Forecasting Household Response in Database Marketing:
A Latent Trait Approach

Subom Rhee
Stonehill College

Gary J. Russell
University of Iowa

October 2008


Subom Rhee is Assistant Professor of Business at the Department of Business Administration, Stonehill College. Gary J. Russell is Henry B. Tippie Research Professor of Marketing at the Tippie College of Business, University of Iowa. The authors thank the Direct Marketing Educational Foundation for providing access to the data used in this study. The authors also thank Professor John Geweke of the Department of Economics, University of Iowa, for many helpful suggestions on model specification and estimation. Please address correspondence to: Professor Subom Rhee, Department of Business Administration, Stonehill College, 320 Washington St, Easton, MA 02357, phone: (508) 565-1854, fax: (508) 565-1444, email: srhee@stonehill.edu.
ABSTRACT

Forecasting Household Response in Database Marketing: A Latent Trait Approach

Database marketers often select households for individual marketing contacts using information on past purchase behavior. One of the most common methods, known as RFM variables approach, ranks households according to three criteria: the recency of the latest purchase event, the long-run frequency of purchases, and the cumulative dollar expenditure. We argue that RFM variables approach is an indirect measure of the latent purchase propensity of the customer. In addition, the use of RFM information in targeting households creates major statistical problems (selection bias and RFM endogeneity) that complicate the calibration of forecasting models. Using a latent trait approach to capture a household’s propensity to purchase a product, we construct a methodology that not only measures directly the latent propensity value of the customer but also avoids the statistical limitations of the RFM variables approach. The result is a general household response forecasting and scoring approach that can be used on any database of customer transactions. We apply our methodology to a database from a charitable organization and show that the forecasting accuracy of the new methodology improves upon the traditional RFM variables approach.

KEYWORDS: Forecasting, Latent Trait Approach, Database Marketing, RFM, Selection Bias, Endogeneity
INTRODUCTION

Database marketing is an increasingly important aspect of the management of traditional catalog retailers (such as Lands’ End) and e-commerce firms (such as Amazon.com). In database marketing, the manager has access to a household database detailing each interaction with the firm over a period of time. The task of the marketing manager is to use the database information to develop predictive models of household purchasing, and then to target segments of households for specific marketing programs (Winer 2001, Berger and Nasr 1998, Hughes 1996).

Scoring Households Using RFM

Firms frequently use household purchase characteristics, collectively known as RFM variables, in selecting the best households for marketing solicitations (Hughes 1996). Recency (R) is defined as the number of periods since the most recent purchase. Frequency (F) is defined as the total number of purchases. Monetary value (M) is defined as the dollar amount that the household has spent to date. Conceptually, RFM variables are used for forecasting because past purchase behavior is often a reliable guide to future purchase behavior (Schmid and Weber 1997, Rossi, McCulloch and Allenby 1996). The predictive power of the three variables is traditionally known as having the rank order: recency is the best predictor, followed by frequency, and then monetary value (David Shepard Associates, Inc. 1999). Forecasting the customer’s response likelihood using RFM variables is widely accepted by database marketers as an easy and useful way of predicting behavior from a customer database.

RFM information is inserted into predictive models. For example, RFM values can be used as independent variables in a probit or logit response model. Additional
procedures drawn from the data mining literature, such as decision trees and neural
networks, can also be used to link RFM values to buying behavior (Hastie, Tibshiramti
and Friedman 2001, Berry and Linoff 2000).

Statistical Problems Induced by RFM

The use of RFM in response modeling appears straightforward on the surface but
important statistical problems arise.

The first problem is known as selection bias. Simply put, selection bias arises when
the researcher uses a non-randomly selected sample to estimate behavioral relationships
(Heckman 1979). If the firm selects households for mailings based on a non-random
selection rule (such as the RFM variables), a study that only analyzes the selected
households generates biased results. This bias arises from the fact that the researcher
does not observe the behavior of non-selected households. Selection bias is a special
type of missing data problem that can only be controlled by formally analyzing the way
that the firm selects customers for marketing solicitations.

The second problem, known as RFM endogeneity, occurs when RFM values not
only represent the past response behavior of the households, but also reflect the past
selection decision of the firm. For instance, if a household is not selected to receive a
marketing offer (and the household has no way to respond to the offer otherwise), the
recency (the number of periods since the last purchase) will be larger and the frequency
and the monetary value will be smaller, than the values of these same variables for a
comparable household who received the solicitation. If the firm consistently ignores the
household for any reason, the RFM values of this household will deteriorate regardless of
the true propensity to respond. In formal statistical terms, it can be shown that RFM
endogeneity yields incorrect parameter estimates in a predictive model due to unobserved correlations between the RFM variables and the error in the model (see, e.g., Davidson and MacKinnon 1993).

The marketing science community has gradually begun to recognize these statistical problems. Industry standard procedures (such as RFM probit regression) and the early model of Bult and Wansbeek (1995) ignore these problems entirely. Jonker et al. (2000) addresses selection bias by relating RFM variables to both household selection and household response. However, because RFM values appear in the specification, issues of endogeneity are not addressed. Studies by Bitran and Mondstein (1996) and by Gonul and Shi (1998) provide an approach to dealing with RFM endogeneity. By replacing observed RFM values with predicted values, these authors construct an instrumental variables methodology (see Davidson and MacKinnon 1993) that corrects for potential parameter biases. In the applications discussed by these authors, households are able to buy products even if a marketing solicitation is not received. Accordingly, parameter biases due to selection bias are not relevant and are not addressed.

