

A problem that occurs frequently in direct marketing is the prediction of the value of a discretizing function of a response variable. For example, to target consumers for a coupon mailing, a retail chain may want to predict whether a prospective customer's grocery expenditures exceed a predetermined threshold. The current approach to this prediction problem is to model the response variable and then apply the discretizing function to the predicted value of the response variable. In contrast with this "indirect" approach, the authors propose a "direct" approach in which they model discretized values of the response variable. They show theoretically that the direct approach achieves better predictive performance than the commonly used indirect approach when the response model is misspecified and the sample size is large. These two conditions are commonly met in direct marketing situations. This result is counterintuitive because the direct approach entails "throwing away" information. However, although both the discretized response data and the continuous data provide biased predictions when a misspecified model is used, the lower information content of the discretized variable helps the bias be smaller. The authors demonstrate the performance of the proposed approach in a simulation experiment, a retail targeting application, and a customer satisfaction application. The key managerial implication of the result is that the current practice of restricting attention to models based on the indirect approach may be suboptimal. Target marketers should expand the set of candidate models to include models based on the proposed direct approach.

A Direct Approach to Predicting Discretized Response in Target Marketing

A supermarket chain is expanding by opening a new store. To attract customers to try the new store, management wants to mail a targeted coupon to households that spend more than \$150 per month on grocery shopping. How should such households be identified? A commonly used approach is estimation of a model of grocery expenditures as a function of demographic characteristics of a customer sample from one of the chain's existing stores. The estimated model is then used to predict grocery expenditures of prospective households on the basis of a rented mailing list that includes demographic characteristics. The resulting pre-

dicted expenditures become the basis for selection of target households from the mailing list, namely, households for which forecast expenditures exceed the threshold of \$150 per month. Levin and Zahavi (1998) describe the use of this approach for several direct marketing applications (see also Malthouse 2002).

We propose a new, direct approach to this targeting problem. In the sample data used for estimation, we convert grocery expenditures of households to a binary variable that takes a value of 1 if the expenditure exceeds the threshold of \$150 per month and a value of 0 otherwise. We then develop a model for the binary indicator variable rather than a model of expenditures. We directly apply the estimated model to the demographic characteristics of each prospective household in the mailing list to predict whether the household is a target (i.e., predict whether the household's expenditure indicator variable is one or zero).

In this article, we present a theory that describes conditions in which the direct approach should perform better than the current indirect approach in predicting which households to target. The primary condition is misspecifica-

*Anand Bodapati is Assistant Professor of Marketing, Anderson Graduate School of Management, University of California, Los Angeles (e-mail: bodapati@ucla.edu). Sachin Gupta is Associate Professor of Marketing, Johnson Graduate School of Management, Cornell University (e-mail: sg248@cornell.edu). The authors are grateful to Vikas Mittal and ACNielsen for data used in this article. They acknowledge helpful comments of the three anonymous *JMR* reviewers, Eric Eisenstein, Vrinda Kadiyali, Angela Lee, and Edward Malthouse. Both authors contributed equally and are listed in alphabetic order.

tion of the response model. We show that when the sample size is large and the response model is incorrectly specified, more accurate predictions can be obtained from a model of the binary indicator variable than from a model of the original expenditure data. This result is counterintuitive because converting a ratio-scaled grocery expenditure variable to a binary indicator variable implies “throwing away” information. However, the result occurs because the bias in predictions that is due to misspecification is smaller for the direct approach. In other words, the lower information content in the binary variable relative to the continuous expenditure variable helps the bias in predictions be smaller.

What are the implications of this result for marketers? Most models of real-world data are incorrectly specified as a result of factors such as omitted variables or incorrect functional form. Omitted variables are almost inevitable because of measurement difficulties or because parameter stability suffers when all relevant predictors are included in the model. Incorrect functional form may occur because of lack of knowledge or deliberate choice of a simple functional form for ease of estimation or interpretation. Consequently, our results imply that if large samples are available, researchers and practitioners who are interested in maximizing predictive performance should include the direct approach in the set of candidate models that they evaluate. To our knowledge, current practice limits consideration to the set of indirect approaches.

The results we present are relevant to a range of target marketing applications. A common characteristic of many target marketing situations is that the decision maker aims to direct one of a small set of discrete actions at each consumer in a population of interest. These models are known as “prospecting scoring models” when applied to new customer acquisition (e.g., Bult and Wansbeek 1995; Malthouse 2002). For example, to target a promotional offer, a book club that wants to attract new members may need to classify prospects into those who frequently purchase books from bookstores and those who do not. Typically, a statistical response model is calibrated on an estimation sample in which both the predictor variables and the response variable are observed. The estimated model is then applied to the prediction sample, in which only predictor variable values are available, to obtain predictions of the response variable (e.g., the expected number of book purchases from bookstores). The predictions are used to classify consumers into different categories on the basis of predetermined thresholds. The thresholds used in the classification scheme usually are based on managerial considerations, such as the profitability of each marketing action and the manager’s risk preference (we discuss this at length subsequently). Note that whereas the response model is conceived in terms of the distribution of the response variable conditional on predictor variable values, the prediction problem is to forecast whether the response variable meets a certain threshold (i.e., the value of a discretizing function of the response variable), not the value of the variable itself.

In the example of the book club, the statistical model can be cast in terms of y , the annual number of books bought at bookstores by a consumer with certain predictor characteristics. In contrast, the managerial problem is predicting for each consumer whether the number of books exceeds a pre-

determined threshold. In other words, the manager is interested in predicting the value of a discretizing function $d(y)$ of y rather than y itself, where the discretizing function is defined by $d(y) = 0$ if $y \in (0, y_{\text{threshold}})$ or $d(y) = 1$ if $y \in [y_{\text{threshold}}, \infty)$, where $y_{\text{threshold}}$ is a predetermined threshold number of books.

The dichotomy between the statistical model and the prediction problem leads to the two competing approaches to estimating the model parameters that we consider. The traditional indirect approach is to take as estimator the maximizer of the likelihood of y in the estimation sample. The proposed direct approach is to take as estimator the maximizer of the likelihood of $d(y)$. The question of interest is, Which of the two approaches leads to higher predictive accuracy in future cross-validation data? In terms of the notation we have introduced, the key result in this article can be stated as follows: If the response model is correctly specified, estimating the model for y and then reconstructing $d(y)$ leads to higher accuracy in the prediction of $d(y)$. However, if the response model is incorrectly specified, estimating the model for $d(y)$ to directly predict $d(y)$ leads to higher predictive accuracy in large samples.

We have organized the remainder of the article as follows: In the next section, we develop theoretical propositions pertaining to the asymptotic predictive fit of the two alternative estimators we have identified. We then report a simulation study that provides insights into conditions that favor performance of the direct approach. We also apply the theory to two real-world data sets. Finally, we discuss the empirical results and our conclusions.

STATISTICAL FRAMEWORK AND THEORY

The Model Family

We denote the real valued response variable as y and the vector of predictor variables as x , where the distribution of x is F_x . The predictor vectors in both the estimation sample D and the prediction sample D_p are taken as independent draws from F_x . The true conditional density of y given x is $\tau_{y|x}(y; x)$. The working model for the conditional density of y given x is $f(y; x, \beta)|_{\beta = \beta^*}$, where f is an assumed functional form, and the value of the parameter vector β^* is unknown. Therefore, the model family considered is the set of functions spanned by considering all permissible values of β . We denote this model family as P_f .

The Prediction Problem

The prediction problem we consider is that of predicting values of $d(y)$, where $d(y)$ is a discretizing, noninvertible function of the response variable y . By “discretizing,” we mean that the function $d(\cdot)$ maps contiguous values of y to the same function value. In other words, it breaks up the space of y -values into a countable (and usually small) set of intervals. We denote the q th interval as $I_q \equiv (l_q, U_q]$, and $d(y) = q$ if $y \in I_q$. Therefore, predicting the value of $d(y)$ from the predictor vector x requires prediction of the interval in which the y -value will fall.

