Balancing Survey Costs with Nonresponse Bias Using Callbacks in Telephone Surveys

Jennifer Czuprynski

April 27, 2000

A Senior Research Project
Dr. John Rasp
Dr. Erich Friedman

Department of Mathematics and Computer Science
Stetson University
Deland, Florida

Table of Contents

Introduction and Literature Review

In a telephone-administered survey, the survey designer is faced with an important and yet difficult decision—how many additional callbacks to make to each household not contacted on the first call. Making too few callbacks results in an unacceptably large nonresponse bias, while making too many callbacks result in larger survey costs. It is the job of the designer to balance the monetary costs of making callbacks with loss of data quality that results from having a biased sample caused by too few callbacks.

In 1949, Politz and Simmons were some of the early pioneers in studying the number of callbacks to be utilized in telephone surveys. They designed a method that required only one call to each household. Interviews are weighted based on the probability of a person being at home when called at a random time. The probability of a person being home was determined by a series of questions regarding when the respondent was home during that previous week (Groves 1989:170). Critics of this method found that it had greater problems than it had benefits. In particular, it does not adjust its findings for people who are not at home at any time during the survey period. The method also assumes that the week of the survey period is a typical week for respondents. One of the largest problems with this method is that it requires respondents accurately to report their activities of the previous week. Yet relying on the memory of respondents may lead to an inaccurate weighting method (Groves 1989:170).

Later methodology developed on callbacks does not require actual questions directed at the respondent's at-home patterns of the week. If the survey is constructed in a manner such that calls are made at random to respondents, then the chance of finding a certain person at home is a constant equal to the proportion of time that he/she spends in his/her home (Bartholomew 1961). Potthoff, Manton, and Woodbury (1993) used the idea that the proportions are constant.

Therefore, they modeled the probability of a person being at home with the binomial distribution. Since the binomial distribution is an individualistic approach, it is combined with the beta distribution to generalize to the entire population. A beta distribution describes a variable that is binomial in nature, according to Potthoff, Manton, and Woodbury, because it "has the correct range; is flexible and mathematically tractable; and is [the] conjugate [prior] to the binomial and negative binomial distributions." However, Bartholomew (1961) challenges the idea that calls are made at random. Oftentimes, he explains, interviewers use information obtained from the first nonresponse call to schedule a time for the following calls. Therefore, it cannot be assumed, in general, that the probability of a person being home is constant. The binomial-beta model fails to work.

If the interviewer calls limitlessly until successfully reaching the respondent, Potthoff, Manton, and Woodbury use the geometric distribution to describe the nature of callbacks. However, Colombo (2000) does not believe that the geometric distribution best describes the nature of callbacks. As Colombo says, the geometric distribution "models behavior at the individual level—it says nothing about how the response probabilities are distributed in the population." Thus, Colombo combines the beta distribution of response probabilities with the geometric distribution for callbacks to model the probability p_k of a successful call across the population:

$$p_k = \int \pi(1-\pi)^{k-1} f(\pi \mid \alpha, \beta) d\pi$$

$$f(\pi \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-\pi)^{\beta-1}$$

where π is the response probability, k is the number of callbacks, and $f(\pi \mid \alpha, \beta)$ is the beta distribution. Creating a histogram of the number of responses obtained at each call and applying either maximum likelihood estimators or method of moments estimators can then determine the

values of alpha and beta.

In an applied setting, survey designers must take survey costs into account at each stage of a survey. This is especially true for callback methods, as Tuchfarber and Klecka (1976) point out. At each callback, the designer of the survey must ask "Is the representativeness being improved enough to justify the costs of making call-backs to all of the numbers not yet reached?" As Bartholomew (1961) explains there are many negative aspects to repeated callbacks, such as delaying the completion of the survey, lowering the interviewers morale with repeated no answers, etc. But Groves (1989:160) considers the cost of each callback the most important drawback to consider. Due to the added amount of paperwork devoted to each household, the extra time of making the callback, the additional time that it takes the supervisor to organize the callbacks, etc., each callback incurs increased survey costs. Groves explains that the cost of surveys is often not a topic considered important by those researching survey methodologies. However, in applied settings, surveys are often conducted on fixed budgets, and the cost of each part of the design must remain within this budget. Groves states that "survey costs and errors are reflections of each other; increasing one reduces the other." Reducing the error due to noncontacts by increasing the number of callbacks made will increase the cost of each interview. A balance must be made between the acceptable noncontact bias and the acceptable callback cost. Elliott et al. (2000) propose that the expected cost E of a completed interview per unit attempted is:

$$E = \sum_{k=1}^{K} p_k[(c_k / q_k) + d_k] + r_k e_k$$

where K is the number of strata based on the number of callbacks made, p_k is the marginal

probability of obtaining an interview at the k^{th} callback, c_k is the relative cost of the k^{th} callback, q_k is the probability of obtaining an interview at the k^{th} callback given that no interview or refusal has been achieved from a previous callback, d_k is the relative cost of an interview at the k^{th} callback, r_k is the marginal probability of a refusal at the k^{th} callback, and e_k is the relative cost of a refusal at the k^{th} callback.

