11-1971

Exploratory Analysis of Marketing Data: A Reply

J. Scott Armstrong
University of Pennsylvania, armstrong@wharton.upenn.edu

Follow this and additional works at: https://repository.upenn.edu/marketing_papers

Part of the Marketing Commons

Exploratory Analysis of Marketing Data: A Reply

Abstract

Response to Crocker's article, addressing that we were attempting to demonstrate that there are situations where trees are expected to be more effective than regression.

Disciplines
Business | Marketing

Exploratory Analysis of Marketing Data: A Reply

J. Scott Armstrong

The major problem is that Crocker misinterprets the intent of our article. Before considering the specific points he raises, let me restate the basic intent: regression analysis is commonly used in the development of predictive models in marketing. Some evidence now exists to suggest that, in certain situations, this practice may not be optimal. An alternative method, tree analysis, may lead to better predictive models in these situations.

We tested the above hypothesis in a situation where regression analysis is often used in an exploratory manner—the site location problem. For example, Kotler [4, p. 442] implied that the Rayco Company was following good practice when it used stepwise regression to develop a model to predict volume at retail locations. And I am aware of at least one oil company which has used stepwise regression for location of service stations.

In our judgment, problems with the data in the site location problem should tend to render the exploratory use of regression analysis less effective than the use of trees. In short, we were attempting to demonstrate that there are situations where trees are expected to be more effective than regression.

I would now like to consider the specific points raised in Crocker's article.

The Use of Dummy Variables

We stated that the use of dummy variables to represent final cells would be cumbersome if one did not know a priori which cells were important. It is admirable that Crocker would find this simple, since there are a few billion possible cells.

Interpretation vs. Prediction

We agree that tree analysis may also be useful in interpretation and that the use of trees followed by the use of regression may be advantageous. Sonquist [7] strongly recommends such a procedure also. Although we were interested solely in prediction, this is a useful comment about interpretation. For those people interested in interpretation it should also be noted that the dependent variable was incorrectly labelled gallons per week. It should have been gallons per month.

"Rules of Thumb" Do Not Include All Factors

A rule of thumb is a rule of thumb is a rule of thumb. Crocker might also have mentioned that a priori knowledge and the amount of autocorrelation are factors which affect the "minimum sample size" for regression. I do not think that Ball's rule of thumb would differ drastically on the downward side—i.e., it would be wise to regard it as a minimum.

Using *A Priori* Information in Exploratory Research

Crocker suggests that the analyst use his a priori information to avoid some of the data problems. I strongly agree on this point and have, myself, advocated such a strategy [1]. We did not intend to advocate the use of exploratory research. Rather, we were examining what happens when people do use the exploratory approach. Exploratory research was defined (p. 488) as research in which little a priori knowledge is used—"one typically specifies only a large list of possible explanatory variables." We did not attempt to generalize beyond exploratory research.

If one were to try to do so, however, it might be useful to set up a polar extreme to exploratory research. This might be designated as "theory-based" research and would involve having the researcher make extensive use of a priori knowledge. In other words, the exploratory end of the continuum asks for as little input from the researcher as possible and the theory-based end asks for as much as possible. Alternative techniques might then be examined in each of these polar extremes as illustrated below:

	Trees	Regression
Exploratory research	A	B
Theory-based research	C	D

A number of hypotheses might be drawn from this figure:

H_1: Trees are superior to regression in exploratory research under certain specified conditions $(A > B)$.

H_2: Trees are superior to regression in theory-based research under certain specified conditions $(C > D)$.

H_3: Theory-based research is preferable to exploratory research when regression is used $(D > B)$. (This is one of the basic beliefs of econometricians.)

H_4: Theory-based research is preferable to exploratory research when trees are used $(C > A)$.

Current Practice vs. Ideal Practice

We defined the problem to hold a priori information constant and minimal. The question still exists as to whether the regression analyst could do a better job of exploring without using much a priori information. Obviously there are an infinite number of ways he could explore. We selected a method before we analyzed the data, and we felt that the most interesting comparison was to examine standard practice rather than some ideal practice (p. 489).