To illustrate that the decision maker’s prediction problem is one of predicting values of a discretizing function of the response variable, we consider the following simplified example: The response variable y_i is the revenue from

prospective customer i , and m is the retail margin per dollar of revenue. The targeting decision that the decision maker faces is whether to incur an acquisition expenditure of c dollars on prospective customer i . It follows that the decision rule for the decision maker is to target prospect i if $y_i m \geq c$. A response model yields a prediction of y_i , \hat{y}_i , and the prediction error around this point estimate. The decision rule for an expected profit maximizing decision maker is to target prospect i if $E(y_i|\hat{y}_i)m \geq c$. Because the decision maker believes that the response model yields unbiased predictions, it follows that $E(y_i|\hat{y}_i) = \hat{y}_i$, leading to the decision rule: Target prospect i if $\hat{y}_i m \geq c$ or if $\hat{y}_i \geq c/m$, where c/m is the breakeven threshold expenditure.

The previous discussion assumes a risk-neutral decision maker who maximizes expected profits. We now show that if the decision maker is not risk neutral, a targeting threshold expenditure different from the breakeven threshold can be derived. The risk preference of the decision maker is represented by the utility function $u(\cdot)$, and the decision rule of a decision maker who maximizes expected utility is to target prospect i if $E_{y_i|\hat{y}_i} u(y_i m - c) > u(0)$. We can rewrite the left-hand side of this decision rule as $u[E(y_i|\hat{y}_i)m - c] + \delta$, where $\delta > 0$, $\delta = 0$, or $\delta < 0$ depending on whether $u(\cdot)$ is convex, linear, or concave. It follows that the decision rule becomes $\hat{y}_i \geq c/m + \{u^{-1}[u(0) - \delta]\}/m$, or $\hat{y}_i \geq (c/m + k)$, where $c/m + k$ is the targeting threshold expenditure.

It is possible that a decision maker may not be able or willing to specify in advance the quantity k , which depends on the utility function, and may rely on examination of the predicted distribution of y to determine the threshold. In this case, a model of the continuous response variable would need to be developed, on the basis of which the manager would determine the effective targeting threshold expenditure. Thereafter, the response variable can be discretized, and the direct approach can be implemented.

The prediction takes the form of a predictive density for $d(y)$. Given the value of the predictor vector x unseen in the estimation data, we need to offer the probability distribution for $d(y)$, that is, $p[d(y)|x]$. Therefore, if the estimate of β is $\hat{\beta}$, the predictive density is

$$(1) \quad p[d(y) = q|x, \hat{\beta}] = \int_{y \in I_q} f(y; x, \hat{\beta}) dy.$$

Given the estimate $\hat{\beta}$ and the prediction data set $D_p \equiv \{(x_n^p, y_n^p)\}_{n=1}^{N_p}$ (which contains $[x, y]$ pairs not used during the estimation process), our measure of the predictive accuracy of $\hat{\beta}$ is the predictive log-likelihood (PLL):

$$(2) \quad \text{PLL}(\hat{\beta}; D_p) \equiv \log \left[\prod_{(x_n, y_n) \in D_p} \int_{y \in I_d(y_n)} f(y_n; x_n, \hat{\beta}) dy \right].$$

The version of predictive log-likelihood we consider is the “plug-in” version, in which the PLL is considered at just the maximum likelihood point estimate. This version does not take into account the uncertainty or standard errors in the maximum likelihood estimate $\hat{\beta}$. A way to take into account the uncertainty is to consider integrated predictive log-likelihood (IPLL), which is constructed by computing the expected value of the PLL over the posterior density of β . Typically, the posterior density is asymptotically Gaussian

with mean $\hat{\beta}$ and variance I^{-1} , where I^{-1} is the covariance matrix for the standard errors (which we discuss in a subsequent section). We denote the Gaussian density function as $N(\beta; \hat{\beta}, I^{-1})$. We then write the IPLL formally as follows:

$$(3) \quad \text{IPLL}(\hat{\beta}; D_p) \equiv \int \log \left[\prod_{(x_n, y_n) \in D_p} \int_{y \in I_d(y_n)} f(y_n; x_n, \hat{\beta}) dy \right] N(\beta; \hat{\beta}, I^{-1}) d\beta.$$

We state the various theoretical results we have developed for the PLL; however, the results also apply to the IPLL in slightly modified form because for maximum likelihood estimates, we typically have $p\text{lim } \hat{\beta} = \beta^*$.

Estimation of β

Having decided on the model family P_f , the task is to identify the member of P_f that is most consistent with the observed data in the estimation sample $D \equiv \{(x_n, y_n)\}_{n=1}^N$. The estimation sample size is N . We denote the most consistent member by $f(y; x, \beta^*)$. We consider two competing approaches for estimating β^* . The first approach is to estimate β^* by maximizing the log-likelihood of y in the observed data. We denote this estimator as $\hat{\beta}_y$:

$$(4) \quad \hat{\beta}_y \equiv \arg \max_{\beta} \log \left[\prod_{(x_n, y_n) \in D} f(y_n; x_n, \beta) \right] \\ \equiv \arg \max_{\beta} \text{LLY}(\beta).$$

The second approach is to estimate β^* by maximizing the log-likelihood of $d(y)$ in the observed data. We denote this estimator as $\hat{\beta}_{d(y)}$:

$$(5) \quad \hat{\beta}_{d(y)} \equiv \arg \max_{\beta} \log \left[\prod_{(x_n, y_n) \in D} \int_{y \in I_d(y_n)} f(y_n; x_n, \beta) dy \right] \\ \equiv \arg \max_{\beta} \text{LLDY}(\beta).$$

In Equation 5, we maximize what is akin to the “grouped likelihood” (see, e.g., Lindsey 1997), where y -values that result in the same $d(y)$ fall into the same group. The question of interest is, Will $\hat{\beta}_y$ or $\hat{\beta}_{d(y)}$ perform better in terms of predictive accuracy? We attempt to provide insights into this choice by investigating the properties of the two estimators.

We show that for large enough N , (1) if $\tau_{y|x} \in P_f$ (i.e., the model is correctly specified), then relative to $\hat{\beta}_{d(y)}$, $\hat{\beta}_y$ has higher expected predictive accuracy, as measured by Equation 2, and thus $\hat{\beta}_y$ is preferred, and (2) if $\tau_{y|x} \notin P_f$ (i.e., the model is misspecified), then $\hat{\beta}_y$ has poorer expected predictive accuracy in general. Therefore, $\hat{\beta}_{d(y)}$ is preferred if large estimation samples are available. We develop these results subsequently. We use the notation > 0 to denote positive definiteness: $A > 0$ means that $v^T A v > 0 \forall v \neq 0$, and $A - B > 0$ means that the matrix difference between A and B is positive definite.

The Correct Specification Case: $\tau_{y|x} \in P_f$

We denote the expected Fisher information matrices at β^* as follows:

$$(6) \quad I_y \equiv N E_x \text{Var}_{y|x} \nabla_{\beta} \text{LLY}(\beta) |_{\beta = \beta^*}, \text{ and}$$

$$(7) \quad I_{d(y)} \equiv N E_x \text{Var}_{y|x} \nabla_{\beta} \text{LLDY}(\beta) |_{\beta = \beta^*}.$$

Lemma 1 (consistency and asymptotic normality): If $I_y > 0$ and $I_{d(y)} > 0$, then for large sample size N , $\hat{\beta}_y \sim N(\beta^*, I_y^{-1})$, and $\hat{\beta}_{d(y)} \sim N(\beta^*, I_{d(y)}^{-1})$.

The proof of this lemma follows from standard first-order asymptotic theory for the X -random case (see, e.g., Barndorff-Nielsen and Cox 1994; Gourieroux and Monfort 1995).

Lemma 2 (difference in precision): Under the assumptions of Lemma 1 and if $d(y)$ is not a sufficient statistic for y , then $I_{d(y)}^{-1} - I_y^{-1} > 0$, almost surely.

(For the proof, see the Appendix.)

Taken together, Lemmas 1 and 2 provide an important result: If the discretization of $d(y)$ is not so severe as to make the model unidentifiable, both of the competing estimation approaches recover the true parameter β^* . However, except under sufficiency of $d(y)$ for y , the estimator $\hat{\beta}_y$ will have lower variance (i.e., higher precision) than the estimator $\hat{\beta}_{d(y)}$.

