Wilkinson's Tests and Econometric Software

B. D. McCullough[1]
Federal Communications Commission
445 12th St. SW, Room 2C-134
Washington, DC 20554
INTERNET: bmccullo@fcc.gov

[1]Thanks for comments to seminar participants at various bureaus of the Federal Communications Commission, and to the developers for correcting some oversights. The views expressed are those of the author and not necessarily those of the Commission.

ABSTRACT

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.

key words: 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 (1989), Veall (1991), McCullough (1997), and McKenzie (1998) are exceptions.

In part, this lack of benchmarking during reviews may be attributed to the fact that, until quite recently (Rogers, et al, 1998), no one ever collected the various benchmarks in a single place. 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 (1985) *Statistics Quiz: Problems which reveal deficiencies in statistical programs*, which is discussed in detail in Sawitzki (1994a). These tests

have been profitably employed by Sawitzki (1994b), who uncovered errors in SAS, SPSS, and S-PLUS, among other packages, and also by Bankhofer and Hilbert (1996a, 1996b). 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 threefold. 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 which claim to be general-purpose, and not suited to specialized econometric packages; 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 which MacKie-Mason (1992) evaluated for user-friendliness: E-Views v3.0, LIMDEP v7.0 for Windows 95, RATS v4.3, SHAZAM v8.0, and TSP v4.4.

2 The Data

Table 1 displays the data set "Nasty," whose values are not unreasonable. The values for BIG are less than the U.S. population, while the values of HUGE are the same order of magnitude as the national debt.

LABEL$	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 1: Data Set NASTY.DAT

3 The Tests

II. Real Numbers

TEST IIA.

Print ROUND with only one digit. Note, 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. 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.

As another test of consistent rounding, recall 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.

Y1 = INT(2.6*7 - 0.2)	$\Rightarrow$	INT(18.0)	$\Rightarrow$	18
Y2 = 2-INT(EXP(LOG(SQRT(2)*SQRT(2))))	$\Rightarrow$	2-INT(2.0)	$\Rightarrow$	0
Y3 = INT(3-EXP(LOG(SQRT(2)*SQRT(2))))	$\Rightarrow$	INT(3.0-2.0)	$\Rightarrow$	1

If a package returns 0 for Y2, this suggests that the program evaluates $\exp[\ln(\sqrt{2}\sqrt{2})] = 2.0$. A program which consistently makes this evaluation will return 1 for Y3. A program which returns 0 for Y2 and 0 for Y3 is a program for which $\exp[\ln(\sqrt{2}\sqrt{2})]$ sometimes equals 2 and sometimes doesn't. Results of Test IIA are presented in Table 2.

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

Table 2: Results of Test IIA

While all programs pass the rounding test, each program inconsistently evaluates arithmetic expressions. Besides the obvious effects on direct calculations of inconsistent rounding, such a program cannot be trusted for logical comparisons. To see this, simply consider

```
IF (EXP(LOG(SQRT(x)*SQRT(x)).EQ.x) THEN
  A=B
ELSE
  A=C
ENDIF
```

Will A be set equal to B or to C? In fact, for many packages which inconsistently round numbers, the answer will depend on the particular value of x.

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 somtimes be surprising, as indicated by Figure 1, which shows the correct graph as produced by RATS and the graph produced by E-Views. 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, without warning.

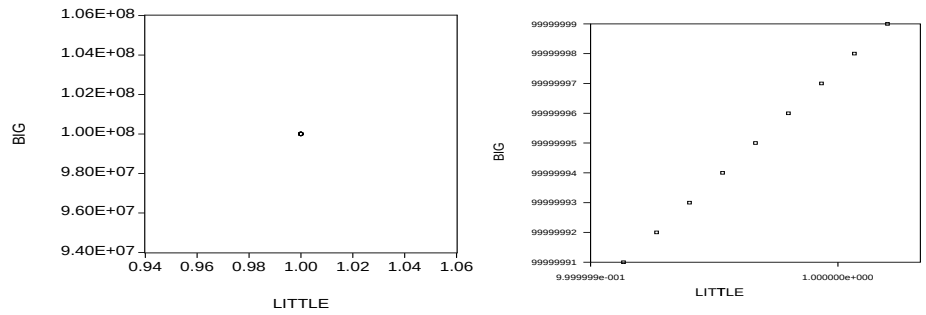

Figure 1: Test IIB Results for E-Views (left) and RATS (right).

Plot X against ZERO. The answer should be a vertical line. Some packages, such as LIMDEP, fail this test because they are unable to scale the horizontal axis for these data.

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

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

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 results 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.

