

Wilkinson's tests and econometric software

B.D. McCullough

Department of Decision Sciences, Drexel University, Philadelphia, PA 19104, USA
E-mail: bdmccullough@drexel.edu

The Wilkinson Tests, entry-level tests for assessing the numerical accuracy of statistical computations, have been applied to statistical software packages. Some software developers, having failed these tests, have corrected deficiencies in subsequent versions. Thus these tests have had a meliorative impact on the state of statistical software. These same tests are applied to several econometrics packages. Many deficiencies are noted.

Keywords: Numerical accuracy, software reliability

1. Introduction

The primary purpose of econometric software is to crunch numbers. Regrettably, the primary consideration in evaluating econometric software is not how well the software fulfills its primary purpose. What does matter is how easy it is to get an answer out of the package; whether the answer is accurate is of almost no importance. Reviews of econometric software typically make no mention whatsoever of accuracy, though Vinod [14,15], Veall [13], McCullough [7,8], and MacKenzie [5] are exceptions.

In part, this lack of benchmarking during reviews may be attributed to the fact that, until quite recently, no one ever collected the various benchmarks in a single place (see the NIST StRD at www.nist.gov/itl/div898/strd). A reviewer may well have been aware of a few scattered benchmarks, but including only a few benchmarks is hardly worth the effort. This view is supported by the fact that while the statistics profession has a long history of concern with the reliability of its software, software reviews in statistical journals rarely mention numerical accuracy. However, there is one collection of tests which has been widely applied in the statistics literature: Wilkinson's [16] *Statistics Quiz: Problems which reveal deficiencies in statistical programs*, which is discussed in detail in Sawitzki [10]. These tests have been profitably employed by Sawitzki [11], who uncovered errors in SAS, SPSS, and S-PLUS, among other packages, and also by Bankhofer and Hilbert [1,2]. To date these tests have not been applied to econometric software.

The Wilkinson tests are not meant to be realistic: they were purposefully designed to expose specific errors in statistical packages. Their elegance is three-fold.

- First, they are simple. Therefore, they can reasonably be applied to most any statistical or software package.

- Second, the flaws they are designed to expose have well-known solutions. That is, these are tests which any package *could* pass. If a software package fails a particular test, there exists a known method of obtaining the correct answer.
- Third, they examine the maintained assumptions of the software we use, and which we rarely pause to question. When we have our program read a file, we assume that it is read correctly. When we graph a variable, we assume that the graph accurately represents the data. When we calculate a number, we assume the calculation is accurate and that missing values are “correctly” accounted for.

Statistics Quiz presents six suites of tests: reading an ASCII file; real numbers; missing data; regression; analysis of variance; and operating on a database. The first and last suites are for packages that claim to be general-purpose, and not suited to specialized econometric packages; also, analysis of variance is not much used in economics. The other three suites are relevant to econometric software, and so we apply them to more recent versions of several of the packages that MacKie-Mason [6] evaluated for user-friendliness: E-Views v3.0, LIMDEP v7.0 for Windows 95, RATS v4.3, SHAZAM v8.0, and TSP v4.4.

In applying these tests, we distinguish between “miserable failure” and “catastrophic failure” [9]. A program fails miserably when it refuses to complete the task and the user is informed that the task has not been completed. By contrast, a program fails catastrophically when it incorrectly completes a task without advising the user that something is amiss.

2. The data

Table 1 displays the data set “Nasty”, whose values are all well within the range of representable numbers for 32-bit double precision. The values for BIG are less than the US population, while the values of HUGE are the same order of magnitude as the national debt.

Table 1
Data Set NASTY.DAT

LABLE\$	X	ZERO	MISS	BIG	LITTLE	HUGE	TINY	ROUND
ONE	1	0	.	99999991	0.99999991	1.0E12	1.0E-12	0.5
TWO	2	0	.	99999992	0.99999992	2.0E12	2.0E-12	1.5
THREE	3	0	.	99999993	0.99999993	3.0E12	3.0E-12	2.5
FOUR	4	0	.	99999994	0.99999994	4.0E12	4.0E-12	3.5
FIVE	5	0	.	99999995	0.99999995	5.0E12	5.0E-12	4.5
SIX	6	0	.	99999996	0.99999996	6.0E12	6.0E-12	5.5
SEVEN	7	0	.	99999997	0.99999997	7.0E12	7.0E-12	6.5
EIGHT	8	0	.	99999998	0.99999998	8.0E12	8.0E-12	7.5
NINE	9	0	.	99999999	0.99999999	9.0E12	9.0E-12	8.5

Table 2
Results of Test IIA

	E-Views	LIMDEP	RATS	SHAZAM	TSP
print ROUND	p	p	p	p	p

3. The tests

3.1. Wilkinson's Test II. Real numbers

Test IIA. Print ROUND with only one digit. This does not mean truncate or round to one digit and then print; it means print displaying only one digit, so that the rounding is done by the program rather than the user. The use of FORMAT statements may be necessary. The answer should be the numbers from 1 to 9. All packages pass. See Table 2.

