

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.

Consequently, it is not surprising that two popular sources of data used by direct marketers are reported on the basis of geographic location: the credit history and demographic data collected by consolidators such as Experian and Claritas, and the census data collected by the U.S. government. In the former case, the Federal Trade Commission (FTC) requires these consolidators to protect the privacy of individual consumers by aggregating the individual-level data of all residents of a particular zip code or “zip +4” (the 5-digit or 9-digit code given to each address in the United States). This practice masks sensitive consumer information, and hereafter we refer to these as “masked data.” Census data are similarly masked, because they are made available in the form of aggregates for each census tract. In both cases, the masked data continue to be useful even after aggregation for the reasons given in the preceding paragraph.

Direct marketers generally ignore the masked nature of these data when developing their targeted marketing campaigns. They match the prospect’s zip code against the demographic variables available in the masked dataset and then treat the masked variables as if these are actually individual-level data. Using a hierarchical Bayes variance components (HBVC) model, our approach improves on this in two ways: (1) We directly incorporate the zip code, thereby allowing us to account for unobserved heterogeneity across zip codes; and (2) we use a hierarchical model to appropriately use the information at the level at which it is collected, i.e., at the level of the zip code rather than of the individual. We also show that the conventional modeling practice, in ignoring the data masking, understates the uncertainty with regard to its parameter values.

Our methodology extends beyond zip codes, which are just one example of a larger class of variables. We introduce the idea of a *massively categorical variable*, a variable such as zip code that takes on a very large number of categorical values, and demonstrate how the HBVC model uses massively categorical variables directly as explanatory variables. This type of variable is in no way limited to being used in choice models, and the approach taken here easily extends to other applications. Other examples of massively categorical variables include using actor, director, or distributor

as independent variables when predicting the performance of a movie; and using SEC codes as predictors when modeling brand equity as a component of market capitalization.

Our research tackles one further, important problem. In the previous literature, the issue of a broadly applicable framework for determining the financial gain of using superior methodologies or datasets has not been addressed. This makes determining the value of the gain from using different models difficult. We develop a decision theoretic framework to measure the competing models on a monetary basis in addition to a statistical one, as we want to judge whether a significant practical difference exists between the modeling techniques. Our approach can be used to explore an array of financial considerations (including ones in which the benefit is nonlinear in the number of positive responses) to make the comparisons as robust as possible. In the nonlinear case, we develop an algorithm based on the Monte Carlo Markov chain output to implement the decision rules that determine the optimal number of prospects to contact. We conclude the analysis by demonstrating how to determine an organization’s willingness to pay for additional data.

2. Previous Literature

Using geography as a basis for understanding marketing issues has recently received attention from several authors. Two examples include Bronnenberg and Sismeiro (2002) and Yang and Allenby (2002). The methodology used in these papers makes use of the behavior in surrounding cells to assist in predicting the behavior in a cell of interest. The former does this strictly on the basis of geography, whereas the latter uses several different social and geographic bases. These methods are similar to autoregressive models, but use geography rather than time as a basis for forming correlations between cells.

Our approach is different in that we do not use an autoregressive type of scheme. Instead, it is closer in spirit to the method of Hoch et al. (1995), who used a hierarchical model to demonstrate the importance of using geographic location and demographic data in determining price elasticities. The advantages

of our approach to the problem are: (1) It is broadly applicable to a far wider class of categorical regressors than geographic ones; (2) in situations where zip codes are isolated because all or many of the surrounding zip codes are not represented in the data, the spatial models are less effective; and (3) We are able to include hierarchical regressors, whereas it is unclear how one would include these directly in spatial models. Hierarchical variables have two important uses: First, they improve the predictive power of the model; second, the parameter values allow marketers to better understand their target market, assisting with segmentation decisions. We leave it to future research to determine the relative effectiveness of the two approaches or to incorporate them into a single model.

Very little academic research has been devoted to the target selection problem in direct marketing. In a notable exception, Bult and Wansbeek (1995) develop a profit maximization approach to select prospects for a mailing campaign. They demonstrate how to determine appropriate cutoff values assuming a constant marginal benefit from positive responses. They also make assumptions about the distribution of independent variables and various R^2 values in developing these rules. Wilcox and Hsu (2000) demonstrate the importance of accounting for uncertainty in parameter values when predicting outcomes in Logit models. Our work extends that done in both of these papers by making less restrictive assumptions than Bult and Wansbeek (1995), allowing for a broad range of profit functions, and accounting fully for the distribution of the parameters.

3. The Prediction Model

In our application, we address the task of determining the likelihood of enrollment of prospective students at a large university in Texas so that the institution can better target its recruiting efforts. Information is gathered on the previous year's students in order to predict the likelihood of enrollment of the current prospects.¹ The data available for this analysis include

¹ The university data used in this paper are representative of the general class of clone models used in direct marketing. For an in-depth analysis of students' choice of college, the reader is referred to Manski and Wise (1983).

the prospect's zip code (and, separately, demographic information for each zip code), the intended major of the student, and information collected by the university on the frequency and nature of contacts between the university and the prospect. This is discussed in detail in §5.

Organizations that want to conduct targeted marketing campaigns currently have numerous sources of information available to facilitate their predictions. These sources can be broadly split into two types. The first type of data is collected by the firm on its interactions with individual patrons. Marketing groups take great care to coordinate the data collection efforts among departments and make substantial expenditures to retain accurate records. In general, these data are often powerful because they are specific to individual patrons and are unique to that firm, allowing them to differentiate themselves from competitors. Unfortunately, many times they are not available. For example, historical data do not exist when a firm wants to extend its reach to new prospects. In our application, the university collects information on the frequency and type of interactions between students and the university, which we call "visitation data."

The second type of data is purchased from an outside firm. Meeting the need for comprehensive information, data consolidators and others make supplementary databases available for a fee. Although these sources tend to improve predictions about how people will respond to an offer, they present methodological challenges as well. The variables contained in these databases are commonly masked through aggregation to protect the privacy of individual consumers, making statistical models of individual behavior more difficult to construct. (These are the previously described masked data.) We propose that the conventional modeling technique not only fails to extract all of the information contained in supplementary databases, but also that some of the information can be revealed without purchasing any data at all, establishing a higher baseline from which to judge the value of acquiring additional data.

The masked data are constructed to be as useful as possible while still protecting the individuals' privacy. An implicit assumption for this to be true is that a reasonably high level of homogeneity exists among

the individuals within the unit of aggregation. Zip codes, for example, make a good basis of aggregation because, as commonplace observation suggests, people tend to live near others similar to themselves. Taking this idea further, we propose that the prospects' zip codes themselves may be useful because they convey a meaningful group membership. Similarities not observed in the supplementary data exist among the prospects within a zip code, and we should take advantage of these associations whether we have the supplementary data or not.

A final modeling challenge arises when analysts need to combine information from multiple sources. The difficulty occurs either when some of the information is masked while some is not, or when multiple sources are masked but at different levels of aggregation. Given the previous discussion, for example, an analyst may want to augment the organization's own data on individual prospects with a masked source purchased from a consolidator.² The standard technique used to construct a complete database is simply to append the consolidator's data to the organization's own source on the basis of each individual's zip code. The entire record is then treated as individual-level data in the subsequent analysis, ignoring that some of the data are masked.

The conventional modeling technique can be expressed in the following manner. Assume that the relationship between the variable of interest y_i and the explanatory variables for each individual is given by

$$y_i = \begin{bmatrix} X_i \\ W_{z_i} \end{bmatrix}^T \begin{bmatrix} \hat{\alpha} \\ \hat{\gamma} \end{bmatrix} + \varepsilon_i, \varepsilon_i \sim N(0, 1)$$

for $i = 1, \dots, N$, (1)

where, in our case, y_i represents a student's latent utility, and a student enrolls in the university if his or her latent utility exceeds zero.³ The index i corresponds to

² The related problem of two masked sources arises, for example, when an analyst wants to combine a consolidator's data source that is masked through zip code-level aggregation with census bureau data that are masked through census-block aggregation.

³ Although our application involves a binomial choice problem modeled using Probit, the problem may occur in any regression structure and can be solved in the manner outlined without loss of generality.

an individual, and the index z_i denotes the zip code of the i th individual. The vector X_i^T represents the explanatory variables for individual i that are found in the organization's own database, and these are unique to each individual. The vector W_{z_i} represents the masked variables found in the consolidator's data. Every individual residing in a zip code shares the same record of information, and the masked data are repeatedly used in the regression when multiple people reside in a zip code. The vectors $\hat{\alpha}$ and $\hat{\gamma}$ are the parameters of interest in estimation.

