Weight Adjustment Methods using Multilevel Propensity Models and Random Forests

Ronaldo Iachan[1], Maria Prosviryakova[1], Kurt Peters[2], Lauren Restivo[1]
[1]ICF International, 530 Gaither Road Suite 500, Rockville, MD 20850
[2]ICF International, 126 College Street Suite 2, Burlington, VT 05401

Abstract
Using two national surveys as an example, this paper focuses on the selection of predictors for non-response analysis and weight adjustments. Weighting classes for non-response adjustments are formed using a set of core variables that are correlated with response behavior and with survey outcomes. The intent is to minimize the potential of non-response bias by balancing it against acceptable increases in weighting variances due to weighting variability. The choice of variables to use when defining weighting classes is determined by measures of variable importance for predicting response. This paper compares two methods for variable selection and weighting class adjustments. One method uses response propensity models estimated with mixed effects logistic regression. The other method uses recursive partitioning, and more specifically, Random Forest algorithms. Mixed effects models are run using SAS Proc Glimmix that allows the modeling of random effects. This paper assesses the methods for evaluating variable importance and the resulting bias reduction achieved through non-response weighting adjustments.

Key Words: non-response analysis, mixed effect models, multilevel models, Random Forest, Glimmix, weighting classes

1. Introduction

Limiting both time and cost concerns, the use of multistage modeling has become a staple in large quantitative research projects (Khan & Shaw, 2011). These surveys necessarily suffer from the potential bias due to unit non-response, a bias potential that seems of increasing concern as response rates have declined for most studies (Borgoni & Berrington, 2011; Brick, 2013; Brick & Montaquila, 2009). One way to reduce the bias potential is to adjust for nonresponse using weighting class methods (Khan & Shaw, 2011; Zhu, 2014). The aim of this paper is to describe and compare two approaches for developing non-response weight adjustment classes and computing weight adjustments using two national studies as examples. These two families of methods were also examined in Iachan, Harding & Peters (2014). The first variable selection method uses multilevel logistic regression models for response indicators, and the second method uses recursive partitioning methods (e.g., Random Forests). This paper also discusses the choice of variables for weighting class adjustments and illustrates this process using the two study examples.

1.1 Overview of the Studies

1.1.1 A health study examining patients with HIV

This health study seeks to obtain information on patients with HIV in order to gain knowledge about their needs and experiences. It utilizes a cross-sectional, multistage structure involving a three-stage sampling design. The first stage of the sample selects primary sampling units (PSUs) defined as states and municipalities and referred to as "project areas". The second stage involves sampling medical facilities, the secondary sampling units (SSUs), from the project areas. Finally, the last stage involves sampling patients from these medical facilities in stage 2.

At the project area level, the design is considered to be a two-stage sampling design with the selection of patients and facilities at the two stages. Basic demographic information on all sampled patients (e.g. age, race, gender, etc.) and facility level data (e.g. facility type and size) were collected. Interview data collected for all selected patients were augmented by chart abstractions and facility level data.

1.1.2 A school based study on risk behaviors

The second example study is a national school survey which seeks to obtain information from students on a number of health indicators. A three-stage sampling design is utilized. First, the sampling frame is structured by units, also known as PSUs, defined by county or group of contiguous counties, and stratified by minority composition (percent Black and percent Hispanic). The second stage of the sampling selects schools grouped by size (small or large), and the third stage sample selects school classes at random. Finally, all students in the selected classrooms are invited to participate in the study. This school study also utilized survey questionnaires as well as a wide range of school characteristics.

As seen below, response rates vary by demographic groups in both surveys. It should be noted that for the patient health survey, response rates may be lower if individuals believe the questions are intrusive or sensitive (Borgoni & Berrington, 2013). This survey asks questions that can be seen as invasive including questions about injection drug use and sexual behavior.

2. Methods

Next we describe two broad families of methods that can be used to select predictors for weighting class non-response adjustments. Both methods seek to choose variables that are correlated with response behavior and survey outcome within classes.

