Articles

Estimation of Survey Cost Parameters Using Paradata

James Wagner*

Keywords: survey costs

DOI: 10.29115/SP-2018-0036

Survey Practice

Vol. 12, Issue 1, 2019

In many survey situations, detailed cost parameters are difficult to estimate. This is especially true in surveys involving interviewers. Overall costs may be easily estimated since interviewer hours, materials, and incentives are relatively easy to track. But costs at a more granular level – for example, hours spent travelling, identifying non-sample units, or engaged in other activities – may be difficult to track. This occurs for a number of reasons. Often, cost information and paradata are collected in separate systems; or the cost information that is collected may not be at a sufficiently detailed level in order to evaluate the costs of particular subtasks. It might be possible to gather these cost data via a special study, but this is usually a very expensive approach. It may also be possible to ask for more detailed reporting from interviewers and other staff. However, this approach might lead to reduced efficiency. This paper proposes the use of regression models estimated from paradata as a method for estimating detailed cost parameters related to interviewer effort. An example of this method is shown from the National Survey of Family Growth 2011-2018. This method was used to evaluate the costs of two treatment arms in an experimental study. The method is also used to monitor interviewer effort over the course of the field period.

INTRODUCTION

Costs are a constraint on the quality of survey data. The quality of the data (usually measured by mean squared error) can only be maximized for a fixed budget when the costs of the various possible design options are known. For example, nonresponse error might be minimized when a large incentive is offered. However, a large incentive will limit the number of interviews that can be obtained, thereby increasing sampling error. Choosing appropriate modes, incentives, call limits, and other design features presupposes knowledge of the costs of these various choices (Groves 1989).

Estimates of costs are very important for the design of efficient surveys. However, these costs are not always directly measured at the required level. An important example is drawn from face-to-face surveys. In this setting, the cost of each call attempt is not directly measured. Instead, interviewers report the amount of time worked each day. This time might be reported in subcategories (e.g. administrative time, travel time, interviewing, sampling, etc.), but it is not

* **Institution:** University of Michigan (MI) **Department:** Survey Research Center **ORCID iD:** 0000-0002-8823-3110

recorded at the call attempt level.

There are other examples of costs that are not directly measured. The costs of training are not always directly measured. Training interviewers is a complex task that involves many team members who work on multiple activities. The time they spend training interviewers may not be easily separated from time spent on other activities. Another example is the sending of email messages. The cost of actually sending the message might be zero; however, there are also the costs of preparing the message, identifying to whom it should be sent, and loading data into systems to carry out these procedures. These costs may be difficult to measure. When costs are not directly measured, we sometimes use proxy measures of cost. For example, Gfroerer et al. (2002) use the number of attempts as a proxy indicator of cost, with more attempts signaling higher costs. This paper will present a method for estimating costs – that are not directly measured – using paradata.

BACKGROUND

Estimating detailed survey costs can be difficult. In face-to-face surveys, the cost of a call attempt is not directly observed. From call record data, we can know the number of attempts and how many attempts with various outcomes are made (noncontact, refusal, interview, etc.). We do not have direct information about how long each attempt takes. Instead, we might have the total hours worked on a day.

We would expect that attempts with different outcomes would take different amounts of time. A noncontact takes less time than a refusal, and refusals generally take less time than interviews. In this setting, a simple average of the amount of time each call attempt takes might not be very useful for comparing two designs. Some studies have taken this approach (Andresen et al. 2008; Pruchno and Hayden 2000). Unless the mixture of outcome types is expected to be the same across the two designs, the average time of all attempts might be biased in one direction or the other. For example, if we introduce a design change that increases contact rates (and thereby reduces the number of call attempts by removing many of the noncontact attempts), this would have less impact than a design change aimed at reducing refusal rates that removes a similar number of refusals since refusals generally take more time than noncontacts.

Instead of direct information about the length of each call attempt, from timesheet data we know how many hours an interviewer worked on a particular day. We may also know how the total hours worked in a day are divided into subcategories (such as administrative functions, travel, and interviewing), but we do not have how long each call attempt might have taken.

It is often the case that the timesheet and call record data are stored in two different systems. Call record data summarize call attempts and are stored in sample management systems. Each record represents a call attempt and

includes the date and time of the call attempt (sometimes entered by the system, i.e. "timestamps," and sometimes entered by the interviewer). An outcome code is also often stored. This outcome code can be used to categorize cases into labelled groups such as "interview," "noncontact," and "refusal." Of course, the number and type of categories can vary a lot depending upon the organization, the survey, and other factors. Timesheets are typically reported at the day level but also might be broken down into subcategories within a day.

