STATISTICS

Matching and Prediction in the Social Sciences

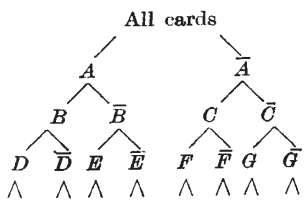
Matching processes are constantly in use in the social sciences. Often their use is explicit, as in studies of the relative effects of different experiences or procedures. In other cases they are implicit, as in the use of specific controls for the selection of consumer panels or in quota surveys. They are also implicit in many extrapolation processes. Even so, the variables used for matching processes are in most cases of non-empirical origin, and are selected on the basis of either custom or hunch. The result may well be that the matching is inadequate, sometimes completely so.

Empirical procedures can, of course, be used. The principle of the empirical method is that, for a variable to be relevant in matching, it must be correlated to some appreciable degree with whatever it is that is being studied (for example, a specific attitude, buying behaviour). If all the proposed matching variables are arranged in a correlation matrix, multiple correlation methods[1] can be used to select from them that combination of variables which gives the highest multiple correlation with whatever is being studied (that is, the criterion). These variables are then used as the basis for matching, and this usually proceeds through a form of prediction, perhaps involving the use of a regression equation.

This sort of work is, however, most arduous. Matching on the empirical principle requires that each fresh criterion must have its own set of matching variables. Also, the development of a correlation matrix is specially inadequate where many of the available variables involve a 'yes' or 'no' response.

Several years ago I developed a technique of the latter kind and have used it in various inquiries[2,3]. But a much simpler and more adequate technique can be used, and I have applied this in more recent studies[4].

The method involves matching on the principle of simple classification. Suppose the matching is being done to allow the study of the effect of different stimuli on criterion X. The following steps are involved. (*a*) Numerous proposed matching variables are included in the testing phase. (*b*) Each of them is then analysed against score on criterion X, and that one which has the greatest association with, or power to predict, the criterion score is selected as the first-stage predictor (or matching variable). This variable will then serve to split the record cards of the informants into at least two groups, A and $\bar{A}$. (*c*) Next, this process is repeated with all the cards in pack A, and the variable with the highest predictive power is selected as the second-stage predictor within that pack. Call this one B, and let it split the A pack into groups B and $\bar{B}$. (*d*) Now repeat the process with pack $\bar{A}$ and select the best predictor within that pack. Call this C and let it split $\bar{A}$ into C and $\bar{C}$.



The vital point here is that the second-stage predictor in pack A is not necessarily the same as the second-stage predictor in pack $\bar{A}$. In other words, what matters in one group may well differ from what matters in another. Multiple correlation procedures do not allow for this. (*e*) The splitting process is continued, so far as it can be taken, on the pattern shown above. The final set of categories or sub-groups defines the matching composite.

There is no need to employ only two-way splits of card packs. Three or more splits can be used at any stage.

The device for selecting any one predictor in the composite is a simplified version of the chi square calculation. The cards are analysed on a four-cell system: 'high' versus 'low' scores on the criterion test and 'yes' versus 'no' on the item being tested as a possible matching variable. The matching power of this trial item is given in the form of a direct count of the number of cases in the four cells which deviate from what might be expected on a null hypothesis (that is, a hypothesis of no association or predictive power at all). Matching power so derived is more accurate than when gauged from most correlation coefficients. It is on the basis of this selection procedure that the final matching composite is built up as described above.

The use of the selected composite of matching variables to achieve matching is also very straightforward. Suppose that in the study of criterion X (for example, buying behaviour), the record cards for sample I are to be matched to those for sample II. (*a*) Split each sample into sub-groups on the above pattern. Suppose this yields 40 sub-groups (for each sample). (*b*) Now equate the number of cards in each sub-group (in the sample I pack) to its equivalent sub-group in the sample II pack. This can be done on random principles, but the equating process must not involve throwing out more than a few record cards. Discarding of cards can be all but avoided if the device is used of multiplying all the pack II sub-groups by a constant sufficient to raise their numbers well above the general level of the numbers in pack I. This constant may be 2 or 3 or 4 depending on the relative sizes of the two samples. (*c*) Alternatively, the average criterion score for each sub-group in pack I can be weighted (separately) by the number of cards in the equivalent sub-group in pack II. In this way, there need be no loss of cards at all, though slight statistical problems may arise.

I have been using the 'classification system' of matching since 1957. There are points at which it is similar to the technique described in the recent communication by Williams and Lance[5].

W. A. Belson

Research Techniques Unit,
London School of Economics and
Political Science,
Houghton Street,
London, W.C.2.
Jan. 8.

[1] Garrett, H. E., "Statistics in Psychology and Education" (1947).
[2] Belson, W. A., *App. Statistics*, 5 (3), 195 (1956).
[3] Belson, W. A., *Pub. Opin. Quart.*, 22 (1), 11 (1958).
[4] Belson, W. A., Rep. British Broadcasting Corporation (in preparation).
[5] Williams, W. T., and Lance, G. N., *Nature*, 182, 1755 (1958).