This is a test of the package's ability to round numbers. Some compilers round numbers inconsistently or use uncommon rounding methods such as round-to-even. Letting R be the rounding function, round-to-even has the interesting property that $R(1.5) = R(2.5)$, for example. An econometric package created with such a compiler may do so, too. Wilkinson presents another test in this section, which is relegated to the appendix for discussion.

Test IIB. Plot HUGE against TINY in a scatterplot. Plot BIG against LITTLE. In each case the answer should be a 45-degree line. The results can sometimes be surprising, as indicated by Fig. 1, which shows the correct graph produced by RATS and the graph produced by E-Views, which failed catastrophically. A common cause of such a result is that computation is in double precision while the graphics routine is in single precision. For such packages, data points may be incorrectly placed or omitted completely, making graphical analysis of data problematic, e.g., visual detection of outliers.

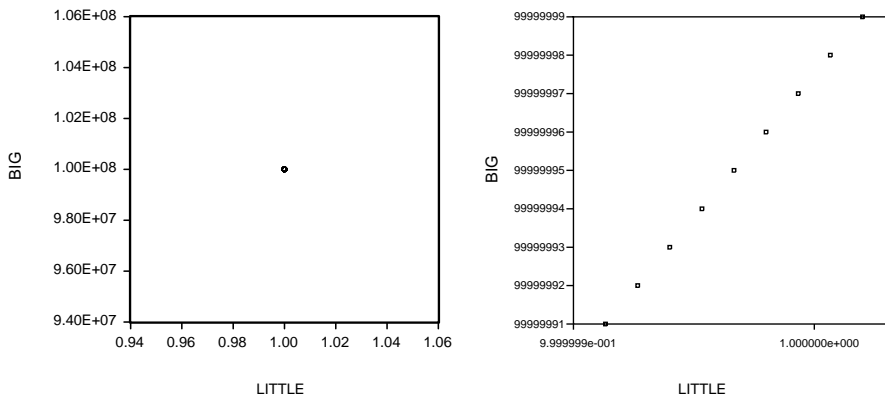


Fig. 1. Test IIB Results for E-Views (left) and RATS (right).

Table 3
Results of Test IIB. †Refused to produce a graph

	E-Views	LIMDEP	RATS	SHAZAM	TSP
HUGE v. TINY	p	p	p	p	p
BIG v. LITTLE	F	p	p	F	p
X v. ZERO	p	†	p	p	p

Table 4
Results of Test IIC – calculate the variance correct answers (indicated by 'p'): NA for ZERO and MISS, 2.738 for all others. †refused to calculate

	E-Views	LIMDEP	RATS	SHAZAM	TSP
X	p	p	p	p	p
ZERO	†	p	p	p	0
MISS	†	†	†	0	†
BIG	p	p	p	2.424	p
LITTLE	p	p	p	2.870	p
HUGE	p	p	p	p	p
TINY	p	p	p	p	p
ROUND	p	p	p	p	p

Plot X against ZERO. The answer should be a vertical line. Some packages, such as LIMDEP, refuse to produce a graph because they are unable to scale the horizontal axis for these data. Since the user is advised that the graph is not produced, this is a miserable failure, rather than a catastrophic failure. SHAZAM's basic plotting routine did not correctly plot big vs. little. (SHAZAM can produce publication-quality graphics by an interface to the "GNU PLOT" program, and these plots pass all the tests.) See Table 3.

Test IIC. Compute basic statistics on each variable. The means should be the fifth value of each variable. Standard deviations should be "undefined" or missing for MISS, zero for ZERO, and 2.738612788 (times 10 to a power) for all other variables (in the table the powers of ten are omitted). Generally, calculation of the means is correct with the following exceptions: SHAZAM returns zero for MISS, when the correct result is 'undefined', while E-Views refuses to calculate for MISS and ZERO. The standard deviation calculations produce some interesting results, as seen in Table 4.

