USING SELECTIVE SAMPLING FOR BINARY CHOICE MODELS TO REDUCE SURVEY COSTS

BAS DONKERS, PHILIP HANS FRANSES AND PETER VERHOEF

ERIM REPORT SERIES *RESEARCH IN MANAGEMENT*	
ERIM Report Series reference number	ERS-2001-67MKT
Publication	November 2001
Number of pages	24
Email address corresponding author	donkers@few.eur.nl
Address	Erasmus Research Institute of Management (ERIM) Rotterdam School of Management / Faculteit Bedrijfskunde Erasmus Universiteit Rotterdam P.O. Box 1738 3000 DR Rotterdam, The Netherlands Phone: +31 10 408 1182 Fax: +31 10 408 9640 Email: info@erim.eur.nl Internet: www.erim.eur.nl

Bibliographic data and classifications of all the ERIM reports are also available on the ERIM website: www.erim.eur.nl

ERASMUS **R**ESEARCH **I**NSTITUTE OF **M**ANAGEMENT

REPORT SERIES

RESEARCH IN MANAGEMENT

BIBLIOGRAPHIC DATA AND CLASSIFICATIONS		
Abstract	Marketing problems sometimes concern the analysis of dichotomous variables, like for example ``buy'' and ``not buy'' and ``respond'' and ``not respond''. It can happen that one outcome strongly outnumbers the other, for example when many households do not respond (to a direct mailing, for example). Standard econometric methods would imply the collection of many data to obtain precise estimates and this can be rather costly. To cut back costs, we propose to implement a non-random sampling scheme and to correct for the subsequent sample selection bias in the econometric model. In this paper we put forward the relevant method, which does not lead to a loss in precision. Our illustration suggests an opportunity to collect 60\% less data points.	
Library of Congress Classification (LCC)	5001-6182	Business
	5410-5417.5	Marketing
	HA 31.2	Sampling: social sciences
Journal of Economic Literature (JEL)	M	Business Administration and Business Economics
	M 31 C 44	Marketing Statistical Decision Theory
	C 19	Econometric and Statistical Methods: Other
European Business Schools Library Group (EBSLG)	85 A	Business General
	280 G 255 A	Managing the marketing function Decision theory (general)
	250 D	Statistical Analysis
Gemeenschappelijke Onderwerpsontsluiting (GOO)		
Classification GOO	85.00	Bedrijfskunde, Organisatiekunde: algemeen
	85.40 85.03	Marketing Methoden en technieken, operations research
	83.03	Methoden en technieken
Keywords GOO	Bedrijfskunde / Bedrijfseconomie	
	Marketing / Besliskunde	
	Steekproeven, Ecometrische modellen	
Free keywords	Outcome-dependent sampling, sample size, survey design, survey costs, binary outcomes, logit model	

Using Selective Sampling for Binary Choice Models to Reduce Survey Costs*

Bas Donkers[†]

Philip Hans Franses[‡]

Peter Verhoef[§]

*We thank Michel Wedel and seminar participants at the Tinbergen Institute, Tilburg University and the 2001 Marketing Science Conference in Wiesbaden for helpful comments.

[†]Corresponding author, Department of Marketing, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, e-mail: donkers@few.eur.nl, Phone +31-10-4082411, Fax +31-10-4089169.

[‡]Department of Marketing and Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, e-mail: franses@few.eur.nl, Phone +31-10-4081273.

[§]Department of Marketing, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, e-mail: verhoef@few.eur.nl, Phone +31-10-4082809.

Using Selective Sampling for Binary Choice Models to Reduce Survey Costs

Abstract

Marketing problems sometimes concern the analysis of dichotomous variables, like for example "buy" and "not buy" and "respond" and "not respond". It can happen that one outcome strongly outnumbers the other, for example when many households do not respond (to a direct mailing, for example). Standard econometric methods would imply the collection of many data to obtain precise estimates and this can be rather costly. To cut back costs, we propose to implement a non-random sampling scheme and to correct for the subsequent sample selection bias in the econometric model. In this paper we put forward the relevant method, which does not lead to a loss in precision. Our illustration suggests an opportunity to collect 60% less data points.

Keywords: Outcome-dependent sampling, sample size, survey design, survey costs, binary outcomes, logit model.

Introduction

Dichotomous outcomes and their determinants, such as market entry ("shall we enter a market or not?"), customer retention ("does the customer stay with us or not?"), or the adoption of a new product ("does the household adopt our product, yes or no?"), are commonly studied in marketing research, see, for example, Baum and Korn (1999), Bolton, Kannan and Bramlett (2000) and Frambach et al. (1998). In empirical work it can happen that one outcome is much more often observed than the other outcome. Examples in marketing research concern low defection rates within industries such as financial services, low response rates to direct mailings and low adoption rates of new products. In studying these dichotomous outcomes one is usually interested in the antecedents ("what causes a customer to leave us?") or in the prediction of the event ("how many households will adopt our product?"). In order to estimate models relating the antecedents to these outcomes or to fit a prediction model, marketing researchers do need a sufficient number of observations, also on the infrequently observed outcome. To gather these observations, researchers will need large samples, at least when they use a random sampling scheme. Consider the following example. Assume that the incidence of the infrequently occurring outcome in the population of interest is 5 percent and suppose that the researcher would want to have at least 100 observations on each outcome. Using a random sampling scheme and assuming no response bias, the sample should then contain at least 2000 observations. A telephone interview of approximately 15 minutes length would then cost approximately $ 60,000 (Aaker, Kumar and Day 2001, p. 237). However, there are ways to reduce survey costs by using smaller samples and relying on modified methods for the analysis. This is the topic of the present paper.