Lemma 3 (difference in expected predictive accuracy under correct specification): Under the assumptions of Lemmas 1 and 2, $E[\text{PLL}(\hat{\beta}_y; D_p) - \text{PLL}(\hat{\beta}_{d(y)}; D_p)] > 0$.

(For the proof, see the Appendix.)

Lemma 3 offers a useful result for marketing practitioners: If the model is correctly specified in that the chosen model family P_f contains the true density, it is better to estimate on the basis of the raw responses y rather than on the discretized responses $d(y)$ even though the problem is prediction of only the discretized responses.

The Misspecification Case: $\tau_{y|x} \notin P_f$

We rely on asymptotic theory for maximum likelihood estimation when the model family is misspecified (e.g., Gallant 1997; White 1982) to obtain the primary implications for the estimation problems we consider herein. The basic result in the asymptotic theory under misspecification is that the maximum likelihood estimate converges to the parameter value that yields the lowest Kullback-Liebler distance between the model's likelihood and the true likelihood for the form of the data being analyzed. Therefore, the maximizer of the likelihood of y under P_f will converge to a point β_y^* , such that the likelihood of y at $\beta = \beta_y^*$ is closest (among all members of the family P_f) to the true likelihood of y (i.e., the likelihood of y under $\tau_{y|x}$). Correspondingly, the maximizer of the likelihood of $d(y)$ under P_f will converge to a point $\beta_{d(y)}^*$, such that the likelihood of $d(y)$ at $\beta = \beta_{d(y)}^*$ is closest (among all members of the family P_f) to the true likelihood of $d(y)$. Furthermore, the two estimates will be asymptotically normal, with variances given by a generalization of the inverse Fisher information matrices (precise statements on the asymptotic results are given in the Appendix). Notably, in the asymptotic theory for the misspecification case, in general, we have

$$(8) \quad \beta_y^* \neq \beta_{d(y)}^*.$$

In other words, as the sample size grows to infinity, $\hat{\beta}_y$ and $\hat{\beta}_{d(y)}$ converge to different values. An exception occurs again if $d(y)$ is sufficient for y . Note that this nonequality in the

misspecification case is in contrast to the case of correct specification, in which $\hat{\beta}_y$ and $\hat{\beta}_{d(y)}$ converge to the same value. The mathematical arguments behind this inequality are presented in the Appendix. We now present the most important result of this section:

Lemma 4 (difference in expected predictive accuracy under misspecification): If $\hat{\beta}_{d(y)}$ is fully identified in that it is the unique maximizer of the likelihood of $d(y)$ under P_f , and if $\beta_y^* \neq \beta_{d(y)}^*$, then for large N , $E[\text{PLL}(\hat{\beta}_{d(y)}; D_p) - \text{PLL}(\hat{\beta}_y; D_p)] > 0$.

(For the proof, see the Appendix.)

Lemma 4 should be of interest to marketing practitioners because it offers the following prescription: If the model is incorrectly specified in that the chosen model family P_f fails to contain the true density $\tau_{y|x}$, it is better to estimate β^* by means of the discretized responses $d(y)$ rather than the raw responses y , even though this approach entails operating on a reduction of the data.

Remarks on the Theory

We now restate the central results of the theory at an informal level. In Lemmas 3 and 4, we show that $\hat{\beta}_{d(y)}$ is worse than $\hat{\beta}_y$ for predictive accuracy in the correct specification case ($\tau_{y|x} \in P_f$) but better in the misspecification case ($\tau_{y|x} \notin P_f$). Why is there this reversal in expected prediction accuracies? Consider the misspecification case first. At the heart of Lemma 4 is the fact that, because there is no difference between fit in the estimation sample and fit in the prediction sample when only large samples are considered, $\hat{\beta}_{d(y)}$ is essentially constructed to be the maximizer of predictive fit among all elements in P_f . Furthermore, because $\beta_{d(y)}$ does not coincide with β_y^* , it will do better than $\hat{\beta}_y$. Neither $\hat{\beta}_{d(y)}$ nor $\hat{\beta}_y$ will give an unbiased reconstruction of the true density of $d(y)$ under $\tau_{y|x}$, but $\hat{\beta}_{d(y)}$ will give the less biased reconstruction. However, in the correct specification case, when $\tau_{y|x} \in P_f$, both $\hat{\beta}_{d(y)}$ and $\hat{\beta}_y$ give unbiased reconstructions of the true density. However, $\hat{\beta}_{d(y)}$ has higher variance and is therefore more unstable, which diminishes its predictive accuracy.

APPLICATIONS OF THE THEORY

We supplement the theoretical results of the previous section with additional insights from a simulation experiment. We also assess the practical significance of our results in finite samples by application to two real-world marketing data sets. The first application is for retail targeting, and the second application is the prediction of top box scores in customer satisfaction measurement. The second application illustrates a large set of ratings-scale measurement problems, such as purchase intention measurement, for which prediction of the top box score is of particular interest.

Simulation Experiment

Simulation experiments offer the advantage of researchers being able to control the various constructs that appear in the theoretical results. In this section, our goal is to understand how Lemma 4 will manifest itself in empirical estimates. Lemma 4 indicates that if the model is misspecified, for a large enough sample size, the direct approach will outperform the indirect approach in terms of PLL. For empirical applications, there are several notable questions. First, how large a misspecification is needed before a sizable superior performance for the direct method

is observed? Presumably, the lower the misspecification, the smaller is the performance advantage of the direct method over the indirect method. Second, how large a sample size is needed before the asymptotic qualification “large enough sample size” is observed to hold true? Third, what is the effect of skewness of the error in the generating model on the performance improvement? Last, what is the sensitivity of results to the choice of discretizing function?

Data-Generating Process

The response variable y is generated from the following process:

$$(9) \quad y = \beta_0 + \beta_1 x^2 + \varepsilon,$$

where $\beta_0 = 6$, $\beta_1 = 1$, and $x \sim N(\mu_x, 1)$. We manipulate the distribution of the error term as a factor in the simulation; in the base case, we consider $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. We also consider the case in which ε has a skewed distribution. The following discretizing function generates $d(y)$:

$$(10) \quad d(y) = \begin{cases} 1, & \text{if } y \in (-\infty, y_1] \\ 2, & \text{if } y \in (y_1, \infty] \end{cases}.$$

We also manipulate the discretizing function in the simulation; in the base case, we choose y_1 such that 50% of the density of y lies to the right of y_1 .

Estimating Model and Misspecification Measure

We considered the case in which it is assumed that the y -values come from the following linear regression model:

$$(11) \quad y = \beta_0 + \beta_1 x + \varepsilon.$$

This model is misspecified because the true model, given by Equation 9, is not nested in the assumed model form of Equation 11. In Equation 11, the likelihood for $d(y)$ is

$$(12) \quad p[d(y) = q|x, \beta] = \int_{y \in I_q} f(y; x, \beta) dy.$$

We specify the density of $d(y)$ as an ordered probit with x as the predictor variable. Note that misspecification arises in our simulation in two ways: (1) by the inclusion of x instead of x^2 as the predictor variable, and (2) when the density of ε in Equation 9 is not Gaussian, specification of Equation 11 as a probit model is incorrect.

To relate the extent of misspecification empirically with the extent of performance improvement obtained by means of the direct method, a measure of misspecification is needed that captures the extent of mismatch between Equations 9 and 11. As White (1994) discusses, a category of commonly used measures is based on the Kullback–Liebler distance between the true model and the assumed model. In

the case of two Gaussian regression models, the Kullback–Liebler distance between the two models corresponds closely to the difference between the residual variances in the two models. The greater the misspecification, the greater is the residual variance in the misspecified model relative to the residual variance in the true model. This, in turn, implies that the greater the misspecification, the greater is the ratio of R_c^2 to R_m^2 , where R_c^2 and R_m^2 are the R^2 values for the correct and the misspecified models, respectively. Accordingly, we consider the ratio of R_c^2 to R_m^2 as a factor (F2) in our simulation experiment.