Thus, when determining the number of callbacks, one must compare the benefit obtained from each call with the cost of the callback process. The benefit of each call can be measured by how much it reduces the noncontact bias. Stopping callbacks at a certain number means that members of a sample are excluded from the survey data simply because they were not home when interviewers called them. As Colombo (2000) points out, "non sampling error is unquantifiable. The difference between the respondents and the nonrespondents could be so large that the survey data would be almost useless." While it is unlikely for such extreme differences to exist between the two groups, it is important to determine how much the respondents and the nonrespondents differ in order to determine the accuracy of the data generated. This is what Colombo calls the "paradoxical role of callbacks." If the population has a homogeneous response distribution, "callbacks will have little impact on bias but do result in an increase in the number of interviews." However, if the population has a heterogeneous response distribution, "callbacks could have a substantial effect on reducing nonresponse bias, but they are unlikely to result in many additional responses. Colombo's paradox creates problems for survey designers because the distribution of response probabilities is unknown prior to designing the survey. Thus, the designer of a survey does not know the benefit of performing more callbacks.

Development of The Model

There are obviously many questions to consider when determining how many callbacks are to be made in a telephone-administered survey. For this study, the main research question is: What is the best number of callbacks to be used in a telephone-administered survey that minimizes costs while also minimizing the error due to noncontact bias.

In general, surveys are conducted to determine how many members of a defined population, such as the County of Volusia, Florida have specific attributes. For example, a survey might be concerned with how many United States residents favor a certain bill that is being debated in Congress or how many people participate in volunteer activities in their community, etc. The survey administrators desire to estimate the true population values for this attribute using only a small sample of the population as an estimate. The survey administrators seek to develop a sample that has a small level of variance in order to create a more accurate estimator. Keeping in mind constraints on cost and bias, we much determine an estimator that has a minimum variance, low cost, and low levels of bias. In doing so, we will be able to generate a methods for determining the best sample size and the best number of callbacks to be made in a survey with these constraints in mind. In the following pages, I will be developing a simple model and then expanding on that model to create models that better reflect the actual population

Model 1

First, consider a simple case where each person in the population being sampled has the same probability of being contacted, π. Assume that only one call is made to each household, where n people are called and only n_1 people are reached. Let, p represent the probability of a

person having the desired attribute, where p is independent of π. Whether or not a person has the desired attribute is measured using a 0/1 variable represented by x_i.

In this scenario, we would expect to contact μ respondents that have the desired attribute, where

$$\mu = \pi pn$$

However, since the actual population values for π and p are unknown, they must be estimated, using

$$\hat{\pi} = \frac{n_1}{n}$$

and

$$\hat{p} = \frac{\sum_{i=1}^{n_1} x_i}{n_1}$$

where $\sum_{i=1}^{n_1} x_i$ is the number of people that were reached who have the attribute. Since p and π are assumed to be independent, we can treat n_1 as a fixed value rather than as a random variable. Inserting these estimates into μ, as found in Appendix A, yields the equation

$$\hat{\mu} = \sum_{i=1}^{n_1} x_i$$

as an estimate for μ. This is simply the number of people contacted that have the attribute.

In order to find a low-biased, low-cost survey method, we need to minimize the variance of the estimator for μ. subject to a constraint on survey costs and a constraint on the amount of nonresponse bias, both set by the survey administrator. The variance of the estimator is:

$$V(\hat{\mu}) = \sum_{i=1}^{n_1} x_i^2 - \frac{1}{n_1}(\sum_{i=1}^{n_1} x_i)^2$$

The cost c is a linear equation of fixed costs, the costs of the interviews obtained, and the costs of the refusals or noncontacts. Letting c_0 represent the fixed costs, c_1 represent the cost of an interview, and c_2 represent the cost of a refusal or noncontact yields the equation

$$c = c_0 + c_1 n_1 + c_2 (n - n_1)$$

The bias B is the difference between the expected value of the estimator and the actual population value,

$$B = | E(\hat{\mu}) - \mu | = | p(n_1 - \pi n) |$$

This equation illustrates that there will be bias unless $n_1 = \pi n$, or unless the estimator for π is equal to the actual π.

Thus, we want to minimize the variance subject to a cost constraint c_f and a bias constraint ε. To minimize a function subject to constraints, we can use Lagrange Multipliers. In this instance, since everything is linear, we can also use linear programming, though this method requires using the actual data, while Lagrange Multipliers allows a theoretical estimation. Since the bias equation involves absolute values, the there are two cases to consider:

$$\frac{n_1}{n} > \pi \text{ and } \frac{n_1}{n} < \pi .$$

Using Lagrange Multipliers yields estimates for n and n_1:

$$n = \frac{c_f - c_0}{c_2} - \frac{\varepsilon}{p + p\pi} + \frac{\pi c_f - \pi c_0}{c_2(1+\pi)}$$

and

$$n_1 = \frac{c_f}{c_1} - \frac{c_0}{c_1} - \frac{c_2}{c_1}\frac{\varepsilon}{\pi p} \quad \text{, when } \frac{n_1}{n} > \pi$$

and

$$n = \frac{c_f}{c_2} - \frac{c_0}{c_2} + \frac{(c_1 - c_2)\varepsilon}{p\pi(c_1 - c_2) + c_2 p} - \frac{(c_1 - c_2)\pi c_f}{c_2\pi(c_1 - c_2) + c_2^2} + \frac{(c_1 - c_2)\pi c_0}{c_2\pi(c_1 - c_2) + c_2}$$

and

$$n_1 = \frac{\varepsilon c_2}{p\pi(c_1 - c_2) + c_2 p} + \frac{\pi c_f}{\pi(c_1 - c_2) + c_2} - \frac{\pi c_0}{\pi(c_1 - c_2) + c_2} \text{, when } \frac{n_1}{n} < \pi$$

Since these estimators are needed prior to the actual gathering of the data, estimates for π and p can be estimated from previous data that has been gathered.