The question is whether we can model the relationship between these two sources of data. That is, can we model the number of hours worked in a day, week, or some other time interval, based on the information in the call records—the number of call attempts or the number of call attempts in each of several different outcome types? This paper proposes using a statistical modeling strategy to estimate hours at the call attempt level. Specifically, we propose to regress the hours worked during some unit of time—day, week, or something else—on counts of calls of different types. The coefficients in such a model could be used to predict the total number of hours required under a different "mixture" of outcome types. Formula (1) gives a general expression for this approach:

$$Hours_{it} = \sum_{p=1}^{P} \beta_p x_{itp} + \varepsilon_{it} \quad (1)$$

The variable *Hours* represents the number of hours worked by an interviewer (indexed *i*) in a specified amount of time (e.g. one week, indexed by *t*). There are *P* predictors, which could include a vector of 1s if an intercept is desired.

There are some issues that need to be resolved in order to implement such an approach. The first issue is that all not all hours that interviewers work are spent making call attempts. A potential solution is to include some information (variables) about what else is being performed, which might be the number of days worked or it might be the number of shifts worked or it might even be as simple as including an intercept in the model, but there need to be some indicators of other nonattempt effort that can account for time that is not spent making call attempts. For example, interviewers may have a weekly meeting with their supervisors to discuss current issues they are facing and strategies for dealing with those issues.

Another problem that needs to be resolved is that the hours and the call records need to be summarized to the same level. The appropriate level of aggregation depends upon the particular situation. We found problems with trying to aggregate data from face-to-face surveys at the day level. For example, we found that interviewers sometimes report call records or hours on the wrong day (Wagner, Olson, and Edgar 2017). This leads to mismatches. Although these mismatches may represent a small proportion of days (less than 5%), they are an important category of matches—usually days with shorter hours worked or fewer call records made. Correcting those mismatches is very intensive. It is simpler, although not perfect, to aggregate the data from both the call records

and the hours up to the week level. Under this approach, we sum all the hours worked each week by each interviewer and then count all of the call records with similar outcomes made in that same week by that interviewer. We also count the number of days worked (using the timesheet) and the number of area segments visited. Both of these allow us to parameterize the model for noncall attempt effort as mentioned previously.

The final issue has to do with model specification. The resulting estimates may vary based upon which variables are included as predictors in the model. For example, defining different groupings of outcome types might lead to different results. Imagine a situation where in reality there are two types of calls. One type takes a long time, and the other takes a short amount of time. For the purposes of the example, let us say the first type takes an average of 5 minutes, and the second type takes an average of 20 minutes. If each type is half the total volume of call attempts, then the average length of the combined types is 12.5 minutes. Misspecifying the model by providing a single predictor that groups together these two types of calls and getting an estimate of 12.5 is not a problem, until we introduce a design change that alters the mix of calls. If the proposed design change would lead to a reduction in the *shorter* calls by one half, that is, to 25% of all call attempts, then the predicted average of 12.5 would be way off. The new average would be 16.25 minutes (.25 x 5 + 0.75 x 20). The estimate from before the design change would be too short in this example. On the other hand, if the model were correctly specified, that is, if we had estimates for each of the two types of calls (i.e. 5 and 20 minutes), then the model would provide accurate predictions of time savings due to reducing one of the types of calls. In summary, it is important to get the right categories of outcomes. It is desirable to define outcome categories that are homogenous with respect to the amount of time they take and that are most likely to directly reflect impact on effort resulting from design changes. Given that model specification may be an issue, it is recommended to try several different specifications to see how sensitive the results are to the different specifications.

DATA AND METHOD

The specific example that we will use is drawn from the National Survey of Family Growth (NSFG). The NSFG is a survey about fertility and family formation. It is a large face-to-face survey conducted nationally. The NSFG is a study of women and men ages 15 to 44, and since 2015, the eligible ages have been expanded to 15 to 49 (see https://www.cdc.gov/nchs/nsfg/index.htm for additional details on the NSFG). About 53% of households contain an eligible person. An important step in the interviewing process is therefore identifying eligible households. This step is known as "screening." Once an eligible person is identified, a "main" interview is attempted.

The data include call record data and timesheet data which are recorded in two separate systems. Call records are recorded at different frequencies, but usually immediately following the attempt (Wagner, Olson, and Edgar 2017).

Timesheets are recorded daily or sometimes less frequently with every two weeks being the minimum level of reporting.

