Assessing the reliability of web-based statistical software

A. M. Kitchen, R. Drachenberg, J. Symanzik

Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, UT 84322, U.S.A.

Summary

Statistical reference datasets from the National Institute of Standards and Technology were used to evaluate the accuracy and precision of two web-based statistical packages, WebStat (found at http://www.stat.sc.edu/webstat/) and Statlets (found at http://www.statlets.com/statletsindex.htm). This evaluation revealed that both packages performed reasonably well in the analysis of lower difficulty datasets, with decreasing accuracy as difficulty increases. The decrease in accuracy for datasets of higher difficulty can often be contributed to the dataset storage format (which seems to be single precision in Statlets and is double precision in WebStat). For most statistical analysis needs, the level of accuracy found in WebStat and Statlets would likely be unsatisfactory. Limitations of the packages are discussed and comparisons of accuracy with commercially available statistical packages are presented. WebStat and Statlets are user-friendly packages that are amenable to use as teaching tools. Thus we advise that the statistics packages we evaluated be restricted to use in teaching situations. They should not be used for the analysis of datasets with higher difficulty levels.

Keywords: Numerical performance, Statistical reference datasets, Statlets, StRD, WebStat.

I. Introduction

The reliability of commercially available statistical software packages such as SAS, SPSS, S-Plus, Microsoft Excel, and Mathematica 4 has been previously assessed (McCullough 1998, McCullough 1999a, McCullough and Wilson 1999, McCullough 2000). These evaluations have been carried out because it is essential to be aware of the level of accuracy and precision associated with numerical output from a statistical package. These tests have revealed mixed competencies of various statistical packages in entry level and intermediate level tests in three areas: estimation, random number generation, and statistical distributions (e.g. Sawitzki 1994a, b; McCullough 1999a, McCullough and Wilson 1999). As software packages often do not disclose the algorithms or methods of implementation used in procedures, the reliability of software packages is frequently not readily apparent from information within the package. Thus a comparison of output to benchmark results has been the preferred method used to evaluate the accuracy and precision of software packages. In this paper we evaluate the accuracy and precision of two web-based packages, WebStat and Statlets. While these web-based packages are most commonly used for teaching purposes, students that are exposed to a particular software package in the classroom are likely to continue to use this package. Thus, a comparison of the results of web-based packages to the results of commercial packages is warranted.

The Statistical Reference Datasets Project was developed by the Statistical Engineering Division and the Mathematical and Computational Sciences Division within the Information Technology Laboratory of the American National Institute of Standards and Technology (NIST). These datasets can be found at http://www.nist.gov/itl/div898/strd. The statistical reference datasets (StRD) were compiled for the express purpose of facilitating statisticians in evaluating statistical software packages (Rogers et al. 1998), but they have also been used in validation of econometric and spreadsheet software (McCullough 1998, 1999a, 1999b, McCullough and Vinod 1999, McCullough and Wilson 1999, Vinod 2000). There are four areas covered by the StRD: univariate summary statistics, one-way analysis of variance, linear regression, and nonlinear regression. Each area includes problems of lower (l), average (a) and higher difficulty (h). The difficulty level is determined by the sources of inaccuracy: truncation error, cancellation error, and accumulation error. Truncation error relates to the inexact binary representation error in storing decimal numbers. Cancellation error results from the 'stiffness', i.e., the number of constant leading digits in the datasets. Since the total number of arithmetic computations is proportional to the size of a dataset, the accumulation error may increase as the number of observations increases due to the accumulation of small errors.

Cancellation error and accumulation error have been discussed in more detail in Simon and Lesage (1989). Note that most of the ANOVA datasets and

several of the univariate summary statistics datasets from the StRD are based on the construction principle outlined by Simon and Lesage (1989).

The StRD provides 'certified values' to 15 digits for linear procedures and 11 digits for nonlinear procedures. These values were produced using multiple precision computer arithmetic to 500 digits for linear procedures and quadruple precision for nonlinear least squares problems, thus reducing rounding error to a minimum. For an interpretation of results obtained from the use of StRD on PCs where only double precision instead of multiple precision is available, see McCollough (1999c). Gentle (1998) summarizes facts about digital representations of numeric data. According to the IEEE Standard 754, the mantissa of a floating-point number with base 2 must consist of 24 bits for single precision and 53 bits for double precision. Since $\lfloor \log_{10} 2^{24} \rfloor = \lfloor 7.22 \rfloor = 7$ and $\lfloor \log_{10} 2^{53} \rfloor = \lfloor 15.95 \rfloor = 15$, we can correctly store about 7 decimal digits in single precision and about 15 decimal digits in double precision, given that there exists a finite binary representation of these numbers.

