Exploratory Analysis of Marketing Data: Trees vs. Regression

J. Scott Armstrong
Assistant Professor of Marketing, The Wharton School
and James G. Andress
Consultant at Booz, Allen, and Hamilton, Inc., Chicago

Reprinted with permission from *Journal of Marketing Research*, (1970), 487-492.

This article compares the predictive ability of models developed by two different statistical methods, tree analysis and regression analysis. Each was used in an exploratory study to develop a model to make predictions for a specific marketing situation.

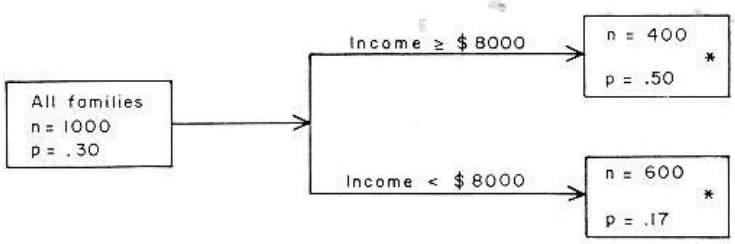
The Statistical Methods

The regression model is well known and no description is provided here. Tree analysis, however, is less well known. To add to the confusion, it has been labeled in a number o£ ways – e.g., multiple classification, multilevel cross-tabulations, or configurational analysis. Whatever the names, the basic idea is to classify objects in cells so that the objects *in the cells* are similar to one another yet different from the objects *in other cells*. Similarity is judged by the score on a given dependent or criterion variable (which differentiates this method from cluster or factor analysis, where the similarity is based only upon scores on a set of descriptive variables).

Tree analysis is an extension to *n* variables of the simple cross-classification approach. Consider the following example: a researcher is studying the factors which determine whether a family owns two or more automobiles. He finds that income may be used to classify respondents. Illustrative results for his sample are provided in Figure 1.

Figure 1

EXAMPLE OF TREE ANALYSIS WITH ONE EXPLANATORY VARIABLE[a]

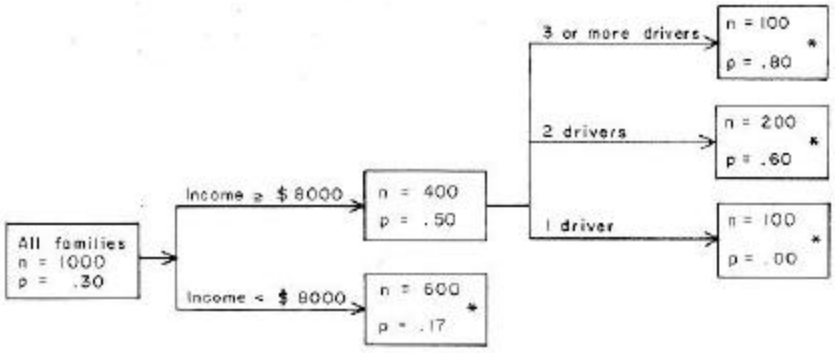


[a] n = sample size
p = proportion of sample having 2 or more cars
* = designates final cells.

He then decides that the number of drivers in the family may also be important for high-income families. This variable is introduced into Figure 2.

Figure 2

EXAMPLE OF TREE ANALYSIS WITH TWO EXPLANATORY VARIABLES[a]



[a] n = sample size
p = proportion of sample having 2 or more cars
* = designates final cells.

He is now able to predict the probability that a family has two or more cars by having information about the number of drivers and the income of the family. This probability is estimated to be equal to the proportion of families in that cell which have two or more cars.

The above example could be handled in an equivalent manner by regression analysis *if* one knew what cells were important. Dummy variables could be used to represent each of the "final cells" in Figures 1 or 2. That is, if the family fell into a particular cell, it would receive a one; otherwise it would receive a zero. The number of variables would be the number of final cells less one (to prevent redundancy) [12]. While equivalent results are obtained, this use of the regression model is very cumbersome.

More commonly, of course, the regression mQde1 is used to measure relationships rather than cell means. In terms of the above example, the regression model would use income and number of drivers as variables (rather than using membership in the final cells). It explains how car ownership changes as income changes and as number of drivers changes. It is in this latter sense that we treat the regression model in this article.