A reduction in sample size could be achieved by selecting respondents based on the outcome of interest. To continue the example above, this permits the researcher to collect information on fewer than the 1900 observations with the frequently observed outcome. Such a sampling scheme is called outcome-dependent sampling,

which is some form of stratified sampling, where the strata are also related to the dependent variable. For example, in customer defection studies, the switchers could be selected more often (Ganesh, Arnold and Reynolds 2000). However, such a selection clearly invalidates the standard assumptions necessary for parameter estimation in binary choice models, like the logit and probit model (Franses and Paap 2001; Greene 2000). Additional problems occur if respondents are not only selected on the outcomes but also on infrequently occurring values of explanatory variables. For example, when considering new product adoption of high technology products in consumer markets, such as cellular telephones, one would probably overrepresent young consumers in the sample.

Fortunately, there do exist estimation methods that correct for selective sampling as discussed above. Most of these estimation methods have been developed in biometrics, where information acquisition often includes expensive testing. A survey of the issues involved is presented in Cosslett (1993), while Imbens and Lancaster (1996) and Scott and Wild (1997) discuss two recently developed estimation techniques. In this paper, we describe how marketing researchers can use these methodologies to correct for outcome-dependent sampling. Moreover, we describe how this methodology can be used to reduce survey cost significantly, which is in the interest of marketing researchers in both academia and practice. Our main proposal concerns implementing a non-random sampling scheme and to apply a corrective estimation method to account for the subsequent sample selection bias.

This paper contributes to the marketing research literature as follows. Although sampling is of particular interest for marketing researchers, scanning the top marketing journals in the last two decades does not reveal many studies on this issue. In their overview of sampling issues, Sudman and Blair (2001) only refer to studies that have appeared more than a decade ago. The reference list of chapters on sampling methods in standard marketing research textbooks (like, Aaker, Kumar and Day 2001) reveals the same phenomenon. Nowadays companies measure variables for huge samples (see Mittal and Kamakura 2001), and the costs of gathering these data can still be quite substantial. Moreover, it may also be very time-consuming to

estimate models based on these large datasets (Blattberg, Glazer and Little 1994). Researchers are generally reluctant to leave out data as it reduces efficiency. However, the method we propose in this paper, reduces sampling costs significantly and does not lead to a loss in efficiency.

The structure of our paper is as follows. In the next section, we describe how outcome-dependent sampling is related to more familiar sampling schemes, such as random sampling and stratified sampling. Then an estimation technique that can handle outcome-dependent sampling is described and its implications for more efficient sampling schemes are discussed. Next, the proposed methodology is applied to the analysis of antecedents of customer retention. Our results show that the described methodology allows for a 60% reduction in sample size, which may lead to a significant reduction in costs. We end with a short summary and the limitations of the proposed methodology.

On selective sampling

Within sampling theory one generally distinguishes between probability and non-probability sampling methods (Lehmann, Gupta and Steckel 1998). Examples of probability sampling methods are simple random sampling, stratified sampling and cluster sampling. Outcome-dependent sampling can be considered as a special case of stratified sampling methods. Within stratified sampling, strata or segments are defined which have different characteristics. Usually, stratified sampling methods are based on characteristics that may affect an outcome. Stated differently, strata are usually defined on the X variables, which is called exogenous stratification. For example, in the case of a financial service provider, researchers may stratify their sample on variables, such as relationship age and the number of services purchased, which may affect customer retention (Verhoef, Franses and Hoekstra 2001).

Instead of stratifying on the X variables, outcome-dependent sampling stratifies the sample on the Y-variable, so stratification is endogenous. For example, a researcher studying factors that affect adoption of mobile phones could stratify

his/her sample on the adoption decision. Stratified sampling can be classified into proportionate and disproportionate stratified sampling schemes. With proportionate stratified sampling, each stratum is proportional to the size of the stratum. In disproportionate sampling the size of the stratum is based on something other than the sample size (Lehmann, Gupta and Steckel 1998). A well-known example of disproportionate sampling in marketing research is the retailpanel of AC Nielsen, in which large stores are overrepresented and small stores are underrepresented.

Outcome-dependent sampling can also be proportionate and disproportionate. In this paper we focus on disproportionate stratified sampling, where the dependent variable is binary in nature. We limit our scope for two reasons. First, when the dependent variable is continuous, the analysis of outcome-dependent samples depends heavily on distributional assumptions. Second, a proportionately stratified sample is a special case of a disproportionate sample. Moreover, substantial efficiency gains can only be expected from disproportionately stratified samples. Especially when the dependent variable has unevenly distributed outcomes, large efficiency gains can be realized by overrepresenting the infrequently observed outcome in the sample. These efficiency gains may decrease sample size and survey costs dramatically.

