Econ Journal Watch,
Volume 1, Number 3,
December 2004, pp 539-545.

CORRESPONDENCE

Editors,

Deirdre McCloskey has complained about economists' mistaking statistical significance for significance. I thought the horse she was beating was dead already. But now she and Stephen Ziliak (in "Size Matters" *EJW* August 2004) demonstrate the beast is still breathing. For some reason McCloskey's logic hasn't penetrated.

Sometimes one good example succeeds where logic fails. It doesn't have to be a "real" example; a hypothetical should serve.

So I propose that every year, from a large population like that of the USA, a million forty-year-olds are selected at random and assigned, randomly, half to an experimental cohort and half to a control cohort, the former prescribed (and enforced) a 325 milligram aspirin tablet per day, the other half a placebo, both pills coated to avoid differences in taste or texture. As they die their age at death is recorded, and whether their death was related to coronary artery disease. Eventually we have recorded a few tens of millions of deaths and can identify any discrepancy in age of death, or cause of death, between the experimentals and the controls. I propose that the data support a positive benefit from the daily aspirin at a .0000001 level of confidence. If that's not convincing, keep going for another decade or two and the level goes to .0000000001. (Let me ignore how we enforce the regimen, and whether over half a century changes in diet, lifestyle, environment, etc., might cloud the results.)

What do we know about the benefits of a daily aspirin? McCloskey and Ziliak would say, "almost nothing." I'd say, only slightly different, "that there may possibly be some value in taking an aspirin per day, worth looking into." When we do look into it we may find that the daily aspirin's value, thought positive, is worth something less than the (small) cost of the daily aspirin, or even less than the trouble of pouring a glass of water and remembering the daily aspirin. Of course it may make a whale of a difference. The McCloskey point is that the .00000001 or the .0000000001 doesn't tell us whether to take an aspirin.

Of course, there will be those among us who decide that aspirin is cheap and any positive result is sufficient, irrespective of magnitude. But if that's one's attitude, there is yet no demonstration that an aspirin a day makes a significant difference. We want the experimenters to look at their results and resolve the issue: Does aspirin make enough difference to take seriously? The .0000000001 doesn't tell me the answer.

Thomas C. Schelling
University of Maryland

ZILIAK AND MCCLOSKEY REPOND:

Editors,

We were pleased to read the letter by Tom Schelling, showing decisively with a new example that statistical significance is neither necessary nor sufficient for proving *economic* significance. Tom notes that the beast—The Standard Error—is still breathing. We can add a little precision: now it dominates in 80% or more of the published papers. We've got to kill it if economics is to progress.

We asked our old friend William Kruskal, a distinguished statistician and past president of the American Statistical Association, "How could the confusion of statistical significance for economic significance proliferate?" He replied, "Well, I guess it's a cheap way to get marketable results." Indeed. An economist couldn't put it better. Orley Ashenfelter told us recently that he tended to agree that the procedure is nonsense, but that young people have to run their careers. It's a wonderment that economists, even Orley, who teaches the stuff, do not yet feel ashamed.

Results of Tom's own study of empirical methods in the 1998 "Committee on Journals" suggest dissatisfaction with the publication standards of the *American Economic Review*. A third of the respondents to the Committee's survey said of the *Review* "there is too little empirical data." "Hardly any [said] too much . . . three-fifths [of the respondents said there is] too little policy focus [too little emphasis on effect size and real-world importance], hardly any [said] too much."

Significance testing is of course a big part of an empirical publication in the *AER*. Over 95% of the empirical papers published in it rely on significance testing to *some* degree; to rely on it to any degree is a plain error in elementary statistics. Schelling's Committee did not look into significance testing. But given the response of the profession to the Committee's findings (namely, none), and to our 1996 paper in the *JEL* at about the same time (next to none), we suspect nothing would have changed had a question on significance been added. Tom urges us to keep trying.

