INSTITUTE FOR SOCIAL RESEARCH
SURVEY RESEARCH CENTER
UNIVERSITY OF MICHIGAN

A Multivariate Stopping Rule for Survey Data Collection

Xinyu Zhang[1], James Wagner[1,2], Michael R. Elliott[1,2,3], Brady T. West[1,2], and Stephanie Coffey[4]

[1] Survey Research Center, Institute for Social Research, Univ. of MI-Ann Arbor

[2]Joint Program in Survey Methodology, Univ. of MD-College Park

[3]Dept. of Biostatistics, School of Public Health, Univ. of MI-Ann Arbor

[4]U.S. Census Bureau

November 2, 2022

Acknowledgements

This work was supported by a grant from the National Institutes for Health (#1R01AG058599-01; PI: James Wagner)

Thank you to Katharine Abraham, Brady T. West, Sipeng Wang, and Yuting Chen for their valuable comments on the early version of the paper in the PhD seminar

Responsive survey design

Declining response rates and rising costs of data collection are two concerning trends in surveys

Use incoming data to make near real-time design decisions during data collection to reduce survey costs or to increase data quality

Collect and analyze paradata to guide design decisions and direct field resources

Stopping rule

A set of criteria that specify when to stop data collection, e.g.,

- Target response rate
- Target number of interviews
- Budget for data collection

In responsive survey designs, a stopping rule is specified by a function of incoming data to improve data quality or to reduce survey costs

Existing stopping rules

Discontinue nonresponse follow-up (project level)

- Stop collecting data when a stable condition (phase capacity) is detected in the last design phase
- Require sufficient funds for follow-up to detect the condition
- Univariate tests (e.g., Rao, Glickman, and Glynn, 2008; Wagner and Raghunathan, 2010; Lewis, 2017) and multivariate tests (e.g., Lewis, 2019)

Stop effort on a subset of nonrespondents (case level)

- Follow up on unresolved cases that are not stopped
- Two-phase sampling for nonresponse (e.g., Hansen and Hurwitz, 1946; Elliott, Little, and Lewitzky, 2000) and a univariate stopping rule aimed at optimizing the cost-error tradeoff (Wagner et al., 2021)

Motivation

Most of the existing stopping rules are univariate

- Lewis (2019) is an exception but does not consider costs

In multipurpose surveys, there may be data quality objectives that must be met for certain estimates with constraints on costs

We propose a multivariate stopping rule that accounts for survey costs and the data quality of multiple estimates

Setup of the proposed stopping rule

Nonresponse follow-up

- At a given point in time d, n_0 cases are interviewed and $n - n_0$ cases are unresolved
- Stop effort on a set of cases S from $n - n_0$ unresolved cases
- Other unresolved cases that are not stopped will be followed up after time d

Consider the data quality for survey variables $Y_1, Y_2, \ldots,$ and Y_P

Objective: optimize the tradeoff between costs and the mean squared errors of these P estimates of sample means

A univariate stopping rule

Consider the data quality for one survey variable Y_p

No cases are stopped

$\hat{C}_o$: estimated total costs (cost component)

$\widehat{MSE}_{p,o}$: estimated mean squared error of $\hat{\bar{Y}}_p$(data quality component)

After stopping effort on a set of cases S

$\hat{C}_{-S}$: estimated costs (cost component)

$\widehat{MSE}_{p,-S}$: estimated mean squared error of $\hat{\bar{Y}}_{p,-S}$ after stopping a set of cases S (data quality component)

A univariate stopping rule (cont'd)

Alternative metrics (A general goal: more gains in cost savings and less losses in the mean squared error)

1. Sum: $\left(1 - \frac{\widehat{MSE}_{p,-S}}{\widehat{MSE}_{p,\mathrm{o}}}\right) + \left(1 - \frac{\hat{C}_{-S}}{\hat{C}_o}\right)$
   - A higher positive value is preferred
   - Evaluate the tradeoff on the additive scale