As might be expected, outcome-dependent sampling affects the estimation results of statistical models. Fortunately, in case of binary outcomes, the consequences of outcome-dependent sampling are well understood and we will discuss these issues in the next section. Besides stratifying on the X- or Y-variables, a researcher can also stratify on both the X and the Y variables simultaneously. This can be done in a proportionate or disproportionate way. A disproportionate stratification on both the X- and the Y variables is recommended, when both the X- and the Y-variables are unevenly distributed and the X-variable(s) has a substantial impact on the Y-variable. We discuss this sampling strategy with two illustrative examples. The first example concerns the adoption of mobile phones in the beginning of the product life cycle. For this product it is well known that market penetration is high among young individuals. When data are collected through a survey on a random sample from the population, it might turn out that only few young respondents

do not have a mobile phone. This suggests that more information on young individuals who do not have a mobile phone will facilitate the identification of other drivers than age for buying a mobile phone. Collecting this additional information using a random sampling scheme would be costly, while also only a small part of the additional observations would provide the relevant information. In case of such a skewed distribution, a random sample of only those young individuals, who do not own a mobile phone, will be small (and cheap), while it contains about the same information. The second example concerns our illustration below, which is based on real-life data. An insurance company is interested in the drivers of customer retention, as this is seen as one of the main drivers of customer lifetime value (Rust, Zeithaml and Lemon 2000). In general, retention rates are rather high 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. As it turns out, both customer defection and relationship age are unevenly distributed within the total customer database. Even when defectors are overrepresented, only a small number of customers with relationships shorter than 1 year who leave the company would be observed. This would complicate the estimation of the effect of relationship age and thus also the effect of other variables. Hence, again the researcher would benefit from overrepresenting defecting customers and customers with short relationships.

Insert Figure 1 about here

Based on this overview, a classification of stratified sampling methods can be made which is new to marketing research. This classification is given in Figure 1. This scheme clearly shows the wide variety of sampling methods available to the researcher. In the left column of the figure we have the regularly described sampling schemes where sampling on the dependent variable is random. The two columns on the right describe the endogenously stratified sampling that we propose to use. In particular the third column is of interest, as here large efficiency gains might be

possible. For future reference we designate the cells in this column with A, B and C. The use of proportionate stratification for the dependent variable, the middle column in Figure 1, complicates estimation as much as disproportionate stratification, while efficiency gains will be much smaller. Proportionate stratification on Y is therefore not very interesting.

In the next section we discuss in detail the estimation methodology that is needed to correct for outcome-dependent sampling schemes for binary choice models. First we give a short description of the consequences of disproportionate outcome-dependent sampling for parameter estimation in the logit model. A detailed, but also more technical description of the consequences of outcome-dependent sampling schemes, is given in Cosslett (1993). The most important consequence of an outcome-dependent sampling, scheme where the infrequently observed outcome is overrepresented, is that when estimation is done without correcting, the probability of observing the rare outcome is overestimated. For the logit model this implies that disproportionate outcome-dependent sampling in combination with random or proportionately stratified sampling on X, cases A and B in Figure 1, results in a bias in the intercept. For other binary choice models, such as the probit model, all parameter estimates are affected. When there is disproportionate sampling on X and Y, case C in Figure 1, two situations have to be distinguished. When the stratification of X and Y is independent, again only the intercept will be affected. However, when stratification is on combinations of X and Y, such as on young individuals without a mobile phone, all parameter estimates will be affected.

Methodology

In this section we discuss how full maximum likelihood (ML) estimates can be obtained from a sample that is stratified on the variable of interest and possibly on exogenous variables. The methodology therefore treats the sampling schemes in the most right column of Figure 1. To estimate the parameters using ML, we have to determine the likelihood function of the data, given information on the sampling

scheme that is used.

We first introduce some notation and other preliminaries. We then turn to discussing the likelihood of binary choice models when the sampling scheme for data collection is outcome dependent. We start with the situation where stratification on x is allowed, but assumed independent of the stratification on y. We focus on the case of random sampling on x (case A). The non-randomness that results from stratification on x can be corrected for by a weighting scheme, see, for example, Nowell and Stanley (1991). We continue with the generalization of the derived likelihood to the case where stratification is based on combinations of x and y (cases B and C). The section ends with a discussion on how matters simplify when the logit model is used to model the binary outcome and some practical implications for sampling schemes.

Preliminaries

We only consider the case of dichotomous outcomes, so the dependent variable y_i takes the value of 1 or 0. Hence, 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, $s = 1, \ldots, S$. These strata can be based on only a subset of the explanatory variables. For example, one could stratify on gender and age, without restricting the other variables. The strata are allowed to overlap. An example of overlapping strata results from the following sampling scheme. First a random sample of the population is drawn, so this stratum contains all values on x. An additional sample is based on a part of the population that one wants to overrepresent, so in this stratum certain values of x are excluded. Here one can think of the overrepresentation of young individuals in the analysis of mobile phone usage. More so, when only few young people do not use a mobile phone, one may even want to collect additional observations on these individuals. These two sampling schemes fit in the cell C of Figure 1, where they differ on whether the strata are defined independently for x_i and y_i or not.

The sampling strategies are presented within the following setup. There is a

finite population of N individuals that represent a random sample from the joint distribution of (y_i, x_i), where x_i is a vector of exogenous explanatory variables. All that is known about the finite population is that there are N_{sj} individuals with $y_i = j$ in stratum s, $j = 0, 1$ and $s = 1, \ldots, S$. Within each stratum of the x's, a random sample of size $n_{sj} \leq N_{sj}$ is drawn from the stratummembers with $y_i = j$. The data consist of y_i and x_i which are recorded for these observations.