The conventional technique simply concatenates the individually measured variables X with the masked variables W and thereafter treats all variables as if they were measured appropriately. We refer to this as the *null model* and note that the Bayesian estimation of it, using sufficiently diffuse prior distributions, yields parameter estimates of $\hat{\alpha}$ and $\hat{\gamma}$ that are equivalent to those found through maximum likelihood estimation. For a description of how to estimate a Bayesian Probit model, see Albert and Chib (1993); for a general introduction to the MCMC methods used in this paper, see Gelman et al. (1995).

The null model can be criticized in a few respects. First, we do not truly have N unique, individual-specific observations of the explanatory variables W , but rather only Z zip code-specific observations ($Z < N$).⁴ Subsequently, greater uncertainty exists in the posterior distributions of the parameter estimates than is found through the null model. Next, the null model assumes that the observed component $W_{z_i} \hat{\gamma}$ explains all of the variation in student responses that can possibly be explained by the masked data, overlooking the possibility of an unobserved component of variation. As systematic differences exist among people across zip codes, the zip codes themselves may reveal something useful about the prospects not captured in the data. We should take advantage of this indicator of group membership if it is meaningful, but we would expect the null model to compete reasonably well with the alternative if it is not.

⁴ We do observe N individual-specific explanatory variables (X_i^T) and responses (y_i) in the university's data.

In contrast to the standard practice, we develop a hierarchical Bayesian variance components model⁵ to solve the problems created by using masked variables. We are the first to describe the problem of masked data, and our analysis extends the use of established hierarchical Bayesian modeling to this new area of research. The variance components approach allows us to treat each source of data, whether it is masked or not, at the appropriate level of aggregation. Furthermore, the variance decomposition helps us to understand the value of each source of information. The adaptive shrinkage inherent in our Bayesian specification ensures that we get useful parameter estimates even when a zip code contains a few individuals.

The HBVC approach to the estimation problem is given by

$$\begin{aligned} y_i &= X_i^T \alpha + \beta_{z_i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1) \quad \text{for } i = 1, \dots, N \\ \beta_z &= W_z^T \gamma + \nu_z, \quad \nu_z \sim N(0, V_{\beta_z}) \\ &\text{for } z = 1, \dots, Z. \end{aligned} \quad (2)$$

The indices i and z_i and the vector of data X_i^T represent the same constructs that they do in the null model. We need some new notation to reflect our treatment of the masked variables and introduce the subscript z to correspond with the zip codes. The subscripts z and z_i can be thought of as labels that both refer to zip codes, with z being used to refer to the zip code in general and z_i being used to refer to the zip code in which the i th prospect resides. The vector W_z^T represents the masked demographic variables that are taken from the consolidator's data for zip code z , and the scalar parameter β_z represents the estimated zip code effect. The observed variation of β_z across zip codes is described by the term $W_z^T \gamma$, where the parameter vector γ is found through the hierarchical regression of zip code effects onto the

demographic data. The masked variables are treated as only Z unique observations, which differ from their treatment as N observations in the null model. The unobserved variation is captured through the parameter ν_z .

It is straightforward to compare the distributions of the effects associated with the masked variables in the two models. The conditional variance of $\hat{\gamma}$ in the null model is $\sum_{i=1}^N (V_{\gamma_0}^{-1} + W_{z_i}^T W_{z_i})^{-1}$, whereas the conditional variance of γ in the HBVC model is $\sum_{z=1}^Z (V_{\gamma_0}^{-1} + W_z^T W_z)^{-1}$. The $N - Z$ repeated records ensure that $W_{z_i}^T W_{z_i} > W_z^T W_z$ and that the distribution of the null model is tighter than that of the HBVC model. Since we might find the marginally significant parameters in the null model to be in fact statistically insignificant if the repeated records are not treated as unique observations, the null specification can lead to misguided inference.

We now turn our attention to getting as much as we can out of the firm's data when no additional information has been bought. The underlying objective is to establish an adequate baseline from which we can assess the value of purchasing supplementary data. In the preceding exposition, we suggest that knowing where a prospect resides is useful because zip codes convey a meaningful association. This knowledge should be useful even when no further data have been bought. The university always possesses its prospects' zip codes, and the open question is how to take advantage of the prospects' group membership. This membership can be thought of as a categorical variable with a very large number of categories, too many to treat them as standard dummy variables.

We can amend the HBVC model to use the massively categorical variables directly as explanatory variables, allowing us to estimate the zip code effects without purchasing supplementary data. This model should explain less variation than one with the benefit of supplementary data, but the massively categorical variables should replicate some of the information provided in the supplement. The model would be revised as

$$\begin{aligned} y_i &= X_i^T \alpha + \beta_{z_i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1) \quad \text{for } i = 1, \dots, N \\ \beta_z &\sim N(0, V_{\beta_z}) \quad \text{for } z = 1, \dots, Z, \end{aligned} \quad (3)$$

⁵ Ainslie and Rossi (1998) is an HBVC framework in a choice model that is similar in nature but different in application. Whereas they break down the heterogeneity of consumer preferences into components that are common across and unique to supermarket categories, we use the HBVC approach to break the variance of enrollment probabilities into components associated with different sources of data.

when no data are purchased from the consolidator, but we use the massively categorical variable zip code directly as an explanatory variable. The heterogeneous zip code effects β_z are restricted to be distributed about zero for identification. Reducing X_i^T to a vector of ones results in a model that includes only the massively categorical variables as explanatory variables.

In our particular dataset, in addition to knowing the prospects' zip codes, the university also knows their intended major. Adding another component to the model for this massively categorical variable can be accomplished even though zip codes and majors are not generally nested.⁶ A fully comprehensive model uses all of the data and is expressed as

$$y_i = X_i^T \alpha + \beta_{z_i} + \beta_{m_i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1)$$

for $i = 1, \dots, N$

$$\beta_z = W_z^T \gamma + \nu_z, \quad \nu_z \sim N(0, V_{\beta_z}) \quad \text{for } z = 1, \dots, Z$$

$$\beta_m \sim N(0, V_{\beta_m}) \quad \text{for } m = 1, \dots, M. \quad (4)$$

The subscripts m and m_i , respectively, refer to the intended major and the intended major of the i th prospect. The heterogeneous major effects β_m are directly analogous to the zip code effects in the previous model. While intended majors are particular to this application, their inclusion in the model demonstrates how multiple massively categorical, masked, and unmasked variables can be simultaneously incorporated in the same model.

4. The Direct Marketer's Decisions

We now move on to determining how to evaluate different modeling or data choices under different scenarios. In general, the direct marketer wants to contact the set of prospects that maximizes its expected gain, balancing the possible money gained from greater enrollment against the certain money lost from contacting additional prospects. Without

⁶Suppose we are using both the prospects' zip codes and their intended majors as massively categorical variables. Now consider an English major applicant in zip code 13021. Neither are all other residents of zip code 13021 necessarily English majors, nor do all other English majors necessarily reside in zip code 13021.

loss of generality in terms of the applicability of these techniques to the broad range of problems encountered by direct marketers, for the remainder of this section we refer to the university's particular problem of targeting prospective enrollees.

Let the function $h(s)$ for $s \in \{0, 1, 2, \dots\}$, which we refer to as the *benefit function*, represent the benefit to the university when s individuals choose to enroll. Define $\delta(s) \equiv h(s) - h(s-1)$ for $s \in \{1, 2, 3, \dots\}$ as the marginal benefit of the s th enrollee. In order to develop a general argument, we merely assume that the marginal benefit of enrollment is non-increasing: $\delta(s+1) \leq \delta(s)$ for $s \in \{1, 2, 3, \dots\}$. Let c represent the marginal cost of contacting prospects and assume that it is constant. Financial considerations determine the values of both $h(s)$ and c before any statistical analysis is undertaken.

Suppose the university is trying to determine which prospects to contact from a group of M individuals. Let R_i be a Bernoulli random variable that represents the response of the i th prospect if he or she is contacted, where θ_i is the chance of a positive response. The sum $S_m = \sum_{i=1}^m R_i$ represents the number of positive responses when prospects 1 to m are contacted. If the probabilities of enrollment were the same for all individuals, S_m would simply follow a binomial distribution.

Given the heterogeneous probabilities of enrollment, we have to use a recursive formula to obtain the distribution of S_m . We find the probability mass function of S_m , when contacting $m \in \{1, \dots, M\}$ prospects, through the recursive step

$$p_{S_m}(s) = \theta_m p_{S_{m-1}}(s-1) + (1-\theta_m) p_{S_{m-1}}(s)$$

for $s \in \{0, \dots, m\}$, (5)

and the end condition

$$p_{S_0}(s) = \begin{cases} 1 & \text{for } s = 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $p_{S_m}(s) \equiv \Pr\{S_m = s\}$. The terms of Equation (5) result from the two possible ways that s positive responses can arise from contacting m prospects: Either (1) $s-1$ positive responses are produced from the first $m-1$ contacts and the m th prospect responds positively, or (2) s positive responses are produced from the first $m-1$ contacts and the m th prospect does not respond positively.

