An Exploration into the Use of Paradata for Nonresponse Adjustment in a Health Survey

Aaron Maitland[1], Carolina Casas Cordero[2], Frauke Kreuter[2]
[1]National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782
[2]University of Maryland, 1218 Lefrak Hall, College Park, MD 20742

Abstract

Nonresponse is a growing concern to survey researchers. One strategy for reducing the effect of nonresponse is statistical adjustment of survey estimates. Methods available for nonresponse adjustment include post-stratification and ratio adjustment. Another approach starts with populating the sample frame with a set of auxiliary variables that are correlated with both the probability of response and the survey variables of interest. The recently released 2006 National Health Interview Survey (NHIS) paradata file contains information, collected during the process of recruiting sample households to participate, that could be potentially useful for nonresponse adjustment. Information collected includes indicators of respondent reluctance and strategies for recruiting sample members. The goal of this paper is to test hypotheses about correlations between paradata variables and survey outcomes.

Key Words: nonresponse, paradata, call record

1. Introduction[1]

Survey researchers have been experiencing a decline in respondents' willingness to cooperate with survey requests over the past several years (De Leeuw and De Heer 2002). This primarily concerns survey data users because the decline in response rates increases the potential for nonresponse bias. Nonresponse bias occurs when the response rate is less than perfect and the nonrespondents systematically differ from the respondents on the survey variables of interest. In general then, researchers have two strategies to minimize the effect of declining response rates. The first strategy is to increase response rates. If one were to successfully recruit all sample individuals to participate, there could not be any nonresponse bias. However, surveys rarely gain participation from all sample individuals.

A second strategy is to find weighting variables that can be used to statistically adjust survey estimates during post-processing of the survey data. Nonresponse adjustment variables have a couple of requirements. First, the variables must be available for both respondents and nonrespondents. Additionally, the variables need to be correlated with both the sample individual's probability of responding and the survey variables of interest. In fact, a simulation study by Little and Vartivarian (2005) concluded that nonresponse adjustment variables were successful at reducing bias only when they were highly correlated ($\rho \approx .8$) with both response propensity and the survey variables of interest.

Traditional adjustment variables are found on most sampling frames. For example, variables such as region and urbanicity are frequently used in nonresponse adjustments. It is not surprising that these adjustment variables often have small effects on survey estimates, given their limited scope. More recently, survey researchers have been exploring the use of a new set of variables (Groves, Wagner, and Peytcheva 2007, Kreuter, Lemay, and Casas-Cordero 2007, Peytchev and Olson 2007). This new set of variables consists of paradata or information that is collected about sample households during the process of data collection. This information is potentially more relevant to the topics addressed in the survey interview.

The purpose of this paper is to investigate the magnitude of the correlations between paradata variables that we are considering as adjustment variables (z), the probability of participating in a survey (p), and the survey variables of interest (y).

[1] The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

2. Data and Methods

The data for this paper come from the 2006 National Health Interview Survey (NHIS). The NHIS is a multi-purpose health survey consisting of several data sets. The paradata file and family files contained all of the data that were used in the analyses for this paper.

1.1 Sampling

The NHIS is a cross-sectional survey. It is an in-person survey with a multistage area probability sample of households. At the first stage, 428 primary sampling units (PSU's) are drawn from approximately 1,900 geographically defined PSU's from the 50 States and the District of Columbia. Area segments are then selected within PSU's. Procedures are followed to oversample blacks, Hispanics, and Asians.

1.2 Paradata

The 2006 NHIS Paradata File is a public use file that contains information about all 44,264 families in the NHIS sample. We selected 31,142 of these families where at least one in-person contact was made with a sample unit member. Most of the cases (9,994) excluded from the analyses were screened out of the sample for reasons such as the household was occupied entirely by Armed Forces adults or race/ethnicity screening requirements. We excluded these cases because the survey variables of interest were not measured for any of these cases. The remainder of the excluded cases (3,128) consisted of families where there was no evidence of an in-person contact attempt. This could have occurred with cases where contact was never made or because interviewers failed to record in-person attempts in the Contact History Instrument (CHI) and subsequent attempts were made by telephone. We chose to analyze cases with at least one in-person contact attempt so that it was at least theoretically possible for the interviewers to have observed the paradata variables of interest. The overall response rate for this sample was approximately 90%.

