

Optimal Sample Size Allocation to Mixed Modes: A Case Study Using the Residential Energy Consumption Survey

Peter Frechtel and Phillip S. Kott

RTI International (a registered trademark and a trade name of Research Triangle Institute),
Research Triangle Park, NC, USA

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference

Introduction

In the current era of declining response rates, reduced federal budgets, and societal changes occurring in short time frames, many US government agencies are seeking new ways to collect and analyze data, without a severe impact on data quality. Many of these new methods are cheaper, faster, or both. For example, some of the newer methods involve the merging of traditionally-collected survey data with administrative data, data from internet panels, or social media posts.

This paper discusses a method that reduces the costs of administering a survey without a severe impact on data quality. The method uses double sampling and regression estimation. It is designed for surveys where some variables are expensive to collect while other variables that are closely related to the expensive variables are less expensive to collect.

The method consists of the following steps:

- 1) Allocate some of the sample to the expensive mode (e.g., face-to-face interview) and some of the sample to the inexpensive mode (e.g., mail, phone, or web).
- 2) Collect the data (all variables with the expensive mode; only the inexpensive variables with the inexpensive mode).
- 3) Fit a model with the expensive variable(s) as the dependent variable and the inexpensive variable(s) as the independent variables.
- 4) Construct estimators for the expensive variables using the actual responses for the expensive mode and the model-predicted responses for the inexpensive mode.

If the expensive variables can be estimated with precision using the inexpensive variables as covariates, then the estimates related to the expensive variables will be made more precise by using the inexpensive variables, when compared to the precision obtained by direct estimates that use only the expensive sample.

This scenario, where some variables are expensive to collect and some are not, occurs frequently. Examples include:

- In the National Health and Nutrition Examination Survey (NHANES), respondents are asked demographic, socioeconomic, dietary, and health-related questions in a face-to-face interview. Some respondents also undergo physical examinations where medical, dental, and physiological measurements are taken, and laboratory tests are administered. The variables collected in the interview can be viewed as the inexpensive variables, and the variables collected in the physical examinations can be viewed as the expensive variables.
- In the National Survey on Drug Use and Health (NSDUH), as in NHANES, respondents are asked numerous questions, including many on drug use and health, in a face-to-face interview. From 2008-2012, a subsample of adult respondents were asked more detailed questions related to mental health disorders, where the interviewers were trained clinicians. The study based on the subsample was called the Mental Health Surveillance Study (MHSS). The MHSS variables can be treated as the expensive ones and the variables from the main interview can be treated as the inexpensive ones.

- In the Residential Energy Consumption Survey (RECS), respondents are asked about the energy characteristics of their place of residence. Before 2015, all interviews were conducted face-to-face, and trained interviewers measured the square footage of the residence by following a set of strict and detailed guidelines. However, in 2015, 57% of the unit respondents answered a questionnaire by web or mail and were not asked to measure the square footage of their residence. The variables other than measured square footage are the inexpensive variables, and the measured square footage is the expensive variable.

The primary question addressed in this paper is this: given a fixed budget, what is the optimal sample allocation between the inexpensive mode and the expensive mode? It stands to reason that if the expensive variables can be modeled well using the inexpensive variables, and the expensive mode is much more expensive than the inexpensive mode, then the optimal allocation would involve a few expensive-mode surveys and many inexpensive-mode surveys. Conversely, a poor model and modes that have similar costs would call for most, if not all, of the surveys to be administered using the expensive mode.

We will use data from the 2005 RECS to investigate this topic. Part of the motivation for using RECS data is that the scenario mentioned above was encountered directly during the 2015 RECS. The original plan was to collect all the data face-to-face via a computer-assisted personal interview (CAPI), but when that turned out to be too difficult and too expensive, the face-to-face mode was abandoned, and data from respondents in the remaining sample were collected using web and mail (but without the questions about measured square footage). The issue of optimal allocation arose when we wondered what would have happened if this were planned. Would we have collected more data using web and mail, or less?