Our measure of misspecification considers only the R^2 ratio of the two models. However, practitioners might judge the same R^2 ratio differently in various contexts, depending on the R^2 value for the correctly specified model. For example, a decrement in R^2 from .50 for the true model to .25 for the misspecified model may be considered a “worse” misspecification than a decrement from .01 to .005, though the ratio of R_c^2 to R_m^2 is 2 in both cases. Accordingly, we introduce R_c^2 , the R^2 value for the correct model, as factor F1 in our experiment.

Experiment Design

Table 1 describes the five factors in the simulation experiment. The factors F1 and F2 jointly characterize the extent of misspecification, from the viewpoint of practitioners and from the notion of Kullback–Liebler distance. We manipulated the two factors orthogonally. Note that $R_c^2/R_m^2 = 2$ indicates a high level of misspecification, whereas a value of 1.05 indicates modest misspecification. When $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, the values of R_c^2 and R_c^2/R_m^2 are altered by varying the values of the parameters σ_ε^2 and μ_x in Equation 9. For Equations 9 and 11, the following can be analytically derived:

$$(13) \quad R_c^2 = \frac{\beta^2(2 + 4\mu_x^2)}{\sigma_\varepsilon^2 + \beta^2(2 + 4\mu_x^2)}, \text{ and}$$

$$(14) \quad R_c^2/R_m^2 = 1 + \frac{1}{2\mu_x^2}.$$

(Derivations of these results are available from the authors.) These expressions can be inverted to produce the values of parameters σ_ε^2 and μ_x that yield the desired levels of R_c^2 and R_c^2/R_m^2 (see Table 2). When ε is chi-square distributed (see F4), we appropriately transform the variable to achieve the values of σ_ε^2 , specified in Table 2.

Because the full factorial of the five factors is large and computationally costly to explore, we choose a fractional factorial design that enables us to explore all the main effects of interest. Fully crossing all levels of F1, F2, and F3

Table 1
FACTORS AND LEVELS IN SIMULATION EXPERIMENT

	Factor	Levels
F1	R_c^2 , the fit of the correctly specified model	.05, .2, .35, .5, .65, and .8
F2	R_c^2/R_m^2 , deterioration in fit of assumed Equation 11 relative to the true Equation 9	1.05, 1.10, 1.2, 1.4, 1.6, 1.8, and 2
F3	Size of the estimation sample	100; 200; 500; 1000; 10,000; and 20,000
F4	Distribution of ε	$N(0, \sigma_\varepsilon^2)$, $\chi^2(1)$, $\chi^2(3)$, $\chi^2(7)$,
F5	Location of cutoff in $d(y)$	Density to the right of cutoff value is 25%, 50%, 75%

Table 2
VALUES OF σ_x^2 AND μ_x TO OBTAIN LEVELS OF F1 AND F2 IN SIMULATION EXPERIMENT

$F2: R_c^2/R_m^2$	μ_x	σ_x^2						
1.05	3.162	798.00	168.00	78.00	42.00	22.62	10.50	
1.10	2.236	418.00	88.00	40.86	22.00	11.85	5.50	
1.20	1.581	228.00	48.00	22.29	12.00	6.46	3.00	
1.40	1.118	133.00	28.00	13.00	7.00	3.77	1.75	
1.60	.913	101.33	21.33	9.90	5.33	2.87	1.33	
1.80	.791	85.50	18.00	8.36	4.50	2.42	1.13	
2.00	.707	76.00	16.00	7.43	4.00	2.15	1.00	
F1: R^2 of the correctly specified model (R_c^2)		.05	.20	.35	.50	.65	.80	

Notes: R_m^2 is the R^2 of the misspecified model.

at fixed levels of F4 (fixed at $\epsilon \sim N[0, \sigma_\epsilon^2]$) and F5 (fixed at 50%) yields 294 cells in the design. The three remaining levels of F4 (ϵ is distributed $\chi^2[1], \chi^2[3], \chi^2[7]$) are crossed with all levels of F1 and F3 and selected levels of F2, holding F5 fixed at 50%, which yields 324 cells. Similarly, the two remaining levels of F5 (25% and 75%) are crossed with all levels of F1 and F3 and selected levels of F2, holding F4 fixed ($\epsilon \sim N[0, \sigma_\epsilon^2]$), which results in 216 cells. Thus, we explore a total of 834 cells in the simulation. In each of the 834 cells, we generate 25 estimation data sets and 25 validation data sets. A validation data set is always of sample size 1000.

Results

We evaluate the performance of the direct approach relative to the indirect approach by using three measures of predictive performance: difference in (1) expected PLL, (2) expected correct classification rate, and (3) expected profits from the targeting scheme. Although our theoretical predictions pertain to PLL, Criteria 2 and 3 are of greater interest from a managerial perspective. The expected profits criterion requires explanation. Under the assumption that the underlying continuous variable y represents dollar expenditures of prospective customers, where m is the contribution margin per dollar of expenditure and c is the cost of targeting each consumer, we can compute the expected profits from both the direct and the indirect methods. We define the subset of consumers in the holdout sample who are predicted by the two approaches to spend more than the threshold level of expenditure as S_{Direct} and $S_{Indirect}$, respectively. When y_i is the observed expenditure of consumer i , the expected profit from the direct approach is

$$\sum_{i \in S_{Direct}} (y_i m - c)$$

and

$$\sum_{i \in S_{Indirect}} (y_i m - c)$$

from the indirect approach. Thus, the percentage difference in profits between the direct and indirect approaches can be expressed as follows:

$$\begin{aligned} (15) \quad \Delta EP &= \frac{\sum_{i \in S_{Direct}} (y_i m - c) - \sum_{i \in S_{Indirect}} (y_i m - c)}{\sum_{i \in S_{Indirect}} (y_i m - c)} \\ &= \frac{\sum_{i \in S_{Direct}} (y_i - c/m) - \sum_{i \in S_{Indirect}} (y_i - c/m)}{\sum_{i \in S_{Indirect}} (y_i - c/m)}, \end{aligned}$$

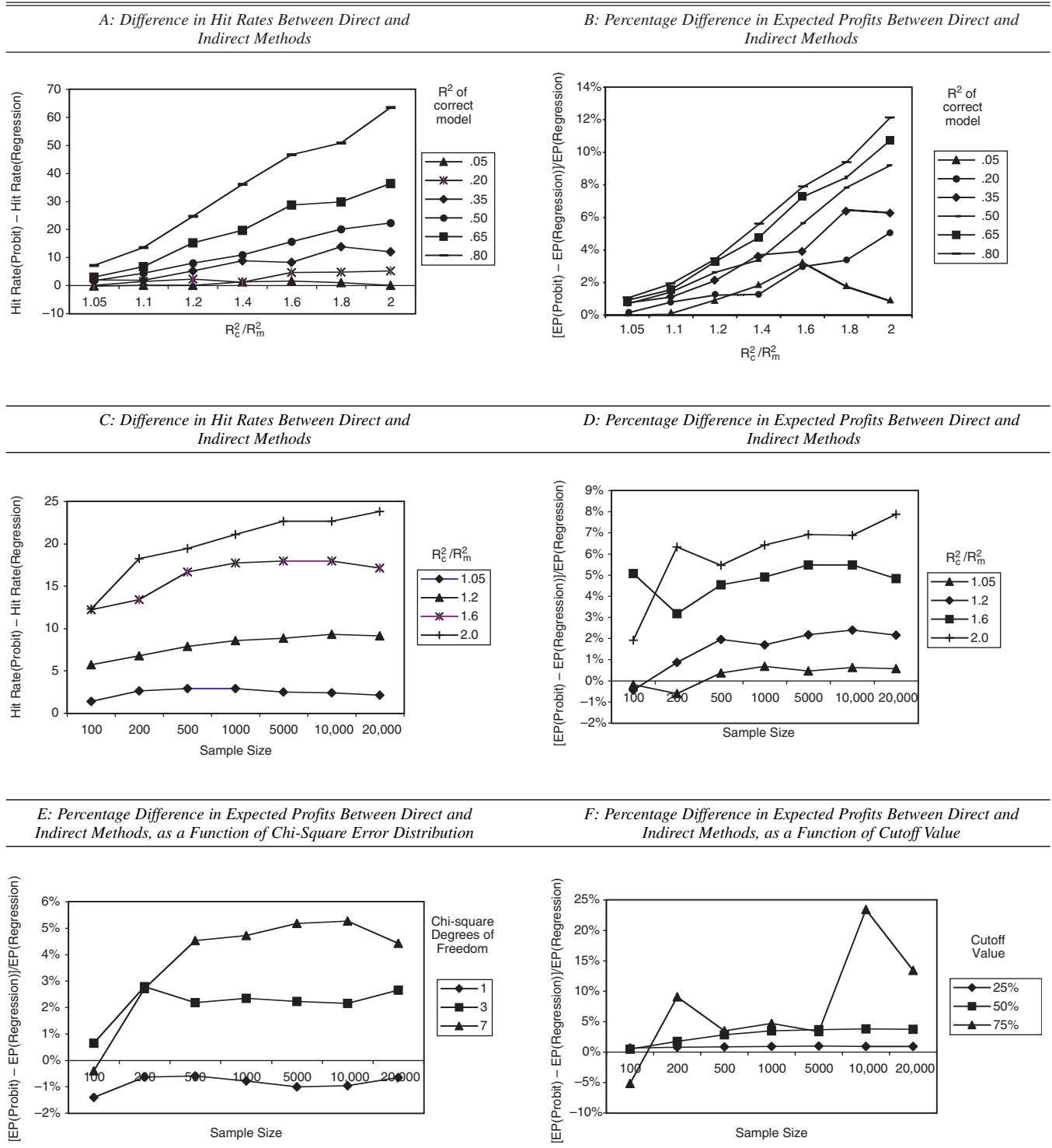
where c/m represents the threshold expenditure. Note that ΔEP depends not only on the ability of the model to slot consumers into the two groups accurately but also on the magnitude of the underlying continuous expenditure variable y .

