Weighting Adjustments for Panel Nonresponse

Qixuan Chen[1], Andrew Gelman[2], Melissa Tracy[3],

Fran H. Norris[4], and Sandro Galea[3]

28 Aug 2012

[1]*Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA*

[2]*Department of Statistics, Columbia University, New York, NY 10027, USA*

[3]*Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY 10032, USA*

[4]*Department of Psychiatry, Dartmouth Medical School, Hanover, NH 03755, USA*

Email: qc2138@columbia.edu

Summary

Although similar to weighting for unit nonresponse in cross-sectional surveys, adjustment for panel nonresponse needs to incorporate information about nonrespondents collected in the early waves of the panel. We review different weighting adjustments for panel nonresponse and discuss methods for incorporating complex survey design variables into the weighting adjustments. We propose a cross-classified method for panel survey data with complex sampling design by first grouping respondents and nonrespondents with similar estimated response propensities to form response propensity strata and then cross-classifying the response propensity strata with the design variables. Our simulation study shows that when design variables are not related to nonresponse, the cross-classified method yields survey estimates that have bias and root mean squared error similar to the estimates weighted by reciprocals of the response propensities and the response propensity stratification method. When design variables are related to nonresponse, the cross-classified method yields estimates with smaller bias than both the other two methods if design variables are not included as covariates in the response propensity regression, but is comparable in the bias to the other two methods if the response propensity model is correctly specified. We apply these methods to the Galveston Bay Recovery Study, a panel study of trajectories of wellness in a community following a disaster.

Key words: adjustment cells; design variables; generalized raking; panel surveys; response propensity.

1 Introduction

Panel surveys collect similar measurements on the same sample at multiple points of time (Duncan and Kalton, 1987). This unique design of panel surveys is appealing to social and public health scientists in monitoring changes over time for a cohort of individuals. However, as with other longitudinal studies, panel surveys are subjected to dropouts during the follow-up. Some individuals participating in the baseline interview do not respond to one or more of the follow-up interviews.

In cross-sectional surveys, weighting adjustments are often for unit nonresponse when a sampled individual does not respond to the entire survey, and imputation is commonly used to handle item nonresponse for individuals who do not respond to particular questions. Unit and item nonresponse also arise in panel surveys and can be handled similarly using weighting and imputation, respectively. However, the choice between weighting and imputation is more complicated with panel nonresponse. Specifically, with weighting, information collected for panel nonrespondents in other waves is discarded, which results in a waste of costly collected data. On the other hand, with imputation, missing responses in the entire wave need to be imputed, which causes concerns about fabrication of large amount of information and attenuation of covariance between variables in the same wave or in the same variable across multiple waves of the panel. Further discussion of the weighting and imputation for panel nonresponse can be found elsewhere (Kalton et al., 1985; Kalton and Miller, 1986; Kalton, 1986; Lepkowski, 1989). Although imputation is more efficient than weighting with some well chosen imputation models, it can be challenging to construct imputation for panel nonresponse taking into account both cross-sectional and longitudinal dependence among the variables. Therefore, weighting adjustments are frequently used to handle panel nonresponse, for example in the Survey of Income and Program Participation (SIPP; Rizzo et al., 1994). In this paper, we examine weighting adjustments methods for panel nonresponse in both simple random samples and complex survey samples.

Two new complications arise when preparing weights for panel nonresponse. First, in a cross-sectional survey, only a limited amount of information (e.g. strata and primary sampling units) is typically available on unit nonrespondents. In contrast, nonrespondents in a wave of a panel survey are also characterized by a large amount of information from responses in the previous waves. Second, multiple sets of weighting adjustments are needed for different analysis purposes. For example, for a three wave panel, there are up to four response patterns, which can be written as 111, 110, 101, and 100, with 1 denoting response and 0 denoting nonresponse in a single wave. Thus three different sets of weighting adjustments are needed for first-wave respondents; these weights correspond to cross-sectional analysis of respondents in the second wave, cross-sectional analysis of respondents in the third wave, and longitudinal analysis of respondents who respond to all the three waves. In this paper, we discuss auxiliary variables that should be included to define weights and review the approach to create multiple sets of weights for different analyses depending on the waves from which data are needed.

2 Methods for Creating Weighting Adjustments

In this section, we consider weighting adjustments for second wave panel nonresponse. We assume that first wave respondents are a simple random sample of the target population.

2.1 *Adjustment Cell Method*

A common method for compensating for panel nonresponse is to form weighting adjustment cells of homogeneous sample units based on the auxiliary variables observed for both respondents and nonrespondents. Continuous variables are first categorized. Variables observed for both respondents and nonrespondents are then cross-classified to form adjustment cells. The inverse response rate in each adjustment cell is assigned as a weight to all the respondents in the cell to compensate for the nonrespondents in that cell. The choice of the auxiliary variables used to form adjustment cells is usually based on prior knowledge on how well a variable predicts nonresponse and outcomes of interest, often augmented by a logit or probit regression predicting response status from candidate adjustment variables. Little and Vartivarian (2005) showed that weighting can reduce nonresponse bias (but can increase variance) when a variable used to form adjustment cells is related to the probability of response. The nonresponse adjustment cell method was used in the SIPP panel by Chapman et al. (1986).

