

BAS DONKERS, PHILIP HANS FRANSES, and PETER C. VERHOEF*

Marketing problems sometimes pertain to the analysis of dichotomous dependent variables, such as “buy” and “not buy” or “respond” and “not respond.” One outcome can strongly outnumber the other, such as when many households do not respond (e.g., to a direct mailing). In such situations, an efficient data-collection strategy is to sample disproportionately more from the smaller group. However, subsequent statistical analysis must account for this sampling strategy. In this article, the authors put forward the econometric method that can correct for the sample selection bias, when this method does not lead to a loss in precision. The authors illustrate the method for synthetic and real-life data and document that reductions of more than 50% in sample sizes can be obtained.

Selective Sampling for Binary Choice Models

Market researchers frequently analyze customers who differ on a dichotomous outcome variable, such as those who try a new product versus those who do not or those who defect versus those who stay (e.g., Baum and Korn 1999; Bolton, Kannan, and Bramlett 2000; Frambach et al. 1998; Ganesh, Arnold, and Reynolds 2000). In many cases, researchers are interested in the antecedents of the underlying decisions and collect data to investigate the possible antecedents. In such situations, one group may be much smaller than the other, and efficiencies can be gained by oversampling the smaller group. However, in doing so, subsequent analysis models need to be modified, which we demonstrate in this article.

Suppose that the incidence of an infrequently occurring outcome is 5% and that a researcher wants at least 100 observations for each outcome. The researcher should collect a sample of 2000 observations, at least according to a random sampling scheme; however, this can be rather expensive. If a researcher were to sample disproportionately more from the smaller group or use outcome-dependent sampling, substantial amounts of money could be saved and efficiency preserved.

Market researchers usually stratify their samples on independent variables (see, e.g., Aaker, Kumar, and Day 2001; Lehmann, Gupta, and Steckel 1998), which is called exogenous stratification. In contrast, in outcome-dependent sampling, the sample is stratified on the y variable. Thus, strati-

fication is endogenous. Disproportionate stratification can also be performed on both the x and the y variables simultaneously. This stratification is recommended when both variables are unevenly distributed and the x variable might have a substantial impact on the y variable.

In our empirical illustration, we present a case in which the latter situation occurs, and we argue that this situation is typical in marketing research. We consider an insurance company that is interested in the drivers of customer retention. In general, defection rates are rather low in the insurance industry. In the first year of a relationship, customer defection is substantially higher than in subsequent years, that is, a defection rate of 7.6% in the first year versus 3.5% during the remainder of the relationship. Customer defection (y) and relationship age (x) are unevenly distributed in the total customer database. Thus, researchers can benefit from overrepresenting defecting customers and customers with short relationships.

A disadvantage of disproportionate stratification on the y variable only, and on the y variable and the x variables simultaneously, is that it affects the estimation results of logit and probit models. Indeed, the standard assumptions required for consistent parameter estimation are violated (Franses and Paap 2001; Greene 2000). Fortunately, there are estimation methods that correct for outcome-dependent sampling (Cosslett 1993; Imbens and Lancaster 1996; Scott and Wild 1997), and we address these methods herein. To use these methods, additional information is needed on the population distribution of the variables used for stratification. Most customer information databases contain such information; when this is not the case, estimates of the population distribution can be derived from observed screening rates.

In a few disciplines, such as biometrics, people seem well aware of correction methods, but in many other disciplines they are not. Given scholars' increasing interest in customer relationship management, in which infrequently observed binary outcomes are quite common, there is value in outlin-

*Bas Donkers is an assistant professor (e-mail: donkers@few.eur.nl); Philip Hans Franses is Professor of Marketing Research, Econometric Institute (e-mail: franses@few.eur.nl); and Peter C. Verhoef is an assistant professor (e-mail: verhoef@few.eur.nl), Department of Marketing and Organization, Erasmus University, Rotterdam. The authors thank the three anonymous *JMR* reviewers, Michel Wedel, and seminar participants at the Tinbergen Institute, Tilburg University, and the 2001 Marketing Science Conference in Wiesbaden for helpful comments.

ing the relevant methods. Given the large range of explanatory variables used in marketing and their importance in the analysis, we pay special attention to the role of sampling on the explanatory variables in combination with outcome-dependent sampling. We show that when the explanatory variables have a skewed distribution, selective sampling on the explanatory variables can reduce sample sizes even further without a loss in estimation precision. In this respect, we extend the work of Scott and Wild (1997) in a direction that is especially relevant to marketing research. Therefore, the objective of this article is to show how consistent parameter estimates of the effect sizes in logit models can be obtained when researchers use data collected by means of outcome-dependent sampling. We show that the necessary adjustments to existing estimation routines are easy to implement, even in the case of selective sampling on both dependent and independent variables.