Latent Trait Scoring Model

This research builds upon existing work by developing an approach to household scoring which corrects for both selection bias and RFM endogeneity. In contrast to earlier studies, we assume that each household has a latent (unobserved) propensity to respond that cannot be adequately captured using RFM variables. Latent trait models have a long history in psychometric studies of psychological constructs such as verbal and quantitative ability (see, e.g., Lord and Novick 1968, Fischer and Molenarr 1995, Langeheine and Rost 1988). These models have also found marketing science
applications in survey research (Balasubramanian and Kamakura 1997), coupon redemption (Bawa et al. 1997) and cross-selling of financial services (Kamakura et al. 1991). Although latent trait models can be regarded as a type of random coefficient heterogeneity model (Allenby and Rossi 1999), they are best viewed as a method of measuring a psychological trait. In this research, we view household scoring as a research procedure designed to estimate a household’s propensity to respond to marketing solicitations by the firm.

Our model is based upon the assumption that both the firm’s selection rule and the household’s response behavior provide indirect indications of the household’s latent propensity. The notion here is that the firm does not select households for mailings using either a census (all households) or a random process (probability sample of households). Instead, the firm selects households using some process which takes into account the likelihood that the household will respond favorably. We do not, however, assume that the selection process is necessarily optimal. The key advantage of this approach is generality: the researcher can estimate a household response model on existing databases in which the firm has attempted to optimize customer contact policy. As we show subsequently, we are able to measure each household’s true propensity to respond and to examine the effectiveness of the firm’s current contact policy.

The remainder of the paper is organized as follows. We first detail our new model, discussing the need to consider both household response and the firm’s household selection rule simultaneously. We demonstrate that our latent trait specification can be formulated as a Hierarchical Bayes model and estimated using Monte Carlo simulation technologies. The new methodology is then applied in an analysis of the customer
database of a non-profit organization. We show that the model provides forecasts with a level of accuracy better than a benchmark RFM probit model. We conclude with a discussion of future research opportunities.

LATENT TRAIT MODEL OF RESPONSE PROPENSITY

We begin by describing the structure of a general model of household choice behavior in a database marketing context. Instead of relying upon RFM variables to measure the propensity of each household to respond to marketing solicitation, we assume the existence of a household-specific latent trait that impacts the household’s probability of responding to a solicitation. This same latent variable is also assumed to impact the firm’s likelihood of targeting the household. By developing the model in this manner, we correct for selection bias (if present) and avoid issues of RFM endogeneity. The result is a general model of household purchase behavior that takes into account potential limitations of the RFM variables approach.

Propensity to Respond

Propensity to respond is defined here as a household characteristic that reflects the household’s inherent interest in the firm’s product offering. We assume that this propensity has two components: a long-run component that varies only by household, and a short-run component that varies over households and time. Implicitly, the long-run component accounts for heterogeneity in response across households, while the short-run component accounts for temporal variation in response propensity within households.

Let \( \tau_{h,t} \) denote the propensity to respond of household \( h \) at time \( t \). We define this construct according to the equation
\[ \tau_{h,t} = \delta_1 \mu_h + \delta_2 Y_{R(h,t-1)} \]  

(1)

where \( Y_{R(h,t-1)} = 1 \) if the household responded to the previous solicitation, and 0 otherwise. Further, we assume that the long-run component is normally distributed across the household population as

\[ \mu_h \sim N( X_h \gamma, 1.0 ) \]  

(2)

where \( X_h \) denotes a set of demographic variables for household \( h \).

This formulation has two key properties. First, \( \tau_{h,t} \) changes over time depending upon whether or not the household purchased at the previous time point. Note that non-purchase at time \( t-1 \) could be due to a rejection of the previous offer. Alternatively, the household may have not been given an opportunity to purchase the product. In our formulation, it is not necessary to distinguish between these two cases. Rather, similar to models of choice inertia found in grocery scanner data applications (Seetharaman and Chintagunta 1998, Jeuland 1979), we assume that the act of purchasing a product at one time point has an impact on future behavior, regardless of why the product was purchased. Note that the parameter on the lagged component (\( \delta_2 \)) can be either positive or negative. Thus, the impact of the short-term component can either enhance or diminish purchasing at the next period.

The normal distribution characterizing the long-run component is intended to allow for heterogeneity in response across households. Note that this distribution has a mean that depends upon demographics and a variance set to one. Intuitively, this formulation states that the long-run component depends in part on demographics, and in part on other (unobserved) household characteristics. Setting the variance of this distribution to one
can be accomplished without loss of generality. This restriction is necessary for model identification and does not impact the fit of the model.

**Modeling the Firm’s Targeting Decision**

We assume that the firm is attempting to optimize its targeting policy using some information in the household database. This information may or may not be RFM variables. We assume that the firm’s rule has some validity and is correlated to some extent with $\tau_{h,t}$, the household’s propensity to respond. *We stress that this assumption does not imply that the firm actually observes* $\tau_{h,t}$. Rather, the expression displayed below for the firm’s selection rule is simply a formal way of stating that households are not necessarily selected at random. During model estimation, the researcher learns the extent to which the current targeting policy is based upon some knowledge of household’s true propensity to respond. In the language of econometric theory, our model of the firm’s targeting policy is a limited information specification – not a structural specification.

