Evaluation of Randomization-Based Estimation and Inference Methods for Survey Data Subject to Nonresponse, Callbacks and Mode Effects October 2009

Randall Powers and John L. Eltinge
Bureau of Labor Statistics, 2 Massachusetts Ave NE Room 1950, Washington, DC 20212

Abstract: In many surveys, field procedures address nonresponse with a combination of callback efforts and changes in the mode of data collection. To analyze the resulting data, one generally needs to account for the relevant features of the underlying population, the sample design, and the nonresponse follow-up plan. This paper suggests a relatively simple approach based on random assignment of sample units to distinct groups. Each group receives different treatments defined by, e.g., varying numbers of callbacks, different contact and interview modes, different response incentives, and different levels of interviewer training. Some properties of the proposed methods are evaluated through a simulation study.

Key Words: Design optimization; Fixed nonresponse; Gold standard; Main-effects model; Responsive design; Subsampling

1. Introduction

In many household and establishment surveys, initial attempts to collect data from a selected sample unit may be unsuccessful. Survey organizations have developed a variety of strategies to address this issue. These strategies generally involve a combination of callbacks; changes in the mode of attempted contact or collection; assignment of the sample case to an interviewer specially trained in working with reluctant respondents; or use of incentives, burden reduction or other special methods to persuade the unit to participate in the survey.

The reference list for this paper covers some of the previous literature related to callbacks and the conversion of reluctant respondents. A detailed literature review is beyond the scope of the current work. However, it is worth noting that the papers included in the reference list exhibit a range of approaches to (a) the specific strategies applied to obtain information from nonresponding sample units; (b) the information used to adjust for residual nonresponse (and related effects of collection mode or other factors) after best efforts to apply the strategy used in (a); and (c) the extent to which subsequent inferences are based on, respectively, models or explicit randomization mechanisms employed in the implementation of (a).

The current work will focus on a relatively simple approach in which sample units are assigned randomly to distinct treatment groups defined by alternative strategies identified in (a); the information in (b) is restricted to data from the original sample design and from the random assignment of treatments; and inferences for (c) are based primarily on the randomization mechanisms in the original sample design and the treatment assignment, as well as some important moment restrictions. Section 2 outlines the primary steps in the proposed method. Detailed development of the properties of the proposed method are beyond the scope of the current work, and will be considered in a separate paper. Section 3 outlines a simulation study to evaluate the properties of a relatively simple form of the proposed method. Section 4 presents numerical results of the simulation study.

2. Notation, Callback Design and Estimators

We consider a finite population U of size N, from which we select a sample S of size M. To simplify notation, we will restrict attention here to selection of S through a equal-probability selection methods. Our goal is to estimate the population mean

$$\bar{Z} = N^{-1}\sum_{i=1}^{N} Z_i \tag{2.1}$$

where Z_i is a characteristic associated with population unit i.

The current paper will restrict attention to two modes for contact and collection of data from our sample units. The first mode, through personal visit, will be treated as a "gold standard" in which the probability of response equals one, and for which the interviewer records the true value Z_i. The second mode is through the telephone; we will treat telephone collection as being subject both to nonresponse and reporting errors. Specifically, we will use a *fixed* nonresponse model, in which each unit i has an integer value $a(i)$ such that unit i will respond to telephone contact attempt $a(i)$, but not earlier. In addition, when unit i does respond to a telephone contact, it reports the error-prone value Y_i, where

$$Y_i = Z_i(1+\delta_i) = Z_i + Z_i\delta_i \tag{2.2}$$

and the terms $Z_i\delta_i$ represent multiplicative errors.

