A Comparison of Weighting Adjustment Methods for Nonresponse

Ronaldo Iachan[1], R. Lee Harding[1], Kurt Peters[2]
[1]ICF International, 530 Gaither Road Suite 500, Rockville, MD 20850
[2]ICF International, 126 College Street Suite 2 Burlington, VT 05401

Abstract
The adjustment of survey data for nonresponse is typically based on weighting class methods. Weighting classes are formed using a set of core variables that are correlated with response behavior and with survey outcomes. The underlying principle is that respondents are more alike within a class than across classes. The choice of variables may be made using response propensity models (i.e., logistic regression models for the response indicator), or with a range of recursive partitioning, tree-based classification methods (e.g., CHAID or CART). Propensity models may also be used more directly to generate propensity scores that are applied to adjust for response probabilities. This research compares these methods using real and simulated data with origins in two kinds of multistage stratified sample surveys: samples of students within schools, and samples of patients within facilities. Comparisons are made along bias, variance and mean squared error of key survey estimates.

1. Introduction

Nonresponse adjustments are applied to sample weights to account for bias unit nonresponse. The idea is to reduce the bias due to unit nonresponse by increasing the weights of responding units with similar characteristics. Nonresponse adjustments come with a trade-off. While the bias of the estimates should be reduced, the variance of the estimates will increase due to the added variation in the weights. Using two studies, we present three methods for adjusting for unit nonresponse. We will look at the additional variance resulting from the nonresponse adjustment and using simulations we will look at the bias of the estimates.

1.1 Overview of the Studies
The first study was a multistage sample of patients within sampled medical facilities across five regions. The first stage of sampling was of facilities. Facilities were sample proportional to the number of patients seen at the facility during a four month period. Patients were sampled such that the combined facility probability of selection and patient probability of selection was equal for all patients.

The National Youth in Transition Database (NYTD) collects data from a national cohort of youth receiving independent living services paid for or provided by each state agency that administers the John H. Chafee Foster Care Independence Program. In the baseline year (2011), each state conducted a census of the 17-year-old cohort, with the Adoption and Foster Care Analysis System (AFCARS) database serving as the frame. Follow-ups are conducted with prior-wave respondents on a biennial basis, with some states choosing to conduct a census of respondents during each follow-up wave and others selecting a sample to survey.

Wave 1, conducted in 2011, serves as the baseline year in which states conducted a census of all 17-year-old youths in the eligible population. There were 15,597 Wave 1 respondents out of 29,105 surveyed. Weighted estimates using the 2011 dataset provide estimates of the 17-year-old NYT population in 2011.

1.2 Overview of the Adjustments

The following section describes approaches for nonresponse adjustment. The first method is a traditional cell weighting adjustment in which weighting classes are created to distribute the weight of nonrespondents over respondents. The weighting classes were defined using crossing variables to form cells. For example if you had two dichotomous variables then the crossing of the variables would create four weighting class cells. The second and third nonresponse adjustment methods rely on a propensity score adjustment using a logistic regression model to estimate response propensities for all sampled cases. For the second method the propensity scores are used to create deciles and the weights of the respondents within the deciles are adjusted to account for the nonrespondents within the deciles. The third method uses the propensity scores directly to adjust the weights. Finally, the fourth and fifth nonresponse adjustment methods use an alternative propensity score adjustment. They use a random forest algorithm to estimate response propensities, with nonresponse adjustments again being applied either directly or by deciles. The first, second and third methods were applied to the patient study. All of the nonresponse adjustment methods were used for the NYTD study.

2. Comparison of Nonresponse Adjustment for Both Studies

2.1 Patient Sample

To obtain information on both the nonresponding and responding patients, the patients were matched to a dataset with demographic variables. The patient sample had a limited number of variables on both respondents and nonrespondents. Missing demographic variable were imputed.

The first method used bivariate analysis (chi-square test) to determine which of the variables were associated with response in each region and overall. The key variables related to nonresponse in the bivariate analysis are presented in Table1.

Table 1: Key Variables Associated with Nonresponse

Regions	Variables
Region 1	Race and Age group
Region 2	Facility Type and Race
Region 3	Facility Size and Age group
Region 4	Facility Size
Region 5	Race
Overall	Facility Size, Race and Gender

A logistic regression model was then fit using the significant variables from the bivariate analysis. The two most significant variables where crossed to create the weighting adjustment cells. The adjustment cells for the overall analysis would have been created facility size and race (African American verses all others).

