EXPLORATORY ANALYSIS OF DATA SET 1

SUSAN R. WILSON

Serially collected data, such as these for the vitamin E diet supplement study on the growth of guinea pigs, are often analysed by researchers using comparisons of groups at a series of time points. As pointed out by MATTHEWS *et al* [1], such an analysis is inadequate in two ways: It may fail to resolve experimentally relevant questions and it may be statistically invalid. They suggest a simple, two-stage remedy. First, a suitable summary of the response of the individual, such as a rate of change or an area under a curve, is identified and calculated for each subject. In the second stage these summary measures are analysed by simple statistical techniques as if they were raw data. From a consultant statistician's view, such an approach has great appeal in being valid, likely to be more relevant to the study questions and relatively simple (in the sense of involving minimal modelling-type assumptions). It is useful to keep in mind this approach when planning experiments. However, as noted by HAND [2], the method may conceal latent problems, and moreover, without the benefit of consultation with the experimenter it is not entirely clear which summary measures are most appropriate. Hence an exploratory data analytic approach to these data was chosen, based on graphical techniques.

From the graphs produced above (accompanying the data), it appears that observation 1 is an outlier in its initial growth pattern, particularly from Week 5 on. Any reasons for this need to be discussed with the experimenter. What is less clear, but more apparent if the panels are superimposed on transparencies using a different colour for each group, is that Groups 2 and 3 are relatively homogeneous, and their values after Week 5 are relatively higher than those for Group 1 (either including or excluding observation 1).

The statistical software XLISP-STAT of TIERNEY [3] has excellent, interactive, dynamic graphics. Features which were used for exploring these data included scatterplots (which have two highlighting techniques, *selecting* and *brushing*), spinning plots and (linear) regression fits with accompanying diagnostic plots. The different plots can interact by *linking* the views. One criticism of this software is that number i on the plot corresponds to observation i+1.

