Assessment of Alternative Weighting Methods for the National Health Interview Survey*

David Shaw† Joe Fred Gonzalez, Jr.‡ Meena Khare‡

Abstract

The National Health Interview Survey (NHIS) is a population-based survey which has collected health information from the U.S. civilian noninstitutionalized population since 1957. Major goals of the NHIS include the production of quality data as well as precise and reliable estimates of health conditions. This paper summarizes the results of alternative weighting techniques for NHIS data. The first approach applied different raking methods to the NHIS interim weights using marginal population totals for age, sex and race/ethnicity, education and income. The second approach consisted of nonresponse adjustments found by predicting the probabilities of household response and altering the weights using a weighting class adjustment. These predicted probabilities were found using two methods applied to the NHIS paradata: logistic regression and recursive partitioning. Raking was then also applied to these nonresponse adjusted weights.

Key Words: survey weighting, raking, nonresponse adjustment, logistic regression, recursive partitioning

1. Introduction

The National Health Interview Survey (NHIS) has been continuously conducted since 1957 and seeks to gain a nationally representative sample of the civilian, noninstitutionalized population of the United States. The NHIS consists of face-to-face personal interviews where the interviewee is asked to provide information about topics in the gamut of health issues as well as acute and chronic conditions the interviewee may suffer from. The 2007 NHIS consists of 75,764 records at the person level, only 75,504 of which were able to be included in the analysis due to the fact that 260 persons met a non-inclusion criterion[9].

In order to gain inference using the NHIS data, responses to the NHIS need to be extrapolated to the population as a whole to ensure the sample was representative of the entire population. Thus, an accurate and reliable weighting procedure is invaluableto any researcher working with this data. Currently, the NHIS uses the procedure outlined in Table 1 to obtain weights for each of the sampled individuals in the survey. The first-stage ratio adjustment helps to bring down standard errors of estimates by including geographical information about the population by region and residence/race-ethnicity within non-self-representing primary sampling units (PSUs). The second-stage ratio adjustment performs a similar task but for the demographic variables of age, sex and race/ethnicity.

The purpose of this paper is to obtain alternative sets of weights by modifying various stages of the current weighting procedure. One modification was done to the nonresponse adjustment (step 2), another modificationwas applied to the second-stage ratio adjustment (step 4) to account for noncoverage, and finally both modifications were applied simultaneously. Socio-demographic variables are useful in reducing standard errors of estimates,

*__Disclaimer:__ The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

†University of Maryland, College Park, Department of Mathematics, College Park, MD 20742

‡National Center for Health Statistics (NCHS), 3311 Toledo Rd, Hyattsville, MD 20782

Table 1: Current NHIS Weighting Procedure

1. Initial value for the weights is the inverse of the probability that an individual was selected
2. Initial value is then multiplied by a household nonresponse adjustment
    - These will be called the **interim** weights
3. All weights corresponding to persons in non-self-representing PSUs are then subjected to a first-stage ratio adjustment
4. All weights are finally subjected to a second-stage ratio adjustment (or post-stratification) to the U.S. population by age, sex and race/ethnicity
    - These will be called the **final** or **post-stratified** weights

especially if values of these demographic variables are known for the entire population. Thus, a raking on age, sex and race/ethnicity is proposed and should perform similarly to the second-stage ratio adjustment, while an additional raking on education and family income levels is proposed to further reduce standard errors in the estimates. An alternative nonresponse adjustment is analyzed as well in the hope that this method will improve the estimates obtained with the current method. All alternative weighting adjustments are alterations of the interim weights. The goal of subsequent analyses will be to match closely the interim weights to the final weights using only demographic variables, as well as investigate new areas of nonresponse adjustment applied to the interim weights.

2. Raking

2.1 Overview

Raking works by iteratively adjusting the sample weights, multiplying each weight by the ratio of the population control total and the corresponding sample total for a given variable. The method has been shown to work well when demographic variables are included, such as age, sex and race/ethnicity, but a drawback of the method is that exact marginal population counts are required. In the following analyses, estimates for these marginal counts are used instead of exact counts. Raking is also beneficial in cases when exact cell counts are unattainable, e.g., when the total number of males in the U.S. is known and the total number of individuals under 18 in the U.S. is known, but perhaps the total number of males under 18 in the U.S. is unknown. More complete descriptions of the procedure and analysis of its effects on estimates are given in the seminal paper [7] whereas practical considerations and the method employed are explained in detail in [3].

