

# Target Selection in Direct Marketing Based on Multiple Contacts

Kirstin M. Derenthal

Institute of Marketing, University of Münster, Am Stadtgraben 13-15, 48143 Münster, Germany, k.derenthal@uni-muenster.de

Edward C. Malthouse

Integrated Marketing Communications, Northwestern University, 1870 Campus Drive, Evanston, IL 60208 US, ecm@northwestern.edu

Direct marketing scoring models predict responses to some contact that will be made in the future, helping the organization decide which customers to target, and are implemented in most analytical CRM software. They are usually estimated from a single “proxy” contact from the past, for which responses have already been observed. This approach is risky because there could be differences between the proxy and future contact, and other exogenous factors could have changed. We propose averaging predictions from multiple scoring models and show analytically, under certain assumptions, that the expected squared difference between the true responses to the future contact and the predicted values from the averaged estimate is less than or equal to the expected squared difference from a single previous contact. The improvement of the aggregated estimate over the single model increases as (1) the variation in effect sizes across contacts increases, (2) the number of averaged contacts increases, and (3) the variance of the effects estimates increases. We incorporate the effects of external factors in our model by weighting the coefficients with a general linear model. Using data from a retail catalog company and a not-for-profit organization, we evaluate our model empirically by testing whether our assumptions hold, examine the extent of variation in slopes and predicted values across models build from various previous contacts, evaluate the amount of improvement over extant models in terms of prediction error and performance as measured by a gains table, and study how improvement depends on the number of averaged contacts.

*Key words:* Customer Relationship Management, direct marketing scoring models, model averaging, bagging

---

## 1. Introduction

Due to the emergence of market niches, the growing importance of customer orientation, the increasing costs of reaching business markets through sales forces and growing importance of measuring the effects of marketing expenditures, more and more companies use direct marketing as an alternative or supplement to traditional advertising. Since the cost per exposure is relatively high compared with traditional advertising media such as magazines, newspapers and TV, it is important to target customers who are likely to respond. Offering products to customers who are not interested can not only weaken the relationship between the customer and the company, it can also cost the company a lot of money. Predictive *scoring models* can help determine who should receive an upcoming marketing contact, such as a direct mailing piece, by estimating the value of a response from a prospective recipient of the contact. Such models can also be used in estimating credit scores and customer lifetime value. They are a core component in most analytical customer relationship management systems.

Scoring models are usually estimated by using data from a *single* previous contact, for example a similar catalog mailed one year before (Malthouse 2003). Catalog mailings and other direct-mail pieces are example *contacts*. We shall sometimes use the term *mailing* even though scoring models are used to predict responses to the more general class of contacts targeted at individuals. Some measure of response to the contact, such as gross demand, is regressed, possibly with a nonlinear model, on the information available at the time of the contact such as recency, frequency and

monetary value. Predictions from the regression model using current customer information provide the scores, and those with higher scores will receive the contact. We call this the *single-proxy-mailing* approach.

In the literature, many different functional forms are used to estimate a scoring model, but most of them follow the single-proxy-mailing approach. For example, Haughton and Oulabi (1997), Levin and Zahavi (2001), Deichmann et al. (2002) model response behavior with decision tree techniques. Linear regression is another common method (Malthouse 1999, 2002). Bult (1993) compare semiparametric versus parametric classification models, Bult and Wansbeek (1995) develop a profit-maximizing approach comparing the performance of CHAID against several parametric models (e.g., logistic regression). Levin and Zahavi (1998) compare logistic, linear, Tobit and two-stage regression. Zadrozny and Elkan (2001) use decision trees and a two-stage regression model to predict donation amounts. Other authors apply neural networks to select customers for a mailing campaign (Zahavi and Levin 1997a,b). Kumar et al. (1995) compare neural networks with logistic regression, Suh et al. (1999) combine these two methods with traditional RFM models to improve predictions. Again, all these models follow the single-proxy-mailing approach.

In the machine learning literature on predictive modeling, Breiman (1996) proposes drawing samples with replacement from the available data (*bootstrap* samples), estimating the same predictive model on each bootstrap sample, and then averaging the results. He calls the approach *bootstrap aggregating* or *bagging* and demonstrates both analytically and empirically that it can improve predictive accuracy. In marketing, Ha et al. (2005) apply bagging with neural networks and show it improves the stability and predictions of scoring models. Nevertheless, their approach is also based on a single-proxy mailing.

The single-proxy-mailing approach implicitly assumes that predictors of response to the future mailing have similar effects as those to the previous one. However, this may not be true because factors influencing customers' purchase behavior might have changed. For example, the merchandise featured in the catalogs could be different. If some of the independent variables are category specific, e.g., frequency for men's clothing, one might expect the number of previous men's clothing purchases to have a larger effect if a particular contact has more of this type of merchandise in it. A proxy catalog featuring trendy clothing may not provide good predictions for a catalog with more traditional styles. Furthermore, some customers may respond differently from last year because a competitor has changed its mailing policies. There might be other exogenous influences such as a recession that could affect predictive models. During a recession, income could have a stronger effect. In the case of not-for-profit organizations, one might expect variables to have different influences after a natural disaster such as a tsunami or hurricane because the urgency of the disaster disrupts normal donation patterns.

Given that any number of factors could be different between some previous contact and a future one, it might be better to average predictions from scoring models build from several previous marketing contacts instead of a single one. This paper examines whether such an average will give better predictions. We prove that the level of improvement depends on the variance of the regression coefficients, both across contacts and due to sampling variation, and the number of contacts used to build the aggregated model. We explicitly capture the effects of external factors by weighting the coefficients with a general linear model. Section 3 discusses the methods used to test the proposed model. Section 4 evaluates our model on data from a retail catalog company and a not-for-profit organization, illustrating how the new method could be used to find the customers/donors with higher expected sales/donations. The method is similar to bagging except that the predictions from multiple previous mailings are averaged rather than those from multiple bootstrap samples. This paper extends the analytical work in Breiman (1996) by (1) allowing covariates to systematically affect the mean across models and (2) deriving formulas explicitly showing how sampling variation and the number of mailings affect the improvement due to aggregation.

Rossi et al. (1996) evaluate the impact of using incremental amounts of purchasing data in predicting consumer's responses to marketing instruments such as coupons. Their results show that if the entire purchase history is used in model estimation, the predictive accuracy is higher than if only one purchase occasion is observed. Heilman et al. (2003) extend this paper to compute the improvement in predictive accuracy by increasing the number of data points in steps of one. However, both papers predict brand choice in a retail environment instead of predicting response, for example the purchase amount, to a marketing contact from a specific company. Furthermore, their approaches are not based on averaging across models, they do not prove why additional information improves predictions and under which conditions it is better to use more than one purchase occasion to build the model. We develop a probabilistic motivation for why averaging across multiple contact points could improve predictive accuracy in Section 2.

## 2. Theoretical Motivation

Assume that an organization makes periodic marketing contacts with its customer base designed to stimulate sales. For example, retail catalog companies send out catalogs and not-for-profit organizations send solicitations for donations. The organization has already made  $K$  contacts and observed responses. Let  $y_{ik}$  denote the sales from customer  $i = 1, \dots, n$  attributed to contact  $k = 1, \dots, K$ , and  $\mathbf{x}_{ik} = (x_{i1k}, \dots, x_{ipk})'$  be a  $p$ -vector of covariates known about customer  $i$  immediately before contact  $k$  (the first element,  $x_{i1k}$ , could be 1 for all  $i$  making the first slope coefficient in a linear model the intercept). The values of  $y_{ik}$  and  $\mathbf{x}_{ik}$  have already been observed.<sup>1</sup> The meanings of the variables are consistent over  $k$ , but the values may change. For example, if variable  $j = 2$  is the logarithm of recency, the value of  $x_{i2k}$  will change over  $k$ , since the customer's recency will change over time. Other covariates could include the log number of previous purchases (frequency), the log total amount of previous purchases (monetary value), interaction variables such as average order size (monetary value / frequency) and frequency and monetary by product category.

The goal of this analysis is to estimate  $y_{i0}$ , the sales attributed to customer  $i$  from some contact  $k = 0$ , which has not yet been made, based on what is known about the customer at the time immediately before the contact,  $\mathbf{x}_{i0}$ . For all  $k \in \{0, \dots, K\}$ , assume

$$y_{ik} = \mathbf{x}'_{ik} \mathbf{b}_k + e_{ik} \quad (1)$$

where  $\mathbf{b}_k$  is a  $p \times 1$  slopes vector with  $E(\mathbf{b}_k) = \beta$ ,  $V(\mathbf{b}_k) = \Sigma_b$  (the subscript  $b$  is a cue indicating that it is the covariance matrix of  $\mathbf{b}$ ), which is a positive definite covariance matrix,  $E(e_{ik}) = 0$ ,  $V(e_{ik}) = \sigma_e^2$ ,  $e_{ik}$  independent of  $\mathbf{b}_k$ , and  $e_{ik}$  independent across  $i$  and  $k$ .