Relative Advantages of Regression vs. Trees

Regression analysis is more powerful than tree analysis. That is, it utilizes the sample observations in a much more efficient manner. Each observation measures each relationship. As a result, regression analysis can handle problems with many variables by using only a moderately large sample size. Ball [1] suggests a rule of thumb that ten observations per variable are a minimum requirement for regression.

Tree analysis, on the other hand, requires large sample sizes. For example, if five explanatory variables are used and if two-way splits are made for each variable, a total sample size of about 1,000 is required for an average cell sample size of 30. Since it typically turns out that the observations are not divided equally,[1] a sample size of much more than 1,000 would be required if it were desired that each cell had at least 30 observations. In comparison, the preceding paragraph suggests that a sample size of only 50 would be adequate for regression.

The power of the regression model is gained at the expense of its more restrictive assumptions: see [7] for a formal summary. As a result of these assumptions, regression analysis encounters difficulties when one or more of the following types of problems are encountered in the data:

[1] Another way of stating this problem is that tree analysis requires larger sample sizes as the degree of multicollinearity rises. The same is true for regression analysis.

1. *Interaction*: e.g., the relationship between X1 and Y is dependent upon the level of X2.
2. *Nonlinear effects*: e.g., the relationship between X1 and Y is dependent upon the level of X1 itself.
3. *Causal priorities*: e.g., X1 causes X which, in turn, causes Y.

A more complete discussion of these data problems may be found in the paper by Morgan and Sonquist [9].

Substantial gains have been made by econometricians in adapting the regression model to handle interaction, nonlinear effects, and causal priorities. To a great extent, however, these procedures (e.g., the specification of a *small* number of explanatory variables, the use of transformations, or the use of simultaneous equations) require a substantial amount of a priori knowledge. The concern in this paper, however, is with exploratory research where one typically specifies only a large list of possible explanatory variables. Under these circumstances, the data problems remain serious.

Tree analysis, while less powerful than regression analysis, makes fewer assumptions about the data. To a great extent, this multiple classification avoids the problems of interaction, nonlinear effect, and causal priorities. If power is not a factor – i.e., if large sample sizes are available – cross-classification would appear to offer a superior method for analyzing data subject to the type of problems discussed here.

Support for the benefits of trees over regression has been provided with simulated data [10] and with actual data [2, 8, 9, 10]. However, the studies which used actual data have been restricted to a comparison of the ability to fit the model to the data. We have been unable to find any published studies which compare the predictive ability of the models.

In summary then, our hypothesis was that trees are superior to regression analysis: (1) when exploratory research is to be carried out, (2) on data having problems with interaction, nonlinearities, or causal priorities, and (3) where large sample sizes are available. The criterion for superiority was based on the ability of the resulting model to make predictions on new data.

Research Design

The problems discussed above occur very frequently with marketing data. The basic plan of the study, then, was to find marketing data for a specific situation and to compare the tree and regression analyses.

The Data

Data from almost 6,000 gas stations of a single brand were obtained from a major oil company operating in the eastern United States. The data were collected by the oil company's salesmen.

The objective of the model was to predict the average weekly gasoline volume for each gas station on the basis of data on 19 explanatory variables. Table 1 summarizes the explanatory variables which were available. As the study was considered to be exploratory research, no attempt was made to discuss the reasons why each variable might be important to the sales of gasoline.

A cursory examination of the data in Table 1 indicates that there may be problems with interaction (e.g., between variables 4 and 14), with nonlinear effects (variable 8), and with causal priorities (variables 7 and 11). As a result, the data appeared to be suitable for testing the hypothesis.

The Statistical Procedures

Two rather well known and widely available programs were used to carry out the data analysis. For cross-classification, a modification of the AID program [11] was used; details are in [6]. For regression, the BMDO2R stepwise regression program [3] was used. Each of these programs is designed for exploratory research to achieve the best fit of a model to the data.

