Wilkinson Tests and gretl

A. Talha Yalta[1] and A. Yasemin Yalta[2]

[1] TOBB University of Economics and Technology
Sogutozu Caddesi No:43; Sogutozu, 06560, Ankara, Turkey
yalta@etu.edu.tr
[2] Hacettepe University, Turkey

Abstract. Applied econometrics has become fully dependent on computers and software tools. It is therefore important that the reliability of various programs providing econometric functionality is vetted within the profession. Here, we report on the results of our verifying the accuracy of Gretl (GNU Regression, Econometrics and Time-series Library) using the Wilkinson tests. Our study was important in the implementation of a number of modifications improving the general accuracy and reliability of this open source econometric package.

Key words: Gretl, econometric software, accuracy testing, open source

1 Introduction

Like science itself, scientific software is a work in progress and it is possible that any such program at any given time contains errors and imperfections. In the case of econometrics, this is well documented by various authors such as Sawitzki [1], McCullough [2, 3] and Yalta [4], who find important flaws and inconsistencies in the programs widely used within the profession. Because it is nearly impossible to test all of the functionality offered by a typical econometric program, such studies generally employ an introductory or an intermediary test suite such as Wilkinson's 1985 "Statistics Quiz" [5], the Statistical Reference Datasets (StRD) by the U.S. National Institute of Standards and Technology (NIST) or McCullough's set of tests [6]. These procedures are based on comparing the output from a sample of econometric functions against the corresponding correct answers or benchmarked values.

Gretl (GNU Regression, Econometrics and Time-series Library) is a sophisticated and cross-platform program for econometric analysis.[3] It is open source and can be freely used, modified and redistributed under the terms of the GNU General Public License (GPLv3). The program has been gaining in popularity

[3] We first became familiar with Gretl several years ago. Although we never got involved in the coding process of the program, we made contributions in the form of testing the numerical accuracy of its various functions, submitting bug reports, and helping its internationalization efforts.

in the recent years and according to the project's web host SourceForge.net, it was downloaded more than 100,000 times in 2008.[4] See Baiocchi and Distaso [8], Mixon and Smith [9], Yalta and Yalta [10], and Rosenblad [11] for reviews of Gretl versions 0.997, 1.51, 1.6.0, and 1.7.3 respectively.

An important tool offered by Gretl but unavailable in most other statistical packages is the StRD linear regression test suite, which automatically assesses the regression results through a series of 11 tests using the reference data sets compiled by the NIST. This function, which is readily available from the "Tools" menu, helps make sure that a given installation of Gretl produces the certified results, thereby increasing reliability. The so called "Wilkinson Tests" is an alternative entry-level testing procedure also useful for assessing econometric software. Just like the StRD, Wilkinson's method has been widely used for testing different software packages. As a result, the objective of this paper is to describe in detail and report on our experience while applying this procedure, which is currently not available in Gretl.

In the next section, we discuss the Wilkinson tests and their effectiveness in exposing flaws in statistical and econometric programs. In section 3, we report on the various accuracy errors, existence of which we discovered in Gretl versions 1.7.9 and earlier. We also discuss how these errors were fixed following our reporting them to the developers as well as the openness of the whole process, which made it possible to understand the nature of the error and verify its correction directly from the source code. Section 4 concludes.

2 The Wilkinson Tests

The Wilkinson procedure is an entry-level test suite for computational accuracy, which was originally released as a booklet by Wilkinson [5] and described in detail by Sawitzki [12]. The tests are deliberately designed to reveal flaws in statistical software using a small but effective data set NASTY shown in Table 1. As discussed by Wilkinson [13], each column in the table is designed to expose a different type of flaw. For example, ZERO is used for testing the conditions likely to cause various zero divide or singularity errors in computational algorithms. MISS contains all missing values, which are important in some areas of economics. BIG and LITTLE have a significant variation in the eighth digit, making them problematic to analyze using a badly designed program. Together with HUGE and TINY, they are also used for revealing the formatting problems

[4] For the popularity of Gretl, also see the econometric study by Lucchetti [7] (available in this volume) finding that the users of Gretl have been steadily increasing at a yearly rate of 43 percent since 2006.

in various output routines. Finally, ROUND tests how the rounding operation is performed for the purpose of printing.

