Modeling and Polishing of Nonresponse Propensity

David Judkins[1], Hongsheng Hao[1], Brandon Barrett[2], Pramod Adhikari[3]
Westat[1]
Emmes[2]
Australian Bureau of Statistics[3]

Keywords: CHAID, logistic regression, raking

Abstract

Two methods for modeling of nonresponse propensity in longitudinal studies are compared in terms of the resulting preservation of marginal baseline distributions among followup respondents and in terms of the relative variance of the weights. These methods are tree-based automatic interaction detection (CHAID) and logistic regression. Additionally, it is demonstrated that sample-based raking of the resulting nonresponse-adjusted weights to baseline distributions can improve face validity with fairly neutral effects on variances for followup cross-sectional estimates.

1. Introduction

Nonresponse adjustment procedures are commonly used in the data weighting process to compensate for missing responses with the purpose of reducing non-sampling error. In this process, the weights of the nonrespondents are distributed to the weights of eligible respondents. A commonly accepted procedure for making such adjustments is to form adjustment classes (or cells) based on differential nonresponse propensities (e.g., Kalton, Lepkowski, and Lin, 1985; Göksel, Judkins, and Mosher, 1992). Various methods are available for modeling the nonresponse propensity and forming the cells. This paper evaluates the effectiveness of two widely-used methods, in much the same as was done by Rizzo et al. (1996) and Folson and Witt (1994). Although Rizzo, et al. found little difference among methods, it was thought worth re-investigating this question in the context of a different longitudinal survey and a newer procedure for incorporating a much larger number of covariates than they had considered. This evaluation was carried out on a longitudinal survey of youth and their parents in support of a program evaluation. The paper also examines the benefits of polishing the nonresponse adjustment with raking to full-sample controls totals.

1.1 Background

In June 2002, the Centers for Disease Control and Prevention (CDC) launched a multimedia advertising campaign designed to encourage physical activity and healthy behaviors among 9- to 13-year-olds in the United States. To assess the impact of the campaign, Westat is conducting the Youth Media Campaign Longitudinal Survey (YMCLS). A panel design that included parent-child dyads was employed (Potter, et al, 2004). Two panels have been selected, one in the spring of 2002, prior to the launch of the campaign, and another in the spring of 2004. To date, each panel has been followed up in the spring of each year. This paper concerns only the first panel. For each panel, a list-assisted random-digit-dialed (RDD) method was used to select a sample of households with children aged 9 to 13 years.

Among screened households with one or two children in this age range, both were selected into the sample with certainty. In households containing three or more children aged 9 to 13 years, two children were randomly selected. A parent/guardian interview was conducted in each household containing an eligible child or children prior to the interview with the sampled child.

1.2 Weighting and Nonresponse Adjustment for the YMC Project

The focus of the YMCSL analyses has been on the child-level estimates, which were produced based on final child-level weights. In the process of deriving the final child weights, nonresponse adjustments are necessary for all levels of sample units where nonresponse has occurred. For the baseline year, nonresponse adjustments were made at the household, parent, and the child levels. For each of the followup years, further adjustments were made at parent and child levels to compensate for parent and child attrition from the prior year. This paper focuses on the 2004 weighting process. In this process, parent and child nonresponse adjustments were applied sequentially to the 2003 child final longitudinal weights for the 2003 responding children.

Table 1 shows the 2004 parent and child weighted response rates, and the 2004 overall and 2002 to 2004 cumulative response rates. Both parent and child 2004 response rates are conditional rates, where the 2004 parent rate is conditioned on 2003 child response, and the 2004 child rate is conditioned on 2004 parent response.The 2004 overall response rate is the product of the 2004 parent and child conditional response rates. It can be seen from the table that the sample attrition from 2003 to 2004 was mainly due to parent nonresponse, and

that children had a fairly high response rate once their parents responded. Thus, the largest adjustment on the 2004 weights was to be made at the parent level. The child nonresponse adjustments were to be made only for those children whose parents responded in 2004.

Table 1. Youth Media Campaign Longitudinal Survey Panel 1 2004 Conditional and Cumulative response rates

Response rate	Weighted response rate (%)
2004 parent	84.4
2004 child	97.2
2004 combined parent and child	82.1
Cumulative (2002 to 2004)	31.2

Due to the longitudinal nature of the YMCLS survey, a large number of variables were available as potential predictors for modeling 2004 response propensity. The parent and child variables from the 2003 extended interviews can be used for both parent and child nonresponse adjustments in 2004, and the 2004 parent extended interviews variables can be added to that set for child nonresponse adjustments in 2004.

