A COMPARISON OF TWO METHODS TO ADJUST WEIGHTS FOR NON-RESPONSE: PROPENSITY MODELING AND WEIGHTING CLASS ADJUSTMENTS

Barbara Lepidus Carlson and Stephen Williams
Mathematica Policy Research, Inc., Princeton, New Jersey 08543-2393

KEY WORDS: non-response; weighting; propensity modeling; Community Tracking Study

Abstract: The most common method used to adjust sampling weights for non-response involves forming weighting classes or cells of homogeneous sample members. Within each cell, the weights of the respondents are inflated to account for the non-respondents. Some problems inherent in this method are cells with too few respondents, adjustment or inflation factors that are too high, and potentially large differences in these adjustment factors from one cell to the next. While there are ways to deal with some of these problems, they generally involve the risk of increasing the mean square error, either through bias or in the variance of the weights. A relatively new approach involves developing logistic regression models to predict response, using a potentially much broader set of predictive variables than can be used in the weighting class methodology. The inverse of the response propensity resulting from the application of such a model can then be used as the adjustment factor to the weights. We applied both of these methods when computing weights for round two of the Community Tracking Study Household Survey. This paper explores the differences resulting from these two methods.

Introduction

Weighting for Non-response – Basic sampling weights reflect the probabilities of selection of the sampling units. An important step in the creation of analysis weights for survey data is the adjustment of the sampling weights to account for non-responding sample units. Typically, eligibility of some of the sample elements are not resolved and interviews are not obtained for some of the eligible sample members. As nonresponse increases, so does the potential for serious biases in the survey results. The most effective solution is to obtain the highest possible response rates. For nonresponse that cannot be avoided, however, we adjust the sampling weights such that the resulting bias is reduced.

There are a number of ways to make such adjustments, but the general goal is to “weight up” the responders so that they account for the non-responders as well as themselves. The idea is to do this in such a way as to minimize the bias of estimates due to non-response while also minimizing the variance of the estimates. Because high variability in the analysis weights can cause variability in the estimates, we keep this in mind when making adjustments to the weights.

The more information we know about nonrespondents, the more effectively we can adjust the weights; that is, adjusting the weights of respondents that have similar characteristics to nonrespondents in order to offset for the missing reports. Often, especially in cross-sectional surveys, we have very little information about the nonrespondents. In these situations, weight adjustments typically are limited to information on the sampling frame, final disposition codes, and extraneous control totals. So-called weighting class adjustment is a reasonable method when relatively few variables are available for the weighting classes (sometimes just design strata).

As the amount of information about nonrespondents increases and as response rates decline, the use of response propensity models becomes more attractive. Response propensity models, using logistic regression, are becoming more in favor when the bias potential has a serious impact on the usefulness of survey results. This method can reasonably be viewed as a simple extension of the weighting class method from categorical to a response surface.

In this study we identified variables related to response rates and developed logistic models that estimate the propensity of individual households to respond, depending on their characteristics.

Weighting Class Methods – One common method for adjusting weights for non-response is to create homogeneous weighting classes or cells of sample members—both respondents and non-respondents. Ideally, these cells would be homogeneous with respect to the main analytical variables as well as the propensity for responding. The cells are formed by cross-tabulating a set of these variables. Within each cell, the responder weights are increased to take on the weights of the non-responders, under the assumption that they are alike. Unfortunately, there are some limitations to this method. First, the information that is used to form these cells must be available for both responders and non-responders. While there is often information available on all sample members when a sample is drawn from a list frame, that is generally not the case for random-digit-dial (RDD) telephone survey samples or area probability samples. Second, in order to ensure relatively stable adjustments, there are some rules that are usually used when forming the cells, such as ensuring a certain number of responders per cell (say, at least 20), and a certain ratio of responders to non-responders in the cell (say, fewer non-responders than responders so that the adjustment factor is less than two).

Propensity Modeling – An alternative and increasingly popular method for adjusting for non-response is propensity modeling; that is, creating a logistic regression model that predicts the likelihood of response versus non-response. This model makes use of any and all available and relevant data on the right-hand side of the equation. This model is then applied to the responders, and a log probability of responding is generated for each case. The weighting adjustment factor is then calculated as the inverse of this probability. While you are still limited to those variables available for both responders and non-responders (as you are with the weighting class methodology), you are not as limited by the cell sizes and ratios inherent in the other methodology, and therefore are not as restricted in terms of the number and type of variables. Furthermore, the adjustment factors will tend to be more stable and "smoother"; that is, with the weighting cell approach, there is more danger of having drastically different adjustment factors across cells while having identical adjustment factors within cells. Using a modeling approach reduces this problem.