The second and third nonresponse adjustment methods used the same logistic regression model from the first. The second method used the propensity deciles from the model to create weighting classes and the third method used the predicted propensities to adjust the weights.

To quantify variability of different adjustment methods, CVs of the weights were computed in each region and overall. The CVs are presented in table 2.

Table 2: Coefficients of Variation of the Nonresponse Adjusted Weights

Regions	Sample Size	Base Weight	Method 1	Method 2	Method 3
Region 1	745	38.3	62.7	46.3	47.2
Region 2	394	0.84	31.7	34.1	34.1
Region 3	397	0.15	23.5	16.1	19.7
Region 4	698	29.0	24.0	24.0	24.7
Region 5	403	5.7	5.7	5.8	12.4
Overall	2,637	54.5	61.9	64.6	64.1

When comparing the CVs across all methods and all regions the results are not completely clear. We expected to see the CVs increasing from nonresponse adjustment method 1 to method 3. In region 1 and 3 the second and third method produces better results than the first. It appears that the creation of weighting class cells based on race and age group in region 1 and the creation of weighting class cells based on facility size and age group in region 2 is creating more variation. It could be that there is less variation in the response propensities then we expected. In regions 2 and 4 as well as overall the CVs are relatively flat across all three methods.

2.2 The National Youth in Transition Database (NYTD)

The NYTD study used all five nonresponse adjustment methods.

1. Traditional Cell Adjustments
2. Propensity Score Adjustment from a Logistic Regression Model using Deciles
3. Propensity Score Adjustment from a Logistic Regression Model using the Predicted Propensities
4. Propensity Score Adjustment from a Random Forest Algorithm using Deciles
5. Propensity Score Adjustment from a Random Forest Algorithm using the Predicted Propensities

The resulting nonresponse adjustments are compared in terms of the variance inflation factor (i.e., the design effect due to unequal weighting) and, with the exception of the random forest weights, estimated bias. Bias was estimated using a re-sampling procedure that simulated nonresponse among the set of responding youths in each wave. To do so, on each iteration, a number of the responding records (proportional to the number of responding records in the original sample) were randomly set to "non-response." This created a smaller replicate sample in which the response rate matched the rate actually observed for that wave, but for which survey outcomes were known for all records. Simulating nonresponse in this way assumes a missing completely at random (MCAR) response mechanism, in which response behavior is associated with neither the covariates (i.e., frame variables) used to compute nonresponse adjustments nor survey outcomes.

Next, the cell weighting, logistic/direct and logistic/decile nonresponse adjustments were computed for the replicate sample. The difference between the "true" score for four key survey outcomes (computed from the complete set of respondents for that wave) and the weighted estimate (using each of the preceding nonresponse weights) was computed for the replicate. The simulation ran for 100 iterations, after which the mean difference between the true and estimated score for each outcome was taken as an estimate of the bias of each nonresponse adjustment.

2.2.1 Cell Weighting Nonresponse Adjustment

Nonresponse adjustment class dimensions were drawn from a list of potential response covariates. For Wave 1, these were 32 AFCARS frame variables that were available for both respondents and non-respondents, as well as sex, race (five levels), and Hispanic origin. All potential response covariates were dichotomized: For categorical covariates, this occasionally required collapsing levels; for continuous

covariates, a median split was applied. Next, missing values for the covariates were imputed using a recursive hot-deck algorithm seeded with a sort list of state by sex.

Following imputation, the potential covariates were tested for association with response (yes vs. no) using (2 × 2) Pearson Chi-Square tests, using a liberal alpha level of .10 when testing associations for significance. The goal was to select up to four significant response covariates to define the adjustment classes while also ensuring that each class contained at least a minimum number of respondents. Specifically, up to four significant response covariates were selected, in descending order of significance, to define the most granular nonresponse adjustment class (i.e., with four dichotomous dimensions, yielding a maximum of 24 = 16 independent adjustment cells).

Slicing the response data at such a granular level often results in empty cells; however, each cell must contain at least one respondent to carry the weight of the non-respondents in that cell. Moreover, allowing only one respondent to represent a potentially large number of non-respondents leads to large weights that increase the weighting variance and lower the precision of weighted estimates. For this reason, a minimum of three respondents were required in every cell of an adjustment class for it to be used. If this was not the case, the least-significant response covariate was dropped from the adjustment class definition (reducing the number of cells by a factor of 2) and the collapsed class was retested. This process was repeated until a suitable adjustment class was found, or until all response covariates were dropped, leaving only the complete dataset to define the (one-dimensional) adjustment class.