Raising the price of *t* is the ticket. Fisher's 1925 publication of Student's table of *t* was slowed down by a weak-armed assistant of Gosset, a young man who could only barely turn the crank of the calculating machine. Gosset, that "Student," was a mathematically savvy chemist working as the chief brewer for the Guinness Corporation, and had to do a lot of the cranking himself. In the 1920s it took a lot of effort, in other words, to get an exact value of *t* for each degree of freedom. The high price did not stop the Significance Mistake from coming. But it came, slowly at first, and now like an avalanche. Eighty years on, it is essentially free to find a *t*. Credit the desk-top computer. "The cheapest way to get marketable results," as Kruskal said.

To unblock the journal referees and editors, and break out of what Morris Altman calls "a steady-state low-level equilibrium," we propose to try real market forces, as against the monopolized traditions of journal editing at present. If some auto mechanics used unsafe procedures to fix brakes—comforting themselves by saying that after all a young auto mechanic has to run his career—it would be a service to public safety and to the good mechanics to publicize the names of the good ones. Market forces would reward them. We propose an Angie's List of economists who understand the insignificance of statistical significance. We will ask a carefully constructed set of major economists and econometricians (every editor of every major journal, for example) to state publicly and for publication by return-addressed postcard their support for the following propositions:

1. Economists should prefer confidence intervals to other methods of reporting sampling variance;

2. Sampling variance is sometimes interesting, but a low value of it is not the same thing as scientific importance;
3. Economic significance is the chief scientific issue in economics; an arbitrary level of sampling significance is no substitute for it;
4. Fit is not a good all-purpose measure of scientific validity, and should be deemphasized in favor of inquiry into other measures of importance.

We will publish the names and responses, or lack of response, of everyone asked. On reflection any economist and econometrician who understands basic statistics will of course agree with the implied standards. In the decades since 1919 that we and scores of others have been making this point, no one has ever been able to *defend* the practice of significance testing. True, many people have gotten angry at the challenge. But no one has actually *met* it.

Stephen T. Ziliak
Roosevelt University

Deirdre N. McCloskey
University of Illinois-Chicago and
Erasmus University, the Netherlands

Editors,

Ziliak and McCloskey's indictment of economic studies misusing statistics (*EJW* August 2004) does not go far enough. A critical question is missing from their survey. Namely, does the study use out-of-sample data to test its model. This is not the same as testing against the correct null hypothesis, an issue ZM deal with.

Any new scientific theory (H1) has to go through two separate processes to be accepted as a replacement for the status quo (H0). First is its creation, second is its testing to see if it *predicts* better than H0.

A theory can be created from a mathematical derivation as in physics, purely from empirical data as in most other sciences, or some combination of both. In a statistical model, even if you have run all your tests and calculated your coefficients absolutely properly you have only accomplished the first step, creating H1. Getting a statistically significant result against the correct H0 merely suggests you are on the right track. To truly test and verify H1 you need to test it against H0 with data *different* and *independent* from that from which H1 was created.

In econometrics this verification can only come from testing H1 on data not used to create it. While it would be sufficient to show that H1 was wrong because the author calculated wrongly in any of the many ways ZM list, this is not a necessary condition to reject H1.

An alternative to fitting coefficients is to stipulate them before doing any testing. Coefficient values could be stipulated by what makes economic sense. However, even in this case it's possible that some implicit data fitting is occurring because any practitioner is going to have some familiarity with the data ahead of testing just from becoming conversant in the field. Even if you are careful about implicit fitting in this case the ultimate test of H1 can still only occur once H1 is tested on sufficient data that occurs after H1 is created. This may seem like an overly cautious standard but it is the only way to guard against implicit fitting.

In practice if a model H1 is fitted on data from time period t_0-t_1(in-sample) it then needs to be tested on data from time period t_1-present (out-of-sample). This doesn't mean refit H1 on out-of-sample data but test it to see if it predicts better than H0 in this time period. You can create an H1 with the highest t-values, R2, and what have you from in-sample data but it's completely meaningless until you see if H1 out predicts H0 in out-of-sample data.