Web-based statistical packages such as WebStat (West 1997, West and Ogden 1997, 1998, West et al. 1998), Statlets and XploRe (Kötter, 1997a, b, Müller, 1998, Schmelzer et al. 1996) have been available for a number of years, and are used for statistical data analysis and as teaching tools. They are available to a wide variety of users, including students, teachers, and other data analyzers. Using the "Advanced Search" and "links to this URL" feature available at http://hotbot.lycos.com/ , on April 16, 2001, we found approximately 60 pages that contain links to the WebStat URL http://www.stat.sc.edu/webstat, and more than 400 pages that contain the old WebStat URL http://www.stat.sc.edu/~west/webstat. We found more than 500 pages that contain the Statlets URL http://www.statlets.com.

The accuracy and precision of such packages, however, have not been evaluated so far. In this paper we present an evaluation of two web-based packages, WebStat and Statlets, and compare their accuracy and precision to that obtained when using commercially available statistical software packages. We found that results obtained from the web-based packages were highly variable and often lacking in accuracy and precision. In addition, the packages had various other limitations, including processes involved with data entry and data transformations. We conclude that the packages lack the reliability needed for most statistical applications, and should be limited to teaching applications.

II. Methods

We assessed two web-based statistical packages; WebStat 2.0 (West 1997, West and Ogden 1997, 1998, West et al. 1998), and Statlets (NWP Associates, Inc., Englewood Cliffs, N.J.) by comparing the results of univariate summary statistics, analysis of variance, and linear regression operations with certified values stated in the statistical reference dataset collection (StRD). We did not evaluate random number generation and statistical distributions since these features were not available in WebStat and Statlets when this analysis was conducted. In addition, nonlinear regression analysis was not evaluated because these packages were unable to carry out these procedures. We used Netscape Navigator 4.73 with Java runtime engine version 1.3.1_01 under Microsoft Windows 98 for all our evaluations. While the host computer can affect the least significant bit, depending on the number of bits carried on to the numeric coprocessor, the evaluation of an algorithm is completely dependent on the Java runtime engine that the browser is using.

We used the simplest or most obvious methods within the software to carry out each analysis. For example, to carry out a simple linear regression we used "simple linear regression," rather than using multiple regression with only one independent variable. This was done for the purpose of analyzing the data using the method that would be most commonly used. However, a polynomial regression analysis was conducted on one dataset, *Wampler1*, using both the most obvious and an alternate method in order to compare the results. The most obvious method for polynomial regression involves the use of the polynomial regression option that is available in Statlets (see results in Table 3). Note that this option is limited to a single explanatory variable. For the alternative method, we used a multiple regression analysis by importing columns of x^2 etc. These results are reported in the text. In addition, it is possible to generate powers of variables within the Statlets program, but we chose to calculate the values to full precision in Microsoft Excel and type them in by hand due to errors found in the calculations performed within Statlets.

The mean and standard deviations obtained from the web-based packages when performing univariate summary statistic analyses were compared to the reference dataset results. Results evaluated from one-way analyses of variance involved the F-ratio. Results of linear regression analyses that were evaluated involved coefficients, standard errors, and R-squared values.

The base 10 log relative error (LRE) (McCullough and Wilson 1999) was calculated to compare results obtained from WebStat and Statlets with the StRD result for all procedures. The LRE value is an estimate of the number of correct digits in the result as compared to certified values from the reference dataset. Where x is the value obtained from the statistical package under evaluation, and c is the certified value from the StRD reference datasets, the number of correct digits in x can be calculated by the log relative error as follows:

$$\lambda = - \log_{10}(|x - c| / |c|)$$

Any LRE between zero and one was set to zero following McCullough and Wilson (1999). In cases where there are multiple LRE values for any one analysis (e.g. for λ_{β}, λ_{σ}, λ_{R}), the lowest LRE is reported. The LRE gives us a concise and comparable measure of the accuracy of the statistical packages under review. The precision of the statistical package in terms of number of decimal places reported in the output is also of interest. The precision required from the output of a package will vary upon the application of the analysis, but it is important to note that the level of precision of reported values can affect the accuracy of the results in two ways. If, for example, the certified value is 0.004, and the estimated value is 0.0036 but becomes rounded to 0.004, its accuracy is overstated. Conversely, if the certified value is 0.0024 and the estimated value is 0.0024, a rounding to 0.002 will result in an understatement of the accuracy of the result.