Similar formulas can be developed for surveys that include more than one call. The formula for μ after a maximum of k calls per household is

$$\mu = \pi p n + (1 - \pi)\pi p n + \ldots + (1 - \pi)^{k-1}\pi p\, n$$

The estimator for π remains the same as in the previous case because it is still the probability of finding someone at home in a given call. This may not be the best estimator for π, but it has the nice feature that it can be calculated after just one call. Thus, after one call to each household,

the survey administrator can determine how many more calls should be made. Additionally, the estimator for p is still simply the number of people in the sample reached that have the attribute divided by the number of people that were reached. Thus,

$$\hat{\pi} = \frac{n_1}{n}$$

and

$$\hat{p} = \frac{\sum_{i=1}^{n_1+n_2+\ldots+n_k} x_i}{n_1 + n_2 + \ldots + n_k}$$

Therefore, as shown in Appendix B, the estimator for μ is

$$\hat{\mu} = (n_1 + \ldots + (1 - \frac{n_1}{n})^{k-1} n_1) \frac{\sum_{i=1}^{n_1+\ldots+n_k} x_i}{n_1 + \ldots + n_k}$$

and the variance of $\hat{\mu}$ is

$$V(\hat{\mu}) = (n_1 + \ldots + (1 - \frac{n_1}{n})^{k-1} n_1)^2 \frac{1}{n_1 + \ldots + n_k} (\sum_{i=1}^{n_1+\ldots+n_k} x_i^2 - \frac{1}{n_1 + \ldots + n_k} (\sum_{i=1}^{n_1+\ldots+n_k} x_i)^2)$$

Assuming that all callbacks have the same costs no matter if it is the second call or the fifteenth call, the cost equation is

$$c = c_0 + c_1 n_1 + c_2 (n_2 + \ldots + n_k) + c_3 (n - n_1) + c_4 (n - n_1 - \ldots - n_k)$$

where c_0 is the fixed costs, c_1 is the cost of the first call, c_2 is the cost of callbacks that obtain interviews, c_3 is the cost of refusals and noncontacts on the first call, and c_4 is the cost of refusals

and noncontacts during the callback process. The bias equation is the difference between the expected value of the estimator and μ.

$$B = |\ (\hat{\pi}pn - \pi pn) + \ldots + ((1-\hat{\pi})^{k-1}\hat{\pi}pn - (1-\pi)^{k-1}\pi pn)\ |$$

Using the same methods as above, we could minimize the variance estimator subject to the sample size n and the maximum number of calls per household k. This would tell us the best numbers of the variables within our control to yield a low-bias, low cost survey. However, there are difficulties minimizing with respect to k because k is not only a variable but also an index representing the number of groups of n corresponding to a callback stage. This minimization cannot be done analytically and thus, would have to be done numerically.

Model 2

Now we consider a slightly more complicated model in which there are two groups of people in the population—those who are always home (group 1) and those who are sometimes home (group 2). In order to make estimations in this scenario, at least two calls must be made. Let m_1 and m_2 be the proportion of the population in each group and π_1 and π_2 be the probability of a person being home in each group. Assuming that p, the proportion of people in the population with the desired attribute, is independent of the at home probabilities π_1 and π_2, p is the same for both groups. Since m_1 and m_2 equal the whole population, $m_2 = 1 - m_1$. Also, since the people in group 1 are always home, the probability of reaching them at home is $\pi_1 = 1$. Now assume that n people are called, n_1 people are reached on the first call and n_2 people are reached on the second call. In one call we would expect to reach every person in group one and the

proportion of group two that is home. In the second call we would then expect to reach the proportion of group two that was not home on the first call, but home on the second call. Thus, in two calls, we would expect to reach μ people that have the desired attribute, where

$$\mu = m_1 p * 1 * n + (1 - m_1) p\pi_2 n + (1 - m_1) p(1 - \pi_2)\pi_2 n$$

Since the proportion of the sample reached on the first call is all of the proportion in group 1 and the proportion of group 2 that were home on the first call,

$$\frac{n_1}{n} = m_1 + (1 - m_1)\pi_2 ,$$

and the proportion reached on the second call are the proportion of group 2 that were missed on the first call and reached on the second call,

$$\frac{n_2}{n} = (1 - m_1)(1 - \pi_2)\pi_2 ,$$

the estimates for m_1, π_2, and p, as derived in Appendix C, are:

$$\hat{m}_1 = \frac{-nn_1 + n_1^2 + nn_2}{n(n - n_1 - n_2)}$$

$$\hat{\pi}_2 = \frac{n_2}{n - n_1}$$

Again the proportion in the population that have the desired attribute is estimated by the number of people reached that have the attribute divided by the total number of people reached,

$$\hat{p} = \frac{\sum_{i=1}^{n_1+n_2} x_i}{n_1 + n_2}$$

Thus, the estimator for μ is

$$\hat{\mu} = \sum_{i=1}^{n_1+n_2} x_i$$

which is again simply the number of people contacted that have the attribute. The variance of this estimator is

$$V(\hat{\mu}) = \sum_{i=1}^{n_1+n_2} x_i - \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1+n_2} x_i\right)^2$$

The costs of the survey are a linear equation of fixed costs c_0, costs of the survey on first call c_1, cost of the survey on the second call c_2, and cost of the nonresponse c_3. Thus,

$$c = c_0 + c_1 n_1 + c_2 n_2 + c_3 (n - n_1 - n_2)$$

The bias equation, the difference between the expected value of the estimator and μ, is

$$B = | pn_1 - m_1 pn - (1 - m_1) p\pi_2 n) + (pn_2 - (1 - m_1) p(1 - \pi_2)\pi_2 n |$$

We can use Lagrange Multipliers here to find the minimum variance subject to cost and bias constraints by taking derivatives with respect to n, n_1, and n_2. Taking the derivatives of these equations cannot be done analytically due to the fact that n_1 and n_2 are both variables and indexes. These equations must be solved numerically.