We begin by looking at the decision problem as if the prospects' probability of enrollment were known with certainty. The expected benefit from contacting the first m prospects, conditional on $\theta = \{\theta_1, \dots, \theta_M\}$, is

$$\begin{aligned} E[h(S_m) | \theta] &= \sum_{s=0}^m h(s) p_{S_m}(s) \\ &= E[\theta_m h(S_{m-1} + 1) + (1 - \theta_m) h(S_{m-1}) | \theta] \\ &\quad \text{for } m \in \{0, \dots, M\}. \end{aligned} \quad (6)$$

Given this relationship, the expected marginal benefit from contacting to the m th prospect is

$$\begin{aligned} \Delta(m) &\equiv E[h(S_m) - h(S_{m-1}) | \theta] \\ &= \sum_{s=0}^{m-1} \theta_m \delta(s+1) p_{S_{m-1}}(s) \quad \text{for } m \in \{1, \dots, M\}. \end{aligned} \quad (7)$$

(Derivation one in the appendix.) Expressed in words, the expected marginal benefit of the m th enrollee is the expected benefit of an additional enrollee given that the first $m-1$ contacts resulted in s enrollments, $\theta_m \delta(s+1)$, weighted by the chance that the first $m-1$ contacts resulted in s enrollments.

Using the expected marginal benefit, we can develop a decision rule that sequentially sorts through the prospects to maximize the university's expected profit. The university should contact the m th prospect if the expected marginal profit from doing so is positive, specifically if $\Delta(m) - c > 0$. Expressing this inequality in terms of the prospects' probabilities of enrollment, the university should contact the m th prospect if

$$\theta_m - \Psi(m) \geq 0,$$

where

$$\begin{aligned} \Psi(m) &\equiv \frac{c}{\sum_{s=0}^{m-1} \delta(s+1) p_{S_{m-1}}(s)} \\ &\quad \text{for prospects } m \in \{1, \dots, M\}. \end{aligned} \quad (8)$$

Following Bult and Wansbeek (1995), we refer to $\Psi(m)$ as the *cutoff function* because it contains the cutoff values against which the prospects are assessed.

The cutoff function does not depend on the probability of enrollment of the prospect under consideration, but it does generally depend on the probabilities of enrollment of the $m-1$ prospects previously considered. (As will be shown in the examples, the previously considered prospects do not play a role when the marginal benefit of enrollment is constant.) Since $\Psi(m)$ does not depend on θ_m , ordering the prospects from the highest probability of enrollment to the lowest ensures that θ_m exceeds $\Psi(m)$ by the greatest possible amount for every m . The ordering also ensures that if the m th prospect exceeds the cutoff value, all of the previously considered prospects will exceed it too and we not need reconsider them. Furthermore, since $\Psi(m)$ is a weakly increasing function of the number of prospects contacted (proof in the appendix), once the m th prospect falls below its cutoff value, every subsequent prospect will fall below theirs too and we can stop considering additional prospects at this point. We now turn to developing some specific cutoff functions.

Example One: A Constant Marginal Benefit from Enrollment

We begin by considering the standard benefit function applied in direct marketing problems, a form that corresponds with the one used in Bult and Wansbeek's (1995) analysis. Suppose the marginal benefit of enrollment is constant regardless of the number of enrollees. Specifically, $h(s) = bs$ and $\delta(s) = b$ for $s \in \{0, 1, 2, \dots\}$, where the constant b represents the marginal benefit of an additional enrollee and the variable s represents the number of people who actually enroll. Since $\delta(s) = b$ for $s \in \{1, 2, 3, \dots\}$, we can directly see that the cutoff function is $\Psi(m) = c/b$ for $m \in \{1, 2, \dots, M\}$. This results in the familiar decision rule of contacting the m th prospect if his or her probability of enrollment is greater than the ratio of the marginal cost to the marginal benefit.

Example Two: A Capacity Constraint

Next we examine a nonlinear benefit function. Suppose a physical constraint, perhaps the amount of classroom space, exists such that the benefit of additional enrollees is less to the university once the

number of enrollees surpasses a given level. Assume the marginal benefit of an enrollee is

$$\delta(s) = \begin{cases} b & \text{for } s < q \\ b - a & \text{for } s \geq q, \end{cases}$$

where $b > a > 0$. In this case, the cutoff function is $\Psi(m) = c/(b - a\Pr\{S_{m-1} > q\})$ for $m \in \{1, 2, \dots, M\}$ (derivation two in the appendix), which depends on probabilities of enrollment of the $m - 1$ previously considered prospects through the term $\Pr\{S_{m-1} > q\}$.

Example Three: A Diminishing Marginal Benefit from Enrollment

Suppose the university faces a universally diminishing marginal benefit. Assume the benefit function is $h(0) = 0$ and $h(s) = b \sum_{j=1}^s (1 - d)^{j-1}$ for $s \in \{1, 2, 3, \dots\}$, where $b > 0$ and $0 < d < 1$. The marginal benefit of the first enrollee is b and the rate at which the marginal benefit diminishes over additional enrollees is d . This problem simplifies to the constant marginal benefit case when d equals zero. An example of such a cutoff function might be a case in which willingness to pay or contribute is proportional to the probability of responding. In a catalog marketing application, it is often the case that high probability customers tend to purchase more. The benefit function implies that $\delta(s + 1) = (1 - d)\delta(s)$ for $s \in \{1, 2, 3, \dots\}$. A recursive relationship exists in the cutoff function such that $\Psi(m + 1) = 1/(1 - d\theta_m)\Psi(m)$ for $m \in \{1, \dots, M\}$ (derivation three in appendix) and $\Psi(1) = c/b$. As a result, the cutoff function for the m th prospect is

$$\Psi(m) = \begin{cases} c/b & \text{for } m = 1 \\ c/(b \prod_{i=1}^{m-1} (1 - \theta_i d)) & \text{for } m \in \{2, \dots, M\}. \end{cases}$$

Our knowledge regarding the probabilities of enrollment is not perfect. Following standard decision theoretic arguments,⁷ we account for the parameter uncertainty by examining the posterior expected profit of contacting a prospect. The posterior expected difference between the m th prospect's probability of enrollment and the cutoff value is $E_{\pi(\theta|data)}[\theta_m - \Psi(m)] = \tilde{\theta}_m - \tilde{\Psi}(m)$, where $\tilde{\theta}_m$ and $\tilde{\Psi}(m)$

are the posterior means of θ_m and $\Psi(m)$, respectively. Because $\Psi(m)$ does not depend on θ_m , ordering the prospects on the basis of their posterior means ensures that $\tilde{\theta}_m$ exceeds $\tilde{\Psi}(m)$ by the greatest possible amount for every m . The ordering also ensures that if the m th prospect makes the cutoff, all the previously considered prospects will make it, too, and we do not need reconsider them. Furthermore, since $\tilde{\Psi}(m)$ is a weakly increasing function of the number of prospects contacted, once the m th prospect falls below its cutoff value, every subsequent prospect will fall below theirs too and we can stop considering additional prospects at this point.

We return to the examples to clarify how the decision rules are implemented. The decision rule in the first example is straightforward to implement because it is a comparison of the posterior mean $\tilde{\theta}_m$ against a fixed and known cutoff value. The decision rules in the second and third examples are more difficult to implement because $\Psi(m)$ is a nonlinear function of θ . Finding $\tilde{\Psi}(m)$ requires taking the expectation over the joint posterior distribution of $\{\theta_1, \dots, \theta_{m-1}\}$. We use the following algorithm based on the MCMC draws to accomplish:

1. After obtaining T draws of the vector $\theta^{(t)}$ from the sampler, compute the posterior mean $\tilde{\theta}$ and reorder the prospects from the highest probability of enrollment to the lowest. Let m represent the position of the prospects in this reordered list.
2. Using the replicates $\theta_1^{(t)}, \dots, \theta_{m-1}^{(t)}$ of the now reordered prospects, compute the M cutoff values $\Psi(m)^{(t)}$ for each of the T draws.
3. Compute the M posterior means of $\tilde{\Psi}(m)$ and implement the decision rule to find the optimal number of prospects to be contacted.

For the second example, step 2 of the algorithm is executed using the recursive relationship of the probabilities described in Equation (5). For the third example, step 2 of the algorithm is executed using the recursive relationship of $\Psi(m)$.

