Adjustments to Student Weights to Account for Student Nonresponse in the National Assessment of Educational Progress

John Riddles[1], John Burke[1], Gonzalo Rivero[1], Keith Rust[1]
[1]Westat, 1600 Research Blvd., Rockville, MD 20850

Abstract

The National Assessment of Educational Progress (NAEP) conducts regular assessments in mathematics and reading among samples of students in grades 4, 8, and 12. Since the resulting missing data are not missing at random, nonresponse adjustments are made via a process of weighting class adjustments. Using 2015 NAEP data we investigated whether improvements could be made to the process, analyzing the relationship of student and school characteristics to both student achievement and nonresponse. The aim was to establish an effective procedure for creating weighting classes that could be used for future assessment cycles. We determined the appropriate variables and their relevant grouping and cut points through the use of conditional inference trees, creating the trees using recursive partitioning with inference-based splits and unbiased variable selection. We describe the approaches used and illustrate the procedures by showing some of our findings.

Key Words: Survey weights; non-ignorable nonresponse; model assessment; conditional inference trees; decision trees

1. Introduction

The National Assessment of Educational Progress (NAEP), conducted by the National Center for Education Statistics, is an annual survey of the knowledge and skills of U.S. students, and has been conducted since the early 1970s. Students are assessed at grades 4, 8, and 12 (although not ever grade is assessed each year). The subjects assessed vary by year, but include reading, mathematics, science, writing, U.S. History, Civics, Geography, Economics, Technology and Engineering Literacy, Music, and Visual Art. Since 1990, in alternate years (currently odd-numbered years) assessments in reading and mathematics are conducted at grades 4 and 8, with sample sizes large enough to yield estimates for public schools at the level of state, and for certain large urban school districts. The total sample sizes for such assessments are of the order of 150,000 students for each grade and subject.

NAEP uses a two-stage sample design, selecting schools and then students within each selected school. Nonresponse can occur at both the school and student levels, and weighting adjustments are made to compensate for this nonresponse. Student nonresponse occurs in all assessments, but is typically quite low, being of the order of 6 percent at grade 4, 8 percent at grade 8, and 15 percent at grade 12. Nevertheless, it is well-recognized that students are not missing at random, and hence the nonresponse adjustments applied have the potential to reduce the student nonresponse bias that would otherwise be present.

Student nonresponse adjustments are performed by creating weighting classes, and inflating the weights of the responding students in each class so that the sum of the survey weights of the responding students in the class is equal to the sum of the original weights of all of the eligible students in the class. Hence, the key to reducing nonresponse bias is the effectiveness of creating weighting classes using variables that are related both to student achievement and to the likelihood of responding. These classes are routinely created each year, using variables that are known to be related to achievement and are believed, based on historical data, to be related to the propensity to respond.

However, for many years there has been no evaluation of whether there are better choices of variables for creating weighting classes, or the way in which they are combined for this purpose. While regularly conducted nonresponse bias analyses consistently demonstrate that the current approach is somewhat effective in reducing nonresponse bias, it is not clear that the current procedure remains optimal for this purpose.

Therefore, we have embarked upon a research project to investigate, from the ground up, what are the most promising variables for use in creating effective nonresponse adjustment classes and how those classes should be formed. In addition to a limited number of individual student characteristics, a sizeable number of school characteristics, for the school attended by each sampled student, are available. We are utilizing data from the 2015 assessments in mathematics and reading for this research. These data contain not only the student and school characteristics of the responding and nonresponding students, but also the achievement data, in mathematics or reading, for the responding students. Thus these data can be used to investigate both which variables are related to student achievement (among respondents), as well as which are related to the propensity to respond.

Both test scores and nonresponse were modeled using conditional inference trees (Hothorn 2006), which are decision trees that are grown more conservatively than most other decision tree models, and more importantly, provide unbiased variable selection. The purpose of the test score model was to determine which variables are related to student achievement. In turn, the significant variables are treated as candidates in the nonresponse model. The use of conditional inference trees to model nonresponse was previously investigated by Lohr et al but was limited to a simulation study (Lohr 2015). Here we wish to examine how well it can be applied in a more practical situation.