For modeling response propensity, the response variable was parent or child response status in 2004. The independent variables included parent attitudes towards children's activities, some critical child attitude and physical activity variables from the prior year, and some background characteristics such as geography, race/ethnicity, parent education, and family income.

In the weighting process, two methods of forming the nonresponse adjustment classes were tested: one was a tree-based algorithm (Chi-square Automatic Interaction Detector—CHAID), and the other was logistic regression modeling. The outcomes of the two methods were then compared, and one method was chosen based on the comparison. After the nonresponse adjustment, a sample-based raking adjustment was applied on the nonresponse-adjusted child weight with the purpose of "polishing" the results.

2. Nonresponse Adjustment Based on CHAID (Tree-Based Algorithms)

CHAID is a commonly used tree-based algorithm that is used to study the relationship between a dependent variable and a series of predictor variables. There are many different implementations of the original idea by Morgan and Sonquist (1963). We used the SPSS version (SPSS, 1993). A CHAID model resembles the root structure of a tree (see Figure 1) with layers of nodes and vertical connectors. At the top is a single node reflecting the entire set. Each layer below that reflects a splitting of nodes at the layer above it. CHAID modeling selects the splitting variable for each node from the set of predictors in a sequential process with independent decisions across nodes. The SPSS version we used has automatic procedures for stopping as well as GUI interface support for manual pruning and the ability to force the choice of splitting variable in selected nodes. This tree-based algorithm has a couple of salient advantages relative to logistic regression modeling. The first is that CHAID is better for modeling phenomena with deep interactions among predictor variables. The second is that there is no need to impute missing covariates. For example, in SPSS CHAID, missing values are treated the same as any legitimate reported value if the variable is unordered, and is automatically associated with that level of an ordinal variable with the closest matching success rate for the variable being modeled.

In the CHAID runs for this project, care was taken to achieve a balance between bias reduction and variance increase due to weighting. The CHAID option setup was established to ensure adequate cell sizes and to avoid very large adjustment factors. In general, the minimum cell size allowed was 30 respondents, and the maximum adjustment factor was constrained to be below 2. The option for manual pruning based on user judgment was employed.

In the parent model, 196 variables from the prior year (2003) parent and child interviews were included as potential predictors in the CHAID modeling, and 36 variables were selected by the algorithm to form 49 parent nonresponse adjustment cells. The most important characteristics used to define cells for parent response propensity were: parent education, whether or not the parent's child[*] reported running for physical activities in the past seven days back in 2003, whether or not the child's other parent was living in the household, and the child's level of organized activity in 2003. The CHAID diagram for the parent model is shown in Figure 1.

The child nonresponse adjustments were made for the child nonresponse given parent response in 2004. In the child model, 230 variables from the prior year (2003) parent and child interviews and current year (2004) parent interviews were included in the CHAID modeling, and 22 were selected to form 27 child nonresponse adjustment cells. The most important characteristics used to define cells for children response propensity were: child's age, child's level of free time activity, and how often parent set limits on the amount of time child played video games. (The CHAID diagram for the child model is not shown to save space.)

[*] If two siblings were in sample, one of the two children was randomly chosen to provide child-level 2003 predictors for parent's 2004 response status.