For our evaluation, we rounded the certified StRD results to 6 and 8 digits respectively and compared those rounded results with the results obtained from Statlets and WebStat. For example, for *Mavro* (see Table 1), Statlets reported a mean of 2.00186. The exact result from the StRD is 2.001856. However, this becomes 2.00186 when rounded to 6 significant digits, i.e., the same result as reported by Statlets. Thus, this relates to the maximum LRE of 6.0 for Statlets.

II.1. WebStat 2.0

WebStat is a freely available data analysis software package for use over the World Wide Web (http://www.stat.sc.edu/webstat/). It is written in the form of a Java applet and is designed to run on any of the three major platforms (Mac, PC, Unix). WebStat was created to provide a statistical analysis tool to users not familiar with the more commonly used commercially available statistical analysis packages. These packages often require knowledge of languages such as Splus, SAS, Minitab, etc., which are mainly specific to statisticians. Students and other potential users are commonly not proficient with these languages, and therefore may not be able to use the procedures. By using Java and the World Wide Web, WebStat is amenable for use by a broad range of statistical software users. In addition, WebStat is part of the commercial electronic textbook CyberStats by CyberGnostics, Inc. (http://www.cyberk.com/index.html) and thus will be the first statistical tool many students will work with.

WebStat is equipped with univariate summary statistics, one-way analysis of variance, and simple and multiple linear regression procedures, and these were thus attempted as part of our evaluation. Nonlinear regression analysis procedures are not available in this package.

II.2. Statlets

Statlets is a web-based statistical software package found at http://www.statlets.com/statletsindex.htm. For our analyses, we used a version of Statlets that is available free of charge through the web. This free web-based package serves as a demo for the commercial version that can analyze 20,000 rows by 100 columns. The free version limits the amount of data that can be analyzed to 100 rows by 10 columns. Statlets is available in two forms. It can be accessed as a group of standalone applets or as a menu driven integrated package.

For our purposes, we analyzed the free menu driven version available through the above URL by clicking on "internet access" and then choosing Version 1.1B – March 4, 1998. Then we chose the Menu version because this is more versatile than the standalone applets. A dataset can be loaded into the clipboard and various applets can then perform numerous analyses. Statlets can perform numerical and graphical summary statistics, hypothesis testing, regression analysis, analysis of variance and others.

III. Results

The StRD provides datasets of lower (l), average (a), and higher (h) level of difficulty. For univariate summary statistics and ANOVA tests, the level of difficulty is based on the stiffness, i.e., the number of constant leading digits in a dataset, and the number of observations. As the number of observations increases, the total number of arithmetic operations also increases. This may lead to an accumulation of small errors, i.e., possibly a larger accumulation error. With increasing stiffness, i.e., with more constant leading digits, the accurate computation of standard deviations becomes more difficult as pointed out in Simon and Lesage (1989). In general, as the stiffness and the number of observations increase, so does the difficulty level of a dataset.

Two linear regression datasets that are used for fitting a line through the origin have been assigned an average level of difficulty since NIST encountered several software packages that produced negative R-squared values and incorrect F-statistics for these datasets. Linear regression datasets of higher level of difficulty are multicollinear and therefore mostly test matrix inversion functions.

III.1. WebStat 2.0

We found that WebStat performed reasonably well in the analysis of lower difficulty datasets over all procedures. However, as the difficulty associated with the dataset increased, the accuracy of the results decreased markedly.

Overall, WebStat performed well in univariate summary statistic analyses (under the constraints of maximal 8 decimal digits of output) with a range of LRE values of all variables excluding one falling between 7.2 and 8.0 (see Table 1). Behind the name of each dataset, Table 1 provides the difficulty level (l, a, h) and stiffness, i.e., the number of constant leading digits, of each dataset. A low LRE value of 2.7 was determined for the standard deviation of the dataset *NumAcc4*. This was the only higher difficulty dataset provided in the reference dataset collection for univariate summary statistics. All datasets for univariate summary statistics from the reference dataset collection could be run on WebStat. Obtaining a LRE value of 8.0 for datasets with high stiffness such as *NumAcc3* which has 7 common leading digits suggests that the internal calculations in WebStat are done with double precision. The use of double precision for computations and single precision in the output has been confirmed by West (2002, private communication).

Table 1. LRE values of univariate summary statistic analyses.