After estimating  $y_{i0}$ , contact  $k = 0$  will be made with a subgroup of customers, usually those with the largest predicted values.<sup>2</sup> The current approach to estimating  $y_{i0}$  is to select a single proxy contact from the past  $k \in \{1, \dots, K\}$  for which the dependent variable has been observed, estimate slope vector  $\mathbf{b}_k$ , and predict  $y_{i0}$  with  $\mathbf{b}_k$ :

$$\hat{y}_0^k = \mathbf{x}'_0 \mathbf{b}_k.$$

This single-proxy-contact approach implicitly assumes that  $\mathbf{b}_0 = \mathbf{b}_k$ , but as we discussed in the introduction this may not be true. We show that if the effects of predictor variables vary stochastically over time, then more reliable estimates of  $y_{i0}$  are made from an average of estimates from previous contacts.

<sup>1</sup> More precisely, they have been observed for  $n_k \leq n$  customers, since not all  $n$  customers may have received contact  $k$ .

<sup>2</sup> Sometimes the effect on lifetime value is added to this before selecting the names, but this does not change the necessity of making accurate predictions of  $y_{i0}$ , which is the goal of this paper. Elsner et al. (2004) and Gnl and Hofstede (2006), for example, optimize mailing decisions over multiple periods and include the timing of campaigns.

We propose the following ‘‘aggregated’’ estimator as an alternative, which is the average of the  $K$  individual models:

$$\bar{\mathbf{b}} = \frac{1}{K} \sum_{k=1}^K \mathbf{b}_k \quad \text{and} \quad \hat{y}_0^a = \mathbf{x}'_0 \bar{\mathbf{b}} = \frac{1}{K} \sum_{k=1}^K \hat{y}_0^k. \quad (2)$$

The superscript  $a$  indicates aggregated. To streamline the motivation for why the aggregated estimate will outperform the single proxy contact, we use ‘‘population’’ slopes  $\mathbf{b}_k$  rather than some estimate computed from  $y_{ik}$  and  $\mathbf{x}_{ik}$ . The more complex case where estimates are used will be addressed in the appendix. Under the assumptions stated around equation (1),

$$\begin{aligned} E(\bar{\mathbf{b}}) &= \boldsymbol{\beta}, \quad E(\hat{y}_0^k) = E(\hat{y}_0^a) = \mathbf{x}'_0 \boldsymbol{\beta}, \quad V(\bar{\mathbf{b}}) = \boldsymbol{\Sigma}_b / K, \\ V(\hat{y}_0^k) &= \mathbf{x}'_0 \boldsymbol{\Sigma}_b \mathbf{x}_0, \quad \text{and} \quad V(\hat{y}_0^a) = (\mathbf{x}'_0 \boldsymbol{\Sigma}_b \mathbf{x}_0) / K. \end{aligned} \quad (3)$$

We can now quantify how much better the aggregated estimate should perform compared with one from a single proxy mailing, by comparing the expected squared residuals from the two models:

$$\begin{aligned} E(y_0 - \hat{y}_0^k)^2 - E(y_0 - \hat{y}_0^a)^2 &= E(y_0^2) - 2E(y_0 \hat{y}_0^k) + E(\hat{y}_0^{k2}) \\ &\quad - E(y_0^2) + 2E(y_0 \hat{y}_0^a) - E(\hat{y}_0^{a2}) \\ &= -2(\mathbf{x}'_0 \boldsymbol{\beta})^2 + E(\hat{y}_0^{k2}) + 2(\mathbf{x}'_0 \boldsymbol{\beta})^2 - E(\hat{y}_0^{a2}) \\ &= E(\hat{y}_0^{k2}) - E(\hat{y}_0^{a2}) \\ &= V(\hat{y}_0^k) + E(\hat{y}_0^k)^2 - V(\hat{y}_0^a) - E(\hat{y}_0^a)^2 \\ &= V(\hat{y}_0^k) - V(\hat{y}_0^a) \\ &= \mathbf{x}'_0 \boldsymbol{\Sigma}_b \mathbf{x}_0 - \mathbf{x}'_0 \frac{\boldsymbol{\Sigma}_b}{K} \mathbf{x}_0 \\ &= \left(1 - \frac{1}{K}\right) \mathbf{x}'_0 \boldsymbol{\Sigma}_b \mathbf{x}_0 \geq 0. \end{aligned} \quad (4)$$

The last equation is a quadratic form and is greater than 0 when  $K > 1$  and  $\boldsymbol{\Sigma}_b$  is positive definite. A symmetric matrix is positive definite if and only if its eigenvalues are all positive (Strang 1980, p. 250), which is true whenever the matrix is of full rank. The cross-product terms in the first line can be replaced by the product of individual expectations because of the independence of  $\mathbf{b}_k$  for all  $k$ .

Thus, the expected squared residuals from the aggregated model will not be larger than those from a single proxy mailing and the amount of the improvement increases with the size of  $K$  and the variability of the slopes ( $\boldsymbol{\Sigma}_b$ ) across mailings. The function  $(1 - 1/K)$  equals 0 when  $K = 1$ , indicating no improvement, and then approaches 1 as  $K \rightarrow \infty$ . This suggests that a large fraction of the improvement can be realized by averaging even a small number of previous scoring models, e.g., for  $K = 5$ ,  $1 - 1/5 = 80\%$  and for  $K = 10$ ,  $1 - 1/10 = 90\%$ . In the appendix, (12) gives a more general result for slopes estimated from observed data rather than ‘‘population’’ values, showing that the amount of improvement also depends on the variance of the slopes estimates: the more variable the estimates (i.e.,  $V(\hat{\boldsymbol{\beta}})$  for some estimate  $\hat{\boldsymbol{\beta}}$ ), the greater the improvement.

Our formulation of the problem in equation (1) is identical to a random (mixed) coefficient model where the slope for a particular mailing is a random variable with mean  $\boldsymbol{\beta}$  and variance  $\boldsymbol{\Sigma}_b$ . Instead of estimating separate regression models and averaging the resulting estimates as we have recommended here, one could alternatively estimate one large mixed model and use the estimates of the fixed-effects  $\boldsymbol{\beta}$ . Unfortunately, it is not computationally feasible, at the present time, to estimate mixed models using hundreds of thousands of cases and there does not seem to be a theoretical reason to prefer mixed-model estimates from substantailly smaller samples to separate estimates from giant samples. We also test this empirically in section 4.6 below.

The result in Equation (5) depends on several assumptions. We now discuss the effects of relaxing certain assumptions. First, the specification of the slopes in (1) is equivalent to the following model:

$$\mathbf{b}_k = \boldsymbol{\beta} + \boldsymbol{\delta}_k, \quad (6)$$

where  $\boldsymbol{\beta}$  is a systematic component, constant for all  $k$ , and  $\boldsymbol{\delta}_k$  is a random component with  $E(\boldsymbol{\delta}_k) = 0$  and  $V(\boldsymbol{\delta}_k) = \boldsymbol{\Sigma}_b$ . When this model holds,  $\bar{\mathbf{b}}$  is a reasonable estimator of  $\boldsymbol{\beta}$  (e.g., in the Appendix we show  $\bar{\mathbf{b}}$  is an unbiased estimate of  $\boldsymbol{\beta}$ ). Now suppose that  $\boldsymbol{\beta}$  is not constant across all contacts. To address this possibility, we allow for the systematic component to be a function of other observable factors, such as the month in which a contact occurs, characteristics of the contact itself, the presence of a recession, etc. If  $\mathbf{w}_k$  be a vector describing these factors for contact  $k$ , then

$$b_{jk} = \mathbf{w}'_k \boldsymbol{\alpha}_j + \delta_{jk}, \quad (7)$$

where  $\boldsymbol{\alpha}_j$  is a vector of fixed effects for covariate  $j$ . Notice that (6) is a special case of (7) if  $w_{1k} = 1$  and  $\alpha_j = 0$  for  $j \geq 2$ . One can perform the general linear hypothesis test to determine if  $\alpha_j = 0$  for all  $j \geq 2$  and thus the simple-average model in (6) suffices.

A second assumption is that the slope for a contact is independent of the slopes for other contacts. If the slopes are dependent across contacts, the variances in (3), the cross-product terms in the first line of (5), and thus the amount of improvement could be larger or smaller, depending on the signs of the covariances. We investigate the extent to which this assumption holds empirically.