For a given positive integer C, the fixed values $a(i)$ lead to a partition of the full population

$$U = \bigcup_{c=1}^{C+1} U_c \tag{2.3}$$

where for $c = 1, \ldots C$, U_c is the subpopulation of units that will first respond to telephone call attempt c. In addition, U_{C+1} is the subpopulation of units that will not respond to any of the first C telephone call attempts. For each $c = 1, \ldots C+1$, define N_c to be the size of the subpopulation U_c, let $\pi_c = N_c / N$,

$$(\bar{Z}_c, \bar{\delta}_c, \bar{\delta}_c^{*}) = N_c^{-1}\sum_{i \in U_c}(Z_i, \delta_i, Z_i\delta_i) \tag{2.4}$$

In addition, we assume that for each c,

$$\bar{\delta}_c^{*} \approx \bar{Z}_c \bar{\delta}_c \tag{2.5}$$

Condition (2.5) is a finite-population variant on the common model-based assumption that the multiplicative error factors are independent of the underlying true values.

Conditional upon selection of the sample S, we assign each sample unit to a group $D = 0,1,\ldots C$. For each sample unit in group D, we will make up to D attempts to collect data through the telephone. If we do not receive a telephone response by the end of attempt D, we will collect data through the "gold standard" personal-visit method on attempt $D+1$. For each $D = 0,1,\ldots C$ and $c = 1,\ldots D$, we define S_D to be the set of sample units assigned to group D, S_{Dc} to be the group of units in S_D that respond on telephone attempt c, $S_{D,D+1}$ to be the group of units in S_D that do not respond to any of the first D call attempts. In addition, define M_D to be the size of S_D, M_{Dc} to be the size of S_{Dc}, and $M_{D,D+1}$ to be the size of $S_{D,D+1}$.

Under the stated conditions, we also define

$$\bar{\pi}_c^* = (\sum_{D=c}^{C+1} M_D)^{-1} \sum_{D=c}^{C+1} M_{Dc} \qquad (2.6)$$

which is approximately unbiased for π_c,

$$\hat{\bar{Y}}_c = (\sum_{D=c}^{C+1} M_{Dc})^{-1} \sum_{D=c}^{C+1} M_{Dc} \hat{\bar{Y}}_{Dc} \qquad (2.7)$$

which is approximately unbiased for $\bar{Z}_c + \bar{\delta}_c^*$, and

$$\hat{\bar{Z}}_D^* = M_{D,D+1}^{-1} \sum_{i \in S_{D,D+1}} Z_i \qquad (2.8)$$

which is approximately unbiased for $(\sum_{c=D+1}^{C+1} \pi_c)^{-1} \sum_{c=D+1}^{C+1} \pi_c \bar{Z}_c$.

Taken together, expressions (2.6)-(2.8) define a set of nonlinear estimating equations for the unknown subpopulation quantities π_c, $\bar{Z}_c$ and $\bar{\delta}_c^*$. Under the additional condition that the multiplicative error terms $\bar{\delta}_c^*$ are approximately constant across the first C subpopulations, standard nonlinear least-squares procedures lead to point estimators that we will denote $\hat{\pi}_c$, $\hat{\bar{Z}}_c$ and $\hat{\bar{\delta}}_1^*$.

3. Design of the Simulation Study

For this study, we defined the true subpopulation proportions to be $\pi_c = (1-p)^{c-1} p$ for each $c = 1,\ldots C$, and $\pi_{C+1} = 1 - \sum_{c=1}^{C} \pi_c$. We considered separate cases with p equal to 0.1, 0.2 and 0.5, respectively. In addition, the true values Z_i were distributed with a mean of 0 and a variance of 10. Five different cases for our measurement errors δ_i with differing means and variances were considered. The first four cases involve nonzero means. This is important because some of the literature on mode effects indicates that for some survey items, measurement errors may have nonzero means for inexpensive collection modes like the telephone. Case A involves a fairly well behaved mean and variance. For Case B, the variance was increased, whereas for Case C, the mean was increased. For Case D, our mean varies across the number of callback attempts made. For Case E, a measurement error mean of 0 was considered. (See Table 0.)