2. Preliminaries

The 2015 assessments include tests scores in Reading and Math for grades 4, 8, and 12. Public schools are grouped by state, and other schools are grouped into four categories: Department of Defense Education Activity, Bureau of Indian Education, Catholic, and non-Catholic private. School-level characteristics include the median household income for the zip code in which the school resides, the racial demographics of the school, and the urbanicity of the school. Also, for public schools, there is a charter school indicator variable. Student-level characteristics include age, gender, race, disability status, and status as an English language learner.

Some degree of preprocessing was needed to reduce potentially spurious information in the data. For one, the urbanicity variable includes thirteen categories. Excluding one category, DoD International, these can be grouped into four larger categories: city, suburb, town, and rural, each containing three sub-categories based on size, e.g., large city, midsize city, and small city. For each of these four categories, we examined whether or not it was worthwhile to collapse the sub-categories. This was done by comparing the test scores between the sub-categories using an appropriate test. If each sub-category had at least ten observations, they were compared using an F-test; otherwise, a permutation test was performed. In both cases, sub-categories were combined if the p-value was greater than .05.

Another variable that we examined was the age of the student in months at the time the test was taken. This is a discrete interval variable, which we further discretized into two categories: younger vs. older. The cutoff used to determine which of the two categories a student is a member of was found by fitting a conditional inference tree that regresses age on test. This is equivalent to selecting the cutoff which minimizes the p-value of a permutation test for comparing the younger vs older students in terms of test scores.

3. Methodology

In this study, we created nonresponse adjustment classes through the application of a conditional inference tree learning algorithm. Candidate variables were pre-selected by examining their ability to predict student achievement. Class imbalance was addressed via a two-step procedure where we fit a conditional inference tree on a balanced sampled to determine the adjustment classes, followed by an estimation of propensities within each class using the original unbalanced sample.

3.1 Conditional Inference Trees for Nonresponse Modeling

Decision tree learning is a popular methodology in both regression and classification and is known to be reasonably flexible, providing good predictive power in many situations. In contrast to many “black box” methods that are popular in the machine learning community, they are generally considered to be highly interpretable, which was a driving factor in our decision to use decision trees to model nonresponse. In particular, the resulting model consists of a number of classes defined by various cutoffs points across numerical covariates as well membership in categories for categorical variables, and in most cases, predictions are taken to be constant within each class. Trees are fit by recursively splitting across covariates, maximizing a measure of the difference in responses between the resulting classes. The method was originally developed to tame the combinatorial explosion of possible interaction effects when adding variables to a model (Morgan 1963). Different tree learning algorithms are defined by how they measure differences in response, whether they split on one variable at a time versus multiple variables, the handling of missing data, stopping criteria, etc.

The most popular decision tree learning algorithms are CHAID (Magidson 1994) and C4.5 (Quinlan 1996). However, these algorithms, along with most decision tree learning algorithms, have a number of drawbacks. For one, there is a tendency toward overfitting, where spurious trends are fit in the lower levels of the tree resulting in large trees with high variance. To manage this issue, pruning algorithms are often applied to produce smaller trees that represent the salient features in the data model while hopefully excluding spurious trends. Another common drawback of many decision tree algorithms is variable selection bias. Since trees seek to maximize the difference in responses between classes by examining all possible splits, they favor variables for which there are many ways to split unless the specific learning algorithm controls for this. These variables mainly include categorical variables with many categories. Additionally, if missingness is treated as an additional category in lieu of imputation or case deletion, the algorithms will sometimes favor covariates with more missing values.

Conditional inference trees were introduced by Hothorn et al in an attempt to address these concerns (Hothorn 2006). The first step of the algorithm is to perform an omnibus test that is the basis for the algorithm's stopping criterion. This tests against the null hypothesis that the covariates are independent of the response conditional on the currently grown tree, and if the p-value is above a user-specified threshold, no further splits are made. If the p-value is small, the variable with the strongest association with the response is selected. Split points are determined by maximizing the test statistic of a permutation test.

