Bayesian Approaches to Sequential Selection of Survey Design Protocols

James Wagner[1], Trivellore Raghunathan[2]
[1]Program in Survey Methodology, University of Michigan, Ann Arbor, MI
[2]Program in Survey Methodology & Dept of Biostatistics, University of Michigan, Ann Arbor, MI

Abstract

Surveys sequentially choose design protocols for contacting and interviewing cases. Most surveys follow very general protocols that minimally differentiate cases from each other. We develop a series of of models to help identify most efficient strategies for cases conditional on their fixed characteristics and history of previous attempts. These models use data from the sampling frame and call records to derive posterior probabilities of contact and interview. Data from an ongoing RDD survey are analyzed from this perspective and the results are presented.

KEY WORDS: Adaptive Design, Survey Design Protocols

1. Introduction

Quality survey data collected on a representative sample from a well-defined population are essential for empirical social science research. However, response rates have been steadily declining, especially in random digit dial surveys, calling into question the validity of social science inferences. Survey methodologists have proposed and experimentally evaluated many design features in an effort to improve response rates. There is a large literature, for example, on incentives. In these studies, the treatment is often given in the same amount or manner to all sampled units. The outcome of the experiment is a difference in response rates. These studies rarely investigate the impact on the quality of survey estimates.

The design features of surveys are analogous to the sequence of treatments and dosage in clinical trials. In the past, clinical trials often considered a single treatment or dosage in comparison to another treatment or dosage. However, there is an alternative model developing which statistically evaluates the impact of a sequence of multiple treatments and dosages. These sequences can be tailored to the characteristics of the individuals, including their history of previous treatment. This approach, adaptive treatment regimes, parallels more closely clinical practice. Proponents of this approach argue that defining a single treatment and dosage for all patients may not produce optimal results. These researchers contend that the treatments should be tailored to characteristics of individuals and their past history of treatment.

Surveys may also benefit from this approach. By considering the information available on cases before attempting contact and the information that is developed while in the process of attempting to interview, it may be possible to identify approaches that perform better (in terms of cost and response) than fixed design features. Resources may be more effectively deployed if this information is used adaptively to change the protocols in order to maximize the contact, screening, and interview propensities of sample subjects.

2. Background

Response rates have been suffering a long decline (de Leeuw and de Heer, 2002; Curtin, Presser, and Singer, 2005). This decline has been accompanied by increasing concerns that survey results will be biased. Although several recent articles have found that response rates are not necessarily linked directly to nonresponse bias (Keeter et al., 2000; Curtin et al., 2000; Merkle and Edelman, 2002), most surveys only have knowledge of their response rates and no knowledge of the differences between responders and nonresponders. They are, therefore, unable to directly address concerns about nonresponse bias.

Survey designers have responded by expending more effort to increase, or simply maintain, response rates. These adaptations have led to higher survey costs and increasing uncertainty about design parameters when new surveys are being developed.

In response to this growing problem, survey researchers have looked at all aspects of both the survey request and the decision to participate with the goal of finding ways to improve response rates (Groves and Couper, 1998). There is an expansive literature, for example, on the use of incentives (see Singer et al., 1999, for a review). Many other design features of surveys -- including prenotification letters, call scheduling, and wording of introductions (Link and Mokdad, 2005; de Leeuw et al., 2005; Greenberg and Stokes, 1990; Weeks et al., 1987; Houtkoup-Steenstra and van den Bergh, 2000) -- have been tested experimentally as means to improve response rates. Many of these studies have failed to find consistent results. Link and Mokdad (2005), for instance, review disparate results on the impact of prenotification letters in RDD surveys.

There is very little research into how these various design features interact with each other. Do different features produce different effects when they are combined together in different ways? Does the sequencing of these design features change their effect? Additionally, there is very little literature investigating whether these design features interact with demographic characteristics of respondents to produce different results for various subgroups (Groves, 2005).