To model this process, we assume that the firm’s decision to target a household depends upon the attractiveness of the household to the firm. Define the attractiveness of a household $h$ to the firm at time $t$ as $U_{S(h,t)}$. We assume that the firm makes a product offer to this household ($Y_{S(h,t)} = 1$) if $U_{S(h,t)}$ is greater than zero. Otherwise, $Y_{S(h,t)} = 0$. Hence, $U_{S(h,t)}$ is a latent variable that drives the observed targeting policy of the firm.

To complete the specification, we assume that the deterministic part of $U_{S(h,t)}$ is a linear function of the household’s propensity to respond $\tau_{h,t}$, defined in equation (1). This leads to the model

$$U_{S(h,t)} = \alpha_0 + \alpha_1 \mu_h + \alpha_2 Y_{R(h,t-1)} + \epsilon_{S_{h,t}}, \quad \epsilon_{S_{h,t}} \sim N(0,1) \quad (3)$$
where the normally distributed error $\varepsilon_{S_{h,t}} \sim N(0,1)$ has mean zero and variance one.

The assumption that the variance of the error is equal to one is necessary for model identification; it has no impact on model fit.

We again emphasize that this expression does not imply that the firm knows the household propensity to respond as measured by $\tau_{h,t}$. All that this expression states is that the attractiveness of a household to the firm is correlated to some extent with the household’s propensity to respond. This specification of the household selection process is identical to that of a probit model for the binary variable $Y_{S(h,t)}$. Intuitively, this model allows for the possibility that households that are selected are likely to be better prospects for the firm.

**Modeling Household Response**

In an analogous fashion, we assume that the household’s decision to respond to a product offer depends upon the attractiveness (or utility) of the offering. Define the attractiveness of a marketing offering to household $h$ at time $t$ as $U_{R(h,t)}$. We assume that the household buys the product ($Y_{R(h,t)} = 1$) if $U_{R(h,t)}$ is greater than zero. Otherwise, $Y_{R(h,t)} = 0$. Hence, $U_{R(h,t)}$ is a latent variable that determines the response behavior of the household.

Given our definition of the household response propensity, we assume that

$$U_{R(h,t)} = \beta_0 + \beta_1 \mu_h + \beta_2 Y_{R(h,t-1)} + \varepsilon_{R_{h,t}}, \quad \varepsilon_{R_{h,t}} \sim N(0,1) \tag{4}$$

where the normally distributed error $\varepsilon_{R_{h,t}} \sim N(0,1)$ has mean zero and variance one.

Again, for model identification reasons, we can set the variance of the error to one without loss of generality. Intuitively, this specification amounts to the assumption that
the deterministic part of $U_{R(h,t)}$ is a linear function of the household’s propensity to respond $\tau_{t,h}$ defined in equation (1).

Because the error in this expression is normally distributed, the model for the household purchase variable is a probit model, conditional upon the long-run and short-run elements of the propensity to respond construct. It should be noted that this model is only applied to households who are targeted by the firm. Households who do not receive a product offer cannot buy the product. For these households, $Y_{R(h,t)}$ must be equal to 0. Stated differently, we can only estimate the response model over the set of households at a particular time point who receive a product offer from the firm.

Properties of the Errors

To complete the specification of the model, we make two key assumptions about the errors in the firm targeting equation and the household response equation. First, we assume that these errors are mutually independent at each time point. Second, we assume that these errors are independent over time.

The first assumption amounts to the notion of conditional independence. The intuition is that the household’s propensity to respond drives both firm behavior and household behavior. Consequently, conditional on the values of $\mu_h$ and $Y_{R(h,t-1)}$, we can assume that the selection and response error terms in this model are independent. Since different values of $\mu_h$ and $Y_{R(h,t-1)}$ lead to different values of selection and response, our model implies a natural correlation between observed selection and observed response across the household population. Put another way, conditional independence allows for a simpler representation of the choice process without sacrificing the reality that selection and response are correlated. Conditional independence is a key element of model
construction both in psychometrics (Lord and Novick 1968, Fischer and Molenarr 1995, Langeheine and Rost 1988) and in marketing science (e.g., Kamakura and Russell 1989, Rossi et al. 1996).

The second assumption is necessary to prevent endogeneity issues from entering the model through the lagged response variable. Lagged response $Y_{R(h,t-1)}$ is already modeled in the system by the selection and response equations at time $t-1$. Given the value of propensity to respond at the previous period, the probability that $Y_{R(h,t-1)}$ equals one is a product of the probability of selection and the probability of response in period $t-1$. Consequently, in the context of the response model, the observed lagged response is only correlated with the error terms in previous time periods (i.e., periods $t-2$, $t-3$, $t-4$ etc.). Since the error terms are assumed independent over time, the lagged response $Y_{R(h,t-1)}$ cannot be correlated with the error terms in the current period ($\varepsilon_{S,h,t}$ or $\varepsilon_{R,h,t}$). Thus, the inclusion of $Y_{R(h,t-1)}$ in the model does not create endogeneity problems.

It is important to notice that the model developed here does not suffer from the problems of selection bias and endogeneity noted in our earlier discussion of RFM models. Both $\mu_h$ and $Y_{R(h,t-1)}$ are independent of the errors in the selection and response models (equations (3) and (4)), thus eliminating endogeneity from the specification. Moreover, as explained by Heckman (1979), biases in parameter estimation due to selection bias are due entirely to a non-zero correlation between the errors of the selection and response equations. Because the errors $\varepsilon_{S,h,t}$ and $\varepsilon_{R,h,t}$ are contemporaneously independent, selection bias is not present in the estimates generated from our model.
Model Estimation

The proposed model (equations (2), (3) and (4)) is a two-equation probit system with an underlying latent variable measuring the response propensity of each household. (The definition of the propensity to respond construct (equation (1)) is used to motivate the structure of the model, but the $\delta_1$ and $\delta_2$ coefficients are not explicitly estimated by our algorithm.) We calibrate the model by formulating the estimation problem using Hierarchical Bayes concepts and employing Markov Chain Monte Carlo (MCMC) technology to simulate draws from the posterior distribution of parameters (Gelman et al. 1996). Details on the algorithm are presented in the Appendix.