In Panels A and B, respectively, of Figure 1, we show the mean difference in hit rates and the percentage difference in expected profits between the probit model and the regression model as a function of levels of misspecification, controlling for differences in R_c^2 . Larger values indicate superior performance of the probit. Panels A and B are based on average results for the two largest sample sizes: 10,000 and 20,000. Consistent with our intuitive expectation, the figures indicate that performance advantage of the direct approach increases with higher levels of misspecification. At low levels of misspecification (e.g., when $R_c^2/R_m^2 = 1.05$), the performance advantage is negative or modest. We also find that the performance advantage increases with fit of the correctly specified model.

In Panels C and D of Figure 1, we show the mean difference in hit rates and percentage difference in expected profits as a function of sample sizes for selected levels of misspecification. The results are averages over all levels of F1. Consistent with our expectation, the performance advantage of the direct approach improves with sample size. Notably, the indirect approach performs better in terms of expected profits only at sample sizes as small as 200, for mild levels of misspecification.

In Panel E of Figure 1, we show the results for the three levels of F4 in which the error term is assumed to be chi-square distributed. The coefficient of skewness of a chi-square variable is inversely related to the square root of the degrees of freedom. We find that the performance improve-

Figure 1
DIFFERENCE IN PERFORMANCE OF DIRECT AND INDIRECT METHODS IN SIMULATION EXPERIMENT



ment of the direct approach is lower when the skewness of the error term is high. Thus, for symmetric error distributions, such as the normal, the performance improvement is largest. In Panel F, we show the results for F5, the location of the cutoff value. We find that performance improvement of the direct approach increases with the cutoff value.

For marketing practitioners, our simulation results imply that conditions most favorable to the direct approach occur when a well-specified model fits well, but a misspecified model suffers large deterioration in fit. This might be the case in modeling situations in which the behavior of the response variable is well understood because of previous

experience with the phenomenon, but data on one or more important predictor variables are unavailable in a given application. Furthermore, symmetric distributions of the response variable and higher cutoff values favor the proposed method.

Retail Targeting

We return to the retail problem of identifying households that spend above-average amounts on grocery shopping. The retailer could offer a targeted incentive, such as a dollars-off coupon, to attract these desirable customers to try its new store. The targeting model is developed on expenditure and demographic data of customers in the retailer's current markets. To illustrate the prescriptions of the model, we used household panel data from Sioux Falls, S.Dak., and Springfield, Mo., collected by ACNielsen. For a sample of 7520 households in these markets, total grocery and drug store expenditures were available for a period of approximately 20 months. We used three key demographic characteristics of the households (annual household income, size of the household, and home ownership) to develop the targeting model. The three variables are often available in demographically scored mailing lists.

The model for predicting household expenditure, y , is a linear multiple regression model with the following predictor variables: household income, $\log(\text{household income})$, household size, $\log(\text{household size})$, and a dummy variable for home ownership. We estimated the model using ordinary least squares. We set the threshold expenditure level to determine whether to target a household at \$150 per month, which translates to \$3,000 for the period our data span. The threshold represents a breakeven expenditure, such that households that spend more than \$150 per month generate positive expected profits, and households that spend less generate negative expected profits. Assuming that Gaussian errors are in the expenditure model, the corresponding model for the discretizing variable $d(y)$ is then a binary probit with the same five variables as predictors. We estimated this model using maximum likelihood. Parameter estimates for both models on the full sample of 7520 households are shown in Table 3.

In the regression model, coefficients of all predictor variables have the expected signs and are significant, except for $\log(\text{income})$, which is insignificant. In the probit model, the linear effect of income is positive but not significant; however, the nonlinear effect of income is positive and significant. Furthermore, household size has a significant, negative coefficient, and $\log(\text{household size})$ has a significant, posi-

tive effect. These two coefficients jointly imply that the probability of expenditures greater than \$3,000 increases initially with household size and then diminishes, within the range of our data, which is a phenomenon we consider plausible.

Next, we examine predictive validity. We partition our full data into K blocks, such that each block has n/K observations, where n is the sample size in the full data set. For $k = 1, \dots, K$, we take the k th block as the prediction sample and pool the observations in the remaining $(K - 1)$ blocks to be the estimation sample. We present results for $K = 10$. We used the estimation sample to compute $\hat{\beta}_y$ and $\hat{\beta}_{d(y)}$ for that sample. We then computed the PLL and correct classification rate in the prediction sample for the estimates $\hat{\beta}_y$ and $\hat{\beta}_{d(y)}$ to determine which estimate was more consistent with the observed $d(y)$ in the prediction sample.

Table 4 shows that in each of the ten prediction samples, the PLL is higher for the probit model; the average difference is significant ($p < .001$) on the basis of a t-test. In nine of ten samples, the correct classification rate is higher for the probit model, and across the ten prediction samples, the correct classification rate of the probit model at 66.1% is marginally higher than the regression model's correct classification rate of 64.5%. We used McNemar's test (Siegal and Castellan 1998) to determine whether the number of discordant pairs (i.e., probit correct, regression incorrect and probit incorrect, regression correct) is equal. This test rejects the null hypothesis ($\chi^2 = 14.77, p < .001$) of equality. In Table 4, we also show ΔEP for each of the ten holdout samples; ΔEP is positive for each of the ten samples with an average value of 30%.

To examine model misspecification due to omission of relevant variables or incorrect functional form, we used regression specification error tests (RESET) (Ramsey 1969). The idea behind RESET tests is that when a relevant variable is omitted from the model, the error term of the incorrect model incorporates the influence of the omitted variable (Kennedy 1993). If some variables Z can be used as proxies for the omitted variables, a specification error test can be formed by examining Z 's relationship to the incorrect model's error term. The test has also been found to be powerful for detecting nonlinearities (Godfrey 1988). The RESET test adds Z to the set of regressors and then tests Z 's set of coefficient estimates against the zero vector by means of an F-test. Typically, Z comprises the squares, cubes, and fourth powers of the predicted dependent variable. When applied to the retail expenditure data, all three tests are highly significant ($p < .001$), indicating that the model for y is misspecified. The superior predictive performance of the

Table 3
RESULTS FOR RETAIL TARGETING PROBLEM

Coefficient	$\hat{\beta}_{d(y)}$		$\hat{\beta}_y$	
	Mean	Standard Error	Mean	Standard Error
Intercept	-.969	.110	727.86	229.69
Income (\$000)	.001	.002	13.66	3.77
$\log(\text{income})$.119	.049	123.68	100.44
Household size	-.186	.040	231.02	83.67
$\log(\text{household size})$.878	.107	814.19	225.96
Ownership of dwelling = 1	.581	.039	1241.64	84.14

Notes: The response variable y is household grocery and drugstore expenditures.