In ZM's wonderful example of Milton Friedman's brief foray into metallurgy, the out-of-sample test of his theory was done by actually testing the newly created metal. In the physical sciences this concept boils down to can the experiment be reproduced. If and only if the experiment can be independently verified is H1 accepted. In 1989 when Martin Fleischmann and Stanley Pons reported their experiments which showed cold fusion, the scientific world did not immediately accept the implications. Certainly scientists were dubious because of the conflict with known theory, but it was also that the idea had to await additional independent experimental verification which never came.

Unlike physical sciences where you can create more data by running more experiments or in biology by getting more samples or subjects, in economics and finance you can only wait for the passage of time if you've used all your data as in-sample data. Unfortunately there are no short cuts so the wise econometrician partitions the data into in-sample and out-of-sample before any hypothesizing.

Say in the case where you have been careful in segregating your data, creating an H1 that's both economically and statistically sensible and now when you run your test on out-of-sample data the model fails. At this point it is not good enough to go back to the drawing board and come up with some tweaks in your in-sample data and retest. Once the test is made on out-of-sample data the out-of-sample data now becomes part of the in-sample data for any future hypothesizing. Any additional testing must occur on new data. If you use the important ideas ZM discuss beforehand it makes it more likely H1 will prove out but going through these steps is no guarantee of success and certainly not sufficient to claim H1 has been verified.

In applied finance, good quantitative traders know these rules well. These traders also know not to play other games like selecting your in-sample data from a time period after your out-of-sample data. To ignore these rules means creating models that are likely to lose money once they are put into use and perhaps having to find a new line of work. Financial markets are very unforgiving.

For the work of economists to be as rigorous as other scientists it needs to be tested on data outside their samples used in their specific research. A new theory cannot be said to be explanatory until it is shown to predict better than alternatives. Fleischmann and Pons became pariahs in their fields not because they made an error in their experiments but because they acted as though their findings should be accepted before they were verified.

Bob Gelfond
CEO and Founder
MagiQ Technologies, Inc.

ZILIAK AND MCCLOSKEY REPOND:

We agree with Robert Gelfond that a test on out-of-sample properties is a good argument. We do not agree, though, that it is necessary, a test that "any new scientific theory must go through." If this was what scientific progress meant there would be no progress in the historical sciences, in which the data are all in, such as evolutionary biology or geology or cosmology or history itself. Out-of-sample prediction, like a test on other sorts of data entirely or a consideration of logical coherence or a *gedankenexperiment*, is a useful option in scientific rhetoric. But there's no formula for scientific argument, and no timeless rules. We don't think that prediction is a rhetorical gold standard in science against which all arguments are to be tested. As we say: that way lies the end of geology and history.

We are sorry that Gelfond accepts the conventional calumnies on Fleischmann and Pons, and suggest that he have a look at a book by Eugene F. Mallove, *Fire From Ice: Searching for the Truth Behind Cold Fusion* (Wiley, 1991). We think it was thugs at Cal Tech and MIT, outraged that mere chemists would claim to have done some physics, that settled the issue in the minds of readers of *The New York Times*. But hundreds of articles are still published each year on the phenomenon Fleischmann and Pons discovered. *The Times* is not a good source for history of science.

Nor do we agree—and this is what really matters here—that "getting a statistically significant result against the correct H0 suggests you are on the right track." Our main point is that it suggests nothing of the kind. Gelfond speaks of "both economically and statistically sensible." If by "sensible" he means "passing conventional levels of significance, modulo sample size," he's missing our main point. It is: arbitrary levels of significance don't matter for science at all. The test of a financial model is its impact on a trader's bank balance—an economic criterion—not any statistic in itself. We note, alas, that this Good Old Chicago School point has recently given way in finance to more and more sophisticated—but financially and scientifically irrelevant—tests of "significance."

So we worry then that Gelfond has not grasped our simple, central point. But that's not unusual. It's been our experience for years and years and years!

Econ Journal Watch welcomes letters commenting on the journal or articles therein.
Please send correspondence to editor@econjournalwatch.org, *Subject line:* EJW Correspondence.