Expanding this example to k calls will allow us to estimate the best number of calls to make. In this scenario for k calls, the actual μ would be:

$$\mu = m_1 pn + (1-m_1)p\pi_2 n + (1-m_1)p(1-\pi_2)\pi_2 n + \ldots$$
$$+ (1-m_1)p(1-\pi_2)^{k-1}\pi_2 n$$

The estimates for m_1 and π_2 can remain the same as before, though as discussed previously, this may not be the best estimator. This estimator does however have the benefit of being able to be calculated after only the second call is made. The estimate for p is

$$\hat{p} = \frac{\sum_{i=1}^{n_1+\ldots+n_k} x_i}{n_1 + \ldots + n_k}$$

As derived in Appendix D, substituting for the estimators yields

$$\hat{\mu} = ([\frac{n(n-n_1)-nn_2}{(n-n_1-n_2)} + (1-[\frac{n(n-n_1)-nn_2}{n(n-n_1-n_2)}])(\frac{n_2}{n-n_1})n + \ldots$$
$$+ (1-\frac{n(n-n_1)-nn_2}{n(n-n_1-n_2)})(1-\frac{n_2}{n-n_1})^{k-1}(\frac{n_2}{n-n_1})n)(\frac{\sum_{i=1}^{n_1+\ldots+n_k} x_i}{n_1+\ldots+n_k})$$

Taking the variance of this estimator yields the equation

$$V(\hat{\mu}) = ([\frac{n(n-n_1)-nn_2}{(n-n_1-n_2)} + (1-[\frac{n(n-n_1)-nn_2}{n(n-n_1-n_2)}])(\frac{n_2}{n-n_1})n + \ldots$$
$$+ (1-\frac{n(n-n_1)-nn_2}{n(n-n_1-n_2)})(1-\frac{n_2}{n-n_1})^{k-1}(\frac{n_2}{n-n_1})n)^2(\frac{1}{n_1+\ldots+n_k})^2(\sum_{i=1}^{n_1+\ldots+n_k} x_i^2 - \frac{(\sum_{i=1}^{n_1+\ldots+n_k} x_i)^2}{n_1+\ldots+n_k})$$

Assuming that all of the callbacks cost the same, the cost equation is a linear combination of fixed costs, costs of obtaining interview on 1st call, costs of obtaining interview during callbacks, costs of nonresponse on 1st call, and costs of nonresponse during callbacks.

Letting c_0 be the fixed costs, c_1 be the cost of obtaining an interview on the first call, c_2 be the costs of obtaining an interview during the callback process, c_3 be the costs of a noncontact on the first call, and c_4 be the costs of noncontacts during the callback process, the equation for cost is

$$c = c_0 + c_1 n_1 + c_2 (n_2 + \ldots + n_k) + c_3 (n - n_1) + c_4 (n - n_1 - \ldots - n_k)$$

The bias equation is

$$B = pn \,|\, (\hat{m}_1 - m_1) + ((1 - \hat{m}_1)\hat{\pi}_2 - (1 - m_1)\pi_2) + \ldots + ((1 - \hat{m}_1)(1 - \hat{\pi}_2)^{k-1}\hat{\pi}_2 - (1 - m_1)(1 - \pi_2)^{k-1}\pi_2 \,|$$

From here, we can minimize the variance with respect to k and n subject to the cost constraint c_f and the bias constraint ε. However, like before, this cannot be solved analytically due to the fact that k is a number and an index. Thus, numerical methods must be utilized.

The cases could be extended to multiple groups of the population with varying degrees of at-home probabilities, such as always home, sometimes home, never home, or always home, often home, sometimes home, rarely home, never home, etc. In each case, similar equations could be developed as above. However, this process becomes tedious, as can be seen from the above derivations.

Model 3

A more accurate approach models the response probabilities as a probability distribution function as opposed to fixed groups of people. This method allows for the variation that is inherent in true populations, where people cannot generally be molded into fixed categories. Lohr suggests modeling response probabilities using logistic regression, since response can be measured as a 0/1 variable. Thus, whether or not a person responds, R_i, is a function of some other demographic variable, x_i. For example, it may be that people with higher incomes are more likely to respond. Thus, the response probability can be estimated as

$$P(R_i = 1) = \frac{e^{\beta_0 + \beta_1 x_i + \ldots}}{1 + e^{\beta_0 + \beta_1 x_i + \ldots}}$$

where the β's are parameters that estimate the response probability. While this model does not require x to have any certain distribution, in practice, x will probably be a discrete variable since in survey even continuous items, such as age and income, are often broken into discrete categories for measurement during the survey process.

The question then becomes how to estimate the β parameter given values for x. There are three situations that need to be considered. In the first case, x's are known for each individual prior to the survey. In telephone surveys, this is always the case with at least the telephone numbers, which could be used to define a place of residence. This is the simplest case because no estimation needs to be made for the x's. For the second case, the population values of x are known prior to the survey, but not the actual x values. Demographic information is often fairly easy to obtain through sources like the Census and other databases, so this is likely to be

the case in most surveys. This case is harder because assumptions have to be made about x's. This case also introduces error that is not in the first case because the x's are estimated for nonrespondents and not exact as they were before. In the third case, nothing is known about x. In this case, it is impossible to estimate the response probabilities of nonrespondents because there is no information of which to draw. If this is the case, a different model might need to be developed.

