QUIRK'S

Marketing Research Review

SURVEY OF ANALYSIS METHODS

Part I: Key Driver Analysis

By: Rajan Sambandam

Editor's note: Rajan Sambandam is vice president/ research at The Response Center, a Fort Washington, Pa., research firm. He can be reached at 215-641-2251 or at rsambandam@response-center.com

Practical marketing research deals with two major problems; identifying key drivers and developing segments. In the first article of this two-part series we will look at key driver analysis and in the second part we will look at segmentation.

Key driver analysis is a broad term used to cover a variety of analytical techniques. It always involves at least one dependent or criterion variable and one or (typically) multiple independent or predictor variables whose effect on the dependent variable needs to be understood. The dependent variable is usually a measure on which the manager is trying to improve the organization's performance. Examples include overall satisfaction, loyalty, value and likelihood to recommend.

When conducting a key driver analysis, there is a very important question that needs to be considered:

- Is the objective of the analysis explanation or prediction?

Answering this question before starting the analysis is very useful because it not only helps in choosing the analytical method to be used but also, to some extent, the choice of variables. When the objective of the analysis is explanation, we try to identify a group of independent variables that can explain variations in the dependent variable and that are actionable. For example, overall satisfaction with a firm can be explained by attribute satisfaction scores. By improving the performance on those attributes identified as key drivers, overall satisfaction can be improved. If the predictors used are not actionable, then the purpose of the analysis is defeated.

In the case of prediction, we try to identify variables that can best predict an outcome. This is different from explanation because the independent variables here do not have to be actionable, since we are not trying to change the dependent variable. As long as the independent variables can be measured, predictions can be made. For example in the financial services industry, it is important to be able to predict (rather than change) the credit worthiness of a prospective customer from the customer's profile.

Beyond the issue of explanation versus prediction, there are two other questions that help in the choice of analytical technique to be used:

- Is there one, or more than one, dependent variable?

- Is the relationship being modeled linear or non-linear?

In the remainder of this article we will discuss analytical methods that would be appropriate if one or both of these questions is answered in the affirmative.

SINGLE DEPENDENT VARIABLE

Scaled Values

Key driver analyses often use a single dependent variable and the most commonly used method is multiple regression analysis. A single scaled dependent variable is explained using multiple independent variables. Typically, the scale for the dependent variable ranges from 5 points to 10 points and is usually an overall measure such as satisfaction or likelihood to recommend.

The independent variables are some measures of attribute satisfaction usually measured on the same scale as the dependent variable, but not necessarily. There are two main parts to the output that are of interest to the manager. The overall fit of the model and the relative importance

The overall fit of the model is often expressed as R^2 or the total variance in the dependent variable that can be explained by the independent variables in the model. R^2 values range from 0 to 1, with higher values indicating better fit. For attitudinal research, values in the range of 0.4 – 0.6 are often considered to be good. Relative importance of the independent variables is expressed in the form of coefficients or beta weights. A weight of 0.4 associated with a variable means that a unit change in that variable can lead to a 0.4 unit change in the dependent variable. Thus, beta weights are used to identify the variables that have the most impact on the dependent variable.

While regression models are quite robust and have been used for many years they do have some drawbacks. The biggest (and perhaps most common) is the problem of multicollinearity. This is a condition where the independent variables have very high correlations among them and hence their impact on the dependent variable is distorted. Different approaches can be taken to address this problem.

A data reduction technique such as factor analysis can be used to create factors out of the variables that are highly correlated. Then the factor scores (which are uncorrelated with each other) can be used as independent variables in the regression analysis. Of course, this would make interpretation of the coefficients harder than when individual variables are used.

Another method of combating multicollinearity is to identify and eliminate redundant variables before running the regression. But this can be an arbitrary solution that may lead to the elimination of important variables. Other solutions such as ridge regression have also been used. But, if in fact the independent variables truly are related to each other, then suppressing the relationship would be a distortion of reality. In this situation other methods, such as structural equation modeling, that use multiple dependent variables may be more helpful and will be discussed later in this article.

Categorical Values

What if the dependent variable to be used is not scaled, but categorical? This situation arises frequently in loyalty research and examples include classifications such as customer/non-customer and active/inactive/non-customer. Using regression analysis would not be appropriate because of the scaling of the dependent variable. Instead, a classification method such as linear discriminant analysis (or its equivalent, logistic regression) is required. This method can identify the key drivers and also provide the means to classify data not used in the analysis into the appropriate categories.

Key driver analyses with categorical dependent variables are often used for both explanation and prediction. An example of the former is when a health care organization is trying to determine the reasons behind its customers dis-enrolling from the health plan. Once these reasons are identified, the company can take steps to address the problems and reduce dis-enrollment.

An example of the latter is when a bank is trying to predict whom it should offer the new type of account it is introducing in the market. Rather than trying to change the characteristics of the consumers, it seeks to identify consumers with the right combination of characteristics that would indicate profitability.

MULTIPLE DEPENDENT VARIABLES

As mentioned above, one problem with multiple regression models is that relationships between independent variables cannot be incorporated. It is possible to overcome this by running a series of regression models. For example, if respondents answer multiple modules in a questionnaire relating to customer service, pricing etc., individual models can be run for each module. Following this an overall model that uses the dependent variables from each model as independents can be run. However, this process can be both cumbersome and statistically inefficient.

A better approach would be to use structural equation modeling techniques such as LISREL or EQS. In these methods, a single model can be specified with as many variables and relationships as desired and all the importance weights can be calculated at once. This can be done for both scaled and binary variables.