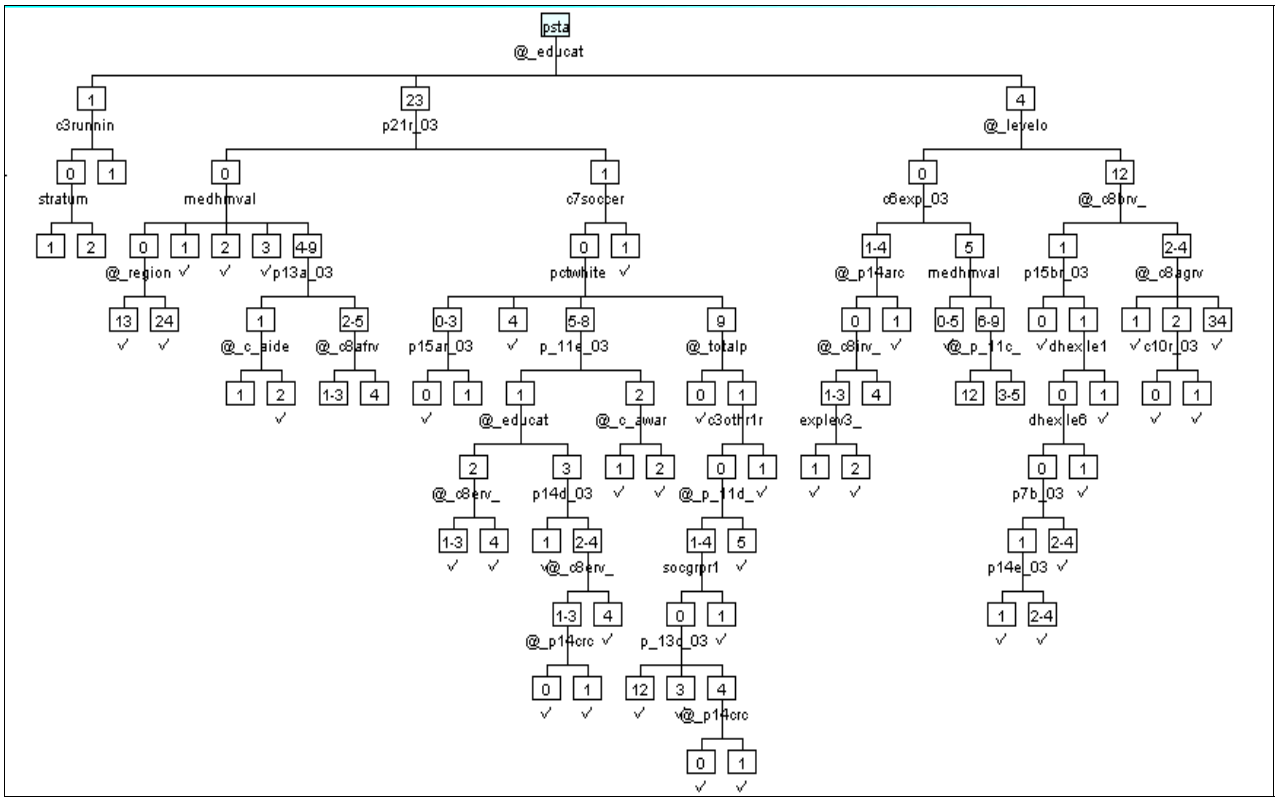

Figure 1. CHAID Tree Diagram for the Parent Model

3. Nonresponse Adjustment Using Logistic Regression Modeling

Just as tree-based procedures have a couple of widely acknowledged advantages over logistic regression procedures, there are a couple of advantages for logistic regression procedures. First, they can extract more information from continuous covariates (at least if the shape of the relationship is known). Second, they can estimate a larger number of main effects than tree-based procedures. We tried to improve these advantages by creating an algorithm that would fit richer logistic regression models.

We did this by taking advantage of recent advances in SAS such as the SQL procedure for capturing variable names from ODS model outputs (SAS, 1999). With these tools, we created a SAS macro to automatically fit a logistic model for predicting response status in terms of a very large set of covariates, including second-order polynomials and two-way interactions. This macro relies on an index file that holds information about all the variables associated with a survey. The index file records (among other attributes) for every variable whether it is stored as a character or numeric variable and whether its categories are ordered. This automatic modeling macro fits a sequence of models with SAS procedures, using ODS output and SQL along the way to build richer and richer models. The procedure is quite robust, requiring very little human intervention. This macro is thus well suited for high-throughput modeling operations. Also, as often happens in survey work, if late edits to disposition codes or survey variables are made, it is almost effortless to refit the models needed for nonresponse adjustment.

The basic steps of the macro are as follows:

1. Use a fast forward stepwise search with generous selection criteria through all eligible ordinal predictors in SAS REG to establish the most important predictors.

2. Narrow the set down by a forward step-wise search in SAS LOGISTIC among those identified by the search in SAS REG.

3. Form extra variables in SAS datasteps to define all second-order polynomials and two-way interactions for all ordinal variables that are “super” significant or are identified in advance as unusually important for theoretical reasons.