I have developed a model for the first case, in which the x's are known in advance. Consider a model in which the specific attribute desired can be measured as a 0/1 variable, such as whether or not a person will favor or object to a certain bill. This attribute can also be measured as a logistic regression model based on the x variable, as was the response probability. Letting Y_i be whether or not the respondent has the attribute,

$$P(Y_i = 1) = \frac{e^{\gamma_0 + \gamma_1 x_i + \ldots}}{1 + e^{\gamma_0 + \gamma_1 x_i + \ldots}}$$

We can use the method of maximum likelihood to determine estimations for the values of the β's and the γ's. The method of maximum likelihood is a method to derive minimum-variance unbiased estimators for parameters by likelihood maximizing the probability of obtaining the observed sample by selecting appropriate values for the parameters, using joint probability or joint density functions.

For this model, the likelihood is the product of the likelihood of being selected in the sample, the likelihood of being contacted, and the likelihood of having the attribute given that the person was already contacted. The likelihood of being selected is dependent upon the distribution of the population according to the demographic x value. In the case where x is

broken into categories, the likelihood of being selected may be represented as a multinomial distribution. Thus, in the case where x is binomial, and η is the proportion of the population in a certain category, such as males, the distribution would be

$$\eta^{x_i}(1-\eta)^{1-x_i}$$

Letting k represent the number of times an individual was called and was not reached, the likelihood of being contacted, as derived in Appendix E, is the product of likelihood of not being contacted to the power of the number of times a person was called and not reached and the likelihood of being reached to the power of whether or not the person was ever contacted:

$$[1-\frac{e^{\beta_0+\beta_1 x_i+\ldots}}{1+e^{\beta_0+\beta_1 x_i+\ldots}}]^k[\frac{e^{\beta_0+\beta_1 x_i+\ldots}}{1+e^{\beta_0+\beta_1 x_i+\ldots}}]^{R_i}$$

Because the second part of the function is raised to the R_i, if the person does not respond this part of the equation will be equal to one, and thus will not affect the overall equation. The likelihood of having the attribute given that the person was contacted is the product of the likelihood of having the attribute raised to whether or not the person responds and has the attribute and the likelihood of not having the attribute raised to whether or not the person responds and has the attribute:

$$=[\frac{e^{\gamma_0+\gamma_1 x_i+\ldots}}{1+e^{\gamma_0+\gamma_1 x_i+\ldots}}]^{Y_i R_i}[1-\frac{e^{\gamma_0+\gamma_1 x_i+\ldots}}{1+e^{\gamma_0+\gamma_1 x_i+\ldots}}]^{(1-Y_i)R_i}$$

Because both parts of the equation are raised to R_i, this part of the equation will become mute if the person did not respond. Thus, the likelihood of a person responding and having the attribute is:

$$L=\eta^{x_i}(1-\eta)^{1-x_i} \ *[1-\frac{e^{\beta_0+\beta_1 x_i+\ldots}}{1+e^{\beta_0+\beta_1 x_i+\ldots}}]^k[\frac{e^{\beta_0+\beta_1 x_i+\ldots}}{1+e^{\beta_0+\beta_1 x_i+\ldots}}]^{R_i} *[\frac{e^{\gamma_0+\gamma_1 x_i+\ldots}}{1+e^{\gamma_0+\gamma_1 x_i+\ldots}}]^{Y_i R_i}[1-\frac{e^{\gamma_0+\gamma_1 x_i+\ldots}}{1+e^{\gamma_0+\gamma_1 x_i+\ldots}}]^{(1-Y_i)R_i}$$

This is the likelihood for individuals. In order to generalize the likelihood function to the whole population, we take the product of this function over the sample population. Thus, the likelihood function that we wish to maximize is:

$$L=\prod_{i=1}^{n}[\eta^{x_i}(1-\eta)^{(1-x_i)}[\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}]^k[1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}]^{R_i}[\frac{e^{\gamma_0+\gamma_1 x_i+\ldots}}{1+e^{\gamma_0+\gamma_1 x_i+\ldots}}]^{Y_i R_i}[1-\frac{e^{\gamma_0+\gamma_1 x_i+\ldots}}{1+e^{\gamma_0+\gamma_1 x_i+\ldots}}]^{(1-Y_i)R_i}]$$

To maximize a function, we need to take the derivative. However, the likelihood function above is difficult to take the derivative of because of the multiple times that the product rule would have to be employed. Since ln[L] is a monotonically increasing function of L, both ln[L] and L are maximized at the same point. Thus, it will be easier for us to take the derivative of the ln[L] because the natural logarithm function changes products into sums, a much easier equation to take the derivative of

$$\ln[L]=\sum_{i=1}^{n}(\ln(\eta^{x_i}(1-\eta)^{1-x_i}))+\ln(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}})^k+\ln(1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}})^{R_i}+\ln(\frac{e^{\gamma_0+\gamma_1 x_i}}{1+e^{\gamma_0+\gamma_1 x_i}})^{Y_i R_i}$$
$$+\ln(1-\frac{e^{\gamma_0+\gamma_1 x_i}}{1+e^{\gamma_0+\gamma_1 x_i}})^{(1-Y_i)R_i})$$