2. Ratio: $\left(\frac{\widehat{MSE}_{p,-S}}{\widehat{MSE}_{p,\mathrm{o}}}\right) / \left(\frac{\hat{C}_{-S}}{\hat{C}_o}\right) \propto \frac{\widehat{MSE}_{p,-S}}{\hat{C}_{-S}}$
   - Inconsistent judgements (a smaller ratio is preferred if $\hat{C}_{-S}$ is fixed; however, a larger ratio is preferred if $\widehat{MSE}_{p,-S}$ is fixed)

3. Product: $\left(\frac{\widehat{MSE}_{p,-S}}{\widehat{MSE}_{p,\mathrm{o}}}\right)\left(\frac{\hat{C}_{-S}}{\hat{C}_o}\right) \propto \hat{C}_{-S}\widehat{MSE}_{p,-S}$
   - A lower value that is also less than $\hat{C}_o\widehat{MSE}_{p,o}$ is preferred
   - Evaluate the tradeoff on the multiplicative scale
   - Hard to interpret

A toy example

Let $\widehat{MSE}_{p,o} = 0.1$ and $\hat{C}_o = 100$

Inputs		Metric		
$\widehat{MSE}_{p,-S}$	$\hat{C}_{-S}$	Sum (rank)	Ratio[1]	Product (rank)
0.11	90	0 (3)	0.00122	9.90 (3)
0.11	88	0.020 (2)	0.00125	9.68 (2)
0.12	90	-0.100 (4)	0.00133	10.80 (5)
0.12	60	0.200 (1)	0.00200	7.20 (1)
0.121	89.2	-0.102 (5)	0.00136	10.79 (4)
0.18	80	-0.600 (6)	0.00225	14.40 (6)

Note. [1]Values are not ranked for ratio due to its inconsistent judgements

A multivariate stopping rule

Consider the data quality for multiple survey variables $Y_1, Y_2, \ldots, Y_P$

Data quality component:

$$\sum_{p=1}^{P} w_p \widehat{MSE}_{p,-S}$$

where $w_1, \ldots, w_P$ are prespecified estimate-level weights subject to constraints $\sum_{p=1}^{P} w_p = 1$ and $w_p \geq 0, p = 1, \ldots, P$

All variables are standardized by z-score scaling to ensure that they are on the same scale

A multivariate stopping rule (cont'd)

Consider an objective function ψ_{-S} after stopping set S:

$$\psi_{-S} = \hat{C}_{-S} \sum_{p=1}^{P} w_p \widehat{MSE}_{p,-S}$$

Computationally expensive (or prohibitive) to find the exact minimal solution when the number of unresolved cases is over 50

- There are $2^{n-n_0} - 1$ possible sets for stopping

Approximation approach

- Stop cases step by step (it is feasible to stop one case with the best cost-error tradeoff at each step)

A multivariate stopping rule (cont'd)

Assume simple random sampling, the objective function for stopping a case, j, ψ_{-j}:

$$\psi_{-j} = \hat{C}_{-j} \sum_{p=1}^{P} w_p \widehat{MSE}_{p,-j}$$

where $\hat{C}_{-j} = \hat{C} - \hat{C}_j$ is the estimated remaining costs after stopping effort on case j, $\widehat{MSE}_{p,-j} = \hat{B}_{p,-j}^2 + \left(\frac{n}{n-1}\right)\hat{V}_p$ is the estimated mean squared error of $\hat{\bar{Y}}_{p,-j}$, $p = 1, \ldots, P$

Select a case that minimizes ψ_{-j} (denoted as ψ_{-S_1}), where S_1 is an initial set of one case to stop

A multivariate stopping rule (cont'd)