Methods

The RECS is administered by the Energy Information Administration (EIA). The estimation of mean measured square footage is not of primary interest to EIA; it is more interested in using the measured square footage values in models that estimate the amount of energy consumed by heating, cooling, various appliances, etc. Nevertheless, mean measured square footage is the focus of this study. The rationale is that the bias and variance of the mean measured square footage are associated with the bias and variance of the predicted values derived using that statistic.

The method we used to find the optimal allocation employed four major steps, mentioned briefly here and covered in more detail immediately below:

- 1) Assume a fixed budget for data collection. We set the budget at \$200,000.
- 2) Assume per-complete-case costs for both the inexpensive and expensive modes. We assumed that the inexpensive mode costs \$25/case and the expensive mode costs \$500/case.
- 3) Use theory to plot the variance of the mean as a function of various values for the expensive sample size and the inexpensive sample size.
- 4) Run a simulation study to confirm that empirical results match the theory.

The first two steps were straightforward, although an interesting next step would be to run a sensitivity analysis: how much does the optimal allocation change if the relative costs change?

The third step involves some derivations described in more detail in Appendix A. These derivations are based in part on an extension of Equation (20) in Legg & Fuller (2009). The basic steps are fairly straightforward, though:

- a. Using data from the expensive model, fit a regression model with the expensive variable as the dependent variable and a subset of the inexpensive variables as independent variables. Save the predicted value and the residual for each observation.
- b. Calculate the variances of both the predicted values and the residuals. The variance of the predicted values is $\sigma_{\hat{h}at}^2$ and the variance of the residuals is σ_{ϵ}^2 .
- c. Given the fixed budget and the complete-case costs from the first and second major steps, compile a list of possible values for the full (i.e., combined) sample and the subsample (collected using the expensive mode). Label the former n and the latter m . Given our fixed budget and cost assumptions, we have the equation $500m + 25(n - m) = 200,000$. So we let m vary from 100 to 400, and calculated the value for n as $8,000 - 19m$.

- d. Calculate the part of the variance that varies with m and n , $\frac{\sigma_{hat}^2}{n} + \frac{\sigma_{\epsilon}^2}{m}$, for each value of m and n from Step (c).
- e. Graph the variance as a function of the subsample size m or the sample size n (given the fixed cost, one can be derived from the other as shown above). According to Appendix A, the optimal sample size ratio should be $\frac{m}{n} = \frac{\sigma_{\epsilon}^2/\sqrt{500}}{\sigma_{hat}^2/\sqrt{25}}$.

The fourth major step (i.e., the simulation study) was completed because the derivations in Appendix A assume that the model used in Step (a) is correct, and this may not be the case. So the goal was to ground-truth the results derived from theory in the third major step. To implement the simulation, for each combination of m and n , we drew a full sample of size n from the complete 2005 RECS data, and a subsample of size m , and repeated this 10,000 times. For each iteration of the simulation, the point estimate of the mean was

$$\bar{y} = \frac{1}{n} (\sum_{i \in M} y_i + \sum_{j \in (N-M)} \hat{y}_j),$$

where n is the full sample size, M is the set of observations in the subsample, N is the set of observations in the full sample, and \hat{y}_j is the model-predicted measured total square footage for sample unit j .

We then calculated the mean squared error of the 10,000 point estimates and the bias of those point estimates. We expect the mean squared error to be lowest around the optimal sample size ratio based on theory, $\frac{m}{n} = \frac{\sigma_{\epsilon}^2/\sqrt{500}}{\sigma_{hat}^2/\sqrt{25}}$.

