Published in *Journal of Advertising Research*, 10 (4), 1970, 27-30.

How to Avoid Exploratory Research

J. Scott Armstrong

Studies in marketing research often start with data rather than with a theory. This exploratory or inductive approach is at odds with the more preferred scientific method where the theory precedes the data in any single research study. (See, for example, the discussion by Francis, 1957)

Because exploratory research is common, however, one might argue that it is of some value. A number of researchers have claimed that the exploratory approach leads to new and useful theories. But there is also the danger that the research will produce false leads or useless theories.

An attempt is made in this paper to illustrate the dangers inherent in the exploratory approach. The question of whether the potential benefits are large enough to outweigh the dangers is left to the reader.

One of the more popular techniques which has been used for exploratory research is regression analysis, in particular, a stepwise version. The following example illustrates the exploratory use of stepwise regression.

Tom Swift, a researcher for the International Caribou Chip Co., was asked to study the international market for caribou chips. Management was interested in some explanations as to what caused variation in sales levels among countries – e.g., why were sales higher in France than in Germany?

The company had reliable data on industry sales for caribou chips for 31 countries and Swift conducted a search for possible explanatory variables. Fortunately, the quality and accessibility of international data have been improving rapidly over recent years (e.g., see Russett, 1964). As a result, it was possible to obtain data on 30 variables which might have some relationship to sales of caribou chips. The task then was to determine which of these 30 variables were, in fact, most closely related to the sales level in each country.

The sales rates in each country were regressed against the explanatory variables by means of a stepwise regression program, using the UCLA Biomedical O2R program (Dixon, 1967). This is a "step-up program" – i.e., the variable having the highest gross correlation with the dependent variable is first entered into the regression. Next, the variable with the highest partial correlation (controlling for the variable already .in the regression is entered), etc. Only variables whose coefficient had a t-statistic of greater than 2.0 were retained in the model. (Swift interpreted this as implying that there was-less than a five per cent chance that the true coefficient was zero.)

Three of the 31 countries proved to be "outliers" on the initial regression analysis. The deviation between actual and predicted sales for each of these three countries was so large that there was less than a 10 per cent probability of such an event occurring by chance.

A careful examination of these three outliers led to the conclusion that they were not comparable with the 28 other countries, and they were dropped from the analysis. The regression was then reanalyzed. The effect of dropping the three countries was to increase the coefficient of determination, R^2, by about 0.15.

The second, and final, regression run is presented in Figure 1. The last three variables did not contribute substantially to the fit of the model (in total they increased R^2 by about 0.10), but Tom was following his prespecified rule of including all variables where the t-statistic was greater than 2.0.

Figure 1
Regression Analysis of Caribou Chip Sales By Country (N = 28)

$Y = -.24 + .43x_1 - .33x_2 + .29x_3 + .84x_4 - .31x_5 - .25x_6 + .20x_7 - .18x_8$
t-statistics (5.5) (3.9) (3.9) (8.7) (3.7) (3.0) (2.5) (2.1)

where Y = Sales of Caribou Chips
x_1 = Rate of population change
x_2 = Proportion population between 15 and 64 years of age
x_3 = Literacy rate
x_4 = National income
x_5 = Urbanization
x_6 = Yearly rainfall index
x_7 = Temperature index
x_8 = Number of household goods

$R^2 = 0.85$ (adjusted for loss of degrees of freedom in estimating parameters)

Tom was rather careful about the analysis. He had followed what he believed to be conventional practice, and the methods were objective in the sense that he had followed prespecified statistical criteria. He had also limited his data-fitting to running only two versions of the model.

Tom was pleased with the results the adjusted R^2 of 0.85 for cross-sectional data appeared to be impressive, and some of the t-statistics were very high. The likelihood that he was on to some key relationships in the international market seemed to be very great.

While a rationale was being prepared to explain each of the interesting relationships, same unfortunate news was received. In submitting the data for computer analysis, a mistake had been made: The data which were actually analyzed were not the international data. Instead, they were random normal deviates collected from Rand's book of random numbers. (The names of the variables had been assigned prior to the analysis of the data.)

The example above shows how the use of regression analysis may produce false leads which, if followed, may lead to further unproductive research.

Admittedly, the example is rather extreme in making its point; stepwise regression was used and, in addition, the number of variables to choose from is large relative to the number of observations.

It requires only a modest search of the literature, however, to produce similar studies. For example, Harvey's (1953) study was a good deal more extreme than the Tom Swift study. He explained differences among the sales of 22 matched pairs of navels by choosing from over 500 variables.