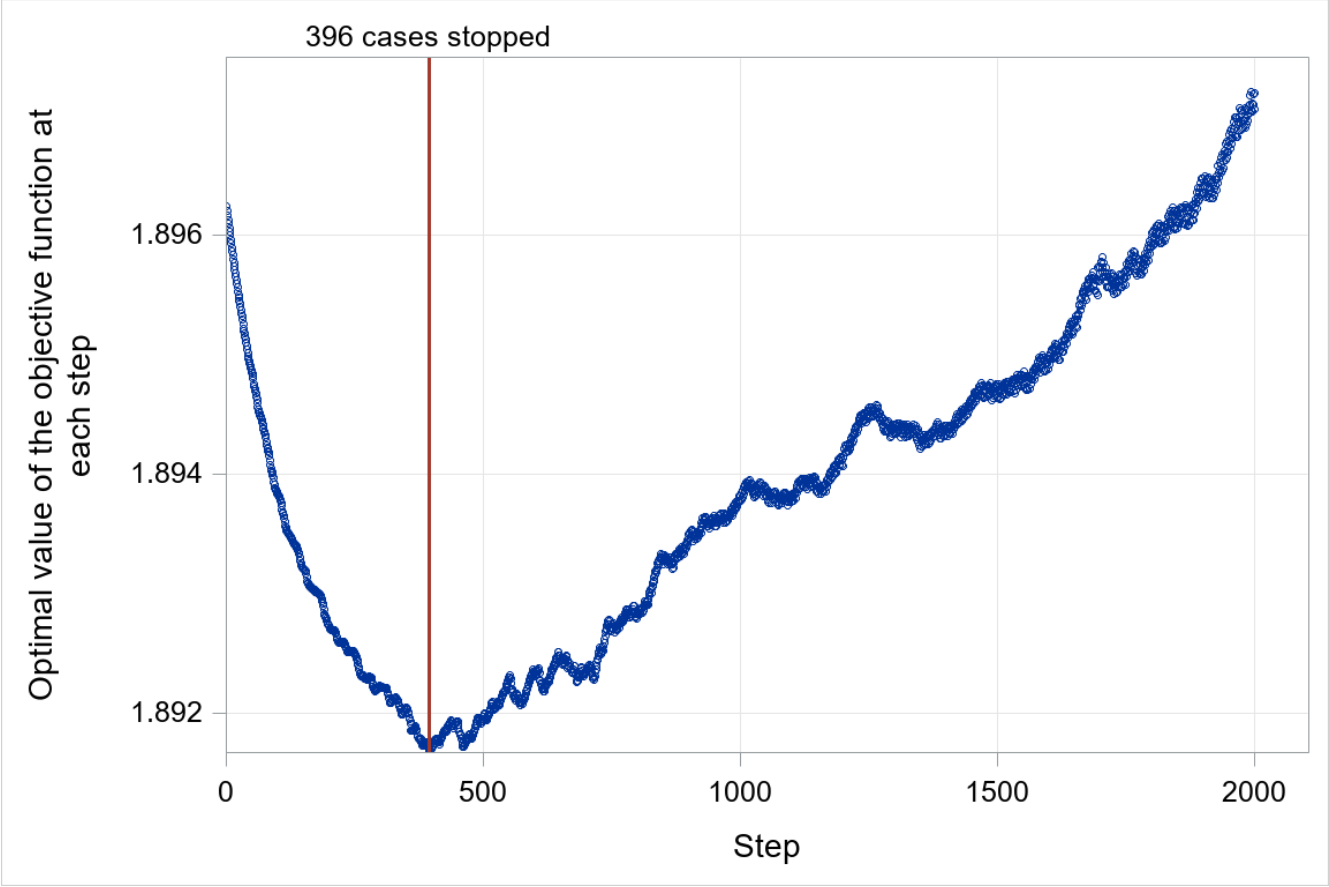
The objective function for stopping another case, j, $\psi_{-S_1,-j}$:

$$\psi_{-S_1,-j} = \hat{C}_{-S_1,-j} \sum_{p=1}^{P} w_p \widehat{MSE}_{p,-S_1,-j}$$

where $\hat{C}_{-S_1,-j} = \hat{C} - \sum_{i \in S_1} \hat{C}_i - \hat{C}_j$ is the estimated remaining costs after stopping effort on set S_1 and case j, $\widehat{MSE}_{p,-S_1,-j} = \hat{B}^2_{p,-S_1,-j} + \left(\frac{n}{n-2}\right)\hat{V}_p$ is the estimated mean squared error of $\hat{\bar{Y}}_{p,-S_1,-j}$, $p = 1, \ldots, P$

