

**Comments on “Size Matters: The Standard Error of Regressions  
In the American Economic Review”**

**By**

**Stephen T. Ziliak and Deirdre N. McCloskey  
(paper presented at ASSA Meetings,  
San Diego, California, Jan., 2004)**

**By**

**Arnold Zellner  
University of Chicago  
1101 East 58th Street  
Chicago, IL 60637**

January, 2004

Comments on “Size Matters: The Standard Error of Regressions  
In the American Economic Review”

By

Stephen T. Ziliak and Deirdre N. McCloskey

(paper presented at ASSA Meetings,  
San Diego, California, Jan., 2004)

By

Arnold Zellner

University of Chicago\*

---

\* Research financed in part by the National Science Foundation and funds from the H.G.B. Alexander Endowment Fund, Graduate School of Business, University of Chicago.

e-mail: [arnold.zellner@gsb.uchicago.edu](mailto:arnold.zellner@gsb.uchicago.edu)

home page: <http://gsbwww.uchicago.edu/fac/arnold.zellner/more>

The authors are to be congratulated for bringing a very important problem to our attention, namely the poor quality of testing procedures used by many economists in their papers published in 1990 issues of a world's leading economic journal, *The American Economic Review*. That this problem has also been found and reported in an earlier paper by the authors relating to articles published in the 1980 issues of the same journal and that there does not appear to be much, if any, improvement in the quality of testing between the 1980s and the 1990s is indeed shocking. They are very critical of tests of significance and urge in their conclusions that a researcher's paper should concentrate on "... the size of the effect it is trying to measure ..." (p.27) and not on "... irrelevant tests of the coefficient's statistical significance ..." (p. 27). Also, throughout their paper they rightly, in my opinion, emphasize a difference between "economic" and "statistical" significance.

I note that the authors cited my 1984 paper in which testing procedures employed in 1978 issues of several leading economic and econometric journals were investigated with the finding that they were rather unsatisfactory in many respects, a finding quite similar to those mentioned in the current Ziliak-McCloskey paper. After reviewing the types of hypotheses (sharp, non-sharp, combined sharp and non-sharp, etc.) that were considered by the researchers in my sample of 1978 papers, I explicitly considered the issue of whether a separate theory of testing is needed and the statistician Good's (1980) suggestion "to roll together significance testing and estimation in a single process." In connection with this issue, I wrote,

"This suggestion seems questionable and fails to take account of a fundamental difference between estimation and significance testing problems. In the former, no particular value or values of the parameters are singled out for special attention, whereas in the latter, particular values, often zero, are given special attention. This is not to deny that estimation problems are often erroneously treated as testing problems. However, as Jeffreys (1963) states, 'Every quantitative law in physics implies a series of significance tests that have rejected numerous possible modifications of the law' (p. 403). In econometrics, too, most economic theories such as Friedman's theory of the consumption function imply hypotheses which can and have been tested. For example, is the elasticity of permanent consumption with respect to permanent income equal to one?"

(pp. 275-276). Many other important economic testing problems have been analyzed in the literature, as is well known and revealed in our surveys of the literature. Thus, testing is an important part of economics and other sciences.

In connection with the state of testing in economic research, I was happy to learn that the authors are in agreement with my statement, Zellner (1984, p. 280) "...there is room for improvement in the analyses of hypotheses in economics and econometrics" but was unhappy that they did not include the following sentence, namely, "In the following sections Bayesian results, which may be helpful in this regard, will be presented." I showed, building on much previous research by others and myself, that the Bayesian posterior odds approach to analyzing hypotheses is much to be preferred to the Fisher and Neyman-Pearson approaches to analyzing all kinds of hypotheses or models. As I pointed out, for a range of testing problems, the posterior odds is a monotonic function of Fisher's p-values that have been used and misused in evaluating alternative hypotheses. Also, it was shown that posterior odds have good sampling properties and can reflect individuals' prior views quite readily. Further, exact finite sample odds can be computed for many testing problems, e.g. time series problems, for which exact finite sample, sampling theory tests are unavailable. Finally, Jeffreys, Schwarz, and others have provided large sample posterior odds for a very wide range of testing problems that are very easy to implement. That is, the posterior odds relating to two hypotheses, denoted by  $K_{12}$ , satisfies the following relation in large samples:

$$-2 \ln K_{12} \doteq -2 \ln LR - k \ln n \doteq x_v^2 - k \ln n$$

or, 
$$K_{12} \doteq n^{k/2} \exp\{-x_v^2 / 2\}$$

where  $n$  = sample size,  $LR$  = likelihood ratio test statistic,  $v$  is the degrees of freedom, and  $k$  is the additional number of parameters under the alternative hypothesis. Thus, it is seen that in this large sample approximate expression the likelihood ratio plays a role as well as the number of additional parameters under the alternative and the sample size,  $n$ . Note that for a given value of the chi square statistic, as  $n$  grows in size, the posterior odds,  $K_{12}$  grows in size. Thus, a given p-value has different implications, depending on the size of the sample, as has been recognized in the literature. Further, in a particular

example relating to a normal mean testing problem, with  $n = 21$ , I have calculated the following relation from the above relation for the odds:

$$\ln K_{12} = 2.52 + 0.877 \ln P + \hat{u} \quad r^2 = .9998$$

where  $P$  is the Fisher  $p$  value. It is seen that the elasticity of the posterior odds with respect to the  $p$  value is 0.877; that is, there is a relation connecting the odds and the  $p$ -value that gives it some informational content. However, the relation will be different for different values of  $n$  and  $k$ . Further, whereas the posterior odds relating to the two hypotheses has a direct relation to our relative beliefs in the two hypotheses, the  $p$  value does not have such a clear-cut relation to the relative beliefs in the two hypotheses. Further, in the article, it was pointed out that if a loss structure is available, then one can act so as to minimize expected loss in choosing between or among hypotheses.

To illustrate with an example, years ago Paul Evans reported his research results in the Money Workshop on his analysis of the German hyperinflation using two models, one based on rational expectations and the other on adaptive expectations, one theory implying that a particular parameter is equal to zero while the other theory implied that this parameter is equal to 0.7. His estimate of the parameter was a little larger than 0.3, with a large sample standard error somewhat larger than .3. He tested the hypothesis that the parameter's value is equal to zero and accepted it at the 5% significance level. Milton Friedman stopped the speaker with the remark that his estimate of the parameter is almost half way between 0 and 0.7 and then, looking at me, he asked Paul Evans, "Why don't you compute the posterior odds on 0 vs. 0.7?" I responded, "Milton, if you want him to do that, I shall be happy to show him how." Over the weekend, Paul learned how to compute and interpret posterior odds and found that, as Milton had suggested, that the evidence was inconclusive with respect to choosing between the two hypotheses, a result published in Evans (1978). Note that this involved assigning probabilities measuring degrees of belief in the two hypotheses and then using the information in the data to determine how it changed his initial beliefs.

Now it may be asked why are intelligent, highly trained empirical workers doing so poorly in testing hypotheses? My answer is that they are all mixed up on the methodologies of testing. Most of them don't know which concept of probability they are using, have a hard time interpreting  $p$ -values, don't know what power functions are and

don't know how to use them, particularly because they have no idea of what values of the parameters to use, and don't know how to choose significance levels as the sample size changes. Last, if they are interested in several hypotheses, they have no idea of how to do a joint test or to interpret and do simultaneous tests. Add to this the advice given by some to "test, test, test" without attention paid to the dependence of alternative tests and effects of pretests on subsequent inferences and one begins to understand why researchers are so mixed up about the theory of testing. And not only economic researchers are mixed up about these matters, but researchers in many other disciplines. Witness the wild discussions of the hypotheses of the existence or non-existence of cold fusion in the physics literature.