Selection on y_i

The simplest endogenous sampling scheme is where first the response category, $y_i = 0$ or $y_i = 1$, is selected randomly with probabilities n_0/n and n_1/n, and then the observation is selected randomly from the subpopulation with y_i in the response category that is chosen. 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 $P^*(y_i = j|x_i)$ the same probability for the population of interest. Of course, $P^*(y_i = j_i|x_i)$ is what we are interested in to learn about. It can be shown that under this sampling scheme the following equality holds

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

with

$$\mu_j = (n_j/n)/P^*(y = j) \tag{2}$$

the ratio of the probability that an observation falls into class j in the sample, n_j/n, and the probability that an observation falls into class j in the population, $P^*(y = j)$. The probability of observing $y_i = j$ in the data is larger than in the population when observations with $y_i = j$ are sampled with a larger probability than with which they occur in the population, that is, $P(y = j|x) \geq P^*(y = j|x)$ whenever $(n_j/n) \geq P^*(y = j)$.

The major advantage of this simple 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, one can treat the sample

with predetermined group sizes as if it was constructed with the alternative sampling scheme, based on conditioning on sufficient statistics. As the likelihood is not based on the actual sampling scheme, this likelihood is called the pseudo-likelihood. ML estimation based on the pseudo-likelihood always results in consistent estimates. For the familiar logit model the resulting parameter estimates are consistent and efficient. When one would estimate the parameters without correcting for the outcome-dependent sampling scheme, all parameter estimates will be biased, except for the logit model, where only the intercept is affected.

Selection on y_i and x_i simultaneously

The pseudo-likelihood for the sampling scheme where selection depends on the outcome in combination with the strata of the exogenous variables is now straightforward to derive. As the samples from the different strata are drawn independently, the total pseudo-likelihood of all observations equals the product of the pseudo-likelihoods of the different strata. For each stratum individually, we can obtain the likelihood using (1), where the population of interest is the stratum under consideration. The equivalent of (1) when the population of interest is a single stratum is

$$P(y_i = j|x_i, \text{stratum}_i = s) = \frac{\mu_{sj} P^*(y_i = j|x_i)}{\mu_{s0} P^*(y_i = 0|x_i) + \mu_{s1} P^*(y_i = 1|x_i)}, \tag{3}$$

The full sample pseudo-likelihood is

$$\begin{aligned} L(\theta) &= \prod_{s=1}^{S} \prod_{j=0}^{1} \prod_{i=1}^{n_{sj}} P(y_i = j|x_i, \text{stratum}_i = s) \\ &= \prod_{s=1}^{S} \prod_{j=0}^{1} \prod_{i=1}^{n_{sj}} \frac{\mu_{sj} P^*(y_i = j|x_i)}{\mu_{s0} P^*(y_i = 0|x_i) + \mu_{s1} P^*(y_i = 1|x_i)}, \end{aligned} \tag{4}$$

In case we use the logit model, maximization of $L(\theta)$ results in consistent estimates, but not necessarily in efficient estimates. Efficient estimates are obtained when there is a stratum-specific intercept for each stratum of x_i in the data, which will often be the case. When this is not the case, efficient estimates can be obtained from an iterative procedure, see Scott and Wild (1997). Notice also that even though

within a stratum certain x's might be fixed, their effects will be identified from the variation across the different strata.

The logit model

So far we have presented a general formulation of the likelihood. However, when the logit model is used, the model can be easily estimated for different sampling schemes using various statistical packages. To see how this is implemented, consider the probability of observing $y_i = 1$ in the standard logit model (Greene 2000):

$$P^*(y_i = 1|x_i) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} \tag{5}$$

To obtain the pseudo-model in (3) we need to correct this probability with parameters μ_{s0} and μ_{s1}, resulting in

$$\begin{aligned} &P(y_i = 1|x_i, \text{stratum}_i = s) = \\ &= \frac{\mu_{s1} \exp(x_i'\beta)/[1 + \exp(x_i'\beta)]}{\mu_{s0}/[1 + \exp(x_i'\beta)] + \mu_{s1} \exp(x_i'\beta)/[1 + \exp(x_i'\beta)]} \\ &= \frac{\mu_{s1} \exp(x_i'\beta)}{\mu_{s0} + \mu_{s1} \exp(x_i'\beta)} \\ &= \frac{\mu_{s1}/\mu_{s0} \exp(x_i'\beta)}{1 + \mu_{s1}/\mu_{s0} \exp(x_i'\beta)} \end{aligned} \tag{6}$$

Let $\hat{\mu}_{s0} = n_{s0}/N_{s0}$ and $\hat{\mu}_{s1} = n_{s1}/N_{s1}$ denote estimates of μ_{s0} and μ_{s1}, then the probability in the pseudo-likelihood is

$$\begin{aligned} P(y_i = 1|x_i, \text{stratum}_i = s) =& \frac{\hat{\mu}_{s1}/\hat{\mu}_{s0} \exp(x_i'\beta)}{1 + \hat{\mu}_{s1}/\hat{\mu}_{s0} \exp(x_i'\beta)} \\ =& \frac{\exp(x_i'\beta + \ln(\hat{\mu}_{s1}/\hat{\mu}_{s0}))}{1 + \exp(x_i'\beta + \ln(\hat{\mu}_{s1}/\hat{\mu}_{s0}))} \end{aligned} \tag{7}$$

The pseudo-likelihood estimator of the logit model is therefore easily obtained by adding a correction of $\ln(\hat{\mu}_{s1}/\hat{\mu}_{s0})$ to the index $x_i'\beta$ for each observation. This can be done in standard statistical packages that can include offsets or estimate models under parameter restrictions, allowing the parameter of the correction to be fixed.