What This Paper Investigates – In this paper, we investigate the impact of non-response propensity modeling as a useful tool for adjusting weights in round two of the Community Tracking Survey. Because some cases in round two of the survey were associated with completed round one interviews, we had substantially more data on which to model response propensity for those cases. We could then make use of this extra information when adjusting for unresolved screening cases and non-responding eligible households.

Methods

CTS Survey – The Community Tracking Study (CTS) is a national study of the rapidly changing health care market and the effects of these changes on people. Funded by the Robert Wood Johnson Foundation, the study is being conducted by the Center for Studying Health System Change. (Information about other aspects of the CTS is available at www.hschange.com.) Mathematica Policy Research, Inc. is the primary contractor for the household survey component. The third round of the household survey was completed in September 2001. The sample in the first round (1996-97) consisted of two independently drawn random-digit dial (RDD) national samples (one clustered within 60 randomly selected sites and one un-clustered) and a small in-person component selected within 12 of the 60 sites. For the RDD sample in the second round (1998-99), we selected a certain percentage of telephone numbers that were associated with round one completed interviews, a certain percentage of telephone numbers that were associated with round one sample members not resulting in completes, and then new sample (telephone numbers that existed in round one but which were not selected, and those that did not exist in round one). While the sampling unit was the telephone number (or address), the data collection unit was the "family insurance unit" (which is a family unit that would typically be covered together under a policy), and the analytical unit was the person or the family insurance unit.

Table 1 shows the counts of the telephone numbers released in the RDD component and the outcome categories. The Site Sample is the clustered sample of 60 sites and the Supplemental Sample is the smaller un-clustered component. Overall, the un-weighted household-level response rate for the RDD component was 65.9 percent. For the re-interview component—those telephone numbers associated with round one completes—this rate is 83.2 percent. As you can see from Table 1, the re-interview sample comprised only 24 percent of the telephone numbers released in round two, but represented almost 50 percent of the round two respondents. The re-interview cases had lower rates of non-response and ineligibility than the other cases released.

Weighting Classes Used – When creating the weighting classes for the RDD component, we used the following variables: round one disposition (round one complete, round one non-complete, not selected in round one, or not in existence in round one), site (primary sampling unit for the clustered sample), and stratum (or substratum for the clustered sample). Some cells were combined due to inadequate size. Approximately 320 cells in the site sample and 20 in the supplemental sample were used.

Development of Propensity Models– Potential variables for the models were identified from the round one questionnaire, the round two sample management files, and variables and methods used by other studies. We focused specifically on the Round 2 (R2) households that completed interviews in Round 1 (R1)—the re-interview component. A screening model was developed to adjust for the RDD cases for which we were unable to resolve eligibility. (A telephone number was considered eligible if it was a working residential number.) Note that the percentage of unresolved cases is small for this re-interview group. We also developed an interview response model, for those re-interview telephone numbers determined to be eligible. We identified variables related to interview response rates and developed logistic models that estimate the propensity of individual households to complete an interview, depending on their characteristics.

The models were run on household-level (or telephone number-level) files. Once potential variables were identified, we reviewed frequency tables to identify correlations between household-level variables and survey disposition. The list of initial candidate variables and levels was long, approximately 200. We then reduced the candidate variables, identified significant interactions, and estimated the model parameters for four models: Response Model and Screening Model for each of the two samples (site and supplement).

We attempted to capture those variables that were statistically significant (Chi-square) and showed the widest range of response rates. To reduce the number of model variables and avoid small-count cells, we combined categories with similar screening response rates and in as logical a fashion as feasible. We first used stepwise weighted regression in an attempt to reduce the candidate variables to 10 or 20 before we considered possible interactions or logistic regression. Backward stepwise solution in SAS was used, setting the significance at 0.10.