It is the case that Jeffreys wrote his book, *Theory of Probability* to instruct his fellow scientists on how to analyze their data appropriately. That is, he provided his fellow scientists with Bayesian estimation, testing, and prediction methods after carefully considering R.A. Fisher's maximum likelihood approach to estimation and his p-value approach to testing. He showed that large sample Bayesian procedures are compatible with the ML approach in estimation but soundly rejected p-values for testing. He explained the importance of associating probabilities with hypotheses in order to get a good theory of testing and noted that in other testing approaches, probabilities representing degrees of belief in alternative hypotheses are not used and thus these approaches do not measure the extent to which data used in testing change beliefs in alternative hypotheses. Also, while he lauded Neyman and Pearson for their explicit recognition of alternative hypotheses, he criticized them for their inability to use the power function without introducing prior information and for their frequentist definition of probability that does not permit probabilities to be associated with degrees of belief in hypotheses. Needless to say, Fisher and Neyman and Pearson did not accept Jeffreys' criticisms. In my paper, mentioned above, I showed by example how use of Bayesian posterior odds helped resolve some of these issues. And very recently, J.O. Berger (2003) titled his Fisher Lecture, "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" In his abstract, he wrote:

"Ronald Fisher advocated testing using p-values. Harold Jeffreys proposed use of objective posterior probabilities of hypotheses and Jerzy Neyman recommended testing

with fixed error probabilities. Each was quite critical of the other approaches. Most troubling for statistics and science is that the three approaches can lead to quite different practical conclusions.” (p.1).

In his paper, Berger attempts to provide a unifying framework for the above three approaches to testing using a “conditional frequentist” approach. He points out that in a simple normal location problem, the three approaches give quite different results. He states, “The discrepancy between the numbers reported by Fisher and Jeffreys are dramatic in both cases, while the numbers reported by Fisher and Neyman are dramatic primarily in the second case. [The first case involves a value of the z statistic = 2.3 and the second = 2.9.]” (pp. 1-2).

That these dramatic differences occur in the simplest normal testing problems indicates that it is important that statistical researchers and others provide results on the relative theoretical and practical merits of alternative testing approaches. While much more can be said about these issues, I shall just remark that it is little wonder that applied researchers are somewhat confused about testing and, according to the evidence provided by Ziliak and McCloskey, are not testing very well.

Last, at least three most important issues have not as yet been mentioned, namely (i) quality of data and (ii) adequate functional forms for data and (iii) finite sample tests vs. large sample tests. On quality of data, I like to recall D. Gale Johnson’s work to determine whether a Cobb-Douglas function was appropriate for pre-World War I agricultural output. He fitted the function and got a perfect fit,  $R^2 = 1$  ! He explained that he found the function that had been used to create the output data. Many other features of data generation are often overlooked in testing using cross section or panel data. For example are the data generated by simple random sampling or by a cluster sampling design? If the latter, the data may not be independent across respondents and it is well known that a departure from independence can have an important influence on testing procedures’ properties. Another problem with such data is missing observations. When Don Rubin addressed our group on imputation techniques, I asked him what percentage of the points in survey data are imputed. He surprised me by saying that probably about 20% or more are imputed, a fact that may have an important impact on properties of testing procedures. Also, during my thesis days, a “few” years ago, when I worked with

seasonally adjusted quarterly personal consumption and income data, I wrote to Washington, DC to obtain the seasonally unadjusted data and was told that for income it did not exist! When I wrote asking for an explanation, I was told that proprietors' incomes were just available once a year from tax records and interpolated over the quarters of the year without a seasonal component. Thus, there was no seasonally unadjusted quarterly personal disposable income series available! How all this affected my estimates, standard errors, tests, etc., is still a mystery. Of course, replication with improved data is most desirable. Finally, the ugly fact of response bias must be briefly mentioned since it is a frequent important problem in many areas and not dealt with adequately in many cases.

