. 2 (December 1986)

180.

Why Economic Historians Should Stop Relying on Statistical Tests of Significance, and Lead Economists and Historians Into the Promised Land

Donald N. McCloskey
University of Iowa*

I have recently become a nuisance at conferences and in referee reports about statistical significance. The profession deserves an explanation.[1]

I have taken to asking people who use the notion of statistical significance whether they know what they are doing. Quite a few don't.

Step back for a minute and think through what a "significant" coefficient means. It means that the *sampling* problem has been solved, or at any rate solved well enough to satisfy conventional standards. (John W. Tukey has recently given some reasons for doubting the conventional standards: "Sunset Salvo," *The American Statistician*, February 1986, pp. 72-76.) In other words, *the sample is large enough to assure that if you took another sample it would give roughly the same result.* The sampling variance, which is the population's variance divided by the square root of the sample size, has been driven down to some nice, low figure. As John Venn put it in 1888, at a time when our procedures were a mere twinkle in the statistician's eye, the coefficient (or the mean or the difference between two means or the estimated variance or the R-squared or whatever other statistic we are examining) would probably be "permanent." We would probably come up with the same estimate again.

But a permanent coefficient is not necessarily an important coefficient. It could well be that unusually high corn yields in little Iowa would raise the income of the United States a little, and that a proper regression analysis of income on the Iowa corn yields would show this. A large enough sample of years would make the relationship register, and would make it keep on registering in successive samples. (Never mind what "successive samples" of *years* could possibly mean: that's another problem with statistical significance, a philosophical one I'd like to put aside.) The coefficient would be statistically significant. Yet that it registers and would keep on registering does not mean that it is important. "Statistically significant" does not mean "substantively significant."

What matters is oomph. Oomph is what we seek. A variable has oomph when its coefficient is large, its variance high, and its character exogenous. A small coefficient on an endogenous variable that does not move around can be statistically significant, but it is not worth remembering. Oomph is what we mean when we talk about money being "important" for explaining income per person. The Iowa corn yield certainly does affect average national income, but has little oomph because the coefficient is low. Likewise, the existence of oxygen in the atmosphere certainly does affect combustion, but it does not vary enough to give it oomph in an explanation of why the house burned down. The stock of money in the hands of Iowa Citians certainly does determine their expenditures, but because it is entirely endogenous it has no oomph.

Statistical significance, which now guides a large part of the intellectual life of economists, has nothing to do with oomph. It implies, to repeat, that you have acquired some control over sampling error as a source of doubt. Sampling error, though, is seldom the main source

* The author is a Professor of Economics and Professor of History at the University of Iowa.

1 Cf. *The Rhetoric of Economics* (Madison: University of Wisconsin Press, 1986), Chps. 8 and 9, especially 9; and "The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests", *American Economic Review* 75 (May 1985): 201-205.

of doubt. The main source of doubt is whether a variable matters, or whether it matters to such-and-such a degree: what matters is whether foreign prices affected American prices under the gold standard *significantly* (that is, with oomph), or whether American wages affected migration from Europe significantly, or whether social security wealth affected capital accumulation significantly. *Statistical* significance will not reveal this substantive significance, this blessed oomph.

The best way to see the point is to suppose that you really do know what the coefficient is. For sure. God has told you, with no nonsense about confidence intervals; sampling error is zero. The t-statistic is infinite. Well, then: Has the variable got oomph? *You don't yet know.* To find out you have to ask and answer other questions, having nothing to do with statistical significance, such as whether the coefficient is large (how large? Large enough to matter in some conversation of scholars or policy-makers); or whether the variable could vary enough to produce effects you consider important. For most scientific questions the answer that across successive samples having a nice, random character the coefficient would be permanent (or statistically significant) is only mildly interesting.

"Mildly interesting" is not the same as "not interesting at all." Occasionally an economist will have a genuine sample and because of its small size will have a genuine worry about the sampling problem. But mainly our problems have nothing to do with sampling error. They have to do with other statistical problems (bias, for example: see Leamer's "Let's Take the Con Out of Econometrics") or, most commonly, with oomphelimity.

At this point I need to treat some objections:

[The regressor is confident.] "Statistical significance is an approximate test of what you call 'oomph.'"

Educate me. Tell me how the permanence of an estimate over successive samples tells how important the variable is. To be sure, large coefficients will *ceteris paribus* have larger significance. But why not look directly at the size, and ask directly whether it is large enough to matter? Why be approximate and irrelevant when you can be precise and relevant? Why put the coefficient through an irrelevant transformation? The calculation of statistical significance fools people into thinking they've solved the central intellectual problem, namely, how important a variable is. But the calculation can't do it. It must be done by us: we must decide how large is large. Tables of t tell us how large is large *with respect to the permanence in sampling.* (Yet even they do not tell us where to set the null hypothesis for the test; this again is a question of substance, not of statistics.) The test does not tell us how large is large with respect to the economic argument in question.