The nonresponse adjustment cell method requires the sample size to be large enough in each cell to obtain a stable response rate estimate. When the number of variables used to form adjustment cells is large, some of the adjustment cells can be small. As a result, some cells may contain few or even no respondents, and the response rates may vary a lot in different cells. Therefore, adjacent adjustment cells with similar estimated response rates are often collapsed to ensure a certain number of respondents (say, at least 20) and a certain ratio of respondents to nonrespondents (say, more than one) in each cell. Alternatively, the response propensity method (discussed below) can be used when the number of variables related to the response status is large.

Adjustment cells can also be formed using nonparametric methods such as regression trees. The chi-square automatic interaction detector (CHAID) categorical search algorithm (Kass, 1980) splits a data set progressively via a tree structure by choosing variables that maximize a chi-square criterion in each split, and can be used to divide sampled units into nonresponse adjustment cells. This method was applied to the SIPP by Rizzo et al. (1996), to the Medical Expenditure Panel Survey (MEPS) by Cohen et al. (1999), and to the British Household Panel Survey by Lynn (2003). SEARCH analysis, another version of the CHAID method, was used to create weighting adjustments by maximizing the variation explained in each split in the SIPP (Kalton et al., 1985; Kalton and Miller, 1986; Lepkowski et al., 1989).

2.2 *Response Propensity Method*

Another method frequently used to handle nonresponse in sample surveys uses propensity weighting and is a straightforward extension of the propensity score theory of Rosenbaum and Rubin (1983) incorporated into survey nonresponse problems by David et al. (1983). Traditionally, response propensities are estimated by fitting a logit or probit regression of the panel response status. With the large amount of information collected previously for both respondents and nonrespondents, an initial screening is often performed to reduce the number of variables to a more manageable size, by examining bivariate associations between each of the auxiliary variables and the panel response status. A subset of variables that are associated with nonresponse are identified, and the multivariable logit or probit regression is then performed on this subset, often with additional steps such as stepwise selection and inclusion of interactions. The response propensity for each respondent is estimated based on the selected model, and the weighting adjustments for the second wave respondents are set to the inverses of the response propensities. The response propensity weighting method was applied to the SIPP by Rizzo et al. (1996), and to the MEPS by Sommers et al. (2004) and Wun et al. (2007).

Compared to the adjustment cell method, propensity weighting is cleaner with continuous variables and allows the use of a larger set of auxiliary variables. However, propensity weighting has two potential limitations. First, the effect of weighting adjustments in reducing nonresponse bias largely relies on correct specification of the response propensity regression. If this model is misspecified, the nonresponse adjusted estimators of the population quantities are likely to be biased. To remedy this problem, Giommi (1984) proposed kernel smoothing and da Silva and Opsomer (2009) proposed local polynomial regression to estimate the response propensities. Secondly, some respondents can have very low estimated response propensities and thus receive very large weighting adjustment factors, which in turn can inflate the variance of the nonresponse-adjusted estimators of the population quantities. A common remedy is to trim or compress large weights (Potter, 1990; Kish, 1992; Meng et al., 2010). Alternatively, Little (1986) proposed a response propensity stratification method. The response propensity stratification method forms adjustment cells based on the estimated response propensities. Specifically, the estimated response propensities are first ordered; respondents and nonrespondents with similar estimated response propensities are grouped to form adjustment cells; and the respondents in each cell are weighted by the inverse of observed response rate in that cell. Since the estimated response propensities are used only for the purpose of forming adjustment cells, the response propensity stratification method relies less on correct specification of the response propensity regression model. Furthermore, the large weighting adjustments due to small estimated response propensities can be avoided by placing appropriate cutpoints in forming adjustment cells. Rosenbaum and Rubin (1983) suggested that five adjustment cells may provide the most effective bias reduction. Common approach is to define the adjustment cells using the quintiles of the distribution of the estimated response propensities.

2.3 *Generalized Raking Method*

Poststratification is used extensively in surveys to adjust for sample weights so that the estimated joint distribution of a set of poststratifying variables matches the known population joint distribution. When the population joint distribution is not available, raking is used to match the marginal distributions of a survey sample to the known population margins via iterative poststratification (Little and Rubin, 2002). The raking method for poststratification can be extended to handle panel nonresponse and is referred to as generalized raking or sample based raking (Kalton and Kasprzyk, 1986). Generalized raking is applied to the second wave respondents to make the marginal distributions of a set of variables match to the corresponding distributions among the first wave respondents. The raking variables are chosen to be predictors of response status, which are either obtained from prior knowledge or are identified through response propensity regression model. Deville and Särndal (1992) discussed several distance measures that can be used to derive raking adjustments. Rizzo et al. (1996) applied the generalized raking method to the SIPP panel using the CALMAR software (Deville et al., 1993).

2.4 *Auxiliary Variables for Weighting Adjustments*

Rizzo et al. (1996) suggested that the choice of auxiliary variables could be more important than the choice of methods for creating weighting adjustments. In that sense the most important property of a method is its ability to facilitate inclusion of a larger, more comprehensive set of variables without creating adjustments that are too noisy to be useful. The auxiliary variables used for weighting adjustments should be predictors of panel response status, thus including these variables in creating weighting adjustments can generally reduce nonresponse bias (Little, 1986; Kalton and Brick, 2000). In addition to the responses to the questions in the previous waves of the panel, variables measuring previous wave data quality can also be predictors of panel response status. Kalton et al. (1990) found that first wave respondents who were less cooperative were more likely to be the second wave nonrespondents in the American Changing Lives Survey. Rizzo et al. (1994) constructed an index of the number of imputed items in the first wave as a measure of cooperation and showed that the number of imputed items was a predictor of panel response propensity in the SIPP panel. Loosveldt et al. (2002) also found that the amount and patterns of item nonresponse in the first wave were related to the second wave nonresponse. Finally, Meekins and Sangster (2004) showed that call history variables, such as number of calls and whether the first wave respondents ever a refusal, had great predictive power for the second wave panel response status, because being hard to reach a sampled individual in the first wave interview can be considered as a negative reaction to the request to participate in the survey and thus increases the odds of nonresponse in later waves.

