International Journal of Statistical Sciences *ISSN 1683–5603*
Vol.5, 2006, pp 109-120

Use of Statistical Packages in Statistics Teaching and Research: An Appraisal

Debasis Sengupta
Indian Statistical Institute
Kolkata, India

[Received March 12, 2004; Accepted June 16, 2006]

Abstract

The availability of statistical packages has opened up new possibilities in statistics teaching. There are both advantages and drawbacks of using packages as a part of teaching both at the introductory and advanced levels. After a brief discussion of these issues, attempt will be made to identify ways in which software can be used to supplement teaching. Emphasis will be given on case studies, demonstration, and practical difficulties faced by a teacher who has limited exposure to the use of software.

1 Introduction

The emergence of statistical software has given rise to many possibilities. There are obvious and not-so-obvious advantages that the software provide in respect of teaching and research. On the other hand there are also apprehensions about the software. The purpose of these article to discuss these issues in the light of some practical experience.

2 Potential advantages of using software in teaching

2.1 Scope for practical demonstrations

Use of software makes it possible to use a wider range of data-analytic tools over relatively large sets of data. Thus, teachers can use realistic examples in the classroom to illustrate various techniques.

For example, the usefulness of frequency distribution can be demonstrated by means of a data set on arsenic content of water samples from 113 boreholes distributed over various locations in Bangladesh. The data were collected by the British

Geological Survey and are freely downloadable from its website.[1] A part of this data set is given in Table 1. Here, `As` represents arsenic concentration in micrograms per litre (concentrations smaller than .5 micrograms per litre are reported as 0), `Fe` is Iron concentration in milligrams per litre, `Lat` is latitude in degrees, `Depth` is depth of the borehole in meters and `Division` is the administrative division. This data set will be revisited in this article in order to illustrate various other points.

The frequency distribution given in Table 2 shows that samples from about 14% of the locations had arsenic concentration above 50 micrograms per litre (the maximum acceptable level according to the Bangladesh government),[2] and another 15% had more than 10 micrograms per litre (the maximum acceptable level according to the World Health Organization).[3] In this case a histogram representing the relative frequency density would not be very useful, since there is a large concentration of extremely small values.