Although Wilkinson's method is based on an artificial data set, the software defects it is designed to reveal are real as shown by a number of studies such as Sawitzki [1], Bankhofer and Hilbert [14, 15], McCullough [3], and Choi and Kiefer [16]. These studies employ the Wilkinson tests in order to assess the reliability of many statistical and econometric programs, each time exposing deficiencies in fundamental statistical operations such as computing sample standard deviations or graphing. Wilkinson himself argues that the data set is not as extreme as it may seem since, for example, "the values of BIG are less than the U.S. population (and) HUGE has values in the same order as the magnitude of the U.S. deficit." McCullough [17] explains that the procedure has three virtues: First, it is simple and easily applied to most econometric packages. Second, the flaws it is designed to reveal have known solutions so that any program could pass. Third, it questions the functionality that we take for granted such as correctly reading a data file or properly handling the missing values.

The Wilkinson tests are organized in four groups focusing on data management (IA, IB), descriptive statistics (IIA–IIF), missing values (IIIA–IIIC), and linear regression (IVA–IVD) respectively. The first two tests involve reading a custom ASCII data file which includes formatted data likely to be produced by different programs. In the second group of tests, the program first prints ROUND with only one digit. Afterwards, three separate graphs plotting BIG against LITTLE, HUGE against TINY and X against ZERO are produced. This is followed by calculating various summary statistics as well as a correlation matrix and Spearman correlations on all the variables. None of these computations should imply a problem for a well designed program. In Test IIE, X is tabulated against X using BIG as a case weight. This is a strictly statistical procedure not available in most econometric programs including Gretl. Finally, Test IIF involves regressing BIG on X in order to check whether the correct answer BIG=99999990+1X is returned. Test IIIA and IIIB assess the handling of the missing values by running the operations "`IF MISS=3 THEN TEST=1 ELSE TEST=0`" and "`IF MISS=<missing> THEN MISS=MISS+1`" respectively. The answer is 2s or missing values for the first case and all missing values for the second case. Similarly, Test IIIC tabulates MISS against ZERO. The program should return one cell with 9 cases in it. The fourth group of tests first extends the data set by the powers $X1 = X^1, \ldots, X9 = X^9$ and runs a series of four regressions. In Test IVA, X is regressed on X1 through X9. Here, R^2 should be unity since this is a perfect fit. Test IVB regresses X on X and a constant with the obvious solution X=0+1X. This is followed by a regression of X on BIG and LITTLE to test whether the program will warn about the singularity

Table 1. The Data Set NASTY

X	ZERO	MISS	BIG	LITTLE	HUGE	TINY	ROUND
1	0	NA	99999991	0.99999991	1e+012	1e-012	0.5
2	0	NA	99999992	0.99999992	2e+012	2e-012	1.5
3	0	NA	99999993	0.99999993	3e+012	3e-012	2.5
4	0	NA	99999994	0.99999994	4e+012	4e-012	3.5
5	0	NA	99999995	0.99999995	5e+012	5e-012	4.5
6	0	NA	99999996	0.99999996	6e+012	6e-012	5.5
7	0	NA	99999997	0.99999997	7e+012	7e-012	6.5
8	0	NA	99999998	0.99999998	8e+012	8e-012	7.5
9	0	NA	99999999	0.99999999	9e+012	9e-012	8.5

problem. Finally, ZERO is regressed on a constant and X, with the expectation of a warning about ZERO having no variance or a regression output where both the correlation and total sum of squares are given as 0.

3 The Performance of Gretl

During our testing of Gretl, we found a number of errors, which mainly affect the display of the data and the presentation of various computation results.[5] The first problem that we found was in Test IB and was regarding the lack of a formatting function for a "display data" window showing more than one variables. The lack of this functionality has the potential to mislead the users. For example, the default format Gretl uses is printing such values with 6 significant digits. In the case of the Wilkinson data set, this leads to an output wrongly implying that LITTLE is constant for observations 1 to 5, while BIG is constant through 5 to 9. We reported this problem to Gretl developers and within three days there was an update adding the "reformat" function to all data display windows.