Greenberg and Stokes' (1990) article on call scheduling for telephone surveys is an example of methodological research that did attempt to define a protocol based on the history of previous attempts to contact and interview a case. However, they only considered the timing of the call. They did not consider tailoring other aspects of the design. Further, they only made use of information about previous calls. They did not make use of other information on the frame, such as Census data for associated geographies. In addition, the callback rules generated by their strategy, since they did not differentiate cases more, were difficult to implement. They required 39% of the calls be placed during the day on the first day of the survey and that 19% of the calls be placed on the second day in the evening. Finally, these rules were generated seventeen years ago. The telephone system, patterns of telephone usage, and even patterns of being at home have changed substantially since then.

Groves, Singer, and Corning (2000) have proposed a theory (which they call Leverage-Saliency theory) that attempts to describe the process potential respondents use in deciding whether to participate in a survey. They argue that potential respondents place different values on the various aspects of the survey request. Surveys that make salient the aspects which are most important to each potential respondent will be more successful. For example, if the potential respondent is less interested in the topic, emphasizing the incentive payment may be a more effective strategy. The notion of tailoring introductions follows logically from this theoretical supposition (Groves and Couper, 1998; Morton-Williams, 1993).

The theory suggests that a one-size-fits-all approach should create inefficiencies in design. For example, reaching an answering machine several times without leaving a message might make potential respondents angry and decrease interview probabilities for households that make use of caller-ID. On the other hand, for households without caller-ID, calling several times before leaving a message, if contact has not been established, might be the most efficient strategy. The challenge in this example would be to identify correlates of the use of caller-ID. Given that different potential respondents will react to different design features in different – potentially even opposite – ways, it makes sense to consider tailoring all design features, including the ways in which they are combined, to the identified characteristics of respondents.

While survey methodology has yet to develop much research along these lines, these sorts of questions have been considered in relation to clinical trials. The result has been the development of new techniques broadly identified as adaptive or dynamic treatment regimes (Murphy, 2003; Murphy, 2005; Thall, Millikan, Sung, 2000; Lavori and Dawson, 2000; Lavori and Dawson, 2004; Collins et al., 2004; Collins et al., 2005). Under this emerging model, treatments are considered to be multi-course and the object is to identify the optimal combination and sequence of treatments for inducing the desired response. These regimes are often tailored not only to the outcomes of previous treatments, but to the characteristics of the patient, as well. Dynamic treatment regimes are tailored to individuals based on their efficacy and potential adverse outcomes, such as toxicity. Such regimes allow the dosage level and type of treatment to vary with time using rules specified before the beginning of treatment; these rules are based on time-varying measurements of subject-specific need.

Murphy (2003) denotes the model in the following useful manner. Each time interval j in {1,2,…,K} requires a treatment denoted a_j. The status at the beginning of time interval j is $\boldsymbol{S_j}$. $\boldsymbol{S_j}$ is a vector of predictors of the outcome for treatments available at time interval j. This may include fixed attributes of the individual being treated and time varying measures on the individual, including the history of previous treatments. The outcome of the analysis is a set of decision rules, one for each time point conditional on $\boldsymbol{S_j}$, such that the probability of response to the treatment is maximized for all individuals.

A relevant example of this approach is provided by Thall, Millikan, and Sung (2000). They consider competing, multi-course treatments for prostate cancer. In order to determine the best sequence of treatments, they estimate the salvage probabilities that one treatment has when another treatment has already failed. They also estimate the cross-resistance between consecutive treatments. Murphy (2003), in an investigation of interventions to improve reading skills, also considers how these optimal regimes might vary for demographic subgroups.

These authors note that these adaptive treatment regimes can be more useful since they more closely approximate clinical practice. In the "real world," when a treatment fails to produce a response, the clinician does not generally stop treatment; rather, the clinician will usually change the dosage or treatment in hopes of producing the desired result.

The statistical methods and design approaches developed for adaptive treatment regimes could benefit surveys. If we consider design features as treatments and completing the interview as the desired response, then the model translates directly into the conceptualizations of adaptive treatment regimes. Conditional on fixed covariates and previous treatments, the task is to find the most efficient next step.

