One man's meat:

Part 3—It's all good grist that comes to our mill

By F. I. Musk*

This is the third of a series of papers which attempts to define one computer user's philosophy. The need for more sophisticated methods of data analysis requires a return to the concept of providing information for decision.

One of the features of the last decade has been the relative failure of industrial statistics to make the impact expected in the early fifties. The cause cannot be the feud that divided statisticians, the echoes of which still reverberate. It is not that the journals are filled with heavy, esoteric methodology. They are, but so are those on computing and operational research. Yet where practical computing and operational research can be and are conducted on a much lower plane than the average journal article, can be and are practised successfully by the non-specialist, statisticians have always been prone to hedge themselves about with (for example) the dangers of departing from normality, the robustness or otherwise of tests and restrictions on the size of samples (they always want more). Despite the phenomenal sales of Moroney (1951), and the lay understanding of statistical technique that must spring from them, statisticians basically dislike laymen. Statisticians are a Mecca, to which one should turn at dawn and dusk. They are pedagogues who teach deviation, but castigate the poor experimenter who deviates from the narrow path, their path, to consider some curious and interesting phenomenon.

I am being unfair. There can be penal dangers in non-adherence to an experimental design, in poor coverage of the experimental space, in extrapolation, in deduction from a sample of one. It is probable that statistics are not being used to their full bent for the same reason that critical path analysis has assumed a popularity in practice that linear programming never had. They are built upon a structure of relatively difficult mathematics. Beneath them there is not solid ground, and who knows where the ignorant enquirer may fetch up if he falls into an unexpected hole? But statistics are used to great effect in medicine, insurance, biology, and are studied in depth in academic circles. Why not in industry?

There are two aspects of statistics in an industrial setting which, perhaps, demarcate them from those in other fields. Firstly, there is an extreme shortage of statisticians in industry. If problem solvers outnumber statisticians by perhaps fifty to one, they cannot be expected to queue for service, and they do not. Consequently, masses of interesting data run to waste with the effluent. Secondly, much industrial data is so incredibly coarse as to make derisory the use of the more sophisticated statistical methods.

* *Courtaulds Ltd., Matlock Road, Coventry.*

Regression analysis

I well remember seeking the views of two prominent statisticians on our intention to write a regression program for a large number of variables. The attitude of one was that regression analysis was very old, uninspiring stuff. The other felt that we were building a sledgehammer, assuming we had nuts to crack. We did build, not a sledgehammer, but a coarse-grained sieve, which was useful only to dispel such variables as were patently non-significant. This, in itself, was one step forward, but there were difficulties.

The input to our first regression program was by means of cards to magnetic tape, allowing for up to 64 independent variables, and up to 36 dependent variables. The data sets had to be fed to the computer in strict order, and at the time of input, it had to be determined which fields represented dependent, and which independent variables. True, the computer could calculate, untouched by human hand, the regression for each dependent variable in turn with respect to all the independent variables, but if in the light of the resulting analysis, changes were required in the variables, all the data had to be repunched, and it was obvious that this was not taking us very far very fast.

Occasioned in the first place by the need to avoid the waste of time and work in repunching, a more general form of input was devised, which would be sorted if need be and placed on tape in data set order. A new program would then, according to the current regression model, select any of the variables earmarked as dependent variables, and any which were independent variables. But if this program could select and reject, why should it not also calculate? If it could calculate, we could add, subtract, multiply or divide one variable with another, through all readings of these variables. If this were so (enthusiasm mounting) we could take powers, logarithms, roots. We could transform variables by constants. We could perform these operations sequentially, without limit, yielding from our original variables, the most complex variable structure necessary. In other words, we would be able to move from linear regression to surface fitting. This we did.

Important objections can be raised against the introduction of coarse plant data to regression analysis, even where some attempt has been made, within the confines of production requirements, to cover the

experimental space reasonably well. Many of the initial independent variables will have little or no effect upon the factor to be measured, but these can be suppressed by standard stepwise procedures. However, since it is unlikely that the measured variables are the prime variables, it is useful if the measured variables can be combined into "true" effects. A scrutiny of the correlations emerging in the analysis is useful here. It is better still if these correlations can be supported by scatter diagrams for each pair of variables, and so we have added this facility. Even better indications arise if we can introduce the variables to a multifactor analysis. We had developed factor analysis for use in comparing relationships in tests of fabrics, where subjective data (loft, drape, handle) arose, as well as objective data (weight, flexural rigidity, bending modulus), and we found it useful to link this routine with the regression by causing it to accept the same standard form of input.

Matrix operations

I cannot say that we have been required to perform many types of matrix operation. Most frequent is matrix inversion, but since we were supplied, as supporting software, with several subroutines for matrix operations (and we added the rest) we were able to link these together, and to the standard form of input, by embedding them all into a matrix interpretive scheme in such a manner that a continuous stream of matrix operations can be generated, and results printed at any stage, by a short macro-program. All matrices entering the computer, or generated at any stage, are automatically labelled and stored on magnetic tape, and so we had also to provide a means of deleting them.

Linked together by a common form of input and transformation facility, these three routines, regression, factor analysis and the matrix scheme, constitute the main elements of our analysis suite, with the grandiloquent title of COSMOS, or Courtaulds' Own System for Matrix Operations and Statistics. It will never be complete, for the possibilities of sophistication are endless. It is intended to provide the experimenter with the greatest measure of assistance possible without calling on the services of a statistician. But, early and late, statistician there must be, and no computer can take his place. It can speed the work of analysis, it can eliminate non-significant matter, it can perform routine testing, but it cannot displace him entirely. What it can do is to release the calls upon his time, and the duration of such calls. He becomes a consultant rather than a first-aid man, and so his range over the research area can be widened.

Output from such analyses requires simplification and explanation. Preferable to a matrix printout, or a column of coefficients and standard deviations, is a printout of an equation or equations in standard form, with degrees of significance listed, where possible, in English. A development is "This might mean either that variable X_5 has hardly any effect upon the value of Y_1, or that the experimental space has not been adequately covered". If the local statistician is there for consultation, then why not use also the great texts? For example, "In this connection, Fisher (or Pearson or Hald) says . . . (and here follows a quotation from a standard text)". This is an interesting field, which is under study.

Let me now return to our philosophy. Information for management can be simplified by grouping, and this, with marshalling and summing, is normally all that is done with sales data. By the use of CRESTS (see Craig, 1966), both numerical information of this kind and the various types of coded information (product code, customer code, country code) can be translated into the standard input form for COSMOS, yielding the possibility of analysis (including graph plotting from the scatter diagram subroutine) of a much more sophisticated type. On the other hand, data derived from laboratory experiments, pilot plants and factories, can, as well as providing experimental results, of itself or combined with data from other sources be treated by CRESTS to produce tabulations, as required. The existence of COSMOS can therefore be said to strengthen the validity of both the first and second tenets of our computer philosophy. Bill O'Brien dilates upon COSMOS in the paper which follows.

References

CRAIG, T. K. (1966). "CRESTS—Courtaulds Rapid Extract, Sort and Tabulate System", *The Computer Journal*, Vol. 9, p. 3.
MORONEY, M. J. (1951). *Facts From Figures*. Penguin Books Ltd.