Propensity Models versus Weighting Class Approaches to Nonresponse Adjustment: A Methodological Comparison

Peter Siegel, James R. Chromy, and Elizabeth Copello
RTI International, 3040 Cornwallis Road, PO Box 12194, Research Triangle Park, NC 27709

Abstract

Statistical adjustment of nonresponse is a deep and pervasive issue for sample surveys. Contemporary statistical methods offer two broad classes of approach to nonresponse adjustment. One is the use of a traditional weighting cell approach. More recently, response propensity modeling, using, typically, logistic regression, has been developed as a further approach to nonresponse adjustment. Additionally, RTI's General Exponential Model (GEM) generalizes weight adjustments and includes poststratification and weight trimming. Data from the Education Longitudinal Study of 2002 (ELS:2002) are used to compare the results of the weighting class method, raking, a logistic regression propensity model, and GEM when used for nonresponse adjustment. For the one-dimensional case where each unit is in one unique cell, it can be shown that the four methods will produce similar results; the logistic propensity model approach produces slightly different results since it does not require calibration to weighted control totals for the selected sample. Expanding to many variables and multiple dimensions, marginal totals, variances, and weight distributions are compared for raking, logistic response propensity, and two GEM special cases. The weighting class method is limited to the cell model.

Keywords: Nonresponse Adjustment, Generalized Exponential Model (GEM), Weighting Class, Propensity Modeling, Raking

1. Introduction

The focus of this paper is on an empirical investigation and comparison of methods rather than on development of new methods. The motivation for the investigation arose from concerns about comparability of statistical results when different nonresponse adjustment methods are introduced in panel surveys or in repeated surveys in general. The examples chosen to compare some alternative nonresponse adjustment methods are simplified by reducing the number of dimensions considered and by limiting the investigation to adjusting for person level nonresponse.

The use of weights in sample surveys is a generally accepted practice. The foundations for weighting sample data by the inverse of the sample selection probabilities is presented by Horvitz and Thompson (1952) for probability proportional to size without replacement sampling. The use of design-based weights as specified by Horvitz and Thompson is unambiguous and replicable by other survey statisticians with knowledge of the selection probabilities. The final survey weights actually used for most surveys involve not only the design-based weight prescribed by Horvitz and Thompson, but also additional factors to account for nonresponse, poststratification, and, in some cases, to limit the impact of extreme weights. The methods used to develop these additional factors depend on the availability of specific auxiliary data at the unit level or at the summary level and may be developed, in good faith, with slightly different results obtained by different practitioners.

2. Methods Studied

Four weight adjustment approaches were tested and compared as they apply to nonresponse adjustment:

(1) Weighting class adjustments are made by partitioning the sample into mutually exclusive groups called weighting classes and adjusting the sample weights in each group by a single adjustment factor so that the sum of the weights of respondents equals the sum of the weights of respondents and nonrespondents.

(2) Raking is an iterative procedure where weighting class type adjustments are first performed in one dimension and then in another until convergence is reached. The method can be extended to two or more dimensions and is sometimes called iterative proportional fitting (IPF). Raking controls at the margin level for each dimension.

(3) Response propensity modeling uses logistic regression and auxiliary data which are available for both respondents and nonrespondents to predict the response propensity of each sample member. The inverse of the respondent's predicted response propensity is the weight adjustment.

(4) The Generalized Exponential Model (GEM) developed by Folsom and Singh (2000) is a unified approach to nonresponse adjustment, poststratification, and extreme weight reduction. It is based on a generalization of Deville and Särndal's logit model

(Deville and Särndal, 1992). The GEM approach controls at the margins, and adjustment factors can be constrained individually.

Weighting class, raking, and GEM methods can be applied to poststratification as well as nonresponse adjustment. In poststratification, control totals are obtained from external sources believed to be the truth or at least much more precise than those based on the current survey sample. Control totals for nonresponse adjustment are generated from the selected sample. The logistic regression modeling approach analyzes the selected sample and uses response as the dependent variable. Logistic regression does not naturally extend to poststratification.