The university's willingness to pay for the supplementary data (WTP) is the difference between the posterior expected gain from the decisions made in light of all available data, including the consolidator's supplement, and the posterior expected gain from the decisions made in light of only the university's own

⁷ See Berger (1985, §4.4) for a more thorough discussion.

source. The supplementary data improve our knowledge by allowing us to observe a component of variation ($W_z^T \gamma$) in the zip code effects, thereby reducing our uncertainty about them. The data's worth is formally expressed as

$$WTP = E_{\pi(\theta_i|X,W)} [E[h(S_{m^*}) - m^*c | \theta]] - E_{\pi(\theta_i|X)} [E[h(S_{m^{**}}) - m^{**}c | \theta]], \quad (9)$$

where m^* and m^{**} denote the optimal decisions made by the university, respectively, in light of the augmented and unaugmented databases. As the equation suggests, a valuable supplementary database necessarily results in helping the university make better decisions.

A few additional points should be made in concluding this section. First, we have assumed that the direct marketer knows the parameter values associated with the decision problem. For example, the marginal cost is assumed to be the constant value c in these problems. This usually is a good approximation; for example, the cost of producing and mailing a brochure, within reasonable bounds, remains constant and is well understood by the marketer. If the parameters are not perfectly known, a sensitivity analysis is relatively easy to perform over a range of values to aid in the decision making. Second, we have assumed specific functional forms for the marginal benefit function $\delta(s)$ in the examples. If an alternative functional form is preferred, the direct marketer either can find an analytical solution of the cutoff function or can numerically calculate it using Equation (8).

5. The Empirical Results

We base our study on data that were collected by a private, southern U.S. university that plans to assess its prospects at the inquiry stage of the admissions process. The students have requested information about the university at this point, but they have not yet decided whether they will apply for admission. We split the dataset randomly into halves, using 37,551 prospects for model building and 34,179 prospects for out-of-sample model comparisons. These students reside in 7,279 zip codes, and each has declared an interest in one of 128 majors.

Both the zip codes and the intended majors are used as massively categorical variables, and the students' responses of "undecided" and "no response" for the intended major are treated separately.

Table 1 describes the explanatory variables that we use in the various prediction models. The university has the opportunity to purchase a supplementary data source that contains over 200 variables that are masked by zip code. From the many variables that are available in this source, we select the four that produce the best prediction results in the null model. While the issue of variable selection is outside of the scope of this paper, we note that adding more variables than these generally diminishes the out-of-sample performance. The university also collects some information about the individual prospects on its own, and we examine its decision problem under two scenarios—one in which it collects campus visitation data and one in which it does not—to more generally reflect the conditions that confront analysts. This approach to the problem results in four

Table 1 Description of the Data

The Massively Categorical Variables	
Z	The zip codes in which the prospective students reside
M	The intended majors of the students
The University's Data (Available for Each Individual—Referred to as X)	
<i>Available in All Scenarios</i>	
RES	A dichotomous indicator of whether the student is an in-state resident (1 = Yes)
<i>Available in the Campus Visitation Data Scenarios</i>	
VISIT	An indicator of whether the student visited campus (1 = Yes)
OPENHOUSE	An indicator of whether the student attended a campus open house (1 = Yes)
CONTACTS	The number of times that the student contacted the university
The Consolidator's Data (Available for each Zip code—Referred to as W)	
COL_ED	The proportion of college educated households in the zip code
FEM_MARRY	The proportion of married females in the zip code
BLT_50	The proportion of structures built in the 1950s in the zip code
BLT_80	The proportion of structures built from 1980 to 1985 in the zip code

information sets based on (1) whether the university is able to collect campus visitation data, and (2) whether the university purchases supplementary data from a consolidator. We compare the HBVC against the null model in each of the four possible cases to demonstrate the incremental benefit of using the HBVC model.

All models were run initially for 10,000 iterations as a run-in period, then a further 10,000 iterations for obtaining posterior distributions. We tested convergence in several ways, including comparing results between earlier and later portions of each run and visual checks, and are confident that we were very conservative in the lengths of burn-ins and runs selected.

5.1. The Statistical Comparison of the HBVC Model Against the Null

We begin the empirical analysis by applying the standard Bayesian hypothesis testing procedure to the problem of model selection. This procedure does not require nested models, produces results that are easy to interpret, and is coherent with foundational principles. A Bayes factor, which globally compares the models' performance, summarizes the evidence provided by the data for the alternative model against the null. We use a numerical method proposed by Newton and Raftery (1994) to evaluate the Bayes factor, and the computations are based on the MCMC output produced through model estimation.

We report both the Bayes factors and log-marginal likelihoods in Table 2. The Bayes factors are expressed in terms of $2 \log B$ to put them on a scale that is common with familiar measures such as the deviance or the likelihood ratio test statistics. A value of zero

suggests equal evidence for both models, and a value greater than ten suggests "very strong evidence" in favor of the HBVC model. Raftery presents a full calibration of values between zero and ten in Gilks et al. (1996, p. 165). We find that the data provide very strong evidence for the HBVC model in every case when the same amount of information is used to estimate both models. The smallest value, 586, well exceeds a standard of 10. Moreover, the data provide very strong evidence that the HBVC models without supplementary information are superior to the corresponding null models with it. This suggests that using a superior modeling technique can be more important than purchasing more information. Finally, we see that having supplementary information makes a statistically significant difference in performance, as the HBVC models with the benefit of the data are superior to those without it.

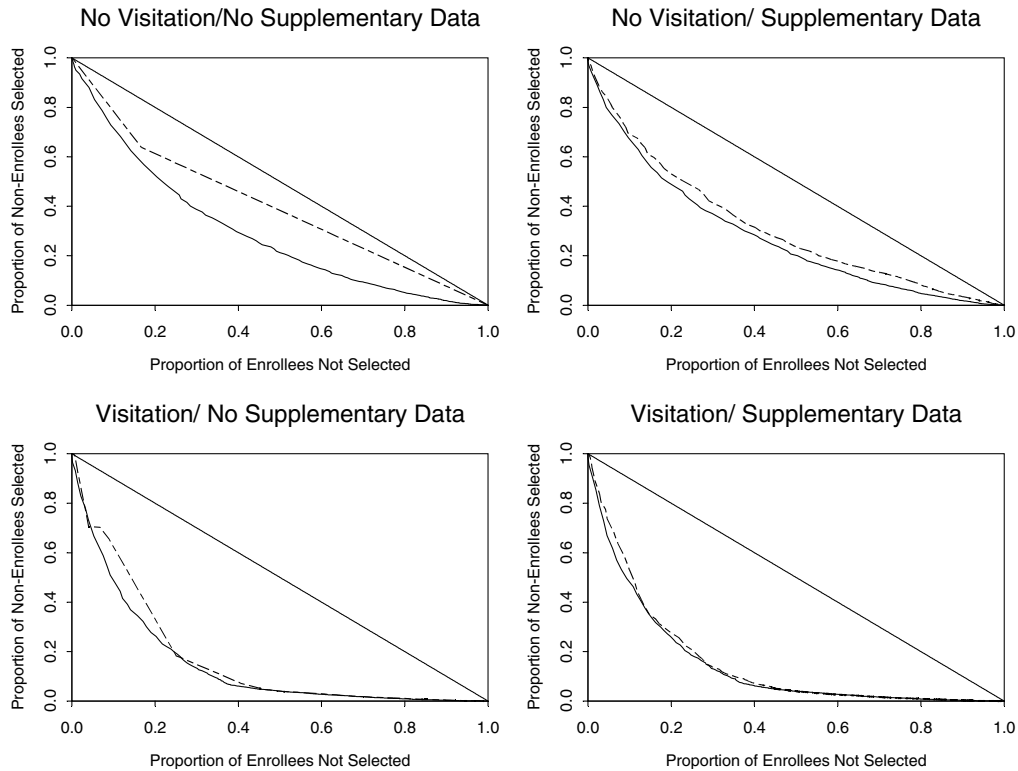
This first statistic was calculated in-sample. For the remainder of this section, all statistics are calculated out-of-sample, as this ensures that improved performance is not simply a result of over-fitting. In addition to assessing the overall out-of-sample fit of the model, we also are interested judging how well it orders the individual prospects. Being able to attain the best possible ordering from the information at hand is crucial in direct marketing because the prospects lowest on the list will be dropped from consideration. For example, even though the data suggest it doesn't fit as well, we still might prefer the null model if it provides a better ordering but systematically overstates the likelihood of enrollment. With this objective in mind, we use the holdout data to construct *receiver operating curves*, or ROC charts, to compare the various models. An ROC chart is constructed by repeatedly dividing the prospects into two groups on the basis of their probabilities of enrollment.⁸ Prospects with probabilities below the cutoff are placed in one group, and prospects with probabilities above the cutoff are placed in the other. After this has been done for all cutoff points between zero and one, we then graph the number of enrollees not selected against the number of non-enrollees selected

⁸ We use the posterior means $\tilde{\theta}$ to divide the prospects because of their role in the decision rule.