The omnibus test is a built-in stopping criterion, and when properly applied, precludes the need for pruning, and the use of permutation tests provides unbiased variable selection. There are other candidate tree algorithms that provide the same or similar guarantees. The most comparable algorithm is GUIDE (Loh 2002), which also provides unbiased variable selection and early stopping. The primary difference is its use of chi-square tests instead of permutation tests. However, simulations performed by other authors provide good indication that conditional inference trees offer superior predictive performance on average (Hothorn 2006).

3.1.1 Conditional Inference Tree Implementations

Analyses were performed using the R statistical language/environment. There are multiple packages available for R that implement conditional inference trees, the first of which was `party`, which was followed by `partykit` (Hothorn 2015). A newer package, `rpms` (Toth 2017), provides an implementation that accounts for complex survey designs. Unlike `party` and `partykit`, `rpms` supports probability-weighted samples. Although the feature was not used in our study, `rpms` also allows the user to fit regressions on leaf nodes, whereas `party` and `partykit` assume constant predictions within leaf nodes. Yet `party` and `partykit` do offer a few advantages. These packages support regression and classification, whereas `rpms` only supports regression. Further, `party` and `partykit` handle missing covariates via surrogate splits, where cases with missing values for a covariate which is split upon are classified by examining another covariate, the surrogate. In `rpms`, cases with missing covariates are deleted, which reduces the effective sample size and may lead to bias. None of these packages handle missing response variables, nor do they allow user-specified loss functions.

In this study, we used the `rpms` package primarily because it supports probability-weighted samples, which was a key component in handling the class imbalance problem that will be discussed in Section 3.3.

3.2 Variable Selection

When modeling nonresponse, it is generally important to exclude variables that are independent of the response variable (test scores in our case). Otherwise, their inclusion will inflate the variance of the nonresponse-adjusted model predictions without a corresponding reduction in bias. As such, we first sought to determine which variables were notably predictive of student test scores. As with nonresponse, test scores were modeled using conditional inference trees, and variables that were not selected by the algorithm for the test score model were necessarily omitted from the nonresponse model. Due to expectations from prior experience, exceptions were made for gender and student race, which were always allowed entry into the nonresponse model, although the tree algorithm associated with the nonresponse model was allowed to throw these variables out if it was determined that they were uninformative with respect to response propensity.

It should be noted that we were not overly concerned about the ability of this model to predict test scores. Our goal here was simply to determine, within a consistent framework, which variables have a significant relationship with student achievement. One reason we used decision trees for the test score model was that it allows us to detect variables which may have a strong interaction effect but a weak main effect. Additionally, decision trees tend to exclude covariates that are confounded with other covariates. In general, this exclusion is desired, as it reduces variance without a large increase in bias. However, there may be cases where two covariates are confounded with respect to test scores but not with respect to the response indicator, in which case the variable would be unfairly excluded by our procedure. For this reason, in future work we will consider examining covariates in isolation from other covariates.

The test score model was fit only on students who responded. Since the data was not missing at random, this introduced some bias in test score estimation, but we were hesitant to use imputation lest we introduce trends that are merely relics of our imputation methodology, and since our purpose here was variable selection, bias in estimating test scores was not a major concern.

As an additional step, we considered using a variable selection procedure to filter out highly uninformative variables before fitting the response model. We used random forests, which are commonly used for this purpose (Breiman 2001, Genuer 2010). However, we found that their use had little impact on results, as they would generally omit the same variables that would have been omitted by the conditional inference tree anyway, and so they were not used in the final model. However, this may be because we used CART-based random forests. In the future, it may be worth evaluating the utility of random forests based on conditional inference trees (cforests) for the purpose of variable selection.

3.3 Addressing Class Imbalance

Student response rates are very high, over 90% in most cases. That is, we have a highly imbalanced response indicator variable. In such cases, models often aggressively try to correctly predict the majority class to maximize accuracy at the cost of a high misclassification rate for the minority class. This is known as the “class imbalance” problem. If one is merely interested in overall accuracy, then this “problem” is not actually an issue. A model may tend toward a constant prediction, placing all respondents into the majority class, but this simply indicates that a constant prediction provides the highest accuracy given the constraints of the model. Either this accuracy suffices for one’s purposes, or it serves as indication that an alternative model should be considered. However, in our case, this does not complement our goal of minimizing nonresponse bias, so the class imbalance problem was addressed.

