

Resources for the Testing and Enhancement of Statistical Software

Micah Altman
Harvard University
micah_altman@harvard.edu

Michael McDonald
Vanderbilt University
mmcdonal@weber.ucsd.edu

Soon after the development of the mainframe computer, Longley (1967) criticized regression programs using it for being dramatically inaccurate. Approximately every ten years thereafter, each new generation of statistical software has been similarly faulted.

In a startling article, McCullough & Vinod (1999) argue that econometric packages can give "horrendously inaccurate" results and that these inaccuracies have gone largely unnoticed (p. 635-7). Moreover, they argue that in consequence of these inaccuracies, past inferences are in question, and future work must document and archive statistical software alongside statistical models (p. 660-662).

In contrast, political scientists writing about quantitative analysis tend not to discuss issues of accuracy in the implementation of statistical models and algorithms. Few of our textbooks, even those geared toward the most sophisticated and computationally intensive techniques, mention issues of implementation accuracy and numerical stability. When political scientists discuss accuracy in computer-intensive quantitative analysis, they are relatively sanguine about the issues of accurate implementation.

In an ongoing research project, we measure the accuracy of statistical abstractions as implemented in statistical packages popular among political methodologists, such as Gauss, Stata, SST and Excel. We evaluate the use of these abstractions in the context of evaluating complex statistical procedures, such as Gary King's (1997) solution to ecological inference. Our working paper, and an extensive list of resources, is available from our web-site:

http://data.fas.harvard.edu/numerical_stability/

To test the accuracy of statistical packages, we implement a series of benchmarks proposed by the National Institute for Standards and Technology (NIST). Like McCullough and Vinod (1999), we find that statistical packages can produce dramatically different results. For example, consider the NIST data set Hahn1. This data comes from a study of thermal expansion in copper, in which non-linear regression was used to estimate the coefficients of the following equation:

$$y = \frac{\beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3}{1 + \beta_5 x + \beta_6 x^2 + \beta_7 x^3} + \epsilon$$

Using three common statistical packages and the NIST data, we estimated this equation. We report these estimates with three significant digits, along with the correct estimates (calculated by NIST using multiple precision). Surprisingly, the coefficients and standard errors are all different from one another, sometimes in opposite directions and sometimes off by orders of magnitude.

Table 1a: Will the True Answer Please...

	Program A	Program B	Program C
β_1	-1.70e+00	1.08e+00	-8.21e+03
β_2	5.75e-02	-1.23e-01	-2.25e+05
β_3	1.27e-03	4.09e-03	-7.05e+06
β_4	-2.01e-07	-1.43e-06	2.27e+07
β_5	-2.92e-03	-5.76e-03	2.31e+06
β_6	9.47e-05	2.41e-04	8.01e+07
β_7	-4.46e-08	-1.23e-07	1.09e+06

Table 1b: Stand Up

	Correct Values	
	β	Standard Error
β_1	1.08e+00	1.71e-01
β_2	-1.22e-01	1.20e-02
β_3	4.09e-03	2.25e-04
β_4	-1.43e-06	2.76e-07
β_5	-5.76e-03	2.47e-04
β_6	2.41e-04	1.04e-05
β_7	-1.23e-07	1.30e-08

It turns out that Hahn1 is especially troublesome for non-linear least squares solvers. Even so, we find that some statistical packages can give unreliable estimates on many benchmark tests. Others, however, perform well in dealing with the standard types of problems that most users face. It is possible to find, and use, packages that produce trustworthy univariate statistics, linear regressions, and anova estimates. For many users, finding the right package will be sufficient.

