An Overview of the SEARCH Program

The ISR program SEARCH was created for analysis of survey data with samples of 1,000 or more, and a clear dependent variable – a behavior or situation to be explained or predicted. And it was to look for non-linearities and non-additivities (interaction effects) assumed away by multiple regression. When the dependent "variable" is a dichotomy, one can call the result a decision tree, and perhaps improve a chi-square.

How is SEARCH related to other "data mining" approaches? Years ago some sociologists suggested starting with a k-way table, broken down by all possible predictors, and then searching for the least costly elimination of each explanatory classification in terms of chi-square. That is dangerous to me, likely to make idiosyncratic decisions, compared to the sequential splitting in SEARCH where the first splits are based on large subsamples.

Then people interested in physiology looked at neural networks and studied the learning process by which successful outcomes strengthened the synapses involved. The basic neural structure was imposed, not selected. In most other areas of behavior or condition, we are still selecting the explanatory structure, not studying its dynamics.

Others searched for clusters of variables, without any particular dependent variable. All these are extensively discussed in Tan, Pang-ning, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*, Addison Wesley, Boston, 2006, and in Witten, Ian H., Eibe Frank, and Mark A. Hall, *Data Mining*, 3rd Edition, 2011.

When some predictors have several classes, why restrict the splits to dichotomies? Partly it is to avoid giving too much preference to predictors with many classes. But mainly, and this is why SEARCH worked, because very few divisions on any variable exhaust its predictive value. For those with a real ranking, such as income brackets, one tries the first group versus the rest, then the first two against the rest, etc., but this does not give unfair advantage over real dichotomies like gender, since the alternatives are not totally independent. Without clear ranking, like region, religion, etc, we test each category against all the rest. And if the dependent variable is a set of categories (e.g., method of getting to work, or voted Republican, voted Democrat, didn't vote) the criterion for best split is to create an increase in chi-square instead of reduction in error variance.

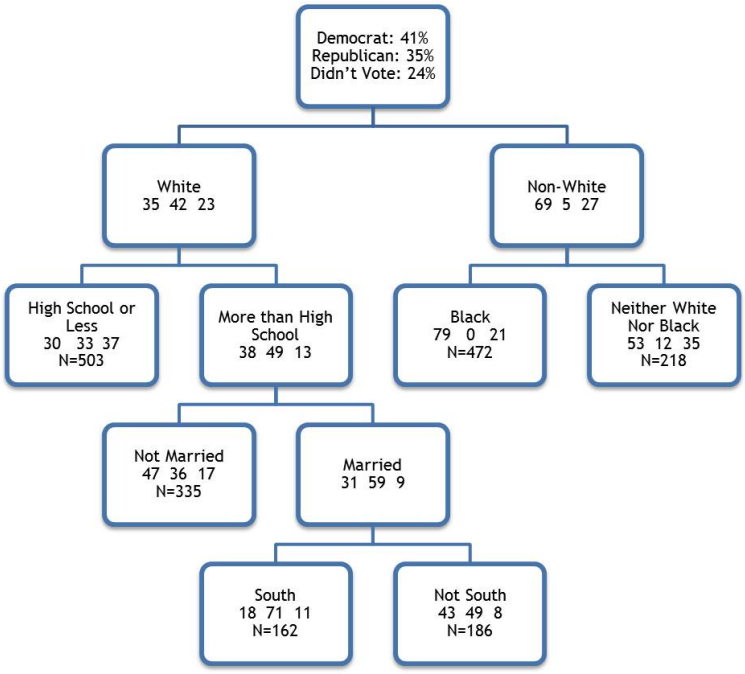
To avoid overdoing it, stopping rules are needed: don't split any group with fewer than 25 cases; reduction in overall error variance must be 0.5%, etc. You can rely on the default options.

When data were scarce, the traditional wisdom was that one set up a model, and tested it with some data. Restrictive assumptions were made to save degrees of freedom for

testing – each predictor was assumed to have a linear effect, and their separate effects were assumed to be additive. Hence, “multiple regression” was developed. Because most sampling error forumlas have the square root of the degrees of freedom (cases minus variables) in the denominator, with *k* predictors, a sample of 2*k* would be enough.

A different approach was used when experimental designs were used, i.e., a decomposition of the variance associated with each predictor and each interaction. Take two predictors, *x* and *w*, the subgroup means on *y*, the dependent variable, would leave the sums of squares around each mean total less than the sum around the grand mean. And with a two-way table, the sum of squares around the *xw* means would be still less, the added gain attributable to interactions. However, with many predictors, and higher order interactions, the number of possible models becomes very large.

Combine this problem of selecting the best model, with the appearance of large rich data sets, many plausible models, and with one further fact: the explanatory power of any predictor is exhausted by a few subgroups. Hence, a sequential splitting could handle non-linearities and interaction effects, providing a root/tree diagram. Such a diagram is optimal for decision-making, since the first divisions are based on large numbers, whereas clustering approaches make risky decisions right at the start (see below).



So SEARCH takes each predictor and asks which binary split would most reduce the error variance (or in the case of a categorical dependent variable most increase the chi-square), splitting the top group from the rest, then the top two, etc. When the predictor

has no natural ranking, as in region, or occupation, the program compares each subgroup with all the others. One could think of ranking a categorical predictor on the basis of its overall relation to the criterion, but that would provide an unfair advantage.

What about tests of significance? First, there is no overall measure of the effect of any one predicator, and second the search procedure has used up “degrees of freedom.” So for those concerned with whether the results can be extrapolated, the only pure solution is to impose the resulting prediction tree on a fresh, independen set of data, preferably from a different time and place. Few samples allow such a division, the exception being the Panel Study of Income Dynamics, which designates four subsamples. Most surveys are “clustered” and random subsamples are sure to include more than one case of a cluster, violating the independence assumption.

So, users of SEARCH need only specify the dependent variable and the predictors, and which predictors have no natural rank, letting the program make the “default” decisions about when to stop and what to print out. When the dataset has a case-weight that can be specified, in order to account for different sampling rates and/or response rates, the tree diagram is cumbersome, but a hierarchical table can easily be edited into a publishable result.

When some predictors are numerical, one converts them into four or five brackets, which loses almost none of their explanatory power. In cases where some predictors are logically prior (exogenous), you may want to use them first, and then analyze the residuals using explanatory variables with uncertain causal direction, as we did in *Productive Americans* (1966). When the dependent variable has many zeroes, it may pay to do a two-stage analysis (first whether zero, then amount for the non-zeroes). And when one explanatory variable is of major concern, one can do a covariance analysis, using the simple regression instead of subgroup means (it will be dominated by differences in level rather than slope). Finally, when a very large sample is available, it makes sense to first take a random selection of 5,000 or so, run SEARCH, then take interesting subgroups, and select 5,000 of each of them for further analysis. Finally, it would be wise to start with an area with already published regression analyses and see whether SEARCH provides new insights.

For further explanations, see James Morgan, “History and Potential of Binary Sebgmentation for Exploratory Data Analysis,” *Journal of Data Science,* 3(2005) readable online at www.sinica.edu.tw/jds. Full documentation and the program itself are available at www.isr.umich.edu/src/smp/search.

Questions or suggestions: Jim Morgan at jnmorgan@umich.edu or Peter Solenberger, who has updated SEARCH to fit well with most of the data management programs. pws@umich.edu.