Summary Statistics							
		Statlets	**WebStat**	**SAS v6.12***	**SPSS v7.5***	**S-PLUS v4.0***	**Excel 97****
PiDigits (l,0)#	$\lambda_{\bar{x}}$	-	8.0	15	14.7	15	15
	λ_s	-	8.0	15	15	15	15
Lottery (l,0)	$\lambda_{\bar{x}}$	-	7.3	15	15	15	15
	λ_s	-	7.4	15	15	15	15
Lew (l,0)	$\lambda_{\bar{x}}$	-	8.0	15	15	15	15
	λ_s	-	7.2	15	13.2	15	15
Mavro (l,3)	$\lambda_{\bar{x}}$	6.0	8.0	15	15	15	15
	λ_s	6.0	7.4	13.1	12.1	13.1	9.4
Michelso (l,0)	$\lambda_{\bar{x}}$	6.0	8.0	15	15	15	15
	λ_s	6.0	7.6	13.8	12.4	13.8	8.3
NumAcc1 (l,7)	$\lambda_{\bar{x}}$	6.0	8.0	15	15	15	15
	λ_s	0.0	8.0	15	15	15	15
NumAcc2 (a,1)	$\lambda_{\bar{x}}$	-	8.0	14.0	15	14.0	14.0
	λ_s	-	8.0	14.2	15	15	11.6
NumAcc3 (a,7)	$\lambda_{\bar{x}}$	-	8.0	15	15	15	15
	λ_s	-	8.0	9.5	9.5	9.5	1.2
NumAcc4 (h,8)	$\lambda_{\bar{x}}$	-	8.0	14.0	15	14.0	14.0
	λ_s	-	2.7	8.3	8.3	8.3	0

#(difficulty, stiffness)
* McCullough 1999
** McCullough and Wilson 1999

When examining the results of one-way analyses of variance performed in WebStat (see Table 2), we found that there was a high level of variability in the accuracy of the results, with some datasets resulting in very accurate results (a LRE of 8.0 was obtained for six datasets), and others producing results of very low accuracy (e.g. LRE = 0 for *SmLs09*). It should be noted that no negative sums of squares were ever reported.

Table 2. LRE values of analysis of variance analyses.

Analysis of Variance							
		Statlets	**WebStat**	**SAS v6.12***	**SPSS v7.5***	**S-PLUS v4.0***	**Excel 97****
SiRstv (l,3) #	λ_F	3.4	8.0	8.3	9.6	13.3	8.5
SmnLsg01 (l,1)	λ_F	-	8.0	13.3	15	14.5	14.3
SmnLsg02 (l,1)	λ_F	-	8.0	11.4	15	14.3	12.5
SmnLsg03 (l,1)	λ_F	-	8.0	11.8	12.7	12.9	12.6
AtmWtAg (a,7)	λ_F	0	1.9	0	no result	9.7	1.8
SmnLsg04 (a,7)	λ_F	-	8.0	0	0	10.4	1.7
SmnLsg05 (a,7)	λ_F	-	8.0	0	0	10.2	1.1
SmnLsg06 (a,7)	λ_F	-	6.5	0	0	10.2	0
SmnLsg07 (h,13)	λ_F	-	1.1	0	0	4.6	0
SmnLsg08 (h,13)	λ_F	-	1.8	0	0	2.7	0
SmnLsg09 (h,13)	λ_F	-	0	0	0	0	0

#(difficulty, stiffness)

* McCullough 1999

** McCullough and Wilson 1999

WebStat performed reasonably well on linear regression analyses (λ_b = 8.0, λ_σ = 7.7, λ_R = 5.2), again under the constraint that at most 8 decimal digits of output are reported (see Table 3). However, due to the fact that only one dataset (*Norris*) could be run on WebStat, our results in this area are not conclusive. Although multiple linear regression is a feature included in the WebStat package, the feature was not operational; no calculations were made when requested. Thus, only simple linear regression analyses were undertaken. In addition, WebStat results from the datasets *NoInt1* and *NoInt2* were not reported as the analyses required the regression line to go through the origin. This was not an option in the WebStat package.

Table 3. LRE values of linear regression analyses.