This approach more closely parallels actual practice than experimental methods which consider only one or two design features. In practice, survey field operations do change their treatment regimes, if only in an ad hoc manner, as study objectives become endangered; for example, by raising incentives when response rate goals are not met or by allowing interviewers to use their judgment about the next step. These ad hoc decisions are likely to be implemented in a less-than-optimal fashion.

3. Data Analysis

We have undertaken the examination of observational data from the Survey of Consumer Attitudes (SCA), an ongoing RDD survey conducted each month by the Survey Research Center at the University of Michigan. Our goal is to locate the next protocol with the highest probability of contact conditional on the fixed covariates of cases as well as previous protocols administered to those cases.

SCA collects 300 RDD interviews per month. The main statistic produced by the survey is the Index of Consumer Sentiment. The data we analyze here include the history of all calls made, including the time, date, interviewer, offer of incentives, and the result obtained. The sample was developed using the Genesys sampling system. Genesys links telephone exchanges to Census data to provide “context” data for all sampled telephone numbers. These data include information about the age, income, and race distributions of the population associated with the exchange of the sampled telephone number as well as information about urbanicity, housing, and other characteristics of the estimated geography. These types of data have been used by others to estimate contact, screening, and interview probabilities for use in weighting adjustments (Lu, Hall, and Williams, 2002; Johnson et al., 2006).

One limitation of an analysis of observational data is that we can only consider the methods that were actually employed by the process that created these data. The variation in the data will be limited by the calling rules and design features that are the current practice. We cannot estimate outside the range of these data. Fortunately, very general rules have been employed for the collection of these data, and, hence, the variation in the data is quite large.

Following the dynamic treatment regime approach in clinical trials, we hypothesize that there may be interactions between demographic characteristics of sample members and the design features used which impact the probability of contact. We also hypothesize that the sequence and combination of protocols may lead to differing contact probabilities.

In order to assess these interactions, we adopt the following two-stage strategy. In the first stage, we use the history prior to the call attempt *j* and the context data, to match the respondents and nonrespondents. We do so by creating strata based on propensity scores. In the second stage, we fit propensity models for contact at call attempt *j* conditional on the specific protocol used at this call attempt. This model is estimated within each propensity stratum. This second-stage propensity model is used to determine the configuration of the protocols that yields the maximum predicted probability for all subjects in each stratum. This configuration is perhaps the most efficient protocol that would have maximized the propensity of contact. We use the posterior distribution of the coefficients from this second-stage model to assess the probability that a given strategy is the contact propensity maximizing strategy.

In the first stage, we used the context data and information about previously used protocols to create propensity strata (Rosenbaum and Rubin, 1983). Let X_{ij}^- denote a $k_j \times 1$ vector of covariates for subject *i* prior to the call attempt *j* or at the end of the previous call attempt. Let R_{ij}^- denote the contact status (1: contact, 0: no contact) on the corresponding subject for the call attempt *j*. The propensity scores were estimated using a logistic regression model of the following form: $p_{ij}^- = logit(\Pr(R_{ij}^- = 1 \mid X_{ij}^-, \alpha_j)) = X_{ij}^- \alpha_j$ where p_{ij}^- is the propensity score to match the respondents and nonrespondents prior to the next call attempt. The list of predictor variables (X_{ij}^-) are listed in Table 1. The predicted probabilities were then divided into quintiles to create matched strata.

Table 1. Contact Propensity Predictor Variables (X_{ij}^{-})

Context Variables	Previous Call Information
Listed/Letter Sent	Call 1: Weekday Day
% Exchange Listed	Call 1: Weekend
Household Density	Call 1: Ans Mach
Median Yrs Education	Call 1: Incentive Offer
Log(Income)	Call 1: Left Message
Census Region 2	Call 2+: Weekday Day
Census Region 3	Call 2+: Weekend
Census Region 4	Call 2+: Answering Machine
% 18-24	Call 2+: 1 Day after Previous Call
% 25-34	Call 2+: 2 Days after Previous Call
% 35-44	Call 2+: 3 Days after Previous Call
% 45-54	Call 2+: 4 Days after Previous Call
% 55-64	Call 2+: 5+ Days after Previous Call
% 65+	Call 2+: Incentive Offer
% White	Call 2+: Left Message
% Black	
% Hispanic	
% Owner Occupied	