4. Numerical Results

When dealing with nonlinear estimators, we need to be careful to determine whether our optimization procedures are actually converging for as many cases as possible. In Table 1, we display convergence rates (out of 1000 replications) for *c=2* and *c=5*, separately for estimation using PROC NLP and PROC NLIN respectively. For PROC NLP, the LSQ statement was used, meaning the least squares were computed. The cov=2 option was used to compute the covariance matrices. The maximum number of iterations was specified as 100. For PROC NLIN, the Newton iterative method was specified, and the maximum number of iterations was also set at 100. The rows report the convergence rates for Cases A through E respectively. From this, we can see that PROC NLP performed much better, and PROC NLIN had a high number of nonconvergers. Table 2 indicates further convergence problems for PROC NLIN for *c=2*, both for *p=0.2* and p=0.5.

In Table 3, we report the mean and standard deviation of our estimators z_1 through z_3 as well as delta for Case A with *p=0.1*. For Case A, we see that the z's tend to be reasonably close to the true mean of 0, whereas the true mean of δ should be 1. Our resulting value for the mean of delta and the large standard deviation show that our method of moment estimator for δ is relatively unstable. For Case B (see Table 4), we see a pattern similar to that for Case A: The z estimators are approximately unbiased, and the δ estimator is still unstable.

Cases C, D and E (Tables 5, 6, and 7, respectively) showed similar results. For each of these cases, the mean estimators for the z values are all approximately unbiased, despite the presence of a nonzero mean of the measurement errors δ . In other words, combination of data from the error-prone telephone interviews and the gold-standard interviews has led to approximately unbiased estimators for the subpopulation means. All of the cases still display substantial variability in the delta estimator. Changing the probability to 0.5 for Case E (Table 8) gave us a more modest but still problematic standard deviation for delta. When we look at a result for *c=5* (Table 9) for p=0.1, we can see that the value δ_1 improves considerably compared to *c=2* in Table 7.

We wished to examine the cause of these large standard deviations. Were there a large number of replicates that were poorly behaved and contributing to the large standard deviation, or was it a small number of poorly behaved replicates that were in a sense so bad that they alone were the cause of the high standard deviation? To do this, we looked at the quantiles for δ_1 for the converging runs in three different runs. We did this for Case E, where the true mean for delta=0. In Table 10, the second through the fourth columns cover three distinct forms of Case E, with different values of C and p. The rows in the table show the tail and central quantiles for the delta estimators. You'll notice that for each of the three cases, the 10th through the 90th percentiles are reasonably well behaved, but the first and 99th percentiles are fairly extreme, Thus, the tails are what are responsible for the high standard deviation.

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics. The authors thank Alan Dorfman, Steve Miller, and Daniell Toth for many helpful comments on an earlier version of this work.

References

Bethlehem, J. G. and Kersten, H. M. P. (1985). On the treatment of nonresponse in sample surveys. *Journal of Official Statistics*, 1, 287-300

Biemer, Paul and Wang, Kevin (2007). Using callback models to adjust for nonignorable nonresponse in face-to-face surveys. *ASA Proceedings of the Joint Statistical Meetings*, 2889-2896.

Croner, Charles M., Williams, Paul D. and Hsiung, Sue (1985). Callback response in the National Health Interview Survey. *ASA Proceedings of the Section on Survey Research Methods*, 164-169.

Deming, W.E. (1953). On a probability mechanism to attain an economic balance between the resultant error of response and the bias of nonresponse. *Journal of the American Statistical Association* 48, 743-772.

Drew, J. H. and Fuller, W. A. (1980). Modeling nonresponse in surveys with callbacks. *ASA Proceedings of the Section on Survey Research Methods*, 639-642.

Drew, J. H. and Ray, G. (1984). Some empirical investigations of nonresponse in surveys with callbacks. *ASA Proceedings of the Section on Survey Research Methods*, 560-565.

Elliott, Michael R., Lewitzky, Steven and Little, Roderick J. A. (1997). Subsampling callbacks to reduce survey costs. *ASA Proceedings of the Section on Survey Research Methods*, 490-495.