2.2 Methodology

For the NHIS weights, a post-stratification is already done on age, sex and race/ethnicity, thus an improved weighting procedure will require additional demographic variables that are correlated with variables of interest in order to lower standard errors of survey estimates. In addition to age, sex and race/ethnicity, individual education level and family income will be utilized in the same way NHIS uses post-stratification, but raking will be necessary

as only estimates of marginal population totals are known for these additional variables, while estimates of the cell counts are unknown. It is possible that a number of variables of interest in NHIS (e.g., type of insurance coverage, frequency of hospital visits, etc.) will be correlated with education, income or both and including these variables may help to improve the current weighting procedure. The abbreviations of raking 1 through raking 6 will correspond to the following rakings:

1. Raked on Age, Sex, Race/Ethnicity (A/S/RE)
2. A/S/RE, Education
3. A/S/RE, Coarse Income
4. A/S/RE, Education, Coarse Income
5. A/S/RE, Fine Income
6. A/S/RE, Education, Fine Income

Raking 1 should be nearly identical to the post-stratified weights discussed in the introduction. There will be a slight disparity however, due to the final weights being adjusted for nonresponse as well as utilizing an additional ratio adjustment. Most analyses of the weights will not include this raking except where relevant as most interest lies in the inclusion of the education and income variables. The difference between being raked on coarse income and fine income is the groupings used to determine the control counts. The control counts for age, sex, race/ethnicity, education and income were obtained from the 2007 Current Population Survey (CPS)[6] and used as estimates for the marginal population totals.

2.3 Imputation

While the person data file is complete for age, sex and race/ethnicity, education and income variables suffered from missing values that needed remedy in order to get accurate counts to perform the raking. Without some form of imputation for these missing values, the raking would not be possible.

2.3.1 Education

A total of 2,286 observations, roughly 3% of the individuals sampled, had a missing value for education level, all of which were recorded by the data collector as having a value of either "refused", "not ascertained" or "don't know." The small amount of missing data for this variable did not warrant any exotic imputation methods, thus a naïve hot-deck imputation was done. In other words, missing values for education were assigned a random value where the assignments were done with respect to the distribution of the control counts for each education group. The goal was to match as closely as possible the distribution of the population counts for education before any raking was done. This imputation was necessary for the raking to converge, and more elaborate and intelligent approaches will clearly accomplish the same goal with more assumptions being made to model the missingness mechanism.

2.3.2 Income

A total of 11,589 observations, roughly 15% of the individuals sampled, were missing or undetermined for the family income variable. Another drawback is that the income values

present in the NHIS public use data fell within one of four coarse groups for income and thus give a rougher picture of family income than may be desired. The scope of this missingness indicates that a more robust strategy than naïve hot-deck should be employed to impute the missing income values. The multiply imputed income dataset that is available from the data release[9] was used to place those missing observations into income groups, in this case given by the more refined groups for income. Five imputations were performed, meaning raking was done five separate times to obtain five different sets of weights. Each set of weights that was adjusted for income using raking required use of this imputed income variable and thus each quantity of interest was estimated five times and averaged over all five to arrive at a final estimate. Computation of the standard errors of these estimates proceeded in the normal fashion, with the inclusion of an additional error term due to the variation in the different imputations[14]. Finally, when the distributions of the different sets of weights were compared, the average over the five sets of weights corresponding to each of the five imputations was used.

3. Nonresponse

3.1 Overview

In order to perform an adequate nonresponse adjustment, some useful information must be provided for both the respondents and the nonrespondents. It has been posited that nonresponse results from two broad factors that are able to be used by researchers: survey design variables and interviewer characteristics[2, 10]. The first source of variation is immutable once the survey is designed and carried out, and difficulties can arise when using the design as the only adjustment for nonresponse. A correction for nonresponse that takes into account interviewer characteristics can be done by utilizing paradata recorded at the family level. Paradata includes information recorded about the interview process into a Contact History Instrument (CHI). Interviewers enter information about the contact attempt as well as the mien of the interviewee during the attempt by recording any "doorstep concerns" or modes of contact used and subsequent outcomes[2].

To use the paradata to correct for nonresponse, a weighting class adjustment was proposed by attempting to model the mechanism of nonresponse and alter the weights accordingly. Two methods were used to predict the probability of response: logistic regression and recursive partitioning. Once a predicted probability of response is obtained, its inverse is calculated and truncated at a value of two[8, 12]. This inflation factor is applied to the sample weights in order to correct for the non-inclusion in analyses of those individuals who did not respond to the survey.

3.2 Logistic Regression

3.2.1 Method

Logistic regression[13] is used to obtain a propensity of response in order to obtain nonresponse adjustments. Fitting an adequate model to the paradata can prove to be a daunting task, and various other methods are needed in selecting the best variables for the model. The ten predictors having the highest absolute correlation with the nonresponse variable were used. Using correlation to obtain predictors for the model is an ad-hoc method to reduce the number of variables under consideration and will not necessarily result in an optimal solution. However, with the scope of the problem this method of paring down the variable space seemed to perform well. A further paring was done using stepwise model

selection where the Bayesian Information Criterion (BIC) was used to obtain the model with the best tradeoff in deviance and number of model parameters. The BIC is defined as

$$BIC = -2 \log L(y, \theta_{\text{red}}) + p \log n$$

where p is the number of parameters in the current model, n is the total number of observations and $L : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is the likelihood function taking as arguments the data as well as the parameters in the current model. The parameters $\theta_{\text{red}} \in \mathbb{R}^p$ are the reduced (or current) model parameters, the name used to emphasize the overall goal of getting as close to a "perfect" fit as possible using the fewest parameters. This amounts to minimizing the BIC by finding the highest value of the likelihood possible using the fewest model parameters.