4. Perform more stepwise searches with REG and LOGISTIC among the main effects previously identified and the new polynomials and interactions.

5. Force back in main effects for significant interactions and give preference to main effects in subsequent pruning operations.

6. In parallel to these operations, use SAS LOGISTIC to conduct forward stepwise searches among unordered variables.

7. Combine best separate models involving only ordered or only unordered predictors and perform additional pruning.

8. Check to see whether these forcing and pruning operations have resulted in a model with less explanatory model than earlier simpler models, and if so, revert to earlier simpler models.

The macro can be made more or less greedy by choice of pruning parameters, taking advantage of SAS options to specify p-values for variable entry and retention in stepwise selection algorithms. The algorithm has some limitations. It is easily overwhelmed by a large number of significant unordered categorical variables. Ideally, these significant unordered variables would probably be added one at a time to the best model composed of ordered variables, but all these steps take time. If the nonresponse rate is near zero, the SAS LOGISTIC calls may have trouble converging, but this can be fixed by using Fisher's scoring on the first few iteration of each model and Newton-Raphson algorithms for subsequent iterations. Nevertheless, convergence issues are always a potential problem with using SAS LOGISTIC. On the other hand, the algorithm has some definite strengths. In particular, because of its reliance on forward searches instead of backward searches, it is not affected by colinearity among either ordered or unordered predictors. If two variables are highly collinear, the macro will simply ignore the second of the two. Theoretically, it could run into difficulties if an ordinal predictor was collinear with an unordered predictor, but this is unlikely if variables are properly classified in advance.

As when comparing any logistic regression model with tree-based models, it seems clear that the logistic procedure will extract more information from continuous covariates and shallow interactions. There is no way for it to detect three-way or higher interactions, but it can detect many more two-way interactions than a tree-based model.

The results may not be easy to interpret. Interactions may appear without main effects. If there are three collinear variables, there is no guarantee that the most parsimonious model will be selected. In fact, neither tree-based nor greedy logistic modeling procedures such as the one tested here create models that are at all easy to interpret when they are run with hundreds of eligible covariates. However, if the only purpose of the modeling is to estimate subject-level nonresponse propensities, then interpretability is clearly a secondary issue.

In this application, 235 parent and child variables from the 2003 round were included as potential predictors for predicting parental nonresponse. To these 25 parent variables from the 2004 round were added for a total of 260 for predicting child nonresponse. For the parent model (Table 2), 19 variables were in the final model, of which 6 were interactions. For the child model (not shown to save space), 10 variables were in the final model, all of which were main effects.

Once the models were formed, the issue still remained of how to use them in the nonresponse adjustments. One possibility is to set the nonresponse adjustment factor for a given individual equal to the inverse of the predicted response propensity from the model for that individual (Cassel, Särndal and Wretman (1983). However, the procedure adopted for this study was to stratify the sample by the predicted nonresponse propensities and then to calculate weighted empirical response rates within each stratum (more frequently known as a cell). This has several advantages. The first is simple but very useful. If this procedure is followed, then the sum of the nonresponse-adjusted weights for respondents will equal the sum of the incoming weights (baseweights or nonresponse-adjusted weights from prior waves) across respondents and nonrespondents. This equality means that results about large sample consistency of base weights with zero nonresponse also apply to the nonresponse adjusted weights with additional assumptions of model validity. More importantly, if the model is incorrect (as seems inevitable), this procedure is more robust as indicated by rows 11, 12, and 14 of Table V of Little and Vartivarian (2003).

The only question is how many cells to form. Literature going back to Cochran (1968) and supported by Aigner, Goldberger and Kalton (1975) suggests that five strata should be adequate for most purposes. Experiments with both five and ten strata were carried out and it was found that five strata were, indeed, preferable. Changing from 5 to 10 strata increased design effects by about 2 percent while leaving bias essentially flat.