The convergence of the MCMC algorithm was checked using a procedure developed by Geweke (2001). In Geweke’s approach, a second simulation, which uses a different (non-standard) logic to draw the simulated values, is conducted following the initial MCMC analysis of the data. Because Geweke (2001) proves that the initial and the new simulation constitute Markov chains with the same stationary point, the researcher is able to check convergence by verifying that the posterior means and variances from the two simulations are the same. The results reported in this article passed this stringent convergence test.

APPLICATION

To understand the properties of the propensity to respond model, we use a customer database from a non-profit organization. Our intention here is to contrast the latent trait approach to a predictive model based solely upon RFM. It is important to understand that the propensity to respond model does not make any use of traditional RFM variables.
This difference is important because it allows us to compare the industry standard RFM approach to a formulation which ignores RFM variables. From a substantive point of view, this application is also designed to show that the propensity to respond model yields insights into the response characteristics of different types of mail solicitations used by the firm and the operating characteristics of the firm’s current household selection policy.

**Data Description**

The data consist of the transaction records of a non-profit organization that uses direct mail to solicit contributions from donors. Data are taken from the period October 1986 through June 1995. There is one record per past donor. Each record contains information on donor ID, postal code, donation history, and solicitation dates. Since the contribution codes and solicitation codes match, each contribution can be traced to a specific solicitation type and date.

We selected a random sample of 1,065 households for our analysis. Since we need a start-up time period to define RFM values, the final calibration sample contains twenty solicitations during the period from July 1991 to October 1994. The holdout sample for verification of the results contains four solicitations potentially available to these households during the period from November 1994 to March 1995. Overall, households receive mailings from as few as two times to as many as eleven times across twenty time periods.

A preliminary analysis of household donation behavior showed that the amount of money donated by household varies little over time. For example, if a given household donates $5 on one occasion, the household is very likely to donate $5 on every donation
occasion. This fact allows us to regard the amount of the donation as a stable characteristic of the household, and concentrate only on the probability that the household decides to make a donation. Thus, the use of our model – a model which focuses only on the incidence of selection and response ($Y_{S(h,t)}$ and $Y_{R(h,t)}$) – is entirely appropriate for this application.

There are four major solicitation types, types A, B, C, and a miscellaneous type. Type A is the major solicitation type that shows the most frequent and regular mailings every three to six months. Type B includes the holiday mailings of December and January. Types C and miscellaneous are less frequent than types A and B and do not show a regular pattern of mailing. We recode these solicitations as Type A and Type Non-A. In some cases, the type of the solicitation sent to household is not recorded in the dataset. These unknown types are called Type Unknown. By separating out the Type Unknown solicitations, we are able to study the characteristics of Type A and Type Non-A without making unwarranted assumptions. As we show subsequently, the selection and response characteristics of Type A and Type Non-A solicitations are decidedly different.

A postal code dataset is used to obtain a demographic description of each household. This dataset includes postal code, income index, percentage of households occupied by white, black, and Hispanic persons, percentage of households with one or more children under 18, persons per household, household median age, and median years of school for people aged 25 or more. Including gender information taken from the donation record, there are a total of nine demographic features. In order to improve the convergence of our estimation algorithm, we used principal components analysis to create a set of nine
uncorrelated demographic variables. All nine principal component variables are used in the analysis.

Latent Trait Models

In our analysis of the donation dataset, we consider two variants of the propensity to respond model in this application. The most general model, called the Dynamic Model, takes the general form

\[
U_{S(h,t)} = \alpha_{0k} + \alpha_{1k} \mu_h + \alpha_{2k} Y_{R(h,t-1)} + \varepsilon_{S,h,t}, \quad \varepsilon_{S,h,t} \sim \mathcal{N}(0,1)
\]

\[
U_{R(h,t)} = \beta_{0k} + \beta_{1k} \mu_h + \beta_{2k} Y_{R(h,t-1)} + \varepsilon_{R,h,t}, \quad \varepsilon_{R,h,t} \sim \mathcal{N}(0,1)
\]

where \(k\) denotes type of solicitation (Type A, Type Non A, or Unknown), and the errors are mutually independent (contemporaneously and for all possible leads and lags). This is the model discussed earlier.

We also estimate a restricted model, called the Long-Run Model, in which the coefficients on lagged choice (\(\alpha_{2k}\) and \(\beta_{2k}\) for all solicitation types \(k\)) are set to zero. This second model has the form

\[
U_{S(h,t)} = \alpha_{0k} + \alpha_{1k} \mu_h + \varepsilon_{S,h,t}, \quad \varepsilon_{S,h,t} \sim \mathcal{N}(0,1)
\]

\[
U_{R(h,t)} = \beta_{0k} + \beta_{1k} \mu_h + \varepsilon_{R,h,t}, \quad \varepsilon_{R,h,t} \sim \mathcal{N}(0,1)
\]

Note, in particular, that the implied selection and response probabilities in equations (7)-(8) vary across households \(h\), but do not vary over time \(t\). This model is useful for two reasons. First, it allows us to judge whether the flexibility provided by Dynamic Model leads to better forecasting performance. Second, it serves as a contrast to the RFM approach, which implicitly assumes that a household’s response propensity varies continuously over time.
Traditional RFM Model

To benchmark the latent trait model, we consider a standard RFM probit model. This uncorrected RFM probit is the model specification which is typically used by industry consultants.