Table 2 Model Comparisons by Bayes Factors

Information Available		Log-Marginal Likelihood		Bayes Factor
		Null	HBVC	HBVC vs. Null
University's Source	Supplementary Source			
Without visitation data	Does not purchase	-5429	-4921	1016
	Does purchase	-5271	-4839	864
With visitation data	Does not purchase	-4253	-3960	586
	Does purchase	-4196	-3889	614

Figure 1 ROC Charts Comparing the HBVC Versus Null Models



for each division. (We note that the cutoff points used in the ROC chart are not determined by the cut-off functions previously described. They are merely a series of fixed points used to divide the prospects.)

Not only is the ROC chart a robust comparison because it is constructed over all possible cutoff points, but it also is very easy to interpret. The better the model, the more the curve will move toward the bottom-left-hand corner of the chart; or, put another way, the more the area under the curve will tend toward zero. For more information on ROC charts and their interpretation, see Swets (1995). In Figure 1, we construct ROC charts to compare the HBVC model (solid line) against the null model (dotted line) under the four information conditions. In Table 3, we report the area under each curve, which can be thought of as a summary measure describing the model's ability to order prospects.

Summarizing the results, we find that the HBVC model always provides a better ordering of the

prospects when the same amount of information is used to estimate both models. This claim is based both on a visual inspection of the ROC charts and on a direct comparison of the summary measures. For example, in the case where neither the visitation nor supplementary data are available, the area under the curve for the HBVC model is only 0.292, whereas for the null model it is 0.402. We also find a greater difference exists between the models when the amount of information used to estimate them is lower. This

Table 3 Model Comparisons by ROC Area Summaries

Information Available		ROC Area	
University's Source	Supplementary Source	Null	HBVC
Without visitation data	Does not purchase	0.402	0.292
	Does purchase	0.310	0.275
With visitation data	Does not purchase	0.169	0.148
	Does purchase	0.155	0.144

Table 4 Posterior Distributions of the Full Models

Source of Information	Variable	Null Model			HBVC Model		
		Mean	SD	HPD Length	Mean	SD	HPD Length
Common to all models	INTERCEPT	-2.104	0.018	0.064	-2.613	0.067	0.264
	RES	0.184	0.018	0.070	0.154	0.027	0.107
Campus visitation data	CONTACTS	0.162	0.011	0.040	0.193	0.013	0.051
	OPENHOUSE	0.165	0.008	0.032	0.175	0.009	0.036
	CAMPUS VISITS	0.238	0.007	0.029	0.252	0.008	0.033
Supplementary data	COL_ED	0.151	0.015	0.055	0.151	0.032	0.124
	FEM_MARRY	0.052	0.016	0.062	0.090	0.038	0.149
	BLT_50	0.058	0.021	0.075	0.019	0.034	0.133
	BLT_80	0.013	0.019	0.072	0.151	0.032	0.124
Random effects	Z				0.416		
	M				0.378		

implies using a superior modeling technique becomes more important when the university has a limited amount of information with which to work. In several panels, the difference between lines appears small. However, even small differences in the ROC charts can lead to large differences in profitability for a direct marketer, as will be demonstrated in the next section. Finally, based upon the ROC summary measures in Table 3, we claim that the HBVC models without supplementary information provide a better overall ordering of the prospects than do the null models with it. In all respects, the ROC charts confirm what we found in the model selection hypothesis testing.

The null model leads the researcher to be overly confident of where the parameter values lie. The null model repeatedly uses the masked data as if they represent the individuals' actual characteristics, and it understates the posterior variance as a result. To demonstrate this, we summarize the posterior distributions of both models, which are estimated using all of the available information, in Table 4. Three statistics are presented for all variables: the mean of the posterior distribution, its standard deviation, and the size of the 95% highest posterior distribution (HPD) width. The last of these statistics is a good measure of parameter dispersion for non-symmetric distributions.

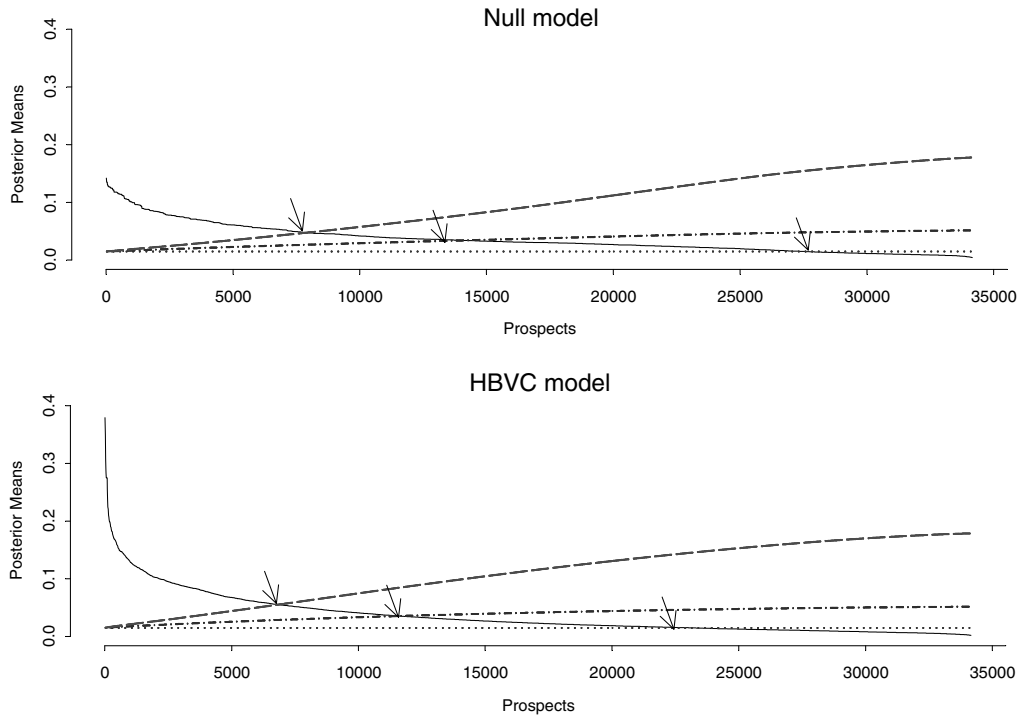
The posterior means are similar across the null and HBVC models; the biggest exceptions to this are the

housing variables BLT_50 and BLT_80. These variables have a high negative correlation (-0.67 in the null model, -0.55 in the HBVC) that may account for the disparity in the parameter estimates. The posterior standard deviations and HPD widths for the visitation variables, which are not masked, also are very similar. This is to be expected because the models are essentially identical for these variables. The posterior standard deviations and HPD widths of masked variables, on the other hand, are much tighter in the null model than they are in the HBVC model. The excessive confidence found in the null model can lead to misguided inference, because statistically insignificant parameters may erroneously be found significant.

5.2. The Empirical Decision Analysis

Having shown that the HBVC model is statistically superior to the null, we now examine how choosing a model affects the university's decisions, and we confirm that a significant practical difference in the generated profit exists between models. Once more, all these calculations are based on the model's ability to predict out-of-sample. We begin the discussion by illustrating how the university identifies which prospects to contact. In Figure 2, we graph the posterior means of the probabilities of enrollment and the cutoff values against the ordered prospects. To construct this diagram, we have assumed the diminishing marginal benefit function from the third

Figure 2 Probabilities of Enrollment and Cutoff Values



example⁹ and use the estimation results from the “No Visitation/With Supplementary Data” case. We graph the cutoff function for several values of d to show how the discount rate affects the university’s decision. The arrows represent the appropriate cutoff score for $d = 0.04\%$ (the leftmost arrow), 0.02% and 0 (the rightmost arrow)

A simple geometric interpretation of the diagram explains the university’s decision process. The point at which the probabilities of enrollment and the cutoff curves intersect determines the optimal number of prospects to be contacted, as the university should contact the first m^* prospects such that the expected marginal benefit of enrollment exceeds the marginal cost contacting the prospect. Reducing the rate at which additional enrollees are discounted flattens the

slope of the cutoff function, and the university contacts more prospects. At $d = 0$ (the constant marginal benefit case) the cutoff function becomes a line of slope zero. With regard to the other financial conditions, reducing the ratio of the marginal cost to the marginal benefit shifts the cutoff function downward, and again the university contacts more prospects. At $c/b = 0$, every prospect with a positive chance of enrollment is contacted.