A third assumption is that of a linear functional form in (1). If we were to relax this assumption and allow  $\hat{y}^k$  to be a function of  $\mathbf{x}$  from a more general class, say bounded and continuous, would there be any benefit to averaging  $\hat{y}^k$ ? (Breiman 1996, § 4.1, especially equation (4.2)) suggests that the averaged predictor will offer an improvement.<sup>3</sup> Thus, we have reason to believe that averaging will improve the stability of a large class of models.

### 3. Methods

This section discusses the methods used to test the analytical results derived earlier. We use two data sets from the retail catalog and not-for-profit industries. Each data set consists of a contact-history file, recording the contacts (catalogs or solicitations) mailed to each customer over an extended period of time, and a transaction file, recording all purchases or donations and the specific contact that generated the transaction. For a large number of contacts, the data will therefore tell us which customers received the contact, summaries of purchase history (e.g., RFM) at the time of the mailing (denoted by  $\mathbf{x}_{ik}$  earlier), and how much each recipient spent or donated, if anything ( $y_{ik}$ ).

We build a variety of models to test our approach varying in terms of the predictor variables and functional form. We consider two sets of predictor variables, one small and the other large. The small set of predictors includes only RFM, which typically explain a large fraction of the variation that can be explained from such transaction data. With these three variables, the effects of multicollinearity will be modest. The large set of predictor variables, labeled RFM+, will add other feature variables that are often used in such predictive models, including day of first purchase / donation, average order / donation amount, frequency and monetary value (FM) during the most recent year, FM from 1 year ago, FM from more than 2 years ago and three interaction terms

<sup>3</sup> An alternative justification for why averaging should improve nonlinear models is that bounded and continuous functions can be approximated by the linear functions in (1) by transforming the  $\mathbf{x}$  variables prior to estimation (Hastie et al. 2001, §5.1):  $\hat{y} = f(\mathbf{x}) = \sum_{m=1}^M b_m h_m(\mathbf{x})$ , where  $h_m : \mathcal{X}^p \rightarrow \mathcal{R}$  is a basis function. Some familiar basis functions for  $p = 1$  include Taylor series polynomials, orthogonal polynomials, and regression splines.

measuring frequency across years.<sup>4</sup> As we show, the predictions resulting from the expanded set of predictor variables are slightly improved over the RFM model, but the effects of multicollinearity are greater.

We test two functional forms for the models. Our derivations have assumed linear regression models, where the dependent variable is a linear function of the independent variables. The first functional form is linear as in (1). The dependent variable is the amount spent or donated, which usually has a bimodal distribution with one mode at 0 for non-responders and the other mode having a log-normal shape. We compute the logarithm of the dependent variable (plus 1 to avoid computing the logarithm of 0) prior to estimation to reduce heteroscedasticity and the influence of outliers. Likewise, the distributions of RFM are usually right skewed with outliers and we use log transformations to symmetrize the distributions and address problems with the relationship being nonlinear.

We test a second, nonlinear functional form to demonstrate, as expected, that our averaging approach applies more generally to nonlinear models. The second functional form is sometimes called a *two-step* model. As indicated above, the distribution of the dependent variable is usually bimodal. The two-step model first uses logistic regression to predict whether the customer will respond to the mailing, i.e.,  $y = 0$  versus  $y > 0$ . For the responders only, a linear regression model predicts the amount spent / donated. After both models have been estimated, the predictions from the two models are multiplied to give the final score.

More precisely, let  $\pi_{ik}$  be the probability that customer  $i$  responds to contact  $k$ . We estimate the following model with logistic regression

$$\log\left(\frac{\pi_{ik}}{1 - \pi_{ik}}\right) = \mathbf{x}'_{ik} \mathbf{c}_k,$$

where  $\mathbf{c}_k$  is a vector of slope coefficients. Using only observations where  $y_{ik} > 0$  we estimate a linear regression predicting  $\log(y_{ik} + 1)$  from the predictor variables producing slope vector  $\mathbf{b}_k$ . Final scores for a future customer are nonlinear functions of the covariates  $\mathbf{x}_0$  given by

$$\hat{y}_0^k = \frac{\exp(\mathbf{x}'_0 \mathbf{b}_k)}{1 + \exp(-\mathbf{x}'_0 \mathbf{c}_k)} \quad (8)$$

The exponent in the numerator converts the log-amount used as the dependent variable in the regression into raw amounts. The aggregate estimate from the two-step model is

$$\hat{y}_0^a = \frac{1}{K} \sum_{k=1}^K \hat{y}_0^k$$

We evaluate the models using two criteria: prediction error and performance measured by a gains table. Prediction error is the average of the squared errors (using a log scale) on the future mailing ( $k = 0$ ):

$$\mathbf{PE}_k = \mathbf{Avg}_i (y_{i0} - \hat{y}_{i0}^k)^2 = \mathbf{Avg}_i (y_{i0} - \mathbf{x}'_{i0} \mathbf{b}_k)^2.$$

The average of  $\mathbf{PE}_k$  over  $k$  is an overall measure of the prediction errors across single-proxy-contact models:

$$\overline{\mathbf{PE}} = \frac{1}{K} \sum_{k=1}^K \mathbf{PE}_k. \quad (9)$$

<sup>4</sup>The first interaction term is computed by multiplying frequency during the most recent year and frequency from 1 year ago, the second one is a binary variable taking the value 1 if a customer has ordered / donated both during the most recent year and 1 year ago, the third interaction term is a binary variable taking the value 1 if the customer has ordered / donated during the most recent year, 1 year ago and two years ago.

**Table 1** Descriptive statistics for the mailings.

	Base Period Mailings				Future Mailing	
	K	n	n Lower	Range of	Mail	n
		Minimum	Quartile	Mail Dates	Date	
Catalog	51	92,160	108,655	1/02–12/04	4/05	169,768
Renewal	36	527	164,996	7/02–6/05	12/05	256,836
Renewal	36	527	164,996	7/02–6/05	3/06	649,684
Thank You	36	7,525	14,941	7/02–6/05	12/05	43,010

The prediction error of the aggregate models is

$$\mathbf{PE}_a = \mathbf{Avg}_i (y_{i0} - \hat{y}_{i0}^a)^2 = \mathbf{Avg}_i (y_{i0} - \mathbf{x}'_{i0} \bar{\mathbf{b}})^2. \quad (10)$$

These are out-of-sample estimates of the prediction error because none of the data for model 0 were used in estimating any of the  $\mathbf{b}_k$ .

The second measure of model performance is the cumulative average amount at the second and fourth deciles of a gains table. Predicted values will ultimately be used to select which customers should receive contact 0. The manager's real objective is to maximize the total amount of profit generated by selecting a certain number of people. We take the second and fourth deciles as points of comparison because the second decile is a point where only the top 20% of customers are contacts, while the fourth decile requires separating middle-quality customers. More precisely, the cumulative average amount at the second decile for some set of predictions  $\hat{y}$  is

$$\mu_{0.2} = \mathbf{Avg}_{i \in \mathcal{I}_{0.2}} y_{i0} \quad (11)$$

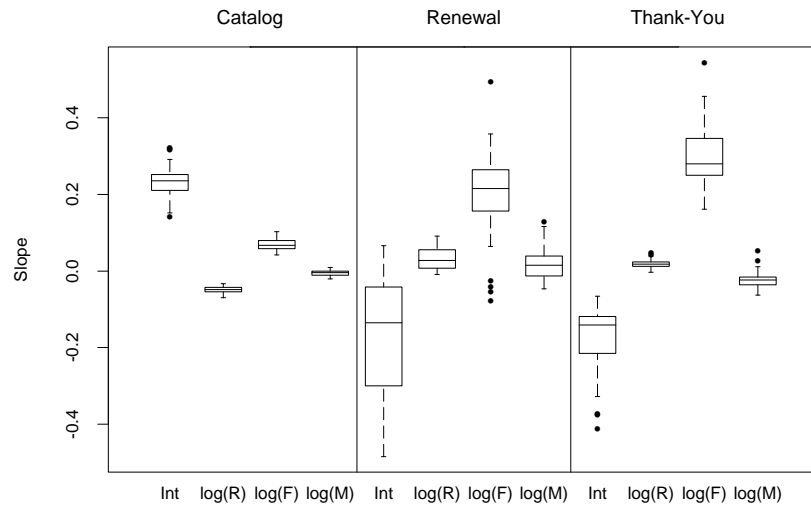
where  $\mathcal{I}_{0.2}$  is the set of all  $i$  such that  $\hat{y}_i$  is greater than the 20th percentile of  $\hat{y}$ . The cumulative average amount for the fourth decile is defined in a similar way.