3 Use of Complex Survey Design Variables

A panel survey typically has differing sample weights for the sampled individuals even in the first wave, as result of unequal probabilities of selection or response. When sample weights are related to panel nonresponse, it is necessary to include them in the panel nonresponse weighting adjustments. Similarly, other survey design variables, such as strata and primary sampling units (PSUs), also should be used in the weighting adjustments if panel nonresponse propensities vary in different strata and PSUs. To use the adjustment cell method, instead of weighting respondents by the reciprocal of response rates, respondents in each adjustment cell are weighted by the inverse of the weighted response rate in that cell. The weighted response rate in a cell is calculated by diving the sum of weighted counts of respondents by the sum of weighted counts of respondents and nonrespondents combined. In the use of response propensity method, instead of fitting an unweighted logistic or probit regression model, a weighted regression model with the sample weights is used to estimate response propensities, where survey design variables are incorporated into the model fitting. Finally, to use the generalized raking method, the panel respondents' marginal distributions for each of the raking variables computed using the nonresponse adjusted weights are forced to equal to the corresponding distributions among the first wave respondents computed using the sample weights.

Little and Vartivarian (2003) argued that using the weighted response rates to incorporate design variables yielded biased estimates of population quantities if design variables were related to nonresponse and was unnecessary if design variables were unrelated to nonresponse. Instead, they suggested to cross-classify design variables with other auxiliary variables to create adjustment cells or to include design variables as covariates in the response propensity model. Grau et al. (2006) followed the recommendation of Little and Vartivarian (2003) and compared the weighting adjustments using weighted logistic regression and unweighted logistic regression with and without design variables as covariates in the Community Tracking Study Household and Physician Surveys. They did not find significant advantage of one method over the others and argued that any difference in the variance estimates of population quantities among different methods might be outweighed by the large variance due to the large adjustment factors using individual response propensity. Wun et al. (2007) also concluded that whether including sample weight as a covariate in the logistic regression does not make much difference in modeling the response propensity in the MEPS. Because there is no consensus as to the best way to incorporate complex survey design variables, some national panel survey studies used weighted logistic regression, e.g. the SIPP (Rizzo et al., 1996), and others used unweighted logistic regression, e.g. the third National Health and Nutrition Examination Survey (Ezzati-Rice and Khare, 1994), for weighting adjustments.

4 Weighting Adjustments for Multiple Wave Panel Nonresponse

Often more than one wave of the panel has panel nonrespondents. Three types of panel nonresponse are distinguished: attrition, reentry, and late entry (Little and David, 1983). Much of non-response in most of panel surveys are due to attrition, where a sampled individual who drops out does not return in any of the later waves of interviews. Little and David (1983) proposed a method for developing weighting adjustments to compensate for attrition nonresponse. For simplicity, we consider a three wave panel. Let $r = (r_1, r_2, r_3)$ denote the panel response status in each of the three waves, with 1 for response and 0 for nonresponse. Let z denote the set of design variables, which are observed for all the sampled units, and $x = (x_1, x_2, x_3)$ denote survey variables collected in each of the three waves of interviews, where x_k are observed when $r_k = 1$ and x_k are missing when $r_k = 0$ ($k = 1, 2, 3$). Little and David then run the following sets of logistic or probit regressions:

(1) r_1 on z, using all sampled units,

(2) r_2 on z and x_1, using sampled units with $r_1 = 1$,

(3) r_3 on z, x_1 and x_2, using sampled units with $r_1 = r_2 = 1$.

The respondents in the first wave are weighted by w_1, where $w_1^{-1} = \hat{p}(r_1 = 1|z)$. The respondents in the second wave are weighted by $w_2 = w_1 w_{2.1}$, where $w_{2.1}^{-1} = \hat{p}(r_2 = 1|z, x_1, r_1 = 1)$. Finally, the respondents in the third wave are weighted by $w_3 = w_2 w_{3.12}$, where $w_{3.12}^{-1} = \hat{p}(r_3 = 1|z, x_1, x_2, r_1 = r_2 = 1)$. The weight w_1 is indeed the unit nonresponse weighting adjustment and is usually incorporated into the sample weights for the sample in the initial wave of interview. The weight w_2 can be used for the cross-sectional analysis of respondents in the second wave as well as the longitudinal analysis of sampled units who respond to both wave 1 and wave 2. Similarly, w_3 can be used for the cross-sectional analysis of respondents in wave 3 as well as the longitudinal analysis of respondents to all three waves.

Little and David further modified the method to compensate for non-attrition nonresponse. Suppose for the first two waves of the panel, we observe four patterns of response: responding to both waves ($r_1 = 1, r_2 = 1$), responding to wave 1 only ($r_1 = 1, r_2 = 0$), responding to wave 2 only ($r_1 = 0, r_2 = 1$), and responding to neither wave ($r_1 = 0, r_2 = 0$). The weighting adjustments are based on the following three regressions:

(1) r_1 on z, using all sampled units,

(2) r_2 on z and x_1, using sampled units with $r_1 = 1$,

(3) r_2 on z, using sampled units with $r_1 = 0$.