For the second-stage model, let S_{ij} denote a vector of variables describing the protocol (treatment) used at call attempt *j*. Here we define protocols as the set of design features that are combined together for each attempt at contact. We then estimated the contact probabilities at call *j*, within each propensity stratum conditional on the specific protocol used during call *j*. Let R_{ij} be the outcome at call *j* for subject *i*, taking the value 1 if the attempt was successful and 0 otherwise.

We estimated the probability of contact using logistic regression, $logit(\Pr(R_{ij} = 1 \mid S_{ij})) = S_{ij}\beta_j$. Table 2 shows the predictor variables S_{ij} that were used to define the protocols in these models. There are three call windows, six lag times between calls, and three approaches to answering machines giving a total of 54 (3x6x3=54) possible strategies. A message left on a previous call is considered to be part of the strategy for the next call since that is the call most likely to be impacted by a message. In other words, a message left on the first call is a part of the protocol for the second call.

We estimated these models for each of the propensity strata. Past data were used to estimate the posterior distribution of β_j. This approach allows us to identify different highest probability strategies for cases with different fixed characteristics (including previous effort). In this approach, the previous effort is not incorporated in the S_{ij} , but in the X_{ij}^{-} used for the propensity stratification models previously described. Our task is to identify the set of S_{ij} that maximize the probability of contact within each propensity stratum.

Table 2. Predictor Variables Defining the Protocols (S_{ij})

Variable	Description
Timing of Call	Day of week and time of day; 3 different windows
Time between calls	Same day, 1, 2, 3, 4, or 5 or more days
Answering Machine/Incentive	No Message, Message, Message with Incentive

The posterior distribution of the coefficients in the logistic model was simulated using the Metropolis algorithm. A normal prior on β_j was assumed. Three chains of 10,000 iterations were run with 1,000 burn-in iterations. We used the Gelman-Rubin statistic to judge the convergence of these chains. This statistic was 1.05 at its highest and was generally less than or equal to 1.01.

Figure 1 shows the average posterior probability (light-colored bars above the x-axis) for each of these protocols on the second call for three of the propensity strata. In the lowest propensity stratum, there is a great deal of variation in the probabilities, suggesting that the protocol does matter. In the highest stratum, on the other hand, there is very little variation in the success of the various protocols. This provides some confirmation of the Leverage-Saliency theory. The same protocol, administered to two different propensity strata, can have very different results. Thus, a tailored approach should improve the efficiency of the effort to contact sampled units.

The dark blue bars below the x-axis show the proportion of cases that received each protocol. It seems that the protocols that were actually used are tailored to the average or highest propensity cases. A strategy that is tailored to the lowest propensity stratum might do better in that stratum.

It is encouraging to note that there are simple changes that could be made to produce efficiencies in establishing contact. These changes involve changing the timing of calls and the delays between calls. Since these changes sometimes go in the opposite direction for cases in different propensity strata, there is hope that they would

Figure 1. Average Posterior Probability of Contact on the Second Call for Each Protocol