The BIC indicates when a model achieves an adequate amount of parsimony, but a statistic to describe how well the model fits the data is also necessary. The residual deviance performs this task and is defined as

$$D = -2 \log \frac{L(y, \theta_{\text{red}})}{L(y, \theta_{\text{full}})}$$

where $\theta_{\text{full}} \in \mathbb{R}^n$ are the parameters for the saturated model. The saturated model fits the data exactly as it possesses as many parameters as there are observations. Thus reducing the deviance amounts to obtaining a fit that is as close to an exact fit as possible. If this can be acquired with relatively few parameters, an adequate model is obtained. However, there is nothing to say what exactly constitutes a "low" deviance, therefore the null deviance is defined as

$$D_0 = -2 \log \frac{L(y, \theta_0)}{L(y, \theta_{\text{full}})}$$

where $\theta_0 \in \mathbb{R}$ is simply an intercept for the relevant response variable. If the null deviance is reduced significantly with the proposed model, an adequate fit is obtained, though this certainly depends on structure of the model and the response variable.

3.2.2 Goodness of Fit

Using the ten variables with highest absolute correlation with nonresponse and stepwise model selection with BIC as a criterion resulted in poor performance of the model as not all of the coefficients could be determined, most likely due to a lack of identifiability in the model as many of the predictors were categorical. Those variables whose coefficients could not be determined were dropped so that a model is obtained with which predictions can be made. This final model is dependent on the variables `reluc22r`, `unable3r`, `ctstat3`, `mode_p` and `totcount`, definitions of which are supplied in Table 2.

There are some serious drawbacks using the logistic regression model, and the fit could be improved vastly. First, all coefficients in the model are significant at the .001 level, though one of the coefficients seems counterintuitive. For the categorical variable `unable3r`, when it is equal to 1 the model intercept decreases by 4.41, but when it is equal to 0 the model intercept only decreases by 2.59. Thus, no matter what value was entered for `unable3r`, the predicted probability of response will still decrease. In other words, if the interviewer enters a value of 0 for "# times respondent is reluctant," it will still decrease the probability of response even though it seems like this should not be the case, though the reduction is still less than if the interviewer had entered 1 for the variable, meaning the respondent was reluctant 1 or more times.

With this approach 10,406 observations were ignored due to missingness in the covariates, though 10,011 of these values should be ignored regardless because no interview

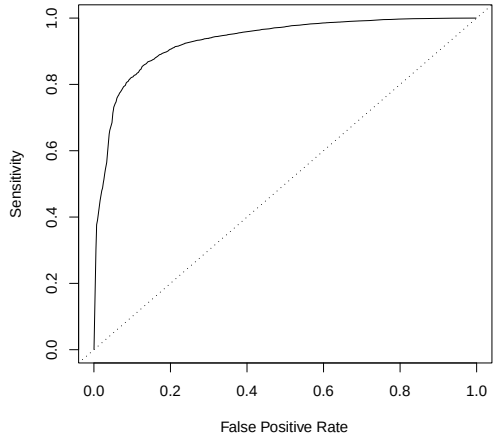
Variable	Description
reluc01r	# times "not interested/don't want to be bothered" was entered
strat05r	# times "called household" entered
unable3r	# times "respondent is reluctant" entered
reluc22r	# times "no concerns" entered
reluc03r	# times "interview takes too much time" entered
unable8r	# times "other" entered for reason unable to conduct interview
ctstat1	# contacts made with sample unit members
ctstat3	# noncontacts
totcount	total count of CHI records for this case
mode_p	# personal visit attempts for this case

Table 2: Definitions of pertinent variables in the paradata that are used for both nonresponse adjustments. Most variables dealing with "# times *X* was entered" were actually categorical variables with 3 categories: "0", "1 or greater", "not recorded'.'

was conducted as the result of some non-inclusion criterion, such as the individual selected was institutionalized or in the military. However, the missingness in any of the variables (ctstat3, mode_p or totcount) caused 395 additional observations to be ignored that may contain valuable information. One potential remedy is to treat the missing values of these variables as zeros. This might be a large assumption, but it seems logical to assume that if no value was recorded for "# of noncontacts" then the interviewer may not have experienced a noncontact to record. A similar argument can be made for "# personal visit attempts" and "total count of CHI records." Though rational, this assumption is baseless and relying too heavily on it may cause a misinterpretation of the results. Also of note is that for many of the observations with a missing value for any of these variables, the values of the other variables are also missing. In other words, if ctstat3 is missing for a particular record, mode_p and totcount are also missing for that record. This leads to the potential solution of treating these variables as factors and treating the missing values as separate levels of the factors. For example, ctstat3 ("# of noncontacts") can take any value from 0 to 63, but as this number increases the number of families with the corresponding number of noncontacts decreases rapidly. Categories are created to best split the data among respondents and nonrespondents. One drawback here is that the "best" splitting is unknown and would require some knowledge of the response variable which could cause problems when trying to predict the response variable using this information. This idea, however, is explored further in the next section. All of these remedies require assumptions that may not necessarily be grounded in reality, so the default method of ignoring the missing values was used.