So far we have discussed parameter estimation and it turns out that when the logit model is used, efficient estimates are obtained straightforwardly. However,

without standard errors these parameter estimates are difficult to interpret. The regular standard errors that result from maximum likelihood estimation based on the pseudo-likelihood can be used, but they are conservative, that is, the true standard errors will be smaller. In the technical appendix we show how exact standard errors can be computed, again in the case where there are stratum-specific constants in the model.

Practical considerations

The optimal ratio of observations with $y_i = 1$ and $y_i = 0$ depends on the application at hand. Theoretical and simulation results concerning the optimal sample composition are presented by Lancaster and Imbens (1991) and indicate that an equal split of the sample into 50%-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 1's.

These results can be understood as follows. As information on the relationship of interest is based on the simultaneous variation of the dependent and the 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 $F_n(y_i = 1) = 0.5$. Splitting the sample in equal parts also makes sense by symmetry considerations, as there is no *a priori* reason to have more observations from one type than from the other.

Illustration

In this section we illustrate our methodology on real-life data in marketing research. We will show that selective sampling, followed by an appropriate estimation method, enables one to interview 60% less individuals.

Data

The data 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 intermediairies. They sell all types of insurance policies, ranging from fire and theft insurance to life insurance. The company aims at having close relationships with its customers. As the company does not have intermediaries who can signal customer dissatisfaction, the company is highly interested in the determinants of customer satisfaction and the role customer satisfaction plays in a customer's decision to leave the company. Moreover, they are especially interested in the behavior of customers who are with the company for less than one year, as these customers apparently leave the company more often than customers with longer relationships (7.6% versus 3.5%).

An exogenously stratified sample of 2300 customers has been collected in 1999 to obtain information about customer satisfaction. These customers have been reinterviewed in 2000 and for 1374 customers we observe answers to the questions we need. From these 1374 customers only 53 customers (3.9%) have left the company since the first interview in 1999. As the company realized that this was only a rather small number of observations on inactive customers to perform statistical analysis, 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 twelve 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 1. Short_Duration is a dummy variable indicating 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, measured on a single 5-point Likert scale.

Insert Table 1 about here

In this illustration 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. This higher defection rate shows up in the average of the dummy variable Short_Duration in Table 1, which is higher for the customers who have left the company.

Estimation results for various sampling schemes

The total sample of observations consists of a random sample and a non-random sample of customers that have left the insurance company. To get some feeling for the consequences that outcome-dependent sampling may have on the precision of the estimated coefficients, we present in Table 2 the estimation results of a logit model with the decision to leave the company as the dependent variable and satisfaction and a dummy variable indicating a short relationship as explanatory variables. The different columns in this table represent different sampling schemes that (could) have been used. The first row presents the number of observations that is used for estimation. The results when the outcome-dependent sampling scheme is not corrected for are presented in Table 3, but first we discuss the correct estimation results.

Insert Table 2 about here

Column (1) in Table 2 presents the estimation results of the logit model for the original sample, which is a random sample from the population of interest. Both explanatory variables have highly significant parameter estimates, so the amount of information contained in the original sample is sufficient to obtain statistically significant estimates of the parameters of interest. To see what happens when we add additional observations, we re-estimate the parameters for three other samples. First, we investigate what the consequences are when the company had focused on defecting customers with short relationship durations. In the additional sample

there are six such observations. The parameter estimates that result when these six observations are added are presented in column (2) of the table. This sampling scheme fits into cell C of Figure 1, where selection is on x and y. Notice that to obtain consistent estimates of all parameters, we of course need to use the estimation techniques we discussed. This results in a substantial improvement of the estimation precision of the differential defection rate in the first year.

The company, however, did not focus especially on defecting customers with short durations, but on all defecting customers, so the 30 additional observations that are collected are a random sample of the defecting customers. Column (3) presents the estimation results when all additional observations are included. This sampling scheme fits into cell A of Figure 1, as there is no selection on x. Finally, we consider all (53+30) customers who leave the company and a random sample of 466 customers who did stay with the company. This sample mimics an sample of 40% of the size of the original sample that could result from an outcome-dependent sampling scheme and results in more efficient estimates than the original random sample. This sampling scheme also fits in cell A of Figure 1.

The estimation results in Table 2 are all corrected for the outcome-dependent sampling schemes that are used. In this table, column (2) presents the estimation results of the model on the original sample and the customers in the additional sample who have a short relationship. Thus, there are only 6 observations added to the original sample of 1374 observations. The standard errors of the parameter estimates when these 6 observations would have been randomly drawn are on average $\sqrt{1374/(1374+6)} = 0.998$ times the original ones, which amounts to a decrease of 0.2%. However, what we observe from Table 2 is a decrease of 14% in the estimated standard error for the effect of Short_Duration, which would be equivalent to an increase in the size of a random sample with 30%. Also the precision of the satisfaction parameter increases with more than 0.2%. Naturally, the large gain in efficiency of the effect of Short_Duration could have been expected, as the observations that were added are particularly informative about the effect of this variable.

Column (3) presents the estimation results based on the original and additional sample combined. These are the most efficient estimates we can obtain with our data, as they are based on all the available observations. Comparing with the original sample, one would expect a decrease in all standard errors of about 1% when the additional observations would have constituted a random draw from the whole population. Compared to the results in column (2), we now see that the precision on the effect of satisfaction has also increased substantially. Compared with the random sample, standard errors all have decreased by more than 16%.