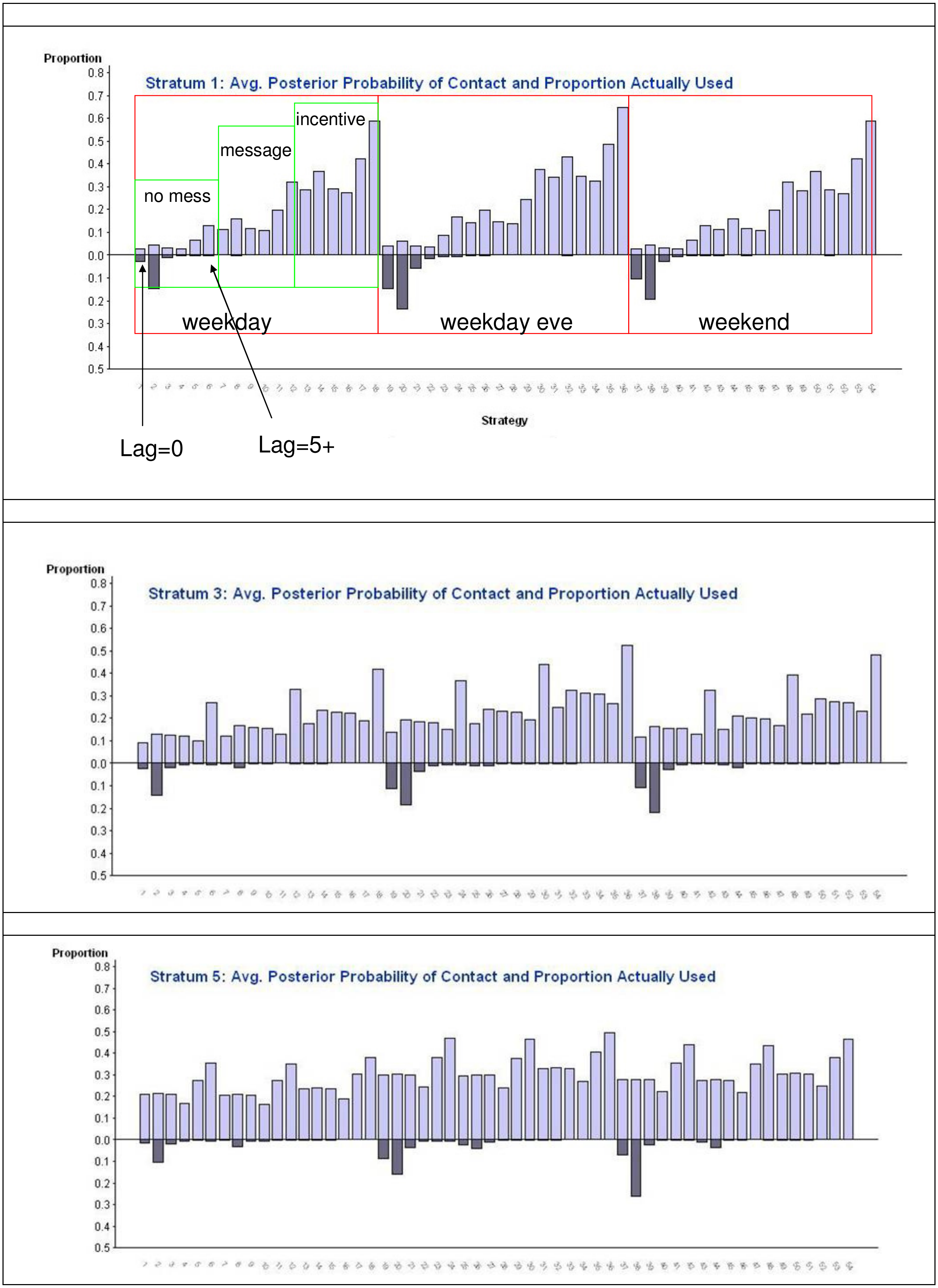