Elliott, Michael R., Little, Roderick J. A. and Lewitzky, Steve (2000). Subsampling callbacks to improve survey efficiency. *Journal of the American Statistical Association*, 95, 730-738.

Groves, Robert M. and Heeringa, Steven G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs
Journal of the Royal Statistical Society, Series A: Statistics in Society, 169, 439-457.

Hansen, M.H. and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association* 41, 517-529.

Harpuder, Brian E. and Stec, Jeffery A. (1999). Achieving an optimum number of callback attempts: Cost-saving vs. non-response error due to non-contacts in RDD surveys. *ASA Proceedings of the Section on Survey Research Methods*, 913-918.

Kalsbeek, William D., Botman, Steven L. and Massey, James T. (1989). Cost-efficiency and the number of allowable callbacks in the National Health Interview Survey. *ASA Proceedings of the Section on Survey Research Methods*, 434-439

Kuk, Anthony Y. C., Mak, T. K. and Li, W. K. (2001). Estimation procedures for categorical survey data with nonignorable nonresponse. *Communications in Statistics: Theory and Methods*, 30, 643-663.

Marckwardt, Albert M. (1984). Response rates, callbacks and coverage: The WFS experience (No. 55). *World Fertility Survey Scientific Reports.*

Merkle, Daniel M., Bauman, Sandra L. and Lavrakas, Paul J. (1993). The impact of callbacks on survey estimates in an annual RDD survey. *ASA Proceedings of the Section on Survey Research Methods*, 1070-1075.

Park, Hyeonah, Na, Seongryong and Jeon, Jongwoo (2008). Estimation using response probability under callbacks. *Statistics & Probability Letters*, 78, 1735-1741.

Polizt, A. and Simmons, W. (1949). An attempt to get the “not at homes” into the sample without callbacks. *Journal of the American Statistical Association* 44, 9-16.

Potthoff, Richard F., Manton, Kenneth G. and Woodbury, Max A. (1993).
Correcting for nonavailability bias in surveys by weighting based on number of callbacks. Journal of the American Statistical Association, 88, 1197-1207.

Proctor, Charles H. (1977). Two direct approaches to survey nonresponse: Estimating a proportion with callbacks and allocating effort to raise response rate
ASA Proceedings of the Social Statistics Section, 284-290.

SAS Institute Inc. (2009a). SAS Online Documentation 9.2, Cary, NC: SAS Institute Inc.
http://support.sas.com/documentation/cdl/en/ormpug/59679/HTML/default/nlp.htm
Retrieved on September 25, 2009.

SAS Institute Inc. (2009b). SAS Online Documentation 9.2, Cary, NC: SAS Institute Inc.
http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/nlin_toc.htm
Retrieved on September 25, 2009.

Stec, Jeffery A., Lavrakas, Paul J. and Shuttles, Charles W. (2004). Gaining efficiencies in scheduling callbacks in large RDD national surveys. *ASA Proceedings of the Joint Statistical Meetings*, 4430-4437.

Stokes, S. Lynne and Greenberg, Betsy S. (1990). A priority system to improve callback success in telephone surveys. *ASA Proceedings of the Section on Survey Research Methods*, 742-747.

Table 0: Five Cases Considered

Means and Variances of Measurement Errors

Case	Mean	Variance
A	1	1
B	1	10
C	10	1
D	cx5	1
E	0	1

Table 1: Convergence Rates for p=0.1:

Convergence Rates
(out of 1000 replications) for p=0.1;
Default stopping rule: 100 iterations

Case	C=2		C=3		C=5	
	NLP	NLIN	NLP	NLIN	NLP	NLIN
A	0.993	0.936	0.993	0.924	1.000	0.898
B	0.992	0.889	0.991	0.847	1.000	0.787
C	0.992	0.901	0.982	0.838	1.000	0.714
D	0.991	0.895	0.981	0.845	1.000	0.743
E	0.997	0.953	0.993	0.946	1.000	0.953