For both the third and fourth major steps, a model had to be fit to estimate measured square footage as a function of the inexpensive variables. Before that, the 2005 RECS public-use dataset was downloaded and pre-processed using the following steps:

- *Download the data.* The data are available, as of 1/16/2018, from the website <https://www.eia.gov/consumption/residential/data/2005/index.php?view=microdata>. The dataset contains 4,382 observations.
- *Subset the data.* Only the 2,044 cases meeting the following criteria were kept:
 - TYPEHUQ = 2 (single-family, detached housing units)
 - ZTOTSQFT = 0 and ZSQFTEST = 0 (neither measured square footage nor respondent-estimated square footage were imputed)
 - STORIES \neq 50 (the STORIES variable was something other than "some other type")
- *Recode the dependent and independent variables for the model.* The model used to predict measured square footage was very similar to the model used to impute square footage for the 2015 RECS (U.S. Energy Information Administration, 2017). The dependent variable was the square root of measured square footage: $z = \sqrt{TOTSQFT}$. The independent variables were:
 - The square root of respondent-estimated square footage, SQRT_SQFTEST.
 - A recoded version of STORIES, called STORIES_R:
 - 1 = one story
 - 2 = two stories
 - 2.5 = split level
 - 3 = three or four stories
 - TOTROOMS, the total number of rooms in the housing unit.
 - A recoded version of the attic variables (ATTIC, ATTICHEAT, ATTCCOOL, and ATTICFIN), called ATTICCOM:
 - 0 = no attic
 - 1 = attic, completely heated and/or cooled, and/or finished
 - 2 = attic that is not finished or completely heated/cooled
 - CELLAR, whether or not the housing unit contained a basement.
 - A recoded version of the garage variables (GARGLOC, GARAGE1C, GARAGE2C, and GARAGE3C), called GARGCOM:
 - 0 = no garage

- 1 = one-car garage
- 2 = two-car garage
- 3 = three-car garage
- URBRUR, a self-reported urban/rural variable (1 = City, 2= Town, 3 = Suburbs, 4 = Rural).

After the model was fit, we had to undo the square-root transformation to create predicted values and residuals. This was done as follows:

- 1) Calculate the overall adjustment factor $a = \frac{\sum_{i \in M} y_i}{\sum_{i \in M} \hat{z}_i^2}$. This is used to account for the fact that the expected value of the square of the model-predicted value is not equal to the square of the predicted value from the model. That is, $E(z^2) \neq (E(z))^2$.
- 2) The predicted measured square footage for observation i , $\hat{y}_i = a\hat{z}_i^2$.
- 3) The residual for observation i is $e_i = y_i - \hat{y}_i$.

The variances of the predicted values (the \hat{y}_i) and residuals (the e_i) were used in the third major step to calculate $\sigma_{\hat{y}}^2$ and σ_e^2 respectively, and the predicted values were used in the fourth major step (i.e., the simulation) to calculate \bar{y} .

Results

Table 1 provides statistics associated with the model used to predict the "expensive" variable TOTSQFT as a function of the "inexpensive" variables. This model was fit using 2,044 cases as described in the previous section. The model produced values of $\sigma_{\hat{y}}^2 = 1.7$ million, $\sigma_e^2 = 1.2$ million, and an optimal sample size ratio of 0.16. The optimal sample size ratio suggests that the ideal allocation would be approximately $m = 316$ and $n = 2000$. (In practice, the analyst will typically not have access to the full dataset in order to estimate $\sigma_{\hat{y}}^2$ and σ_e^2 .)

Table 1. Covariate-level statistics for model where $\sqrt{TOTSQFT}$ is regressed on "inexpensive" variables.

Neither the weights nor the sample design variables were used. $N = 2,044$, $R^2 = 0.62$. The adjustment factor a used to correct the squared predicted values was 1.03.

Independent Variable	Coefficient Estimate	Standard Error	T-Test: $\beta = 0$	P-Value
$\sqrt{SQFTEST}$	0.32	0.03	9.51	< 0.001
# Stories (1, 2, 2.5 (Split-Level), 3 or more)	3.76	0.48	7.83	< 0.001
# Rooms	1.04	0.17	6.19	< 0.001
No Attic	-12.23	0.49	-24.82	< 0.001
No basement	-8.00	0.43	-18.44	< 0.001
Size of garage (0 = no garage, 1 = 1-car, 2 = 2-car, 3 = 3-car)	3.17	0.24	13.40	< 0.001
Urban/rural				
1 = city	-3.70	0.52	-7.13	< 0.001
2 = town	-2.86	0.61	-4.71	< 0.001
3 = suburbs	-1.51	0.61	-2.48	0.013
4 = rural	reference cell			