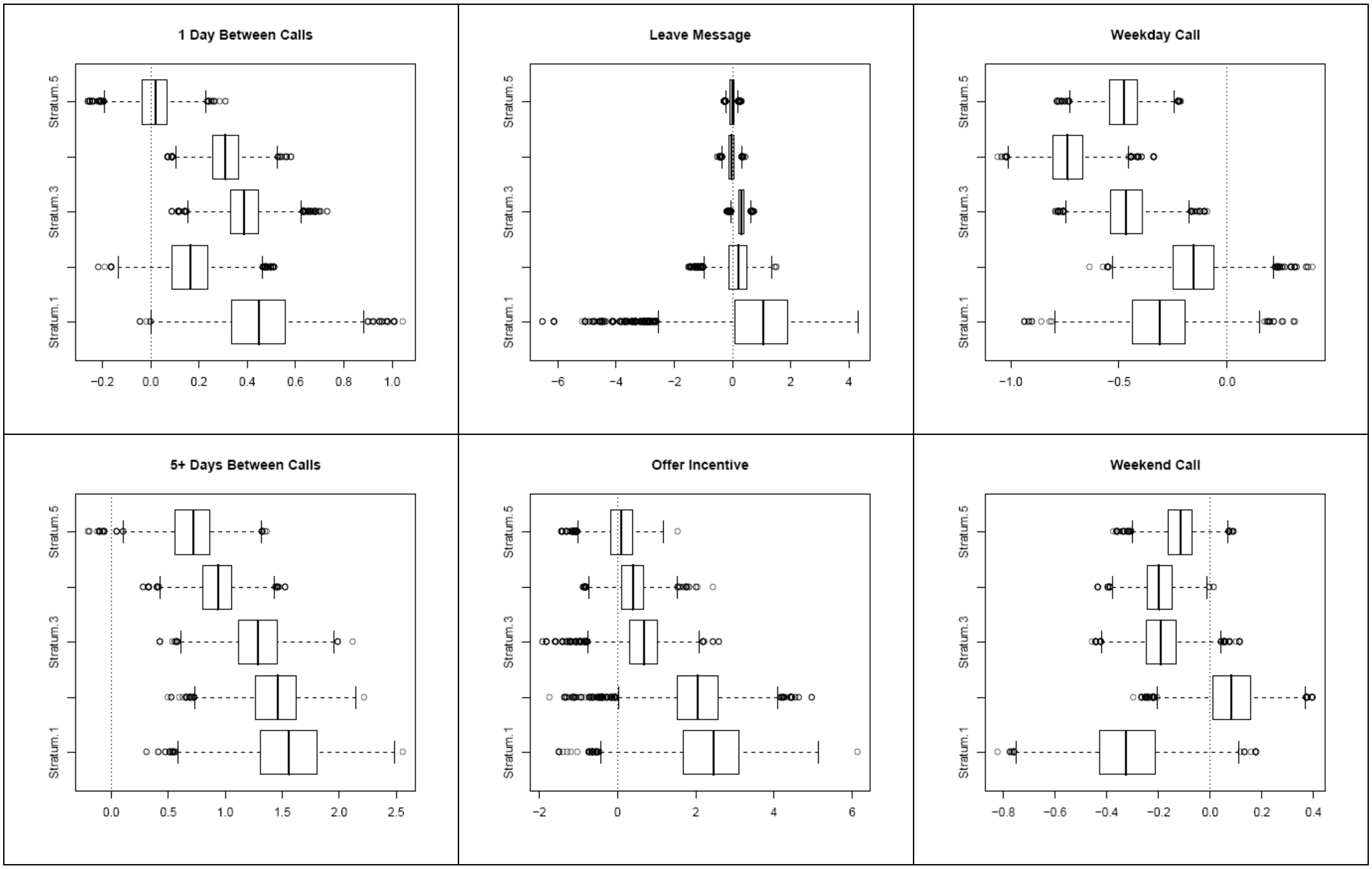