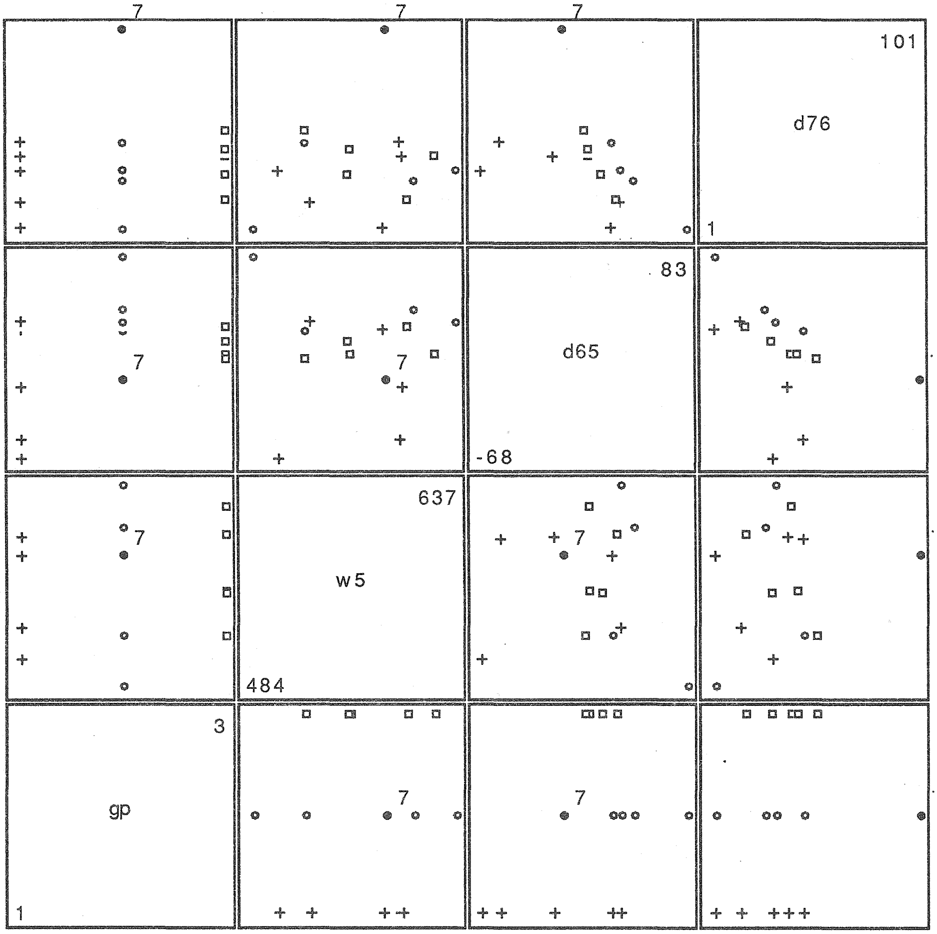


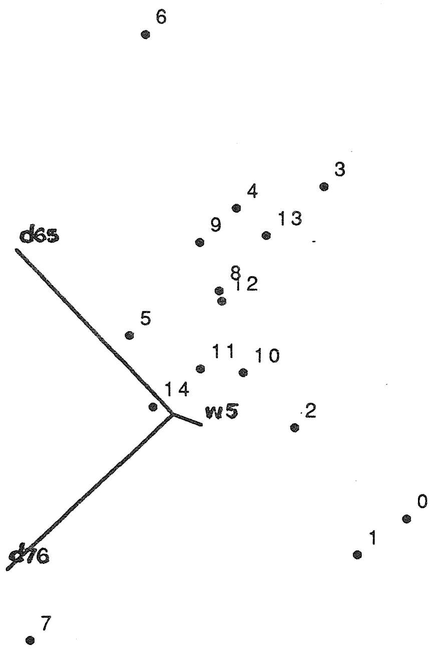
FIGURE 1

A Scatterplot Matrix for the Three Groups of Guinea Pigs

(*Crosses* : Group1 *Circles* : Group 2 *Squares* : Group 3
Variables are described in the text.)
An apparent outlier is highlighted: Observation 8 (Plot no. is 7)

Since treatment started at the beginning of Week 5, with measurements taken at the ends of Weeks 5, 6 and 7, for ease of presentation just the measurements taken then are considered here. Figure 1 gives a scatterplot matrix for the values of the three Groups of guinea pigs on the weight measures for Week 5 (w5), the difference in weight measures between Weeks 6 and 5 (d65) and the difference in the weight measures between Weeks 7 and 6 (d76). Observation 8 from Group 2 appears to be different. Figure 2 gives one of the spin-plot views. Four of the observations appear to be different from the remaining "cluster", namely observations 1, 2 as well as 7, 8. (This also appeared with spinning the plot of values for Weeks 5, 6 and 7.)

FIGURE 2
Spin-plot View
(Number i corresponds to Observation i+1
Variables are described in the text.)



Observations 1 and 2 both had large decreases in weight between the ends of Weeks 5 and 6, followed by relatively substantial recovery by the end of Week 7. Observation 7 had substantial weight gain between the ends of Weeks 5 and 6 followed by little change to the end of Week 7. Observation 8 had little change between the ends of Weeks 5 and 6 followed by a substantial weight gain between the ends of Weeks 6 and 7.

Results from regression analyses, based on prediction of weight by the end of Week 7, given the values at the end of Weeks 6 and 5, suggest that Groups 2 and 3 have significantly higher values than those for Group 1, but values for Groups 2 and 3 are indistinguishable from one another, and moreover that observation 1 was very influential in the determination of such a conclusion.

The above, exploratory, approach raises many questions to be put to the experimenter, and illustrates the necessity of having very clearly defined initial hypotheses with appropriately collected data. In any investigation such hypotheses should be evaluated first. Then exploratory techniques can be applied, and they may well raise more questions which may even need further experiments for their resolution. The XLISP-STAT software is one of the new and very useful statistical environments for relatively easily carrying out appropriate exploration. Also, the software is available free of charge; see *Appendix B* of TIERNEY [3] for details.

REFERENCES

[1] MATTHEWS, J.N.S., ALTMAN, D.G., CAMPBELL, M.J. and ROYSTON, P. (1990) Analysis of serial measurements in medical research. *British Medical Journal* **300**, 230-235.

[2] HAND, D.J. (1991) How to analyse your repeated measures data. *Ibid*.

[3] TIERNEY, L. (1990) *LISP-STAT : An Object-Oriented Environment for Statistical Computing and Dynamic Graphics.* John Wiley, New York.

Centre for Mathematics and Its Applications
The Australian National University
GPO Box 4, Canberra, ACT 2601