The respondents in wave 1 ($r_1 = 1$) are weighted by w_1, where $w_1^{-1} = \hat{p}(r_1 = 1|z)$. The respondents in wave 2 are weighted by $w_{2.1}$, where $w_{2.1}^{-1} = \hat{p}(r_2 = 1|z, x_1, r_1 = 1)$ using model (2) when $r_1 = 1$, and $w_{2.1}^{-1} = \hat{p}(r_2 = 1|z, r_1 = 0)$ using model (3) when $r_1 = 0$. The adjusted weights w_1 and $w_{2.1}$ are used for the cross-sectional analysis of respondents in wave 1 and wave 2, respectively. For the longitudinal analysis involving responding units in both wave 1 and wave 2, $w_2 = w_1 w_{2.1}$ should be used. However, one can argue to switch the order of wave 1 and wave 2, and calculate the adjusted weights for respondents in wave 2 using a regression of r_2 on z and for respondents in wave 1 using a regression of r_1 conditioning on z, x_2 and r_2. This will result in totally different adjusted weights. The lack of uniqueness can be circumvented by turning nonattrition patterns into attrition patterns either through imputation or discarding interviews that fall outside the attrition patterns (Kalton et al., 1985; Lepkowski, 1989) .

5 Application to the GBRS Study

5.1 Description of the Data

The Galveston Bay Recovery Survey (GBRS) was conducted after Hurricane Ike struck the Galveston Bay area in Texas on September 13-14, 2008 (Tracy et al., 2011). Hurricane Ike was the third costliest hurricane to ever make landfall in the United States, causing 195 deaths and resulting in $29.6 billion in damage (Berg, 2009). The goal of the GBRS is to characterize trajectories and determinants of post-disaster mental health outcomes using a three wave panel survey. The study population consists of residents living in Galveston County and Chalmers County, who were present in the county when Hurricane Ike hit and had been living in the area for at least one month prior to the storm. The two-county area was divided into five damage geographic strata, with differing sampling rates to oversample the areas that were expected to be more affected by the storm. Eighty area segments composed of Census blocks were then selected proportional to Census 2000 number of occupied households. Six hundred and fifty-eight individuals participated in the baseline survey, with 239 from Stratum 1, 68 from Stratum 2, 123 from Stratum 3, 33 from Stratum 4, and 195 from Stratum 5. Two follow-up interviews were conducted approximately two and twelve months after

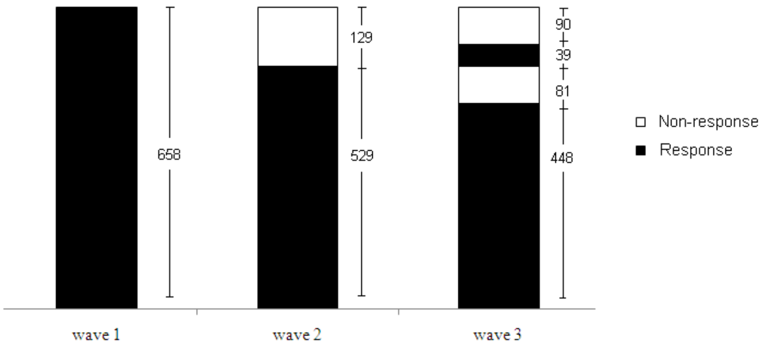

Figure 1: Patterns of panel nonresponse in the Galveston Bay Recovery Study (conditional on response to wave 1), showing the expected pattern of increasing dropouts with some wave 2 non-respondents returning in wave 3.

the baseline interview, with 529 participating in Wave 2 and 487 participating in Wave 3. Of the Wave 3 participants, 39 had not participated in Wave 2, leaving 448 of the baseline sample with data in all three waves of the panel survey. Study sample failing to respond to Wave 2 or 3 are panel nonrespondents. Figure 1 shows the patterns of panel nonresponse in the GBRS study.

5.2 Weighting Adjustments

Weighting adjustments were used to compensate for panel nonresponse in the GBRS study. As the number of observed auxiliary variables is large and many of these variables are continuous, response propensity method was used. Next, we present the step-by-step procedure to construct weighting adjustments for the 529 Wave 2 respondents.

The auxiliary variables used for the weighting adjustments include: design variables (strata, segments, sample weights), survey variables in the baseline interview, and participants' cooperation variables in the baseline interview (ever a refusal, number of calls). Since some of the survey variables were subject to item nonresponse, multiple imputation was utilized to replace the missing values. Five imputed complete data sets were generated. Variables representing the item nonresponse in the baseline interview were also created, including the number of item nonresponses and missing data indicators for each survey variable with more than 20 missing observations. This data preparation step resulted in a data set with approximate

150 auxiliary variables.

Before attempting the response propensity modeling, an initial screening analysis of the auxiliary variables was performed to reduce the large number of variables to a more manageable set. With the panel response status as the dependent variable, survey weighted logistic regression was used to examine the association between each auxiliary variable and the panel response status. With a moderate sample size (n = 658), variables having *p*-values less than or equal to 0.2 were retained for multivariable analysis. For auxiliary variables with item nonresponse, *p*-values were calculated based on the 5 imputed data sets, taking into account both within-imputation and between-imputation variations using Rubin's rules (Rubin, 1987). The screening analysis reduced the number of auxiliary variables from 150 to 31.