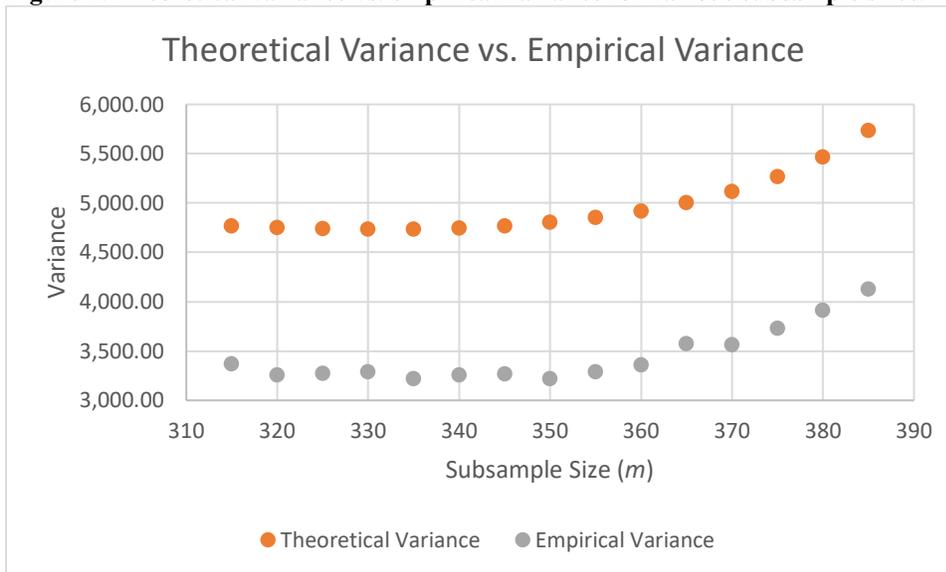
Table 2 compares the theoretical variance of the estimated mean with the empirical variances from the simulation. The two track well, as shown in Figure 1. The two parallel each other, but do not overlap because the theoretical variance has no finite population correction adjustment, but the simulation had a very large sampling fraction which is reflected in the empirical variances. The biases resulting from the method are very small.

Table 2. Theoretical variance vs. empirical variance for various subsample sizes.*

<i>m</i>	<i>n</i>	Theoretical Variance ()	Simulation Results		
			Empirical MSE	Empirical Bias	Empirical Variance
315	2,015	4,767.68	3,366.69	-0.89	3,365.90
320	1,920	4,749.37	3,254.47	0.55	3,254.17
325	1,825	4,737.39	3,274.20	0.32	3,274.10
330	1,730	4,732.38	3,288.26	0.07	3,288.26
335	1,635	4,735.16	3,220.07	-0.15	3,220.05
340	1,540	4,746.78	3,259.18	-1.03	3,258.13
345	1,445	4,768.60	3,267.33	-0.24	3,267.27
350	1,350	4,802.36	3,219.34	0.05	3,219.34
355	1,255	4,850.38	3,289.80	-0.25	3,289.74
360	1,160	4,915.74	3,358.12	-0.31	3,358.03
365	1,065	5,002.67	3,571.72	1.13	3,570.45
370	970	5,117.06	3,560.20	0.28	3,560.12
375	875	5,267.40	3,731.74	-0.68	3,731.28
380	780	5,466.31	3,911.92	0.29	3,911.84
385	685	5,733.49	4,128.89	-0.53	4,128.61

*The mean measured square footage across all 2,044 observations, which was used to estimate the MSE and bias, was 3,051.88.

Figure 1. Theoretical variance vs. empirical variance for various subsample sizes.



