The Role of Package Driven Statistics Courses

Alan J Lee and George A F Seber - Auckland, New Zealand

1. Some history

In 1974, one of the authors (GAFS) introduced a terminating first year service course in statistics, paper 26.181, for non-mathematics students. In the following years we observed that the service course 26.181 catered not only for students majoring in other subjects, but also for a substantial number of mathematics students who preferred a more practical approach than the traditional one. The numbers were also steadily growing (about 1500 in 1990). At that stage we realised that 26.181 provided a potential source of advancing statistics students who might be interested in taking a second year follow-up course. In the 1980s Alan Lee (and probably most of us of that vintage) had been strongly influenced by works on exploratory data analysis by such people as Tukey, McNeil, Velleman and Hoaglin. Under Lee's direction a second year data analysis course 26.281 was launched in 1981. His aim was to provide further training in practical statistics and data analysis without requiring too much mathematical knowledge or statistical theory. He realised that the students needed easy computer access for a realistic approach to data analysis. Suitable access to a mainframe was out of the question at the time, but micros seemed a viable alternative. A suitable statistical package was therefore developed called STATCALC (Lee et al., 1984) which had partly evolved from several programs on exploratory data analysis adapted from McNeil (1977) by Ross Ihaka. The package runs on IBM and Macintosh personal computers, the latter being currently used in our department. Lee and Peter Mullins also wrote a manual to go with the package. The manual, with its extensive tutorial section, also serves as the text for the course.

It should be realised that the course is not simply a methods course with a package appended to provide some hands-on experience. Rather, the course is "driven" by the package. The emphasis is on model fitting leading into inference, with residuals, diagnostic checks, plots and transformations readily on tap. There is also a heavy reliance on case studies and practical demonstrations using Kodak Datashow to display the computer screen in a lecture. Practice and theory go hand in hand and the students are encouraged to experiment for themselves. STATCALC is very user-friendly with single line commands and is aimed at students with little knowledge of computers. The course is popular and steadily growing (about 260 students at present).

2. General philosophy

All statistics instructors realise that there are certain topics that students find difficult. One such is the concept of a model. Even more difficult is the related idea of the sampling distribution of a parameter estimate. This concept is fundamental to an understanding of inference and, in our experience, is not well understood by students. For this reason we like to teach a segment on exploratory data analysis (EDA) at the beginning of the course. One benefit of teaching EDA before conventional inferential statistics is that it is possible to introduce some of the key ideas behind modelling without requiring the students to grapple with difficult inferential ideas at the same time.

A data model (as described for example in McNeil, 1977) is presented as a device for summarising data. A canonical example is when we have a set of "regression data" consisting of pairs of obsevsations (x_i,y_i), $i = 1,2,...,n$. A model for this data is

$$y_i = a + bx_i + r_i$$

where the constants a and b determine the line that summarises the data. The success of the model as a summary is judged by the extent to which the "residuals" r_i exhibit no pattern, as the aim is to explain all of the structure in the data by the line. The model fitting process (i.e. the choice of a and b) can be either least squares or some robust method. This gives plenty of opportunity for discussing the role of robust methods at a non-technical level. If the methods fail to give a good fit (i.e. if the residuals still exhibit pattern, due perhaps to the true relationship between x and y being non-linear), we may have to transform the data and/or fit a more complicated model. This cycle of fitting, making diagnostic checks and if necessary refitting, is repeated until a satisfactory description of the data is achieved.

Thus the essentials of the model building process can be taught without having to burden the student with too much technical detail. The tools necessary to enable students to do something like the above on a variety of data structures (single batches, groups of batches, two-way tables, time series, simple and multiple regression data) can be introduced in a non-technical manner using intuitive explanations and plenty of examples. There is a heavy reliance on pictures and graphical methods.

The role of the computer in all this is simple: it makes the above approach possible. No student is going to fit models (even straight lines by least squares), compute residuals, plot them, transform and refit if the only technology available is a pocket calculator and pencil and paper. We need to equip them with tools to make these tasks effortless, and set their minds free to think about the data, rather than the mechanics of pressing buttons. This means providing an adequate number of computer terminals running easy-to-use software, backed up by trained tutors that can give on-the-spot help to the more computer-phobic students. It is vital that the frustration level in the students be kept as close to zero as possible. After all, we are selling the computers as an easy-to-use tool, and nothing destroys the credibility of this claim faster than malfunctioning or hard-to-use hardware and software. Micros tend to be better than mainframes and terminals in terms of frustration minimisation. They are also less prone to go down for "routine maintenance".

The EDA segment of the course requires plenty of practical work, and the students emerge being reasonably well-drilled in the basics of model fitting. They are also familiar with a range of simple data structures. In addition, they are also led to appreciate the need for inference. How else are they to decide if a line has slope sufficiently close to zero to indicate no relationship, or if the means of two sets of data are sufficiently similar to indicate that treatment A and treatment B are of similar efficacy? Thus EDA leads naturally to the more traditional confirmatory part of the course.