Finally, the ROC curve and concordance table in Figure 3.2.2 show the model's ability to predict response in the given data, with 93% concordance indicating an adequate fit but leaving room for improvement. The ROC curve seems to have an area-under-curve close to one, but again this could be improved with a better fit. Investigating the fit, the null deviance is almost halved from 25,988 to 13,810 using only 7 additional degrees of freedom compared with the null model. The expected value of the residual deviance should be equal to the degrees of freedom in the model, therefore in practice if the residual deviance for a binomial model is not close to the degrees of freedom a dispersion parameter may need to be included and estimated[13, 15]. The final model had 34,048 degrees of freedom indicating under-dispersion and a possible need for estimation of a dispersion parameter, but when included the fit and results were largely unchanged, thus the simpler model was

preferred. When analyzing binary responses using logistic regression, the deviance has no guarantee of being chi-square-distributed as the theory suggests[15], and the fairly significant reduction in null deviance with such a large problem in scope is able to illuminate some merits of the model, but it is merely an adequate model and alternate methods would be preferred.



Concordant	93.0%
Discordant	6.84%
Tied	.167%

Figure 1: ROC curve and concordance table for the predicted probability of nonresponse obtained through logistic regression. The area-under-curve is quite high, though there is room for improvement. Similarly, 93.0% concordance is quite high but again could be improved upon using various alternative methods.

In future analyses using logistic regression, exploration of two-way interactions may prove interesting but could suffer from the same identifiability problems and would be decidedly less parsimonious than the current model. Using "in-house" data—data in which many of the variables are given in more detail—could allow for categorical variables to be treated as linear terms, and parsimony and identifiability would be more easily achieved at the expense of potentially worsening the fit depending on how the values are distributed. The handling of missing values when using logistic regression leaves much to be desired as currently the values are simply ignored and any remedy will require strong assumptions about the missingness as well as the data. Finally, inflation factors obtained with logistic regression tend to be overly variable and spread out, essentially creating one-person "groups" based on propensity scores that are too fine to account for overall nonresponse in the survey. This has the undesired effect of over-inflating many of the weights. One solution to this is to place the inflation factors into groups defined by the quintiles of their distribution and proceed as before, but with the quintile value used in place of the inflation factor. Modeling nonresponse from the paradata using logistic regression has its shortcomings, thus alternative approaches to handling nonresponse should be investigated.

3.3 Recursive Partitioning

3.3.1 Method

Using logistic regression to model nonresponse is subject to the various pitfalls described above, and still another problem with the approach is the high variability of the predicted probabilities. While one solution to this problem is to split the predicted probabilities into separate classes[12], other methods can be used to achieve the same goal. Recursive par-

titioning has the desired properties and uses a technique similar to CHi-square Automatic Interaction Detection (CHAID)[8, 11] to correct for nonresponse by using predictors to split the respondents and nonrespondents into groups that are as homogeneous as possible[5].

Recursive partitioning works by searching through all possible splits of the data with respect to a given set of predictors and determining which of these is the local optimum. The number of ways to find a local optimum are manifold, one way defining a measure of impurity of a distribution of responses at each node and searching for the split which achieves the minimum of this value. Impurity measures can include different types of entropy as well as deviance, the one used in subsequent analyses being the Gini index and defined as

$$H_i = 1 - \sum_k p_{ik}^2$$

where i indicates the index for the current node at which the split is being determined and p_{ik} is the estimated proportion of the k^{th} class of the response variable at node i[5, 15]. The algorithm that determines the best-fitting tree terminates when either less than 20 observations are being considered at a test node, less than 6 observations fall within a terminal node or when the Gini index can no longer be significantly reduced. For nonresponse adjustments, the predicted probability of response is simply the proportion of respondents at the corresponding terminal node.

As in logistic regression, recursive partitioning too has a tendency to overfit when the number of predictors is large, thus similar methods for achieving parsimony should be considered when selecting variables. One such method is pruning where the "cost" R of a current tree is altered to include the size of the tree. Thus, a new cost R_α is defined as

$$R_\alpha = R + \alpha \cdot n$$

where n is the number of terminal nodes in the tree. Ten-fold cross-validation is used to find the optimal value of α and determine a subtree which minimizes R_α[15]. This procedure is strikingly similar to the stepwise model selection used for the logistic regression model.

3.3.2 Goodness of Fit

To grow the decision tree modeling the response mechanism, all variables in the paradata that did not suffer from a large number of missing values were considered at each node for splitting. The final pruned tree is given in Figure 2 and shows each splitting criterion and the predicted probabilities of response at each terminal node. For categorical variables (`reluc01r, reluc03r, reluc22r, strat05r, unable3r, unable8r`) a value of "a" in the tree means the corresponding value of that variable in the paradata was not recorded, a value of "b" means the corresponding value was 0, and a value of "c" means the corresponding value was 1 or greater.