As	Fe	Lat	Depth	Division	As	Fe	Lat	Depth	Division
1	0.06	26.479	9	Rajshahi	13.6	0.749	24.444	118	Dhaka
2.2	6.21	25.7	15	Rajshahi	40.8	0.403	23.167	133	Khulna
1.1	1.48	24.088	37	Dhaka	2.7	0.192	24.454	NA	Dhaka
1.1	0.24	26.083	18	Rajshahi	3.4	0.02	23.173	24	Khulna
0	1.38	26.054	37	Rajshahi	120	2.61	24.047	20	Dhaka
1.7	0.26	24.91	26	Dhaka	32.5	1.13	23.059	24	Khulna
0	0.17	25.903	30	Rajshahi	7.5	2.84	24.605	34	Dhaka
0.8	0.293	25.901	30	Rajshahi	0	0.028	22.812	274	Khulna
0	0.027	25.665	18	Rajshahi	77.6	7.33	24.24	27	Dhaka
0.8	1.65	25.653	20	Rajshahi	11.8	0.019	22.718	183	Khulna
0.9	0.224	25.288	21	Rajshahi	16.8	4.43	24.108	27	Dhaka
0	0.356	26.1	24	Rajshahi	1.4	0.094	22.586	18	Khulna
1.9	0.837	25.9	18	Rajshahi	7.7	1.62	24.197	29	Dhaka
0	0.271	26.203	NA	Rajshahi	8.1	0.086	22.663	38	Khulna
0	0.018	25.931	27	Rajshahi	70.6	5.16	23.926	NA	Dhaka
0	0.988	25.752	18	Rajshahi	0.7	0.022	22.548	274	Khulna
4	9.07	25.913	14	Rajshahi	7.5	0.071	23.734	137	Dhaka
0	0.182	25.965	18	Rajshahi	29	2.26	23.765	22	Dhaka
0.6	2.13	25.653	18	Rajshahi	61.8	4.43	23.754	127	Dhaka
0	0.084	25.532	18	Rajshahi	215	5.67	23.599	84	Dhaka
0	0.054	25.318	18	Rajshahi	200	10.6	23.21	20	Dhaka
0	0.016	25.134	12	Rajshahi	144	6.92	23.005	20	Dhaka
0	0.976	25.109	15	Rajshahi	3.1	1.17	23.167	238	Dhaka
0	0.04	24.804	18	Rajshahi	3.4	0.156	22.7	610	Barisal
0	0.023	24.933	18	Rajshahi	10	0.112	22.355	274	Barisal
0.5	0.08	25.05	15	Rajshahi	401	4.43	23.51	26	Dhaka
0.5	0.354	24.793	30	Rajshahi	7.5	0.096	23.769	53	Dhaka
5.6	0.078	22.365	NA	Barisal	1.2	33.6	22.1467	265	Chittagong
1	1.24	24.8	30	Rajshahi	4.2	0.211	22.1924	146	Chittagong
1.4	0.086	22.365	37	Barisal	0.7	12.4	22.1467	68	Chittagong
0.6	0.03	24.684	61	Rajshahi	3.7	1.66	23.4453	32	Chittagong
14.6	0.426	22.365	27	Barisal	444	10.1	23.536	32	Chittagong
0.5	0.174	24.461	30	Rajshahi	107	2.78	23.2499	23	Chittagong
1.7	0.014	24.374	37	Rajshahi	2.5	4.02	23.2484	107	Chittagong
0	0.03	24.304	18	Rajshahi	234	7.06	23.2342	168	Chittagong
5.5	0.032	24.414	27	Rajshahi	100	0.665	22.8334	11	Chittagong
0	0.037	24.506	18	Rajshahi	111	0.49	23.0053	13	Chittagong
0.9	0.036	24.179	27	Rajshahi	275	1.32	22.7761	25	Chittagong
12.3	3.99	24.507	15	Rajshahi	4.4	0.033	22.4735	110	Chittagong
15.4	0.397	24.463	18	Rajshahi	3.4	0.202	22.3704	146	Chittagong
3.5	0.044	24.15	18	Rajshahi	2	0.365	22.2234	43	Chittagong
1.9	0.151	24.006	30	Rajshahi	4.6	0.527	22.0721	31	Chittagong
1.1	0.061	23.94	18	Rajshahi	2	0.128	21.4349	10	Chittagong
28.7	3.75	23.951	20	Rajshahi	1.2	0.307	21.2432	52	Chittagong
82.5	15.8	25.156	27	Dhaka	0.9	0.139	22.4815	56.96	Chittagong
8.4	0.015	24.005	113	Khulna	0.6	0.065	22.4815	160.92	Chittagong
34.9	0.791	25.117	27	Dhaka	1.1	0.047	22.4815	261.94	Chittagong
0.8	0.017	24.019	NA	Khulna	8.7	0.088	21.517	7.32	Chittagong
0	0.953	23.909	49	Khulna	1.1	0.29	21.8128	542.68	Chittagong
27.8	19.6	25.018	36	Rajshahi	1.1	0.089	22.4817	329.26	Chittagong
17.7	0.007	23.761	27	Khulna	11.3	0.249	23.8659	18.3	Chittagong
1.4	0.316	24.758	20	Dhaka	261	0.134	23.7627	26.22	Chittagong
5.7	0.02	23.529	18	Khulna	3.4	0.161	24.1945	30.5	Sylhet
4.2	0.074	24.748	107	Dhaka	5.2	0.211	24.3776	36.59	Sylhet
35.6	0.914	23.547	122	Khulna	2.3	0.197	24.4696	30.5	Sylhet
1.8	0.697	24.866	65	Dhaka	11.2	2.26	24.8795	30.5	Sylhet
17	0.494	23.488	21	Khulna					

Table 1: Arsenic data (As concentration smaller than 0.5 micrograms per litre is reported as 0). Source — http://www.bgs.ac.uk/arsenic/bangladesh/Data/BWDBSurveyData.csv.