## 4. Results

### 4.1. Data Source and Preliminary Analyses

The section describes the two data sets used in this empirical evaluation, summarizes the slope estimates and diagnostics, and discusses the implications of these analyses on the improvement in predictive accuracy based on the analytical results derived earlier. We begin with the data sets, summarized in Table 1. The first data set comes from a small European mail-order company selling through catalogs (first row). We have a complete record of everyone who received a catalog and their transaction history beginning in 2000. The company mailed 25 catalogs during 2003 and 26 in 2004. For each of these 25+26=51 mailings, we constructed a separate data set with predictor variables (e.g., RFM) taking their values as of the time of the mailing and a dependent variable indicating the amount spent in response to the mailing. We then estimated separate regression models for each of the data sets and averaged the slopes using (2). Very large sample sizes were available for estimating each of the models. The aggregate model and each of the 51 "single-proxy-mailing" models were then used to predict responses to a catalog sent during April, 2005. This target catalog was selected so that there would not be any overlap in the response windows between it and the 51 mailings in 2003/2004.

The second dataset comes from a US not-for-profit organization sending out solicitations to potential donors. Each month, the organization sends out two types of mailings. The first is a "renewal mailing." Those who make a donation are then sent a "thank-you" mailing acknowledging their contribution and asking for another one. Thus, the organization sends out 24 mailings per year. We know the contact and donation history from July, 2001 through June, 2006 and use a similar testing procedure as with the catalog company above. We constructed 36 data sets for

**Figure 1** Boxplots showing the slopes from RFM models across 51 catalog mailings (2 years) for the catalog company, 36 renewal and thank-you mailings (3 years) for the not-for-profit organization.**Table 2** Descriptive statistics for the slopes from RFM models across 51 catalog mailings (2 years) for the catalog company, 36 renewal and thank-you mailings (3 years) for the not-for-profit organization.

	Catalog		Renewal		Thank-You	
	Mean	Std Dev	Mean	Std	Mean	Std Dev
Intercept	0.2321	0.0400	-0.1665	0.1578	-0.1749	0.0898
Log(Recency)	-0.0495	0.0085	0.0328	0.0285	0.0194	0.0100
Log(Freq)	0.0693	0.0146	0.1979	0.1179	0.3005	0.0831
Log(Monetary)	-0.0053	0.0069	0.0176	0.0412	-0.0231	0.0218

each of the two mailing types for the months July, 2002 through June, 2005 with RFM as of the time of the mailing and the donation amount to the particular mailing. To ensure that the RFM values are comparable, we used only “recent” frequency and monetary occurring during the year prior to the mailing. As with the catalog data set, we estimated separate scoring models for each of the  $36 \times 2 = 72$  mailings. Since renewal and thank-you mailings seem to have fundamentally different modeling issues (e.g., all recipients of thank-you mailings have small values of recency), we estimated two separate aggregate estimators. We then attempted to predict responses to the December-2005 and March-2006 mailings. We selected these two future mailings because, as we show in our analysis, the holiday season (November and December) is quite different from the rest of the year, represented by March.

Before examining how well the models predict the future mailing, we first study the slope estimates in view of equations (5) and (12) to anticipate the improvement in performance. Recall that improvement in prediction error of the aggregate model over the single-proxy models depends on the variation in slopes both across contacts and estimates: the more variable the slopes, the greater the improvement. Figure 1 and Table 2 describe the distributions of slopes across the 51 catalog mailings, 36 renewal mailings, and 36 thank-you mailings. Across the predictor variables, the renewal mailing has the most variability in slopes across mailings and we therefore expect the aggregated estimates to offer more improvement for this group than for the others.

**Table 3** *F*-values from significance tests on the coefficients.

	Catalog		Renewal		Thank-You	
	Year	Month	Year	Month	Year	Month
Wilks' $\Lambda$	<b>12.7</b>	2.4	<b>6.4</b>	<b>3.7</b>	<b>6.2</b>	<b>3.0</b>
Intercept	3.1	3.3	<b>25.7</b>	<b>7.1</b>	0.1	<b>11.8</b>
Log(Recency)	0.0	3.7	<b>10.2</b>	<b>8.1</b>	<b>17.5</b>	<b>4.1</b>
Log(Freq)	<i>5.6</i>	2.5	<i>7.9</i>	<b>6.4</b>	<b>20.1</b>	<b>9.7</b>
Log(Monetary)	<i>4.5</i>	0.0	1.9	<b>9.4</b>	<b>15.6</b>	1.6

**Bold** values are significant at the 0.01 level and *italic* are significant at only the 0.05 level.

In equation (12), the variance of the estimated slopes, which increases with the amount of multicollinearity, also produces greater improvement. It is usually undesirable to have unstable estimates and multicollinearity. Bagging and our approach improve estimates that vary across samples, while those that are already stable offer less opportunity for improvement. To test the effects of multicollinearity, we use two sets of predictor variables for the catalog data set: RFM alone and RFM+, with FM this year, the year before, etc. The extent of multicollinearity can be summarized by condition indices<sup>5</sup> for each of the 51 models. The condition indices of the RFM models range from 4.6 to 4.9, while those for the RFM+ models range from 56.3 to 68.9. Condition indices under 10 indicate “no serious problem with multicollinearity” and those above  $\sqrt{1000} \approx 32$  indicate “severe multicollinearity” (Montgomery and Peck 1982, pp. 301–2). The improvement should be greater with RFM+ models than with RFM models.

Next, we test whether there is systematic variation in slopes across months or years with MANOVA models. Let  $w_{1k} = 1$  for all  $k$  (the intercept),  $w_{2k}$  equal the two-digit year number mailing  $k$ ,  $w_{3k} = 1$  if the month of mail  $k$  is January and 0 otherwise,  $\dots$ ,  $w_{13,k} = 1$  for November and 0 otherwise. We estimate the main-effect model

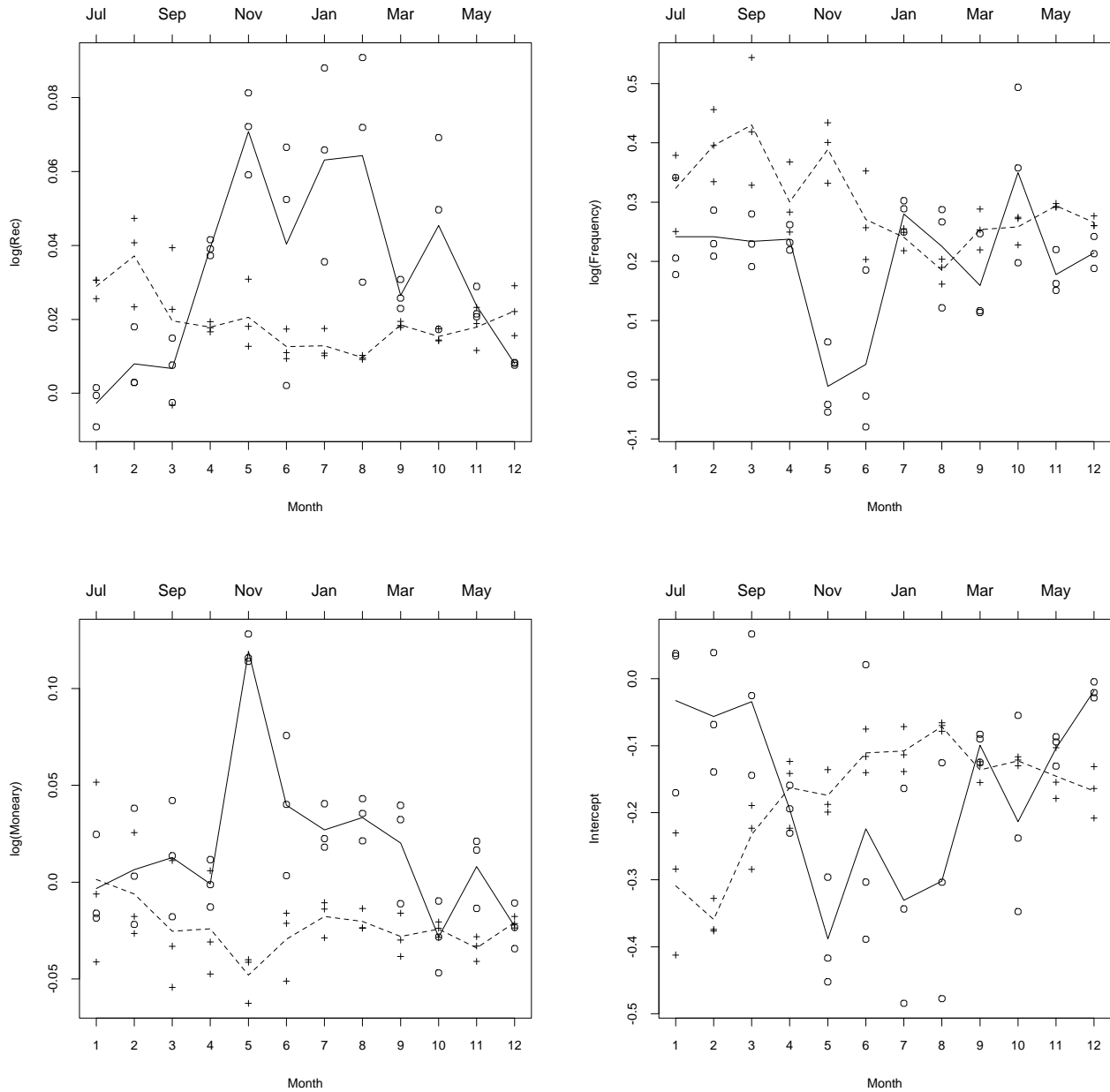
$$\mathbf{b}_k = \mathbf{w}'_k \boldsymbol{\alpha} + \boldsymbol{\delta},$$