Table 1

COMPLETE SET OF POTENTIAL EXPLANATORY VARIABLES

Variable number	*Variable name*	*Description of categories*
1	Type of outlet	City service station; truck stop; toll road station; car dealer; other
2	Type of area	Industrial; shopping center; other commercial; residential; interstate; open rural
3	Street location	Far corner less than 5,000 cars/day traffic; far corner over 5,000 cars/day; near corner less than 5,000 cars/day; near corner over 5,000 cars/day; not on corner
4	Total traffic	Under 5,000 cars/day; 5,001 to 10,000; 10,001 to 15,000; over 15,000
5	Frontage	Number of feet along street
6	Number of streets with pump islands	One street only; two streets
7	Number of pumps	1; 2; 3; 4; 5 or more
8	Building age	Old; postwar; modern
9	Building type	Conventional; ranch; colonial; contemporary; non-standard
10	Canopy?	No; yes
11	Number of bays	Actual number; 9 or more – 9
12	Rating of general appearance (by salesmen)	Very poor; poor; fair; good; excellent
13	Stamps?	No; yes
14	Open 24 hours?	No; yes
15	Price sign?	No; yes
16	State inspection?	No; yes
17	Rent trailers?	No; yes
18	Motel also?	No; yes
19	Restaurant also?	No; yes

The AID program does essentially the same job that could be done on a card sorter by trial and error. The use of the sorter to split the sample, however, is hardly economical for exploratory research. AID, on the other hand, can examine a large number of feasible two-way splits and take the best one – that is, the one which leads to the greatest reduction in variance. In terms of the example used earlier, a decision would be made by AID as to which variable, income or number of drivers, and which breakpoint (e.g., $6,000; $10,000; $8,000; etc.) will do most to reduce the unexplained variance.

The BMD02R regression program is a step-up version. The first variable selected is the one with the highest correlation to the dependent variable. The next is the one with the highest partial correlation (i.e., correcting for the variable previously entered into the model), etc.

Each program requires the user to prespecify a number of statistical decision rules. We attempted to follow those rules which would most likely be used by a researcher who might employ each technique.

For AID, two key decisions were required. First, and most important, a minimum permissible cell size of 30 was specified. Sonquist and Morgan [11] indicate that a cell of this size or larger should be adequate. The second decision was to specify a minimum reduction in variance of 0.5%. This implies that a split will not be made unless the unexplained variance in gasoline sales is reduced by at least 0.5%. This rule seemed to be in line with previous work published on AID.

For the regression ana1ysis, rules were selected in an effort to maximize the adjusted R^2 (i.e., adjusted for the loss in degrees of freedom due to the variables included in the model). To do this, it is necessary to exclude all variables whose t-statistic is less than 1.0 [4].[2] It is also common to include all variables whose t-statistic is greater than 1.0, although this does not *ensure* that the adjusted R^2 will be maximized. This combination of a necessary condition and of common practice yielded the following rule: "include all those and only those variables which have a t-statistic of at least 1.0."

In a sense, then, what we are presenting here is a contest between two types of researchers. Will the researcher who follows standard rules for tree analysis do better or worse in developing a predictive model than the researcher who follows standard rules for step-wise regression?

Developing The Models

The predictive model developed from AID and that from BMD02R were each obtained from data on the same 2,717 gas stations. These stations, designated as the analysis sample, had been randomly selected from the original 5,717 stations. The remaining 3,000 stations were designated as the validation sample. Each of the models, AID and BMD02R, was permitted to select freely from the variables listed in Table 1. In short, each model was developed from the same information.

The results of the AID analysis are presented in Figure 3. This model accounted for 32% of the variance in the analysis sample. Predictions may be made from this mo del by finding the cell into which a gas station falls and then estimating its volume as equal to the mean value for that cell.

The results from the regression (BMD02R) are presented in Table 2. This equation uses 19 variables out of a possible 37. There were 37 variables to choose from, rather than the original 19, since it was necessary to create dummy variables to express the category-type data [12]. For example, five variables were required to express the six types of areas (variable 2 in Table 1). This model accounted for 33% of the variance in the analysis sample. Predictions may be obtained by taking the measures for a station and putting them into the equation in Table 2.

While the regression and tree models were about equal in their ability to exp lain variations in the analysis sample, we do not have much confidence in such a comparison. Each of these models involves a great deal of data manipulation. As a result, the comparison of the models was carried out on a new set of data as described below.