Stepwise selection was then used to identify important predictors of panel response status by considering the interrelationship among the selected auxiliary variables. A significance level of 0.05 was used. The stepwise selection was conducted by combing information across the 5 imputed data sets with repeated use of Rubin's rule in each step of variable selection (Wood, 2008; Chen, 2009). Interaction terms between any two significant variables were also assessed. Nine main effects and one interaction term were retained in the final response propensity model. The items retained were: AAQ total score, had spells or attacks when suddenly felt anxious, displaced from home for more than 1 week, financial loss as a result of Ike, suicide plan in lifetime, without any resource for more than 1 week, lifetime GAD severity, postdisaster emotional support, quality of life, and the interaction between had spells or attacks when suddenly felt anxious and without any resource for more than 1 week.

Three weighting adjustments approaches were compared. The first approach was the response propensity method, using the reciprocal of the predicted response probability calculated from the response propensity model. The second approach was response propensity stratification, grouping the respondents and nonrespondents using the deciles of the predicted response propensity and using the reciprocal of the weighted response rate in each adjustment cell. The third approach followed the suggestion of Little and Vartivarian (2003) but used a slightly different strategy. We first grouped the respondents and nonrespondents using the quartiles of the predicted response propensity, and then cross-classified the response propensity group with the 5 strata and the median sample weight to form new adjustment cells. We call this modified strategy as the cross-classified method. Since stratum 2-4 contains small numbers of subjects, some of the resulted cross-classified cells have few observations. As a result, adjacent small adjustment cells with similar re-

Weighting method	Mean	Std	Min	Median	Max	1+CV^2
Reciprocal of predicted response propensities	1.33	0.92	1.00	1.13	15.90	1.48
Response propensity stratification	1.32	0.55	1.02	1.10	3.39	1.17
Cross-classified method	1.24	0.22	1.00	1.17	2.11	1.03

Table 1: Distributional summaries of unit-level weights obtained by three different adjustment methods.

sponse rates were collapsed to achieve at least 20 sample units in each cell. The reciprocal of the response rate in each newly collapsed adjustment cell was used for weighting adjustments for respondents in that cell.

Panel non-response weighting adjustments are used to reduce bias in the estimation of finite population quantities. However, as a trade-off, weighting adjustments can also increase the variability in the estimation. Although we are not claiming that lower variance of weights is always better, high variance of weights often results in a loss of precision in the estimation of survey quantities. The statistic (1+CV^2) is a useful index of the loss of precision by the use of weighting adjustments, where CV is the coefficient of variation of the weighting adjustments (Kish, 1992). Table 1 shows that adjustment method using reciprocals of response propensities has the largest value of (1+CV^2), primarily because of the presence of some outlying adjustments (such as the maximum value of 15.90). The cross-classified adjustment method has the smallest value of (1+CV^2) and the smallest maximum value. Although both the adjustments using reciprocals of predicted response propensities and the response propensity stratifiction method are commonly used in practice, we chose the cross-classified weighting adjustments for the GBRS study, following Little and Vartivarian (2003).

As a final step, the baseline sample weights that incorporating unequal probability of selection and unit nonresponse were multiplied by the cross-classified weighting adjustments and post-stratified to obtain the final Wave 2 adjusted weights using raking method. The post-stratification was conducted using American Community Survey based on the stratifying variables: age, gender, marital status, race/ethnicity, whether born in the United States, education, employment status, and household income. Analysis results not shown here indicate that the estimates using the Wave 2 sample with weighting adjustments that incorporate both the panel non-response and post-stratification adjustments and the estimates using the complete Wave 1 sample are similar for the baseline characteristics.

6 Simulation Study

In panel surveys, the number of auxiliary variables that predict panel response status is often large and many of these auxiliary variables are continuous. Hence the response propensity method using logistic regression is more commonly used in the weighting adjustments than the other methods. As we discussed in section 3, Little and Vartivarian (2003) suggested to include design variables as covariates in the response propensity model. In fitting such a response propensity model, important interaction terms between the design variables and other auxiliary variables should be included in the model as well. However, the number of interaction terms can be large when many auxiliary variables are asscoaited with the panel response status. Such a model fitting can be complicated and the resulted response propensity estimates can be unstable, especially when sample size is moderate. Alternatively, we consider a modified strategy, the cross-classified method, as we used in the GBRS study. We first fit a response propensity model with the auxiliary variables as the only covariates. We then grouped sample units according to their estimated response propensities. The final weighting adjustment cells were formed by cross-classifying the response propensity strata with the design variables. We used a simulation study to compare the bias reduction and variance inflation of the cross-classified strategy with the direct use of response propensity method and the response propensity stratification method.