Table 2. Logistic regression model for 2004 parental response

Effect±	Estimate		Effect	Estimate
Intercept	2.3372		Parent education	0.9334
Did an unspecified physical activity in past 7 days	0.3914		At least one session of organized activity in the previous week	-0.5772
Campaign Exposure in previous year	0.4203		Household income	-0.7729
In "Greenbelt Families" PRIZM cluster	-0.6203		In "Gray Collars" PRIZM cluster	-1.0934
In "River City USA" PRIZM cluster	1.9871		In "Single City Blues" PRIZM cluster	-1.468
Interaction: Played soccer in previous week *Agreement with "There are lots of places in my neighborhood where I can do physical activities"	1.0608		Interaction: Name of advertisement*Agreement with "I'm too busy to do more physical activities than I do"	-1.0852
Agreement with "Kids my age think physical activity is fun"	0.51		Interaction: Child lives in household with parent*Household income	0.9626
Interaction: Agreement with "There are lots of places in my neighborhood where I can do physical activities * Agreement with "I could find a physical activity to do that I enjoy"	2.478		Interaction: Agreement with "If I did physical activities on most days, it would help me spend more time with my friends" *At least one session of organized activity in previous week	0.9972
Agreement with "I could find a physical activity to do that I enjoy"	-2.1952		Squared value of Agreement with "There are lots of places in my neighborhood where I can do physical activities"	-2.0838
Total days participated in all free-time activities	-0.7585		Ran for physical activity in past 7 days	0.3868

± All measures of agreement are on five-point Likert scales.

4. Comparison: CHAID Versus Logistic Regression

There is little obvious agreement between the two models. Parental education is the first split in the CHAID parental nonresponse model and strongly significant in the logistic regression model, income is salient in both parental models, and child age is important in both child nonresponse models, but many variables appear in just one model. It is hard to get any intuitive feeling for which model is more reasonable because neither one is very transparent. The deep interactions, truncated variable names, and general density of Figure 1 makes the CHAID model hard to grasp intuitively. On the other hand, the interactions and colinearities among predictors in the logistic regression make it just as hard to grasp intuitively.

We judged the competition between the two nonresponse adjustment methods by comparing the bias, variance, and mean square error (MSE) of estimates of the 2003 population based on nonresponse-adjusted weights of 2004 respondents. To do this we tabulated 2003 variables on 2004 respondents using each alternate set of nonresponse-adjusted weights. The following 2003 variables were used in the comparisons: 1) the 2003 child outcomes, including previous week's free-time activity, organized activity, and a number of mean attitude scores; 2) demographic, socioeconomic, and geographic characteristics, such as gender, age, race/ethnicity, parental education, income, urbanicity, census region, and metropolitan status; and 3) the child outcomes by subgroups of the population characteristics. The statistics were mostly percentages, but there were also means of Likert scales and means of count variables (such as number of exercise sessions in week). The full set of comparisons for the whole survey population and for the subgroups resulted in a total of 692 comparisons of bias, variance, and MSE.

For the bias comparisons, the biases using 2004 nonresponse adjusted longitudinal weights based on each method were estimated as

$$bias = \hat{y}_{04} - \hat{y}_{03}$$

where $\hat{y}_{04}$ were the estimates of the 2003 outcome using the 2004 nonresponse adjusted longitudinal weights for 2004 respondents, and $\hat{y}_{03}$ were the estimates of the 2003 outcome using the 2003 final longitudinal weights for 2003 respondents.

For the variance comparisons, the variances were estimated using jackknife replication, and variance ratios were computed as the variances for the 2004 estimates over variances for the 2003 estimates:

$$\text{variance ratio} = \frac{\text{var}(\hat{y}_{04})}{\text{var}(\hat{y}_{03})}$$

The mean square errors for the 2003 estimates based on the 2004 respondents were estimated by

$$MSE = bias^2 + \text{var}(\hat{y}_{04})$$

For most of these estimates, the magnitude of the variance term was substantially larger than that of the bias term. Thus most mean square errors were dominated by the variance component.

Figure 2 is a scatter plot of estimated biases using the two nonresponse adjustment methods for the 692 estimates.[†] It can be seen that most of the points are closely clustered around the diagonal line, which suggests that the magnitudes of the biases in the 2004-based estimates using the two methods are very similar. The standard error scatter plot in Figure 3 shows a similar pattern with even smaller differences between the two methods. As a result, the mean square errors plotted in Figure 4 also have a similar pattern.