Class imbalance may be addressed in one of several ways. Oversampling the minority class or undersampling the majority class are commonly used techniques. These techniques are unfortunately problematic for conditional inference trees since they affect the test statistics, and current implementations do not adjust for the artificial increase/decrease in sample size, which in turn affects the significance tests used by the conditional inference tree algorithm. It may be possible to correct this bias by adjusting the significance level of the test, but we have found that it is very difficult to fine-tune, and it may introduce some numerical instability.

Another option is to use a loss function that accounts for class imbalance. Common choices are balanced accuracy and the F1 score. Given time constraints, we were not able to implement a conditional inference tree under these loss functions, but it is certainly an option for future work.

Yet another option is to apply weights to our sample in a manner that balances the two classes but does not affect the sample size with respect to the significance tests. We chose this option, but one drawback is that it leads to propensity estimates that are biased toward .5, which is to be expected under a sample weighted in this manner. To address this issue, we used a weighted sample to build a tree and attain adjustment classes and then used an unweighted sample to estimate the propensity within each adjustment class. This allows us to properly capture features of the minority class, non-respondents in our case, without strongly biasing our propensity estimates.

3.4 Clustering of States

Initial results from the nonresponse model were not promising. Separate models were fit for each state and private school class, and the vast majority of them were constant, i.e., no variables were determined to be significantly related to nonresponse. One possible solution was to instead fit a single model instead of separate models for each, thus increasing the effective sample size for the model. However, we did not necessarily want to preclude the use of state-level information, so the state variable was included as a covariate in the model. Unfortunately, it was computationally infeasible to get a model fit, likely due to the great number of possible ways to split on the state variable. As an alternative, we modeled nonresponse by fitting a conditional inference tree where we regressed the state variable on the response indicator. There were still just as many ways to split on that variable, but the regression was not complicated by other covariates. The resulting cells were clusters of states. This state cluster variable was then used as a covariate in the nonresponse model that includes all other covariates. This reduced the number of categories for state in the model, which in turn allowed us to fit the model in a reasonable amount of time.

Separate models were fit for the six state-subject combinations. For each, between 3 and 10 state clusters were identified.

4. Results

In this section, we examine which variables are key with respect to student achievement and nonresponse. Additionally, we evaluate the effect our nonresponse model has on sample balance in lieu of examining error in propensity estimation, since we cannot directly observe propensities.

4.1 Test Score Model Evaluation

Although predictive power of our test score model was not our goal, we did attempt to verify the reasonableness of the test score model. Figure 1 compares actual vs. predicted test scores for our model. For this figure, separate models are fit for each state (for public schools) and private school classification. The plots show that the models seem to slightly overestimate low scores and underestimate high scores, but overall it indicates that the models are reasonable enough for our purpose.