So far we have discussed the gains in efficiency that can be achieved by adding endogenously sampled observations. However, the most important question is whether we can cut back on survey costs without loosing estimation precision, by using outcome-dependent sampling schemes. This is what we investigate in column (4), where we use a sample with a size that is only 40% of the size of the original sample. Although the sample size is substantially smaller than the original sample, we see that the precision of all parameters is higher than the precision from the parameter estimates in the original sample. The company could therefore have saved about 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.

The estimation results in Table 2 are based on the estimation algorithm that corrects for outcome-dependent sampling. However, what are the consequences of treating an endogenously selected sample as if it were a random sample? When selection is only based on the endogenous variable, only the intercept in the model would be substantially influenced. When selection is also based on explanatory variables, the parameters for the variables related to the selection are also affected. Table 3 presents the estimation results for each of the samples in Table 2 without correcting for the nonrandom sampling scheme, so estimation is done using the regular logit model. As the estimation methodology that corrects for outcome-dependent sampling also uses information on population frequencies, some small differences occur in all the parameter estimates and their standard deviations. However, we will focus

on the systematic differences between the parameter estimates in Tables 2 and 3.

Insert Table 3 about here

For the parameter estimates from the random sample presented in column (1), there is clearly no effect of the correction. For the other columns we observe substantial and predictable changes in the parameter estimates. The largest changes in the parameter estimates are observed in the estimates of the intercept. As we go more to the right in Table 3, the proportion of customers in the sample that leave the company increases. Consequently, without correcting for this the estimated intercept increases. As there is no selection on customer satisfaction, there are no large changes in the estimated effect, at least when the changes are compared to the standard error or the estimate. For the effect of short relationships on the probability to leave the company, there is a clear difference between the parameter estimate in column (2) and the other columns. This stems from the overrepresentation of the defecting customers with a short relationship. As there are more observations with a short relationship that leave the company in this sample, the effect of having a short relationship duration is highly overestimated. The difference between the parameter estimates in Table 2 and 3 is about $\ln(\frac{12+6}{12})$, which corresponds to the correction that is needed.

When such inappropriate parameter estimates are used for predicting customer defection probabilities, the predicted probabilities are inadequate. The predicted probability of a customer leaving the company in his first year, when he has average satisfaction, is 6.5% when the model is estimated correctly. In contrast with this, when we use the incorrect parameter estimates in column (2) of Table 3, it is 9.1%, or almost 50% higher. Such a difference might have severe implications when the firm would use these estimates to choose between a customer acquisition program and a customer retention program. The value of an acquired customer would be highly underestimated, as the probability that this customer leaves the company is substantially overestimated.

Conclusion

Binary outcomes in real-life marketing research applications can be unevenly distributed over the two possible outcomes. In this paper we discussed how binary choice models can be consistently estimated on samples that have been constructed based on the realizations of the outcome variable, resulting in what is called an endogenously stratified or outcome-dependent sample. Such estimation techniques allow researchers to collect smaller samples that still contain sufficient information for precise estimation of the model of interest.

We illustrated the use of outcome-dependent sampling in combination with the required estimation technique on a sample of customers of an insurance company, where only few customers leave the company. Only 3.9% of the customers leave the company in the random sample that has been collected. To obtain more information, additional customers that left the company in the last year have been interviewed. These observations can only be used when the estimation method corrects for this non-random sampling scheme. We show that these additional observations result in an increase in estimation precision that exceeds the expected gain in precision of additional random observations by far. The large reduction in the amount of data needed, which is possible when outcome-dependent sampling schemes are used, becomes clear when we construct a sample that consists of only 40% of the number of observations from the originally obtained random sample. The parameter estimates based on this sample are more precise, even though data collection costs are reduced with about 60%. When our data would have been based on a telephone interview of about 15 minutes, total costs savings would have been as much as $ 25,000.

Future research could focus on generalising the estimation methodology to ordered or unordered discrete outcomes.

Appendix

In this technical appendix we present the formulae to obtain the correct standard errors for the parameter estimates, based on Scott and Wild (1997). The correct covariance matrix of the parameter estimates will be derived for the situation with stratum-specific constants. We refer to Scott and Wild (1997) for the more general case.

For the regular ML estimator, the covariance matrix can be obtained from the matrix of second order derivatives of the log-likelihood. When $\mathcal{T}$ denotes the matrix of second order derivatives of the log likelihood, then $-\mathcal{T}^{-1}$ is a consistent estimate of the covariance matrix. For our pseudo maximum likelihood estimator, the correct standard errors will also be based on the matrix of the second order derivatives of the pseudo likelihood, evaluated at the parameter estimates.