Table 4
PERFORMANCE OF DIRECT AND INDIRECT APPROACHES IN TEN CROSS-VALIDATION SAMPLES IN RETAIL TARGETING APPLICATION

Sample Number	Direct Approach			Indirect Approach			$\Delta EP\%$
	-PLL	CC Rate Percentage	Expected Profits (\$)	-PLL	CC Rate Percentage	Expected Profits (\$)	
1	482.8	65.0	309.6	490.2	63.3	227.6	36.0
2	471.4	68.1	345.8	482.5	65.4	237.0	45.9
3	470.5	67.3	317.6	481.6	65.0	221.5	43.4
4	474.0	66.4	455.6	478.6	64.6	387.7	17.5
5	482.9	65.0	243.8	491.2	65.8	215.4	13.2
6	484.5	63.8	307.3	492.2	62.2	219.2	40.2
7	477.5	66.8	270.8	483.6	64.4	199.1	36.0
8	480.6	65.6	414.4	487.0	64.5	330.1	25.5
9	484.2	65.3	275.9	494.7	63.9	194.8	41.6
10	468.9	67.6	392.0	479.3	66.2	331.7	18.2
	477.7	66.1	3332.9	486.1	64.5	2564.1	30.0

Notes: CC rate = correct classification rate; $\Delta EP\%$ = percentage difference in expected profits between direct and indirect methods; -PLL = negative predictive log-likelihood. The sample size of each estimation sample is 6768, and the sample size of each holdout sample is 752.

probit model is consistent with the theory described in this article.

Predicting Customer Satisfaction

Several studies of customer satisfaction model overall customer satisfaction as a function of performance of the product or service on various attributes (Oliver 1997). Overall satisfaction is considered an antecedent to repurchase intentions, which influence customer retention and firm profitability. Typically, overall satisfaction and perceived performance on attributes is measured in customer surveys by means of ratings scales. In practice, interest often centers on customers who check the top two boxes on the overall satisfaction ratings scale, that is, customers who are “completely satisfied.” The reason for this is that highly satisfied customers have been found to have a disproportionately higher probability of repurchase (Heskett et al. 1994; Mittal and Kamakura 2001, p. 139). Thus, it is worthwhile to develop predictive models of a discretized function of the overall satisfaction measure, namely, top boxes versus the rest of the scale.

To illustrate, we used survey data collected by a major U.S. automobile manufacturer to measure customers’ satisfaction with their service experience at the dealership. Customers were measured on ten-point rating scales immediately after they purchased a new vehicle. The usable sample consists of data from 100,345 respondents. We used reported satisfaction on seven attributes of the service experience (shown in Table 5) at the dealership to predict overall satisfaction with the service. We used the natural logarithm of the attribute perceptions to capture the possibility of diminishing returns (Anderson and Sullivan 1993). We controlled for age and sex of the respondent in the model.

Following common practice in the analysis of satisfaction data, we used a regression model for y and assumed that there were Gaussian errors. Consequently, we used a probit model to predict the probability of observing a top-two box score (i.e., 9 or 10 on the ten-point ratings scale) versus a score of less than 9. In Table 5, we show coefficient estimates for the regression and probit models based on the entire sample of respondents. All estimates are statistically significant. As we expected, satisfaction with each service

Table 5
RESULTS FOR CUSTOMER SATISFACTION APPLICATION

Coefficient	$\hat{\beta}_{d(y)}$		$\hat{\beta}_y$	
	Mean	Standard Error	Mean	Standard Error
Intercept	-18.40	.030	-1.54	.024
Time to complete service ^a	.22	.004	.25	.002
Honesty and sincerity ^a	4.40	.010	1.69	.015
All requested work complete ^a	.24	.004	.29	.003
Time elapsed from arrival to write-up ^a	1.10	.024	.12	.016
Quality of work performed ^a	.53	.014	.65	.010
Ability to schedule appointment ^a	.69	.021	.22	.014
Explained service work performed ^a	.10	.019	-.12	.013
Age of respondent	.01	.002	.02	.001
Sex of respondent (male = 1)	-.03	.013	-.04	.006
Mean PLL($\hat{\beta}; D_p$) across ten holdout samples		-1996.09		-2495.99

^aMeasured on a ten-point satisfaction scale. Attribute perceptions enter model in logarithmic form.

Notes: The response variable y is overall satisfaction with service experience at the automobile dealership.

attribute is positively related to satisfaction with overall service experience, with the exception of the coefficient for “explained service work performed,” which has an unexpected negative sign in the model for y .

We assessed the ability of both models to predict a top-two box score in ten holdout samples (as we described previously). The difference in mean PLL favors $\hat{\beta}_{d(y)}$ and is highly significant on the basis of a t-test ($p < .001$). Note that when a top-two box score from the model is predicted for y , the cutoff is taken to be 8.5. The correct classification rate across the ten holdout samples, shown in Table 6, is 93.14% for the probit, which is marginally higher than the correct classification rate for the regression model at 92.33%. Furthermore, in each of the ten holdout samples, the probit model does better in terms of PLL and correct classification. The difference in the number of discordant pairs in Table 6 is highly significant based on McNemar’s test ($\chi^2 = 222.40$, $p < .001$).

The RESET tests using squares, cubes, and fourth powers of the predicted dependent variable are highly significant ($p < .001$), indicating that the model is misspecified. Therefore, the superiority of $\hat{\beta}_{d(y)}$ in predictions is consistent with the theory in this article.

CONCLUSIONS AND MANAGERIAL IMPLICATIONS

Why does $\hat{\beta}_{d(y)}$ perform better under misspecification? Predictive accuracy has a bias component and a variance component. Under correct specification, both estimators have the same bias component, but $\hat{\beta}_{d(y)}$ has a worse variance component. Under misspecification, $\hat{\beta}_{d(y)}$ has a better bias component than $\hat{\beta}_y$, but $\hat{\beta}_{d(y)}$ has a worse variance component than $\hat{\beta}_y$. In large samples, the variance component diminishes in importance; only the bias component matters because the variance difference is of order $O(1/n)$, and the bias difference is of order $O(1)$. Therefore, the prescription of this theory is that in large samples in real data, it is better to use $\hat{\beta}_{d(y)}$.

How should marketing practitioners use our results? In direct marketing situations of the kind we consider, current practice is to compare a set of candidate model specifications of the response variable y in terms of accuracy of predicting $d(y)$ in a holdout sample. Our results imply that under misspecification, each of the candidate models of y may be dominated by the corresponding model for $d(y)$ in terms of predictions. Because practitioners do not know whether the candidate models are misspecified, it follows that models for $d(y)$ corresponding to each candidate model for y should be tested, in addition to testing the candidate models for y . In summary, our recommendation to marketing practitioners is to expand the set of candidate models that are evaluated to include the corresponding models for

$d(y)$. This will increase the average predictive performance of the selected response model.

The problem we have studied arises in many contexts in marketing, two of which we illustrate in this article. Another relevant context is the prediction of purchase intentions for a new product concept. Typically, the response variable is purchase intentions measured on an attitudinal scale by a survey, and the predictor variables are respondents’ evaluations of attributes of the new product. Commercial market research firms often focus on the distribution of a discretizing function of the measured purchase intentions. A popular example is the “top box” method, in which interest centers on whether a respondent checks the top box (i.e., he or she will definitely buy the product). In forecasting sales based on the measured purchase intentions, the assumption is often made that the percentage of all respondents who will actually buy is the percentage that checked the top box. Thus, the theoretical predictions of this article should be pertinent to that situation as well.