where month is treated as a categorical variable (11 dummies) and year as a continuous variable (since there are only two years for the catalog company year is essentially a dummy variable). Table 3 gives results from significance tests. The first row gives the  $F$  values from a multivariate test of the null hypothesis of no overall effect for month or year and the bottom rows give the  $F$  values (computed from type III sums of squares) for univariate tests. For the catalog company, there are no significant differences between the years and there are some differences across months that are only marginally significant. These are indications that the accounting for month or year effects for the catalog will not produce any improvement.

There are many highly significant differences for the renewal and thank-you charity mailings both for month and year. This suggests that the accounting for systematic difference across months and years could improve the aggregated estimate for charity mailings. Figure 2 shows scatterplots of the slope estimates against month. The three plus signs for each month show the estimated slopes for the three years of the thank-you mailing and the circles show the slopes for the renewal mailing. The solid lines show the simple average values for the renewal mailing and the dashed lines show the average values for the thank-you mailing. The scatterplot to the upper right reveals that frequency has a substantially smaller effect on the renewal mailing during November and December than during the rest of the year, while the upper left scatterplot shows recency has a stronger effect in Winter. The lower left scatterplot shows that monetary value is a more important predictor in November, indicating that generous donors tend to donate even higher amounts when it comes

<sup>5</sup> The condition index is the square root of the largest eigenvalue divided by the smallest eigenvalue of the predictor-variable correlation matrix.

**Figure 2** Scatterplots of the estimated slopes for the not-for-profit organization. The solid lines and circles show aggregated and individual slopes for the renewal mailing, the dashed lines and plus signs show aggregated and individual slopes for the thank-you mailing.



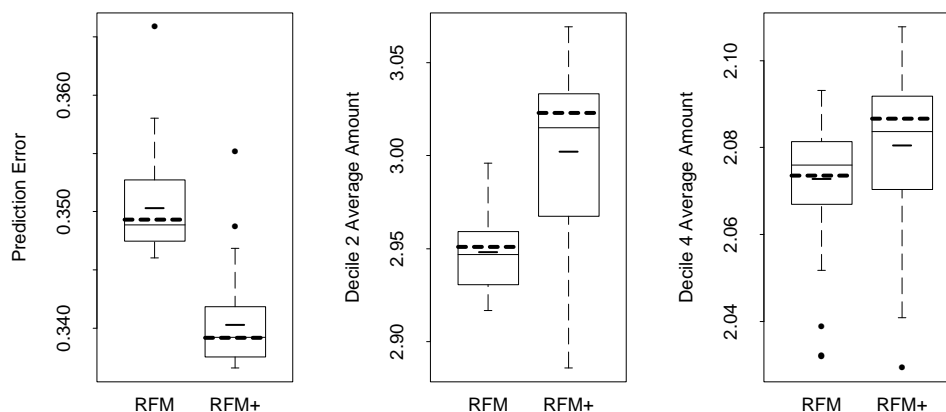
to holiday season. Accounting for difference in seasons should improve estimates for the charity mailing.

**4.2. Catalog Company**

The diagnostics in section 4.1 suggest only modest opportunity for improvement applying the aggregated estimate to the RFM models, but larger improvement for the multicollinear RFM+ models. We now test these predictions by using the single-proxy and aggregated models to estimate responses to a mailing in the “future,” sent in April, 2005. Measures of performance for the catalog

data are shown in Figure 3. The boxplots<sup>6</sup> show the distributions of the performance measures across the 51 single-proxy-mailing models.

**Figure 3** Boxplots showing predictions of catalog April,2005 catalog of RFM and RFM+ models across two previous years (51 mailings). The dashed lines show values for the aggregated estimates and the short solid line shows the average across all 51 mailings.



The left-most plot shows prediction error of the future mailing computed using equations (9) and (10). Smaller values of prediction error indicate better models. For the RFM model, note the following:

1. The line in the middle of the box shows the median prediction error and a short line has been added to the plot indicating the average prediction error across the 51 models. There is an outlier, indicating a very poor model. The distribution is slightly right skewed, and the average prediction error is slightly larger than the median.
2. The dashed line shows the prediction error for the aggregated model, using the estimator in (2). The prediction error for the aggregated estimator is slightly lower than the average prediction error across the 51 single-proxy-mailing models, but slightly worse than the median.
3. Although more than half of the single-proxy-mailing models have smaller prediction error than the aggregated estimator, in practice one does not know if a particular single-proxy model will be a good or bad one, when predicting responses to some mailing that has not yet been made. Using the aggregated estimator is therefore a less risky approach than the single-proxy approach. Recall that in practice averages from a gains table are multiplied by the size of the mailing, which could be millions. So, a difference of a few cents between models, after multiplying by millions, is a non-trivial amount.

The next two plots show performance in a gains table at the second and fourth deciles, as defined in (11). Here, higher amounts indicate better models. In each of the two graphs, the second boxplot shows the results for the RFM+ model having 14 predictor variables. Note the following about performance measured with gains tables and the RFM+ models:

<sup>6</sup> Boxplots are normally composed of a box with a line in the box, lines extending from the box (whiskers), and dots indicating outliers. The ends of the box show the lower and upper quartiles and the line in the box shows the median. Whiskers indicate the range of “most” of the distribution, i.e., all but the outliers. In addition, the dashed lines show the aggregated estimate value, the short line in the box shows the average, and the dotted line shows estimates from the GLM-aggregated model.

1. On all three criteria, the RFM+ models outperform the RFM models, indicating that the additional variables improve the explanatory power.

2. The aggregated estimator (dashed lines) using the RFM+ variables is better than the median (long solid line in box) and mean (short solid line) of the single-proxy models.

3. Across all performance criteria and models, some single-proxy models do better, but as indicated above, one does not know whether a single-proxy model is good or bad until after doing the future mailing. The aggregate estimate is less risky. This conclusion is true for both the RFM and RFM+ models, confirming that the amount of improvement in (5) does not depend on the number of predictor variables.

4. There is substantially more variation in gains-table performance (but not prediction error) across single-proxy models using the RFM+ variables compared with the RFM alone. This could be due to the higher degree of multicollinearity in the RFM+ data. At the same time, the improvement of the aggregated model over the average single-proxy model (dashed line versus the short solid line) is greater for the RFM+ models than for the RFM models. This is consistent with (12) — we observe results confirming that the amount of improvement also depends on the variability of the estimates. High-variance estimates are instable, but can be improved by averaging across multiple ones.

### 4.3. Not-For-Profit Organization

We now examine the amount of improvement for the not-for-profit organization. The diagnostics in section 4.1 suggested that we should see more improvement in the renewal mailings than the thank-you mailings since there is more variation across contacts (Figure 1). The systematic variation across months in Figure 2 suggests that the GLM estimate should outperform the simple average across the 36 mailings. The results for the not-for-profit renewal mailings are shown in Figure 4. All three models have only RFM independent variables. First we discuss the results of the linear models.