Linear Regression							
		Statlets	**WebStat**	**SAS v6.12***	**SPSS v7.5***	**S-PLUS v4.0***	**Excel 97****
Norris (l)	λ_β	6.0	8.0	12.3	12.3	12.5	12.1
simple	λ_σ	6.0	7.7	11.9	10.2	14.1	13.8
	λ_R	5.2	5.2	11.6	9.9	13.8	nr
Pontius (l)	λ_β	0.0	-	11.4	12.5	12.7	11.2
polynomial	λ_σ	0.0	-	9.2	8.9	13.2	14.3
	λ_R	0.0	-	8.9	8.6	12.9	nr
NoInt1 (a)	λ_β	6.0	-	14.7	14.7	14.7	14.7
simple w/o	λ_σ	6.0	-	15	12.5	14.4	15
intercept	λ_R	4.5	-	11.6	12.8	14.0	nr
NoInt2 (a)	λ_β	6.0	-	15	15	15	15
simple w/o	λ_σ	6.0	-	14.9	14.3	15	15
intercept	λ_R	4.3	-	15	13.0	14.9	nr
Filip (h)	λ_β	ns	-	ns	ns	7.1	0
polynomial	λ_σ	ns	-	ns	ns	7.0	0
	λ_R	ns	-	ns	ns	7.8	nr
Longley (h)	λ_β	0.0	-	8.6	12.1	13.0	7.4
multiple	λ_σ	6.0	-	10.3	13.3	14.2	8.6
	λ_R	6.0	-	10.8	13.2	14.1	nr
Wampler1 (h)	λ_β	0.0	-	8.3	6.6	9.8	7.0
polynomial	λ_σ	0.0	-	15	6.6	15	7.2
	λ_R	0.0	-	15	15	15	nr
Wampler2 (h)	λ_β	0.0	-	10.0	9.7	13.5	9.7
polynomial	λ_σ	0.0	-	15	9.7	15	11.8
	λ_R	6.0	-	15	15	15	nr
Wampler3 (h)	λ_β	0.0	-	7.0	7.4	9.2	6.6
polynomial	λ_σ	0.0	-	10.9	10.6	13.5	11.2
	λ_R	0.0	-	10.8	10.8	15	nr
Wampler4 (h)	λ_β	0.0	-	7.0	7.4	7.5	6.6
polynomial	λ_σ	1.9	-	11.5	10.8	13.6	11.2
	λ_R	0.0	-	14.8	14.2	15	nr
Wampler5 (h)	λ_β	0.0	-	7.0	5.8	5.5	6.6
polynomial	λ_σ	0.0	-	11.5	10.8	13.5	11.2
	λ_R	0.0	-	15	15	15	nr

* McCullough 1999 nr = not reported

** McCullough and Wilson 1999 ns = no solution

In addition to these numerical results, there are some noteworthy limitations and problems associated with the WebStat package. The precision of WebStat was restrictive for some analyses as WebStat performs analyses in double precision but the reported output consists of at most 8 decimal digits, i.e., only single precision. This has an adverse effect on accuracy (with a maximum achievable LRE value of 8.0) although the 'acceptable' level of precision will vary from user to user.

There are limitations in the package in procedures associated with the entry of data into the datasheet. WebStat will only insert a complete dataset when copying and pasting datasets of less than approximately 3000 observations. For larger datasets, only a portion of the dataset is inserted into the spreadsheet. To analyze datasets larger than 3000 observations, it is necessary to import the data from a URL. Importing data from a file is also possible but is limited to users employing a HotJava browser.

WebStat performs four forms of data transformations; square, square root, log 10, and natural log. The package computed values accurately up to 8 digits for square root, log 10, and natural log transformations when compared to results of computations completed in S-Plus and Excel, but reported inaccurate values for certain square transformations (see Table 4). These results are not immediately explainable, in particular when double precision is used for internal calculations. In addition, WebStat extended its reported precision of 8 digits for the inaccurate square transformations, but did not extend the reported precision for other transformations. Limitations were also imposed on data entry of some large numbers. For example, the last two entries of x values in Table 4 were entered as 2.0E11 and 2.0E12.

Table 4: Values before and after a square transformation in WebStat.

x	x^2
2	4
20	400
200	40000
2000	4000000
20000	4.0E8
200000	4.0E10
2000000	3.999999983616E12
2.0E7	4.00000001507328E14
2.0E8	4.0000001090256896E16
2.0E9	3.9999999372269978E18
2.0E10	4.0000000801635094E20
1.99999996E11	3.999999911278523E22
1.99999999E12	4.0E24

III.2. Statlets

Statlets performed reasonably well with univariate summary statistics although the freely available version could only run three of the nine datasets (see Table 1). The other six datasets contained more than 100 observations each, so were outside the capabilities of this version. Statlets reports only 6 digits in its output. It accurately returned five or six digits in all but one case. For the dataset *NumAcc1* with 7 common leading digits, Statlets returned a value of zero for the standard deviation instead of one.