The remainder of this article is structured as follows: In the next section, we present an estimation technique that accounts for the effect of outcome-dependent sampling based on the work of Scott and Wild (1997). We also discuss implications for more efficient sampling schemes. We then present a simulation study to show how the potential reduction in sample sizes is related to characteristics of the population. Next, we apply the proposed methodology to the analysis of real-life data on the antecedents of customer retention. For this example, we find that the described methodology allows for a 60% reduction in sample size. We conclude with a short summary and the limitations of the proposed methodology.

METHODOLOGY

Models for binary dependent variables are usually analyzed by means of maximum-likelihood procedures. In this section, we discuss how full maximum-likelihood estimates can be obtained from a sample that is stratified on the dependent variable of interest and possibly on exogenous variables. To estimate the parameters, we need to determine the likelihood function of the data, given information on the sampling scheme that is used. We introduce some notation and other preliminaries and then turn to discussing the likelihood of binary choice models when the sampling scheme for data collection is outcome dependent. We start with a situation in which stratification on x is allowed but assumed independent of stratification on y . We continue with a case in which stratification is based on combinations of x and y . The section ends with a discussion of any simplifications that arise when the logit model is used and some implications for the use of outcome-dependent sampling schemes in practice.

Preliminaries

We assume that the dependent variable y_i has the value of 1 or 0. Thus, y_i can describe a firm's decision to enter a market or a customer's decision to stay with the company. For the exogenous variables collected in x_i , we define strata, $s = 1, \dots, S$. The strata can be based on only a subset of the explanatory variables. In general, stratification on a subset of the variables occurs when the stratification is used to reduce survey costs.

We present the sampling strategies in the following setup: There is a finite population of N people, and all that is known about the population is that there are N_{sj} people with

$y_i = j$ in stratum s , where $j = 0, 1$ and $s = 1, \dots, S$. In general, this finite population is the population that is of interest to researchers, which in a direct marketing or customer relationship management application is often the current customer base, but it can also include all potential customers. In each stratum, a random sample of size $n_{sj} \leq N_{sj}$ is drawn from the stratum members. The data consist of y_i and x_i , which are recorded for these observations. To correct for outcome-dependent sampling, information on the population frequencies (N_{sj}) is required. When the population of interest is the current customer base, such information is readily available from the customer information database. In other situations, good estimates for these can be obtained from screening rates obtained during data collection.¹

Selection on y_i

The simplest endogenous sampling scheme is that during which a response category, $y_i = 0$ or $y_i = 1$, is selected randomly with probabilities n_0/n and n_1/n in the first stage, and the observation is drawn randomly from the subpopulation with y_i in the selected response category. With this sampling scheme, the sample sizes for each response category are random. Let $P(y_i = j|x_i)$ denote the probability of observing outcome j given x in the sample, and let $P^*(y_i = j|x_i)$ denote the same probability for the population of interest. We are interested in learning about these population probabilities. In this fully randomized sampling scheme, it holds that

$$(1) \quad P(y_i = j|x_i) = \frac{\mu_j P^*(y_i = j|x_i)}{\mu_0 P^*(y_i = 0|x_i) + \mu_1 P^*(y_i = 1|x_i)},$$

where $\mu_j = (n_j/n)/P^*(y_i = j)$ represents the ratio of the probability that an observation belongs to class j in the sample (n_j/n) to the probability that an observation falls into class j in the population $P^*(y_i = j)$. The probability of observing $y_i = j$ in the data is greater than in the population when observations with $y_i = j$ are sampled with a greater probability than they occur in the population; that is, $P(y_i = j|x_i) > P^*(y_i = j|x_i)$ whenever $(n_j/n) > P^*(y_i = j)$. In Equation 1, this is equivalent to $\mu_j > 1$.