The second issue that we encountered was in Test IIA and was due to the fact that the Gretl commands `print`, `printf`, and `sprintf` rounded numerical values using "unbiased rounding" or the "round-to-even" method. As a result, attempting to print ROUND with only one digit returned $\{0, 2, 2, 4, 4, 6, 6, 8, 8\}$ instead of the correct answer of printing the numbers from 1 to 9. This problem was because of not following the principle that rounding for the purpose of printing is not about calculation but about the presentation of the results. Gretl developers acknowledged this issue and corrected it within just four days.

Thirdly, in Test IIB, Gretl failed to plot BIG against LITTLE accurately and also refused to plot X against ZERO returning an error message. These were

[5] This report is abbreviated. See Yalta [18] for a more detailed discussion.

due to a number of communication problems with the Gnuplot program, which handle the plotting functionality of Gretl. Within six days of our reporting of the errors, the Gretl source files handling the graphing behavior were revised. As can be seen in Figure 1 below, Gretl now correctly shows a 45 degree line and a vertical line for the two plots.

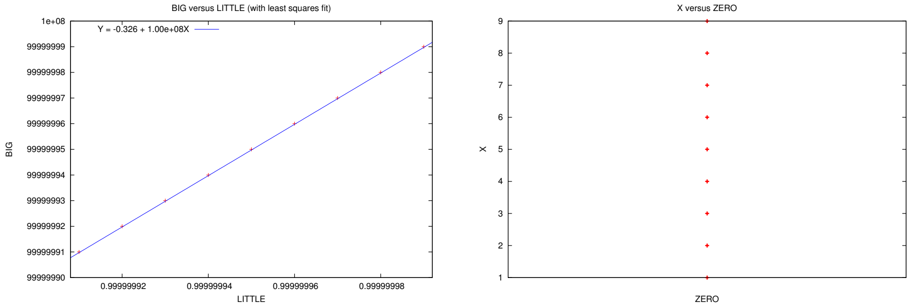

Fig. 1. Gretl's "BIG versus LITTLE" and "X versus ZERO" Plots After Corrections

A fourth error that we came across was while attempting to compute the various descriptive statistics for the eight series in Test IIC. Gretl performed these calculations accurately, however, printed out after rounding the standard deviation of TINY as effectively 0 instead of the correct value 2.7386e+012. This error was also fixed within 24 hours of our reporting it.

Finally, we discovered in Test IID that Gretl produced an erratic output for the calculation of Spearman's rank correlations between ZERO and the other variables. This problem, which was due to numerically printing the special "Not Available DouBLe" (NADBL) value was corrected by the Gretl developers in a matter of 24 hours as well. Gretl passed the remaining 10 tests successfully by computing the results accurately and showing the correct behavior as required by these tests.

In addition to a number of worthwhile improvements, an important benefit of our testing Gretl using the Wilkinson tests was being able to observe how the access to the programming code made it possible to see the cause, facilitated a rapid fix and enabled the verification of the various corrections. Table 2 shows some revision details on the various Gretl source files updated within just a few days after our exposing of the various software defects. It is possible to examine all these changes using SourceForge's "viewvc" interface available at `http://gretl.cvs.sourceforge.net/viewvc/gretl/`.