Weighting class methods are the simplest to implement and to explain. Adjustments are either based on a single dimension or are performed at the cell level (fully interacted model) for multi-way table controls. When alternative methods are applied at the fully interacted model level, they reduce to a weighting class approach as is shown in the following sections.

Raking or iterative proportional fitting is designed to control marginal distributions only and continues until the cell level adjustments stabilize (Oh and Scheuren 1983). If applied in a single dimension (or at the cell level), it reduces to the weighting class method.

Logistic regression or response propensity methods fit a logistic regression model to the selected sample in order to predict the probability of responding. Variables used as predictors in the logistic regression must be known for all members of the selected sample (both respondents and nonrespondents). Although the predictor variables can be continuous or categorical, for comparison with other methods, this research considered only categorical predictors.

Deville and Saarndal (1992) proposed the following weight adjustment factor which allows setting bounds on the adjustment lower and upper bounds:

$$a_k(\lambda) = \frac{l(u-1) + u(1-l)e^{Ax_k'\lambda}}{(u-1) + (1-l)e^{Ax_k'\lambda}}$$

where $l < 1 < u$ and $A = (u-l)/[(u-1)(1-l)]$. The parameters, *u* and *l,* are user-specified bounds on the adjustment factors. The column vector, λ, represents the model parameters corresponding to the covariate vector, *x*. The model parameters are obtained for poststratification by requiring that

$$\sum_{respondents} x_k d_k a_k(\lambda) = T_x$$

where T_x is a vector of poststratification totals

Two special cases are used in this report. The first was identified in the Deville-Särndal paper. As $l \to 0$ and $u \to \infty$, $a_k(\lambda) \to e^{x_k'\lambda}$. This solution corresponds to an exponential model and in the limit yields the same results as the raking method.

Folsom and Singh's GEM generalized the Deville-Saarndal calibration method by allowing unit-specific bounds on the adjustment factors and by adding a centering factor, c_k, between l_k and u_k, which need not be 1.

$$a_k(\lambda) = \frac{l_k(u_k - c_k) + u_k(c_k - l_k)e^{A_k x_k'\lambda}}{(u_k - c_k) + (c_k - l_k)e^{A_k x_k'\lambda}}$$

with $A_k = (u_k - l_k)/[(u_k - c_k)(c_k - l_k)]$. This model can be applied to either poststratification or nonresponse adjustment. For nonresponse adjustment, model parameters are obtained by solving

$$\sum_{respondents} x_k d_k a_k(\lambda) = \widetilde{T}_x$$

where $\widetilde{T}_x$ is a vector of sums based on the selected sample (using the design weights before adjustment). The second special case presented in this report is based on the GEM model when allowing $l_k = 1$, $c_k = 2$, and $u_k \to \infty$, then $a_k(\lambda) \to 1 + e^{x_k\lambda}$; i.e., the GEM solution approaches the solution obtained by fitting a logistic regression model. .

Results from both special cases of the GEM model are presented below and compared with results from other nonresponse adjustment approaches.

3. Other Comparative Studies

Two empirical studies completed in 1994 used panel data from the Survey of Income and Program Participation (SIPP). SIPP was using a weighting class approach for nonresponse adjustment, and Folsom and Witt (1994) compared it to inverse response propensity weighting via generalized raking. They had mixed results and were not able to show any superiority for the response propensity approach over the weighting class approach. Rizzo, Kalton, Brick, and Petroni (1994) compared SIPP's weighting class approach with six alternative weighting schemes and concluded that the different methods produced similar estimates, the weights from the different methods were highly correlated with each other, and the variability of the weights was similar for all the weighting schemes.

Also, Kalton and Flores-Cervantes (2003) compared eight weighting techniques: cell weighting, raking, linear weighting, GREG weighting, logistic regression weighting, a mixture of cell weighting and another method, logit weighting, and truncated linear weighting. Each adjustment method was briefly described, and its application was illustrated with a simple example. The results were generally compared across methods. They noted that "the choice of auxiliary variables and of the mode in which they are employed in the adjustments may be of more significance than the choice of a particular method."