With respect to functional form, note, as I have pointed out earlier, see Zellner (1997), that those who have studied the relation between murder rates and execution rates have generally assumed a log-log relation. In such a relation, with a negative coefficient for the log of the execution rate, as the execution rate goes to zero, the model predicts that the murder rate goes to infinity, *cet. par.* Maybe a semi-log relation would be better. For some additional points regarding the testing of the murder rate and execution rate relation, see Ehrlich's (1996) and subsequent papers for an extensive and cogent discussion of various testing methodologies that have been employed. Further, some in testing the proportionality hypotheses between permanent consumption and permanent income have used linear relations to investigate this hypothesis, e.g.  $c = a + by + u$  without noticing that since  $0 < c < y$ , the dependent variable is truncated and thus it and the error term  $u$  can not be normally distributed with doubly infinite ranges. Also, under the alternative hypothesis,  $a > 0$ , at low levels of permanent income, the  $Ec/y$  ratio can be bigger than 1! Tests' results are affected by careless specification of models. See Zellner and Moulton (1985) for empirical results on testing the proportionality hypothesis using alternative functional forms.

Third, in the case of time series problems, simultaneous equations models, and a number of other problems, asymptotically justified tests are employed because finite sample sampling theory tests are not available. Many have recognized that these large sample tests can be very inaccurate in small sample situations. Thus, it is very fortunate

that exact finite sample posterior odds are available for comparing, testing, and/or combining alternative models.

In closing, I hope that this very useful paper by Ziliak and McCloskey will be followed by further research to compare further Bayesian and non-Bayesian testing procedures to determine which perform better and are more useful for applied researchers in economics and other areas of scientific research. From the evidence that I have seen, it is my belief that traditional Bayesian and new Bayesian information theoretic approaches, see, e.g., Zellner and Tobias (1999) and Zellner (1997, 2002, 2003) will meet the needs of researchers and permit them to incorporate significant economic size considerations in their testing procedures and to reach significantly better conclusions.

### References

- Berger, J.O. (2003), "Fisher Lecture: Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" (with discussion), *Statistical Science*, 18, No. 1, 1-32.
- Ehrlich, I. (1996), "Crime, Punishment, and the Market for Offenses," *J. of Economic Perspectives*, 10, No. 1, 43-67.
- Evans, P. (1978), "Time Series Analysis of the German Hyperinflation," *International Economic Review*, 19, February, 195-209.
- Jeffreys, H. (1963), "Review of L.J. Savage et al., *The Foundations of Statistical Inference* (1962)," *Technometrics*, 5, 407-410.
- \_\_\_\_\_ (1998), *Theory of Probability*, reprint of 3<sup>rd</sup> revised edition, first edition, 1939, in *Oxford Classic Texts in the Physical Sciences*, Oxford: Oxford U. Press.
- Zellner, A. (1984), "Posterior Odds Ratios for Regression Hypotheses: General Considerations and Some Specific Results," published in A. Zellner, *Basic Issues in Econometrics*, Chicago: U. of Chicago Press, 275-305.
- \_\_\_\_\_ (1997), *Bayesian Analysis in Econometrics and Statistics*, invited contribution to M. Perlman and M. Blaug, eds., *Economists of the Twentieth Century Series*, Cheltenham, UK and Lyme, US: Edward Elgar Publishing Ltd.
- \_\_\_\_\_ (2002), "Information Processing and Bayesian Analysis," *J. of Econometrics*, 107, 41-50.

\_\_\_\_\_ (2003), "Some Historical Aspects of Bayesian Information Processing,"  
invited paper presented at the American Statistical Association Meeting, August,  
San Francisco.

\_\_\_\_\_ and Moulton, B.M. (1985), "Bayesian Regression Diagnostics with  
Applications to International Consumption and Income Data," *J. of Econometrics*,  
29, 187-211.

\_\_\_\_\_ and Tobias, J. (1999), "Further Results on the Bayesian Method of Moments  
Analysis of the Multiple Regression Model," *International Economic Review*, 42,  
February, 121-140.