To elucidate the process of prediction, consider two families who had values recorded in the paradata given in Table 3. Recall the meanings of the variables are given in Table 2. The ellipses indicate that the value of these variables do not affect in any way the predicted probability which can be either a benefit or a drawback depending on the problem. The predicted probability for Family 1 is the proportion of families in the dataset who have `reluc01r` equal to "MISSING" or "1," `ctstat1` less than 0.5 and `strat05r` equal to "0" or "1" which for this dataset is 0.1489. Similarly for Family 2, following the tree in Figure 2 to the rightmost node shows that this family will have a predicted probability of response of 0.9806. Similar to the logistic regression approach, inflation factors are

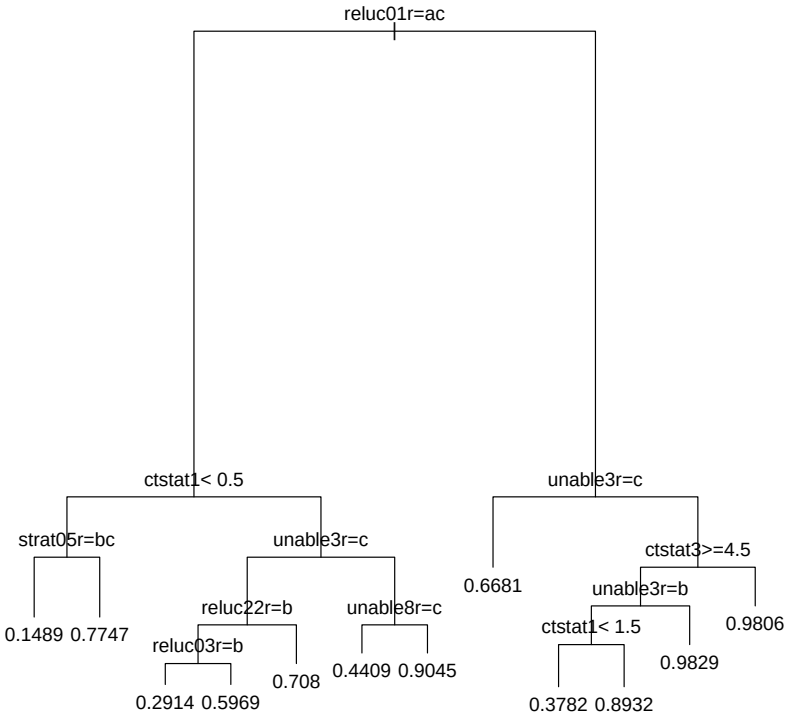


Figure 2: Dendrogram of fit for predicted probability of nonresponse obtained from recursive partitioning. The predicted probabilities are found by traversing the tree, making a decision at each node corresponding to the values of the predictors of the given observation. The variable descriptions are given in Table 2, where here a value of "a" means the corresponding variable of that variable in the paradata was not recorded, a value of "b" means the value was 0 and a value of "c" means the value was greater than or equal to 1. A summary of how prediction is done with this tree is given in Table 3.

truncated at a value of two so as not to allow single individuals to be over-representative of a certain population.

Figure 3 shows the difference between the distribution of the nonresponse inflation factors for both logistic regression (left) and recursive partitioning (right). The inflation factors arising from logistic regression indicate a larger spread than thos obtained from recursive partitioning, and as explained in section 3.2.2 this can potentially cause problems. The recursive partitioning factors have a similar, yet more discretized distribution than the logistic regression factors. Additionally, more weight seems to be at lower values for the recursive partitioning, meaning that there is less occurrence of small groups of individuals representing large populations. A final potential benefit of recursive partitioning is that it can be interpreted in a fairly straightforward manner due to its use of rules defined by predictors as opposed to linear combinations of the predictors.

Though recursive partitioning paints a relatively more optimistic picture than logistic regression, there are still some drawbacks. The model relies on local optima to define the

Variable Name	Family 1	Family 2
reluc01r	MISSING	0
strat05r	1	...
unable3r	...	0
reluc22r	...	...
reluc03r	...	...
unable8r	...	...
ctstat1	0	...
ctstat3	...	2
Pr(Response)	0.149	0.981
1/Pr(Response)	6.716	1.020
Inflation Factor	2.000	1.020

Table 3: Example calculation of inflation factors using the tree in Figure 2. Ellipses ('...') indicate the corresponding variable can take any value.

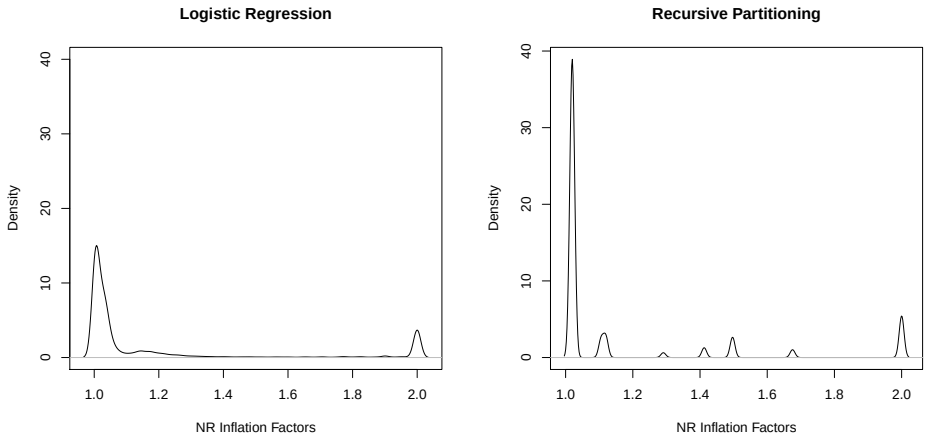


Figure 3: Distributions of nonresponse inflation factors. Logistic regression tends to give inflation factors that are more "spread out" while recursive partitioning gives more discretized values.