6.1 Design of the Simulation Study

The simulation was done using the data of the 529 respondents in the first two waves of the GBRS study. Let Y denote the hurricane Ike-related post traumatic stress disorder (PTSD) checklist score, the continuous survey outcome variable measured in Wave 2, which is only observed for Wave 2 respondents. Let X be age, the auxiliary variable collected in Wave 1, which is measured for both the Wave 2 respondents and nonrespondents. Age was standardized to have zero mean and unit standard deviation. A new stratum variable was created by combing Stratum 2-4 in the GBRS study into the new Stratum 2 and renaming Stratum 5 in the GBRS study to the new Stratum 3, so that the number of subjects in each new stratum is similar. S_1 is an indicator variable that equals 1 for subjects in Stratum 1, and 0 otherwise, and S_2 is a similar indicator for Stratum 2. The Wave 2 response indicators, R_i ($i = 1, \ldots, 529$), were generated using

Response model	α_0	α_1	α_2	α_3	α_4	α_5
X	1	1	0	0	0	0
X, S_1, S_2	1	1	1	1	0	0
X, S_1, S_2, XS_1	1	1	1	1	1	0
X, S_1, S_2, XS_2	1	1	1	1	0	1
X, S_1, S_2, XS_1, XS_2	1	1	1	1	1	1

Table 2: The coefficients for the predictors in the different response models used in our simulation study.

the following logistic regression model:

$$\text{logit}\,\Pr(R_i = 1|X_i, S_{1i}, S_{2i}) = \alpha_0 + \alpha_1 X_i + \alpha_2 S_{i1} + \alpha_3 S_{i2} + \alpha_4 X_i S_{i1} + \alpha_5 X_i S_{i2}.$$

We considered five response models with the $\boldsymbol{\alpha}$ coefficients listed in Table 2. These models result in $70\% - 80\%$ of response rates in the second wave of the sample.

Six weighting adjustments were compared: (1) M0: setting all weights to the reciprocal of the overall response rate; (2) M1: reciprocals of predicted response propensities by fitting a logistic response propensity model with X as the only covariate; (3) M2: reciprocals of predicted response propensities by fitting a logistic regression model with X, S_1, S_2, and their interaction terms as the covariates; (4) M3: response propensity stratification (5 groups) method using the predicted response propensities in M1; (5) M4: response propensity stratification (5 groups) method using the predicted response propensities in M2; and (6) M5: cross-classified method by cross-classifying the response propensity strata in M3 with the stratum indicators S_1 and S_2.

For each response model, we replicated 1000 simulations and compared the estimates of mean Y using the six weighting adjustments in terms of empirical bias and root mean squared error (RMSE). The empirical bias and RMSE are defined as,

$$\begin{aligned} \text{Bias} &= \frac{1}{1000}\sum_{t=1}^{1000}(\hat{\mu}^{(t)} - \tilde{\mu}), \\ \text{RMSE} &= \sqrt{\frac{1}{1000}\sum_{t=1}^{1000}(\hat{\mu}^{(t)} - \tilde{\mu})^2}, \end{aligned}$$

where, $\hat{\mu}^{(t)}$ is the inverse probability weighted estimate of mean Y (Basu, 1971) in the tth replicate of simulation using the Wave 2 respondents and the adjusted weights $\boldsymbol{w}$, defined as

$$\hat{\mu} = \frac{\sum_{h=1}^{3} \sum_{j=1}^{m_h} y_{hj} w_{hj}}{\sum_{h=1}^{3} \sum_{j=1}^{m_h} w_{hj}},$$

with m_h be the number of Wave 2 respondents in Strata h, and y_{hj} and w_{hj} be the value of Y and the adjusted weight for subject j in Strata h, respectively. And $\tilde{\mu}$ is the the estimate of mean Y using the 529 complete observations before generating any nonresponse, defined as

$$\tilde{\mu} = \frac{\sum_{h=1}^{3} \sum_{j=1}^{n_h} y_{hj}}{529},$$

with n_h be the total number of Wave 2 respondents and nonrespondents combined in Strata h. In the simulation, for simplicity, we ignored the sample weights in the GBRS study. For both bias and RMSE, a smaller absolute value means better weighting adjustments.

6.2 Simulation Results

Table 3 shows the empirical bias and RMSE of the estimates of mean Y using the six weighting adjustments. In the response model with X as the only covariate, the empirical bias is close to 0 in both M1 (the correct model) and M2 (the over-fitted model), and the stratification in M3 and M4 slightly increases the bias but reduces the RMSE. The proposed cross-classified method M5 yields similar bias and RMSE to the other approaches. In the response model with (X, S_1, S_2) as the covariates, both M1 and M3 have large empirical bias and RMSE because of the misspecification in the response propensity model, while M2 and M4 have small bias and RMSE even though their response propensity model is over-fitted. This suggests that a larger model is more favorable than a smaller model in estimating response propensities. If important predictors for nonresponse are omitted, the resulted weighted estimate will be subjected to large bias. On the contrary, a model with nuisance variables leads to a good weighted estimate as long as important variables are included and the nuisance variables do not add much noise in the model prediction. The cross-classified M5 estimate has small bias and RMSE similar to M2.

The simulations in the response models with interaction terms between X and S_1 or S_2 again show that M1 and M3 have larger bias than M2, M4 and M5, due to the model misspecification. M2 has the smallest

Response model		M0	M1	M2	M3	M4	M5
(X)	bias	32.0	-0.4	-0.1	-1.6	0.7	-1.4
	RMSE	42.4	34.4	34.4	33.9	33.6	34.2
(X, S_1, S_2)	bias	35.2	14.7	-0.1	13.7	3.0	-0.1
	RMSE	40.5	27.8	23.0	27.0	23.0	23.2
(X, S_1, S_2, XS_1)	bias	36.4	4.7	0.5	3.3	-3.0	1.8
	RMSE	41.7	27.2	31.5	26.1	26.6	30.9
(X, S_1, S_2, XS_2)	bias	43.5	16.6	-1.0	15.9	2.5	-3.9
	RMSE	48.2	32.3	29.5	31.3	26.7	27.3
$(X, S_1, S_2, XS_1, XS_2)$	bias	45.0	3.6	-0.4	2.9	-1.8	-2.1
	RMSE	49.7	33.5	35.8	31.6	31.5	33.2

