The Algebra of SEARCH

The algebra behind SEARCH is quite simple. Those familiar with analysis of designed experiments know that attributing to various factors their contribution of reducing the error variance (summed squared deviations around means) requires designs with equal numbers in the cells of tables, something that is not true of sample data. The simplest case is a binary split producing two means, and the sums of squared deviations around those means is a measure of gained information. It is large when the two means are most different and both subgroups are large.

To allow more than two subgroups makes comparison of alternative predictors incomparable, favoring those with many subgroups, but sticking to the simple binary splits plays fair. So SEARCH tries many binary splits: For ranked predictors, it tries splits of the lowest group against all the rest, then the lowest two groups, etc. For categorical predictors like race or region, it tries each class against all the others. One reason this works is that a few categories of any predictor exhaust its explanatory power.

Sometimes the effects are additive, as in analysis of happiness where after a first split on the quality of the convoy of social support, both subgroups then split on health, and the resulting four groups each split on income. But when analyzing wage rates, after a split on college graduates or not, the college grads earn more if they have moved away from where they grew up, but the rest earn more if they live in or near a large city, and people do more unpaid do it yourself work if they are married, live in single family sstructures in rural areas, are well educated and have no childern under 2 at home.

Tests of significance take account of degrees of freedom, cases minus things tried, but SEARCH is trying millions of alternatives. Only a test of the results against a fresh set of data can assure that results are real because the first few splits are based on many cases, and attempts on fresh data tend to replicate well at the start.