Range	Frequency	Relative frequency (%)
0-1μg/L	34	30.09
1.1-10μg/L	46	40.71
10.1-50μg/L	17	15.04
50.1-100μg/L	5	4.42
More than 100μg/L	11	9.73

Table 2: Frequency distribution of arsenic data.

2.2 Developing intuition

In a physics laboratory a student can design the conditions of an experiment and examine the outcome rather quickly. This helps one gather experience quickly and thus develop intuition about the underlying physical laws. In statistics the response can be very slow unless the computation is done in a computer. A judicious user can use a statistical software to develop his/her intuitions.

This is made possible by exposure to a variety of outcomes — both expected and unexpected. Consider once again the arsenic data, restricted to the 108 cases where the information is complete (that is, after excluding the five cases where the depth of the borehole is not known). The arsenic content `As` may be regressed linearly on `Fe`, `Lat` and `Depth`. It is well known that inclusion of too many variables in a regression equation may lead to lack of precision of the estimators of some parameters. For the present data, when `As` is regressed on `Lat` alone, the p-value of the regression coefficient is .102. The p-value of this coefficient increases to .127 when `Fe` is included in the model. This increase is expected from the foregoing explanation. However, the p-value drops to .028 as soon as the variable `Depth` is brought into the model. Such an outcome may surprise the analyst if he/she is not aware that the effect of a variable may be masked by the absence of other variables, and inclusion of the missing variable would then make everything fall in place.

Students sometimes think that a regression model fits quite well if the regression coefficients are statistically significant. For the present data set, the regression equation involving the three explanatory variables is as follows.

$$\underset{}{\texttt{As}} = \underset{\substack{(154.5)\\[.018]}}{370.7} - \underset{\substack{(6.332)\\[.028]}}{14.11}\,\texttt{Lat} + \underset{\substack{(1.567)\\[.004]}}{4.571}\,\texttt{Fe} - \underset{\substack{(.0794)\\[.053]}}{.1553}\,\texttt{Depth}. \tag{1}$$

In the above equation the numbers in parentheses indicate the standard errors, while the numbers in square brackets indicate the p-values. The multiple R-square is .124. This shows that there much more variability in the arsenic concentration than what is explained by the explanatory variables — even though most of the coefficients are statistically significant at the 5% level.

The role of high leverage observations are also illustrated through analysis of this data set. Two observations (no. 50 and 85) have very high leverages — .494 and .162. Once these are excluded from the analysis, the fitted equation changes drastically to

$$\texttt{As} = \underset{\substack{(143.1)\\ [.015]}}{353.5} - \underset{\substack{(5.867)\\ [.019]}}{13.97}\,\texttt{Lat} + \underset{\substack{(2.328)\\ [.000]}}{12.79}\,\texttt{Fe} - \underset{\substack{(.0745)\\ [.193]}}{.0976}\,\texttt{Depth}. \tag{2}$$

Incidentally, the multiple R-square jumps up to .269 when the two cases are excluded.

2.3 Follow-up analysis

While a particular analysis answers some questions, it may also give rise to further questions. Sometimes the analyst has to look beyond the domain of statistics for answers to these new questions. For example, the signs of the estimated regression coefficients given in (1) would make sense if one has some understanding of the mechanism of arsenic accumulation in groundwater (see, for instance the Introduction section of [1]).

The effect of the geographical location is also brought out by a one-way analysis of variance after classifying the 113 cases by the administrative division. The divisional effect is highly significant. The divisional averages show that Chittagong (mean arsenic concentration 63.47 μg/L) and Dhaka (63.04) are worst affected, while the problem is not so serious in Khulna (12.26), Barisal (7.00), Sylhet (5.53) and Rajshahi (2.88).

If one looks closely into the two high-leverage cases, the first is found to have come from a rather shallow borehole located in Rajshahi, and the arsenic concentration is unusually high. On the other hand, case 87 comes from a deep borehole in Chittagong with a high iron concentration, but the arsenic concentration is unusually low. The unusual level of the response at the high-leverage points explain why their deletion improves the fit.

Following up on new questions, whether these are statistical or not, help students make important connection between their theoretical knowledge and the problems of real life and gain confidence in their ability to solve problems. The statistical part of this follow-up study is made easier by software.