The major advantage of this simple randomized sampling scheme is that the likelihood is easy to obtain; however, its relevance is much broader. Cosslett (1993) indicates that when population and sample frequencies are known, the sample can be treated with predetermined group sizes as if it were constructed with the fully randomized sampling scheme, based on conditioning on sufficient statistics. Because the likelihood is not based on the actual sampling scheme, the likelihood is called the "pseudolikelihood." Maximum-likelihood estimation based on the pseudolikelihood results in consistent estimates. For the familiar logit model, the resulting parameter estimates are also efficient. When the parameters are estimated without correcting for the outcome-dependent sampling scheme, all parameter estimates will be inconsistent. An exception is the logit model, in which only the estimate for the intercept is affected.

Simultaneous Selection on y_i and x_i

Derivation of the pseudolikelihood for the sampling scheme in which selection depends on the outcome in com-

¹We thank an anonymous reviewer for pointing this out.

bination with the strata of the exogenous variables is now straightforward. Because the samples from the different strata are drawn independently, the total pseudolikelihood of all observations equals the product of the pseudolikelihoods of the different strata. For each individual stratum, we can obtain the likelihood by means of Equation 1, in which the population of interest is the stratum being considered. The equivalent of Equation 1 when the population of interest is a stratum is

$$(2) \quad P(y_i = j|x_i, \text{stratum}_i = s) \\ = \frac{\mu_{sj}P^*(y_i = j|x_i)}{\mu_{s0}P^*(y_i = 0|x_i) + \mu_{s1}P^*(y_i = 1|x_i)}.$$

The full-sample pseudolikelihood now reads as follows:

$$(3) \quad L(\theta) = \prod_{s=1}^S \prod_{j=0}^1 \prod_{i=1}^{n_{sj}} P(y_i = j|x_i, \text{stratum}_i = s) \\ = \prod_{s=1}^S \prod_{j=0}^1 \prod_{i=1}^{n_{sj}} \frac{\mu_{sj}P^*(y_i = j|x_i)}{\mu_{s0}P^*(y_i = 0|x_i) + \mu_{s1}P^*(y_i = 1|x_i)}.$$

If we use the logit model, maximization of $L(\theta)$ results in consistent, but not necessarily efficient, estimates. Efficient estimates are obtained when there is a stratum-specific intercept (a dummy variable) for each stratum of x_i in the data. If there are no stratum-specific intercepts or if another binary choice model (e.g., the probit model) is used, efficient estimates can be obtained from an iterative procedure (see Scott and Wild 1997). Note that even though certain x 's might be fixed within a stratum, their effects will be identified from the variation across the different strata.

The Logit Model

Thus far we have presented a general formulation of the likelihood that can be used for any binary choice model. However, when the familiar logit model is used, the model can be easily estimated for different sampling schemes by means of various statistical packages. Consider the probability of observing $y_i = 1$ in the standard logit model (Greene 2000):

$$(4) \quad P^*(y_i = 1|x_i) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}.$$

To obtain the pseudomodel in Equation 2, we need to correct this probability with parameters μ_{s0} and μ_{s1} . When $\hat{\mu}_{s0} = n_{s0}/N_{s0}$ and $\hat{\mu}_{s1} = n_{s1}/N_{s1}$ are used as estimates of μ_{s0} and μ_{s1} , the probability in the pseudolikelihood is

$$(5) \quad P(y_i = j|x_i, \text{stratum}_i = s) \\ = \frac{\mu_{sj}\exp(x_i'\beta)/[1 + \exp(x_i'\beta)]}{\mu_{s0}/[1 + \exp(x_i'\beta)] + \mu_{s1}\exp(x_i'\beta)/[1 + \exp(x_i'\beta)]} \\ = \frac{\exp[x_i'\beta + \ln(\mu_{sj}/\mu_{s0})]}{1 + \exp[x_i'\beta + \ln(\mu_{s1}/\mu_{s0})]}.$$

Therefore, the pseudolikelihood estimator of the logit model is easily obtained by adding a correction of $\ln(\hat{\mu}_{s1}/\hat{\mu}_{s0})$ to $x_i'\beta$ for each observation. This can be done in standard statistical packages that can include offsets or estimate models

under parameter restrictions, enabling the parameter of the correction to be fixed.

We have discussed parameter estimation; it turns out that when the logit model is used, efficient estimates are obtained straightforwardly. However, without standard errors, the parameter estimates are difficult to interpret. The regular standard errors that result from maximum-likelihood estimation based on the pseudolikelihood can be used, but the true standard errors will be smaller. More details on how to compute the exact standard errors can be found in Scott and Wild's (1997) study.