For MISS, SHAZAM returned zero, and the others refused to perform the calculation. The SHAZAM results for BIG and LITTLE can be explained by its algorithm for computing the variance. Ling [4] and Chan et al. [3] analyzed various algorithms for computing the sample variance. The least reliable of these methods was shown to be the 'calculator formula' (so-called because it is used as a shortcut formula in

many elementary texts)

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 / n}{n - 1} \quad (1)$$

whereas the usual formula is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (2)$$

Equation (1) squares the observations themselves, rather than their deviations from the mean, thus unnecessarily using up the computer's finite precision. Indeed, texts on statistical computing (e.g. [12], use Eq. (1) as an example of 'how not to compute the sample variance'.

Test IID. Compute a correlation matrix for all the variables. The correlations for all variables should be unity, except for ZERO and MISS, which should be "undefined" or missing. If MISS must be removed from the dataset by the user before the correlations can be computed, this indicates that the package does not "handle" missing observations, but simply deletes them. This turned out to be the case for all the packages but SHAZAM. Results are displayed in Table 5.

The common failure for this test was the calculation of a zero correlation between ZERO and all the other variables. This is clearly an incorrect answer. The correlation coefficient is defined

$$\rho_{wz} = \frac{\text{cov}(w, z)}{\sigma_w \sigma_z} \quad (3)$$

where $\text{cov}(w, z)$ is the covariance between w and z and σ_w is the standard deviation of w . Since the standard deviation of ZERO is zero, Eq. (3) has zero in the denominator and so its correlation with any other variable is undefined. Additionally, E-Views and SHAZAM both return correlations greater than unity, which constitutes evidence of an unstable algorithm. SHAZAM computes a correlation of both MISS with itself and ZERO with itself as unity, when the correct answer for both is 'undefined'.

Test IIE. Tabulate X against X, using BIG as a case weight. None of the packages offers this procedure.

Test IIF. Regress BIG on X and a constant. The constant should be 99999990 and the coefficient should be unity. Summary results presented in Table 6 indicate that all programs pass.

Table 5

Results of Test IID. Lower diagonal of correlation matrix: Only incorrect results are displayed

<i>E-Views</i>								
	X	ZERO	BIG	LITTLE	HUGE	TINY	ROUND	
X								
ZERO	0	0						
BIG	1.13	0						
LITTLE	1.01	0	1.14					
HUGE		0	1.13	1.01				
TINY		0	1.13	1.01				
ROUND		0	1.13	1.01				
<i>LIMDEP, RATS, TSP</i>								
	X	ZERO	BIG	LITTLE	HUGE	TINY	ROUND	
X								
ZERO	0	0						
BIG		0						
LITTLE		0						
HUGE		0						
TINY		0						
ROUND		0						
<i>SHAZAM</i>								
	X	ZERO	BIG	LITTLE	HUGE	TINY	ROUND	MISS
X								
ZERO	0	1						
BIG	1.129	0	1.277					0
LITTLE	1.001	0	1.137	1.013				0
HUGE		0	1.30	1.001				0
TINY		0	1.30	1.001				0
ROUND		0	1.30	1.001				0
MISS	0	0						1

Table 6
Results of Test IIF

	E-Views	LIMDEP	RATS	SHAZAM	TSP
Regress BIG on X	p	p	p	p	p

3.2. Wilkinson's Test III. Missing data

Missing values are common in some areas of economics, so it is important to know how they are handled, both in calculations and in logical tests. We draw a distinction between 'handling' missing values and simply excluding all observations.

Test IIIA. Use the data set NASTY on the following transformation:

```
IF MISS = 3
THEN TEST = 1
ELSE TEST = 2
```

TEST should have the value 2 for all cases because MISS does not anywhere equal 3. Another accepted solution is for TEST to be equal to the missing value. Any other

Table 7
Results of Test IIIA

	E-Views	LIMDEP	RATS	SHAZAM	TSP
vectorized	p	p	p	p	NA
do loop	p	p	F	p	p

Table 8
Results of Test IIIB

	E-Views	LIMDEP	RATS	SHAZAM	TSP
vectorized	p	p	p	F	NA
do loop	p	F	p	F	p

answer implies that the software cannot be used for testing logical comparisons when missing values are present.

If the package does not have an ELSE statement, two consecutive IF statements can be used. We distinguish between the vectorized and do-loop versions of the test, where applicable conducting both. They should give the same answer. Sometimes they do not. Results are presented in Table 7. TSP does not offer a vectorized version and returns TEST = 2 for the loop. RATS returns TEST = 1 for the loop and TEST = 2 for the vectorized.