2.4 Believing theorems

Understanding the proof of a theorem does not necessarily mean that one *believes* it. Believing a proven result is an essential part of developing statistical intuition. If real life does not provide adequate opportunity for watching a result in action (as is the case for many statistical theorems), random simulations can act as a substitute. For example, the appropriateness of the asymptotic distribution of a statistic can be

verified through simulations. Insight into rates of convergence may also be gained in this manner. An often neglected quantity in studying rates of convergence is the constant associated with the leading term. Sometimes simulations help one realize that a large value of the constant can be responsible for apparently slow convergence.

2.5 Quick verification of hunches

When one selects a model from a number of competing models using some data-based selection criterion, and proceeds to draw inference on the parameters of the model on the basis of the same data set, then some bias may creep into the estimator. This 'selection bias' may be visualized by considering two competing subset regression models which are intrinsically equally good. One of these two models is bound to be selected via the chosen criteria, — in a particular realization of the concerned random phenomena. This means that this particular realization of the chosen model provides better explanation of the response than many other realizations of the same model. Thus, conditioning on selection alters the explanatory characteristics and thus may lead to bias. This is only a hunch. It is very difficult to prove, even in a simple case, that the estimators of regression coefficients in a particular subset model are biased when conditioned on selection. This may however be established through simulations, as shown by Miller.[4]

What often stands between a 'hunch' and a 'proof' is a piece of mathematical jugglery which has nothing to do with statistics. Simulation through statistical software has the potential to remove this obstacle.

2.6 Checking assumptions and finding alternative models

Using statistical software one can get an idea about the appropriateness of the various assumptions underlying a certain analysis. For example, the plot of the observed vs. fitted values of arsenic concentration from the foregoing analysis (after dropping observations 50 and 85), shown in Figure 1, indicate the presence of considerable heteroscedasticity. In particular, there is larger variation in observed concentration when the predicted concentration is larger. The plot also underscores what should be understood at the outset — that the predicted concentration obtained from a linear model may be negative in some cases, whereas the observed concentration can never be negative.

The problem of heteroscedasticity and the non-negativity of the response may be tackled by using log(As) as the response variable. This transformation is usually ruled out when some response values are 0. However, in this case, the 0 values are in fact censored from the left at .5 micrograms per litre. The normal likelihood for the censored data may be maximized using the EM algorithm.[5] The 'complete data' in this case would include the exact values of these low arsenic concentration. The iterative steps are as follows.

$$\text{E-step} \quad : \quad \log(\texttt{As}_i) = x_i^T\beta - \sigma \exp\left[-\frac{1}{2}\left(\frac{\log(.5) - x_i^T\beta}{\sigma}\right)^2\right] / \Phi\left(\frac{\log(.5) - x_i^T\beta}{\sigma}\right)$$

for each missing observation $\texttt{As}_i$ (β and σ are current estimates),

M-step : β = usual least squared estimator with missing values substituted as above,

σ = root-mean squared value of residuals from above least squares analysis.

In the above expressions, $\texttt{As}_i$ is the arsenic concentration in the ith case, x_i is the vector of the explanatory variables in the ith case and $\Phi(\cdot)$ is the standard normal distribution function. The iterations may begin with the M-step with log(.5) substituted for the missing values. The E- and M-steps can be coded easily in software such as *R* (freely downloadable from the internet).[6] The fitted equation (after dropping cases 50 and 85 and also case 87 which becomes a high-leverage point when the other two cases are dropped) is

$$\log(\texttt{As}) = \underset{\substack{(5.343)\\[.000]}}{31.81} - \underset{\substack{(.2193)\\[.000]}}{1.317}\,\texttt{Lat} + \underset{\substack{(.0928)\\[.000]}}{.5462}\,\texttt{Fe} - \underset{\substack{(.0027)\\[.114]}}{.0044}\,\texttt{Depth}. \tag{3}$$

The multiple R-square is 0.404, which cannot be compared with the previous value,

as the response has been transformed. Figure 2 shows the observed vs. predicted log-concentration of arsenic. It appears that heteroscedasticity is somewhat removed.