Select another case that minimizes $\psi_{-S_1,-j}$ (denoted as ψ_{-S_2})
Repeat this process until a set of cases to stop that approximately minimizes ψ_{-S}

Optimal value of the objective function at each step (first 2000 steps)



Simulation study

2018 Health and Retirement Study (HRS) - Telephone Component

- Longitudinal study of the US population over age 50
- Among 7,415 sampled cases, 5,462 responded
  1,953 nonrespondents and some item missing data are multiply imputed for creating the benchmark estimates
- The field work took 416 days

The stopping rule is implemented once at the end of data collection day 28

2016 and 2018 HRS data* for modeling survey design parameters

- Timesheet data (for interviewer hours for each call outcome)
- Call record data and survey data (for propensity scores at the call attempt level)
- Survey data (for values of survey variables)

*Data were observed by data collection day 28 in the 2018 wave of the HRS

Cumulative effort

Extend the horizon for predicting interviewer hours out to 21 call attempts

We built a multinomial logistic regression model to predict propensity scores of three call attempt outcomes for call attempts from $t_{0,i} + 1$ to 21, for case i, $i = n_0 + 1, \ldots, n$, at call attempt t, $t = t_{0,i} + 1, \ldots 21$, where $t_{0,i}$ is the number of call attempts made by day 28 for case i

Estimated propensity scores for case i at call attempt t are $\hat{p}_{iw,t,i}$, $\hat{p}_{cont,t,i}$, and $\hat{p}_{nocont,t,i}$ for interview, contact but no interview, and no contact, respectively

Cumulative effort (cont'd)

Interviewer hours are not directly measured for each call attempt

Strategy to estimate interviewer hours for each call attempt

- Timesheet data at the interviewer-day level include
  a) interviewer hours
  b) call attempts by mode and outcome

- Fit a multilevel regression model with a random intercept for each interviewer and a random slope for the indicator of any face-to-face attempts for each interviewer on each day

- Coefficients are estimated time per attempt (e.g., estimated average interviewer hours of an interview, contact without an interview, and no contact are $\hat{c}_{iw} = 1.6$, $\hat{c}_{cont} = 0.2$, and $\hat{c}_{nocont} = 0.07$, respectively)

Cumulative effort (cont'd)

For case i, $i = n_0 + 1, \ldots, n$, the estimated cumulative effort is

$$\hat{C}_i = \sum_{t=t_{0,i}+1}^{21} \hat{s}_{t-1,i}(1.6 * \hat{p}_{iw,t,i} + 0.2 * \hat{p}_{cont,t,i} + 0.07 * \hat{p}_{nocon,t,i})$$

where $\hat{s}_{t_{0,i},i} = 1$ is the estimated probability of not being interviewed at call attempt $t_{0,i}$, and
$\hat{s}_{t-1,i}$ is the estimated probability of not being interviewed at call attempt $t - 1$

Selected survey variables of interest

Three survey variables

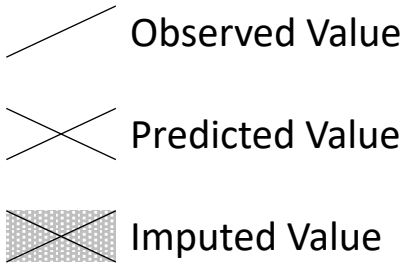
- Self-rated health (SRH; binary – fair or poor/excellent or very good or good)
- Impairment that limits work (ILW; binary – yes/no or too old)
- Functional limitations (FLs; continuous – 0-23)