Test IIIB. Use the data set NASTY on the following calculation:

$$\text{IFMISS} = \langle \text{missing} \rangle \text{ THENMISS} = \text{MISS} + 1$$

The correct answer is $\langle \text{missing} \rangle$, since 1 added to a missing value is still missing. As in the previous test, we distinguish between vectorized and do loop methods. They should give the same answer, but sometimes they do not. Results are presented in Table 8. For the loop, LIMDEP returns MISS = -998. SHAZAM returns MISS = -99998 for both.

3.3. Wilkinson's Test IV. Regression

Test IVA. Using the variable X , compute $X1 = X$, $X2 = X^2$, $X3 = X^3$, ..., $X9 = X^9$. Regress $X1$ on a constant and $X2$ through $X9$. The coefficients, to three significant digits, are: 0.353, 1.14, -0.705, 0.262, -0.0616, 0.00920, -0.000847, 0.0000438, -0.000000974. Since this test is bound to stress the machinery, what is important is not the coefficients but the overall regression. Since this gives a perfect fit, R^2 should be unity. As an added check, the sum of squared residuals should be close to zero as should the integrated squared error evaluated for the estimated coefficients. Results are presented in Table 9. E-Views can perform the the regression only up to $X7$, refusing to compute for the eighth and ninth degree polynomials. Since it refuses to compute rather than return an incorrect answer, this is a miserable failure rather than a catastrophic failure.

Table 9
Results of Test IVA, †only calculated up to X7

	E-Views	LIMDEP	RATS	SHAZAM	TSP
polynomial regression	†	p	p	p	p

Table 10
Results of Test IVB

	E-Views	LIMDEP	RATS	SHAZAM	TSP
regress X on X	p	p	p	p	p

Table 11
Results of Test IVC

	E-Views	LIMDEP	RATS	SHAZAM	TSP
regress X on BIG, LITTLE	p	p	F	p	p

Table 12
Results of Test IVD

	E-Views	LIMDEP	RATS	SHAZAM	TSP
regress ZERO on X	p	p	p	p	p

Test IVB. Regress X on a constant and X. The constant should be exactly zero and the regression coefficient should be unity. Results are summarized in Table 10: all packages pass.

Test IVC. Regress X on a constant, BIG, and LITTLE. The program should inform the user that this is a singular regression. Results are presented in Table 11. RATS returns coefficients and does not warn about the singularity, which is a catastrophic failure.

Test IVD. Regress ZERO on a constant and X. The program should inform the user that ZERO has no variance or should report both the correlation and sum of squares to be zero. Results are presented in Table 12. All packages pass.

4. Conclusions

Wilkinson's Tests have been applied to five econometric packages, uncovering flaws in all five. These flaws include dropping points from a graph, incorrect calculation of the sample variance, correlation coefficients in excess of unity, and incorrect and inconsistent handling of missing values. These econometrics packages fared about as well as did the statistics packages examined by Sawitzki [11]. Some of these statistics packages improved their performance in subsequent versions, and the same can be hoped for these econometric packages. All the packages assessed in this article have since released new versions. It will be interesting to see whether the developers have fixed known errors.

Table 13
Results of Test IIA

	E-Views	LIMDEP	RATS	SHAZAM	TSP
Y1,Y2,Y3	18,0,0	18,0,0	18,0,0	18,0,0	18,0,0

Appendix

Wilkinson's Test IIA continues in the following fashion. As another test of consistent rounding, note that $\sqrt{2}\sqrt{2} = 2$, and $\exp[\ln(2)] = 2$. Compute the following scalars where INT is the greatest integer function (converts reals to integers by throwing away the decimals), and LOG is the natural logarithm. Wilkinson writes that if a statistical package fails this test, "you cannot trust it to compute any function accurately".

$$\begin{aligned}
 Y1 &= \text{INT}(2.6*7 - 0.2) && \Rightarrow && \text{INT}(18.0) && \Rightarrow && 18 \\
 Y2 &= 2 - \text{INT}(\text{EXP}(\text{LOG}(\text{SQRT}(2)*\text{SQRT}(2)))) && \Rightarrow && 2 - \text{INT}(2.0) && \Rightarrow && 0 \\
 Y3 &= \text{INT}(3 - \text{EXP}(\text{LOG}(\text{SQRT}(2)*\text{SQRT}(2)))) && \Rightarrow && \text{INT}(3.0 - 2.0) && \Rightarrow && 1
 \end{aligned}$$

Results of this test are presented in Table 13.