Table 2. Some Revision Details on the Updated Gretl Source Files

TEST	SOURCE FILE	REV.	DATE	TEST	SOURCE FILE	REV.	DATE
IB	lib/src/printout.c	1.375	Dec 20	IIA	lib/src/printscan.c	1.24	Dec 21
IB	gui2/series view.c	1.44	Dec 18	IIA	lib/src/printscan.c	1.25	Dec 22
IB	gui2/series view.c	1.45	Dec 19	IIB	lib/src/graphing.c	1.409	Dec 19
IB	gui2/series view.c	1.46	Dec 19	IIB	lib/src/graphing.c	1.410	Dec 23
IB	gui2/series view.c	1.47	Dec 20	IIB	lib/src/graphing.c	1.412	Dec 23
IB	gui2/series view.c	1.48	Dec 21	IIB	lib/src/graphing.c	1.413	Dec 23
IB	gui2/series view.c	1.49	Dec 22	IIB	lib/src/graphing.c	1.414	Dec 24
IIA	lib/src/printout.c	1.372	Dec 19	IIB	lib/src/graphing.c	1.415	Dec 24
IIA	lib/src/printout.c	1.373	Dec 19	IIB	lib/src/graphing.c	1.416	Dec 24
IIA	lib/src/printout.c	1.375	Dec 20	IIB	lib/src/plotspec.c	1.42	Dec 24
IIA	lib/src/printout.c	1.376	Dec 22	IIB	lib/src/plotspec.c	1.43	Dec 24
IIA	lib/src/printout.c	1.377	Dec 22	IIB	lib/src/gretl_matrix.c	1.393	Dec 23
IIA	lib/src/printout.c	1.378	Dec 22	IIC	lib/src/gretl_matrix.c	1.388	Dec 19
IIA	lib/src/printscan.c	1.21	Dec 19	IID	lib/src/gretl_matrix.c	1.392	Dec 23
IIA	lib/src/printscan.c	1.22	Dec 19	IID	lib/src/graphing.c	1.411	Dec 23
IIA	lib/src/printscan.c	1.23	Dec 20				

4 Final Thoughts

It is not extraordinary or uncommon that a complex econometric program such as Gretl contains errors and imperfections. The important issue is the mechanism through which such problems are addressed by the developers. Earlier studies on the reliability of econometric packages show that software vendors are unequal in their attention to computational accuracy. Sawitzki [1] ran the Wilkinson tests on nine different commercial packages, found a number of errors in all them and reported that the reaction he received from different vendors varied ranging between “cooperative concern and rude comments.” Yalta [4] found that the various numerical issues in the GAUSS software package reported by Knusel [19, 20] and later by Vinod [21] were not fully fixed in seven years and after several major revisions. Yalta and Jenal [22] report that Addinsoft did not fix the grossly erroneous least squares estimator for ARIMA in the XLSTAT statistical program and let the users obtain invalid results by using a defective function. Microsoft Excel is widely used in the field of economics and McCullough and Heiser [23] discuss that errors found in Excel97 were still either not fixed or wrongly fixed in Excel2007. On the other hand, there also exist studies such as Zeileis and Kleiber [24], Keeling and Pavur [25] and McKenzie and Takaoka [26] which report correction of errors or more accurate results in comparison to earlier versions of various econometric packages.

A question worth investigating is whether the transparent and collaborative nature of the open source development model provides some advantages in the process of error correction, resulting in better and more reliable software in a scientific setting. Indeed, McCullough [27] finds that the open source Gnumeric spreadsheet program fixed within a few weeks all the reported flaws surprisingly similar to those found in Excel. Similarly, Kuan [28] examines three pairs of commercial and open source programs and reports generally faster fixing of bugs in the latter. Our experience applying the Wilkinson tests on Gretl was concurrent with these cursory studies. We observed the correction of all the reported flaws after just six days. This is a remarkable performance considering the fact that the developers do not receive any monetary compensation for their contributions to the program. In addition, here it was also possible for us to access the source code and see the exact cause of the problem each time we discovered an error. Furthermore, the open source nature of Gretl also enabled an instant dissemination of the various fixes and enabled our verifying the correction of the errors.

In conclusion, our assessment of Gretl using the Wilkinson tests allowed the detection of several important flaws and resulted in a number of worthwhile revisions. Also, it is our understanding that the availability of the programming code and the absence of commercial concerns can provide an open source econometrics package such as Gretl an advantage in the reliability department.

Acknowledgments

We wish to thank Allin Cottrell for providing prompt and useful answers for our reports and inquiries regarding Gretl and its inner workings.