The HBVC and null models’ influence on the university’s decisions is clear. The HBVC model provides greater discrimination among the prospects because greater variation exists in the probabilities of enrollment. Subsequently, both the probabilities of enrollment and the cutoff curves have a steeper slope, and fewer prospects are contacted when using the HBVC model rather than the null, no matter the discount rate. By itself, the diagram does not suggest that using the HBVC model is more profitable because it does not judge whether the model accurately determines which prospects are best. Nonetheless, when coupled with the ROC chart analysis that shows the HBVC

⁹ We use the benefit function from third example throughout this section because it contains the first example as a limiting case and, as will be shown, the profit can be graphed on a three-dimensional chart.

model provides a better ordering of the prospects, the diagram suggests that the HBVC model cuts out an unprofitable segment of prospects and should be more profitable. We use predictive distributions to assess this notion directly.

The posterior predictive distributions in our analysis are based on the prospects in the holdout sample. We predict the responses of individual prospects, the number of prospects that enroll when the university acts optimally, and the profit arising from each model under various financial constraints. In doing so, we are assessing the models' performance at the point at which decisions are being made. Let \check{S}_{m^*} represent the predicted number of enrollees when the university acts optimally by mailing to m^* of M prospects. The inverted cap on S represents that the predicted number of enrollees, being a response from the holdout prospects, is yet unobserved. In the universally diminishing marginal benefit case, the posterior predictive distribution of profit can be expressed as

$$\text{Profit}(\check{S}_{m^*}) = b \left[\sum_{j=1}^{\check{S}_{m^*}} (1-d)^{j-1} - m^* (c/b) \right].$$

We express the profit in this manner so that we can reduce the number of financial parameters from three to two in our diagrams.

The procedure to obtain a sample from the posterior predictive distribution, \check{S}_{m^*} , for the prospects in the holdout sample entails several steps. First, using both the characteristics of the M prospects in the holdout sample and the T draws of the parameters in the probit model, calculate the T draws of the vector of enrollment probabilities $\check{\theta}^{(t)}$. Second, determine the optimal number of prospects to contact m^* given the probabilities of enrollment. Finally, use a Bernoulli random variable to draw the predicted responses $\check{R}^{(t)}$ of the contacted prospects given the probabilities of enrollment $\check{\theta}^{(t)}$. The predictive value $\check{S}_{m^*}^{(t)}$ is the sum of the m^* responses on the t th iteration of the sampler; the collection of predictive values is a sample from the posterior predictive distribution.

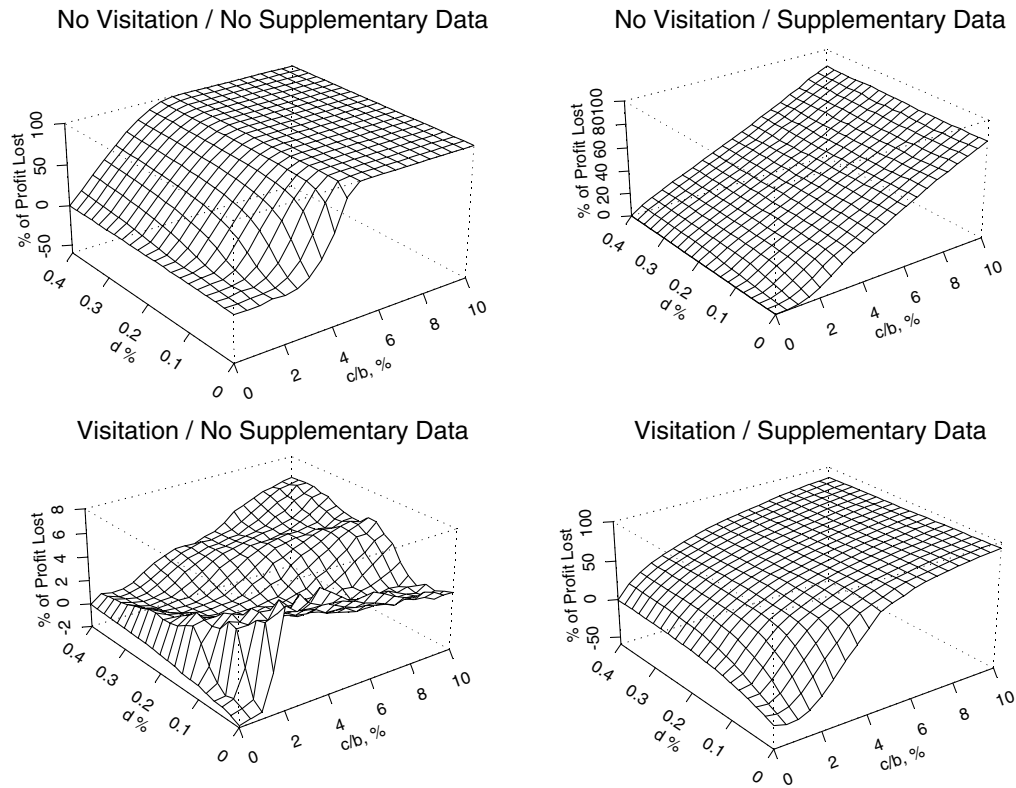
In Figure 3, we graph the expected percent of profit lost from using the null model. (We report the per-

cent of profit lost from using the null model instead of the percent of profit gained from using the HBVC model because the expected total profit from using the null model is zero sometimes.) We assume an array of financial conditions to construct the diagram. The marginal cost to marginal benefit ratio c/b ranges from 0 to 0.1 in increments of 0.005. Near zero, the model choice should have very little influence on the total profit gained because all of the prospects with a positive chance of enrollment are contacted. The discount rate of additional enrollees d ranges from 0 to 0.04% in increments of 0.002%. At zero, the graph describes the degenerate constant marginal benefit case. Our assumptions result in having to compute 1,000 posterior predictive distributions of \check{S}_{m^*} in each of the four information conditions.

The four diagrams show that the HBVC model makes a significant practical difference in the university's profit. Averaged over the four information sets and the array of financial conditions, the expected percent of profit lost is 43.6%. The expected loss from using the null model is greatest when the university has the least amount of information on which to base its predictions, such as when no campus visitation data are available.

We notice an anomaly in the expected profit calculations when $c/b = 0$, particularly near $d = 0$, when both the campus visitation and the supplementary data are available. The expected loss is negative, but clearly it should be zero, as both models recommend that all prospects should be contacted. This suggests that at least one of the models does not accurately predict the amount of profit to be gained. We test the predictions of the model against the actual data by calculating the proportion of profit values simulated from the posterior predictive distribution that are greater than the realized profit (calculated using the actual number of holdout prospects that enroll rather than the predicted number of enrollees). In the area of the diagram where the expected loss is negative, we find that the realized data plausibly could have come from the HBVC model, as 35.8% of the predicted profit values are greater than the actual profit

Figure 3 Expected Percent of Profit Lost from Using the Null Model

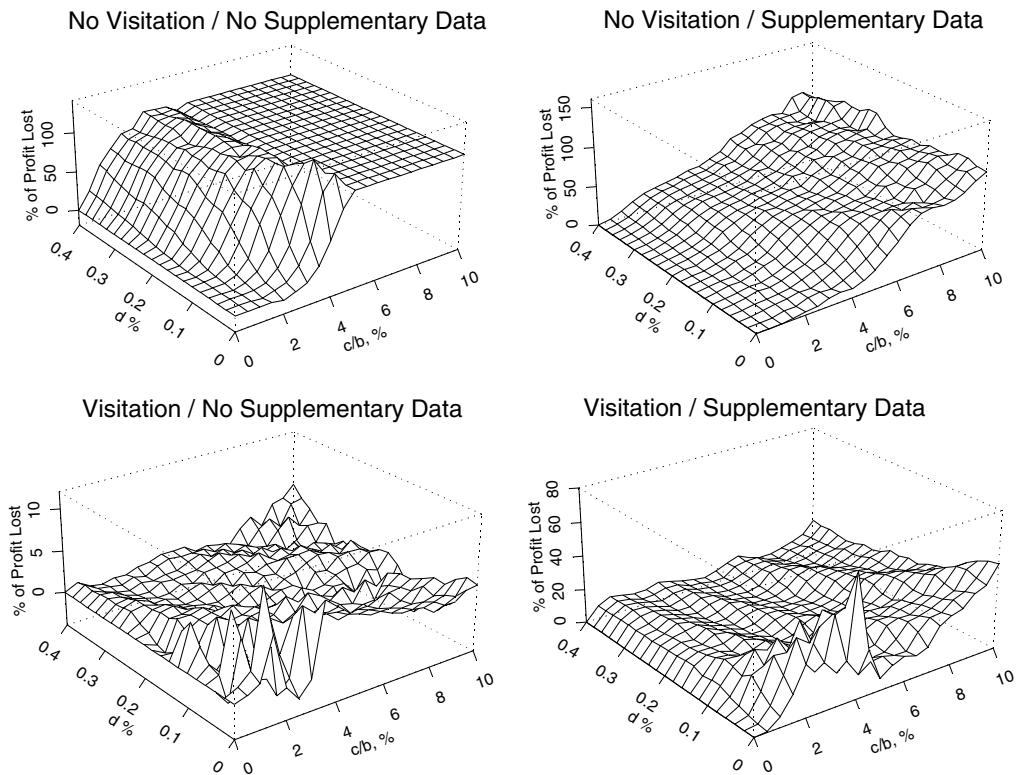


realized from the holdout prospects (a posterior predictive p-value of 0.358). The realized data, on the other hand, do not seem to come from the null model because 100% of the predicted values are greater than the realized value. The null model over predicts the amount of profit generated in this region, which we attribute to the null model's excessive certainty about the parameter values, reflected in the overly tight HPDs in Table 4.