The definitions of the RFM variables used in this research follow standard industry practice (David Sheppard Associates 1999, Hughes 1996). Recency is defined as the number of days since the last donation received by the firm. Frequency is defined as the total number of contributions made in the past up to the solicitation date. Monetary value is defined as the cumulative amount of contributions (in dollars) that the household spent previous to the solicitation date. Since the correlation between the frequency and monetary variables in our data is .95, only the recency and frequency variables are used in the RFM model. For this reason, model coefficients for frequency must be understood to incorporate the impact of monetary value as well.

The traditional RFM model is a binary probit system that forecasts the probability of household response. Formally, we write the utility of household response as

$$U_{R(h,t)} = \beta_0k + \beta_{1k}[\text{Rec}]_{ht} + \beta_{2k}[\text{Freq}]_{ht} + \varepsilon_{R,h,t}$$

where the error $\varepsilon_{R,h,t}$ has a normal distribution with means equal to zero, variance equal to one. The subscript $k$ in this model denotes the type of solicitation sent to the household: Type A, Type Non-A, and Type Unknown. Here, Rec denotes recency and Freq denotes frequency. To connect this model with the observables in the donation dataset, we assume the probability that $U_{R(h,t)} > 0$ is the probability that binary variable $Y_{R(h,t)} = 1$ (household makes donation).
Assessing Forecast Accuracy

We assess model performance by predicting response in the holdout dataset. Since the response behavior is not observed when a household is not selected, prediction is based on the response of the selected households only. To ensure comparability across models, we use the MAD (Mean Absolute Deviation) statistic and the overall accuracy (hit rate for both purchases and non-purchases). To take account of estimation uncertainty, these measures are computed using 2000 simulated draws of parameters from the posterior distribution for the latent trait models and the RFM model estimated by MCMC. The mean of the 2000 prediction measures are reported for all models. This procedure enables the forecast measures to be tested for the statistical difference.

The holdout prediction statistics of the two latent trait models and the RFM model are presented in Table 1. Overall, the latent trait models show performance better than the traditional RFM probit model in both MAD and the Overall Hit. The MAD of the Dynamic and Long-Run Models are not statistically different but the Overall Hit of the Dynamic Model is statistically higher than the Long-Run Model.

Latent Trait Model Coefficients

The pattern of coefficients for Dynamic latent trait model (Table 2) tells an interesting story. Beginning with demographic effects, it is important to note that 6 of the 9 principal component variables capturing demographic effects are statistically insignificant. This indicates that demographics are generally a poor guide to the long-run propensity to respond \( \mu_h \) trait of households. In line with the extensive marketing science
literature on consumer heterogeneity (see Allenby and Rossi (1999) for a review), most of the differences in long-run household buying behavior are unrelated to the set of demographics available for analysis.

The pattern of results among the solicitation types is quite informative. Consider the differences between solicitation Type A (the routine mailings) and solicitation Type Non-A (special mailings, often seasonal). (We do not discuss the Type Unknown solicitations because they are some unknown mixture of all solicitation types.) Note that the selection rules are quite different. Type A is sent to households who are better long-run prospects, but who have not donated recently. In contrast, Type Non-A mailings basically ignore long-run response, instead emphasizing households that have donated recently. Turning to the response coefficients, it is clear that Type A solicitations generate a much stronger long-run response than Type Non-A solicitations. In contrast, the significantly negative coefficient on lagged response for Type Non-A solicitations indicates that households who recently donated are unlikely to donate again.

Graphs of the selection and response curves, shown in Figure 1, reinforce these general points. The horizontal axis, displaying the long-run propensity to respond trait, is restricted to the range of $\mu_h$ values characterizing the households in our dataset. The overall impression conveyed by Figure 1 is that the decision to donate to this charitable organization is closely tied to the value of the propensity to respond measure. This, of course, is how the model is constructed. However, Figure 1 also shows that the decision to mail a solicitation to a household depends very weakly on this trait. Managers,
instead, seem to base mailing policy primarily upon whether the household has donated in the recent past.

Substantively, these results suggest Type A solicitations work better for this charity because managers wait a period of time after a donation before sending out a new solicitation and attempt to target households with long-run interests in the charity. In contrast, the Type Non-A solicitations are mailed out in a more opportunistic fashion, relying more on short-run response. It is possible that the firm views Type A solicitations more in terms of household retention, and the Type Non-A solicitations more in terms of household acquisition. Nevertheless, given the response pattern for the Type Non-A solicitations, the current mailing policy for Type Non-A (send mailings to those who have recently responded) is clearly counterproductive. This observation, along with the fact that the mailing policy of the routine Type A solicitations is weakly linked to the household propensity trait, strongly argues that the firm’s mailing policy could be improved by using the estimated propensity to respond as a guideline for mailing.

**Summary**

RFM variables are best regarded as behavioral indicators of an underlying interest in the firm’s product or service. In this case, the product is the cause of the charitable organization. Although RFM variables provide some information on a household’s propensity to respond trait, RFM variables are also impacted by the mailing policy selected by managers. The latent trait approach is superior in the sense that it separates the household trait (the decision to respond to a solicitation) from behavioral responses to
the trait (the manager’s decision to contact a household). Moreover, this trait is a stable characteristic of the household which cannot be affected by the firm.