Are there better ways to use regression analysis in exploratory research? We were not interested in the question. My personal opinion, however, is that there is little to be gained here. For example, the introduction of interaction terms greatly increases the possibility of obtaining spurious correlations. And the cleverness of researchers enables them to find relationships where they do not exist. In short, I think that standard practice is reasonable for exploratory research.

Maximizing R^2

The confusion on this point will be reduced by distinguishing between "state" and "process." We agree that the final state which maximizes R2 should include all those variables with t-statistics greater than one [3]. More than one final state may meet these requirements [9]. To our knowledge there is no process, short of trying every possible combination, which will guarantee finding the state which maximizes R2. We did not claim that our process would maximize R2, since our interest was in using a procedure representative of current practice.

Should the Criterion Be Dictated by Technique or by the Problem?

We selected the criterion measure prior to data analysis because we felt that the criterion should be selected in light of the site location problem rather than of the techniques available for solving the problem.

Given the criterion selected, the typical researcher would use a standard and convenient program, such as regression. We used minimum variance procedures in developing each prediction model in order to place each on an equal footing. While we do not agree with Crocker's emphasis upon choosing a criterion to fit the technique, our argument would have been strengthened had we used alternative criteria which were appropriate to the site location problem.

The Generality of the Results

We started with a hypothesis drawn from the literature, selected a situation in which we thought the hypothesis was appropriate, and found results consistent with the hypothesis. This should add to one's confidence in the hypothesis. After our study was completed, we found additional evidence to support our hypothesis: study which examined empirical data where interaction was suspected [8] and a study which used artificial data [7].

Our study met the conditions of the problem that we set out to solve. It did not meet those of the problem Crocker would like us to have attacked.

Comments on Summary

We agree that handicaps were imposed on regression in our comparison. That is precisely what the hypothesis says. A certain type of situation has been examined where regression analysis is often used in an exploratory manner. In view of the handicaps imposed by this situation, we felt that tree analysis would lead to better predictions. Given the situation, Crocker has not shown that the procedure used for comparison put regression at a disadvantage.

There seems to be an implication that we were biased against regression and in favor of trees. Bias on the part of the researcher can have some surprising implications [6]. We hope that such was not the case with us. Also, we tried to avoid this possibility by making decisions on the nature of the test and of the criterion prior to examining the data.

Crocker doubts whether comparisons of methods need to be made. We (along with Platt [5] and Chamberlain [2]) feel that the comparison of alternative methods is essential to scientific advancement. Has Crocker given the reader a more complete picture? Or is he too close to the trees (regression) to see the forest? The moral of our study was that regression analysis is not the only game in town—and, in certain situations, it may not be the best game.

References

1. Armstrong, J. Scott. "How to Avoid Exploratory Research," *Journal of Advertising Research*, 10 (August 1970), 27-30.

2. Chamberlain, T. C. 'The Method of Multiple Working Hypotheses," *Science* (February 7, 1890), reprinted in 148 (May 7, 1965), 754-9.

3. Haitovsky, Yoel. "A Note on the Maximization of R^2," *American Statistician*, 23 (February 1969), 20-1.

4. Kotler, Philip. *Marketing Management*. Englewood Cliffs, N.J.: Prentice-Hall, 1967.

5. Platt, John R. "Strong Inference," *Science*, 146 (October 1964), 347-53.

6. Rosenthal, Robert and Ralph L. Rosnow. *Artifact in Behavioral Research*. New York: Academic Press, 1969.

7. Sonquist, John A. *Multivariate Model Building*. Ann Arbor: Survey Research Center, University of Michigan, 1970.

8. Stuckert, R. P. "A Configurational Approach to Prediction," *Sociometry,* 21 (September 1958), 225-37.

9. Weiss, Moshe. "Letter to the Editor," *American Statistician*, 24 (June 1970), 20.