Practical Considerations

The optimal ratio of observations with $y_i = 1$ and $y_i = 0$ depends on the application at hand. Lancaster and Imbens (1991) present theoretical and simulation results regarding the optimal sample composition and indicate that an equal split of the sample into 50% and 50% is often close to optimal. Moreover, as Breslow and Day (1980) and Cramer, Franses, and Slagter (1999) indicate, when a sample has an 80%–20% split among 1–0 observations, there is only little to be gained by adding more 1s.

The results can be understood as follows: Because information on the relationship of interest is based on the simultaneous variation of the dependent and independent variable, it might be useful to maximize the variation in the dependent variable. The sample variation in the dependent variable (measured by its variance) equals $F_n(y_i = 1) \times [1 - F_n(y_i = 1)]$, where $F_n(y_i = 1)$ is the sample fraction of observations with $y_i = 1$. This variation attains its maximum for $F_n(y_i = 1) = .5$. Splitting the sample in equal parts also makes sense given symmetry considerations, because there is no a priori reason to have more observations from one type than the other.

SYNTHETIC DATA

In this section, we present a simulation study that shows how the potential gains vary with the distribution of the dependent variable in the population. We also examine the consequences of stratification on an unevenly distributed binary explanatory variable when it is also used for stratification.

Because our interest lies in the efficiency of sampling schemes, we need a measure of efficiency. Kuhfeld, Tobias, and Garrat (1994) suggest using a D-efficiency measure, which is based on the determinant of the covariance matrix of the parameter estimates. Even though they focus on the efficiency of experimental designs, the measure is also well suited for the comparison of sampling schemes. However, a more convenient measure (which has a one-to-one relationship with D-efficiency) is the D-error measure (Arora and Huber 2001), which does not require the variables to be suitably scaled. The D-error measure is defined as $|\Sigma|^{1/K}$, where Σ is the variance-covariance matrix of the parameter estimates, and K is the number of explanatory variables. Lower values of D-error imply more efficient designs, when a reduction of D-error with $\alpha\%$ enables the use of an $\alpha\%$ smaller sample without loss of efficiency.

We now turn to a limited study that uses synthetic data. The simulated samples always consist of 1000 observations, and these are drawn from a population of 200,000 people. In this population, the fraction of observations having $y_i = 1$ varies from 2.5% to 50%. The data are generated in accor-

dance with the logit model, and this model is also used for estimation. The model used to generate the data consists of two explanatory variables: a binary (discrete) variable, x_d , and a continuously distributed variable, x_c . The continuous variable is $N(0,1)$ distributed, and the distribution of the binary variable varies over the two simulation designs. In the first simulation design, the binary explanatory variable is evenly distributed, and thus each value is equally likely. In this situation, it does not make much sense to select on y and x simultaneously. In the second design, the binary explanatory variable has a skewed distribution: One value occurs nine times more often than the other.

For the first design, in which x is symmetrically distributed, we report the average D-error measure for 500 replications in the first two rows of Table 1. The first row presents the results for a random sample from the population, and the second row presents the results for the outcome-dependent sampling scheme. We report the efficiency gain obtained by the nonrandom sampling scheme in parentheses. This also measures the potential percentage reduction in the required sample size. The last four rows of Table 1 present the average D-error measures and percentage efficiency gains for the second design, in which stratification on x_d is also considered, again based on 500 replications.

Table 1 shows that the use of outcome-dependent sampling schemes can lead to substantial reductions in sample sizes. When an outcome occurs for only 2.5% of the people in the population, sample size reductions of 60% to 80% are feasible, depending on the distribution of x . Efficiency gains obviously are smaller when the dependent variable is less rare. It might be concluded that the use of outcome-dependent sampling schemes is not worthwhile when the infrequent outcome occurs for more than 15% of the population.

The bottom part of Table 1 shows the consequences of stratification on explanatory variables when they have a skewed distribution. The first column shows that stratification on only the x variables results in an efficiency gain of 42% and 71% for stratification on y only. For infrequent outcomes, stratification on y is therefore more powerful than stratification on x . However, as we expected, the most powerful sampling scheme stratifies on x and y simultaneously, leading to an efficiency gain of 79%. Note that when the dependent variable is more evenly distributed, there are still

substantial efficiency gains possible by stratifying on x , as is evident from the last column in Table 1.