4. The Empirical Study

One example of nonresponse can be seen in the base year of ELS:2002. Nonresponse occurred both at the school and the student levels. For purposes of comparing nonresponse adjustment methods, the comparative study was limited to student response among students attending public schools. For this population, a response rate of 87 percent (12,039 respondents from a sample of 13,882 selected students) was achieved. No trimming of extreme weights is done in the initial comparisons of the methods.

Five sets of variables were used to compare the four methods. Each of these five sets is described in the subsections below. For each of the four methods, the mean, minimum, median, and maximum adjustment factor and weight after adjustment were examined, as well as the unequal weighting effect (UWE). The relative root mean squared differences (RRMSD) between methods were also computed as:

$$\text{RRMSD} = \frac{\sqrt{\sum_{n} \frac{(X_i - Y_i)^2}{n}}}{\overline{X}}$$

where X_i is the nonresponse adjusted weight for student i using one adjustment method, Y_i is the nonresponse adjusted weight for student i using a second adjustment method, $\overline{X}$ is the mean weight[1], and *n* is the number of responding students on the file.

In the one variable case, weighting class and raking are operationally identical since no iteration is required. When using two or more variables, the weighting class method can only be applied at the fully interacted or cell model. Most of our interest in considering two or more variables is on controlling to marginal totals for each variable. As noted above, GEM can be run to either be similar to the raking approach (GEM Case 1) or to the logistic propensity model approach (GEM Case 2). These two particular cases were studied in the comparative analysis and are summarized in terms of special limiting and centering factors identified as:

GEM Case 1:

$$l \to 0, u \to \infty, c = 1, a_k(\lambda) \to e^{x_k'\lambda},$$

and GEM Case 2:

$$l \to 1, u \to \infty, c = 2, a_k(\lambda) \to 1 + e^{x_k'\lambda}.$$

Although, the GEM approach is very general in allowing the limiting parameters and the centering parameter to be set at other values and made specific to particular respondents, only these two cases were examined in the comparative analysis[2].

4.1 One Variable

The one variable model was first tested using the variable sex (male and female). A common feature of the weighting class, raking, and GEM methods is that each one calibrates the weights of the respondents to sum to the total of the weights before adjustments for the selected sample for each control total. When applied to a cell or a one-dimensional model, exactly one adjustment factor is applied to each cell by all three of these methods and it must be identical to achieve the calibration property. The logistic response propensity method does not control adjusted weight sums to correspond precisely to initial weights sums even though it comes close. This lack of precise control by the logistic response propensity method is shown in small differences in mean weights as well as in the RRMSDs and the unequal weighting effects.

Table 1. Method Comparsions for One Variable

Method	RRMSD			
	Raking (IPF)	GEM Case 1	Logistic RP	GEM Case 2
Raking (IPF)				
GEM Case 1	0.000000			
Logistic RP	0.000517	0.000517		
GEM Case 2	0.000000	0.000000	0.000517	
UWE	1.5807	1.5807	1.5807	1.5807
Mean weight	263.87	263.87	263.98	263.87

Table 1 verifies that for one variable, raking, GEM case 1, and GEM case 2 produce identical results. The logistic response propensity model produces results that differ from the other three as shown by the pairwise RRMSD's of 0.000517 when the logistic

[1] The mean weight was computed assuming a calibration method which forces the weight sums to equal the total weight sum overall.

[2] Numerically, it is not possible to specify infinity as an upper limit. A value of 10^8 was used for these comparative studies. In practice, a value of about 3.0 is often used for an upper limit and yields about the same results unless unusually extreme weight adjustments are required based on highly variable or very low response rates.

response propensity method is one member of the pair and the mean weight calculations of 263.98 for logistic response propensity and 268.87 for the three other methods. Unequal weighting effects (UWEs) are comparable.

4.2 Two Variables

The variables sex (male and female) and race/ethnicity (Hispanic, Asian, Black, and White/Other) were used for the adjustment using two variables. Table 2 shows the RRMSDs to six decimal places; GEM case 1 shows no difference with raking (RRMSD=0.000000) and GEM case 2 shows only small differences with the logistic response propensity method (RRMSD=0.000067). Larger differences as measured by the RRMSD (RRMSD > 0.001600) are noted for comparisons of the first two methods with the last two methods.