Once a suitable adjustment class was defined, the nonresponse adjustment was computed as the ratio of cases selected to be surveyed in each weighting class cell to the number of responding cases in that cell, $w_1 = n_{selected} / nr_{espnded}$. The mean nonresponse adjustment weight was 1.87 with a minimum of 1.50 and a maximum of 4.64.

2.2.2 Logistic Regression Propensity Score Nonresponse Adjustment

The second approach to adjusting the NYTD data for nonresponse employed a logistic regression model to estimate response propensities (Iannacchione, Milne, & Folsom, 1991). Rather than selecting a discrete number of response covariates to use for defining nonresponse adjustment classes, this approach models response behavior as a function of the full set of potential response covariates. The resulting model is then used to estimate each sampled youth's propensity to respond. For Wave 1, all 39 available frame variables were entered into the model as predictors, with no interactions specified. Table shows the resulting confusion matrix for this model on the Wave 1 data, after classifying records with estimated response above .5 as a predicted response and below .5 as a predicted non-response. Overall, the model correctly classified 60% of the Wave 1 sample.

Table 3. Confusion Matrix for Logistic Regression Response Propensity Model

		Predicted Response	Non-Response
Actual	Response	39% TPR	15% FNR
	Non-Response	25% FPR	21% TNR

Estimated response propensities were attached to each sampled Wave 1 record using this model. For the direct logistic nonresponse adjustment, the nonresponse adjustment was computed as the inverse of the record's estimated response propensity. For the decile logistic nonresponse adjustment, the sample was divided into deciles according to estimated response propensities; then, within each decile, the nonresponse adjustment was computed as the sum of respondents and non-respondents divided by the number of respondents. Table 4 shows the number of respondents and non-respondents in each decile, along with the mean estimated response propensity.

Table 4. Summary of Logistic Regression Response Propensity Score Adjustment by Decile

Decile	*Mean Response Propensity*	*n Respondents*	*n Non-Respondents*	*Nonresponse Adjustment*
1	.33	953	1,957	3.05
2	.42	1,189	1,722	2.45
3	.46	1,371	1,539	2.12
4	.50	1,438	1,473	2.02
5	.53	1,541	1,369	1.89
6	.55	1,625	1,286	1.79
7	.58	1,705	1,205	1.71
8	.61	1,797	1,114	1.62
9	.65	1,891	1,019	1.54
10	.72	2,087	824	1.39

2.2.3 Random Forest Propensity Score Nonresponse Adjustment

The final approach tested for adjusting the NYTD data for nonresponse employed a random forest algorithm to estimate response propensities. The random forest algorithm is an ensemble recursive partitioning method that builds a forest of decision trees (in this case, classification trees), and then aggregates the votes across all trees in the forest to arrive at a predicted outcome (Strobl, Malley, & Tutz, 2009). For our purposes, a random forest (implemented using the randomForest package in R) of 400 trees estimated response propensities based on the same set of 39 Wave 1 frame variables used for the logistic regression model described above. One advantage of the non-parametric random forest approach over the logistic regression approach is that the former does not require explicit specification of higher-order interactions, but will naturally capture these interactions as the ensemble of trees is built.

The resulting confusion matrix for the random forest model is presented in Table 5. In this case, the results are comparable to those obtained using the main-effects logistic regression model described above.

Table 5. Confusion Matrix for Random Forest Response Propensity Model

		Predicted Response	Non-Response
Actual	Response	38% TPR	16% FNR
	Non-Response	24% FPR	23% TNR

Estimated response propensities were attached to each sampled Wave 1 record using this model. For the direct random forest nonresponse adjustment, the nonresponse adjustment was computed as the inverse of the record's estimated response propensity. For the decile random forest nonresponse adjustment, the sample was divided into deciles according to estimated response propensities; then, within each decile, the nonresponse adjustment was computed as the sum of respondents and non-respondents divided by the number of respondents. Table 6 shows the number of respondents and non-respondents in each decile, along with the mean estimated response propensity.