Let $\mathcal{T}$ denote minus the matrix of second order derivatives of the pseudo likelihood, evaluated at the parameter estimates, $\hat{\theta}$

$$\mathcal{T} = -\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta-\hat{\theta}}$$

Let $\mathrm{Var}_{est}(\hat{\theta})$ denote a consistent estimator for the covariance matrix of the parameter estimates, $\hat{\theta}$. This covariance matrix is now obtained as follows. First define $\mathrm{Var}_{est}(\hat{\theta}) = \mathcal{T}^{*-1}$, then $\mathcal{T}^*$ equals

$$\mathcal{T}^* = \mathcal{T} + \sum_{s=1}^{S} \frac{B_s B_s'}{K_s},$$

with

$$B_s = \sum_{x_i \in s} \sum_{j=0,1} \frac{\partial P^*(y_i = j|x_i)}{\partial\theta} I(y_i = j) \text{ and } K_s = A_s^{-1} - W_s,$$

where

$$W_s = \sum_{x_i \in s} P^*(y_i = 1|x_i)\{1 - P^*(y_i = 1|x_i)\},$$

$$A_s = \left(\frac{1}{n_{s0}} + \frac{1}{n_{s1}}\right) - \left(\frac{1}{N_{s0}} + \frac{1}{N_{s1}}\right)$$

Table 1: Descriptive statistics for explanatory variables.

	Original sample		Additional sample
	Active	Inactive	Inactive
N	1321	53	30
Short_Duration (Mean)	0.12	0.23	0.20
Satisfaction (Mean)	3.4	3.1	3.2

Table 2: Estimation results for four different sampling schemes.

Standard errors in parentheses.

	(1)	(2)	(3)	(4)
N	1374	1380	1404	549
Intercept	-0.453	-0.470	-0.648	-0.577
	(0.812)	(0.793)	(0.696)	(0.807)
Satisfaction	-0.903	-0.898	-0.850	-0.894
	(0.252)	(0.246)	(0.215)	(0.246)
Short_Duration	0.699	0.672	0.696	0.684
	(0.335)	(0.294)	(0.278)	(0.311)

(1): Original sample

(2): Original sample and only short durations from additional sample

(3): Original sample and complete additional sample

(4): Sample including all customers leaving but only 40% of the size of the original sample

Table 3: Uncorrected estimation results for four different sampling schemes.
Standard errors in parentheses.

	(1)	(2)	(3)	(4)
N	1374	1380	1404	549
Intercept	-0.453	-0.213	0.073	0.238
	(0.812)	(0.795)	(0.685)	(0.807)
Satisfaction	-0.903	-0.958	-0.903	-0.948
	(0.252)	(0.247)	(0.212)	(0.246)
Short_Duration	0.699	1.162	0.697	0.633
	(0.335)	(0.298)	(0.282)	(0.314)

(1): Original sample

(2): Original sample and only short durations from additional sample

(3): Original sample and complete additional sample

(4): Sample including all customers leaving but only 40% of the size of the original sample

Figure 1:

Classification of Stratified Sampling Schemes

			Sampling on Y		
			Random	Stratification (*Outcome Dependent Sampling*)	
				Proportionate	Disproportionate
Sampling on X	Random				A
	Stratification	Proportionate			B
		Disproportionate			C

References

Aaker, David A., V. Kumar and George S. Day (2001), Marketing Research. Chischester: John Wiley & Sons

Baum, Joel A.C. and Helaine J. Korn (1999), Dynamics of Dyadic Competitive Interaction, Strategic Management Journal, 20, 251-278

Blattberg, Robert C., Rashi Glazer and John D.C. Little (1994), The Marketing Information Revolution. Boston: Harvard Business School Press

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, 95-108

Breslow, Norman E. and Nicholas E. Day (1980), Statistical Methods in Cancer Research. IARC, Lyon

Cosslett, Stephen R. (1993), Estimation from Endogenously Stratified Samples, in G.S. Maddala, C.R. Rao, and H.D. Vinod (eds), Handbook of Statistics, vol 11, 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: 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: Prentice Hall

Imbens, Guido W. and Tony Lancaster (1996), Efficient Estimation and Stratified Sampling, Journal of Econometrics, 74, 289-318

Lancaster, Tony, and Guido W. Imbens (1991), Choice Based Sampling: Inference and Optimality, Department of Economics Working Paper, Brown University, Providence

Lehmann, Donald R., Sunil Gupta and Joel L. Steckel (1998), Marketing Research, Addison Wesley, Reading, Massachusetts

Mittal, Vikas and Wagner A. Kamakura (2001), "Satisfaction, Repurchase Intent and Repurchase Behavior: Investigating the Moderating Effect of Customer Characteristics", Journal of Marketing Research, 38 (1), 131-42

Nowell, Clifford and Linda R. Stanley (1991), Length-Biased Sampling in Mall Intercept Surveys, Journal of Marketing Research, 28, 475-479

Rust, Roland T., Valarie A. Zeithaml and Katherine N. Lemon (2000), Driving Customer Equity, How Customer Lifetime Value is Reshaping Corporate Strategy, Free Press, New York

Scott, Alastair J., and Chris J. Wild (1997), Fitting Regression Models to Case-Control Data by Maximum Likelihood, Biometrika, 84, 57-71

Sudman, Seymour and Edward Blair (2000), Sampling in the Twenty-First Century, Journal of the Academy of Marketing Science, 27 (2), 269-77

Verhoef, Peter C., Philip Hans Franses and Janny C. Hoekstra (2001) The Impact of Satisfction and Payment Equity on Cross Buying: A Dynamic Model for a Multi-Service Provider, Journal of Retailing, 77 (3), 359-78

Publications in the Report Series Research* in Management

ERIM Research Program: "Marketing"

2001

Predicting Customer Potential Value. An application in the insurance industry
Peter C. Verhoef & Bas Donkers
ERS-2001-01-MKT