After simplification shown in Appendix E, this yields

$$\ln(L) = n\ln(1-\eta) + \ln(\frac{\eta}{1-\eta})\sum_{i=1}^{n} x_i - k\sum_{i=1}^{n}\ln(1+e^{\beta_0+\beta_1 x_i}) + \sum_{i=1}^{n} R_i(\beta_0 + \beta_0 x_i) - \sum_{i=1}^{n} R_i \ln(1+e^{\beta_0+\beta_1 x_i})$$
$$+ \sum_{i=1}^{n} Y_i R_i(\gamma_0 + \gamma_1 x_i) - \sum_{i=1}^{n} Y_i R_i \ln(1+e^{\gamma_0+\gamma_1 x_i}) - \sum_{i=1}^{n}(1-Y_i)R_i(\gamma_0 + \gamma_1 x_i)$$

Since we wish to maximize this function in order to determine the β's and the γ's, we need to take the derivative with respect to these variables and set them equal to zero.

$$\frac{\partial \ln L}{\partial \beta_0} = -k\sum_{i=1}^{n}\frac{1}{1+e^{\beta_0+\beta_1 x_i}} + \sum_{i=1}^{n} R_i - \sum_{i=1}^{n} R_i \frac{1}{1+e^{\beta_0+\beta_1 x_i}} = 0$$

$$\frac{\partial \ln L}{\partial \beta_0} = -k\sum_{i=1}^{n}\frac{1}{1+e^{\beta_0+\beta_1 x_i}} + \sum_{i=1}^{n} R_i x_i - \sum_{i=1}^{n} R_i \frac{1}{1+e^{\beta_0+\beta_1 x_i}} x_i$$

$$\frac{\partial \ln(L)}{\partial \gamma_0} = \sum_{i=1}^{n} Y_i R_i - \sum_{i=1}^{n} Y_i R_i \frac{1}{1+e^{\gamma_0+\gamma_1 x_i}} - \sum_{i=1}^{n}(1-Y_i)R_i = 0$$

$$\frac{\partial \ln L}{\partial \gamma_1} = \sum_{i=1}^{n} Y_i R_i x_i - \sum_{i=1}^{n} Y_i R_i \frac{1}{1+e^{\gamma_0+\gamma_1 x_i}} x_i - \sum_{i=1}^{n}(1-Y_i)R_i x_i = 0$$

The values for the parameters cannot be solved analytically. Numerical techniques must be used.

Once the estimators for the β's and the γ's have been found, we can derive μ, the probability of a person in the sample being contacted and having the attribute, and the variance of the estimator for μ. Using the binomial example from above,

$$\mu = P(Y=1 \mid x=1) * P(X=1) + P(Y=1 \mid x=0) * P(x=0)$$

$$\hat{\mu} = \frac{e^{\hat{\gamma}_0+\hat{\gamma}_1}}{1+e^{\hat{\gamma}_0+\hat{\gamma}_1}}\eta + \frac{e^{\hat{\gamma}_0}}{1+e^{\hat{\gamma}_0}}(1-\eta)$$

It is notable that the β's are not in the formula for the estimator of μ. The β's were simply nuisance parameters that needed to be found in order to find the γ's, but they do not really contribute to the end result. The variance of this estimator is:

$$V(\hat{\mu}) = \eta^2 \frac{e^{\hat{\gamma}_0+\hat{\gamma}_1}}{1+e^{\hat{\gamma}_0+\hat{\gamma}_1}}\left(1-\frac{e^{\hat{\gamma}_0+\hat{\gamma}_1}}{1+e^{\hat{\gamma}_0+\hat{\gamma}_1}}\right) + (1-\eta)^2 \frac{e^{\hat{\gamma}_0}}{1+e^{\hat{\gamma}_0}}\left(1-\frac{e^{\hat{\gamma}_0}}{1+e^{\hat{\gamma}_0}}\right)$$

This equation should be minimized with respect to the sample size and the number of callbacks, subject to a cost constraint and a bias constraint, as in the pervious examples.

Application

In order to employ the necessary numerical techniques, I have applied the equations developed in the previous section to data from a telephone survey conducted in September 2000 by the Stetson Institute for Social Research (SISR), housed in the Department of Sociology and Anthropology at Stetson University. A team of faculty researchers conducted the surveys with the aid of student interviewers and supervisors. Residents were sampled with random digit dialing, using a sample purchased from a sample-generating agency. Interviewers were directed to call each number in the sample five times or until the number was removed from the list for another reason. Numbers were removed when either the number was found to be invalid (i.e. business, disconnected, etc.), when a survey was completed, or when interviewers received refusals from respondents. See Appendix F for the Telephone Codes.

Interviewers were instructed to leave a maximum of two messages on the residents' answering machines or voice mail before discarding the number. The interview period lasted until four hundred interviews were completed, approximately three weeks. Surveys were conducted in the evenings during the week and in the afternoons and evenings on the weekends. A sample log sheet can be found in Appendix G. The survey concerned local political issues that would be voted on in the November 2000 elections.

In order to analyze the data, I treated all calls as a noncontact unless the interviewer actually conducted a survey. This is not the most meaningful way to analyze the data because disconnects and businesses should not be treated the same as refusals or as no answers. However, we have used this simplification assumption for the development of these models. Better formulas would need to be developed to account for all of these options.

The sample size for the survey was initially 2715, but this was not large enough to obtain the goal of 400 interviews, so a new sample was created and added to this sample. However, because the households in the second sample were not called five times before the completion of the survey, I did not include this sample in my analysis.

Since the model developed requires the x demographic variable be a variable known in advance, we have to use one of the two x variables that we know – telephone number and zip code. Both of these variables represent the person's place of residence, but the zip code is a more accurate measure of this. I took the zip codes of Volusia County, Florida and divided them into six areas, which I call Daytona, Port Orange, Edgewater, Lake Helen, Deland, and Pierson. The zip code map with these areas mapped out is found in Appendix G. The x values have a multinomial distribution.