	E-Views	LIMDEP	RATS	SHAZAM	TSP
X	2.582	p	p	p	p
ZERO	†	p	p	p	0
MISS	†	†	†	0	†
BIG	2.285	p	p	2.424	p
LITTLE	2.701	p	p	2.870	p
HUGE	2.582	p	p	p	p
TINY	2.582	p	p	p	p
ROUND	2.582	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 for variable MISS

For MISS, SHAZAM returned zero, and the others refused to perform the calculation. E-Views has a tendency to calculate 2.582 instead of the correct 2.738, which would be the case if the number of observations, n, was used in the denominator, rather than $n - 1$. However, the E-Views manuals makes no mention of maximum likelihood calculation for this statistic. The E-Views and SHAZAM results for BIG and LITTLE can be explained by their algorithm for computing the variance. Ling (1974) and Chan, Golub, and Leveque (198?) 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)

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

wherease the usual formula is

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

Forumula 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.* Thisted, 1988) use Formula 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 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

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

where $cov(w,z)$ is the covariance between w and z and and σ_w is the standard deviation of w. Since the standard deviation of ZERO is zero, (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,

	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			

E-Views

	X	ZERO	BIG	LITTLE	HUGE	TINY	ROUND
X							
ZERO	0	0					
BIG		0					
LITTLE		0					
HUGE		0					
TINY		0					
ROUND		0					

LIMDEP, RATS, TSP

	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

SHAZAM

Table 5: Results of Test IID
Only incorrect results are displayed.

which is theoretically impossible, and 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.

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

Table 6: Results of Test IIF

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 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. TSP returns TEST=<missing> for the vectorized and TEST=2 for the loop. For RATS, the IF command is not supposed to be used with series (though this is not obvious from the documentation and will produce an answer). RATS returns TEST=1 for the loop.

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

Table 7: Results of Test IIIA

Test IIIB.

Use the data set NASTY on the following calculation:

```
IF MISS = <missing> THEN MISS = MISS + 1
```

The correct answer is <missing>, 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 again they often do not. For the loop, E-Views returns MISS = 1 while LIMDEP returns MISS = -998. RATS produces a fatal error for the loop, and SHAZAM twice returns MISS = -99998.

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

Table 8: Results of Test IIIB
†exited program when calculation attempted
‡program does not offer vectorized operation

IV. Regression

TEST IVA.

Using the variable X, compute X1 = X, X2 = X^2, X3 = $X^3, \ldots,$ 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	LIMDEP	RATS	SHAZAM	TSP
F†	p	p	p	p

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

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.

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

Table 10: Results of Test IVB

TEST IVC.

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

TEST IVD.

Regress ZERO on a constant and X. The program should inform you that

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

Table 11: Results of Test IVC

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.

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

Table 12: Results of Test IVD

4 Conclusions

Wilkinson's (1985) 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 (1994b). Some of these statistics packages improved their performance in subsequent versions, and the same can be hoped for for these econometric packages.

References

Bankhofer, U. and A. Hilbert (1997), "An Application of Two-Mode Classification to Analze the Statistical Software Market," in *Classification and Knowledge Organisation*, R. Klar and O. Opitz, eds., Springer: Heidelberg, 567-572

Bankhofer, U. and A. Hilbert (1997), "Statistical Software Packages for Windows: A Market Survey" *Statistical Papers*, **38**, 393-407

Chan, T. F., G. H. Golub and R. J. Leveque (1983), "Algorithms for Computing the Sample Variance: Analysis and Recommendations," *American*

Statistician, **37**, 242-247

Ling, R. F. (1974), "Comparison of Several Algorithms for Computing Sample Means and Variances," *Journal of the American Statistical Association*, **69**, 859-866

MacKenzie, C. R. (1998), "MicroFit 4.0," *Journal of Applied Econometrics*, **13**, 77-89

MacKie-Mason, Jeffrey K. "Econometric Software: A User's View," *Journal of Economic Perspectives*, Fall 1992, **6**(4), pp. 165-188.

McCullough, B. D. (1997), "Benchmarking Numerical Accuracy: A Review of RATS v4.2," *Journal of Applied Econometrics*, **12**, 181-190

McCullough, B. D. (1999), "Econometric Software Reliability:EViews, LIMDEP, SHAZAM, and TSP," *Journal of Applied Econometrics*, forthcoming

Sawitzki, Günter (1994), "Testing Numerical Reliability of Data Analysis Systems," *Computational Statistics and Data Analysis* **18**, 269-286

Sawitzki, Günther (1994), "Report on the Numerical Reliability of Data Analysis Systems," *Computational Statistics and Data Analysis* **18**, 289-301

Thisted, R. A. (1988), *Elements of Statistical Computing*, New York: Chapman and Hall

Veall, M. R. (1991), "SHAZAM 6.2: A Review" *Journal of Applied Econometrics*, **6**, 317-320

Vinod, H. D. (1989), "A Review of SORITEC 6.2" *American Statistician*, **43**, 266-269

Wilkinson, Leland (1985), *Statistics Quiz*, SYSTAT: Evanston, IL

Wilkinson, Leland (1994), "Practical Guidelines for Testing Statistical Software," in *Computational Statistics*, P. Dirschedl and R. Ostermann, eds., Physica-Verlag: Berlin, pp. 111-124