Modeling Potenitally Time-Varying Effects of Promotions on Sales
Philip Hans Franses, Richard Paap & Philip A. Sijthoff
ERS-2001-05-MKT

Modeling Consideration Sets and Brand Choice Using Artificial Neural Networks
Björn Vroomen, Philip Hans Franses & Erjen van Nierop
ERS-2001-10-MKT

Firm Size and Export Intensity: A Transaction Costs and Resource-Based Perspective
Ernst Verwaal & Bas Donkers
ERS-2001-12-MKT

Customs-Related Transaction Costs, Firm Size and International Trade Intensity
Ernst Verwaal & Bas Donkers
ERS-2001-13-MKT

The Effectiveness of Different Mechanisms for Integrating Marketing and R & D
Mark A.A.M. Leenders & Berend Wierenga
ERS-2001-20-MKT

Intra-Firm Adoption Decisions: Departmental Adoption of the Common European Currency
Yvonne M. van Everdingen & Berend Wierenga
ERS-2001-21-MKT

Econometric Analysis of the Market Share Attraction Model
Dennis Fok, Philip Hans Franses & Richard Paap
ERS-2001-25-MKT

Buying High Tech Products: An Embeddedness Perspective
Stefan Wuyts, Stefan Stremersch & Philip Hans Franses
ERS-2001-27-MKT

Changing Perceptions and Changing Behavior in Customer Relationships
Peter C. Verhoef, Philip Hans Franses & Bas Donkers
ERS-2001-31-MKT

How and Why Decision Models Influence Marketing Resource Allocations
Gary L. Lilien, Arvind Rangaswamy, Katrin Starke & Gerrit H. van Bruggen
ERS-2001-33-MKT

---

* A complete overview of the ERIM Report Series Research in Management:
http://www.ers.erim.eur.nl

ERIM Research Programs:
LIS Business Processes, Logistics and Information Systems
ORG Organizing for Performance
MKT Marketing
F&A Finance and Accounting
STR Strategy and Entrepreneurship

An Equilibrium-Correction Model for Dynamic Network Data
David Dekker, Philip Hans Franses & David Krackhardt
ERS-2001-39-MKT

Aggegration Methods in International Comparisons: What Have We Learned?
Bert M. Balk
ERS-2001-41-MKT

The Impact of Channel Function Performance on Buyer-Seller Relationships in Marketing Channels
Gerrit H. van Bruggen, Manish Kacker & Chantal Nieuwlaat
ERS-2001-44-MKT

Incorporating Responsiveness to Marketing Efforts when Modeling Brand Choice
Dennis Fok, Philip Hans Franses & Richard Paap
ERS-2001-47-MKT

Competitiveness of Family Businesses: Distinghuising Family Orientation and Business Orientation
Mark A.A.M. Leenders & Eric Waarts
ERS-2001-50-MKT

The Effectiveness of Advertising Matching Purchase Motivation: An Experimental Test
ERS-2001-65-MKT
Joost Loef, Gerrit Antonides & W. Fred van Raaij

Using Selective Sampling for Binary Choice Models to Reduce Survey Costs
ERS-2001-67-MKT
Bas Donkers, Philip Hans Franses & Peter Verhoef

Deriving Target Selction Rules from Edogenously Selected Samples
ERS-2001-68-MKT
Bas Donkers, Jedid-Jah Jonker, Philip Hans Franses & Richard Paap

2000
Impact of the Employee Communication and Perceived External Prestige on Organizational Identification
Ale Smidts, Cees B.M. van Riel & Ad Th.H. Pruyn
ERS-2000-01-MKT

Forecasting Market Shares from Models for Sales
Dennis Fok & Philip Hans Franses
ERS-2000-03-MKT

The Effect of Relational Constructs on Relationship Performance: Does Duration Matter?
Peter C. Verhoef, Philip Hans Franses & Janny C. Hoekstra
ERS-2000-08-MKT

Informants in Organizational Marketing Research: How Many, Who, and How to Aggregate Response?
Gerrit H. van Bruggen, Gary L. Lilien & Manish Kacker
ERS-2000-32-MKT

The Powerful Triangle of Marketing Data, Managerial Judgment, and Marketing Management Support Systems
Gerrit H. van Bruggen, Ale Smidts & Berend Wierenga
ERS-2000-33-MKT

Consumer Perception and Evaluation of Waiting Time: A Field Experiment
Gerrit Antonides, Peter C. Verhoef & Marcel van Aalst
ERS-2000-35-MKT

Broker Positions in Task-Specific Knowledge Networks: Effects on Perceived Performance and Role Stressors in an Account Management System
David Dekker, Frans Stokman & Philip Hans Franses
ERS-2000-37-MKT

Modeling Unobserved Consideration Sets for Household Panel Data
Erjen van Nierop, Richard Paap, Bart Bronnenberg, Philip Hans Franses & Michel Wedel
ERS-2000-42-MKT

A Managerial Perspective on the Logic of Increasing Returns
Erik den Hartigh, Fred Langerak & Harry Commandeur
ERS-2000-48-MKT

The Mediating Effect of NPD-Activities and NPD-Performance on the Relationship between Market Orientation and Organizational Performance
Fred Langerak, Erik Jan Hultink & Henry S.J. Robben
ERS-2000-50-MKT

Sensemaking from actions: Deriving organization members' means and ends from their day-to-day behavior
Johan van Rekom, Cees B.M. van Riel & Berend Wierenga
ERS-2000-52-MKT