The Paradata File includes three different types of variables that were considered as potential nonresponse adjustment variables. All of these variables were recorded by interviewers using the U.S. Census Bureau's Contact History Instrument (CHI). One set of variables describes the effort involved in making contact with the household. These measures range from indications that no one was home to the type of effort that the interviewer was making to contact a household such as driving by a home, observing no one was home, or speaking with a neighbour. It also measures occurrences like barriers to contact and inability to locate a household. Another set of variables measure the cooperation of a household. These are mostly reasons that respondents mentioned for not participating in the interview. For example they might have said they were not interested, too busy, or had privacy concerns. These reasons could have been mentioned on the doorstep, during the interview, or after completion of the interview. Table 2 contains a complete list of the contactability and cooperation paradata variables used in this study.

The other theoretically interesting variable that we considered was whether or not the interview was ever broken off or the case required follow-up due to health reasons. We chose this variable because it was potentially related to several survey variables of interest.

The public use file consists of case level summaries of the contact history for each case. It does not include information on every contact attempt. In other words, the variables on the Paradata File indicate whether or not any of these cooperation, contactability, or break off reasons were ever observed during the entire field period of a case. The file does not indicate how many times these variables were observed or on which contact attempt something was observed.

1.3 Survey Variables of Interest

We chose 30 variables from the NHIS Family File as our variables of interest. This information was collected for 28,230 families in our sample. We selected variables that were measured for all cases on the file. A few variables only pertained to families with children and were excluded from our analyses. The variables generally measured the health status of members of the household, any health difficulties experienced by members of the household, and the utilization of health services by members of the household. The variables used in the analyses include a mixture of dichotomous and count variables. For example, one variable recorded whether any family member was limited in any way whereas another variable recorded the number of family members that were limited in any way. The information was provided by a knowledgeable household respondent.

1.4 Analysis Plan

Our analysis plan was to begin by examining bivariate correlations between the paradata variables that we are considering for adjustment variables (z), survey participation (p), and our survey variables of interest (y). It is important to clarify that the paradata variables and survey participation indicator were collected for all individuals in the sample. The survey variables of interest were only collected for responding households. Throughout this analysis we are assuming that the correlation between the paradata variables and the survey variables of interest are the same for both respondents and nonrespondents. Next, we created some composite variables from the individual paradata variables. We took this approach, because it is possible that the variables may perform better as a composite than individually. One explanation for this is that the individual variables contain a lot of measurement error for reasons such as the interviewer forgetting to record some observations or some of the variables might be difficult to observe. We performed a factor anlaysis based on tetrachoric correlations separately for the contactability and cooperation variables to discover which variables within each set are most closely related. We then estimated factor scores based on the factor model and checked the correlations between these factor scores, response status, and the variables of interest. Last, we ran a logistic regression predicting survey participation using the cooperation and contactability paradata variables. We then output response propensities for each case based on this model and correlated these response propensities with the survey variables of interest.

3. Findings

We began by examining the bivariate correlations between the paradata variables (z), survey participation (p), and the variables of interest (y). We did this for each set of variables described above. The scatterplots in Figure 1a and Figure 1b illustrate the relationship between the z-p correlations and the z-y correlations for both the contactability and the cooperation paradata variables. In general, the range of the z-p correlations was much wider for the cooperation variables. These correlations ranged from .02-.41 and the average correlation was .14. The strongest predictor of response was the paradata variable indicating whether the respondents expressed a lack of interest or that they did not want to be bothered. In contrast, the correlations between the contactability variables and response status were between 0 and .17 and the average correlation was .07. The strongest predictor of response was whether the interviewer observed that the household did not answer the door and there was evidence that someone was home.

As observed in Figure 1a and Figure 1b, the correlations between the paradata variables and the survey variables of interest were much weaker. The strongest z-y correlation observed in the data was .1. The average z-y correlation was .05 for both the contactability and cooperation sets of paradata variables.

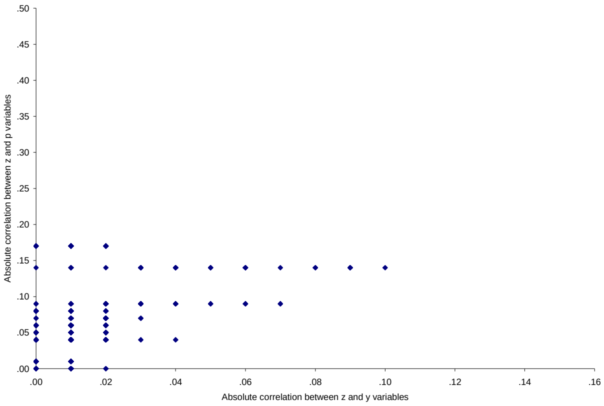

Figure 1a: Relationship between z-y and z-p correlations for contactability variables.