In contrast, nonlinear optimization, maximum likelihood, and simulation problems can be intrinsically hard. For these problems, there is no set of tractable, universally applicable, boundedly accurate, robust, solution techniques. Many statistical packages provide random number generators and statistical distributions that are inadequate for serious simulation. Moreover, even the best algorithms, implemented correctly, can go astray on complex problems, and there is often no way of knowing beforehand that problems will occur, and no definitive test that the algorithm

can perform to verify success afterwards. Therefore, political scientists who work with complex simulations cannot leave the solutions up to the developers of statistical applications. They must develop an understanding of the different algorithms themselves, know where each is likely to be appropriate, and apply multiple techniques to test the robustness of their estimations.

We make three recommendations, in short:

1. Choose accurate and robust software, and provide developers with an incentive to make accurate software. (There is good news, publicizing this research helps to provide these incentives - many of the problems we note have already been fixed).
2. For standard analysis, accurate software may be all that we need. For complex simulation, non-linear estimation, or ugly data, we can often get better results by paying attention to how are solutions are computed.
3. *Worry*. McCullough and Vinod claim that numerical instability pervades the practice of econometrics, and that this throws into question a wide variety of previous results. We too find that numerical instability is present in methods commonly used for political analysis.

But (perhaps) don't worry too much. In our initial examination of two published studies involving complex statistical models, we do find that changing the implementation can change published standard errors, but not by enough to affect substantive conclusions.

The extent of the effects of numerical instability in political science research is largely unknown. We plan further replication studies to investigate this, and particularly to identify prototypes for maximum likelihood estimation benchmarks. To this end, we invite all researchers concerned about the robustness of their results to provide public replication data and code, and we encourage researchers to contact us in this regard.

In addition to our working paper, and to aid political scientists with these issues, we have compiled an introductory annotated bibliography (below), and constructed a web page (http://data.fas.harvard.edu/numerical_stability/) with links to resources in the following areas:

Statistical Software Test Data and Test Matrices: Links to the NIST repository of reference data sets with certified computational results, which enables the objective evaluation of statistical software, and to other repositories of data for testing statistical procedures.

Random Number Generators and Test Suites: Links to tutorials, pre-prints, and software for random number generation. Includes links to George Marsaglia's DIEHARD suite for testing random numbers, and the SPRNG test suite from

the NCSA, as well as links to sources of true (physically generated) random numbers.

Optimization Resources: Links to optimization software that you can modify, and to optimization test-beds where you can submit your problems for analysis.

General Resources for Numerical Algorithms: Links to general resources for numerical algorithms, including repositories from the *Association for Computing Machinery*, and the *Journal of the American Statistical Association*. This also includes links to libraries of high precision statistical distributions.

Annotated Bibliography

Gentle, James E., 1998. *Random Number Generation and Monte Carlo Methods*. New York: Springer-Verlag. [An excellent review of the generation, testing and use of pseudorandom numbers.]

Gill, Phillip E., Walter Murray and Margaret H. Wright, 1981. *Practical Optimization*. San Diego: Academic Press, Inc. [An introduction to optimization that pays attentions to numerical stability.]

Longley, James W. 1967. "An Appraisal of Computer Programs for the Electronic Computer from the Point of View of the User", *Journal of the American Statistical Association*. 62(348): 856-66. [The first article to highlight the dangers of numerical stability in statistical research.]

Higham, Nicholas J., 1996. *Accuracy and Stability of Numerical Algorithms*, Philadelphia: SIAM. [An introduction to the analysis of numerical stability of linear problems.]

Knuth, Donald E., 1997. *The Art of Computer Programming*, (3rd Edition). Reading, MA: Addison-Wesley. [An encyclopedic reference across the fundamentals of computer science. Volume 2, Semi-numerical algorithms contains useful introductions on random number generation and floating point arithmetic.]

McCullough, Brian D., 1998. "Assessing the Reliability of Statistical Software: Part I," *The American Statistician* 52(4): 358-366. [A practical methodology for benchmarking the accuracy of statistical software.]

McCullough, Bruce D. and H. D. Vinod, 1999. "The Numerical Reliability of Econometric Software" *Journal of Economic Literature* 37: 633-665. [A startling argument on the importance of numerical accuracy in econometrics.]