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.