Table 3: Empirical bias and RMSE of the estimates of mean Y using the six weighting adjustments. The cross-classified method (M5) has similar bias and RMSE to the other estimators when strata are not related to nonresponse, but has smaller bias than both the response propensity method (M1) and the response propensity stratification method (M3) when strata are related to nonresponse and the response propensity model is misspecified.

bias, and the stratification in M4 reduces RMSE at a cost of larger bias. The stratification in M3 reduces both the bias and RMSE, but the improvement is limited. This again implies the importance in including design variables as covariates in the response propensity regression as suggested by Little and Vartivarian (2003). The cross-classified M5 estimate is comparable in the bias to M2 and M4 but has smaller RMSE than M2. The over-smoothed estimate M0 is very biased and has high variability in all five response model scenarios.

7 Conclusion

Weighting is widely used to compensate for panel nonresponse in panel surveys. Weighted analyses can be difficult (for example, in constructing standard errors for weighted regressions) but the weights themselves are relatively easy to construct, compared to the complexity of implementing imputation models that account for both cross-sectional and longitudinal dependence between all the variables. However, a careful assessment of the variation of the adjusted weights is needed in order to avoid a serious loss of precision in the survey estimates.

Our review highlights two choices with weighting adjustments. The first choice is the use of auxiliary variables in the adjustment. Compared to item nonresponse in cross-sectional surveys, a large number of auxiliary variables is available for panel nonrespondents. The auxiliary variables used in the adjustment should be predictive of the panel response status and can be responses to the questions in the previous waves of the panel or variables that measure previous wave data quality, such as call history variables and the amount of item nonresponse. When the prior knowledge is not available, a logistic regression or regression tree type analysis can be conducted to identify important auxiliary variables that are predictive of the panel response status.

The second choice is the application of adjustment methods. Adjustment cells are popular when the amount of auxiliary variables that are predictive of the panel nonresponse is small. This method incorporates the interaction effects of the auxiliary variables and avoids strong modeling assumptions in the relations between the auxiliary variables and the response status. However, the direct adjustment cell method only allows a small number of auxiliary variables. On the other hand, the response propensity method is advantageous in allowing continuous variables and the inclusion of a large set of auxiliary variables, but the effect of weighting adjustments in reducing nonresponse bias largely relies on correct specification of the response propensity regression. Moreover, a logistic regression model with many covariates can result in highly variable response propensities and thus adjusted weights. An alternative, the response propensity stratification method, forms adjustment cells based on the estimated response propensities. This approach allows the inclusion of a large amount of auxiliary variables and is less sensitive to the misspecification in the response propensity model. Our simulation shows that the response propensity stratification results in estimates with smaller variability but larger bias than the estimates weighted by the reciprocals of response propensities. Finally, the generalized raking method works well when there is no interaction effects between the auxiliary variables, but convergence can be slow or impossible especially when a large number of auxiliary variables is involved and each variable has multiple categories.

Many panel surveys use complex sampling designs in the first wave of sampling. There is no consensus as to the best way to incorporate complex survey design variables into the weighting adjustments for panel nonresponse. We proposed a cross-classified method that first creates response propensity strata by grouping respondents and nonrespondents with similar estimated response propensities and then cross-classifies the response propensity strata with design variables, such as stratum, PSU, and sampling weights, to form weighting adjustment cells. The simulation supports the conclusion of Little and Vartivarian (2003) that design variables should be included as covariates in the response propensity regression if design variables are related to nonresponse. It also shows that when design variables are not predictive of the response status, the cross-classified weighting adjustment method yields survey estimates with bias and efficiency similar to the estimates weighted by the reciprocals of response propensities or by the response propensity stratification method. However, when design variables are important predictors of the response status, the cross-classified method yields survey estimates with smaller bias than the response propensity method and the response propensity stratification method if the design variables are not included as covariates in the response propensity model, and yields survey estimates with comparable bias to the other two methods if the response propensity model is correctly specified. We conclude that when many auxiliary variables and their interaction terms with design variables are related to nonresponse, the cross-classified method is a good way to create weighting adjustments, because it is simple to implement, avoids the complication in modeling all the interactions, and is robust to the response model misspecification.

All potential important predictors of nonresponse should be included as covariates in the response propensity model, including responses to the questions in the previous waves of the panels, data quality variables, design variables and their interaction terms. Omitting important nonresponse predictors in the response propensity model can result in serious bias and high variability in the survey estimates. Although our limited simulation does not show any adverse effect of including nuisance variables in the response propensity regression, the inclusion of many nuisance variables might increase the variability in the weighting adjustments and thus survey estimates. Caution is needed in selecting covariates for response propensity models. The ultimate goal of weighting is to construct an adjusted dataset that matches the population as closely as possible. As larger variability in the adjusted weights usually implies the larger variability in the resulted survey estimates, the statistic $(1 + \mathrm{CV}^2)$ is often used as an index to measure the loss in efficiency due to the weighting adjustment. Although a method with smaller loss in efficiency is favorable, small-variance-of-weights is not always better as evident by the over-smoothed weighting adjustments in the simulation.

Acknowledgements

This research was partially supported by the National Center for Disaster Mental Health Research (NIMH Grant 5 P60 MH082598), Fran H. Norris, Center Director, Sandro Galea, Research Director and by the

Institute of Education Sciences (Grant R305D090006-09A).