A Comparison Of Predictive Ability

The 3,000 gas stations for the predictive test data were saved to control for spurious relationships which arise in fitting the analysis sample. The ability to predict with this validation sample should provide a good indication of the ability of the model to predict volumes for new sites.

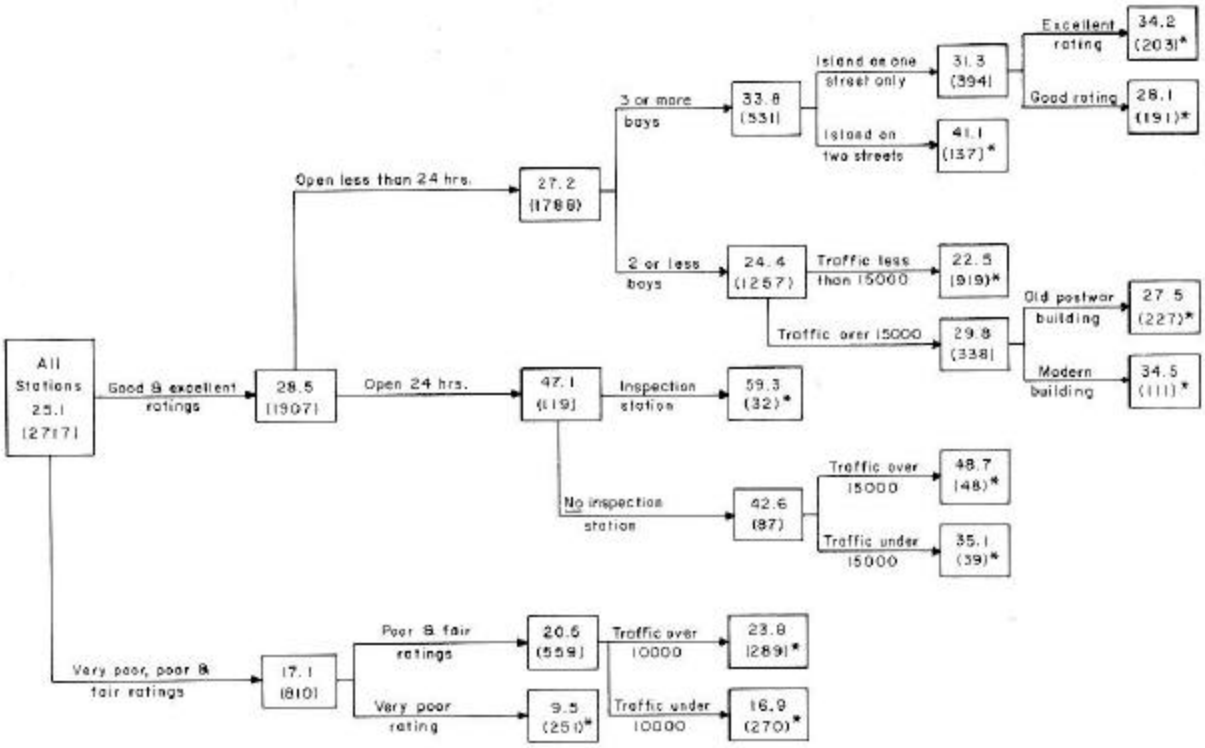
Data for the independent variables were used to obtain a predicted weekly volume of gasoline sales using each model. The difference between actual volume and predicted volume was then found for each gas station. This difference was converted to percentage terms as follows:

$$\left[\frac{\text{actual} - \text{predicted}}{\frac{1}{2}(\text{actual} + \text{predicted})}\right] \times 100.$$

[2] The statement in the text requires one qualification. The highest adjusted R^2 is obtained for only for that set of statistical rules used in the program. Other procedures (e.g., a step-down rather than a step-up program) could lead to different results.

Figure 3

MODEL DEVELOPED BY AID TO PREDICT WEEKLY GASOLINE VOLUME*



* Asterisk designates final cells.

In short, we calculated the percentage error for each gas station using each model. Finally, the mean absolute percentage error was calculated for each model. This latter figure was used as the criterion for the predictive test since it provides a measure which is easily understood (i.e., it is the average error). The use of average sales rather than actual sales in the denominator is merely an attempt to make errors in predicting the scale of operation symmetrical. For example, this criterion shows that if actual sales were 100, then a prediction of 50 is about as good (or as bad) as a prediction of 200. It should be noted, of course, that this criterion is only one of many which could have been selected. It was the only one considered in this study.