Bibliography

[1] Sawitzki, G.: Report on the numerical reliability of data analysis systems. Computational Statistics and Data Analysis **18** (1994) 289–301

[2] McCullough, B.D.: Assessing the reliability of statistical software: Part II. American Statistician **53** (1999) 149–159

[3] McCullough, B.D.: Wilkinson's tests and econometric software. Journal of Economic and Social Measurement **29** (2004) 261–270

[4] Yalta, A.T.: The numerical reliability of gauss 8.0. The American Statistician **61** (2007) 262–268

[5] Wilkinson, L.: Statistics Quiz. 1 edn. SYSTAT, Evanston, IL (1985)

[6] McCullough, B.D.: Assessing the reliability of statistical software: Part I. American Statistician **52** (1998) 358–366

[7] Lucchetti, R.: Who uses gretl? an analysis of the SourceForge download data. In: Proceedings of the 1st Gretl Conference, Bilbao, Spain (Forthcoming)

[8] Baiocchi, G., Distaso, W.: GRETL: Econometric software for the GNU generation. Journal of Applied Econometrics **18** (2003) 105–110

[9] Mixon, J.W., Smith, R.J.: Teaching undergraduate econometrics with GRETL. Journal of Applied Econometrics **21** (2006) 1103–1107

[10] Yalta, A.T., Yalta, A.Y.: GRETL 1.6.0 and its numerical accuracy. Journal of Applied Econometrics **22** (2007) 849–854

[11] Rosenblad, A.: Gretl 1.7.3. Journal of Statistical Software **25** (2008) 19

[12] Sawitzki, G.: Testing numerical reliability of data analysis systems. Computational Statistics and Data Analysis **18** (1994) 269–286

[13] Wilkinson, L.: Practical guidelines for testing statistical software. In Dirschedl, P., Ostermann, R., eds.: Computational Statistics, 25th Conference on Statistical Computing at Schloss Reisenburg, Physica, Verlag (1994)

[14] Bankhofer, U., Hilbert, A.: Statistical software packages for windows - a market survey. Statistical Papers **38** (1997) 377–471

[15] Bankhofer, U., Hilbert, A.: An Application of Two-Mode Classification to Analyze the Statistical Software Market. In: Klar, R., Opitz, O., eds.: Classification and Knowledge Organisation. Springer, Heidelberg (1997) 567–572

[16] Choi, H.S., Kiefer, N.M.: Software evaluation: EasyReg international. International Journal of Forecasting **21** (2005) 609–616

[17] McCullough, B.D.: The Accuracy of Econometric Software. In: Belsley, Kontoghiorghes eds.: Handbook of Computational Econometrics. Wiley (to appear)

[18] Yalta, A.T.: Should economists use open source software for doing research? (2009) [Unpublished Manuscript].

[19] Knüsel, L.: On the accuracy of the statistical distributions in GAUSS. Computational Statistics and Data Analysis **20** (1995) 699–702

[20] Knüsel, L.: Telegrams. Computational Statistics and Data Analysis **21** (1996) 116

[21] Vinod, H.D.: Review of GAUSS for Windows, including its numerical accuracy. Journal of Applied Econometrics **15** (2000) 211–220

[22] Yalta, A.T., Jenal, O.: On the importance of verifying forecasting results. International Journal of Forecasting **25** (2009) forthcoming

[23] McCullough, B.D., Heiser, D.A.: On the accuracy of statistical procedures in Microsoft Excel 2007. Computational Statistics and Data Analysis **52** (2008) forthcoming

[24] Zeileis, A., Kleiber, C.: Validating multiple structural change models – a case study. Journal of Applied Econometrics **20** (2005) 685–690

[25] Keeling, K.B., Pavur, R.J.: A comparative study of the reliability of nine statistical software packages. Computational Statistics and Data Analysis **51** (2007) 3811 – 3831

[26] McKenzie, C.R., Takaoka, S.: Eviews 5.1. Journal of Applied Econometrics **22** (2007) 1145–1152

[27] McCullough, B.D.: Fixing statistical errors in spreadsheet software: The cases of Gnumeric and Excel (2004) [CSDA Statistical Software Newsletter; retrieved December 10, 2008].

[28] Kuan, J.: Open source software as consumer integration into production (2001) [Unpublished Working Paper, Stanford University].