REAL-LIFE DATA

In this section, we apply our proposed methodology to real-life data in marketing research. As we show, selective sampling followed by an appropriate estimation method would have enabled the company to interview up to 60% fewer people without loss of precision. The data in this application emerge from the customer base of an insurance company in the Netherlands. This company is a large direct writer and does not use insurance agents as intermediaries. The company sells all types of insurance policies, ranging from fire and theft insurance to life insurance. The company aims at having close relationships with its customers. Because the company does not have intermediaries who can signal customer dissatisfaction, it is highly interested in the determinants of customer satisfaction and the role satisfaction plays in a customer's decision to leave the company. Moreover, it is especially interested in the behavior of customers who are with the company for less than one year, because these customers leave the company at a more frequent rate than customers with longer relationships (7.6% versus 3.5%). An additional variable of interest is the number of insurance policies a customer purchases. This measure may reflect a customer's loyalty, but it may also indicate higher switching costs.

An exogenously stratified sample of 2300 customers was collected in 1999 to obtain information about customer satisfaction, and the customers were reinterviewed in 2000. For 1374 customers, we observed answers to the questions we had. From the 1374 customers, only 53 (3.9%) have left the company since the first interview in 1999. Because the company realized that this was a rather small number of observations on inactive customers to perform statistical analyses, additional interviews were conducted. More precisely, the company gathered information from an additional random sample of 30 customers who left the company and who were not interviewed in 1999. These customers were randomly selected, conditional on having left the company in the past 12 months. For the analysis of customer retention, the total sample is therefore endogenously stratified. Some descriptive statistics for the original random sample and the additional sample are presented in Table 2. The dummy variable

Table 1
EFFICIENCY MEASURES FOR OUTCOME-DEPENDENT SAMPLING SCHEMES

	Percentage of Population with $y = 1$				
	2.5%	5%	10%	15%	50%
$P(x_d = 1) = 50\%$					
Random	15.8	8.5	6.9	6.2	6.1
Selection on y	6.1 (61%)	6.1 (28%)	6.1 (12%)	6.1 (2%)	6.1 (0%)
$P(x_d = 1) = 10\%$					
Random	29.7	12.5	9.8	8.8	8.6
Selection only on y	8.6 (71%)	8.6 (31%)	8.6 (21%)	8.6 (2%)	8.6 (0%)
Selection on y and x	6.1 (79%)	6.1 (51%)	6.1 (38%)	6.1 (31%)	6.1 (29%)
Selection only on x	17.1 (42%)	9.2 (26%)	7.1 (27%)	6.3 (28%)	6.1 (29%)

Notes: The reduction in sample size, relative to a random sample, is given in parentheses.

Table 2
DESCRIPTIVE STATISTICS

	Original Sample		Additional Sample
	Active	Inactive	Inactive
N	1321	53	30
Short_duration (mean)	.12	.23	.20
Satisfaction (mean)	3.4	3.1	3.2
Number of insurance products (mean)	2.1	1.0	1.2
Dummy 1 insurance (mean)	.55	.96	.83

short_duration indicates whether the customer is with the company for at most one year at the time of the interview (1) or not (0). Satisfaction is a measure of the customer's general satisfaction with the company, based on a set of questions that used five-point Likert scales.

We focus on the effect of customer satisfaction on the decision to leave the company, where special attention is paid to the apparently higher customer defection rate in the first year of a customer's relationship (compare .23 with .12 in Table 2). The total sample of observations consists of a random sample and a nonrandom sample of customers who have left the insurance company. To illustrate the practical consequences of outcome-dependent sampling on the precision of the estimated coefficients, we present the estimation results of a logit model for leaving the company ($y = 1$) in Table 3. Each column pertains to a different sampling scheme.

Column 1, Table 3, presents the estimation results of the logit model for the original random sample. Customer satisfaction and the number of insurance products purchased reduce the probability that a customer leaves the company, and these effects are significant. Customers with a short relationship duration are slightly more likely to leave the company, but this effect is statistically not significant.

The role that outcome-dependent sampling can play in data collection is illustrated in the remaining columns of Table 3. The estimation results in Columns 2 and 3 are corrected for the outcome-dependent sampling scheme with the preceding methods we advocated. As we expected, the parameter estimates do not differ much. However, the uncorrected estimates in Column 4 are substantially different.