Figure 2: Posterior Distribution of Coefficients from the Second Call Model by Propensity Stratum



not require impossible staffing plans in order to be implemented. A rule that said call every case on the first day and then call all nonfinal cases back on the second day, for example, would be difficult to implement for just this reason.

Figure 1 does not clearly illustrate whether the best strategies for achieving contact differ across the strata. Figure 2 does show the differences in strategy by stratum. In this figure, a dotted line indicates where zero is. Coefficients to the right of the line represent protocols that tend to increase the probability of contact.

It seems that leaving a message has a positive impact on the probability of contact in the lowest propensity stratum and a negative impact in the other strata. The posterior probability that a message left on an answering machine will increase the probability of contact for a case in stratum 1 – the lowest probability stratum – (i.e., $\Pr(\beta_{Str1,AM} > 0)$) on the next call is 0.76 in stratum 1. The posterior probability for this protocol is 0.41 in stratum 5, the highest probability stratum. A similar result can be seen for placing a weekend call. This is a better strategy for lower propensity cases (propensity stratum 2).

Table 3 summarizes the posterior probabilities that a coefficient is positive for the second call protocols. For some protocols, we do not see as marked a differentiation in best strategies across the strata. In general, waiting longer between calls seems to have a higher probability of contact.

These results may be refined by matching strategies to specific covariates rather than the propensity score, which is a scalar summarization of all the covariates. Such refinements may help reduce the variability in the estimated coefficients seen in Figure 2.

Table 3. Posterior Probability that a Coefficient From the Second Call Model is Positive

Stratum	Weekday	Weekend	Message	Incentive	Days Between Calls				
					1	2	3	4	5+
1	0.03	0.02	0.76	0.99	1.00	0.54	0.42	0.90	1.00
2	0.14	0.78	0.66	0.99	0.94	0.99	0.59	0.95	1.00
3	0.00	0.02	0.99	0.88	1.00	1.00	0.90	0.58	1.00
4	0.00	0.00	0.42	0.83	1.00	1.00	0.94	0.93	1.00
5	0.00	0.04	0.41	0.61	0.60	0.52	0.08	0.90	1.00

These analyses indicate that there are efficiencies to be gained. Our hypothesis that different protocols will interact with demographic characteristics of potential respondents as well as with previous treatments were at least partially confirmed. In addition, given the variety of best treatments and the nearness (in terms of probability) of next best treatments, these data suggest that a partial set of rules based on the results of these analyses would be operationally feasible.

This is all based on the relatively “thin” data available about RDD telephone numbers before contact. It seems likely that the data available after contact has been established, or in face-to-face surveys, would be even more powerfully applied under this approach.

4. A Learning Algorithm

Our analyses are based on observational data. These analyses do, however, suggest a useful experimental strategy. We propose using the posterior densities of the model coefficients to randomly assign strategies to cases. Strategies would be assigned with probability proportional to the probability that the strategy is the strategy with the highest probability of achieving contact.

So, for example, if in propensity stratum 2, there is a 78% chance that a weekend call would do better than a call placed at another time, then 78% of the cases in this stratum would be randomized to a weekend call. If the results are confirmed by the experiment, then when the posterior distributions are updated with the new data, the proportion of sample assigned to this protocol in the next iteration of the survey will grow.

This approach could be used even when very little or no data are available prior to fielding a survey. Weak priors could be used in order to give higher probability to strategies that are assumed to be better (e.g., weekend and evening calling). A weak prior would allow the experiment to explore strategies.

This approach implies learning at two levels. First, in the short term, it could be used to identify experimentally strategies that are tailored to the specifics of the case. The strategy would hone in on the best strategies as data accumulate. Second, in the long term, the strategy could adapt to a changing social environment in which the best strategy of today may not be the best strategy of tomorrow.

5. Conclusion

Research into survey design methods has focused on particular design features — for example, incentives, prenotification letters, etc. There has been very little research on how these features have varying effects on different subpopulations and almost no research on how different combinations of these design features might impact different subpopulations. We have attempted to demonstrate how this new conceptualization of research into survey methods might be implemented.

We have shown some evidence that different subgroups react differently to different combinations of design features. A further extension of our research would be to identify specific characteristics of respondents that predict variation in contact rates. Another extension would consider the design features from all the attempts as a single protocol rather than a call-by-call approach.

This reconceptualization may also allow us to reintegrate seemingly contradictory results. For instance, contradictory evidence regarding the impact of prenotification letters on response rates may be the result of either differences in sequencing or interactions with demographic characteristics of respondents.

Research along these lines may impact survey practice by providing a framework to choose appropriate design strategies for efficient data collection based on the statistical analysis of past data. Such a framework, in combination with measures of data quality other than the response rate may allow us to focus on bringing in lower responding groups for whom the “average design” is less effective. This has the potential for improving the quality of data for social science research.

References

Collins, L. M., S. A. Murphy, et al. (2004). "A conceptual framework for adaptive preventive interventions." Prev Sci **5**(3): 185-96.

