This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: Evaluation of Econometric Models

Volume Author/Editor: Jan Kmenta and James B. Ramsey, eds.

Volume Publisher: Academic Press

Volume ISBN: 978-0-12-416550-2

Volume URL: http://www.nber.org/books/kmen80-1

Publication Date: 1980

Chapter Title: Some Comments on the Papers by Welsch and Hill

Chapter Author: William S. Krasker

Chapter URL: http://www.nber.org/chapters/c11702

Chapter pages in book: (p. 223 - 226)

Some Comments on the Papers by Welsch and Hill

*WILLIAM S. KRASKER**

UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN

Ever since the introduction of high-speed computers, econometricians have been able to analyze data with great ease. Although this computational revolution has been beneficial on the whole, it has had an unfortunate side effect. One is less likely to uncover erroneous observations and other anomalies in the data when using a computer than when analyzing data by hand.

With large data sets, one has almost no choice but to rely entirely on well-defined programmable diagnostics and estimators. Since our models remain approximations to reality and our methods of data collection are still imperfect, it is important that our diagnostics alert us to aberrant data. Moreover, we would like to have estimators which are not overly sensitive to small departures from the assumptions. In recent years statisticians have begun to study the existing statistical procedures to see which ones have this limited-sensitivity property and to develop alternative techniques where necessary. The papers by Roy Welsch and Bruce Hill provide examples.

Using an empirical example, Welsch focuses on the sensitivity of ordinary least squares estimates to changes in small subsets of the data. The model—which is a large cross section with 14 parameters—provides an excellent illustration of the potential hazards of relying exclusively on least squares.

Welsch considers first the possible nonnormality of the disturbances and, in fact, finds evidence that the distribution is heavy-tailed. This is a step which is often ignored by econometricians, some of whom are unaware that there are straightforward ways to check the validity of the normality

* Present address: Graduate School of Business Administration, Harvard University, Soldiers Field, Boston, Massachusetts 02163.

Copyright © 1980 by Academic Press, Inc.
All rights of reproduction in any form reserved.
ISBN 0-12-416550-8

assumption. On the other hand, Welsch's lack of emphasis on the normality issue is commendable. Even without normal disturbances, the least squares estimator is generally consistent and asymptotically normal, though less efficient than some other (nonlinear) estimators. In a large sample, even the relatively inefficient least squares estimates will often be precise enough to yield useful inferences. The real danger in a large sample is the bias caused by aberrant data or other misspecifications, and it is this problem with which Welsch's main results deal.

Perhaps the simplest way to achieve protection against aberrant data is to use diagnostics which reveal the influential observations. Those observations which have a large effect on the results should be examined closely, keeping in mind the possibility that they were generated by a process not fully accounted for by the model. Welsch discusses and applies some of the tools for unmasking influential observations which he and others have developed, and the results are quite revealing. Certain observations were found to be highly influential. More importantly, because of the large number of explanatory variables, it is likely that some of the influential points would have gone unnoticed in a more traditional analysis.

There is, however, a problem which arises with the preceding diagnostics. Except in cases where the disparate observation is subsequently found to be erroneous (which frequently happens), one does not know how to proceed after the influential points have been located. One hesitates to leave them in the sample; for if they do not obey the same stochastic law as the bulk of the data, they may cause a substantial bias. On the other hand, unnecessarily removing those points may result in a large loss of efficiency.

An alternative is bounded-influence estimation, i.e., using an estimator which is consistent when all the model's assumptions hold but which automatically limits the influence of any small subset of the data. Since disparate observations cannot greatly alter the fit obtained from the bounded-influence estimator, those observations necessarily show up with large residuals. For this reason it is unfortunate that Welsch did not include the residual plot from his bounded-influence regression, which is an important and revealing diagnostic.

Also, one would like to see the asymptotic standard errors from the bounded-influence regression. Besides yielding a measure of the precision of the estimates, they provide a way to test the assumptions. Since both the least squares and the bounded-influence estimators are consistent when all the classical assumptions hold, we have evidence against these assumptions when the two estimates differ "significantly" from each other. The omission of the estimated asymptotic standard errors is perhaps justified, however, since little is known about their properties. Indeed, as Welsch points out, bounded-influence estimation is currently an area of active research.

The paper by Bruce Hill studies robust estimation in the "random model." At the outset, we should mention three aspects of the paper which depart from the framework usually adopted by econometricians. First, Hill's treatment is explicitly Bayesian throughout. As he points out, it is important to be explicit about prior information in the random model, for the robustness of his methods depends crucially on which of the parameters are *a priori* independent. A non-Bayesian approach would not eliminate this problem, but simply ignore it.

Second, most of Hill's paper is restricted to the simple case of analysis-of-variance. Econometricians are accustomed to working only with the more general regression model and may wonder if Hill's restriction to analysis-of-variance in the first half of his paper simplifies the exposition enough to compensate for the loss of generality. On balance, it would seem that the main points of the paper are made adequately by restricting attention to what would be the varying intercept terms in a regression model, so that little would be gained from considering the regression model at the outset.

Finally, when the paper does address the regression case in Section 4, the population being sampled is assumed finite. This assumption is virtually never made in econometric models, even though it is, strictly speaking, sometimes the correct formulation. Of course, it is true that if the finite population is very large, it can be treated as infinite. For example, a survey of households from a population such as the state of Michigan is certainly sufficiently large. But the whole point of the assumptions underlying Hill's analysis is that the sampling will take place within well-defined blocks, such as counties, or towns, or individual apartment buildings. Certainly, in the last case, at least, the infinite-sample approximation would not be justified.

When estimating the parameters of a statistical model, econometricians usually make assumptions and then seek an estimator which is efficient relative to those assumptions. Only rarely do econometricians examine the properties of their estimators when the assumptions hold only approximately. As Hill notes, estimators which use all the information in the sample frequently are not robust, so that one should consider "inefficient" estimators which are less sensitive to those assumptions about which one has the least confidence. Frequently the robust estimators are much more complicated than the classical estimators. However, in Hill's study, the robust alternative procedures are actually simpler than the "efficient" estimators; indeed, this was a major factor in their selection. For example, to use the nonlinear function $H(x)$, which arises in Hill's treatment of the regression model, we must evaluate the functions $W_i(x)$. Alternatively, one can achieve a degree of robustness by using a simple linear approximation to $H(x)$.

Two additional aspects of Hill's view of robustness deserve comment. In several places in his paper he suggests trying a few different shapes for the underlying distributions to determine the sensitivity of the results. However, this can only show robustness to the specific alternative distributions which are considered. Where possible, it is better to show that the results will not change much provided the true distributions are in some neighborhood of the assumed shapes (with respect to some metric on the space of distributions). This has been done in other contexts, though it has not been used in a Bayesian analysis such as Hill's.

Finally, Hill says at the end of his paper that we should build better models, rather than merely seek protection against small departures from the assumed model. This is certainly true, but the fact remains that even the best available model will never be an exact description of reality. The proper interpretation of Hill's statement, in my opinion, is that the use of a robust estimator does not justify neglecting potential improvements in the model. The model should always be as well specified as possible—but one should nevertheless protect oneself against whatever small errors remain.