What may be done to avoid being misled by the data? The obvious answer, suggested earlier, is to start out with as well-specified a theory as possible, and *then* analyze the data. In other wards, try to avoid doing exploratory research. Some people might argue, however, that in situations where the prior knowledge is very poor the best strategy is to get on with the data analysis. For this latter situation, there are some safeguards which may be taken to reduce the likelihood of being misled.

Following are two strategies-first, on how to avoid exploratory research and, second, on how to reduce the risks involved with exploratory research.

On Avoiding Exploratory Research

Perhaps the primary contribution of "econometrics" is that it emphasizes that, *for a given study,* the theory should precede the data. Certainly this is a matter of degree. Even the exploratory approach calls for some prior theory in that certain variables must be selected from the infinite number available. The question, then, is

> To what extent may theory be specified by an a priori analysis? That is, to what extent may one avoid exploratory research?

The history of economics reveals a persistent trend toward a more extensive use of the a priori analysis. The various stages of development seem to be roughly along the following lines:

1. Specification of a small number of causal variables
2. Specification of the direction of each of the relationships
3. Specification of the pattern of causality
4. Specification of the functional form
5. Specification of the magnitude of each relationship.

At the current time, 1 and 2 are accepted as standard requirements for an econometric study; much work is currently being carried out using 3 and 4 while 5 is controversial and seldom employed. Below, each of these stages of a priori analysis is discussed in more detail.

Specification of a Small Number of Causal Variables: The researcher should not rely upon the data to tell him which variables are important. He should try to utilize the a priori analysis to restrict the number of causal variables which he will submit to analysis. He should not accept all of the variables which happen to be available as possible causal variables.

One very common way to reduce the number of possible causal variables is to consult with experts in the given problem area. Another way is to examine previous studies or theories in related areas. Or the researcher may simply have to rely on his own hunches. In some way, however, he should attempt to reduce the number of causal variables to a small set.

What is a "small set" of causal variables depends upon how much information is contained in the data which are to be analyzed. A simple approximation of the amount of information is the number of *independent observations.* The number of variables should be small relative to the number of observations. This advice is especially important in cases where the causal variables are correlated with one another.

Specification of the Direction of Each of the Relationships: If the researcher cannot make an a priori specification of the direction or sign of the relationship, it is probably best to omit that variable from the regression. Specification of the sign may also benefit from discussions with experts or from a knowledge of previous studies.

Specification of the Pattern of Causality: The direction of causality should flow from the independent variables to the dependent variable. If there are also important causal relationships flowing from the dependent to the independent variables, then a series of equations should be used in analyzing the data. (See Johnston, 1963, for a further discussion of this problem.)

Specification of the Functional Form: An examination should be made of the various measurement scales. Is the relationship between the dependent and independent variable constant over the range of data? If not, is it possible to make changes in the scales which would ensure that the relationship is constant?

At the current state of knowledge, and with the computer programs which are available, it is strongly recommended that the researcher select a functional form which allows the various causal effects to be combined in an additive fashion.

This recommendation is really not as restrictive as it might sound. The use of transformations provides a great deal of flexibility to the researcher. (For useful discussions on transformations see Johnston, 1963; Frank, 1966; and Prais and Houthakker, 1955.)

The most common functional form for work in econometrics is the multiplicative or "log-log" model. One example of the multiplicative model is:

$$Y = aI^{B_1} P^{B_2} \qquad (1)$$

It may be transformed into the additive form by taking logs on both sides:

$$\text{Log } Y = \log B + B_1 \log I + B_2 \log P \qquad (2)$$

The model is unit-free and is easy to interpret (as the B 's represent elasticities). In most cases in advertising and marketing research, the multiplicative model would be the appropriate functional form. But the key point to make is that the researcher should try to use the a priori analysis in deciding upon the appropriate functional form.

The Specification of the Magnitude of the Relationships: The specification of the magnitude of a relationship (e.g., an income elasticity of + 1.3) on an a priori basis is still controversial. Theil and Goldberger (1961) provide one of the early examples of such an approach, but there has been little progress in recent years.

In addition to specifying the magnitude, the researcher will also, of course, have to specify his degree of confidence in each estimate if he intends to use the regression model to update his estimates.

Is it reasonable to presume that the researcher has sufficient information with which to completely specify the model? In other words, can he write down the equation solely on the basis of his prior knowledge? It would seem that there are numerous cases where he can. For example, Armstrong (1968) used such'- an approach in the development of a sales forecasting model. The performance of this "a priori model" proved to be rather good in a predictive test.

Conclusion: The exploratory approach tends to be akin to a projective test in psychology. The researcher has much more freedom to read his own ideas into the data. He can experiment with an almost unlimited number of possible models and choose one which fits his needs.