Table 2: Convergence Rates for C=2, p=0.2 and 0.5

Convergence Rates
(out of 1000 replications) for c=2, p=0.2 and 0.5; Default
stopping rule: 100 iterations

	C=2,p=0.2		C=2, p=0.5	
Case	NLP	NLIN	NLP	NLIN
A	0.996	0.956	0.995	0.962
B	0.996	0.885	0.995	0.895
C	0.986	0.921	0.996	0.958
D	0.990	0.887	0.989	0.888
E	0.999	0.964	1.000	0.964

Table 3: Mean and Standard Deviation for C=2, Case A, p=0.1

C=2 Case A: error mean=1, p=0.1, PROC NLP
999 out of 1000 runs converged

Variable	Mean	Standard Deviation
Z_1	-0.1278	3.1867
Z_2	0.0055	3.1395
Z_3	0.0146	3.8087
δ_1	-0.8231	141.1515

Table 4: Mean and Standard Deviation for C=2, Case B, p=0.1

C=2 Case B: error mean=1, p=0.1, PROC NLP
992 out of 1000 runs converged

Variable	Mean	Standard Deviation
Z_1	0.0597	3.1074
Z_2	0.0665	2.7874
Z_3	0.0947	3.5900
δ_1	0.5989	263.8131

Table 5: Mean and Standard Deviation for C=2, Case C, p=0.1

C=2 Case C: error mean=10, p=0.1, PROC NLP
992 out of 1000 runs converged

Variable	Mean	Standard Deviation
Z_1	-0.0299	2.4316
Z_2	-0.0148	2.2753
Z_3	0.2174	3.7847
δ_1	-2.8880	162.9991

Table 6: Mean and Standard Deviation for C=2, Case D, p=0.1

C=2 Case D: error mean=cx5, p=0.1, PROC NLP
991 out of 1000 runs converged

Variable	Mean	Standard Deviation
Z_1	-0.0570	1.8077
Z_2	0.0136	2.9300
Z_3	0.1283	3.5553
δ_1	-0.7474	128.3369

Table 7: Mean and Standard Deviation for C=2, Case E, p=0.1

C=2 Case E: error mean=0, p=0.1, PROC NLP
997 out of 1000 runs converged

Variable	Mean	Standard Deviation
Z_1	-0.0187	3.2039
Z_2	0.1399	3.1008
Z_3	0.0989	3.9023
δ_1	-0.7867	117.6986

Table 8: Mean and Standard Deviation for C=2, Case E, p=0.5

C=2 Case E: error mean=0, p=0.5, PROC NLP
993 out of 1000 runs converged

Variable	Mean	Standard Deviation
Z_1	-0.0041	3.0863
Z_2	0.0985	3.1742
Z_3	-0.0509	3.2976
δ_1	-0.6335	55.5745

Table 9: Mean and Standard Deviation for C=5, Case E, p=0.1

C=5 Case E: error mean=0, p=0.1, PROC NLP
993 out of 1000 case converged

Variable	Mean	Standard Deviation
Z_1	-0.0757	2.7368
Z_2	-0.0306	2.9492
Z_3	-0.1229	2.8704
Z_4	-0.0807	2.7129
Z_5	-0.1145	2.7608
Z_6	0.1095	3.5572
δ_1	-0.0346	6.4159

Table 10: Quantiles for δ_1 for three cases

Quantile	c=2, Case E,
	p=0.1
		C=2, Case E,
		p=0.5
			C=5, Case E,
			p=0.1

0.01	13.0870	2.5691	9.8879
0.10	1.1449	0.5079	1.0707
0.25	0.3271	0.2170	0.3152
0.50	0.0111	-0.0152	0.0002
0.75	.0.2978	-0.2339	-0.3375
0.90	-1.0724	-0.5420	-1.1111
0.99	-21.1403	-12.6719	-7.0577