Most of the analysis of variance datasets exceeded the size limits of this Statlets version and so could not be run (see Table 2). Statlets can only accept the data for ANOVA's in two column format, and will not run the data if it is in table format. That only left the following two datasets to test: *SiRstv* and *AtmWtAg*. For *AtmWtAg*, Statlets calculated both the sum of squares and the mean sum of squares as zero, thereby not being able to produce an F-ratio (thus this was interpreted as an LRE of 0). The last dataset, *SiRstv*, ultimately the only one analyzed by Statlets, returned 3.4 digits of accuracy for the F-ratio. Here we see the decrease in accuracy that results from the assumed single precision storage used by Statlets as the stiffness of the datasets increases. The results from these two datasets lead to the assumption that Statlets only uses single precision for the internal calculations. For the *SiRstv* dataset with 3 common leading digits, the best we could hope for under single precision is 3 to 4 decimal digits of accuracy. For the *AtmWtAg* dataset with 7 common leading digits, the best we could hope for under single precision is 0 to 1 decimal digits of accuracy. In fact, we get 3.4 and 0 digits of accuracy for these two datasets as expected.

The linear regression analyses in Statlets are the most interesting as all the datasets were within the size limitations of this version. However, this is where the performance varied considerably. There were three simple linear regression datasets and Statlets performed very well with these. Two of them, *NoInt1* and *NoInt2*, were run using simple regression, and forcing the y-intercept to go through the origin. The number of accurate digits returned by Statlets for these three datasets ranged from four through six (see Table 3).

For the multiple regression dataset, *Longley*, Statlets reported a strange result. The constant is reported as –348226, whereas the reference dataset reports it as about –3482258, i.e., the Statlets result is of a different magnitude of order 10, resulting in an LRE of 0. For the coefficients b_1 to b_6, the LRE falls between 5.5 and 6. However, according to the rule of reporting the LRE for the worst result, we get $\lambda_\beta = 0$ for the *Longley* dataset.

The StRD website contains seven polynomial regression datasets. Statlets failed miserably for this procedure matching none of the digits in almost all cases (see Table 3). However, it returned an R-square value (with six digits of accuracy) for the dataset *Wampler2*. Although, when analyzing a statistical package, one should use the simplest procedure available, in this case we chose to try a backdoor approach to see what would happen. For the dataset *Wampler1*, we used Microsoft Excel to calculate all the powers of x, and then typed the data into Statlets by hand. We then ran a multiple regression analysis instead of a polynomial regression. Although the numbers themselves were far closer to the certified values, (e.g., 692.582 is closer to 1 than –61912.7) they still matched zero or near zero of the digits in this dataset.

III.2.1. Rounding error in cut-and–pasting operations in Statlets

A severe limitation with Statlets became evident when analyzing the univariate summary statistics dataset *NumAcc1*. Since the easiest and most obvious method of getting data into the Statlets datasheet is to copy and paste, this is the method we used in all cases. This operation involves the data being converted into a string, transferred, and then converted back to a number. We found that the operation produces errors in Statlets. When pasting in the *NumAcc1* dataset, Statlets rounded the three entries (see Table 5). When we entered the data into Statlets by hand, the entries were not rounded and the standard deviation returned by Statlets was 1.0 as expected.

Table 5. Data entered into Statlets by copy/paste and by hand.

Dataset	Copy/pasted	Hand entered
10000001	1E7	10000001
10000003	1E7	10000003
10000002	1E7	10000002

III.3. Comparisons between WebStat and commercially available statistics packages

In analyzing univariate summary statistics, we found that WebStat performed averagely (see Table 1) in comparison to SAS, SPSS, S-Plus, which all had LRE values for means and standard deviations ranging between 8.3 and 15 (McCullough 1999, McCullough and Wilson 1999). When comparing the results of

WebStat to Excel, we saw that Excel had two very low LRE values; one for a dataset with an average difficulty rating (*NumAcc3*, LRE=1.2) and one for a dataset of higher difficulty (*NumAcc4*, LRE=0). In comparison, WebStat had only one very low LRE value, and this occurred in a dataset of higher difficulty (*NumAcc4*, LRE=2.7) (see Table 1). WebStat and Excel demonstrated less reliability when compared to SAS, SPSS, and S-Plus, with one and two LRE values, respectively, which were below 3.

WebStat showed highly variable results in its performance of one-way analyses of variance. Overall, WebStat's performance is comparable to SAS and SPSS (under the constraint that WebStat only reports 8 significant digits compared to the higher precision output used in the other packages) for datasets of lower difficulty (see Table 2). Surprisingly, WebStat performs considerably better for some of the datasets of average difficulty (LRE of 8.0) than SAS and SPSS (LRE of 0). S-Plus performed better than WebStat with datasets at average levels of difficulty, but was only slightly better with datasets at higher levels of difficulty. WebStat also performed favorably in comparison to Excel, which obtained low LRE values for all datasets at average and higher levels of difficulty.

The results of linear regression analyses when using WebStat are limited as the package could only compute results for one dataset, *Norris*. The results from this dataset were reasonably good. However, all commercially available statistics packages also did well with this dataset (see Table 3).