Note the following for the linear models (March and December):

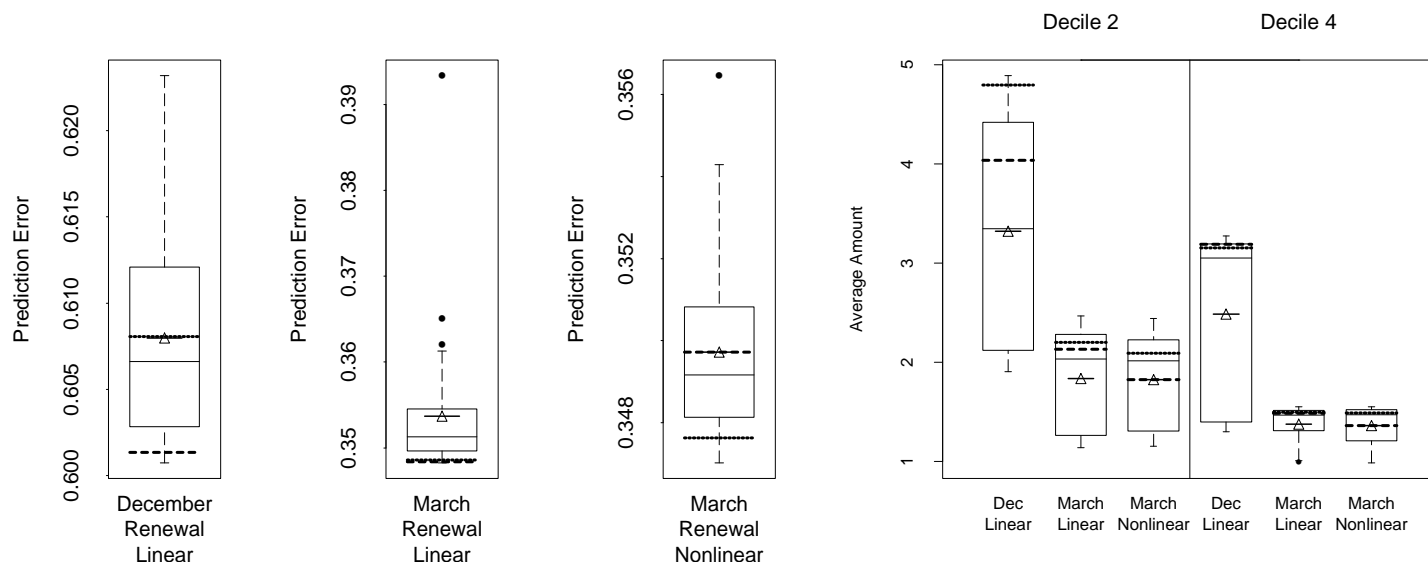
1. As with the catalog mailing, there is substantial variation on all three criteria (prediction error and average amounts at the second and fourth deciles) across single-proxy mailings, especially with the December mailing. Single-proxy models estimated from previous holiday mailings perform well on the December renewal mailing, but proxy mailings from other times of the year do not do so well.

2. The aggregated model (dashed line) is better than the mean and the median of the single models. Note that for the December mailing, the aggregated estimate does about 70 cents per donor better than the average single scoring model in the second and fourth decile. The prediction error is also substantially smaller. This large improvement is due the December slopes being substantially different from other months. Predicting a December mailing with one from another time of the year leads to poor predictions. This also explains why the improvement is smaller for the March mailing. March is a more “normal” month, so most of the mailings are good predictors for this mailing.

3. The GLM estimate (dotted line) improves the predictions of the December mailing by another 80 cents in the second decile over the simple average estimate. This can be explained by the simple average estimate giving equal weight to all months and, in the case of December, most other months are not good proxies.

To demonstrate that our new method also applies more generally to nonlinear models, we test the two-step model for the March renewal mailing and compare the aggregated estimate with the estimates from the single models. We do not run a GLM for the nonlinear model. The boxplots show that aggregation can also improve nonlinear models. The performance in the second decile of the average of the nonlinear models is substantially better than the median performance of a

**Figure 4** Boxplots showing distributions across single-proxy models for December, 2005 and March, 2006 renewal mailings using RFM. The dashed lines show values for the aggregated estimates, the short solid line with a triangle in the middle shows the average across all 36 mailings, and the dotted line shows estimates from the GLM estimates.



single-proxy model. There is variation across the single mailings and the aggregated scoring model outperforms the mean of the single mailings by about 25 cents in the second and 15 cents in the fourth decile.

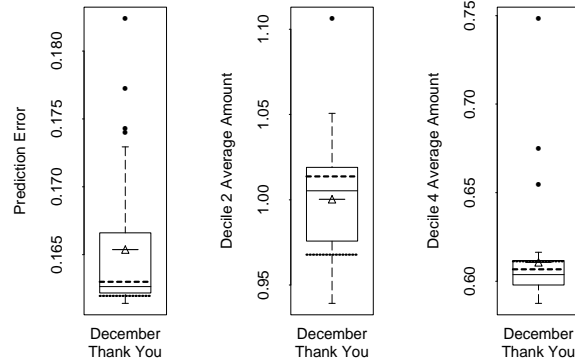
The results for the not-for-profit thank-you mailing are shown in Figure 5. The model only includes RFM independent variables. Compared to the renewal mailing, the variation of the thank-you mailing on all three criteria across single-proxy mailings is lower, as the diagnostic in section 4.1 predicted. The aggregated estimate (dashed line) outperforms the mean (short solid line) of the single-proxy mailings in terms of prediction error and in the second decile. The predictive accuracy of the GLM estimate (dotted line) is higher than of the aggregated estimate. However, in the second decile, the GLM estimate is outperformed by the majority of the single mailings. This can be tracked back to the results of the GLM model that show less significant month effects for the thank-you mailings. In the fourth decile, the GLM estimate beats the aggregate estimate.

Summarizing the results from the not-for-profit organization data, we conclude that the improvement from the aggregated model over the single-proxy models is large if the regression coefficients vary substantially. If there are any month effects, a GLM model does even better than the aggregated estimate. Furthermore, the results demonstrate that our method can improve nonlinear models as well.

#### 4.4. Number of Mailings

In the previous section, we evaluated our model by comparing the amount of improvement in terms of prediction error and performance for different levels of variation in slopes across different industries. We came to the conclusion that, on average, better predictions can be made by aggregating across multiple previous contacts. This leads to the question how many contacts are necessary to get a substantial improvement. Equations (5) and (12) suggest that the level of improvement increases with the number of catalogs according to the function  $(1 - 1/K)$ . The more previous scoring models we average, the larger the amount of improvement will be. We test this aspect of our

**Figure 5** Boxplots showing distributions across single-proxy models for March, 2006 thank-you mailing using RFM. The dashed lines show values for the aggregated estimates, the short solid line with a triangle in the middle shows the average across all 36 mailings, and the dotted line shows estimates from the GLM estimates.



derivations with the March renewal mailing of the not-for-profit organization, since there are fewer seasonal effects than the December mailing. We use different numbers of contacts to predict the March mailing, starting with the June mailing of the previous year. The  $K = 2$  aggregate estimate is the average across the June and May 2005 mailings, the  $K = 3$  estimate is across June, May and April 2005, etc. We average up to 12 previous mailings to predict the March, 2006 mailing.

Figure 6 presents the prediction errors and gains table amounts of the aggregated models depending on the number of contacts used to build them. The results support our theory since the prediction error decreases with the number of mailings. For both deciles, the cumulative average order amount increases with the number of contacts used for aggregation. The shapes of the curves are similar to the shape of the function  $(1 - 1/K)$  from (5), indicating that a large amount of the improvement can be realized by averaging even a small number of previous contacts. As the function  $(1 - 1/K)$  suggests, there is little improvement after about  $K = 6-8$  mailings ( $1 - 1/K$  approaches an asymptote of 1 as  $K$  increases).

#### 4.5. Sample Sizes

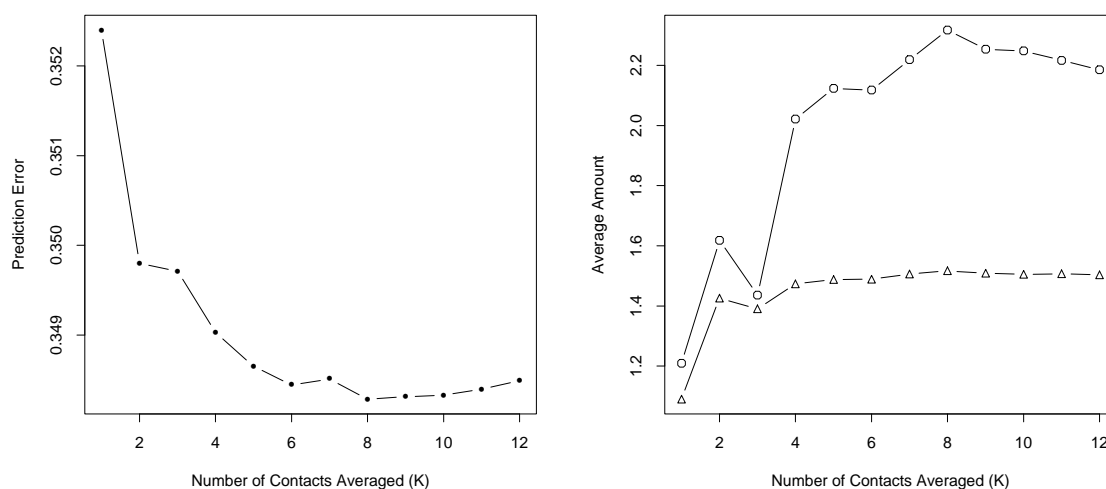
One could question whether the observed improvement in the aggregated method is simply due to estimating the coefficients with larger samples. Equation (12) shows that the improvement depends on three factors: the variability in slopes across mailings, the number of mailings averaged and the variability of the estimates. The third term is clearly a function of sample size, whose effect can be easily seen in the case of simple linear regression where the standard error of the predicted mean value of  $y$  for a given  $x_0$  is

$$\sqrt{\frac{S_e^2}{n} + S_b^2(x_0 - \bar{x})^2},$$

where  $S_e$  is the standard error of the estimate,  $S_b$  is the standard error of the slope (and involves an  $\sqrt{n-1}$  in its denominator), and  $\bar{x}$  is the mean of the  $x$  values. Thus, the accuracy of mean predictions improves at roughly the rate of  $1/\sqrt{n}$ . Table 1 shows that the lower quartile for the sample sizes of *individual* renewal models is 164,996, so most of the individual models are estimated with huge samples and have reliable estimates. The marginal improvement in reliability due to having samples of a couple hundred thousand versus many hundred thousand is small.