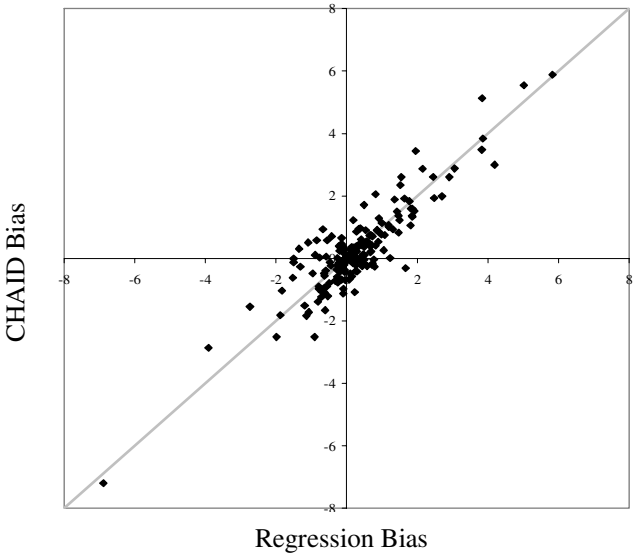

Figure 2. Scatter plot of biases: CHAID versus logistic regression weighting adjustments

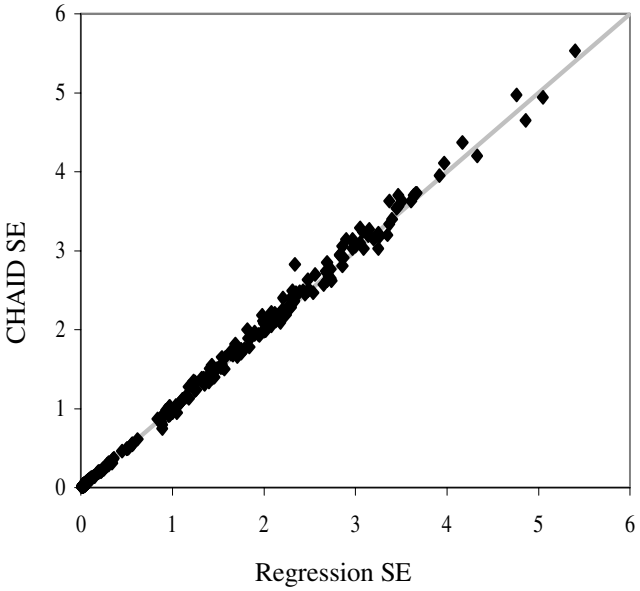

Figure 3. Scatter plot of standard errors CHAID versus logistic regression weighting adjustments

[†] Because different types of statistics were included in the evaluation, it is difficult to define the axes for this chart or to average the biases together in any meaningful way, but the point remains that biases are very similar for the two methods.

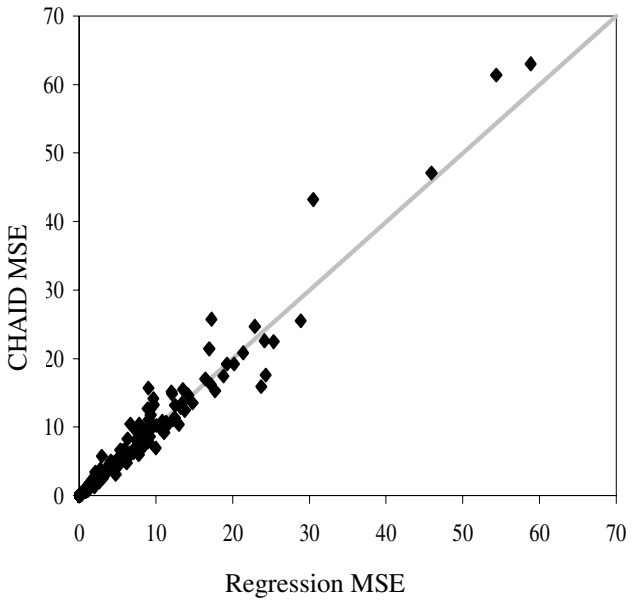

Figure 4. Scatter plot of MSEs: CHAID versus logistic regression weighting adjustments

To examine which method did better in terms of the numbers of comparisons, the number of occurrences that each child nonresponse-adjustment method had less absolute bias, variance, or MSE than the other method was counted. The results are summarized in Table 3. It can be seen from the table that CHAID won in more comparisons than the logistic regression for bias and MSE while logistic regression won on variance. The margins of wins were, however, modest. Moreover, there were many ties,[‡] and as shown in Figures 2 through 4, differences were generally very small. (Average biases, variances, and MSEs would not be meaningful given variety of statistics in comparison.)