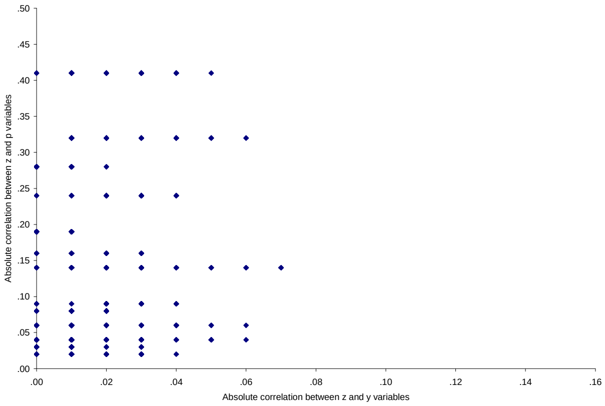

Figure 1b: Relationship between z-y and z-p correlations for cooperation variables.

Another variable we examined was whether or not the interview was ever broken off or the case required follow-up due to health reasons. This absolute correlation between this variable and survey participation was .06. The largest correlation with the survey variables of interest was .10.

It is possible that the variables may perform better as a composite rather than individually. One explanation for this is that the individual variables contain a lot of measurement error either because the interviewer forgot to record a variable or some of the variables might be difficult to observe. Hence, we factor analyzed the contactability and cooperation measures separately to discover which variables are most closely related. We then estimated factor scores based on the factor model. Since the indicators in the factor model are dichotomous measures we based the factor analysis on tetrachoric correlations as suggested by the work of Muthen (1989). We used the Unweighted Least Squares (ULS) method of estimation with a promax rotation of the initial factor solution. The same methods were used for both factor analyses.

The factor analysis of the cooperation variables yielded four factors. The variables loading most heavily on the first factor were those measuring whether or not a household member mentioned that the survey is voluntary, privacy concerns, anti-government concerns, that the survey content does not apply, asked questions about the survey, or mentioned any concerns at all. This first factor is potentially interesting because many of these variables are measuring respondents' reactions to information about the survey. The second factor included variables measuring time concerns such as a household member mentioned they are too busy, the interview takes too much time, respondent broke appointments, or had other scheduling difficulties. The third factor included variables that measured how hostile a household was to the survey request. For example someone in the household might have mentioned that they were too busy or did not want to be bothered, slammed the door on the interviewer, or been hostile or threatened the interviewer. The final factor included variables that indicate the extent of gatekeeper issues such as the interviewer could only talk to a specific household member, other household members tell respondent not to participate, or other family issues were mentioned.

The factor analysis of the contactability variables yielded three factors. The first factor measured the effort that interviewers made at contacting sample households such as the interviewer visited the home and recorded that no one was home, no one was home and an appointment was broken, no one was home and a previous letter was taken, the household did not answer the door when it appeared someone was home, the interviewer drove by the home, the interviewer spoke to a neighbor, or other contact problems were recorded. The second factor measured how difficult it was for an interviewer to locate a household or classify a case. For example the interviewer recorded that the address does not exist, the household was on vacation or away from home, or the interview recorded that a case was ineligible. The third factor measured the extent to which the interviewer faced barriers when accessing a property. For example, the interviewer encountered a locked gate or buzzer entry or the interviewer contacting a doorman or building management.

Table 1 shows the absolute correlations between the factor scores, response status, and the survey variables of interest. The results for the cooperation variables are shown in the top half of the table and the results for the contactability variables are shown in the bottom half. Overall, the cooperation variables are more strongly correlated with response status. However, neither set of factor scores improve the correlations with the survey variables of interest.

Table 1: Correlations between factor scores, survey participation, and survey variables of interest.

Variable set	Correlation between factor scores and survey participation	Correlation between factor scores and survey variables of interest	
		Range	Average
Cooperation variables			
Factor 1: Survey content/privacy	.28	.01-.06	.03
Factor 2: Time concerns	.24	.01-.08	.04
Factor 3: Hostility to survey request	.47	0-.05	.02
Factor 4: Gatekeeper issues	.07	0-.04	.01
Contactability variables			
Factor 1: Contact problems	.18	0-.10	.05
Factor 2: Location/classification issues	.02	0-.04	.02
Factor 3: Barrier issues	.11	0-.06	.03

Finally, we ran a logistic regression model with survey participation status as the dependent variable and all of the individual paradata variables as predictors. From this model we obtained the estimated response propensity for each person. Next, we correlated these response propensities based on the logistic regression model with the survey variables of interest. Once again, the correlations with the survey variables of interest were quite small. The average correlation was only .02 and the largest correlation was only .04.