We summarized weighted and unweighted counts for each of the variables. From these, the screening and response rates were calculated for each variable and variable level. We then reviewed both response rates to identify those variables for which the rates vary the most dramatically. The Chi-square tests for independence identified variables, but the response rate tables were particularly useful for grouping levels for the initial least squares regressions. The weights for national estimates based on the site sample (household level) were used.

The next step was to run contingency tables. The data were weighted (scaled down to sample size) and produced a range of test statistics for independence (including Chi-square). Two sets of tables were run: 1) using the dichotomy of determined or not determined (to be a household) and 2) the dichotomy of eligibles, interviewed or not interviewed. The test values were used to reduce the number of variables and also to identify candidates for interaction variables. This step actually told about the same story as the table of response rates but additionally presented statistical significance. Also, this approach is very similar to the CHAID method used in the National Survey of Family Growth and others. Combined use of contingency and rate tables reduced the approximately 200 variable categories to 52 in the case of the response model and 41 for the screener model.

Next we ran these somewhat reduced sets of variables and some of the potentially important interactions in stepwise regression (SAS). This procedure is often used to reduce the variable list because it is much faster than the iterative solutions of the logistic regression. Using the backward stepwise regression with rejection level set at 0.1, the Response Model variables were reduced from 52 to 21.

A traditional measure of model fit in linear regression is the R^2, or coefficient of determination. This value was small for the models (0.08) as it typically is for binary dependent variables (Hosmer and Lemeshow, 2000). It also is basically useless for measuring model fit for the binary situation. Without the R^2 to assess goodness of model fit and predictive power, numerous diagnostic and test statistics exist with no clear favorite. So far, we are relying only on the maximum likelihood ratio tests, comparing the Chi-square values for the full model versus that for the model with only the intercept. Also, the full model has been compared to models based only on different site/stratum variables as a comparison against weighting classes using such variables. These comparisons have demonstrated the usefulness of the full model, but they are only tests of significance, not tests of goodness of fit. (The Hosmer-Lemeshow goodness of fit tests and a generalized R^2 are recommended by some but have not yet been calculated.)

The variables coming out of this step were used in SAS stepwise Logistic with reject level set at 0.05 to identify the final set of variables and interaction terms. Again, four models were developed, one for screening and one for response, for each of the site and supplemental samples.

The final stage was to enter the variables into the Logistic procedure in SUDAAN, and run for the final parameter estimates, using the design strata. The models were run using this specialized software in order to estimate variances correctly, due to the complex nature of the sample design of the CTS.

The change in model coefficients was minimal, but as expected the estimation variances increased slightly (the conclusions about model effectiveness were unchanged). The results of the four models are presented in Table 2 (site sample and supplement sample, screener and response models).

The fact that 24 variables are highly significant for the Response Model gives us strong evidence of its usefulness. (The Chi-square test showing model significance is not surprising.) Similarly, with 16 significant (0.05) variables, the Screening Model was also found to be useful.

Weighting Steps – For both methods used, there are a series of steps in the weighting process:

First, we create a sampling weight for each telephone number based on its probability of selection. Then we adjust for eligibility determination: first for whether the eligibility of the telephone number was determined (did we determine if it was a working, residential number), and next for whether the survey eligibility of the household was determined (is there at least one eligible family unit; i.e., a family unit containing a civilian adult), among eligible telephone numbers.

Then we adjust for whether the [eligible] household responded. Other adjustments include those for multiple telephone numbers in the household, as well as interruptions in telephone service, and a post-stratification adjustment to the number of telephone and non-telephone households.

Next, we apply these household weights to each family in household to create family unit weights, and then make adjustments at the family level: first for whether the eligibility of the family unit was determined (among families in responding households), and then for whether the [eligible] family unit responded.

Finally, we create person-level weights (adjusting for the probability of selection of one child within each family unit), and then adjust for high person-level item non-response. (We treat as unit non-response, even though there was one informant for all members of family unit.)

After these steps, the RDD site sample and in-person samples were integrated and their weights adjusted to account for dual chances of selection. The final steps involved post-stratification, weight trimming, and re-post-stratification.

The differences between the two methods are reflected in the adjustments for the resolution of telephone number eligibility, the resolution of survey eligibility, and household interview response. For the re-interview sample (round 1 completes), we used a model to make the first of these adjustments, and another model to make the third one. We skipped the interim step, under the assumption that virtually all eligible telephone numbers in this sample were associated with eligible households. So all households with undetermined eligibility were included with known eligibles in the third step.