This raises an interesting concern with using the predictive distributions to compare the models. While we would like to use the difference between the expected posterior predictive profits to compare the models because they are not specific to the responses found in the dataset, the comparison is useful only if both models provide accurate predictions. We note that even though the null model does not tend to over-predict the profit on the whole, a concern still remains about the usefulness of the comparisons.

To alleviate any concern about the predictive accuracy of the models, we graph the realized percent of profit lost from using the null model in Figure 4. The realized profit is a noisier measure than the average predictive profit because its value is based on the actual enrollment decisions of the prospects in the holdout sample rather than their expected responses. On the other hand, it represents the actual difference in profit that would be realized if the university takes the actions prescribed by the models for the holdout sample. The probabilities influence the decision to target a prospect but do not determine the prospect's value in this measure. The realized profit is useful when comparing different modeling techniques because a prospect's value is not inflated by an overly optimistic prediction about whether he or she will enroll, and it is particularly helpful when the underlying assumptions of any of the models are brought into question. In this case, the realized profit diagrams confirm what we found in the expected

Figure 4 Realized Percent of Profit Lost from Using the Null Model



profit ones, while resolving the anomaly at the borders where $c/b = 0$.

6. Conclusions and Future Research

In this paper, we describe the problems that arise from using the masked data sold by consolidators. We develop a flexible statistical model to take full advantage of the information contained in masked data and show that the conventional modeling practice, which ignores the masking, is overly certain about where the parameter values lie. We also introduce the idea of a massively categorical variable and show how it can be used to replicate some of the information contained in a consolidator's database. We examine the competing models over an array of financial conditions in order to make robust comparisons. In the case of a nonlinear benefit function, we demonstrate how to use the MCMC output from parameter estimation

to implement the decision rule that determines how many prospects are contacted.

This research might proceed in several directions. The first promising area is to develop a hierarchical Bayes technique to select variables from the abundant databases that are sold by consolidators. Variable selection continues to be a major concern for direct marketers, and it is clear that conventional methods cannot be used when some of the variables are masked. The second area is to consider incorporating semiparametric or nonparametric techniques (such as those in Bult 1993) within the Bayesian model's hierarchy. Direct marketers often suspect that many variables have complex, nonlinear relationships with the dependent variable, which has led to the widespread use of neural nets and similar methodologies. Semiparametric approaches might allow for such complexity without the disadvantages of neural nets. Finally, we foresee the use of massively categorical variables in developing models of consumer behavior on the Internet. Just as zip codes provide useful information

in our application, Web-page addresses might do the same in others.

Acknowledgments

The authors thank the anonymous reviewers and AE for input that greatly added to the contribution of the paper; Peter Boatwright, Ed Kaplan, Subrata Sen, K. Sudhir, P. B. Seetharaman, Xavier Drèze, and Dick Wittink for comments; participants in seminars held at UCLA, USC, Washington University, and the BAMMCONF for helpful suggestions; and Yale, UCLA, Cornell, and the BAMMCONF for financial assistance.

Appendix A: The Hierarchical Bayes Variance Components Model

The most general model we consider includes both individual-level effects and zip-level effects. The zip-level effects are modeled in a hierarchical structure.

$$y_i = X_i^T \alpha + \beta_{m_i} + \beta_{z_i} + \varepsilon_i \quad \text{for } i = 1, \dots, N \text{ individuals,}$$

where $\varepsilon_i \sim N(0, 1)$,

X_i^T is a 1 by k vector consisting of an intercept and $(k-1)$ individual-level explanatory variables,

α is a k by 1 vector of coefficients describing preferences,

m_i is the major of the i th individual, and

z_i is the zip code of the i th individual.

The effects for the student's intended major are modeled as

$$\beta_m \sim N(0, V_{\beta_m}) \quad \text{for } m = 1, \dots, M \text{ intended majors.}$$

Finally, the zip-level effects are modeled as

$$\beta_z = W_z^T \gamma + v_z \quad \text{for } z = 1, \dots, Z \text{ zip codes,}$$

where $v_z \sim N(0, V_{\beta_z})$, W_z^T is a 1 by d vector of demographic explanatory variables for the z th zip code, and γ is a d by 1 vector of zip code level effects. For identification purposes, there is no intercept (i.e., no vector of 1s) in W_z . When demographics are excluded, this step is treated the same as β_m .

The Prior Distributions

1. $\alpha \sim N_k(\mu_{\alpha 0}, V_{\alpha 0})$, where $\mu_{\alpha 0} = 0_k$ and $V_{\alpha 0} = 10^6 I_k$.
2. $V_{\beta_m} \sim \text{Inv} \chi^2(v_{m0}, v_{m0} \sigma_{m0}^2)$, where $v_{m0} = 4$ and $\sigma_{m0}^2 = 1$.
3. $\gamma \sim N_d(\mu_{\gamma 0}, V_{\gamma 0})$, where $\mu_{\gamma 0} = 0_d$ and $V_{\gamma 0} = 10^6 I_d$.
4. $V_{\beta_z} \sim \text{Inv} \chi^2(v_{z0}, v_{z0} \sigma_{z0}^2)$, where $v_{z0} = 4$ and $\sigma_{z0}^2 = 1$.

The Gibbs Sampler

1. Set $\mu_i = X_i^T \alpha + \beta_{m_i} + \beta_{z_i}$.
2. Draw $y_i | \alpha, \{\beta_m\}, \{\beta_z\}, I_i$ from a truncated $N(\mu_i, 1)$. The draw is truncated to be above 0 if $I_i = 1$ (e.g. if the person enrolls), and to be below otherwise. See Albert and Chib (1993) for details.
3. Set $y_i^{\text{indiv}} = y_i - (\beta_{z_i} + \beta_{m_i})$.
4. Draw $\alpha | y_i^{\text{indiv}}, \{\beta_m\}, \{\beta_z\}$ from $N_k((V_{\alpha 0}^{-1} + X^T X)^{-1}(V_{\alpha 0}^{-1} \mu_{\alpha 0} + X^T y^{\text{indiv}}), (V_{\alpha 0}^{-1} + X^T X)^{-1})$.

5. Set $y_i^{maj} = y_i - (X_i^T \alpha + \beta_{z_i})$ for $i = 1, \dots, N$. Let $\tilde{y}_m^{maj} = \sum_{i=1}^N y_i^{maj} \mathbf{1}_{\{m_i=m\}}$ for $m = 1, \dots, M$.
6. Draw $\beta_m | \tilde{y}_m^{maj}, V_{\beta_m}$ from $N((n_m + V_{\beta_m}^{-1})^{-1} \tilde{y}_m^{maj}, (n_m + V_{\beta_m}^{-1})^{-1})$.
7. Draw $V_{\beta_m} | \{\beta_m\}$ from $\text{Inv} \chi^2(v_{m0} + M, v_{m0} \sigma_{m0}^2 + \sum_{m=1}^M \beta_m^2)$.
8. Set $y_i^{zip} = y_i - (X_i^T \alpha + \beta_{m_i})$ for $i = 1, \dots, N$. $\tilde{y}_z^{zip} = \sum_{i=1}^N y_i^{zip} \mathbf{1}_{\{z_i=z\}}$ for $z = 1, \dots, Z$.
9. Draw $\beta_z | \tilde{y}_z^{zip}, \gamma, V_{\beta_z}$ from $N((V_{\beta_z}^{-1} + n_z)^{-1}(V_{\beta_z}^{-1} W_z^T \gamma + \tilde{y}_z^{zip}), (V_{\beta_z}^{-1} + n_z)^{-1})$.
10. Draw $\gamma | \{\beta_z\}, V_{\beta_z}$ from $N_d((V_{\gamma 0}^{-1} + W^T W)^{-1}(V_{\gamma 0}^{-1} \mu_{\gamma 0} + W^T \beta_z), (V_{\gamma 0}^{-1} + W^T W)^{-1})$.
11. Draw $V_{\beta_z} | \{\beta_z\}, \mu_{\beta_z}$ from inverse $\chi^2(v_{z0} + Z, v_{z0} + \sum_{z=1}^Z (\beta_z - W_z^T \gamma)^2)$.