APPENDIX

Lemma 2: Difference in Precision

Under the assumptions of Lemma 1 and if $d(y)$ is not a sufficient statistic for y , then

$$(A1) \quad \mathbf{I}_{d(y)}^{-1} - \mathbf{I}_y^{-1} > 0 \quad \text{almost surely.}$$

To prove this lemma, we first invoke the linear algebra result that if $A > 0$, then $B > 0$, and if $A - B > 0$, then $B^{-1} - A^{-1} > 0$ (Horn and Johnson 1985). Therefore, to prove the lemma, we need to show only that $\mathbf{I}_y - \mathbf{I}_{d(y)} > 0$. Noting that the predictor vectors x are independent draws from F and that the expected value of the gradient of either of the two log-likelihoods at any observation is the zero vector (Gourieroux and Monfort 1995), we can write the information matrices in Equations 6 and 7 as follows:

$$(A2) \quad \mathbf{I}_y = N E_x \int \frac{1}{f(y; x, \beta)} [\nabla_{\beta} f(y; x, \beta)] [\nabla_{\beta} f(y; x, \beta)]^T dy \Big|_{\beta = \beta^*}$$

$$= N E_x \sum_q \int_{y \in I_q} \frac{1}{f(y; x, \beta)} [\nabla_{\beta} f(y; x, \beta)] [\nabla_{\beta} f(y; x, \beta)]^T dy \Big|_{\beta = \beta^*}, \text{ and}$$

$$(A3) \quad \mathbf{I}_{d(y)} = N E_x \sum_q \frac{1}{p[d(y) = q; x, \beta]} \{ \nabla_{\beta} p[d(y) = q; x, \beta] \} [\nabla_{\beta} p[d(y) = q; x, \beta]]^T \Big|_{\beta = \beta^*}.$$

Note that because $d(y)$ is a discretization of y ,

$$p[d(y) = q; x, \beta] = \int_{y \in I_q} f(y; x, \beta) dy.$$

Assume that the support for the density for y is not a function of the parameters, so that the integration and differenti-

Table 6

CROSS-CLASSIFICATION BY REGRESSION AND PROBIT MODELS IN TEN HOLDOUT SAMPLES OF SATISFACTION DATA

Classification by Probit Model	Classification by Regression Model	
	Correct	Incorrect
Correct	91,548	1918
Incorrect	1098	5781

ation operators are interchangeable. Therefore, the difference between the two information matrices can be written as

$$(A4) \quad \mathbf{I}_y - \mathbf{I}_{d(y)} = N E_x \sum_q \left(\int_{y \in I_q} \frac{1}{f(y; x, \beta)} [\nabla_{\beta} f(y; x, \beta)] \right. \\ \left. [\nabla_{\beta} f(y; x, \beta)]^T dy - \frac{1}{\int_{y \in I_q} f(y; x, \beta) dy} \right. \\ \left. \left\{ \nabla_{\beta} \left[\int_{y \in I_q} f(y; x, \beta) dy \right] \right\} \right. \\ \left. \left\{ \nabla_{\beta} \left[\int_{y \in I_q} f(y; x, \beta) dy \right] \right\}^T \right) \Big|_{\beta = \beta^*}.$$

Consider any vector $v \neq 0$ that is nonstochastic, in the sense that it does not depend on y . Denote the inner product of v and the gradient of the likelihood of y as $v_y \equiv v^T \nabla_{\beta} f(y; x, \beta)$. In this notation, it follows that

$$(A5) \quad v^T \mathbf{I}_y v - v^T \mathbf{I}_{d(y)} v = N E_x \sum_q \left[\int_{y \in I_q} \frac{1}{f(y; x, \beta)} v_y^2 dy \right. \\ \left. - \frac{1}{\int_{y \in I_q} f(y; x, \beta) dy} \right. \\ \left. \left(\int_{y \in I_q} v_y dy \right)^2 \right] \Big|_{\beta = \beta^*} \\ = N E_x \sum_q \left\{ \frac{1}{\int_{y \in I_q} f(y; x, \beta) dy} \right. \\ \left. \int_{y_1 \in I_q} \int_{y_2 \in I_q} \left[\sqrt{\frac{f(y_2; x, \beta)}{f(y_1; x, \beta)}} v_{y_1} \right. \right. \\ \left. \left. - \sqrt{\frac{f(y_1; x, \beta)}{f(y_2; x, \beta)}} v_{y_2} \right]^2 dy_1 dy_2 \right\} \Big|_{\beta = \beta^*} \\ > 0 \text{ almost surely.}$$

Therefore, the matrix $\mathbf{I}_y - \mathbf{I}_{d(y)}$ is positive definite.

It is worth discussing the circumstances in which the preceding expression is equal to zero. If the function $d(y)$ is sufficient for y , then v_{y_2} is always equal to v_{y_1} , and the two information matrices are equal. The more interesting case is when $d(y)$ is not sufficient for y , as is assumed in Lemma 2. In this case, there exist function vectors v and likelihood functions f , where $v_{y_2} = v_{y_1}$ over the sample space. However, the measure for the set of such functions can be shown to be zero for regular measures. Thus, we used the qualification ‘‘almost surely’’ in the statement of the lemma.

Lemma 3: Difference in Expected Predictive Accuracy Under Correct Specification

Under the assumptions of Lemma 1 and Lemma 2,

$$E[\text{PLL}(\hat{\beta}_y; D_p) - \text{PLL}(\hat{\beta}_{d(y)}; D_p)] > 0.$$

The distribution expressions in Lemma 1 imply from second-order asymptotics that for large N (see Greene 2000),

$$(A6) \quad E \text{ PLL}(\hat{\beta}_y; D_p) = \text{PLL}(\beta^*; D_p) \\ + \text{trace} \left[\mathbf{I}_y^{-1} \nabla_{\beta} \nabla_{\beta}^T \text{PLL}(\beta; D_p) \Big|_{\beta = \beta^*} \right], \text{ and}$$

$$(A7) \quad E \text{ PLL}(\hat{\beta}_{d(y)}; D_p) = \text{PLL}(\beta^*; D_p) \\ + \text{trace} \left[\mathbf{I}_{d(y)}^{-1} \nabla_{\beta} \nabla_{\beta}^T \text{PLL}(\beta; D_p) \Big|_{\beta = \beta^*} \right].$$

Denote the Hessian of the predictive accuracy as follows:

$$(A8) \quad \mathbf{H} \equiv \nabla_{\beta} \nabla_{\beta}^T \text{PLL}(\beta; D_p) \Big|_{\beta = \beta^*}.$$

Let the negative Cholesky decomposition of \mathbf{H} be $\mathbf{H} = -\mathbf{H}^{1/2} \mathbf{H}^{1/2^T}$. Let the columns of $\mathbf{H}^{1/2}$ be h_1, h_2, \dots, h_C . Therefore, we can write the difference between the predictive accuracies as follows:

$$(A9) \quad E[\text{PLL}(\hat{\beta}_y; D_p) - \text{PLL}(\hat{\beta}_{d(y)}; D_p)] \\ = \text{trace}[(\mathbf{I}_y^{-1} - \mathbf{I}_{d(y)}^{-1})\mathbf{H}] \\ = \text{trace}[-\mathbf{H}^{1/2^T} (\mathbf{I}_{d(y)}^{-1} - \mathbf{I}_y^{-1}) \mathbf{H}^{1/2}] \\ = \sum_{c=1}^C h_c^T (\mathbf{I}_{d(y)}^{-1} - \mathbf{I}_y^{-1}) h_c \\ > 0.$$

The last step follows from the result of Lemma 2 that $[\mathbf{I}_{d(y)}^{-1} - \mathbf{I}_y^{-1}]$ is positive definite.

Convergence and Asymptotic Normality of $\hat{\beta}_y$ and $\hat{\beta}_{d(y)}$ in the Misspecification Case

Let the minimizers in the set P_f of the Kullback–Liebler distance to the true likelihood of y and $d(y)$ be denoted, respectively, as

$$(A10) \quad \beta_y^* = \arg \min_{\beta} E_X E_{Y|X} \log \left[\frac{\tau_{y|x}(y; x)}{f(y; x, \beta)} \right], \text{ and}$$

$$(A11) \quad \beta_{d(y)}^* = \arg \min_{\beta} E_X E_{Y|X} \log \left[\frac{\int_{y' \in I_{d(y)}} \tau_{y|x}(y'; x) dy'}{\int_{y' \in I_{d(y)}} f(y'; x, \beta) dy'} \right].$$

In Equation A11, we substitute the expressions for the likelihoods of $d(y)$ under $\tau_{y|x}$ and $f(y; x, \beta)$, which are, respectively,

$$\int_{y' \in I_{d(y)}} \tau_{y|x}(y'; x) dy'$$

and

$$\int_{y' \in I_{d(y)}} f(y'; x, \beta) dy'$$

into the usual expression for Kullback–Liebler distance. (Note that in Equations A10 and A11 and everywhere else in this section, the expectation $E_{Y|X}$ is carried with respect to the distribution of y under the true density $\tau_{y|x}$.) Therefore, β_y^* indexes the element in P_f that most closely reconstructs the true likelihood of $d(y)$.