We can now address the idea of probability models, which attempt to summarise not the data (as does a data model) but rather the probabilistic mechanism which generates the data. (The random number generator on the computer is a useful metaphor here in that the probability model describes the "program" of Nature's random number generator.) Since the students are by now familiar with the modelling process, we can concentrate on an informal discussion of inferential issues such as sampling distributions, estimation methods, testing, and so on, while covering the following topics in the course: regression, one and two-way ANOVA, analysis of covariance, industrial statistics, contingency tables and non-parametric methods. The usual methods for performing confirmatory data analysis (estimation, checking of assumptions) closely parallel the methods already introduced more informally in the EDA segment, so they are already familiar when introduced in the confirmatory context.

This two pronged approach seems to work quite well. The emphasis is on example and analogy, since many students do not have sufficient mathematics for a formal presentation of the theory. However, mathematics is not statistics, but only a convenient and powerful language for talking about statistics. The challenge in a course such as this is to find at times an alternative language, and to some extent the language of computing fills this need. Much can be done with simulation experiments. The power of computer graphics can be exploited in ways bounded only by one's imagination and programming inadequacies.

Applied statistics is a practical subject so that most of the learning takes place when the students actually do data analysis. This means that the assignments should be structured to exercise the students in the skills taught in the course, and also reflect the open-ended nature of the subject. No analysis is ever totally definitive. A very important aspect of assignments is presentation. A statistician who cannot communicate his or her findings to the client is of little use, so assignments should take the form of reports from the data analyst (the student) to the client (the instructor). Ideally these should be typed, or at the very least written neatly in acceptable English, with conclusions drawn clearly spelled out in non-technical language. Selected tables, graphs and plots that bolster conclusions should be integrated with the text, or perhaps presented in an appendix. What should not be tolerated is the handing in of (almost) virgin printout, with only minimal scribbled notations to indicate that the embryonic statistician's brain has been in contact with the computer! To encourage good habits, the assignments should be worth a substantial proportion of the final grade. This may cause problems of collusion, but this has not been a big problem in our experience. In addition, formal examinations are useful in that they require students to take a holistic view of the course and tune up their knowledge to an acceptable level of precision. Our examinations feature computer printout which the students are invited to interpret.

3. Another new course

Following the success of 26.281, a third year course in the same spirit was introduced by Alan Lee and Chris Wild in 1988 with the following syllabus: techniques of multiple regression and analysis of variance, multi-way contingency tables and log-linear models, time series and multivariate methods. These are reasonably sophisticated techniques and it is even more of a challenge to describe concepts at this advanced level with a minimum of mathematics.

We use the SAS package throughout the course for several reasons. Most importantly, STATCALC is insufficiently comprehensive to perform all the analyses we require, so we need a fully-featured package such as SAS. Due to the strong presence of the SAS Institute in New Zealand, SAS is widely available outside the confines of academe, so a good working knowledge of SAS is a marketable skill.

MINITAB would be an acceptable alternative, except that its graphics capabilities are not as extensive as those of SAS, and it is not as widely available as SAS outside universities, at least in New Zealand. It is however considerably cheaper, and easier to use than SAS, which cannot be regarded as easy to learn even by its most zealous defenders. The manuals for SAS are rather forbidding, particularly in terms of their number and size, and they are too expensive for students to buy a full set. However, despite these drawbacks, we feel that the advantages of SAS outlined above justify its use. To get round the documentation difficulties, we issue extensive handouts giving examples of all analyses covered in the course, including the relevant SAS code, annotated as appropriate.

4. Role of courses

There is a temptation to regard such data analysis courses as "soft" options for those students who can't do mathematics. This is not the case. Our experience is that such courses require a good level of intelligence rather than mathematical sophistication, although a reasonable mathematical knowledge is needed to handle the material, particularly in the third year. The courses encourage "lateral thinking" and experimentation. Ultimately we prefer our more able students to take both the theoretical and data analysis courses. To encourage this, one of the theoretical third year courses, traditionally on linear models, is being modified this year by Seber to bring it closer in content to the data course. Students will then be able to get both theory and practice for the same techniques.

Why do we need such courses? In the first place students like such courses and these days we have to go with the demand! If they are not "confirmed" mathematicians then such courses give some reality to what they are learning. Secondly, a mix of computing and applied statistics is now desired by many other subjects, for example zoology and marketing. Thirdly, statistical packages are being used extensively by scientists and other researchers with, often, little understanding of what they are getting. Such packages produce all sorts of subsidiary material, some of which may not be applicable. Students and researchers need to be able to interpret the printout sensibly. For example, in the preface of his book on Multivariate Analysis, Seber (1984) writes: "However, the classical normal-based procedures still have a place for a number of reasons. First, some of these procedures are robust under moderate departures from the usual assumptions ... Fifth, many of the current computer packages automatically compute various normal-based statistics and it is important that package users know what these statistics are and how they behave in practice."

No doubt other statisticians throughout the world have been engaged in similar exercises with perhaps little communication with each other. This is true in this part of the world. Perhaps we need to find out what we are all doing in this regard at least in New Zealand and Australia.

References

Lee, A J, McInerney, P J and Mullins, P R (1984) *STATCALC : A Statistics Program for the IBM and Macintosh Personal Computers.* STATCALC Holdings Ltd, Auckland University, New Zealand.

McNeil, D R (1977) *Interactive Data Analysis.* Wiley, New York.

Seber, G A F (1984) *Multivariate Analysis.* Wiley, New York.