The main factor in the discrepancy in LRE values between WebStat and the commercial packages is the difference between the 8 digit single precision output of WebStat compared with the higher precision output of SAS, SPSS, S-Plus, and Excel.

III.4. Comparisons between Statlets and commercially available statistics packages

Statlets showed a poor performance (see Table 1) in the analysis of univariate summary statistics, with LRE values of 6.0 for five variables computed, and an LRE of zero for one variable. The fact that Statlets had an LRE of zero for a datset of low difficulty (*NumAcc1*) means that Statlets does not compare favorably to the commercially available statistical software packages (SAS, SPSS, S-Plus, and Excel) which ranged between 8.3 and 15 in all datasets that could be analyzed by Statlets (McCullough 1999, McCullough and Wilson 1999). Again, it should be noted that we used the freely accessible version of Statlets that only supports 100 rows by 10 columns, thus restricting the number of datasets for our evaluation. The assumed use of single precision in Statlets and the restrictive 6 decimal digits of output, compared to the higher precision output used in the commercially available

statistics packages, is an influential factor here resulting in Statlets having a maximum LRE of 6.0.

Results of analyses of variance in Statlets are limited as the majority of the reference datasets exceeded the size limitations of the data sheet in the package. The one dataset that could be analyzed in Statlets performed poorly in comparison with all commercially available packages (see Table 2). Statlets returned a LRE value for the F-ratio of 3.4 while SAS, SPSS, S-Plus, and Excel returned values between 8.3 and 13.3. This result can be explained by an internal single precision format used in Statlets.

For linear regression analysis, the coefficient associated with the lowest LRE was used to compare with the minimums calculated in the same manner for the commercial software. Statlets performed reasonably well for simple linear regression with LRE's ranging from 4.3 to 6.0 (see Table 3); however all the double precision packages did far better with figures from 9.9 to 15. Polynomial regression in Statlets, as previously discussed, did very poorly with LRE values of zero or near zero in almost every category. In one dataset, *Wampler2*, Statlets calculated the R-squared value with perfect accuracy (which seems to be meaningless since the LRE's for the coefficient and standard error are 0); but SAS, SPSS and S-Plus did as well with LRE's equaling 15.

There was one multiple regression dataset, *Longley*. Although Statlets performed well for the R-squared value (LRE = 6) and the standard error (LRE = 6), it did perform poorly on the coefficient values (LRE = 0). Again, there really is no comparison with the double precision packages. They all out-performed Statlets with LRE's ranging from 7.4 to 14.2 in all categories.

III.5. Comparisons between WebStat and Statlets

When WebStat and Statlets were both able to run the same analyses, WebStat performed better than Statlets in all but one case. In this case, (*Norris*, R^2 value), WebStat and Statlets did equally well. One major disadvantage of Statlets clearly is that it only reports six significant digits in its output, thus producing lower LRE values for many of the datasets than WebStat which reports eight significant digits. WebStat was also able to perform more of the analyses because it does not have the size limitations that the freely accessible version of Statlets has. However, Statlets could perform more linear regression analyses than WebStat, although the performance of Statlets on most of the datasets that WebStat could not run was very poor.

IV. Conclusion

The two web-based software packages WebStat and Statlets that we analyzed in this paper were severely limited in accuracy and precision due to the limited output format of 6 (Statlets) and 8 (WebStat) decimal digits and the base dataset storage used which seems to be single precision only for Statlets. For most statistical analysis needs, the level of accuracy found in WebStat and Statlets would likely be unsatisfactory. High levels of variability in the accuracy of the results for any one procedure in any one package leads us to conclude that the packages lack the reliability needed for many statistical analysis purposes. Due to the ease with which these packages can be used, and the availability of introductory level statistical procedures, however, they are ideal tools for use in introductory statistical classes. If one has to make a choice between WebStat and Statlets, one should select WebStat since it performs considerably better than Statlets in those analyses it can address. The polynomial and multiple linear regression analyses that can be performed in Statlets but not in WebStat produce absolutely unreliable results - thus suggesting one should not use this feature at all in Statlets. Since the main difference between the commercial and demo versions of Statlets is that the commercial version supports larger datasets, we cannot recommend its purchase due to the severe restrictions of the freely available demo version.

Developing web-based statistical software packages is certainly a difficult task. Developers do not only have to create an attractive user interface but also have to rewrite or redevelop statistical functionality. However, even with the most user-friendly web-accessible interface, the true merits of any statistical package lie in its functionality and the accuracy of its results. Future (commercial or non-commercial) web-based statistical packages have to overcome the accuracy problems reported for WebStat and Statlets in this paper to become serious competitors to classical statistical packages such as SAS, SPSS, and S-Plus.