References

Chapman, D.W., Bailey, L., and Kasprzyk, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census. *Survey Methodology*, **12**, 161-180.

Cohen, S.B., DiGaetano, R., and Goksel, H. (1999). Estimation procedures in the 1996 Medical Expenditure Panel Survey household component. *Agency for Health Care Policy and Research, MEPS Methodology Report No. 5*, AHCPR Publication No. 99-0027.

David, M., Little, R. J. A., Samuhel, M. E., and Triest, R. K. (1983). Nonrandom nonresponse models based on the propensity to respond, *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 168-173.

Deville, J.-C., and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Deville, J.-C., Särndal, C.E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, **88**, 1013-1020.

Duncan, G., and Kalton, G. (1987). Issue of Design and Analysis of Survey Across Time. *International Statistics Review*, **51**, 97-117.

da Silva, D.N. and Opsomer J.D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology*, **35**, 165-176.

Giommi, A. (1984). A simple method for estimating individual response probabilities in sampling from finite populations. *Metron*, **42**, 185-200.

Grau, E., Potter, F., Williams, S., and Diaz-Tena, N. (2006). Nonresponse adjustment using logistic regression: to weight or not to weight? *Proceedings of the Survey Research Methods Section, American Statistical Association*, 3073-3080.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, **2**, 303-314.

Kalton, G. and Brick, M. (2000). Weighting in household panel surveys. In Advid, R. (Ed.) *Researching Social and Economic Change. The Uses of Household Panel Studies*. London/New York: Routledge, 96-112.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, **12**, 1-16.

Kalton, G., Lepkowski, J., Montanari, G.E., and Maligalig, D. (1990). Characteristics of second wave non-respondents in a panel survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 462-467.

Kalton, G., Lepkowski, J., and Lin, T.-K. (1985). Compensating for wave nonresponse in the 1979 ISDP research panel. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 372-377.

Kalton, G., and Miller, M.E. (1986). Effects of adjustments for wave nonresponse on panel survey estimates. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 194-199.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, **29**, 119-127.

Kish, L. (1992). Weighting for Unequal Pi. *Journal of Official Statistics*, **8**, 183200.

Lepkowski, J.M. (1989). The treatment of wave nonresponse in panel surveys. *In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P., eds. Panel Surveys*. New York: John Wiley.

Lepkowski, J.M., Kalton, G., and Kasprzyk, D. (1989). Weigting adjustments for patial nonresponse in the 1984 SIPP panel. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 296-301.

Loosveldt, G., Pickery, J., and Billiet, J. (2002). Item nonresponse as a predictor of unit nonresponse in a panel survey. *Journal of Official Statistics*, **18**, 545-557.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, **54**, 139-157.

Little, R.J.A. and David, M. (1983). Weighting adjustments fir non-response in panel surveys. *Bureau of the Census working paper.*

Little, J.A., and Rubin, D.B. (2002). *Statistical analysis with missing data, 2nd edition*. Wiley.

Little, R.J.A. and Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, **22**, 1589-1599.

Little, R.J.A. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, **31**, 161-168.

Lynn, P. (Ed.) (2003). *Quality Profile: British Household Panel Survey Waves 1 to 10: 1991-2000*. Colchester: Institute for Social and Economic Research.

Meekins, B.J. and Sangster, R.L. (2004). Predicting wave nonresponse from prior wave data quality. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 4015-4021.

Potter, F.A. (1990). Study of Procedures to Identify and Trim Extreme Sample Weights. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 225230.

Rizzo, L., Kalton, G., and Brick, M. (1994). Adjusting for panel nonresponse in the Survey of Income and Program Participation. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 422-427.

Rizzo, L., Kalton, G., and Brick, M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, **22**, 43-53.

Rosenbaum, P. R., and Rubin, D. B., (1983), The central role of the propensity score in observational studies for causal effects," *Biometrika*, **70**, 4155.

Sommers, J., Riesz, S., and Kashihara, D. (2004). Response propensity weighting for the Medical Expenditure Panel Survey - Insurance Component (MEPS-IC). *Proceedings of the Survey Research Methods Section, American Statistical Association*, 4410-4417.

Wun, L.-M., Ezzati-Rice, T.M., Diaz-Tena, N., and Greenblatt, J. (2007). On modeling response propensity for dwelling unit (DU) level non-response adjustment in the Medical Expenditure Panel Survey (MEPS). *Statistics in Medicine*, **26**, 1875-1884.

Tracy, M., Norris, F.H., and Galea, S. (2011). Differences in the determinants of posttraumatic stress disorder and depression after a mass traumatic event. *Depression and Anxiety*, **28**, 666-675.

Berg, R (2009). Hurricane Ike Tropical Cyclone Report. *National Hurricane Center*, 01-23.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Wood, A.M., White, I.R., Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, **27**, 3227-3246.

Chen, Q. (2009). Bayesian Predictive Inference for Three Topics in Survey Samples. PhD thesis. University of Michigan, Ann Arbor.

Meng, X. L., Chen, C., Duan, N., and Alegria, M. (2010). Power-shrinkage: An alternative method for dealing with excessive weights. Presentation at Joint Statistical Meetings. `http://andrewgelman.com/movabletype/mlm/meng_JSM_presentation_20090802_8am.pdf`.

Basu, D. (1971). An essay on the logical foundation of survey sampling. Part 1, in *Foundation of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.