The above discussion based on our analytical results suggests that the improvement due to sample size should be small. We also test this empirically by drawing samples of 10,000 from each

**Figure 6** Prediction error and average donation amount in second (○) and fourth decile (△) for March, 2006 thank-you mailing depending on the number of aggregated mailings.



of 10 March renewal mailings, estimate 10 separate regression models and average their slopes. This averaged model is compared with single-proxy models estimated from random samples of 100,000 from the same 10 mailings. The performance of the aggregated model is compared with the performances of the single-proxy models in Figure 7. The averaged model, shown with the dashed line, outperforms nearly all of the single-proxy models on all three performance criteria. On average, the aggregated scoring model outperforms the mean of the single mailings by about 20 cents in the second and 15 cents in the fourth decile. Thus, the improvement in performance is not likely due to sample size alone.

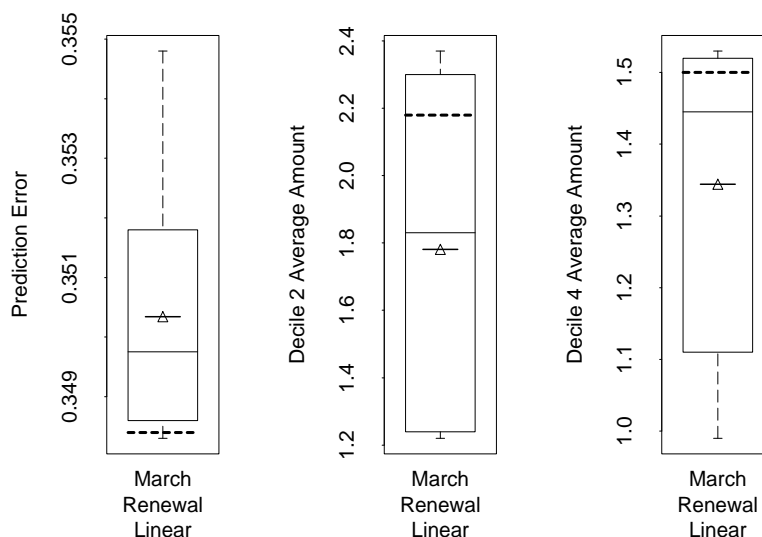
#### 4.6. Comparison with Random Coefficient Model

We indicated above that our formulation was a random coefficient model. Instead of estimating the slopes of individual mailings as random effects and then using the estimated fixed-effect mean in future predictions, we have elected to estimate separate regression models and average the results because this would allow us to use larger sample sizes. We also suggested that there is no obvious theoretical reason to prefer the random coefficient estimation over the average of separate models when such large samples are available. We test this assertion by drawing samples of 10,000 from each of 12 mailings. We build 12 separate regression models and average their regression coefficients and compare the performance of this model with the performance of a mixed model using all 120,000 cases. The results for the two models are almost identical. The prediction error of the simple average is 0.3647 compared with 0.3648 for the random coefficient model. The performances at the second and fourth deciles for the simple average are \$2.59 and \$1.70 compared with \$2.48 and \$1.70 for the random coefficient model. Thus, there is very little difference between the estimation methods.

### 5. Discussion

This paper proposes a new method of estimating direct marketing scoring models. Instead of using a single proxy contact from the past to estimate responses to a future contact, we recommend averaging the predictions from multiple models in the past. We analytically show that the expected squared difference between the true future response value and the aggregated estimate is smaller

**Figure 7** Boxplots showing predictions across 10 single-proxy models build from 100,000 observations for March, 2006 Renewal mailing using RFM. The short solid lines show the average across all 10 mailings. The dashed lines show values for an aggregated model built from 10 models each using 10,000 observations.



than the expected squared difference between the future response and a single proxy mailing. The amount of improvement depends on three factors: the number of previous contacts averaged, the variability in effect sizes across previous contacts, and the sampling variability of the estimators. This paper extends the bagging literature by suggesting averaging across replicates of the study (previous contacts) instead of bootstrap samples, extending the theoretical justification to include an explicit link to the number of averaged models and the variance of the estimates, and allowing the averaged estimator to be a function of contact-specific covariates such as seasonality.

These analytical results are tested with four mailings from two direct marketing organizations in different industries. The empirical results support the predictions from our derivations since, on average, the aggregated model outperforms the single models on both predictive accuracy and performance. We find more improvement for datasets that have a substantially large variation in slopes than for datasets with very stable regression coefficients. Even when the averaged model offers little improvement over the median single-proxy model, there is substantial variation across single-proxy models: some single-proxy models do better in predicting responses to the future mailing while others do worse. Unfortunately, the marketing manager has no way of knowing whether a particular single-proxy model is one of the good ones or bad ones until after the future contact has been made and the responses observed. Our averaged approach is therefore less risky.

Depending on the variability of the regression coefficients, the performance in the second decile is improved by \$0.02 to \$1.44 over single-proxy models. For one mailing sent by the not-for-profit organization, the single models generate \$3.32 on average in the second decile; the aggregated model generates \$4.03 (22% improvement); a GLM model including seasonal effects generates \$4.79 (44% improvement). If the company had a mailing depth of 20%, we would expect the GLM model to generate an additional \$75,000 revenues compared to a single model. The level of improvement diminishes as the mailing depth increases. Direct marketers commonly mail millions of pieces; with mailings of this magnitude, even small improvements translate into large amounts of money. Concerning the number of mailings, we found that a large fraction of the maximum

improvement can be realized by averaging even a small number of previous scoring models, e.g. 5 to 10 contacts/mailings. Although most of our derivations are based on linear regression models, we also tested our method on a nonlinear two-step model. On all performance criteria the aggregated estimator did better than the mean of the single-proxy models, suggesting that any of the extant scoring model approaches can be improved with averaging.

No paper is without limitations. We did not include classifiers predicting binary or nominal-level response variables. Although we focus on predicting continuous dependent variables, we are confident that our method would also work for classification models as well, because Breiman's paper shows that bagging improves accuracy for models predicting a class. Some future research topics include extending the results developed here to classifiers predicting nominal-level response variables. Furthermore, we did not take into account the profit function of the marketer because deciding who should be contacted after calculating the scores is a very complicated decision depending on many different factors such as long-term strategic goals, customer lifetime value, capacity considerations etc.(Malthouse (2003)). In Figure 3 we showed that the single-proxy models using the more multicollinear RFM+ variables produce less consistent gains-table performance values than RFM alone. While the effects of multicollinearity on more traditional measures such as the standard errors of the slopes have been thoroughly studied, little work has been done to study the effects on gains tables. Figure 2 showed strong seasonal effects in slopes. Frequency had smaller effects during the holiday season while recency was stronger. Future work should develop theoretical explanations for why the effects are different.

### Acknowledgments

We thank Manfred Krafft for helpful comments on our manuscript.