CONCLUSION

This study develops a general procedure for estimating the response probabilities of households in database marketing. The proposed approach, based upon a simultaneous selection-response formulation, assumes that each household has a latent propensity to respond that impacts both the firm’s decision to mail a solicitation and the household’s decision to respond to a solicitation. Inclusion of the selection decision of the firm in the model recognizes the potential for selection bias; the propensity to respond construct solves the problem of endogeneity of RFM. Our empirical analysis showed that the Dynamic Model yielded the best forecasting results. The latent trait model generates exogenous measures of the long-run propensity to respond of each household and provides the researcher with a tool to understand the effectiveness of current household solicitation policy.

Contributions of Research

Although recency, frequency and monetary value are intuitively reasonable ways of measuring the attractiveness of households, constructing a predictive model using RFM variables is problematic. The underlying problem is that the RFM variables in a database are functions of both the household’s interest in the product category and the firm’s mailing policy. That is, a household’s RFM profile depends upon characteristics of the both the household and the firm. From a statistical point of view, this confounding of household behavior and firm decision behavior is particularly worrisome.
The proposed model, by explicitly considering the rule of household selection rules in generating the dataset, calibrates a household response model that can be generalized to future datasets. This generalizability is due to the fact that our model is free from the selection bias and RFM endogeneity problems that impact most conventional methodologies. Using the response coefficients from the model output along with knowledge of the household trait value and lagged response behavior $Y_{R(h,t-1)}$, a researcher can predict response behavior in a future scenario in which the firm’s decision rules have changed. A major strength of our model is its ability to recover the true response characteristics of households from a customer database, even when the firm has attempted to optimize the mailing of solicitations. In principle, this is not possible with the RFM approach because the observed RFM profile depends upon both past purchase behavior and whether or not managers selected the household for mailings.

**Limitations and Extensions**

Our model has several limitations, all of which provide avenues for future research. The current model predicts only the incidence of selection and response. The most obvious extension is the construction of a model that predicts both probability of response and the dollar donation (or expenditure). This could be accomplished by changing the response equation to a Tobit model (Davidson and MacKinnon 1993) in which the latent propensity to respond drives both incidence and dollar amount. The current model is also limited to the prediction of one response per household. Clearly, most catalog retailers sell a large array of products. These retailers often develop specialty catalogs emphasizing subsets of the product line, and target these catalogs to various segments in the customer database. By constructing a set of correlated latent
response variables for different product subsets, the model could be further generalized to consider the basket of purchases made by a household. Taken together, these generalizations would permit the analyst to develop a global choice model for use by a multiple-category catalog retailer.
Appendix

ESTIMATION OF DYNAMIC PROPENSITY TO RESPOND MODEL

The Dynamic Propensity to Respond model is formulated in a Hierarchical Bayesian fashion (Gelman et al. 1996) and estimated using Markov Chain Monte Carlo procedures (Robert and Casella 1999). Here, we sketch the procedure used to estimate this model. The Long-Run Propensity to Respond model is estimated in a similar fashion by simply deleting the lagged response variable from both selection and response equations.

Dynamic Response Model

The Dynamic Propensity to Respond model is formulated in the following manner. The household-specific long-run propensity to respond for each household h is considered to be an independent, random draw from the Normal distribution

\[ \mu_h \sim N( X_h \gamma, 1.0 ) \]  

where \( \mu_h \) is the long-run propensity to respond construct, \( X_h \) is a vector of household demographics, and \( \gamma \) is a vector of parameters. We assume, without loss of generality, that the precision of the Normal distribution (inverse of the variance) is equal to one.

At each time point t, the household has two potential observations: a binary variable \( Y_{S(h,t)} \) reporting whether household h was sent a solicitation (= 1) or not (= 0); and a binary variable \( Y_{R(h,t)} \) reporting whether household h made a donation (= 1) or not (= 0). The selection and response models are formulated as two independent utility models, conditional on the long-run propensity to respond construct \( \mu_h \) and upon whether
the household made a donation during the last solicitation \( Y_{R(h,t-1)} \). Formally, we write the utility of selection as

\[
U_{S(h,t)} = \alpha_0 + \alpha_1 \mu_h + \alpha_2 Y_{R(h,t-1)} + \varepsilon_{S_{h,t}}, \quad \varepsilon_{S_{h,t}} \sim N(0,1) \tag{A2}
\]

and the utility of response as

\[
U_{R(h,t)} = \beta_0 + \beta_1 \mu_h + \beta_2 Y_{R(h,t-1)} + \varepsilon_{R_{h,t}}, \quad \varepsilon_{R_{h,t}} \sim N(0,1) \tag{A3}
\]

for all time point \( t \) of each household \( h \). The two errors \( \varepsilon_{S_{h,t}} \) and \( \varepsilon_{R_{h,t}} \) are assumed independent both at time \( t \), and over all possible pairs of past and future time points. In our empirical work, we allow the parameters of (A2) and (A3) to depend on the type of solicitation sent by the firm. However, to simplify the exposition here, we ignore this feature of the model in the equations below.