In 2013–2014, the NSFG implemented an incentive experiment. One of the goals was to see which incentive was more cost effective. To meet this objective, cost estimates for interviews taken under each incentive amount were needed. More details on the experiment are available in Wagner et al. (2017). The problem was that the interviewers could work on both types of samples (i.e. cases that receive incentive amount A and also cases that received incentive amount B). As a result, in their timesheets, they reported hours for which they could have been working both types of sample. Using a hypothetical example, an interviewer might have reported working 6 hours on a specific day. On that day, an interviewer worked cases from each arm of the incentive experiment. However, we do not know how many of those hours were work expended on sample type A and how many were expended on sample type B.

In order to produce separate estimates for each of the incentive amounts, we used the regression approach described in the previous section. We counted calls of different types each interviewer made each week, and then we regressed the hours that interviewers worked each week on these counts of calls of different types. These gave us our initial estimates of how long each type of call takes. Then we counted how many of each type of call were made on each type of sample (i.e. incentive A and B, or $40 and $60). We used these counts to estimate the total hours worked on each type of sample.

In order to test sensitivity to the particular specification used, we also used a simpler model where we calculated the overall average length of time any attempt took. Then we simply counted the total number of attempts each type of sample had and calculated total hours using the simple average and the total count of call attempts.

RESULTS

We begin with the simple model. There were 52,894.4 hours of interviewing time during the experiment. This means the average call attempt length was 0.423 hours or about 25.4 minutes

Table 1 summarizes the results. We know the total number of attempts made under each incentive amount. These counts are multiplied by 0.423 to obtain the estimated hours applied to each type of sample. In order to rescale the cost estimates to a per interview basis, we also show the number of completed interviews.

Table 1. Estimated hours per interview (HPI) of two incentive treatments using simple model.

	Incentive amount	
	$40	**$60**
Total attempts	62,334	62,605
Estimated hours	26,367.30	26,481.90
Completes	2,435	2,725
HPI	10.8	9.7

The hours per interview (HPI) are the total estimated hours divided by the number of completes. This simple model results in estimated savings of 1.1 hours under the higher incentive amount.

Next, we present the results from the more complex model. Table 2 shows the estimated coefficients from our regression model. We created several categories of outcomes to be predictors in the model. First, we wanted to include the distinction between screening and main interviews. We also wanted to treat the attempts that result in an interview differently. The screening interview includes developing a roster of the household and selecting a respondent. The main interview is about 60–80 minutes long. Call attempts that have contact normally take more time than those that do not have contact. Therefore, we distinguish between attempts that had contact (other than interviews) and those that do not. Finally, cases that are judged to be nonsample (e.g. vacant housing units) may take a different amount of time since these units need to have their status verified. They are also treated separately.

In addition to categories of outcomes, we had several other measures of effort that are not directly related to contact attempts. The first of these is an intercept in the model. This intercept captures, in a sense, the hours required to get started working each week. This might include a weekly meeting, talking to supervisors, and completing administrative tasks. A second predictor of this type is the number of days worked. This is developed from the timesheet data. This would capture effort to begin call attempts each day. This could include, for example, time spent planning the day's work, travel from the interviewer's home to sampled area segments, and other such activities. Finally, we counted the number of area segments visited. There may be time associated with travelling between area segments (Wagner and Olson 2018). We would expect more travel the more segments that are visited each day.

Table 2. Estimated coefficients from a model predicting hours worked in a week.

Predictor	Estimated coefficient	Standard error
Intercept	7.85	0.31
Number of days worked	1.64	0.10
Main interviews	1.38	0.04
Screening interviews	0.24	0.02
Main contact	-0.07	0.02
Screening contact	0.11	0.03
Main no contact	0.15	0.01
Screening no contact	0.06	0.00
Nonsample	0.25	0.03
Number of segments visited	0.75	0.05

We note that one of the estimates is negative. Under the interpretation that each coefficient represents the amount of time that each attempt of that type takes, this coefficient is nonsensical. However, in the context of the model, the interpretation is that an interviewer who makes a "main contact" attempt in a week will work 0.07 hours less than an interviewer with the same number of call days worked, segments, visited, and call attempts of the other types who do not have a "main contact" attempt. The explanation for this negative estimate is a combination of misspecification, sampling error, and collinearity between the predictors. In this specific instance, as we will see in Table 3, this negative estimate does not have much impact on estimates.