splits and thus is not guaranteed to find the globally optimal decision tree. One possible solution to this is to use random forests[4] to aggregate many trees that are defined on different subsets of predictors and observations. This can help to reduce error in the model but can reduce the interpretability as well. Another drawback is the dearth of variables used in the final tree. Due to the stopping criteria supplied as well as the pruning, recursive partitioning results in a decision tree which only takes into account a relatively small number of variables, meaning some meaningful variables may be completely ignored. This problem can also be present in logistic regression when trying to achieve parsimony, though it certainly depends on the methods used to select the final variables. With decision trees, random forests can also be used to incorporate as much meaningful information as possible in the model. Recursive partitioning ostensibly has a slight advantage over logistic regression in flexibility and interpretability, thus it is compared with logistic regression in the subsequent section.

4. Performance

4.1 Comparison of Estimates

There are myriad tests that can be run to determine if the new sets of weights are worthwhile. Unfortunately, all of these tests will rely heavily on certain assumptions. Various estimates of variables of interest will be found along with their standard errors, but it is difficult to tell which estimate is "better" due to the lack of a true value for the given variable. Typically a low standard error in the estimate will indicate confidence in the estimate, but the closeness of an estimate to the "truth"—the bias—can be just as valuable. The estimate obtained using the NHIS post-stratified weights for each variable will be assumed to be the "true" value of that variable because the NHIS weights are design-based and give unbiased estimates, therefore, all estimates utilizing alternative weighting strategies will be compared to the NHIS estimates.

In the following analysis, variables of interest were chosen for the population of individuals under the age of sixty-five. Four of these variables pertained to insurance coverage, with definitions of each variable given in [1]. The last three variables deal with self-reported health status. Table 4 shows the results for three of the six different raking adjustments with no nonresponse adjustment compared to the NHIS estimates. The relative standard errors (RSE) for the raked estimates were nearly identical to those for the NHIS estimates so they were not included, and all are below the typical threshold of 0.25. Relative root mean square errors (RRMSE) were also similar for each of the rakings, thus the relative absolute bias (RAB) for each of the weighting methods was calculated. Since the RSEs were close to equal across weighting methods and the bias is a component of the RRMSE, the RAB was used in order to make comparisons and gain insight into which method is performing the "best." Similar tables are available for the nonresponse adjusted weights where those adjusted using logistic regression have RRMSEs and RABs slightly larger than the corresponding statistics for estimates found using recursive partitioning, though in the end these differences are not significant. Both nonresponse adjustments increase all of the RRMSEs and RABs when compared to those that are not adjusted for nonresponse, where the highest increases come from variables whose proportions NHIS estimated to be relatively low (such as "Health Status: Fair/Poor" and the insurance variable "Other"). The nonresponse adjustments make matters worse due to the fact that they were performed on the interim weights in the NHIS which had already included a nonresponse adjustment. Essentially this resulted in two adjustments which produced much larger weights than desired to the fact that inflation factors are increasing through both of these adjustments. Performing future analyses on weights that do not include a nonresponse adjustment will hopefully mitigate this problem.

	NHIS Est.		**Raking 2**		**Raking 3**		**Raking 4**	
Variable (under 65 years old)	Est. (%)	RSE	RRMSE	RAB	RRMSE	RAB	RRMSE	RAB
Uninsured	16.64	.017	.020	.010	.018	.007	.017	.003
Private	66.80	.007	.009	.005	.007	.001	.007	.001
Other (military, Medicare or other gov. care)	2.68	.059	.058	.000	.058	.004	.058	.005
Medicaid	13.88	.023	.028	.013	.025	.006	.024	.004
Health Status: Excellent/Very Good	69.77	.005	.059	.001	.009	.005	.018	.007
Health Status: Good	22.78	.013	.025	.008	.058	.000	.007	.001
Health Status: Fair/Poor	7.44	.021	.020	.010	.028	.013	.058	.004
Average	—	.021	.031	.007	.029	.005	.027	.004

Table 4: Statistics for raking with no nonresponse adjustment. RRMSE is calculated using the NHIS estimate as the "true" value. Relative absolute bias (RAB) is done similarly.

An accurate view of the performance of the weighting methods may be obfuscated by the amount of statistics included in Table 4, thus the average of the RAB over the different variables of interest was computed for each of the weighting methods and ranked from lowest to highest. Averaging over all variables of interest will make analysis slightly easier and is justified due to RAB being a relative statistic. Without a nonresponse adjustment, including education and income (detailed or not) resulted in a lower average RAB over the variables of interest than any of the other raking methods. When nonresponse adjustments are included, including education and income also results in a lower average RAB over the variables of interest, though the detailed income groupings are favored.