Conclusions and Next Steps

The method appears to have worked well. Both the theoretical variance and the empirical variance are minimized at around $m = 335$, which suggests that if precise estimation of the mean total square footage is the primary goal of the survey, and if a CAPI case really is about 20 times as expensive as a web/mail case, then about 20% of the cases should be CAPI and the other 80% should be web/mail. In general, the sample allocation is dependent on both the ability of the model to precisely estimate total measured square footage and the cost of the expensive mode relative to the inexpensive mode.

Although the results in this paper apply to the Residential Energy Consumption Survey, the method used to find the optimal sample allocation is not too complex and can be done for any survey with some expensive variables and some inexpensive variables, as long as earlier data and per-case cost estimates are available. The hope is that the method allows survey practitioners to precisely estimate the means of outcome variables at a fraction of the cost of collecting all data using the "expensive" mode.

The most obvious topic for further research is to incorporate the weights and sample design (weights are in the appendix). Not only were the weights and sample design not used for any of the calculations in this paper, but the sample design will often be different for the expensive and inexpensive modes. If the expensive mode is a face-to-face interview, for example, then geographic clustering will likely be necessary to control data collection costs; and the same may not be true for the inexpensive mode. This suggests that the true optimal allocation for RECS might be even fewer than 20% CAPI cases: a single CAPI case subject to clustering would be worth even less relative to a single web/mail case that is not subject to clustering.

References

Legg, J.C. and Fuller, W.A. (2009). "Two-Phase Sampling," In *Handbook of Statistics 29: Vol 29A, Sample Surveys: Design, Methods, and Applications*, Pfeffermann, D. and Rao, C.R. eds., pp. 55-70. Elsevier: Amsterdam.

U.S. Energy Information Administration (2017). "2015 RECS Square Footage Methodology." Available as of 4/24/2018 at

https://www.eia.gov/consumption/residential/reports/2015/squarefootage/pdf/2015_recs_squarefootage.pdf.

Appendix A
Derivation of Variance Estimator and Optimal Allocation Formula

Let $S = S_1$ be the first-phase sample

S_2 be the second-phase sample

d_k be the first-phase weight of an element k selected for the first-phase sample

w_k be the (combined) weight of an element k selected for the second-phase sample, 0 otherwise

\mathbf{x}_k be a vector of variables collected for elements in the first-phase sample (and unity; mathematically, there is a vector \mathbf{g} such that $\mathbf{g}^T \mathbf{x}_k = 1$ for all k),

y_k be the survey variables collected only for elements in the second-phase sample, and

U be the population of size N .

The traditional regression estimator (for a population y -total: $T_y = \sum_U y_k$) is

$$\begin{aligned} t_y &= \sum_S d_k \mathbf{x}_k^T \mathbf{b}, \text{ where } \mathbf{b} = \left(\sum_S w_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \sum_S w_j \mathbf{x}_j y_j \\ &= \sum_S w_k y_k + \left(\sum_S d_k \mathbf{x}_k^T - \sum_S w_k \mathbf{x}_k^T \right) \mathbf{b} \quad \left(\text{because } \sum_S w_k \mathbf{x}_k^T \mathbf{b} = \sum_S w_k \mathbf{g}^T \mathbf{x}_k \mathbf{x}_k^T \mathbf{b} = \sum_S w_k y_k \right) \end{aligned}$$

The prediction model is $y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k$, where $E(\varepsilon_k | \mathbf{x}_k) = 0$.

The probability-sampling analogue is $y_k = \mathbf{x}_k^T \mathbf{B} + e_k$, where \mathbf{B} is the probability limit of \mathbf{b} , and $e_k = y_k - \mathbf{x}_k^T \mathbf{B}$.