## Appendix

Section 2 derived the motivation for our aggregated estimator. To clarify the exposition, "population" values of the slopes were used in the derivations. In this appendix we derive properties of the OLS estimates rather than the population values. As in (5), we ultimately want to evaluate the improvement,  $E(y_0 - \hat{y}_0^k)^2 - E(y_0 - \hat{y}_0^a)^2$ , but before doing so we need to establish some additional notation and the properties of the OLS estimates. For each mailing  $k = 1, \dots, K$ , suppose that  $n_k$  values of  $\mathbf{x}_{ik}$  and  $y_{ik}$  have been observed with  $\mathbf{X}_k = (\mathbf{x}_{1k}, \dots, \mathbf{x}_{n_k k})'$  and  $\mathbf{y}_k = (y_{1k}, \dots, y_{n_k k})'$ . Assuming  $\mathbf{X}_k' \mathbf{X}_k$  is invertible for all  $k$  (otherwise the generalized inverse can be used below), the OLS estimate of  $\mathbf{b}_k$  is  $\hat{\mathbf{b}}_k = (\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k' \mathbf{y}_k$  and the aggregate estimator is

$$\bar{\mathbf{b}} = \frac{1}{K} \sum_{k=1}^K \hat{\mathbf{b}}_k.$$

Using formulas for the expectations of random vectors,

$$E(\hat{\mathbf{b}}_k) = E(\bar{\mathbf{b}}) = \boldsymbol{\beta},$$

so that both  $\hat{\mathbf{b}}_k$  and  $\bar{\mathbf{b}}$  are unbiased, in that their expected values are both equal to the mean slope across all mailings  $\boldsymbol{\beta}$ . Likewise  $E(\hat{y}_0^k) = E(\mathbf{x}'_0 \hat{\mathbf{b}}_k) = \mathbf{x}'_0 \boldsymbol{\beta}$  and  $E(\hat{y}_0^a) = E(\mathbf{x}'_0 \bar{\mathbf{b}}) = \mathbf{x}'_0 \boldsymbol{\beta}$ .

The variances can be computed in a similar way:

$$\begin{aligned} V(\hat{\mathbf{b}}_k) &= V((\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k' \mathbf{y}_k) \\ &= (\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k' V(\mathbf{X}_k \mathbf{b}_k + \mathbf{e}_k) \mathbf{X}_k (\mathbf{X}_k' \mathbf{X}_k)^{-1} \\ &= \boldsymbol{\Sigma}_b + \sigma_e^2 (\mathbf{X}_k' \mathbf{X}_k)^{-1} \end{aligned}$$

The first term is the variance in slopes across contacts and the second term is recognized as the usual variance of OLS estimates of a regression slope vector.

$$V(\bar{\mathbf{b}}) = V \left[ \frac{1}{K} \sum_{k=1}^K (\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k' \mathbf{y}_k \right] = \frac{\boldsymbol{\Sigma}_b}{K} + \frac{\sigma_e^2}{K^2} \sum_{k=1}^K (\mathbf{X}_k' \mathbf{X}_k)^{-1}.$$

If  $(\mathbf{X}'_k \mathbf{X}_k)^{-1}$  is constant across all  $k$ , i.e.,  $\mathbf{X}'_1 \mathbf{X}_1 = \dots = \mathbf{X}'_K \mathbf{X}_K = \mathbf{X}' \mathbf{X}$  for some  $\mathbf{X}' \mathbf{X}$ , then the previous expression simplifies further:

$$V(\bar{\mathbf{b}}) = \frac{1}{K} [\boldsymbol{\Sigma}_b + \sigma_e^2 (\mathbf{X}' \mathbf{X})^{-1}].$$

The variances of predicted values follow from the variances of the slopes, continuing to assume the common  $(\mathbf{X}' \mathbf{X})^{-1}$ :

$$V(\hat{y}_0^k) = \mathbf{x}'_0 [\boldsymbol{\Sigma}_b + \sigma_e^2 (\mathbf{X}' \mathbf{X})^{-1}] \mathbf{x}_0$$

$$V(\hat{y}_0^a) = \frac{1}{K} \mathbf{x}'_0 [\boldsymbol{\Sigma}_b + \sigma_e^2 (\mathbf{X}' \mathbf{X})^{-1}] \mathbf{x}_0.$$

The expected improvement in prediction error follows from (4):

$$E(y_0 - \hat{y}_0^k)^2 - E(y_0 - \hat{y}_0^a)^2 = V(\hat{y}_0^k) - V(\hat{y}_0^a) = \left(1 - \frac{1}{K}\right) \mathbf{x}'_0 [\boldsymbol{\Sigma}_b + \sigma_e^2 (\mathbf{X}' \mathbf{X})^{-1}] \mathbf{x}_0 \geq 0. \quad (12)$$

This expression is similar to (5), except that it includes the extra term  $\sigma_e^2 (\mathbf{X}' \mathbf{X})^{-1}$ . Improvement is greater when the quadratic form  $\sigma_e^2 \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0$  is large. Let us replace  $\mathbf{X}' \mathbf{X}$  with its Jordan decomposition

$$\sigma_e^2 \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0 = \sigma_e^2 \mathbf{x}'_0 (\boldsymbol{\Gamma} \boldsymbol{\Lambda} \boldsymbol{\Gamma}')^{-1} \mathbf{x}_0 = \sigma_e^2 \mathbf{x}'_0 \boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma}' \mathbf{x}_0 = \sigma_e^2 \mathbf{u}'_0 \boldsymbol{\Lambda}^{-1} \mathbf{u}_0,$$

where  $\boldsymbol{\Gamma}$  is a  $p \times p$  matrix of unit-length eigenvectors,  $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$  contains the eigenvalues, and  $\mathbf{u}_0 = \boldsymbol{\Gamma}' \mathbf{x}_0$  is a rotated version of  $\mathbf{x}_0$  having principal component coordinates. The quadratic form can become larger when some  $\lambda_j \approx 0$  or the principal coordinates  $\mathbf{u}_0$  are large.

## References

- Breiman, L. 1996. Bagging predictors. *Machine Learning* **24**(2) 123–140.
- Bult, J.R. 1993. Semiparametric versus parametric classification models: An application to direct marketing. *Marketing Sci.* **30**(3) 380–390.
- Bult, J.R., T. Wansbeek. 1995. Optimal selection for direct mail. *Marketing Sci.* **14**(4) 378–394.
- Deichmann, J., A. Eshghi, D. Houghton, S. Sayek, N. Teebagy. 2002. Application of multiple adaptive regression splines (mars) in direct response modeling. *J. of Interactive Marketing* **16**(4) 15–27.
- Elsner, R., M. Krafft, A. Huchzermeier. 2004. Optimizing rhenania's direct marketing business through dynamic multilevel modeling (dmlm) in a multicatalog-brand environment. *Marketing Sci.* **23**(2) 192–206.
- Gnl, F.F., F. Ter Hofstede. 2006. How to compute optimal catalog mailing decisions. *Marketing Sci.* **25**(1) 65–74.
- Ha, K., S. Cho, D. MacLachlan. 2005. Response models based on bagging neural networks. *J. of Interactive Marketing* **19**(1) 17–30.
- Hastie, T., R. Tibshirani, J. Friedman. 2001. *The Elements of Statistical Learning*. Springer, New York.
- Houghton, D., S. Oulabi. 1997. Direct marketing modeling with cart and chaid. *J. Direct Marketing* **11**(4) 42–52.
- Heilman, C.M., F. Kaefer, S.D. Ramenofsky. 2003. Determining the appropriate amount of data for classifying consumers for direct marketing purposes. *J. of Interactive Marketing* **17**(3) 5–28.
- Kumar, A., V.R. Rao, H. Soni. 1995. An empirical comparison of neural network and logistic regression models. *Marketing Letters* **6**(4) 251–263.
- Levin, N., J. Zahavi. 1998. Continuous predictive modeling — a comparative analysis. *J. of Interactive Marketing* **12**(2) 5–22.
- Levin, N., J. Zahavi. 2001. Predictive modeling using segmentation. *J. of Interactive Marketing* **15**(2) 2–22.
- Malthouse, E.C. 1999. Ridge regression and direct marketing scoring models. *J. of Interactive Marketing* **13**(4) 10–23.
- Malthouse, E.C. 2002. Performance-based variable selection for scoring models. *J. of Interactive Marketing* **16**(4) 37–50.

- Malthouse, E.C. 2003. Scoring models. D. Iacobucci, B. Calder, eds., *Kellogg on Integrated Marketing*. Wiley, New York, 162–188.
- Montgomery, D.C., E.A. Peck. 1982. *Introduction to Linear Regression Analysis*. Wiley, New York.
- Rossi, P.E., R. McCulloch, G. Allenby. 1996. The value of household information in target marketing. *Marketing Sci.* **15**(3) 321–340.
- Strang, G. 1980. *Linear Algebra and its Applications*. 2nd ed. Academic Press, Orlando.
- Suh, E.H., K.C. Noh, C.K. Suh. 1999. Customer list segmentation using the combined response model. *Expert Systems with Applications* **17** 89–97.
- Zadrozny, B., C. Elkan. 2001. Learning and making decisions when costs and probabilities are both unknown. Tech. rep., Department of Computer Science and Engineering, University of California, San Diego.
- Zahavi, J., N. Levin. 1997a. Applying neural computing to target marketing. *J. Direct Marketing* **11**(1) 5–22.
- Zahavi, J., N. Levin. 1997b. Issues and problems in applying neural computing to target marketing. *J. Direct Marketing* **11**(4) 63–75.