For the round 1 non-completes and cases new to round 2, we combined the first two steps into one weighting class adjustment. The adjustment factor was the ratio of the following two weighted sums. Numerator: phone numbers known to be eligible (known households) plus a certain fraction (using external source) of those with undetermined eligibility. Denominator: households with determined survey eligibility. The third step was also a weighting class adjustment, the ratio of the following two weighted sums. Numerator: all eligible households. Denominator: all responding eligible households.

For the purposes of methodological investigation, we also weighted the round 1 complete cases using only weighting class adjustments.

How They Were Compared

We compared the two sets of weights: a weighting-class-only approach (re-interview and non-re-interview samples) and a weighting-class/propensity-model hybrid approach (re-interview sample used modeling, non-re-interview sample used weighting classes).

Because one of the purported benefits of using the propensity model approach is less variability in the weights, we compared the design effect due to unequal weighting $(1+(cv/100)^2)$ for the different weighting methodologies.

We then looked at estimates (generated in SUDAAN, to get appropriate estimates of the standard error and overall design effect) to see if the different approaches had an impact on the bias. We generated national estimates using eight different weights (site, supplemental samples)*(person-level, family-level) * (cell-only vs. propensity-hybrid methods). While the final CTS weights produced have the in-person component integrated, the results here pertain only to the RDD sample component. The in-person component was not involved in the modeling for non-response propensity.

Estimates were generated using four different variables: the percent of families with an emergency room visit or hospitalization in the last year; the percent of persons who are uninsured, the percent of persons who self-report being in excellent to good health, and the mean number of doctor visits in the last year (separately for children and adults).

We ran these estimates for the entire round two RDD sample and for the subset of those cases for whom a propensity model was run (those who were round one respondents). We also compared these estimates to those made using a completely unadjusted sampling weight; that is, the inverse of the probability of selection with no adjustments for non-response, or any other adjustments.

Results

Table 3 shows the design effects due to unequal weighting for the various weights and weighting methodologies. As one would expect, the sampling weights unadjusted for non-response have the lowest variation. But the propensity method and the weighting-cell method produced estimates with almost identical design effects due to unequal weighting.

Table 4 shows national estimates for the round two RDD sample, for the various weights. Once again, we see very little difference between the estimates using the propensity model weights and those using the weighting cell approach. Furthermore, the unadjusted sampling weight produced estimates very similar to the adjusted weights, with the exception of the percent uninsured. (The unadjusted sampling weights underestimated the percent uninsured.) The re-interview component has a lower estimate of the uninsured, presumably because it is a relatively more stable population than the rest of the sample.

Discussion

We have found very little difference between the propensity method and the weighting cell method when looking at the CTS household survey RDD re-interview sample. The expected benefits of the propensity modeling (less variation in the weights and less bias) were not seen. This is likely due to two main reasons: (1) the number of weighting cells here was so large (over 300) that the weighting cell approach nearly approximated the smoother nature of the propensity modeling approach, and (2) the screener and interview response rates among the household survey re-interview sample was high to start with, allowing for very little variation in the non-response adjustments.

It was difficult to quantify the relative levels of effort for implementation of the two methods. Modeling tends to be more labor-intensive than the weighting cell approach, although this can vary depending on how much collapsing of weighting cells was necessary. For both methods, once the initial work is done, repeating the same methodology in future rounds of the same survey would likely involve comparable labor.

Conclusions and Limitations

Results were almost identical for weighting classes versus propensity models. Although this should not be surprising, it is interesting to note that a parsimonious model containing a dozen or so carefully identified variables could produce essentially the same results as weighting class adjustments using hundreds of weighting classes based only on design features (strata and sampling units).

By comparing the models for the site sample with those for the supplement, we see some differences resulting from different design features, but to a large extent these models are and should be similar (because the two samples represent the same study population). Comparing the between-sample models suggests that the models may be somewhat stable and allow the same variables to be used effectively in subsequent rounds by only updating the parameter estimates.