Column 2 presents the estimation results based on the original and the additional sample combined. These are the most efficient estimates we can obtain with our data, because they are based on all the available observations. Compared with the original sample, we expected a decrease in all standard errors of approximately 2% when the additional observations constituted a random draw from the entire population.² However, compared with Column 1, standard errors have decreased by more than 17%. Note that if a random sampling scheme is used, the sample size would need to be increased to $1.17^2 = 1.37$ times the original size. Thus, the gain in precision from 30 nonrandomly selected observations is equivalent to the gain of more than 500 randomly selected observations.

²On the basis of \sqrt{N} consistency of the estimates, we expect standard errors of the parameter estimates to be $\sqrt{1374/(1374 + 30)} = .979$ times the original ones, amounting to the reported decrease of 2%.

Table 3
ESTIMATION RESULTS

	Samples			
	1 ^a	2 ^b	3 ^c	4 ^d
Sample size (N)	1374	1404	505	505
Intercept	2.312 (1.087)	1.806 (.810)	1.940 (.917)	2.645 (.893)
Satisfaction	-.919 (.261)	-.854 (.223)	-.888 (.256)	-.866 (.254)
Short_duration	.329 (.346)	.363 (.290)	.367 (.297)	-.928 (.292)
Number of insurance products	-2.040 (.646)	-1.756 (.308)	-1.787 (.328)	-.916 (.268)
N × D – measure	125.14	66.99	28.02	—

^aOriginal sample.

^bOriginal sample and additional sample.

^cSampling scheme using simultaneous selection on x and y (see text for details).

^dSame as 3 but not corrected for outcome-dependent sampling.

Notes: Standard errors are in parentheses.

The most important gain of using outcome-dependent sampling schemes is the possible reduction in sample size and survey costs. This is shown in Column 3, Table 3. The purpose of the sample used for estimation in Column 3 is twofold. First, it shows how large the potential gains in a real-life setting can be. Second, it illustrates the flexibility of the methodology by using a rather complicated sampling scheme, which we outline next.

The sampling scheme stratifies the population by means of the strata that result from all possible combinations of the dependent variable, the dummy for a short relationship, and the dummy for having only one insurance policy. First, the customers who defected were all retained in the sample. For the customers who stayed with the company, we selectively reduced the number of observations in some strata. Among customers who stay with the company, many have a long company relationship. Therefore, we substantially reduced the number of observations in the strata with customers with a long relationship. Among these customers, we distinguished between customers with one insurance product and customers with multiple insurance products. In doing so, we observed many more customers who purchase multiple insurance products. The stratum with customers who purchase multiple insurance products is therefore reduced by 85%, and the stratum of customers with only one insurance product is reduced by 70%. Recall that the two strata only differ in the number of insurance products purchased, and both contain only customers who stay with the company and have a long relationship.

The resulting sample is less than 40% of the size of the original sample, but, as is evident in Column 3, the precision of all parameters is greater than the precision from the parameter estimates in the original sample. Therefore, the company could have saved more than 60% of the money it spent on the random sample without giving up the information content of the sample. The only additional efforts are in using a slightly more complicated estimation algorithm.

A general efficiency measure for sampling designs in which different numbers of observations are collected is the D-error measure, which we discussed previously, multiplied

by the number of observations. This measure is reported in the last row of Table 3, and it indicates that the outcome-dependent sampling schemes are up to four times as efficient as the random sampling scheme.

Finally, the consequences of treating an endogenously selected sample as if it were a random sample can be observed from the last column in Table 3. When selection is based only on the endogenous variable, such as the sample in Column 2, only the intercept in the model is substantially influenced. When selection is also based on explanatory variables, the parameters for the variables related to the selection are also affected. To determine the extent to which this happens, Column 4 presents the estimation results for the same sample as that used in Column 3, but we estimate the logit parameters without proper correction.

The most striking difference between the corrected and uncorrected estimates is the negative (and significant) effect of a short relationship for the uncorrected parameter estimates, which seems rather implausible. Because of different estimates, predicted defection rates also differ substantially. For a customer with one insurance policy and average satisfaction, the predicted defection rate when the customer has a short relationship is 8.9% with the correct estimates and 12.2% with the incorrect estimates. This difference becomes much greater for the same customer with a long relationship, for which the correct predicted rate is 5.4%, whereas the incorrect estimates result in a prediction as high as 26.1%. Such differences in retention rates might have severe managerial implications if the firm uses these estimates to value existing customers or to evaluate the costs and benefits of customer acquisition programs.