We observe a mailing to a household (\( Y_{S(h,t)} = 1 \)) when the utility of selection is greater than zero, and no mailing (\( Y_{S(h,t)} = 0 \)) when the utility of selection is less than or equal to zero. In a similar fashion, we observe a donation (\( Y_{R(h,t)} = 1 \)) when the utility of response is greater than zero, and no donation (\( Y_{R(h,t)} = 0 \)) when the utility of response is less than or equal to zero. Equations (A2) and (A3) form a two-equation binary probit system in which selection and response variables are independent, conditional upon the values of \( \mu_h \) and \( Y_{R(h,t-1)} \).

Note that when \( Y_{S(h,t)} = 0 \) (no mailing to the household), then we must observe that \( Y_{R(h,t)} = 0 \) (no donation is made). That is, when the household is not sent a mailing, we observe no response, but do not know whether or not the household would have responded if given the opportunity. For this reason, it is necessary to regard the donation
response as missing whenever $Y_{S(h,t)} = 0$. Accordingly, in the development below, it is understood that equation (A3) is dropped from the model for all combinations of $h$ and $t$ for which $Y_{S(h,t)} = 0$.

**Prior Distributions**

The prior distributions for $\gamma$, $\alpha$ and $\beta$ are assumed to be normal. We choose relatively diffuse priors to allow the observed data to dominate the analysis. Specifically, we assume that $\gamma \sim N[0, .025 I(d)]$, $\alpha \sim N[0, .025 I(3)]$ and $\beta \sim N[0, .025 I(3)]$ where $d$ is the number of demographic variables and $I(z)$ denotes a (square) identity matrix of dimension $z$. Note that we are using the Bayesian convention of writing a normal distribution as $N(m, p)$ where $m$ is the mean and $p$ is the precision (the inverse of the variance).

**Full Conditional Distributions**

After constructing the posterior distribution for the model, we derive the full conditional distributions of the parameters. This leads to the following relations:

$$f[\mu | else] \sim N[mean(\mu), prec(\mu)] \quad (A4)$$

where $mean(\mu) = prec(\mu)^{-1} \{X \gamma + \Sigma_t(\alpha_1 U_S) - \alpha_0 \alpha_1 T - \Sigma_t(\alpha_1 \alpha_2 Y) + \Sigma_t(\beta_1 U_R) - \beta_0 \beta_1 T - \Sigma_t(\beta_1 \beta_2 Y)\}$ and $prec(\mu) = (1 + \alpha_1^2 T + \beta_1^2 T)^{*}I(H)$. Here, $mean(.)$ and $prec(.)$ are the mean and precision of a normal distribution, $X$ is $(H \times d)$ matrix of demographics, $U_S$ and $U_R$ are $(H \times T)$ matrices of utility of selection and utility of response, respectively, $Y$ is $(H \times T)$ matrix of lagged response, and $H$ is the number of households, and $T$ is the number of time periods.

$$f[\gamma | else] \sim N[mean(\gamma), prec(\gamma)] \quad (A5)$$

where $mean(\gamma) = prec(\gamma)^{-1} \{X^t \mu + 0_{(d \times 1)} (.025)\}$, $prec(\gamma) = X^t X + (.025) I(d)$, and $0_{(d \times 1)}$ is $(d \times 1)$ vector of zeros.

$$f[U_S | else] \sim \text{Truncated } N[\bar{U}_S, I(H^*T)] \quad (A6)$$
where the elements of \((H \times T)\) matrix of \(\overline{U}_S\) are obtained from the deterministic elements on the right-hand side of equation (A2). The notation Truncated N\((a,b)\) denotes a truncated normal with upper bound = 0 if \(Y_{S(h,t)} = 0\) and lower bound = 0 if \(Y_{S(h,t)} = 1\).

\[
f[U_R|\text{else}] \sim \text{Truncated N}[\overline{U}_R, I(H^*T)] \tag{A7}
\]

where the elements of \((H \times T)\) matrix of \(\overline{U}_R\) are obtained from the deterministic elements on the right-hand side of equation (A3). The notation Truncated N\((a,b)\) denotes a truncated normal with upper bound = 0 if \(Y_{R(h,t)} = 0\) and lower bound = 0 if \(Y_{R(h,t)} = 1\).

\[
f[\alpha|\text{else}] \sim \text{N}[\text{mean}(\alpha), \text{prec}(\alpha)] \tag{A8}
\]

where \(\text{mean}(\alpha) = \text{prec}(\alpha)^{-1} \{ \mu Y^* U_S^* + 0_{(3 \times 1)} (.025) \}\) and \(\text{prec}(\alpha) = \mu Y^* \mu Y^* + (.025)I(3)\). Here, \(\mu Y^*\) is \((H^*T \times 3)\) matrix that contains a \((H^*T \times 1)\) vector of ones, a vector of \(\mu (H \otimes T)\) and a \((H^*T \times 1)\) vector of \(Y_{R(h,t)}^*\). \(U_S^*\) is \(U_S (H \otimes T)\) vector. \(0_{(3 \times 1)}\) is \((3 \times 1)\) vector of zeroes.

\[
f[\beta|\text{else}] \sim \text{N}[\text{mean}(\beta), \text{prec}(\beta)] \tag{A9}
\]

where \(\text{mean}(\beta) = \text{prec}(\beta)^{-1} \{ \mu Y^* U_R^* + 0_{(3 \times 1)} (.025) \}\) and \(\text{prec}(\beta) = \mu Y^* \mu Y^* + (.025)I(3)\). Here, \(\mu Y^*\) is \((H^*T \times 3)\) matrix that contains a \((H^*T \times 1)\) vector of ones, a vector of \(\mu (H \otimes T)\) and a \((H^*T \times 1)\) vector of \(Y_{R(h,t-1)}^*\). \(U_R^*\) is \(U_R (H \otimes T)\) vector. \(0_{(3 \times 1)}\) is \((3 \times 1)\) vector of zeroes.