Table 2. Method Comparisons for Two Variables Controlled at the Margins

Method	RRMSD			
	Raking (IPF)	GEM Case 1	Logistic RP	GEM Case 2
Raking (IPF)				
GEM Case 1	0.000000			
Logistic RP	0.001623	0.001623		
GEM Case 2	0.001619	0.001619	0.000067	
UWE	1.5695	1.5695	1.5692	1.5692
Mean weight	263.8699	263.8699	263.8740	263.8699

Table 3 shows what happens when the two variables, sex and race/ethnicity, are treated as an eight-cell model. This table can be interpreted the same way as Table 1, although the differences among methods as measured by the RRMSD are smaller in this eight-cell case.

Table 3. Method Comparisons for Two Variables Controlled at the Cell Level

Method	RRMSD			
	Raking (IPF)	GEM Case 1	Logistic RP	GEM Case 2
Raking (IPF)				
GEM Case 1	0.000000			
Logistic RP	0.000069	0.000069		
GEM Case 2	0.000000	0.000000	0.000069	
UWE	1.56961	1.56961	1.56958	1.56961
Mean weight	263.8699	263.8699	263.8753	263.8699

4.3 Four Variables

In addition to sex and race/ethnicity, control variables were added for region (Northeast, Midwest, South, and West) and metropolitan status (urban, suburban, and rural). Table 4 shows the results of method comparisons when all four variables are controlled to the marginal totals. As was the case with the two variable model, the differences between raking and GEM case 1 (RRMSD=0.0051) and logistic response propensity and GEM case 2 (RMSD=0.0015) are smaller than the differences comparing the first two methods with the last two methods (RRMSD>0.0086). As before, the mean weight for the logistic response propensity method is slightly different from the mean weights for the three calibration methods. Unequal weighting effects were reasonably close for all four methods.

Table 4. Method Comparisons for Four Variables Controlled at the Margins

Method	RRMSD			
	Raking (IPF)	GEM Case 1	Logistic RP	GEM Case 2
Raking (IPF)				
GEM Case 1	0.005106			
Logistic RP	0.008619	0.010184		
GEM Case 2	0.008862	0.010337	0.001455	
UWE	1.5953	1.5971	1.5944	1.5956
Mean weight	263.8699	263.8699	263.8793	263.8699

With four variables, a completely interacted cell model would require control for 96 cells. Even with the large samples available from this population, 96 cells cannot be formed and some additional combining of cells would be required. This is a common procedure when using the weighting class approach. The cells would be combined to ensure that each cell has a minimum number of respondents and that no extreme adjustment factors would result. From examining the two-variable case, we can be fairly certain that similar comparative results would occur with a larger number of cell controls used. Since the combining of cells is external to any of the methods being compared, only marginal control models were evaluated for cases with more than two control variables.

4.4 Six Variables

Next, larger models were explored to show how the various methods handle more complex weight adjustments. To choose a larger number of variables, a pool of 23 variables known for both respondents and nonrespondents was included in GEM. Then the six statistically significant variables were kept in the model, and the remaining non-significant variables were dropped from the model. These six variables were sex (male and female), region (Northeast, Midwest, South, and West), number of part-time teachers (0-1; 2-3; 4-6; >6), percentage of students with an IEP (<6; 6-10; 11-15; >15), school level (K-12, PreK-10, 1-12, PreK/1-9/12, PreK-12; middle grades but no elementary; only high school), and 10th-grade enrolment (0-99; 100-249; 250-499, >499).

Table 5 shows the method comparisons for six variables. With more variables, the RRMSDs get larger, but the pattern of relative sizes remains the same with larger RRMSDs when comparing the first two methods (raking and GEM case 1) to the last two methods (logistic response propensity and GEM case 2). The mean weight for the logistic response propensity method is different from the other three. Unequal weighting effects are reasonably close.