In round two, both propensity models and weighting classes were used to adjust the sampling weights for incomplete screening and response. Both methods were somewhat labor intensive. A choice arises for future rounds between using the same hybrid or using only weighting class adjustments. Using only propensity models seems unwise since just frame information is available for those for those that were not re-interview cases. Consider also that even though the model usage would be much simpler in subsequent rounds, using the hybrid would still be more costly than using just weighting classes (the major part of the modeling effort was data mining—identifying the variables to use in the models). However, although little seemed to be gained for the national level estimates, we should not forget that site-level estimates are a major focus of the CTS.

Among the limitations of this evaluations are, first, that the results are survey specific; they cannot be expected to represent other surveys or other survey designs. Furthermore, we only examined a small number of analytic variables. For the analyses involving the entire round two sample (not just the re-interview cases), it should be kept in mind that the "propensity model" is a hybrid approach, where the impact of the propensity model is diluted with the weighting cell approach used for the non-re-interview cases.

The fact that the proportion of unresolved numbers and non-respondents are much greater in the non-re-interview group that used only weighting class adjustments lead us to suspect that the model adjustments do not have much opportunity to have an overall impact. Another obstacle for the model effectiveness is the fact that household composition sometimes changes between survey rounds in the re-interview cases, reducing the correlation between round one covariates and round two outcomes.

Future Research

This paper focused on national estimates. We should also look at weights for making site-specific estimates. The use of the large number of weighting classes should almost ensure that the non-response bias is minimized for national-level estimates. However, a major emphasis of the CTS is to characterize a sub-sample of the sites (communities) in detail. That is, the survey results for each of the twelve "high-intensity" sites are of primary interest. We should not assume that weighting class adjustments at the site level will be nearly as effective as at the national level; there were only 8 to 12 weighting classes within a site compared to hundreds, nationally. On the other hand, the models, structured to be site-specific should perform well at the site level.

We plan to calculate goodness-of-fit statistics for the four models presented here. We might also examine more variables within the household survey. The propensity modeling methodology was also used in round two of the CTS Physician Survey. Because this CTS component involves a list sample, there were more variables available from which to construct the models. We plan to investigate the impact of this approach for the physician survey as well. Of course, this approach could be examined in surveys other than the CTS as well.

References available from the authors.

Table 1. CTS Round 2 – Household Survey – RDD Sample – Telephone Numbers Released

	Round One Completes in Round Two Sample			Total Round Two Sample		
Round 2 Status	Site sample	Supplement	Total	Site sample	Supplement	Total
Eligible responder	11101	1281	12382	23246	2729	25975
Eligible nonresponder	104	13	117	565	66	631
Ineligible family	23	1	24	231	35	266
Undetermined family eligibility	1760	189	1949	8861	978	9839
Ineligible phone number	2388	275	2663	28301	3206	31507
Undetermined phone number	604	63	667	5537	593	6130
Total (=number in screener model)	15980	1822	17802	66741	7607	74348
Excluding ineligible phone (=no. in response model)	13592	1547	15139	38440	4401	42841