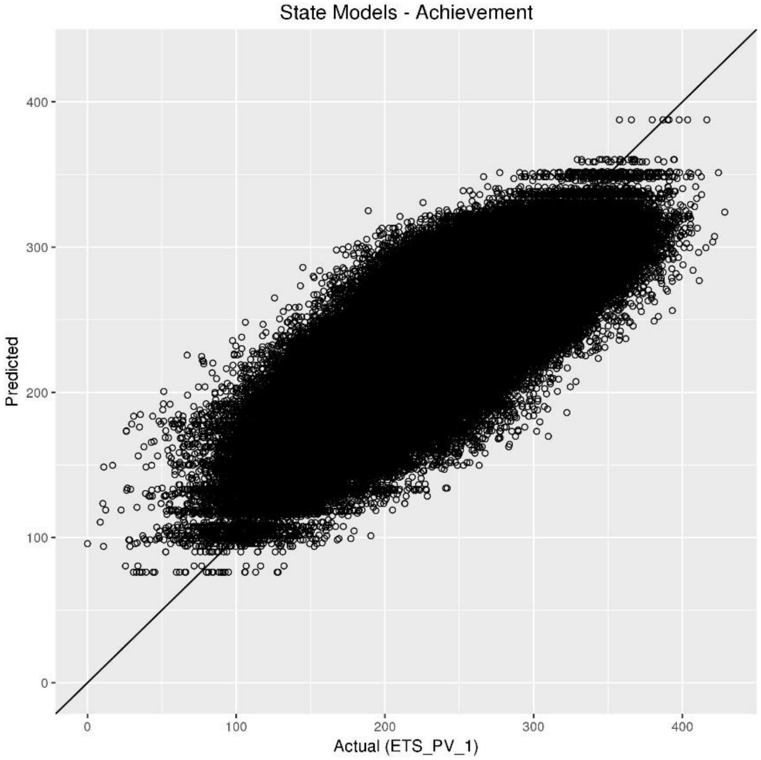

Figure 1: Actual vs. predicted test scores

We also wanted to examine the degree of consistency in which variables were retained by the test score model. However, with only six models, this was difficult to examine. One possible solution is resampling. However, a simpler alternative is to fit separate models for each state and examine how often each variable is kept by the models. Unlike the nonresponse models, which were often constant when separate models were fit for each state, reasonably sized models for test scores were consistently found across states. Upon examination, we discovered that all student-level characteristics, with the exception of age, were included in every test score model, indicating that they are significant predictors of test scores. Student age was retained in 84% percent of models. Median income and urbanicity were each retained in 68% of models, not necessarily together. Individual racial demographics (at the school level) were retained 50-60% of the time, although in every case at least on racial demographic variable was retained. Charter status was only included for 60% of models for public schools.

4.2 Nonresponse Model Evaluation

As previously mentioned, we fit six separate nonresponse models, one for each grade/subject combination. To determine which variables are key in predicting nonresponse, we examined which variables were included in each model, as well as which variable appeared at the root of the tree. Table 1 shows these results. The Racial Demographics row combines all of the school-level racial demographic variables. Variables that do not appear in this table, e.g., gender, were not selected by any of the six models.

Table 1. Key Variables in Nonresponse Models

Variable	**# of models**	**# of times as root**
State Cluster	6	5
Disability Status	6	1
Racial Demographics	5	0
Student Race	2	0
ELL Status	1	0
Median Income	1	0

One can see that state, disability status, and racial variables are by far the most relevant variables by this measure. The trees were all fairly small, including only a handful of variables, so we felt assured that we likely did not overfit the data, although underfitting is a possibility.

We also examined how well propensity adjustments using inverse propensity weighting balance different auxiliary variables between respondents and the full sample. If they are balanced, this does not guarantee that we have corrected for bias with respect to test scores. Regardless, we feel that it can give use some clue as to the potential validity of our methodology. For example, suppose that the ratio of males to females in the full sample is .5, but the ratio among respondents is .8. A propensity method that properly accounts for this imbalance should estimate higher propensities for males than females on average. Under inverse propensity weighting, the males who responded will receive less weight than the females who responded, balancing the ratio toward .5. If the propensity-adjusted gender ratio among respondents deviates significantly from that, this may indicate some remaining bias in the estimation of test scores if gender is a confounding variable.

Table 2 shows the ratio of propensity-adjusted means for respondents over the means for the full sample. In this table, the tree-based weighting classes are compared against previously used weighting classes. Ratios closer to one than in the alternative class adjustment are bolded.

Table 2. Examination of Sample Balance

Variable	Cond. Inference Trees	Prev. Adjustment Classes
Median Income	**0.99**	0.97
% Native American	**0.98**	0.97
% Asian	**0.95**	0.94
% African-American	**0.97**	0.96
% Pacific Islander	0.98	0.98
% Hispanic	1.01	1.01
% 2 or more races	**1.00**	0.99

There does not seem to be a huge difference in this respect between the two sets of adjustment classes, but conditional inference trees are at least comparable to the previously used classes and may be slightly better. Similar statistics were calculated for each level of each categorical variable. They are too numerous to list here, but similar results can be found: our method provided slightly better balance, though the difference was generally not large. This indicates a modest improvement overall. One exception to this is student race, which is less balanced under our methodology. This is a point of concern and warrants further investigation.