4. Discussion

As Groves and Couper (1998) proposed, information about the process of recruiting respondents to participate is important for explaining survey participation rates. Several of the paradata variables investigated in this study were related to survey participation. However, this does not necessarily mean that these variables would be suitable for nonresponse adjustment. None of the paradata variables were very strongly related to the survey variables of interest. In fact, the largest correlations between the paradata variables and the survey variables of interest were approximately .1. As indicated by Little and Vartivarian (2005), these weak correlations suggest that the CHI paradata variables probably would not lead to significant changes in survey estimates if they were part of nonresponse adjustment procedures.

Measurement error is one potential contributor to the weak correlations that we observed in this study. The topic of measurement error in the context of paradata is a relatively unstudied topic and deserves more attention. However, our efforts to smooth out this measurement error through factor analysis made no difference with respect to the correlations we observed.

Finally, paradata have multiple uses. Others have shown that paradata provide useful resources for managing data collection process and studying data quality (e.g. Dahlhammer et al. 2005). However, this study points to the importance of collecting paradata variables that are correlated with the survey variables of interest if finding nonresponse adjustment variables is one of the goals of paradata collection. Instruments such as the CHI could present a unique opportunity to allow the interviewer to collect more health specific information about a household that would be related to the survey variables of interest in health surveys. These variables should be carefully selected, but collecting this proxy information could lead to stronger correlations between potential adjustment variables and the survey variables of interests and ultimately to better nonresponse adjustments. This undertaking would have to balance the potentially beneficial data collection with the need to obtain informed consent from survey respondents and apply appropriate standards for protecting the confidentiality of the data that is collected.

References

Dahlhammer JM, Stussman BJ, Simile CM, Taylor B. Modeling survey contact in the National Health Interview Survey (NHIS). Proceedings of Statistics Canada Symposium: Methodological Challenges for Future Information Needs, 2005.

De Leeuw E, DeHeer W. Trends in household survey nonresponse: A longitudinal and international comparison. Survey Nonresponse, Groves R, Eltinge J, Little R, (eds.). Wiley: New York, 2002; 41-54.

Groves RM, Couper MP. Nonresponse in Household Surveys. Wiley: New York, 1998.

Groves R, Wagner J, Peytcheva E. Use of interviewer judgments about attributes of selected respondents in post-survey adjustment for unit nonresponse: an illustration with the National Survey of Family Growth. Proceedings of the ASA Section on Survey Research Methods, 2007.

Kreuter F, Lemay M, Casas-Cordero C. Using proxy measures of survey outcomes in post survey adjustments: Examples from the European Social Survey (ESS). Proceedings of the ASA Section on Survey Research Methods, 2007.

Little RJ, Vartivarian S. Does Weighting for Nonresponse increase the variance of survey means? Survey Methdology 2005; 31:161-168.

Muthen B. Dichotomous Factor Analysis of Symptom Data. Sociological Methods and Research 1989; 18:19-65.

Peytchev A, Olson K. Using interviewer observations to improve nonresponse adjustments: NES 2004. Proceedings of the ASA Section on Survey Research Methods, 2007.

Table 2: List of contactability and cooperation variables.

Contactability variables	Cooperation variables
No one home	Not interested/Does not want to be bothered
No one home - appointment broken	Too busy
No one home - previous note/letter taken	Interview takes too much time
Household does not answer door - evidence someone is home	Breaks appointments (puts FR off indefinitely)
Drive-by	Scheduling difficulties
Multiple drive-bys	Survey is voluntary
Unable to reach/locked gate/buzzer entry	Privacy concerns
Address does not exist/unable to locate	Anti government concerns
On vacation, away from home/at second home	Does not understand survey/Asks questions about survey
Spoke with neighbor	Survey content does not apply
Building management/doorman contact	Hang up/slams door on FR
Completed case (Type B or C)	Hostile or threatens FR
Other specify	Other household members tell respondent not to participate
	Talk only to specific household member
	Family issues
	No concerns
	Other specify