Collins, L. M., S. A. Murphy, et al. (2005). "A strategy for optimizing and evaluating behavioral interventions." Ann Behav Med **30**(1): 65-73.

Curtin, R., S. Presser, et al. (2000). "The Effects of Response Rate Changes on the Index of Consumer Sentiment." Public Opin Q **64**(4): 413-428.

Curtin, R., S. Presser, et al. (2005). "Changes in Telephone Survey Nonresponse over the Past Quarter Century." Public Opin Q **69**(1): 87-98.

de Leeuw, E. and W. de Heer (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. Survey Nonresponse. R. M. Groves. New York, John Wiley & Sons**:** 41-54.

de Leeuw, E., J. Hox, et al. (2005). "The Influence of Advance Letters on Response in Telephone Surveys: A Meta-Analysis." Working Paper series of the Program in Survey Research and Methodology(10): 1-26.

Greenberg, B. S. and S. L. Stokes (1990). "Developing an Optimal Call Scheduling Strategy for a Telephone Survey." Journal of Official Statistics **6**(4): 421-435.

Groves, R. M. (2005). “Research synthesis: nonresponse rates and nonresponse error in household surveys.” 16th International Workshop on Household Survey Nonresponse, Tällberg, Sweden:28-31.

Groves, Robert M., and Mick Couper (1998). Nonresponse in household interview surveys. New York: Wiley.

Groves, R. M., E. Singer, et al. (2000). "Leverage-Saliency Theory of Survey Participation: Description and an Illustration." Public Opin Q **64**(3): 299-308.

Houtkoop-Steenstra, H. and H. van den Bergh (2000). "Effects of Introductions in Large Scale Telephone Survey Interviews." Sociological Methods Research **28**(3): 281-300.

Johnson, Timothy P., Young I. K. Cho, Richard T. Campbell, and Allyson L. Holbrook (2006). “Using Community-Level Correlates to Evaluate Nonresponse Effects in a Telephone Survey.” Public Opin Q 70 (5):704-719.

Keeter, S., C. Miller, et al. (2000). "Consequences of Reducing Nonresponse in a National Telephone Survey." Public Opin Q **64**(2): 125-148.

Lavori, P. W. and R. Dawson (2004). "Dynamic treatment regimes: practical design considerations." Clin Trials **1**(1): 9-20.

Lavori, P. W., R. Dawson, et al. (2000). "Flexible treatment strategies in chronic disease: clinical and research implications." Biol Psychiatry **48**(6): 605-14.

Link, M. W. and A. Mokdad (2005). "Advance Letters as a Means of Improving Respondent Cooperation in Random Digit Dial Studies: A Multistate Experiment." Public Opin Q **69**(4): 572-587.

Lu, R. C., J. Hall, et al. (2002). "Resolvability, Screening, and Response Models in RDD Surveys: Utilizing Genesys Telephone-Exchange Data." JSM-SRMS Conference Proceedings **2002**: 2198-2202.

Merkle, D. M. and M. Edelman (2002). Nonresponse in Exit Polls: A Comprehensive Analysis. Survey Nonresponse. R. M. Groves. New York, John Wiley & Sons**:** 243-257.

Morton-Williams, J. (1993). Interviewer Approaches, Aldershot, Hants, England

Murphy, S. A. (2003). "Optimal dynamic treatment regimes." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **65**(2): 331-355.

Murphy, S. A. (2005). "An experimental design for the development of adaptive treatment strategies." Stat Med **24**(10): 1455-81.

Rosenbaum, Paul R., and Donald B. Rubin (1983). “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” Biometrika 70 (1):41-55.

Singer, E., J. V. Hoewyk, et al. (1999). "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys." Journal of Official Statistics **15**(2): 217-230.

Thall, P. F., R. E. Millikan, et al. (2000). "Evaluating multiple treatment courses in clinical trials." Stat Med **19**(8): 1011-28.

Weeks, M. F., R. A. Kulka, et al. (1987). "Optimal Call Scheduling for a Telephone Survey." Public Opin Q **51**(4): 540-549.