5. Summary

In this study, we revisited the creation of weighting classes for student nonresponse in the National Assessment of Educational Progress. Decision trees were selected due to their high interpretability and their ability to capture the most salient interaction effects. More specifically, a conditional inference tree algorithm was used to determine weighting classes due to a number of desirable features, particularly unbiased variable selection. Prior to modeling nonresponse, we modeled student achievement in order to determine what variables are predictive of test scores. Class imbalance was addressed via the use of probability weighting during the estimation of nonresponse adjustment cells but an unweighted sample during propensity estimation within classes. States were clustered and then used as a predictor. The clustering noticeably improved the models over fitting separate models to each state or ignoring the state variable.

Results indicate the nonresponse model shows very modest improvement over previously used nonresponse adjustment classes when examining balance across covariates. Key variables include disability status, school-level racial variables, and state cluster.

6. Future Work

While our results do indicate some modest improvement over previously used weighting classes, there is still certainly room for improvement. Firstly, the stopping criteria for the conditional inference trees needs more examination. In particular, we may attempt to optimize the alpha level of the permutation tests used with the conditional inference trees. We took special care to prevent overfitting; however, is our suspicion that we may, in fact, be underfitting, and it would be worth investigating the impact of higher alpha values on predictive power. Alternatively, it is worth investigating other possible stopping criteria and splitting criteria. The use of p-values provides unbiased variable selection, but it is not statistical significance itself that we are concerned about. Rather, we are concerned about the reduction in nonresponse bias. This cannot be directly measured; however, a more direct measure of predictive power would serve as a preferable surrogate over statistical significance. For stopping criteria, cross-validation could be used to adjust the alpha parameter to maximize predictive power under the constraints of the model, but this is a computationally intensive process, and more direct alternatives may exist.

Although we saw improvements in sample balance on average, previously used adjustment classes better balanced student racial characteristics than did the methodology described herein, which is concerning for reasons previously described. To address this, we may force racial characteristics into the model or consider an alternative model, possibly one that directly addresses sample balance, such as in propensity matching methods.

State-level information is not available in all years, so a single model that would exclude state clusters will need to be developed. Since the state cluster was a key variable in our model, special care will be needed to develop a reasonable model in its absence.

In the future, variance estimation will be performed, and resampling methods will be used to evaluate the stability of the models.

Acknowledgements

We are grateful to our discussant, Lynne Stokes, for her very insightful comments on our work.

References

Breiman, Leo (2001). "Random forests." Machine learning 45.1: 5-32.

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot (2010). "Variable selection using random forests." Pattern Recognition Letters 31.14: 2225-2236.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis (2006). "Unbiased recursive partitioning: A conditional inference framework." Journal of Computational and Graphical statistics 15.3: 651-674.

Hothorn, Torsten, Achim Zeileis (2015). “partykit: A Modular Toolkit for Recursive Partytioning in R.” Journal of Machine Learning Research, 16, 3905-3909.

Loh, Wei-Yin (2002). "Regression tress with unbiased variable selection and interaction detection." Statistica Sinica: 361-386.

Lohr, Sharon, Valerie Hsu, and Jill Montaquila (2015). "Using Classification and Regression Trees to Model Survey Nonresponse." Proceedings of the Survey Research Methods Section.

Magidson, Jay (1994). "The chaid approach to segmentation modeling: Chi-squared automatic interaction detection." Advanced methods of marketing research: 118-159.

Morgan, James N., and John A. Sonquist (1963). "Problems in the analysis of survey data, and a proposal." Journal of the American statistical association 58.302: 415-434.

Quinlan, J. Ross (1996). "Bagging, boosting, and C4. 5." AAAI/IAAI, Vol. 1.

Toth, Daniell (2017). “rpms: Recursive Partitioning for Modeling Survey Data.” R package version 0.2.1. https://CRAN.R-project.org/package=rpms