Table 3. Number of occurrences that each method performed better than the other method

Method with better outcome	Absolute bias	Standard error	MSE
Logistic regression	242	142	256
CHAID	311	113	314
Tied	139	437	122
Total comparisons	692	692	692

5. Sample-Based Raking Adjustment

Although either logistic regression or tree-based methods should remove almost all of the bias that is possible to remove by conditioning on measured covariates, it may still happen that there are particularly sensitive variables in a longitudinal survey, for which it is highly desirable to hold estimates constant across waves. Such control may improve face validity, efficiency or both. Sample-based raking is an ideal tool for that purpose. Whether efficiency is improved by such raking will, of course, depend on the extent that these variables are related to prime outcomes of interest. If these variables are only weakly correlated to variables of interest, the sample based raking is likely to increase variances. It is possible to consider using sample-based raking as an alternative to the modeling procedures considered here, as done by Folsom (1991), Folsom and Witt (1994), Rizzo, Kalton and Brick (1996), and Folsom and Singh (2000). However, in this paper, sample-based raking is only considered as a polishing tool. This is because it would not be possible to consider raking on the hundreds of dimensions available on the YMCLS. Some tool had to be used first to extract the important information from the high dimensional set of covariates before raking could be fruitfully applied.

In the YMCLS application, preliminary 2004 nonresponse-adjusted weights were raked to 11-dimensional control totals from the 2003 survey. The variables that defined these dimensions consisted of 2003 outcome measurements (related to physical activity of child), child demographics (age, sex, race, and ethnicity), parental socioeconomic status (education and income), and geography (urbanicity and region) As a result of the raking, variances of totals decreased for 12 of 20 key outcomes, with a mean decrease in variance for these 20 outcomes of 4 percent. Interestingly, some design effects became less than 1.

6. Discussion and Summary

As in the previous work of Rizzo et al (1996) and Folsom and Witt (1994), no clear winner emerged from this evaluation of competing algorithms for modeling nonresponse propensity. We had anticipated that the new logistic regression macro with its ability to automatically use hundred of covariates to form rich models would show a decisive advantage over the tree-based approach. However, the results of this evaluation indicate that the tree-based method is as good as, if not better than, logistic regression with automated second-order searches. We also confirmed that sample-based raking is a useful tool for polishing results.

Whether these findings will carry over to future studies probably depends on the importance of continuous covariates in the nonresponse model and the depth of interactions. Given the closeness of the evaluation results, operational issues may become decisive. A fair amount of effort was required to create the greedy second-order logistic regression macro. The SPSS CHAID procedure also required substantial effort to use in a SAS environment. Long variable names and variable with embedded special characters posed special problems for the

[‡] Ties were to two decimal places for bias and standard error and to four decimal places for MSE.

integration. Whether prior round data have already been imputed is also an important issue since the logistic regression procedure will not consider any covariate with missing data while the CHAID procedure easily makes use of partially reported items from prior rounds.

7. References

Aigner, D.J., Goldberger, A.S. and Kalton, G. (1975). On the explanatory power of dummy variable regressions. *International Economic Review*, 16, 503-509.

Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem, *In Incomplete Data in Sample Surveys, Vol. III*: *Symposium on Incomplete Data, Proceedings* (W.G. Madoro and I. Olkin, Eds.) New York: Academic Press.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 295-313.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the Social Statistics Section of the American Statistical Association*, pp. 197-202.

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 of the American Statistical Association*, pp. 428-433.

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 of the American Statistical Association*, pp. 598-603.

Göksel, H., Judkins, D. R., and Mosher, W. D. (1992). Nonresponse adjustments for a telephone follow-up to a national in-person survey. *Journal of Official Statistics*, 8, 417-433.

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

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

Morgan, J.N. and Sonquist, J.A. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association*, **58**, 415-435.

Potter LD, Duke JC, Nolin MJ, Judkins D, Huhman M. *Evaluation of the CDC VERB Campaign: findings from the Youth Media Campaign Longitudinal Survey, 2002-2003*. Westat: Rockville, Maryland, 2004.

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

SAS (1999). *SAS OnlineDoc®, Version 8*, Cary, NC: SAS Institute Inc.

SPSS (1993). *SPSS for Windows CHAID Release 6.0.* Chicago: Statistical Innovations.