The use of the a priori analysis greatly restricts the freedom of the researcher in massaging the data. It represents an attempt to force the researcher to develop most of the model before he looks at the data.

In summary, it is possible to avoid many of the problems in exploratory research by simply moving away from this type of research and relying more heavily upon an a priori analysis.

What may be done to avoid problems in those cases where exploratory research is necessary? In particular, what may be done when stepwise regression is used in exploratory research?

Probably the most useful step which may be taken is to obtain measures of statistical significance, and there are a number of ways to obtain these measures. The most commonly used approach, the t-statistic of the regression coefficient, yields a rather poor measure when stepwise regression is used. Two alternative approaches – the split-sample technique and the Monte Carlo simulation technique – provide certain advantages for testing the significance of the coefficients from a stepwise regression.

Why is the t-statistic of the coefficient a poor indicator of statistical significance? The use of the t-statistic is based on the assumption that a model was specified in advance of the analysis. But the stepwise regression model is not fixed in advance. On the contrary, the objective is to experiment with a series of models in an attempt to find the "best" one. It is analogous to using 100 hypotheses and finding that five of them are statistically significant at the five per cent level and then using these five significant hypotheses for further work on the assumption that each differs from the null hypothesis. In short, the trial and error approach renders the t-statistics to be almost meaningless when there are many potential causal variables from which to choose.

To avoid drawing misleading inferences about significance when using the exploratory approach, one might use the split-sample technique. The data are randomly split into two (or more) subsamples. One subsample is then used for model building. That is, the exploratory manipulations are performed on only a part of the original sample. The regression from this analysis subsample is then fitted to the other subsample. In this second step, however, only a single model is used -i.e., there are none of the exploratory manipulations. The statistical significance may then be judged from the model based on these virgin data.

In the Tom Swift example, the split sample approach is at somewhat of a disadvantage since the sample size is small in an absolute sense. In such cases, Monte Carlo simulation may be used to examine significance. This approach could utilize random normal deviates.

A series of regression runs then may be made where each run conforms to the same procedure used on the actual data. In other words, a series of runs would be made, each one of which would look like the Tom Swift study. Each run would utilize the same sample size, the same number of variables, the same treatment for outliers, and the same rules for entering or removing variables as were used in the original problem. It would, however, be based on different samples of random data. The researcher could plot the distributions for each coefficient, and the results from a particular study could then be compared against the distributions generated by the Monte Carlo simulation to see if they are unusual.

For an inexpensive approximation of running the simulation study, one might refer to the results of the simulation studies on random data by Ando and Kaufman (1966). Tables 3 and 4 of their study present the distribution of R^2s (for a sample of 50 simulation runs) where the sample size varies from 10 to 200, the number of potential causal variables ranges from 10 to 50, and the number of causal variables included in the model ranges from two to five.

Avoiding Problems

The best way to avoid the problems associated with using regression analysis in exploratory research is not to do exploratory research. More precisely, one should try to reduce the degree of exploratory work in a given study by the extensive use of a priori analysis.

If one does feel compelled to use the exploratory approach, it seems necessary to examine the statistical significance of the estimated relationships. The t-statistics for the coefficients provide rather poor measures of significance, and alternative approaches to estimating statistical significance should be employed.

The split sample approach is useful if the sample size is large relative to the number of variables. Where sample size is small, as in the Tom Swift study, a Monte Carlo analysis may be used. The latter approach would, in all likelihood, have saved Tom.

References

Ando, A. and G. M. Kaufman (1966), "Evaluation of an ad hoc procedure for estimating parameters of some linear models," *Review of Economics and Statistics*, 48, 334-40.

Armstrong, J. Scott (1968), "Long-range forecasting for international markets: The use of causal models," in Robert L. King (ed.), *Marketing and the New Science of Planning.* Chicago: American Marketing Association.

Dixon, W. J., ed. (1967), *BMD: Biomedical Computer Programs*. Berkeley: University of California Press.

Francis, R. G. (1957), "The relation of data to theory," *Rural Sociology*, 22, 258-66.

Frank, Ronald E. (1966), "Use of transformation," *Journal of Marketing Research*, 3, 247-253.

Harvey, John (1953), "The content characteristics of best-selling novels," *Public Opinion Quarterly*, 17, 91-114.

Johnston, J. (1963), *Econometric Methods.* New York: McGraw-Hill.

Prais, S. J. And H. S. Houthakker (1955), *The Analysis of Family Budgets.* Cambridge: Cambridge University Press.

Russett, B. M., et al. (1961), *World Handbook of Political and Social Indicators.* New Haven: Yale University Press.

Theil, H. and A. S. Goldberger (1961), "On pure and mixed statistical estimation in economics," *International Economic Review*, 2, 65-78.