**Markov Chain Monte Carlo Algorithm**

To estimate the parameters of the model, we use the Gibbs Sampler (Gelfand and Smith 1990). Starting with the vector of long-run propensity to respond parameters \(\mu\) in equation (A4), we successively sample the parameters from each equation in turn (i.e., sample the parameters of equation (A4) through equation (A9) in order, then repeat the sequence). The stationary point of this Markov Chain contains the model parameters (Gelman et al. 1996).

For each of the models, we ran a chain of 20,000 simulates. We used the last 5,000 simulates to compute posterior means and variances. The convergence of this algorithm was checked using a procedure developed by Geweke (2001, 2003). In this approach, a
second simulation, which uses a different (non-standard) logic to draw the simulated values, is conducted. Because the underlying theory indicates that the initial and the new simulation constitute Markov chains with the same stationary point, the researcher can check convergence by verifying that the posterior means and variances of the two simulations are the same. The results reported in this article passed this stringent convergence test.
Table 1 – Accuracy of Holdout Data Forecasts

<table>
<thead>
<tr>
<th>Model</th>
<th>MAD</th>
<th>Overall Hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Model</td>
<td>31.30%</td>
<td>78.71%</td>
</tr>
<tr>
<td>Long-Run Model</td>
<td>31.34%</td>
<td>78.57%</td>
</tr>
<tr>
<td>RFM Probit Model</td>
<td>32.84%</td>
<td>75.76%</td>
</tr>
</tbody>
</table>

Notes:
1. MAD of the Latent Trait Models are significantly lower than the RFM Probit Model (p < .01).
2. MAD of the two Latent Trait Models are not significantly different (p < .01).
3. Overall Hit of the Latent Trait Models are significantly higher than the RFM Probit Model (p < .01).
4. Overall Hit of the Dynamic Model is significantly higher than the Long-Run Model (p < .01).
### Table 2 – Dynamic Propensity to Respond Model

#### DEMOGRAPHICS

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Coefficient</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-White</td>
<td>0.0192</td>
<td>0.0255</td>
</tr>
<tr>
<td>High income/education</td>
<td>-0.0679*</td>
<td>0.0301</td>
</tr>
<tr>
<td>White, Children under 18</td>
<td>0.1336*</td>
<td>0.0328</td>
</tr>
<tr>
<td>Black male, Non-Hispanic</td>
<td>-0.0957*</td>
<td>0.0399</td>
</tr>
<tr>
<td>Hispanic male</td>
<td>-0.0184</td>
<td>0.0421</td>
</tr>
<tr>
<td>Household Size, Age</td>
<td>-0.0366</td>
<td>0.0492</td>
</tr>
<tr>
<td>Children under 18, Small Household</td>
<td>-0.1061</td>
<td>0.0735</td>
</tr>
<tr>
<td>Low education/income</td>
<td>0.0892</td>
<td>0.1029</td>
</tr>
<tr>
<td>White and Black (non-ethnic)</td>
<td>-0.0423</td>
<td>0.1989</td>
</tr>
</tbody>
</table>

#### TYPE A SOLICITATION

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Selection</th>
<th>Std Dev</th>
<th>Response</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.6731*</td>
<td>0.0170</td>
<td>-0.2699*</td>
<td>0.0379</td>
</tr>
<tr>
<td>Long-Run Propensity to Respond</td>
<td>0.1077*</td>
<td>0.0212</td>
<td>0.7278*</td>
<td>0.0609</td>
</tr>
<tr>
<td>Lagged Response</td>
<td>-0.5159*</td>
<td>0.0523</td>
<td>0.1175</td>
<td>0.1424</td>
</tr>
</tbody>
</table>

#### TYPE NON-A SOLICITATION

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Selection</th>
<th>Std Dev</th>
<th>Response</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.5125*</td>
<td>0.0140</td>
<td>-1.0307*</td>
<td>0.0321</td>
</tr>
<tr>
<td>Long-Run Propensity to Respond</td>
<td>-0.0340*</td>
<td>0.0149</td>
<td>0.3618*</td>
<td>0.0533</td>
</tr>
<tr>
<td>Lagged Response</td>
<td>0.3959*</td>
<td>0.0489</td>
<td>-0.2449*</td>
<td>0.0989</td>
</tr>
</tbody>
</table>

#### UNKNOWN SOLICITATION TYPE

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Selection</th>
<th>Std Dev</th>
<th>Response</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.1087*</td>
<td>0.0195</td>
<td>-0.2051*</td>
<td>0.0345</td>
</tr>
<tr>
<td>Long-Run Propensity to Respond</td>
<td>0.1750*</td>
<td>0.0352</td>
<td>0.8583*</td>
<td>0.0556</td>
</tr>
<tr>
<td>Lagged Response</td>
<td>1.3628*</td>
<td>0.2328</td>
<td>-0.1429</td>
<td>0.1877</td>
</tr>
</tbody>
</table>

Note: Demographics are principal component variables derived from a postal code demographic dataset. Std Dev indicates posterior standard deviation of the corresponding coefficient distribution. Parameter estimates denoted by an asterisk (*) are more than 2 standard deviations away from zero.
Figure 1 – Selection and Response Functions for Donation Dataset
References