I analyzed the results of the survey question that asked whether or not the respondent would favor a certain bill that would be on the ballot in the upcoming election. Specifically, the question (number 12) asked

"Would you vote for or against this proposal to raise property taxes to buy and preserve environmentally-sensitive lands?"

The choice was either favor or not favor, thus it is a 0/1 variable that can be modeled using logistic regression. The survey is included as Appendix I.

A total of 5933 calls were made and only 287 people were reached, for a low response rate of 4.8%. Of these, 187 (65.2%) favored the bill. On the first call, all 2715 people in sample were called and 131 (4.8%) responses were obtained. The first call accounted for 45.6% of all of the responses. Eighty-nine (67.9%) of the respondents on the first call favored the bill. On the second call, 1454 calls were made. Some of these were cut out due to being contacted, while

others resulted in refusals, businesses, disconnects, and other responses which are not called again. Of these 1454 calls, 118 (8.1%) responses were obtained. This call accounted for 41.1% of the total responses obtained. Thus, in the first and second call, 86.7% of the respondents were reached. Of the respondents reached on the second call, 80 (67.8%) favored the bill. In the third call, 742 calls were made and 22 (3%) responses were obtained, representing 7.7% of the responses gained overall. Seven (31.9%) of these respondents favored the bill. On the fourth call, 541 calls were made and seven (1.3%) responses were obtained, accounting for 2.8% of the total responses. Of these respondents, four (57.1%) favored the bill. On the fifth call, 481 calls were made with only eight (1.7%) respondents, representing 2.8% of the total. Seven (87.5%) of these favored the bill. Nine people were mistakenly called a sixth time, garnering only one (11.1%) response (0.3% of the total). This person did not favor the bill.

Of the 2715 people in the sample, 779 (28.7%) were in Daytona. Of these, 70 (9.0%) responded and 45 (64.3%) favored the bill. There were 602 (22.2%) people in the Port Orange portion of the sample. Fifty-six (9.3%) of these responded and 38 (67.9%) favored the bill. In Edgewater, there were 321 (11.8%) people in the sample. Thirty-four (10.6%) of these responded and 20 (58.8%) favored the bill. Lake Helen contained 662 (24.4%) of the people in the sample, 84 (12.7%) of these responding and 54 (64.3%) of them favoring the bill. In DeLand there were 307 (11.3%) people from the sample, 37 (12.1%) of these responding, and 26 (70.3%) favoring the bill. Pierson contained 44 (1.6%) of the people in the sample, six (13.6%) of these responding and four (66.7%) of them favoring the bill.

Before performing any type of analysis, we must first test to see if the nonresponse is ignorable. According to Sharon Lohr, if the probability of nonresponse depends on the value of the desired attribute, then the nonresponse is called nonignorable and thus must be accounted for.

For example, if the survey seeks to determine how many people volunteer in community organizations, it may be difficult to contact these people who volunteer because they are busy volunteering. Since nonresponse is measured using the γ variables, we are testing the null hypothesis that $\gamma_1=0$ against the alternative hypothesis that $\gamma_1\neq0$. To perform the hypothesis test, we create the model using the likelihood equation derived above. Then we create the model using $\gamma_1=0$. Twice the difference of the ln[L] of the two models is distributed χ^2 with one degree of freedom, where 1 is the number of parameters that differ between the two models. For these data,

$$\chi^2 = |\ln L(\text{full model})-\ln L(\text{partial model})|$$

$$= 2|-1326.186-(-1326.818)|$$

$$=1.263$$

Since 1.263 is less than 3.84, we cannot reject the null hypothesis at the $\alpha = .05$ level. Thus, in this case the nonresponse is not nonignorable and so the sample reached is representative of the population. However, I will consider analyzing it as if the nonresponse is nonignorable in order to demonstrate the process.

For the data analysis I used the Excel built-in optimization package Solver to minimize the variance and estimate values for the β's and the γ's. The probability that a person will favor the tax increase overall is the sum of the probabilities of the individual areas favoring the tax increase weighted according to the proportion of the population they represent. Probabilities of responding are functions of the γ parameters, specifically each area has a probability function of only γ_0 and the γ that corresponds to the area. The others disappear when they are multiplied by

the x value. Thus, letting w represent the proportion of the population representing each group, the probability that a person will favor the tax increase regardless of whether or not he or she responds to the survey is

$$\mu = P(Y=1) = w_1\, P(Y = 1|\text{Daytona}) + w_2\, P(Y=1|\text{Port Orange}) + w_3\, P(Y=1|\text{Edgewater}) + w_4\, P(Y=1|\text{Lake Helen}) + w_5\, P(Y=1|\text{DeLand}) + w_6\, P(Y=1|\text{Pierson})$$

$$\hat{\mu} = w_1 * \frac{e^{\gamma_0+\lambda_1}}{1+e^{\gamma_0+\lambda_1}} + w_2 * \frac{e^{\gamma_0+\lambda_2}}{1+e^{\gamma_0+\lambda_2}} + w_3 * \frac{e^{\gamma_0+\lambda_3}}{1+e^{\gamma_0+\lambda_3}} + w_4 * \frac{e^{\gamma_0+\lambda_4}}{1+e^{\gamma_0+\lambda_4}} + w_5 * \frac{e^{\gamma_0+\lambda_5}}{1+e^{\gamma_0+\lambda_5}} + w_6 * \frac{e^{\gamma_0+\lambda_6}}{1+e^{\gamma_0+\lambda_6}}$$