We need one more piece of notation. We denote the negative expected Hessian matrices of the log-likelihoods as

$$(A12) \quad J_y \equiv -N E_X \nabla_{\beta} \nabla_{\beta}^T LLY(\beta) |_{\beta = \beta_y^*}, \text{ and}$$

$$(A13) \quad J_{d(y)} \equiv -N E_X \nabla_{\beta} \nabla_{\beta}^T LLDY(\beta) |_{\beta = \beta_{d(y)}^*}.$$

We can apply a central result in misspecification theory (see White 1994) here to give the asymptotic distributions of $\hat{\beta}_y$ and $\hat{\beta}_{d(y)}$:

$$(A14) \quad \hat{\beta}_y \sim N(\beta_y^*, J_y^{-1} \mathbf{I}_y J_y^{-1}), \text{ and}$$

$$(A15) \quad \hat{\beta}_{d(y)} \sim N(\beta_{d(y)}^*, J_{d(y)}^{-1} \mathbf{I}_{d(y)} J_{d(y)}^{-1}).$$

We derive these expressions for the distributions under the assumption that the log-likelihood function under misspecification is differentiable. It must not go unmentioned that under misspecification, in general, we have $J_y \neq \mathbf{I}_y$ and $J_{d(y)} \neq \mathbf{I}_{d(y)}$. Compare this with the situation under correct specification, where $J_y = \mathbf{I}_y$ and $J_{d(y)} = \mathbf{I}_{d(y)}$ and Equations A14 and A15 reduce to the distribution expressions in Lemma 1.

The Inequality of β_y^ and $\beta_{d(y)}^*$ in the Misspecification Case*

It is important to note that, in general, β_y^* is not equal to $\beta_{d(y)}^*$. This is simply because β_y^* yields the density in P_f that is closest to the true likelihood of y , whereas $\beta_{d(y)}^*$ yields the density in P_f closest to the true likelihood of $d(y)$. To understand this more formally, consider the Kullback–Liebler identity that the minimizer of the distance to any density is the density itself. Therefore, considering the density of $d(y)$ under P_f at $\beta = \beta_y^*$,

$$(A16) \quad \beta_y^* = \arg \min_{\beta} E_X E_{Y|X} \log \left[\frac{\int_{y' \in I_{d(y)}} f(y'; x, \beta_y^*) dy'}{\int_{y' \in I_{d(y)}} f(y'; x, \beta) dy'} \right].$$

Recall that $\beta_{d(y)}^*$ is given by Equation A11. The expression on the right-hand side of the equality sign in Equation A11 differs from the expression on the right-hand side in Equation A16 in only one respect: The expression for $\beta_{d(y)}^*$ has

$\tau_{y|x}(y'; x)$, whereas the expression for β_y^* has $f(y'; x, \beta_y^*)$. Because the model is misspecified,

$$(A17) \quad \tau_{y|x}(y'; x) \neq f(y'; x, \beta_y^*).$$

Therefore, from Equations A11, A16, and A17, in general, $\beta_y^* \neq \beta_{d(y)}^*$. An exception occurs again if $d(y)$ is sufficient for y .

Lemma 4: Difference in Expected Predictive Accuracy Under Misspecification

If $\hat{\beta}_{d(y)}$ is fully identified in that it is the unique maximizer of the likelihood of $d(y)$ under P_f , and $\beta_y^* \neq \beta_{d(y)}^*$, then for large N

$$E[\text{PLL}(\hat{\beta}_{d(y)}; D_p) - \text{PLL}(\hat{\beta}_y; D_p)] > 0.$$

Because $\hat{\beta}_{d(y)}$ and $\hat{\beta}_y$ converge respectively to $\beta_{d(y)}^*$ and β_y^* , we need to show only that

$$(A18) \quad \text{PLL}(\beta_{d(y)}^*; D_p) - \text{PLL}(\beta_y^*; D_p) > 0.$$

To understand this, consider the maximizer of the predictive accuracy $\text{PLL}()$ as defined in Equation 2:

$$\begin{aligned} \arg \max_{\beta} \text{PLL}(\beta; D_p) &= \arg \max_{\beta} N_p E_X E_{Y|X} \\ &\log \left[\int_{y' \in I_{d(y)}} f(y'; x, \beta) dy' \right] \\ &= \beta_{d(y)}^*. \end{aligned}$$

The last step in the preceding follows closely from the definition of $\beta_{d(y)}^*$. In addition, that $\beta_{d(y)}^*$ is the unique maximizer of the likelihood of $d(y)$ under P_f directly implies that $\beta_{d(y)}^*$ is the unique maximizer of $\text{PLL}()$:

$$\text{PLL}(\beta_{d(y)}^*; D_p) - \text{PLL}(\beta; D_p) > 0 \quad \forall \beta \neq \beta_{d(y)}^*.$$

Therefore, the preceding statement is true, particularly for $\beta = \beta_y^*$.

REFERENCES

- Anderson, Eugene W. and Mary Sullivan (1993), "The Antecedents and Consequences of Customer Satisfaction for Firms," *Marketing Science*, 12 (2), 125–43.
- Barndorff-Nielsen, O.E. and D.R. Cox (1994), *Inference and Asymptotics*. London: Chapman and Hall.
- Bult, J. and T. Wansbeek (1995), "Optimal Selection for Direct Mail," *Marketing Science*, 14 (4), 378–94.
- Gallant, Ronald (1997), *An Introduction to Econometric Theory*. Princeton, NJ: Princeton University Press.
- Godfrey, L.G. (1988), *Misspecification Tests in Econometrics*. Cambridge, UK: Cambridge University Press.
- Gourieroux, Christian and Alain Monfort (1995), *Statistics and Econometric Models*. Cambridge, UK: Cambridge University Press.
- Greene, William H. (2000), *Econometric Analysis*, 4th ed. Upper Saddle River, NJ: Prentice Hall.
- Heskett, James L., Thomas O. Jones, Gary W. Loveman, and W. Earl Sasser (1994), "Putting the Service-Profit Chain to Work," *Harvard Business Review*, 72 (March–April), 164–74.
- Horn, Roger A. and Charles R. Johnson (1985), *Matrix Analysis*. New York: Cambridge University Press.

- Kennedy, Peter (1993), *A Guide to Econometrics*. Cambridge: Massachusetts Institute of Technology Press.
- Levin, Nissan and Jacob Zahavi (1998), "Continuous Predictive Modeling—A Comparative Analysis," *Journal of Interactive Marketing*, 12 (2), 5–22.
- Lindsey, James K. (1996), *Parametric Statistical Inference*. Oxford, UK: Clarendon Press.
- Malthouse, Edward C. (2002), "Performance-Based Variable Selection for Scoring Models," *Journal of Interactive Marketing*, 16 (4), 37–50.
- Mittal, Vikas and Wagner Kamakura (2001), "Satisfaction, Repurchase Intent, and Repurchase Behavior: Investigating the Moderating Effect of Customer Characteristics," *Journal of Marketing Research*, 38 (February), 131–42.
- Oliver, Richard L. (1997), *Satisfaction: A Behavioral Perspective on the Consumer*. New York: McGraw-Hill.
- Ramsey, J.B. (1969), "Tests for Specification Error in Classical Linear Least Squares Regression Analysis," *Journal of the Royal Statistical Society*, (B31), 250–71.
- Siegel, S. and N.J. Castellan Jr. (1998), *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.
- White, Halbert (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–26.
- (1994), *Estimation, Inference, and Specification*. Cambridge, UK: Cambridge University Press.

Copyright of Journal of Marketing Research (JMR) is the property of American Marketing Association and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.