According to Wilkinson, none of these packages can be trusted to perform any calculation because all return 0 instead of 1 for Y3. To the contrary, 0 is the correct answer for Y3, and Wilkinson can be excused for this mistake.

Wilkinson's *Statistics Quiz* was published in 1985, the same year that the IEEE-754 standard for computer arithmetic was released, so Wilkinson cannot be faulted for contradicting IEEE-754. In fact, the results displayed in Table 13 are perfectly consistent with IEEE-754.

To see this, recall that in 32-bit double precision the computer has about 16 digits with which to represent any number. In the case of taking square roots that are not exactly representable in 16 digits, the last digit must be rounded. The square root of 2, being an irrational number, cannot be exactly represented in sixteen digits. So the square root of 2 is 1.414213562373095 to sixteen digits, the next four digits are 0488. Let a = the sixteen digit square root of 2. Then a^2 equals not 2 but $2 + 4.44089\text{E-}16$.

Similarly, $\text{EXP}(\text{LOG}(\text{SQRT}(2)*\text{SQRT}(2)))$ becomes $\text{EXP}(\text{LOG}(2 + 4.44089\text{E-}16))$ which equals something slightly greater than 2, for simplicity call it $2 + \epsilon$.

For Y2, the integer part of $(2 + \epsilon)$ is 2 which, subtracted from 2, yields zero. For Y3, $3 - (2 + \epsilon)$ becomes $1 - \epsilon$ which, when the integer part is taken, becomes zero. Thus, the five packages considered all produce answers that are in accordance with IEEE-754. Hence, rather than indicate a problem with the software, this test suggests that the user be aware of the limitations of finite precision computation.

Acknowledgements

Thanks for comments to J. Doornik, P. Hollinger, and conference participants of the 1999 Boston meeting of the Society for Computational Economics, as well as seminar

participants at various bureaus of the Federal Communications Commission and the software developers. Finally, thanks to M. Cornea for an illuminating discussion of IEEE-754 square root calculations.

References

- [1] U. Bankhofer and A. Hilbert, An Application of Two-Mode Classification to Analyze the Statistical Software Market, in: *Classification and Knowledge Organisation*, R. Klar and O. Opitz, eds, Springer, Heidelberg, 1997, pp. 567–572.
- [2] U. Bankhofer and A. Hilbert, Statistical Software Packages for Windows: A Market Survey, *Statistical Papers* **38** (1997), 393–407.
- [3] T.F. Chan, G.H. Golub and R.J. Leveque, Algorithms for Computing the Sample Variance: Analysis and Recommendations, *American Statistician* **37** (1983), 242–247.
- [4] R.F. Ling, Comparison of Several Algorithms for Computing Sample Means and Variances, *Journal of the American Statistical Association* **69** (1974), 859–866.
- [5] C.R. MacKenzie, MicroFit 4.0, *Journal of Applied Econometrics* **13** (1998), 77–89.
- [6] J. MacKie-Mason, Econometric Software: A User's View, *Journal of Economic Perspectives* **6** (1992), 165–188.
- [7] B.D. McCullough, Benchmarking Numerical Accuracy: A Review of RATS v4.2, *Journal of Applied Econometrics* **12** (1997), 181–190.
- [8] B.D. McCullough, Econometric Software Reliability: EViews, LIMDEP, SHAZAM, and TSP, *Journal of Applied Econometrics* **12** (1997), 191–202.
- [9] W. Murray, Failure, the Causes and Cures, in: *Numerical Methods for Unconstrained Optimization*, W. Murray, ed., Academic Press, NY, 1972, pp. 107–122.
- [10] G. Sawitzki, Testing Numerical Reliability of Data Analysis Systems, *Computational Statistics and Data Analysis* **18** (1994), 269–286.
- [11] G. Sawitzki, Report on the Numerical Reliability of Data Analysis Systems, *Computational Statistics and Data Analysis* **18** (1994), 289–301.
- [12] R.A. Thisted, *Elements of Statistical Computing*, Chapman and Hall, New York, 1988.
- [13] M.R. Veall, SHAZAM 6.2: A Review, *Journal of Applied Econometrics* **6** (1991), 317–320.
- [14] H.D. Vinod, A Review of SORITEC 6.2, *American Statistician* **43** (1989), 266–269.
- [15] H.D. Vinod, Review of GAUSS for Windows Including Its Numerical Accuracy, *Journal of Applied Econometrics* **15** (2000), 211–220.
- [16] L. Wilkinson, *Statistics Quiz*, SYSTAT, Evanston, IL, 1985.

Copyright of Journal of Economic & Social Measurement is the property of IOS Press and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.