One can also fit a binary data regression model for the 'indicator variable' of arsenic concentration being more than 10 micrograms per litre, using the three explanatory variables. The fitted equation (after dropping observations 50, 85 and 87) happens to be

$$\log\left(\frac{Prob(\texttt{As} > 10)}{1 - Prob(\texttt{As} > 10)}\right) = \underset{\substack{(4.943)\\[.011]}}{9.967} - \underset{\substack{(.2038)\\[.007]}}{.4448}\,\texttt{Lat} + \underset{\substack{(.0591)\\[.004]}}{.1447}\,\texttt{Fe} - \underset{\substack{(.0045)\\[.011]}}{.0091}\,\texttt{Depth}. \qquad (4)$$

The binary observation (1 for arsenic concentration higher than the WHO limit and 0 for concentration within the limit) is plotted in Figure 3 against the predicted probability of the response being 1. The fit is reasonable.

3 Potential problems with using software

3.1 Too expensive

The high cost of computer hardware as well as statistical software is a major problem faced by academic institutions in many developing countries. Even though hardware becomes cheaper with newer technology flooding the market, the older hardware quickly becomes unusable because of the shortage of spare parts. Newer versions of software require newer hardware. Thus, one never really has the chance to reduce expenditure.

In this scenario one has to be ready to extract as much mileage as possible from the available resources. For examples, the smartest machines or the the latest version of the best software is not necessary for most of the computational needs of an average statistics teacher or student. A pentium PC with *MS-Excel* can be used to carry out almost all the statistical computing at the bachelor's degree level. Making a not-so-smart software do sophisticated computation is a challenge which should excite the students. The scope of manipulating with vectors makes packages like *MS-Excel* more attractive than writing programs in C or Fortran.

Minitab is widely used in undergraduate teaching, and its website provides a freely downloadable trial version.[7] However, it is very expensive. *Systat* provides a somewhat cheaper alternative.[8]

Another alternative is provided by the *R Project for Statistical Computing* which has created the freely available computing environment called *R*.[9] A versatile statistics toolbox is also available with *R*.[6]

3.2 Need for a teacher to learn more

This is perhaps biggest hurdle. An uninitiated teacher may find it quite difficult to use the various options of a software, not to speak of programming in a general or statistical computing environment. Unexplained keywords and unexpected results only adds to the frustration. I have taken the easiest route in this respect, by learning together with students. The students can do the programming while the teacher can help them interpret the findings. In the process a greater link is achieved between theory and practice.

A teacher also needs a number of real data sets which can be used in the initial phase of learning data analysis. There are plenty of data-oriented books in the market, including some books on data alone.[10,11] These data are freely downloadable from the internet. Students can also be asked to collect primary and secondary data, which is an important experience for them. If they have access to the internet, then they can make use of some excellent resources available in it.[12]

3.3 Allocation of time

When statistical data analysis has to be accommodated in an existing curriculum, it may jeopardize the time table as well as the teacher's preparatory plan. As far as the teacher's time is concerned, the difficulty is mostly restricted to the initial stage. My experience at the Indian Statistical Institute, Kolkata suggests that lecture time is not strained at all by introducing the computational component. Home assignments in courses such as statistical methods, statistical inference, linear models, sample survey and design of experiments, regression techniques, time series analysis, multivariate analysis, large sample techniques, stochastic processes and so on can be made computer-oriented. Some class time has to be set aside for discussion of these assignments. This can be accommodated at the expense of routine algebra or calculus used in some proofs.

3.4 Stunted thought process of students

It is common knowledge that school students who have early access to calculators tend to forget multiplication tables and generally have less time to develop an intuition about numbers. A similar risk is faced by statistics students in college who have easy access to statistical computing. They may be tempted to follow worked out examples too closely, or to check out a 'hunch' too soon — even before trying to guess the answer through intuition or reasoning.

While this possibility cannot be ruled out, access to software is not the issue. The same reliance on worked out matter can also be seen among students who do not have access to computers at all. For many students of this category statistics is a bunch of theorems stated in books.