CONCLUSION

Binary outcomes in real-life marketing research applications can be unevenly distributed over the two possible outcomes. In this article, we discussed how binary choice models can be consistently estimated on outcome-dependent samples. Such estimation techniques enable researchers to collect much smaller samples that contain sufficient information for precise model estimation. A requirement for the application of these techniques is that information is needed about the population distribution of the variables that are used for stratification. Often, this information is available in customer databases. If it is not available, estimates of the population totals combined with some sensitivity analysis could be used. For explanatory variables that are not used for stratification, in general, the corresponding parameter estimates will not vary much.

Our study, which uses synthetic data, indicates that gains of more than 60% can be achieved when one of the two outcomes is observed in only 2.5% of the population. Moreover, when the explanatory variables are unevenly distributed, stratification on both y and x is shown to yield efficiency gains of up to 75%.

We demonstrated the use of outcome-dependent sampling in combination with the required estimation technique on a sample of customers of an insurance company of which only a few customers leave the company. To obtain more information, additional customers who left the company in the last year were interviewed. These observations can only be

used when the estimation method corrects for this nonrandom sampling scheme. We show that a sample with a size of only 40% of the size of the random sample, which would include the additional observations, leads to more precise parameter estimates than the original random sample. To illustrate potential cost savings, if our data had been based on a telephone interview of about 15 minutes, total cost savings would have been as much as \$30,000. Further research could focus on generalizing the estimation methodology to ordered or unordered discrete outcomes.

REFERENCES

- Aaker, David A., V. Kumar, and George S. Day (2001), *Marketing Research*. Chichester, UK: John Wiley & Sons.
- Arora, Neelaj and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," *Journal of Consumer Research*, 28 (2), 273-83.
- Baum, Joel A.C. and Helaine J. Korn (1999), "Dynamics of Dyadic Competitive Interaction," *Strategic Management Journal*, 20 (3), 251-78.
- Bolton, Ruth N., P.K. Kannan, and Matthew D. Bramlett (2000), "Implications of Loyalty Program Membership and Service Experiences for Customer Retention and Value," *Journal of the Academy of Marketing Science*, 28 (1), 95-108.
- Breslow, Norman E. and Nicholas E. Day (1980), *Statistical Methods in Cancer Research*. Lyon, France: International Agency for Research on Cancer.
- Cosslett, Stephen R. (1993), "Estimation from Endogenously Stratified Samples," in *Handbook of Statistics*, Vol. 11, G.S. Maddala, C.R. Rao, and H.D. Vinod, eds. Amsterdam: Elsevier Science Publishers.
- Cramer, Mars, Philip Hans Franses, and Erica Slagter (1999), "Censored Regression Analysis in Large Samples with Many Zero Observations," *Econometric Institute Report EI-9939/A*, Erasmus University, Rotterdam.
- Frambach, Ruud T., Harry Barkema, Bart Nooteboom, and Michel Wedel (1998), "Adoption of a Service Innovation in the Business Market: An Empirical Test of Supply-Side Variables," *Journal of Business Research*, 41 (2), 161-74.
- Franses, Philip Hans and Richard Paap (2001), *Quantitative Models for Marketing Research*. Cambridge, UK: Cambridge University Press.
- Ganesh, Jaishankar, Mark J. Arnold, and Kristy E. Reynolds (2000), "Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers," *Journal of Marketing*, 65 (July), 65-87.
- Greene, William H. (2000), *Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall.
- Imbens, Guido W. and Tony Lancaster (1996), "Efficient Estimation and Stratified Sampling," *Journal of Econometrics*, 74 (2), 289-318.
- Kuhfeld, Warren F., Randall D. Tobias, and Mark Garrat (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, 31 (November), 545-57.
- Lancaster, Tony and Guido W. Imbens (1991), "Choice Based Sampling: Inference and Optimality," working paper, Department of Economics, Brown University.
- Lehmann, Donald R., Sunil Gupta, and Joel L. Steckel (1998), *Marketing Research*. Reading, MA: Addison-Wesley.
- Scott, Alastair J. and Chris J. Wild (1997), "Fitting Regression Models to Case-Control Data by Maximum Likelihood," *Biometrika*, 84 (1), 57-71.