$$= 0.651470927$$

Since a person either will or will not favor the bill, the individual probabilities above can be considered a Bernoulli variable. Thus, the variance of the estimator is

$$V(\hat{\mu}) = w_1^2 * \frac{e^{\gamma_0+\gamma_1}}{1+e^{\gamma_0+\gamma_1}}\left(1-\frac{e^{\gamma_0+\gamma_1}}{1+e^{\gamma_0+\gamma_1}}\right) + w_2^2 * \frac{e^{\gamma_0+\gamma_2}}{1+e^{\gamma_0+\gamma_2}}\left(1-\frac{e^{\gamma_0+\gamma_2}}{1+e^{\gamma_0+\gamma_2}}\right) + w_3^2 * \frac{e^{\gamma_0+\gamma_3}}{1+e^{\gamma_0+\gamma_3}}\left(1-\frac{e^{\gamma_0+\gamma_3}}{1+e^{\gamma_0+\gamma_3}}\right) + w_4^2 * \frac{e^{\gamma_0+\gamma_4}}{1+e^{\gamma_0+\gamma_4}}\left(1-\frac{e^{\gamma_0+\gamma_4}}{1+e^{\gamma_0+\gamma_4}}\right) + w_5^2 * \frac{e^{\gamma_0+\gamma_5}}{1+e^{\gamma_0+\gamma_5}}\left(1-\frac{e^{\gamma_0+\gamma_5}}{1+e^{\gamma_0+\gamma_5}}\right) + w_6^2 * \frac{e^{\gamma_0+\gamma_6}}{1+e^{\gamma_0+\gamma_6}}\left(1-\frac{e^{\gamma_0+\gamma_6}}{1+e^{\gamma_0+\gamma_6}}\right) +$$

$$= 0.049390126$$

This is the variance estimator that we would like to maximize with regard to the number of calls and the sample size subject to constraints on cost and bias. I have used the spreadsheet to

calculate the cost and variance of the different callbacks stages, assuming that the first call costs $1 and the callbacks cost $0.50.

Number of Calls	Cost	Variance
1	$2715	.04774
2	$3419.5	.04763
3	$3789	.04901
4	$4058.5	.04901
5	$4297	.04939

Thus, if we knew a limit on cost, we could see which stage of the callback process fits within that constraint. Looking at bias as well, we would be able to determine the best number of calls to make.

Conclusions

Future work is needed in this area to accurately model the number of calls that should be made subject to the constraints on cost and on nonresponse bias. An actual bias estimate needs to be defined for Model 3 in order to optimize it and find estimates for the number of calls and the sample size. In expanding my model, researchers should try to develop a method to make estimations when the x values are not known in advance but the population means are known, since this is more likely the case. One approach to this would be to adjust the mean, taking the respondents into account, and then assume this x value for all nonrespondents. A more accurate method makes assumptions about the distribution of the x variable within the nonrespondents and creates x variables subject to this distribution.

It would also be more accurate to consider models where the probability of having a desired attribute is not independent of the probability of being contacted. This is more likely the case since people who are active, and thus hard to reach, may for example, vote differently than people who are often home. Thus, these are dependent upon each other. If they are dependent, more callbacks need to be made because the group not reached is less like the group reached. It is, in fact, in this scenario that callbacks are necessary.

Another area that could be further explored is the cost function. In my model the cost function is always a linear function. It may be more accurate to use equations for cost proposed by Elliot et al (2000) or Groves (1989) that were explored within the literature review.

It would also be useful to separate the refusals from the no answers/busy's from the disconnects/businesses since getting these results from different types of reasons and should be analyzed separately and the type of response is handled differently. Only the no answers/busy's are called again, and that is actually what my model attempts to predict.

References

Bartholomew, D.J. "A Method of Allowing for 'Not-At-Home Bias in Sample Surveys." *Applied Statistics* 10 (1961) 52-59.

Colombo, Richard. "A Model for Diagnosing and Reducing Nonresponse Bias." *Journal of Advertising Research* 40 (2000) 85-93. <<http:www.proquest.umi.com>>

Elliott, Michael R., Roderick J.A. Little, and Steve Lewitzky. "Subsampling Callbacks To Improve Survey Efficiency." *Journal of the American Statistical Association* 95:451 (2000) 730-738.

Everett, Diane and John Schorr. *Calling Response Codes.* Stetson Institute for Social Research, Stetson University. 2000.

Everett, Diane and John Schorr. *Daytona-Beach News Journal Political Issues Poll*. Stetson Institute for Social Research, Stetson University. 2000.

Groves, Robert M. *Survey Errors and Survey Costs*. New York: John Wiley & Sons, 1989.

Lohr, Sharon L. *Sampling: Design and Analysis*. Boston: Duxbury Press, 1999.

Map & Globe Stores, Inc. *Street Atlas Of East Central Atlantic Florida: Volusia And Flagler Counties.* Florida: Map and Globe Stores, Inc., c2000.

Potthoff, Richard F., Kenneth G. Manton, and Max A. Woodbury. "Correcting for Nonavailability Bias in Surveys by Weighting Based on Number of Callbacks." *Journal of the American Statistical Association* 88:424 (1993) 1197-1207.

Tuchfarber, Alfred J. and William R. Klecka. *Random Digit Dialing: Lowering the Cost of Victimization Surveys*. Cincinnati: University of Cincinnati, 1976.

Wackerly, Dennis D., William Mendenhall III, and Richard L. Scheaffer. *Mathematical Statistics with Applications.* California: Wadsworth Publishing Company, 1996.