4.2 Comparison of Weight Distributions

Another method of assessing the potential utility of alternative weighting methods is to compare the distributions of the weights themselves. Various statistics can be computed, such as coefficient of variation (CV), median, maximum, inner quartile range (IQR) and standard deviation (SD). The CV, IQR and SD all give an idea of how closely distributed the weights are to one another, so relatively small values for these statistics will be desired as discussed in section 3. The CV and SD are both sensitive to outliers which the IQR attempts to overcome. When analyzing weights, outliers are important to investigate because they indicate that one or a few people are representing an abnormally large segment of the overall population. The median gives an idea as to where the weights are most heavily concentrated without taking into account any of the outliers, whereas the maximum shows the most individuals in the overall population that are represented by one individual. All of these statistics will be beneficial to studying which weighting method is best, and they are included in Table 5.

Weighting Method	CV	Median	Max	IQR	SD
NHIS Weights					
Post-Stratified	0.559	3731	46787	2345	2191
Interim	0.573	3915	42065	2401	2246
Raked					
Raking 1	0.566	3726	46348	2324	2218
Raking 2	0.570	3709	49682	2372	2232
Raking 3	0.566	3729	46797	2331	2219
Raking 4	0.570	3696	48900	2380	2233
Raking 5	0.571	3687	45203	2371	2238
Raking 6	0.575	3654	46015	2422	2252
Nonresponse Adjusted (with no raking)					
Logistic Regression	0.819	3349	92821	2338	3209
Recursive Partitioning	0.783	3405	98274	2317	3066

Table 5: Various statistics for each set of weights obtained. Raking tends to lower the median of the weights while increasing the CV and standard deviation, while adjusting for nonresponse tends to decrease the median and IQR while greatly increasing the CV and standard deviation.

Post-stratified weights achieve the lowest CV and SD, which are to be expected. These weights are more closely distributed due to the multiple ratio adjustments performed. The multiple ratio adjustments are done on fairly homogeneous population groups so the resulting weights obtained from these adjustments will have a smaller variance due to the fact

that the adjustments create weights that have within-group means similar to the means estimated for the whole population within these groups. The closest CV and SD to that of the post-stratified weights are raking 1 and raking 3, while raking 2 and raking 4 are close as well. Raking 1 is simply raked on age, sex and race/ethnicity, and these results indicate that perhaps this is the best we can do at the current time. Raking 3, which includes income, is distributed nearly identically to raking 1, suggesting income may be a viable demographic variable to consider as well. Rakings 1 and 3 both have lower IQR than the post-stratified weights, thus these weightings might have some outliers that could cause problems. This is a potential avenue for improvement. Finally, nonresponse adjustment seems to have some major flaws. The maximum weight for both nonresponse-adjusted weights is roughly doubled from their raked counterparts, indicating a large amount of weight is being placed on a single response. However, this increase in the outliers results in a reduction of the IQR, both values being lower than that of the post-stratified weights. Some investigation into the outliers of the weights could prove beneficial and may result in weights that are closer to the post-stratified weights. Statistics for raked and nonresponse-adjusted weights are similar to those for the nonresponse-adjusted weights that did not include any raking.

Figure 4 also sheds light on the effect raking has on the distribution of the weights. The maximum weight in each case was roughly 46,000, thus the plot is truncated at a value of 15,000 in order to see the distribution of the weights more clearly. Raking 3 achieves a distribution closest to that of the NHIS weights, which is a raking including age, sex, race/ethnicity and income. This could be due to a number of factors. First, income was multiply imputed, meaning the distribution of raked weights including a raking on income is the mean across the five different imputations. This may help to further refine the distribution, but this is not the case in rakings 4 and 6 which also include income. Another reason for this may be the fact that the income variable only includes four categories, meaning the weights are only slightly altered from the weights obtained in raking 1. It seems they are altered in the correct direction though, thus income is a viable alternative demographic variable to include for post-stratification. Raking on the more refined income groupings seems to place higher weight on more observations whereas the other rakings have a smoother distribution. Distribution plots for nonresponse-adjusted weights are similar, but shifted left as higher weight is placed on the outlying observations resulting in less weight being placed on the bulk of the remaining observations.

5. Conclusion

Raking has some potential advantages. Estimates may be more reliable as the weights now match estimated population counts for education and income whereas the current NHIS weights do not. For variables that are assumed to be highly correlated with education and income, this could have some hidden improvements in the estimates that may be difficult to quantify. Some possible outlets are to use cross-validation to get a better idea of what the "true" bias of these raked estimates are as opposed to assuming the post-stratified estimate is the truth. Furthermore, when taking the post-stratified estimates as truth, raking on education and income results in lower RAB than simply raking on age, sex, and race/ethnicity, while raking on income yields weights that are similarly distributed to post-stratified weights with the added benefit of accurately representing income levels in the estimates.