The calculator syndrome and the book syndrome are two sides of the same coin. The teacher's thrust has to be on relating all new knowledge (whether these are gathered from books or software) to the world we can touch, hear and see. A statistical data analyst should know that in a regression problem a high-leverage point is an extreme case. He should be able to identify some obviously high-leverage points simply by inspection — when the data set is not too large. (The high-leverage points of the arsenic data set is a case in point.) The data analyst should also satisfy himself that the outcome of the analysis is meaningful, the coefficients have the appropriate sign, the predictions are in the right range, and so on. No statistical software can impart or impair these skills. The responsibility lies with the analyst.

3.5 Statistical software are not perfect

All statistical software have errors in them. These errors are not independent. The fact is that while competing with one another for the development of newer versions,

the software teams borrow features and ideas from one another. They also borrow errors in the process.

It is a common practice in most (if not all) statistical software to denote a standardized statistic as a *t-ratio* whenever the standard deviation in the denominator is an estimated one. This nomenclature is appropriate in the context of linear regression if normality is assumed. However, a *t-ratio* is also mentioned alongside the estimated parameters in various other contexts such as the generalized linear model, where the scaled statistic is normal only in the asymptotic sense, and no small sample distribution is available! The *t-ratio* is therefore a misnomer which could not have been invented by all the software independently.

4 Research

Issues discussed in Section 2 are relevant not only for teaching but for research as well. Software also helps a researcher visualize sets, surfaces and other mathematical constructs. Availability of fast statistical computing has made possible the widespread use of iterative methods such as the EM-algorithm. It has also inspired the advent of computation-intensive methodology such as resampling techniques, diagnostics, Markov chain monte carlo and other Bayesian computational algorithms.

Unfortunately, many new computation-intensive methods add little if anything to the existing body of knowledge on the subject. As these methods have become fashionable, those who seek the quickest route to publication are tempted to join the bandwagon and do what they were doing elsewhere — churning out fancy methods that nobody will ever use, except for comparing with fancier methods.

5 Concluding remarks

It was once suggested to me that statistical packages may have made statisticians redundant. There is some truth in this statement. Scientists and engineers now have access to a much wider range of statistical tools than ever before. However, the packages can not possibly provide application-specific interpretation. Statisticians can fill in this void. Those who are unable to do so may have become redundant to the users of statistical methods.

The positive aspect of this development is that statisticians need no longer be constrained by the lack of computational resources. There will be no excuse for not demonstrating the effectiveness of a particular method. There will be no credit for inventing minor computational shortcuts. Packages are pushing statisticians to mind their statistics.

References

1. British Geological Survey (2001). *The groundwater arsenic problem in Bangladesh (Phase 2)*, http://www.bgs.ac.uk/arsenic/bangladesh/reports.htm (also available from Graphosman, 3/3-C Purana Paltan, Karim Mansion, 1st Floor, Dhaka 1000, Bangladesh). Data available in http://www.bgs.ac.uk/arsenic/bangladesh/Data/BWDBSurveyData.csv.
2. Government of the People's Republic of Bangladesh, Ministry of Health and Family Welfare, *Arsenic Contamination of Ground Water in Bangladesh* (a briefing paper), http://phys4.harvard.edu/~wilson/arsenic_project_ground_water.html.
3. World Health Organization (1998). *Guidelines for Drinking-Water Quality*, 2nd ed., Vol.2 (Health Criteria and Other Supporting Information), WHO, Geneva.
4. Miller, A.J. (1990). *Subset Selection in Regression*, Chapman & Hall, London, pp.7-9.
5. McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.
6. R Foundation for Statistical Computing, *The R Project for Statistical Computing*, http://www.r-project.org.
7. Minitab Homepage, http://www.minitab.com.
8. Systat Software Inc. Homepage, http://www.systat.com.
9. Lääräkä, Esa, *R : a versatile computing environment for statistical analysis*, Biocenter Oulu Course, http://www.biochem.oulu.fi/BioStat/el1_6up.pdf.
10. Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E., (1994). *A Handbook of Small Data Sets*, Chapman & Hall, London.
(Get data from http://www.stat.ucla.edu/practice/.)
11. Andrews, D.F. and Herzberg, A.M. (1985). *Data*, Springer-Verlag, Berlin. (Data available from http://www.stat.ucla.edu/practice/.)
12. Statistical Science Web, http://www.statsci.org.