Modification: No Hierarchical Regressors

Note that steps 9–11 above do not apply if there are no hierarchical regressors, i.e., when there are no consolidator variables. In that case, to retain identification, the distribution of β_z is distributed about zero, and steps 9–11 are replaced by steps identical to those used for major effects, i.e., steps 6 and 7.

Appendix B: Derivations

Derivation One

We want to derive Equation (7).

$$\begin{aligned} \Delta(m) &\equiv E[h(S_m) - h(S_{m-1}) | \theta] = \sum_{s=0}^m h(s) p_{S_m}(s) - \sum_{s=0}^{m-1} h(s) p_{S_{m-1}}(s) \\ &= \theta_m \sum_{s=0}^m h(s) p_{S_{m-1}}(s-1) + (1 - \theta_m) \sum_{s=0}^m h(s) p_{S_{m-1}}(s) \\ &\quad - \sum_{s=0}^{m-1} h(s) p_{S_{m-1}}(s) \\ &= \theta_m \sum_{s=0}^{m-1} h(s+1) p_{S_{m-1}}(s) + (1 - \theta_m) \sum_{s=0}^{m-1} h(s) p_{S_{m-1}}(s) \\ &\quad - \sum_{s=0}^{m-1} h(s) p_{S_{m-1}}(s) \\ &= \theta_m \sum_{s=0}^{m-1} [h(s+1) - h(s)] p_{S_{m-1}}(s) \\ &= \theta_m \sum_{s=0}^{m-1} \delta(s+1) p_{S_{m-1}}(s) \quad \text{for } m \in \{1, \dots, M\}. \end{aligned}$$

Derivation Two

$$\begin{aligned} \Psi(m) &= \frac{c}{\sum_{s=0}^{m-1} \delta(s+1) p_{S_{m-1}}(s)} \\ &= \frac{c}{\sum_{s=0}^q b^s p_{S_{m-1}}(s) + \sum_{s=q+1}^{m-1} (b-a) p_{S_{m-1}}(s)} \\ &= \frac{c}{b-a} \frac{c}{\sum_{s=0}^{m-1} p_{S_{m-1}}(s)} = \frac{c}{b-a} \frac{c}{\Pr\{S_{m-1} > q\}}. \end{aligned}$$

Derivation Three

The denominator of the cutoff function $\Psi(m+1)$ can be rewritten as

$$\begin{aligned} \sum_{s=0}^m \delta(s+1)p_{s_m}(s) &= \sum_{s=0}^m \delta(s+1)[\theta_m p_{s_{m-1}}(s-1) + (1-\theta_m)p_{s_{m-1}}(s)] \\ &= \theta_m \sum_{s=1}^m \delta(s+1)p_{s_{m-1}}(s-1) + (1-\theta_m) \\ &\quad \cdot \sum_{s=0}^m \delta(s+1)p_{s_{m-1}}(s) \\ &= \theta_m \sum_{s=0}^{m-1} \delta(s+2)p_{s_{m-1}}(s) + (1-\theta_m) \\ &\quad \cdot \sum_{s=0}^{m-1} \delta(s+1)p_{s_{m-1}}(s). \end{aligned}$$

Since $\delta(s+2) = (1-d)\delta(s+1)$ in this example,

$$\begin{aligned} \sum_{s=0}^m \delta(s+1)p_{s_m}(s) &= [\theta_m(1-d) + (1-\theta_m)] \sum_{s=0}^{m-1} \delta(s+1)p_{s_{m-1}}(s) \\ &= (1-d\theta_m) \sum_{s=0}^{m-1} \delta(s+1)p_{s_{m-1}}(s). \end{aligned}$$

Thus,

$$\begin{aligned} \Psi(m+1) &= \frac{1}{(1-d\theta_m) \sum_{s=0}^{m-1} \delta(s+1)p_{s_{m-1}}(s)} \\ &= \frac{1}{(1-d\theta_m)} \Psi(m). \end{aligned}$$

Proof

We want to show that the cutoff function weakly increases as more prospects are contacted. The cutoff function given in Equation (8) can be rewritten as $\Psi(m) = c\theta_m/\Delta(m)$.

$$\begin{aligned} \Delta(m) &= \theta_m \sum_{s=0}^{m-1} \delta(s)p_{s_{m-1}}(s) \\ &= \theta_m \sum_{s=0}^{m-1} \delta(s) \left[\theta_{m-1} p_{s_{m-2}}(s-1) + (1-\theta_{m-1}) p_{s_{m-2}}(s) \right] \\ &= \theta_m \left[\theta_{m-1} \sum_{s=1}^{m-1} \delta(s)p_{s_{m-2}}(s-1) + (1-\theta_{m-1}) \sum_{s=0}^{m-2} \delta(s)p_{s_{m-2}}(s) \right] \\ &= \theta_m \left[\theta_{m-1} \sum_{s=0}^{m-2} \delta(s+1)p_{s_{m-2}}(s) + \frac{(1-\theta_{m-1})}{\theta_{m-1}} \Delta(m-1) \right] \\ &\quad \text{for } m \in \{2, 3, 4, \dots\}. \end{aligned}$$

Since $\delta(s+1) \leq \delta(s)$ by assumption,

$$\begin{aligned} \Delta(m) &\leq \theta_m \left[\theta_{m-1} \sum_{s=0}^{m-2} \delta(s)p_{s_{m-2}}(s) + \frac{(1-\theta_{m-1})}{\theta_{m-1}} \Delta(m-1) \right] \\ &\leq \frac{\theta_m}{\theta_{m-1}} \Delta(m-1). \end{aligned}$$

Thus,

$$\Psi(m) = \frac{c\theta_m}{\Delta(m)} \geq \frac{c\theta_m}{(\theta_m/\theta_{m-1})\Delta(m-1)} = \frac{c\theta_{m-1}}{\Delta(m-1)} = \Psi(m-1),$$

which completes the proof.

References

- Ainslie, A., P. E. Rossi. 1998. Brand choice across multiple categories: A hierarchical error components model. *Marketing Sci.* 17(2) 91–106.
- Albert, J., S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* 88 669–679.
- Allenby, G., N. Arora, J. Ginter. 1998. On the heterogeneity of demand. *J. Marketing Res.* 35(Aug) 384–389.
- Berger, J. O. 1985. *Statistical Decision Theory*. Springer-Verlag, Berlin.
- Bronnenberg, Bart J., Catarina Sismeiro. 2002. Using multimarket data to predict brand performance in markets for which no or poor data exist. *J. Marketing Res.* 39(Feb.) 1–17.
- Bult, J. R. 1993. Semiparametric versus parametric classification models: An application to direct marketing. *J. Marketing Res.* 3(0) 380–390.
- , T. Wansbeek. 1995. Optimal selection for direct mail. *Marketing Sci.* 14 378–394.
- Gelman, A., J. Carlin, H. Stern, D. Rubin. 1995. *Bayesian Data Analysis*. Chapman and Hall, London.
- Gelman, A., Y. Goegebeur, F. Tuerlinckx, I. Van Mechelen. 2000. Diagnostic checks for discrete data regression models using posterior predictive simulations. *J. Roy. Statist. Soc. Ser. C (Appl. Statist.)* 49(2) 247–268.
- Gilks, W., S. Richardson, D. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.
- Hoch, Stephen J., Byung-Do, Kim, Alan L. Montgomery, Peter E. Rossi. 1995. Determinants of store-level price elasticity. *J. Marketing Res.* 32(1) 17–30.
- Manski, C. F., D. A. Wise. 1993. *College Choice in America*. Harvard University Press, Cambridge, MA.
- Newton, M., A. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Roy. Statist. Soc. B* 56(1) 3–48.
- Rossi, P., R. McCulloch, G. Allenby. 1996. Purchase history data in target marketing. *Marketing Sci.* 15 321–340.
- Swets, J. A. 1996. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics*. Lawrence Erlbaum Associates, Mahwah, NJ.
- Wilcox, T., A. Hsu. 2000. Stochastic prediction in multinomial logit models. *Management Sci.* 46(8) 1137–1144.
- Yang, Sha, Greg M. Allenby. 2002. Modeling socially dependent preferences. Working Paper, Ohio State University, Columbus, OH.

This paper was received August 23, 1999, and was with the authors 15 months for 3 revisions; processed by Greg M. Allenby.