By specifying the links between the independent variables, their inherent relationships are acknowledged and thus the problem of multicollinearity is eliminated. But the drawback in this case is that the nature of the relationships needs to be known up front. If this theoretical knowledge is absent, then these methods are not capable of identifying the relationships between the variables.

NON-LINEARITY

All of the methods discussed so far have been traditionally used as linear methods. Linearity implies that each independent variable has a linear (or straight-line) relationship with the dependent variable. But what if the relationship between the independent and dependent variables is non-linear? Research has shown that in many situations, linear models provide reasonable approximations of non-linear relationships and thus tend to be used since they are easier to understand. There are situations however, where the level of non-linearity or the predictive accuracy required is so high that non-linear models may need to be used.

The simplest extensions to linear models use products (or interactions) of independent variables. When two independent variables are multiplied and the product is used as an independent variable in the model, its relationship with the dependent variable is no longer linear. Similarly, other non-linear effects can be obtained by squaring a variable (multiplying it with itself), cubing it or raising it to higher powers. Such models are referred to as polynomial regression models and they have useful properties. For example, squaring a variable can help model a U-shaped relationship such as the one between a fruit juice’s tartness rating and the overall taste rating. Other variations such as logarithmic (or exponential) transformations can also be used if there is a curved relationship between the dependent and independent variables.

The methods described above are not strictly considered to be non-linear methods. In real non-linear models the relationship between the dependent and independent variables is much more complex. It is usually in a product form and linearity cannot be achieved by transforming the variables. Further, the user needs to specify the nature of the non-linear relationship to be modeled. This can be a very important drawback, especially when there are many independent variables. The relationship between the dependent and independent variables can be very complicated, making it extremely hard to specify the type of non-linear model required. A recent development in non-linear models that can help in this regard is the Multivariate Adaptive Regression Splines (MARS) approach that can model non-linear relationships automatically with minimal input from the user.

Non-linear models are particularly useful if prediction rather than explanation is the objective. The reason for this is that the coefficients from a non-linear regression are much harder to interpret than those from a linear regression. The more complicated the model, the harder the coefficients can be to interpret. This is not really a problem for prediction because the issue is only whether an observation's value can be predicted, not so much how the prediction can be accomplished. Hence, if explanation is the objective, it is better to use linear models as much as possible

Artificial Intelligence

The title of artificial intelligence covers several topic areas including artificial neural networks, genetic algorithms, fuzzy logic and expert systems. In this article we will discuss artificial neural networks as they have recently emerged as useful tools in the area of marketing research. Although they have been used for many years in other disciplines, marketing research is only now beginning to realize the potential of these tools. Artificial neural networks were originally conceived as tools that could mathematically emulate the decision making processes of the human brain. Their algorithm is set up in such a way that they "learn" the relationships in the data by looking at one (or a group of) observation(s) at a time.

Neural networks can model arbitrarily complex relationships in the data. This means that the user really doesn't need to know the precise nature of the relationships in the data. If a network of a reasonable size is used as a starting point, it can learn the relationships on its own. Often, the challenge is to stop the network from learning the data too well as this could lead to a problem known as *overfitting*. If this happens, then the model would fit the data on which it is trained extremely well, but would fit new (or test) data poorly.

While complex relationships can be modeled with neural networks, obtaining coefficients or importance weights from them is not straightforward. For this reason, neural networks are much more useful for prediction rather than explanation.

There are many types of neural networks, but the most commonly used distinction is between supervised and unsupervised networks. We will look at supervised networks here and at unsupervised networks in the next article. Supervised neural networks are similar to regression/classification type models in that they have dependent and independent variables.

Back propagating networks are probably the most common supervised learning networks. Typically they contain an input layer, output layer and hidden layer. The input and output layers correspond to the independent and dependent variables in traditional analysis. The hidden layer allows us to model non-linearities. In a back propagating network the input observations are multiplied by random weights and compared to the output. The error or difference in the output is sent back over the network to adjust the weights appropriately. Repeating this process continuously leads to an optimal solution. A holdout (or test) dataset is used to see how well the network can predict observations it has not seen before.

RECENT ADVANCES

Several recent advances have been made in key driver methodology. The first of these relates to regression analysis and is called Hierarchical

Bayes Regression. Consider an example where consumers provide attribute and overall ratings for different companies in the marketplace. Different consumers may rate different companies based on their familiarity with the companies. An overall market-level model can be obtained by combining all of the ratings and running a single regression model across everybody. But if we one could run a separate model for each consumer and then combine all of that information, the resulting coefficients would be much more accurate than what we get from a regular regression analysis. This is what Hierarchical Bayes Regression does and is hence able to produce more accurate information. Of course, this type of analysis can be used only in situations where respondents provide multiple responses.

For classification problems, there have been a series of recent advances such as *stacking, bagging* and *boosting*. In stacking, a variety of different analytical techniques are used to obtain classification information and then the final results are based on the most frequent classification of data points into groups in each of those methods. Bagging is a procedure where the same technique is used on many samples drawn from the same data and the final classifications are made based on the frequencies observed in each sample. Finally, boosting is a method of giving higher weights to observations that are mis-classified and repeating the analysis several times. The final classifications are based on a weighted combination of the results from the various iterations.

CONCLUSION

This survey has touched upon both traditional methods and recent developments in key driver methodology that may be of interest to marketing research professionals. The particular method to be used often hinges on the primary objective - explanation or prediction. Once this determination is made, there are a variety of tools that can be used that include linear and non-linear methods, as well as those that employ multiple dependent variables.