Table 6. Summary of Random Forest Response Propensity Score Adjustment by Decile

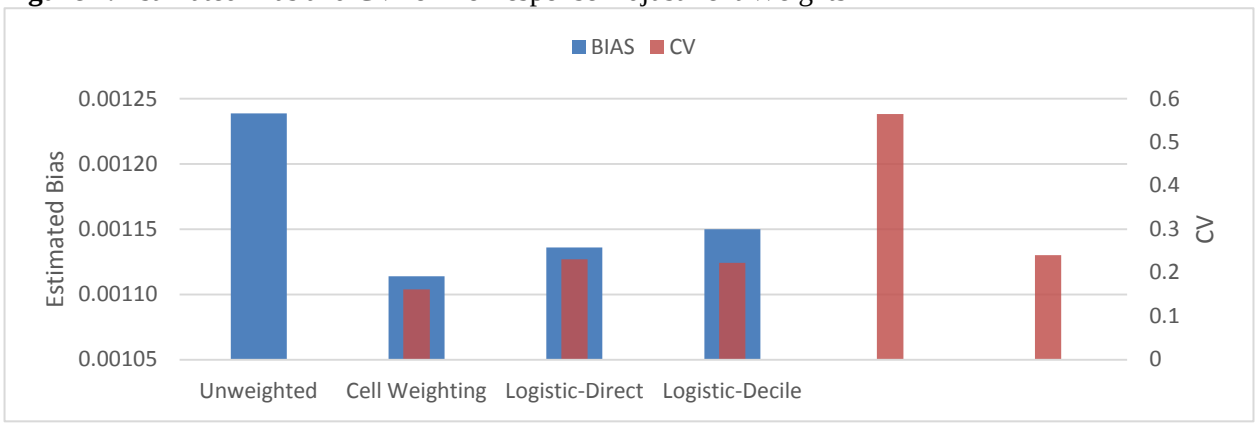
Decile	*Mean Response Propensity*	*n Respondents*	*n Non-Respondents*	*Nonresponse Adjustment*
1	.20	2,029	882	3.30
2	.33	1,710	1,200	2.43
3	.41	1,562	1,360	2.15
4	.48	1,480	1,420	2.04
5	.54	1,353	1,566	1.86
6	.59	1,246	1,663	1.75
7	.64	1,141	1,823	1.63
8	.70	1,061	1,797	1.59
9	.76	966	1,940	1.50
10	.86	960	1,946	1.49

2.2.4 Comparison of Nonresponse Adjustments

To compare the nonresponse adjustments, the two components of mean squared error, bias and variance, of each weight were computed. The variance of each weight is reported in terms of the coefficient of variation (CV), the square of which is the contribution of the weighting variance to the design effect. To estimate bias, a nonresponse simulation was conducted using the Wave 1 response data, as described earlier.

Figure 1 plots the estimated bias, in blue (scale on left axis), overlaid with the CV of each weight from the full response data, in orange (scale on right axis). Although the results are preliminary, they suggest that both the cell weighting and logistic regression propensity score adjustments successfully reduce bias compared to unweighted estimates. The cell weighting approach provided the best balance in terms of bias and variance, yielding a design effect due to weighting of only 1.03. The increased variance in the regression adjustments did not appear to contributing to improved bias reduction. Finally, although bias estimates were not simulated for the random forest propensity score adjustment, these propensities appeared produce the most weighting variance when used directly, suggesting that recursive partitioning models may benefit from the smoothing that results when estimated propensities are grouped into classes (such as deciles).

Figure 1. Estimated Bias and CV for Nonresponse Adjustment Weights



2. Conclusion

As we added more variables to our models we expected to see the increase in variance. The patient study weights did not produce the results we expected. This is likely due to a lack of information for both respondents and nonrespondents. While the patient study yielded mixed results regarding the variance due to the weighting adjustment, the NYTD study results were as expected. The NYTD study had 35 variables on both respondents and nonrespondents. We saw that the cell weighting method with the crossing of at most 4 variables (i.e. 16 weighting class cells) had the least amount of added variance. The deciles adjustments and direct application of the propensity scores from both the logistic models and the random forest models increased the variance. Interestingly, we expected the estimated bias to go down as we applied methods 2 through 5 but that was not the case. It is likely that our method of measuring bias was based on the wrong assumption. We assumed that the nonresponse was missing completely at random which means that the response behavior is associated with neither the covariates (i.e., frame variables) used to compute nonresponse adjustments nor survey outcomes. Further simulations are required using different assumptions regarding the mechanism for nonresponse.

References

Iannacchione, V. G., Milne, J. G., & Folsom, R. E. (1991, August). Response probability weight adjustments using logistic regression. In Proceedings of the Survey Research Methods Section, American Statistical Association (pp. 637-42).

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological methods, 14(4), 323-348.