Table 2. Four Non-Response Propensity Models

Screener Model	Site	Supplement	Response Model	Site	Supplement
Intercept	3.349	4.367	Intercept	2.156	2.129
X1: Hhold is in Non-Metro Area	0.492	0.184	X2: Hhold is in Non-Metro Area	0.264	0.026
*x2b:In Nat Supplement, stratum 1		-0.298	*x2b:In Nat Supplement, stratum 1		0.267
*x3: In Site 5 or Site 12	0.530		*x4: In Site 1,7,8,9,11	-0.367	
*x6: Site 4,6 Str 3 /Site 9, Str 2	0.916		*x6: Site 3,5,10 Stratum 2	-0.277	
X8: Less than 6 R1 Contacts	0.433	0.415	X7: 1 Contact in R1	0.362	0.139
X9: Household has 2 families	-0.320	-0.647	X8: 6+ R1 Contacts	-0.509	-0.567
X12:One-person Household	-0.679	-0.600	X13:Household has 3+ children	0.236	-0.138
X16:Household Income $0-9999	0.343	-0.466	X23:Hhold has 1+ months w/o telephone	-0.509	0.360
X17:Household Income $20,000-39,999	-0.278	-0.928	X24:Hhold has non-pub telephone	0.124	0.305
X20:Hhold has someone cov. by Medicare	0.613	0.782	X26:Hholdr's Age 13-27	-0.643	0.052
X21:Hholdr Age 18-32	-0.502	-0.152	X27:Hholdr's Age 28-37	-0.320	-0.462
X27:Zero Doctor Visits	0.283	0.677	X29:Hholdr's Age 78+	-0.616	-0.250
X29:Someone needs specialist in hhold	0.365	0.140	X31:Hholdr Education 13+ years	0.224	0.289
X33:Hhdr has 2+ part-time jobs or not working	0.200	-0.469	X33:Hholdr is a Proxy	-0.971	1.588
X36:Hhdr Employed by state or local govt	0.429	0.297	X35:R1 Resp's name not given	-0.660	-0.814
X13:More than 5 person hhold	-0.108	-1.250	X37:Zero Doctor Visits	-0.198	0.041
X26:Somewhat satisf or neutral w/ hlth care	-0.083	1.020	X38:11+ Number of Doctor visits	0.142	0.338
X30:Everyone in hhold in excellent health	-0.101	-0.653	X39:No one needs specialist in hhold	-0.186	-0.029
X43:Hhold has 1+ Months without Telephone	-0.085	-1.483	X43:Hhdr has 2+ FT or PT jobs or not workg	-0.042	0.139
			X46:Hhdr self-employed or family bus/farm	-0.444	0.042
			X2526:Hhdr is 13-27 & not white	0.112	-0.975
			X2529:Hhdr is 78+ & not white	0.298	-1.396
			X10:Hhold has 4+ families	-0.348	-2.064
			X12:Household has more than 5 persons	0.121	1.003
			X21:Hhold has someone cov. by Military Insur.	0.064	-0.520
			X36:Hlth care satis: missg, sw satis, very dis	0.003	-0.460
			x48: Hhdr's Firm Size 1-999	0.100	-0.290

Table 3. Design Effects (Unequal Weighting)

			Propensity Weight		Weighting-Cell Weight		Unadjusted Sampling Weight	
Sample	Weight	Unwted n	C.V.	DEFFw	C.V.	DEFFw	C.V.	DEFFw
ALL ROUND 2 RDD SAMPLE								
Site	Family	28027	89.001	1.792	89.141	1.795	50.268	1.253
Supplemental		3251	53.786	1.289	53.768	1.289	11.362	1.013
Site	Person	51780	107.774	2.162	107.915	2.165	50.275	1.253
Supplemental		5982	88.322	1.780	88.421	1.782	10.743	1.012
ROUND 1 COMPLETES ONLY								
Site	Family	13456	59.170	1.350	58.664	1.344	48.600	1.236
Supplemental		1522	23.376	1.055	19.648	1.039	3.612	1.001
Site	Person	25316	76.418	1.584	75.813	1.575	48.399	1.234
Supplemental		2838	52.254	1.273	49.909	1.249	3.574	1.001

Table 4. National Estimates Based on Entire Round Two RDD Sample

Sample	Unweighted	Propensity Model Weight			Weighting-Cell Weight			Unadjusted Sampling Weight		
Component	Sample	Estimate	S.E.	DEFF	Estimate	S.E.	DEFF	Estimate	S.E.	DEFF
Variable = Percent of families with emergency room visit or hospitalization										
Site	28027	39.29	0.64	4.78	39.33	0.64	4.82	39.10	0.48	2.67
Supplemenl	3251	39.47	1.00	1.35	39.50	0.99	1.34	39.36	0.88	1.05
Variable = Percent of persons who are uninsured										
Site	51780	12.45	0.50	11.71	12.39	0.49	11.37	11.07	0.41	8.88
Supplement	5982	12.43	0.77	3.28	12.30	0.77	3.27	11.59	0.57	1.90
Variable = Percent of persons who are in excellent, very good, or good health										
Site	51780	88.12	0.32	5.21	88.06	0.33	5.24	87.59	0.30	4.19
Supplement	5982	87.46	0.55	1.64	87.38	0.55	1.65	87.60	0.47	1.22
Variable = Mean number of doctor visits (children / adults)										
Site	8910/42870	3.00/3.62	0.06/0.03	2.68/2.24	3.00/3.64	0.06/0.03	2.60/2.25	3.16/3.68	0.04/0.03	1.30/2.20
Supplement	1024/4958	3.34/3.57	0.21/0.07	3.19/1.32	3.35/3.59	0.22/0.07	3.20/1.32	3.20/3.58	0.11/0.06	0.99/1.11