Table 5. Method Comparisons for Six Variables Controlled at the Margins

Method	RRMSD			
	Raking (IPF)	GEM Case 1	Logistic RP	GEM Case 2
Raking (IPF)				
GEM Case 1	0.008217			
Logistic RP	0.021364	0.023303		
GEM Case 2	0.022165	0.024069	0.003165	
UWE	1.5952	1.5961	1.6020	1.6025
Mean weight	263.8699	263.8699	263.8395	263.8699

4.5 Eight Variables

As an alternative method for choosing a larger number of variables for a more complex nonresponse adjustment, all 23 variables known for both respondents and nonrespondents were included in a Chi-Squared Automatic Interaction Detection (CHAID), which is a tree analysis. With response as the dependent model variable, eight significant variables were identified and included in each nonresponse adjustment method. The eight variables selected were metropolitan status (three levels), region (four levels), number of full-time teachers (four levels), percentage of full-time teachers certified (three levels), number of part-time teachers (four levels), percentage of students with an IEP (four levels), total enrollment (four levels), and number of class periods (four levels).

Table 6. Method Comparisons for Eight Variables Controlled at the Margins

Method	RRMSD			
	Raking (IPF)	GEM Case 1	Logistic RP	GEM Case 2
Raking (IPF)				
GEM Case 1	0.017508			
Logistic RP	0.020650	0.025503		
GEM Case 2	0.022043	0.027410	0.004712	
UWE	1.6135	1.6120	1.6120	1.6138
Mean weight	263.8699	263.8699	263.8073	263.8699

Table 6 compares methods when the specified eight variables are controlled at the margins. As more variables are added, the computational requirements for obtaining convergence to control totals or for fitting a logistic response propensity model all increase dramatically. The RRMSDs increase or remain high. The pattern of higher values for comparing the first two methods against the last two methods vs. comparing within the two groups remains, but is less pronounced.

5. Control of Unequal Weighting Effects (UWEs)

For the purpose of comparing basic methods, no attempt was made to control unequal weighting effects under any of the methods. As noted above, the differences in UWEs across methods for the same set of control variables were small. In addition, the UWEs increased only moderately when the number of control variables increased. *Ad hoc* methods to control for extreme weights can be applied with any of the methods studied. The GEM methods allow for putting limits on the adjustment factors as part of the specification process even though the options studied used the loosest possible limits. In addition, GEM provides a general method for trimming extreme weights whether they arise from the adjustment process or from the design-based weight structure (see Folsom and Singh 2000).

6. Additional Comments and Conclusions

Weighting class methods are the simplest to apply and work well for small samples or when only a few auxiliary variables may be available. When adding control variables and crossing their categories to form a cell model, empty cells occur frequently even with large samples. This leads to combining of cells and as a result, the exact control for some marginal totals is lost.

Because of the need for combining cells for weighting class approaches, the main focus of this paper has been on comparing methods that achieve controls for marginal totals of each categorical auxiliary variable. Some combining of auxiliary variable categories can also occur when using marginal control models, but more variables can be used simultaneously with known control over marginal totals.

As expected, GEM special case 1 and raking produced nearly identical results. GEM special case 2 and the logistic response propensity produced similar, but not identical, results. Since the logistic response propensity only approximately controls marginal totals and both GEM cases and raking control them explicitly, the logistic response propensity results usually differed from the other three, even in the value of the mean weight.

References

Deville, J.C., and Särndal, C-E. (1992), “Calibration Estimating in Survey Sampling,” *Journal of the American Statistical Association*, *87*: 376–382.

Folsom, R.E., and Singh, A.C. (2000), “The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratification,” *Proceedings of the Section on Survey Research Methods* (pp. 598–603).

Folsom, R.E., and Witt, M.B. (1994), “Testing a New Attrition Nonresponse Adjustment Method for SIPP,” *Proceedings of the Section on Survey Research Methods* (pp. 428–433).

Kalton, G., and Flores-Cervantes, I. (2003), “Weighting Methods,” *Journal of Official Statistics*, 19: 81-97.

Oh, H. L., and Scheuren, F.S. (1983), “Weighting Adjustment for Sampling and Nonresponse Error Reduction,” in *Incomplete Data in Sample Surveys*, W.G. Madow, I. Olkin, and D.B. Rubin (eds). New York: Academic Press.

Rizzo, L., Kalton, G., Brick, M., and Petroni, R. (1994), "Adjusting for Panel Nonresponse in the Survey of Income and Program Participation," *Proceedings of the Section on Survey Research Methods*, 422-427.