However, it should be noted that WebStat did considerably better than Excel for several of the datasets of average and high difficulty level. For most of the datasets of low difficulty level where Excel did well, WebStat also did well under the constraint of only 8 decimal digits of output, i.e., single precision output. However, it has been pointed out by West (2002, private communication) that the next version of WebStat will have a switch such that users can choose the reported precision of their results. It might therefore be worthwhile to reevaluate the accuracy of this upcoming version of WebStat once it becomes available.

Acknowledgements

The authors would like to thank the Co-Editor and an anonymous Associate Editor of Computational Statistics for their helpful comments. Thanks are also due to Webster West for his comments regarding WebStat.

References

Gentle, J. E. (1998), *Numerical linear algebra for applications in statistics*, Springer, New York, NY.

Kötter, T. (1997a), "Interactive interfaces of statistical software for the internet", in *Advances in Statistical Software 6*, W. Bandilla, and F. Faulbaum, eds., Stuttgart: Lucius & Lucius, 153-158.

Kötter, T. (1997b), "Implementation of networked applications for statistical computing", *Bulletin of the International Statistical Institute, 51st Session Istanbul 1997, Proceedings Book* **2**, 47-50.

McCullough, B. D. (1998), "Assessing reliability of statistical software: Part I", *The American Statistician*, **52**, 358-366.

McCullough, B. D. (1999a), "Assessing reliability of statistical software: Part II", *The American Statistician*, **53**, 149-159.

McCullough, B. D. (1999b), "The reliability of econometric software: Eviews, LIMDEP, SHAZAM, and TSP", *Journal of Applied Econometrics*, **14**, 191-202.

McCullough, B. D. (1999c), "Experience with the StRD: application and interpretation", *Computing Science and Statistics*, **31**, 16-21.

McCullough, B. D. (2000), "The accuracy of *Mathematica 4* as a statistical package", *Computational Statistics*, **15**, 279-299.

McCullough, B. D., and Wilson, B. (1999), "On the accuracy of statistical procedures in Microsoft Excel 97", *Computational Statistics & Data Analysis*, **31**(1), 27-37.

McCullough, B. D., and Vinod, H. D. (1999), "The numerical reliability of econometric software", *Journal of Economic Literature*, **37**, 633-665.

Müller, M. (1998), "Computer-assisted statistics teaching in network environments", in *COMPSTAT - Proceedings in Computational Statistics, 13th Symposium held in Bristol, Great Britain, 1998*, R. Payne, and P. Green, eds., Heidelberg: Physica-Verlag, 77-88.

Rogers, J., Filliben, J., Gill, L., Guthrie, W., Lagergren, E., and Vangel, M. (1998), "StRD: statistical reference datasets for assessing the numerical accuracy of statistical software," NIST TN# 1396, National Institute of Standards and Technology, USA.

Sawitzki, G. (1994a), "Testing numerical reliability of data analysis systems", *Computational Statistics & Data Analysis*, **18**, 269-286.

Sawitzki, G. (1994b), "Report on the reliability of data analysis systems", *Computational Statistics & Data Analysis*, **18**, 289-301.

Schmelzer, S., Kötter, T., Klinke, S., and Härdle, W. (1996), "A new generation of a statistical computing environment on the net", in *COMPSTAT - Proceedings in Computational Statistics, 12th Symposium held in Barcelona, Spain, 1996*, A. Prat, ed., Heidelberg: Physica-Verlag, 135-148.

Simon, S. D., and Lesage, J. P. (1989), "Assessing the accuracy of ANOVA calculations in statistical software", *Computational Statistics & Data Analysis*, **8**, 325-332.

Vinod, H. D. (2000), "Review of Gauss for Windows, including its numerical accuracy", *Journal of Applied Econometrics*, **15**, 211-220.

West, R. W. (1997), "Statistical applications for the World Wide Web", *Bulletin of the International Statistical Institute, 51st Session Istanbul 1997, Proceedings Book* **2**, 7-10.

West, R. W., and Ogden, R. T. (1997), "Statistical analysis with WebStat, a Java applet for the World Wide Web", *Journal of Statistical Software*, **2**(3), http://www.jstatsoft.org/v02/i03.

West, R. W., and Ogden, R. T. (1998), "WebStat: An environment for statistical analysis on the World Wide Web", *Computing Science and Statistics*, **29**(1), 307-310.

West, R. W., Ogden, R. T., and Rossini, A. J. (1998), "Statistical tools on the World Wide Web", *The American Statistician*, **52**(3), 257-262.