The downside to requiring the weights to match additional demographic control totals is that RSEs increase as well as CVs and SDs for the weight distributions, ostensibly resulting in less desirable weights. Another problem arises from the fact that the control totals were obtained from CPS and thus also include error. This error was not estimated in the current

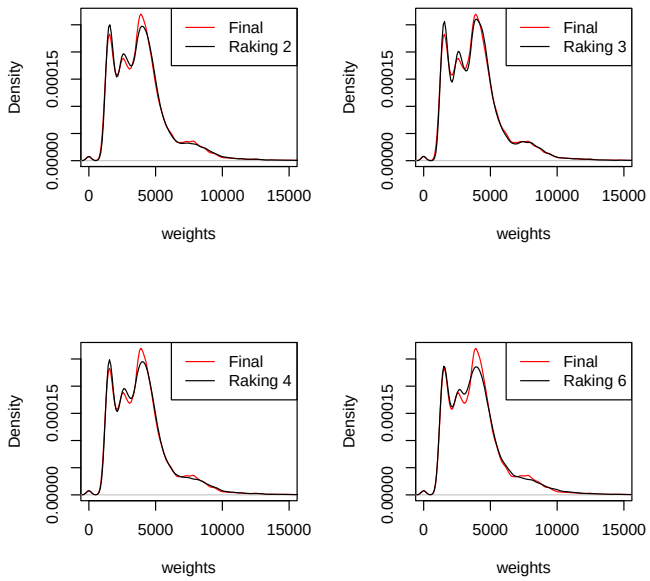


Figure 4: Comparison of raked weight distributions. Raking 3 (age, sex, race-ethnicity, coarse income) seems almost identical to the NHIS post-stratified weights, whereas the other sets of weights tend to place more value at different peaks of the distribution. Rakings 1 and 5 were not included as they were similar to rakings 4 and 6, respectively.

study, and the CPS estimates were taken at face value. Education and income values both had to be imputed, income values heavily so, which propagates yet additional error and further reduces the reliability of the estimates.

Nonresponse adjustment is a foggier issue. When comparing recursive partitioning to logistic regression, recursive partitioning seems to have some advantages. The predicted probabilities are easily interpreted and the method is highly flexible and can support any amount of missing data. Additionally, recursive partitioning results in lower RAB and CV when compared with logistic regression, though this difference is only slight. Both adjustments resulted in weights that were more densely distributed at lower values but with a few outliers, and an investigation into these outliers may prove beneficial. Finally, nonresponse adjustment seeks to answer the question of imputing responses for those individuals that refused to participate in the survey. Not adjusting for this will result in estimates that are misleading.

The obvious drawbacks to nonresponse adjustment are that the estimates obtained are significantly different from those obtained using the NHIS weights which also includes a nonresponse adjustment. This indicates a more robust nonresponse adjustment is necessary. Nonresponse adjustment also has a tendency to over-inflate weights that are already near the maximum value, resulting in the creation as well as the aggravation of outliers.

While still in the nascent stages of altering the weighting procedures performed in the NHIS, it seems that raking has some potential advantages over the current weighting procedure, though there are numerous pitfalls to be addressed. Similarly, more robust nonresponse adjustment procedures applied in addition to the raking could result in estimates that are better than those obtained with the current procedure. Additional analyses such as cross-validation to estimate bias and refine the estimates as well as ratio adjustments

to reduce standard errors and more robust nonresponse adjustments would aid in producing weights with myriad beneficial properties. Currently it is difficult to say if or by how much the current weighting procedure can be improved by using the above techniques, but it never hurts to try.

References

[1] P.F. Adams, P.M. Barnes, and J.L. Vickerie. Summary health statistics for the US population: National Health Interview Survey, 2007. *Vital Health Stat*, 10(238), 2008.

[2] N. Bates, J. Dahlhamer, and E. Singer. Privacy concerns, too busy, or just not interested: Using doorstep concerns to predict survey nonresponse. *Journal of Official Statistics*, 24(4):591–612, 2008.

[3] M.P. Battaglia, D.I. Izrael, D.C. Hoaglin, and M.R. Frankel. Practical considerations in raking survey data. *Survey Practice: Practical Information for Survey Researcher*, June 2009.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

[6] U.S. Census Bureau. Current Population Survey (CPS), Aug 2010. `www.census.gov/cps/`.

[7] W.E. Deming and F.F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11:427–444, 1940.

[8] T. Ezzati and M. Khare. Nonresponse adjustments in a national health survey. In *Proceedings of the American Statistical Association, Survey Methods Section, Alexandria, VA*, pages 339–344. American Statistical Association, 1992.

[9] Centers for Disease Control and Prevention. NHIS – 2007 data release, Aug 2010. `www.cdc.gov/NCHS/nhis/nhis_2007_data_release.htm`.

[10] R.M. Groves and M.P. Couper. *Nonresponse in Household Interview Surveys*. Wiley, New York, 1998.

[11] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119–127, 1980.

[12] R.J.A. Little. Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54:139–157, 1986.

[13] P. McCullagh and J. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC, Boca Raton, 1989.

[14] N. Schenker, T.E. Raghunathan, P. Chiu, D.M. Makuc, G. Zhang, and A.J. Cohen. Multiple imputation of family income and personal earnings in the national health interview survey: Methods and examples. Technical report, National Center for Health Statistics and University of Michigan, Aug 2008.

[15] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S (4th edition)*. Springer, 2002.