Results

The mean absolute percentage error was 57.9% for the regression model but only 41.3 % for the AID model. This difference of over 16% in accuracy would appear to be of some practical significance for decisions involving site location, administrative control, etc. The difference is also statistically significant at the .01 level (using a one-tailed t-test for difference between means).[3]

The results, then, were consistent with our hypothesis and provide some support for the relative advantage of AID over regression for exploratory studies in marketing.

[3] While the parent distribution may not be normal, the f-statistic seems appropriate in this case primarily because the sample size is very large [5, p. 308].

Limitations

This study examined only one situation. As a result, the issue of cross-classification vs. regression is certainly not closed. Further results must be obtained on other types of marketing data in order to assess the extent to which these results may be generalized.

It is also important to note again that this paper was concerned only with exploratory research. No discussion has been presented as to the benefits of the exploratory approach. It was merely assumed to be useful in certain situations.

The results of study were sensitive to our interpretation of the standard rules employed by the researcher. Some experimentation was also carried out with other rules for both AID and BMD02R. The AID program was run again with a minimum cell size of 50, while the regression program was run with a different number of variables entered (e.g., same number of variables as in the AID model) and with different rules for selecting variables (F-level of .05 rather than 1.0 to include and exclude variables). These different models generally led to similar results – i.e., AID superior to regression – but the margin of superiority varied. One regression model, however, was found to be as good as the AID model.

Conclusions

In situations where there are large sample sizes and where the data are subject to interaction, nonlinearities or causal priorities, there is reason to believe that tree analysis will be superior to regression analysis. This study provided support for the use of trees rather than regression for exploratory research in one such situation.

Two predictive models were developed from a sample of 2,717 gas stations. Each used conventional decision rules – following what were thought to be the typical rules-of-thumb. Predictions were then made on the gasoline volume for 3,000 stations (virgin data). The predictions generated by the AID model were substantially more accurate than those generated by the regression model.

We do not, by any means, regard the problem of trees vs. regression as a closed issue. More evidence is required from other situations to test out the generality of the hypothesis tested in this study. Our impression is that the ideal type of data analysis will employ both trees and regression concurrently.

References

1. G. H. Ball, "Data Analysis in the Social Sciences: What About the Details?" *Proceedings of the Fall Joint Computer Conference*, Las Vegas, Nevada, 1965.
2. James M. Carman, "Correlates of Brand Loyalty: Some Positive Results," *Journal of Marketing Research*, 7 (February 1970), 67-76.
3. W. J. Dixon, ed., *BMD: Biomedical Computer Programs*, Los Angeles: University of California Press, 1967
4. Yoel Haitovsky, "A Note on the Maximization of R^2," *American Statistician*, 23 (February 1969), 20-1.
5. William L. Hays, *Statistics for Psychologists*, New York: Holt, Rinehart, and Winston, 1963.
6. Britton Harris, C. Wordley, and D. Seymour, *Program Description: PENAID*, Philadelphia: Institute for Environmental Studies, University of Pennsylvania, 1969.
7. J. Johnston, Econometric Methods, New York: McGraw Hill, 1963.
8. David B. Montgomery and J. Scott Armstrong, "Consumer Response to a Legitimated Brand Appeal," in Johan Arndt, ed., *Insights into Consumer Behavior*, Boston: Allyn & Bacon, 1968.

9. James N. Morgan and John A. Sonquist, "Problems in the Analysis of Survey Data, and a Proposal," *Journal of the American Statistical Association*, 58 (June 1963), 415-35.

10. Laurence I. Press, Miles S. Rogers, and Gerald H. Shure, "An Interactive Technique for the Analysis of Multivariate Data," *Behavioral Science*, 14 (September 1969), 364-70.

11. John A. Sonquist and James N. Morgan, *The Detection of Inter-Action Effects*, Monograph No. 35, Ann Arbor: Survey Research Center, University of Michigan, 1964.

12. Daniel B. Suits, "The Use of Dummy Variables in Regression Equations," *Journal of the American Statistical Association*, 52 (September 1957), 548-51.