The error in the estimator under the model is

$$\begin{aligned} t_y - T_y &\approx \sum_S w_k (\mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k) + \left(\sum_S d_k \mathbf{x}_k^T - \sum_S w_k \mathbf{x}_k^T \right) \boldsymbol{\beta} - \sum_U (\mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k) \\ &\quad \left(\text{because } \left(\sum_S d_k \mathbf{x}_k^T - \sum_S w_k \mathbf{x}_k^T \right) (\mathbf{b} - \boldsymbol{\beta}) / T_y \approx 0 \right) \\ &= \left(\sum_S d_k \mathbf{x}_k^T \boldsymbol{\beta} - \sum_U \mathbf{x}_k^T \boldsymbol{\beta} \right) + \left(\sum_S w_k \varepsilon_k - \sum_U \varepsilon_k \right) \end{aligned}$$

If the first-phase sample in the text is a simple random sample (srs) of size n and the second phase an srs of size m , then the model variance of the estimator ignoring finite population correction is N^2 times the sum of the variance of $\mathbf{x}_k^T \boldsymbol{\beta}$, call it σ_{hat}^2 , divided by n and the variance of ε_k , call it σ_ε^2 , divided by m .

If c_1 is the unit cost for the first-phase sample and c_2 the unit cost for the second-phase sample, then minimizing the variance for a constant cost C is equivalent to minimizing the Lagrangian:

$$L = N^2 \left[\sigma_{hat}^2 / n + \sigma_\varepsilon^2 / m \right] + \lambda (c_1 n + c_2 m - C).$$

Consequently, the optimal sample-size ratio would be $\frac{m}{n} = \frac{\frac{\sigma_\varepsilon}{\sqrt{c_2}}}{\frac{\sigma_{hat}}{\sqrt{c_1}}}$.

A Prediction Estimator (like the one used in the text):

$$\begin{aligned}
 t_y^* &= \sum_{S_2} d_k y_k + \left(\sum_S d_k \mathbf{x}_k^T - \sum_{S_2} d_k \mathbf{x}_k^T \right) \mathbf{b}^*, \\
 &\quad \text{where } \mathbf{b}^* = \left(\sum_{S_2} [w_j - d_j] \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \sum_{S_2} [w_j - d_j] \mathbf{x}_j y_j \\
 &= \sum_{S_2} w_k y_k + \left(\sum_S d_k \mathbf{x}_k^T - \sum_{S_2} w_k \mathbf{x}_k^T \right) \mathbf{b}^*, \\
 &\quad \text{because } \sum_{S_2} [w_k - d_k] \mathbf{x}_k^T \mathbf{b}^* = \sum_{S_2} [w_k - d_k] \mathbf{g}^T \mathbf{x}_k \mathbf{x}_k^T \mathbf{b}^* = \sum_{S_2} [w_k - d_k] y_k
 \end{aligned}$$

(Remember that $w_k = 0$ for k not in S_2 .)

Note that if both phases are srs (as in the text), then $\mathbf{b}^* = \mathbf{b}$.

When \mathbf{x}_k is a scalar x_k (such as when it is a predicted value) and there is no intercept

$$\begin{aligned}
 t_y &= \sum_S d_k x_k b, \quad \text{where } b = \frac{\sum_S w_j y_j}{\sum_S w_j x_j} \\
 &= \sum_S w_k y_k + \left(\sum_S d_k x_k - \sum_S w_k x_k \right) b;
 \end{aligned}$$

and

$$\begin{aligned}
 t_y^* &= \sum_{S_2} d_k y_k + \left(\sum_S d_k x_k - \sum_{S_2} d_k x_k \right) b^*, \quad \text{where } b^* = \frac{\sum_{S_2} [w_j - d_j] y_j}{\sum_{S_2} [w_j - d_j] x_j} \\
 &= \sum_{S_2} w_k y_k + \left(\sum_S d_k x_k - \sum_{S_2} w_k x_k \right) b^*, \quad \text{because } \sum_{S_2} d_k x_k = \sum_{S_2} w_k x_k - \sum_{S_2} [w_j - d_j] x_j
 \end{aligned}$$

Note that when the x_k are themselves determined through a model fitting (i.e., each is a predicted value like the \hat{z}_i^2 in the text), the relevant model for determining σ_{hat}^2 and σ_ε^2 is the simple through-the-origin linear model relating y_k to x_k , not the model used in creating the x_k .