[He looks worried.] "But statistical significance provides a good initial hurdle for the variables. They should at least be statistically significant. Those that survive can be tested later for oomph."

No. There's no reason to make a necessary hurdle out of merely desirable quality -- the quality, remember, of appearing to be permanent within such-and-such bounds, at least so far as sampling error is the problem, as it usually is not. Doing so would be like choosing academic colleagues "first" on the basis of their geniality. Geniality is a desirable quality, Lord knows, but not so desirable that it should head a list of lexicographically ordered "priorities." The procedure would make it impossible to hire a brilliant woman with a slightly sub-par amount of geniality. Anyway, for all the talk of "priorities" in public discourse, lexicographical orderings are irrational. The irrationality is greater when the "later testing" for other qualities is not in fact carried out. In actual, middle-brow econometric practice it seldom is. (See the papers cited earlier for some examples drawn from the *American Economic Review*.) Most economists pack up their statistical package and go home as soon as they find "significant" results "consistent with the hypothesis."

[Beads of sweat appear on his forehead.] "But everyone does it. It must have some survival value in producing good economics. And someone who knows more about statistics than I do must have decided that it is a good practice. After all, the econometrics textbooks and the canned programs and all the papers in the journals are filled with it."

The argument here is from authority. Arguments from authority are not always wrong, though this one seems to be. I do not know why economists and quantitative historians have misread their statistics books. It would make a good paper on the rhetoric of econometrics to trace the literature back to the authoritative turnings. (I record the impression that the reliance on significance in dropping variables is not usually recommended in so many words by textbooks, but in practice has figured more heavily as computers have become cheaper.) One can merely quote authorities in reply, and note that the authorities are of the best sort. Again I refer to the articles mentioned above and the works cited there: for instance, the article by William H. Kruskal (past president of the ASA, etc., etc.), "Statistical Significance" in the *International Encyclopedia of Statistics* (1978; and an earlier version in the *International Encyclopedia of the Social Sciences* (1968)); or the elementary book, *Statistics* (pp. 501, A-23, and *passim*), by David Freedman, Robert Pisani, and Roger Purves (well-known statisticians, youngish turks, etc., etc.). The point has been well known since the early days of modern statistics. Only 3.14159% of economists seem to be aware of it (a short list would include some specialist econometricians such a Griliches and Leamer and a few amateurs such as Arrow and Mayer; I learned it from Eric Gustafson).

[He loosens his tie, sweat dripping from his nose.] "But there's nothing else to do. I want to use statistical procedures. What do you propose to substitute? How will I fill up my days?"

Fill them up with statistical calculations that are to the point. Find out what people consider to be a large coefficient and then see if your data show it. Do sensitivity analysis. Bend over backwards to see how robust your argument is. Encompass your opponent's model with yours, showing how his results follow as special cases of yours. Take collecting "data" seriously (the word means "givens" in Latin: we should prefer "capta," things *taken*). There's plenty of useful econometric work to be done (see Sims, Leamer, Hendry, *et al* among econometricians, and Mosteller, Tukey, Hogg, *et al* among statisticians) that does not rely on the misuse of statistical significance.

[He is quaking nervously and his palms are wet. But at once he stops and cools. A smirk spreads over his face. He has found peace.] "To hell with you. As long as editors keep publishing articles that misuse statistical significance I'm going to keep on submitting them. I've got a career to run."

Shame on you. The argument is immoral. Our custom of forbidding talk about morality is strong among economists, some of whom think that the model of selfish behavior is in fact a set of suggestions about how to behave. But there's no two ways about it: it's immoral to lie, and for a scholar it's a mortal sin. That 96.8584% of editors fall into the group of economists who do not know the difference between statistical and substantive significance does not justify someone who does know the difference in going on pretending she does not.

Scholarship that depends on convenient lies will not last. To put it sharply, it is gradually becoming plain that the econometric work of the past quarter century relying inappropriately on statistical significance (which is most of it, unhappily) has to be done over again.

Economic historians are well placed to do better. We capture our own data, and therefore know that errors in variables are no joke. The intellectual traditions of cliometrics favor self-doubt, which in turn favors a sort of counting called "robust." Above all, we are trying to answer substantive historical questions about particular events, not trying to "test" more or less vague hypotheses about Economic Behavior. □