Correlation matrix

	SRH	ILW	FLs
SRH	1.00		
ILW	-.40	1.00	
FLs	-.48	.59	1.00

Data structure of the simulation study

R	D	S	$Y^{(R)}$	$Y^{(D)}$	$Y^{(P)}$	$Y^{(\bar{S})}$	$Y^{(\bar{S}I)}$	$Y^{(RI)}$
1	1	NA						
.	.							
.	.							
1	1							
1	0	1		?		?		
.	.	.						
.	.	.						
1	0	1						
1	0	0		?				
.	.	.						
.	.	.						
1	0	0						
0	0	1	?	?		?		
.	.	.						
.	.	.						
0	0	1						
0	0	0	?	?		?		
.	.	.						
.	.	.						
0	0	0						



R = Final response status

D = Response status by a selected date

S = Stopping effort recommendation

$Y^{(R)}/Y^{(D)}/Y^{(\bar{S})}$ = Observed values

$Y^{(P)}$ = Predicted values based on $Y^{(D)}$

$Y^{(\bar{S}I)}$ = Observed and imputed values based on $Y^{(\bar{S})}$

$Y^{(RI)}$ = Observed and imputed values based on $Y^{(R)}$

Configuration of estimate-level weights

Scenario #	Weight for SRH	Weight for FLs	Weight for ILW
1	1	0	0
2	0	1	0
3	0	0	1
4	1/2	1/2	0
5	1/2	0	1/2
6	0	1/2	1/2
7	1/4	1/4	1/2
8	1/4	1/2	1/4
9	1/2	1/4	1/4
10	1/3	1/3	1/3

Evaluation criteria

Data quality

- Average absolute percent relative bias (avg absolute %relbias) of the three multiply imputed estimates

  For each $\hat{\bar{y}}^{(SI)}$, $absolute\ \%relbias\left(\hat{\bar{y}}^{(SI)}\right) = \left|\frac{\hat{\bar{y}}^{(SI)}}{\hat{\bar{y}}^{(RI)}} - 1\right| \times 100$

- Average percent relative root mean squared error (avg %relrmse) of the three multiply imputed estimates

  For each $\hat{\bar{y}}^{(SI)}$, $\%relrmse\left(\hat{\bar{y}}^{(SI)}\right) = \left(\frac{rmse\left(\hat{\bar{y}}^{(SI)}\right)}{rmse\left(\hat{\bar{y}}^{(RI)}\right)} - 1\right) \times 100$

Costs

- Estimated percent relative estimated total saved hours (%saved interviewer hours) by using the stopping rule

Simulation results (nonresponse adjusted)

No.	Scenario (Configuration of estimate-level weights in the proposed stopping rule)	%Saved interviewer hours	Avg absolute %relbias	Avg %relrmse
1	SRH:1 FLs:0 ILW:0	5.7	0.4	7.3
2	SRH:0 FLs:1 ILW:0	14.0	0.6	13.4
3	SRH:0 FLs:0 ILW:1	6.4	0.2	2.2
4	SRH:1/2 FLs:1/2 ILW:0	12.4	0.2	1.3
5	SRH:1/2 FLs:0 ILW:1/2	12.0	0.5	7.5
6	SRH:0 FLs:1/2 ILW:1/2	12.3	0.7	17.7
7	SRH:1/4 FLs:1/4 ILW:1/2	11.1	0.9	22.5
8	SRH:1/4 FLs:1/2 ILW:1/4	11.3	0.9	29.4
9	SRH:1/2 FLs:1/4 ILW:1/4	8.7	0.4	7.4
10	SRH:1/3 FLs:1/3 ILW:1/3	12.1	0.8	16.0

Note. Estimated total interviewer hours for the actual data collection (no stopping rule) are18,267.

Future directions

Identify approaches to the configuration of estimate-level weights

Adapt the proposed stopping rule to the optimization problem of maximizing data quality for a given budget

Improve cost predictions at the case level

Identify optimal timing of implementation of the stopping rule

Account for complex sample design

Test the stopping rule experimentally

Thank you!

Email: zhxinyu@umich.edu