Massively Categorical Variables: Revealing the Information in Zip Codes

Thomas J. Steenburgh • Andrew Ainslie • Peder Hans Engebretson
Yale University, New Haven, Connecticut 06520
University of California, Los Angeles, Los Angeles, California 90095
ClearInfo, Denver, Colorado
thomas.steenburgh@yale.edu • andrew.ainslie@anderson.ucla.edu • phe3@earthlink.com

We introduce the idea of a *massively categorical variable*, a variable such as zip code that takes on too many values to treat in the standard manner. We show how to use a massively categorical variable directly as an explanatory variable.

As an application of this concept, we explore several of the issues that analysts confront when trying to develop a direct marketing campaign. We begin by pointing out that the data contained in many of the common sources are masked through aggregation in order to protect consumer privacy. This creates some difficulty when trying to construct models of individual level behavior.

We show how to take full advantage of such data through a hierarchical Bayesian variance components (HBVC) model. The flexibility of our approach allows us to combine several sources of information, some of which may not be aggregated, in a coherent manner. We show that the conventional modeling practice understates the uncertainty with regard to its parameter values.

We explore an array of financial considerations, including ones in which the marginal benefit is non-linear, to make robust model comparisons. To implement the decision rules that determine the optimal number of prospects to contact, we develop an algorithm based on the Monte Carlo Markov chain output from parameter estimation. We conclude the analysis by demonstrating how to determine an organization's willingness to pay for additional data.

(*Direct Marketing; Categorical Variables; Hierarchical Bayes Analysis; Variance Components; Decision Theory.*)

1. Introduction

The main prediction problem faced by direct marketing organizations is to determine as accurately as possible the likelihood that each individual prospect will accept a given offer. This is generally accomplished either by conducting a trial campaign on a subset of the prospects or by examining how a previous group of people responded to an offer in order to find a new set of prospects that is likely to respond positively. Our study is an example of the latter method, which is sometimes referred to as *clone marketing*. Based on the responses of a previous cohort, we show how to identify the set of prospects from the current group that should be targeted.

Direct marketers have long been aware that geographic location can be a key variable in predicting consumer behavior. Geographic location can act as a proxy for demographic variables on which only limited information has been obtained. For example, people living in a poor, rural area of Arkansas exhibit very different buying patterns than those living in an expensive area of New York City. This is due to differences in the consumers' income, education, and family size; their distance from competing retailers; their inventory-carrying capacity; and other characteristics by which residents in these areas systematically differ. The addresses themselves convey useful information about the consumers.

Marketing Science © 2003 INFORMS
Vol. 22, No. 1, Winter 2003 pp. 40–57

0732-2399/03/2201/0040$05.00
1526-548X electronic ISSN