Table 3. Estimated hours per interview (HPI) based upon the model in Table 2

Description	Est. coefficient	$40 count	$40 hours est.	$60 count	$60 hours
Intercept	7.85	900	7,061.00	899	7,053.10
Days worked	1.64	3,882	6,354.40	4,049	6,627.80
Completed main iws	1.38	2,435	3,358.30	2,725	3,758.20
Completed scrn iws	0.24	7,491	1,801.30	7,541	1,813.30
Contact main	-0.07	8,156	-569.5	8,450	-590
Contact scrn	0.11	3,870	433	4,039	451.9
Noncontact main	0.15	10,309	1,548.90	9,681	1,454.50
Noncontact scrn	0.06	28,715	1,761.00	28,962	1,776.20
Nonsample	0.25	1,358	333.3	1,207	296.2
Trips	0.75	5,223	3,939.10	5,612	4,232.40
Est. total hours:			26,020.70		26,873.70
Est. HPI:			10.7		9.9

This more complex model results in estimates that are similar to those resulting from the simple model. The larger incentive results in a reduction in the HP of about 0.8 hours per interview. This is a slightly lower estimate of the savings due to the higher incentive. However, on a large project, three-tenths of an hour difference on the estimated cost of each interview can be important. When factoring in the cost of the incentive, this savings in hours means that

there is a slight cost advantage to the higher incentive amount (see Wagner et al. 2017 for additional details).

DISCUSSION

There are situations where we do not have direct measures of detailed costs for survey design features. Paradata may be a useful input for modeling strategies aimed at producing estimates of these costs.

One issue is model specification. In the example presented here, the difference between the estimates of the HPI resulting from the two different models might seem small, but on a large project, these differences can be important. Three-tenths of an hour, across 20,000 interviews, is a large number of hours. In this case, we believe that the more complex model is more accurate, since we found that the distribution of call attempt types did change across the incentive treatments. The higher incentive resulted in a distribution of call attempt types that included fewer contact attempts relative to the number of interviews.

The modeling approach helps answer questions about costs, but it can be sensitive to model selection. Therefore, careful thought about which predictors to include, and assessing the sensitivity of the results to these choices by trying several different models is an important step.

Further enhancements could be made to the modeling approach. Specifically, additional features could be added to the model, such as characteristics of the sample (at the level of the interviewer day). For example, whether the sample is in a metropolitan or nonmetropolitan area could be an additional predictor. The model could estimate separate intercepts for each interviewer, under the assumption that they each behave differently. This would certainly complicate the estimation of hours spent on each arm of an experiment. The method also could allow for estimating variance. The estimates provided for the experiment did not evaluate whether the differences were statistically significant. This would be a useful extension. Finally, methods other than regression could be used to estimate parameters. Iterative techniques, for example, could be used to identify a set of parameters that provide a solution to the problem subject to constraints (e.g. all coefficients must be greater than zero).

ACKNOWLEDGEMENTS

The National Survey of Family Growth (NSFG) is conducted by the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS), under contract # 200-2010-33976 with University of Michigan's Institute for Social Research with funding from several agencies of the U.S. Department of Health and Human Services, including CDC/NCHS, the National Institute of Child Health and Human Development , the Office of Population Affairs , and others listed on the NSFG webpage (see http://www.cdc.gov/nchs/nsfg/). The views expressed here do not represent those of NCHS or other funding agencies.

Correspondence:
James Wagner
4053 ISR
426 Thompson St
Ann Arbor, MI 48104
Email: jameswag@umich.edu
Phone: 734-647-5600

REFERENCES

Andresen, E.M., C.R. Machuga, M.E. Van Booven, J. Egel, J.T. Chibnall, and R.C. Tait. 2008. "Effects and Costs of Tracing Strategies on Nonresponse Bias in a Survey of Workers with Low-Back Injury." *Public Opinion Quarterly* 72 (1): 40–54.

Gfroerer, J.C., J. Eyerman, and J.R. Chromy, eds. 2002. *Redesigning an Ongoing National Household Survey: Methodological Issues*. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Pruchno, R.A., and J.M. Hayden. 2000. "Interview Modality Effects on Costs and Data Quality in a Sample of Older Women." *Journal of Aging and Health* 12 (1): 3–24.

Wagner, J., and K. Olson. 2018. "An Analysis of Interviewer Travel and Field Outcomes in Two Field Surveys." *Journal of Official Statistics* 34 (1): 211–37.

Wagner, J., K. Olson, and M. Edgar. 2017. "The Utility of GPS Data in Assessing Interviewer Travel Behavior and Errors in Level-of-Effort Paradata." *Survey Research Methods* 11 (3): 219–33.

Wagner, J., B.T. West, H. Guyer, P. Burton, J. Kelley, M.P. Couper, and W.D. Mosher. 2017. "The Effects of a Mid-Data Collection Change in Financial Incentives on Total Survey Error in the National Survey of Family Growth." In *Total Survey Error in Practice*, edited by P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West. New York: Wiley.