2.1 Multilevel logistic regression models to form weighting classes

Multistage cluster samples allows researchers to make inferences about place and individual-level factors across a large geographic area or PSU, while reducing the time and costs of data collection. Multilevel logistic regression models take into account the hierarchical structure of multilevel sampling data by accounting for group level and individual level data across the different levels (Khan & Shaw, 2011; Rabe-Hesketh, Skrondal, & Zheng, 2007; Larsen & Merlo, 2005).Thus, when the data are clustered using different levels or stages, single-level statistical models are no longer viable and a more complex modeling technique is required to develop weighting classes.

2.1 Recursive partitioning method: Random Forest

Recursive partitioning methods can also be used to select the most significant predictors for weighting class non-response adjustments. With these methods, predictor variables are split into boxes or regions in which variables with similar response values are grouped together (Strobl, Malley, & Tutz, 2009). Variables are selected for splitting based on impurity reduction or choosing the variable that has the strongest association with the response variable for the next split (Strobl et al., 2009). We used a random forest algorithm available in R ('RandomForest' package).

3. Analysis Methods

Our non-response analysis can be distilled into the following seven steps performed for the two studies.

- Step 1 involves choosing a set of variables that may be potentially related to non-response propensities for the interview data.
- Step 2 identifies a subset of variables that satisfies two criteria: (a) they do not have much missing data (25% or less) and (b) they are associated with the response indicator in bivariate analysis.
- Step 3 runs Recursive Partitioning Algorithm (RPA) using the subset of variables defined after steps 2a and 2b.
- Step 4 examines the pairwise correlations between the variables in the subset defined in Step 2; for each pair of correlated variables, delete one of the variables in the pair. Expect to end up with 10-15 variables to be used in the multivariate multilevel logistic models in the next step.
- Step 5 fits a multilevel (random effects) model for the response indicator Y (0-1) with Proc Glimmix.
- Step 6 is to form non-response weight adjustment classes using the results of the RPA in steps 3, on one hand, and those of the steps 5 models, on the other hand.
- Finally, the last step involves evaluating the results from the two methods, the multilevel model and the Random Forest algorithm.. We evaluated model performances and compared the most significant variables identified with the two methods.

For the patient health survey, the multilevel model in Step 5 has patients as level 1 and facilities as level 2 units. For the school survey, levels may include students, schools, and PSUs.

The ideal weighting classes in Step 6 are homogeneous in terms of response behavior and outcomes. When constructing weighting classes, it is important to reduce dimensionality for reasons that go beyond model parsimony such as to avoid over-fitting. One should avoid too many weighting cells that may lead to empty or sparse cells (few respondents or non-respondents in the cell) and to limit the variance because too many cells that inflate variances.

4. Results

4.1 HIV-multilevel logistic regression model for health study

We first looked at variables that may be potentially related to non-response propensities for the interview data. Table 1 below shows the potential variables available for respondents and non-respondents. Recall, a collection of data was gathered from a sample of patients from different facilities in the selected project areas or states including basic demographic information as well as information about the facility that they received care from.

Table 1: Potential variables related to patient non-response

#	Variable	Label
1	NRTYPE	Facility Type
2	NREPL	Facility Size
3	NRAGE	Age Group
4	NRGENDER	Gender at birth
5	NRRACE	Race/Ethnicity
6	NRYEAR	Years Post Diagnosis Groups
7	birth_country	Country of Birth
8	dx_status	Diagnostic status (calculated)
9	aids_insurance	Primary reimbursement for medical treatment (AIDS)
10	hiv_insurance	Primary reimbursement for medical treatment (HIV)
11	c_art	Patient received anti-retroviral therapy
12	idu	Adult IDU

#	Variable	Label
13	bldprd_legacy	Adult received clotting factor (LEGACY)
14	sex_idu	Adult heterosexual contact with IDU
15	sex_hiv	Adult heterosexual contact with person with HIV infection
16	transfusion	Adult received transfusion
17	transplant	Adult received transplant or artificial insemination
18	hcw	Health care worker
19	class	HIV class
20	aids_categ	AIDS case definition category
21	nir	Adult no identified risk factor (NIR)
22	oth_risk	Adult other documented risk
23	sex_orientation	Sexual Orientation

Figure 1 is the result of following Step 2, in which we removed variables with more than 25% missing data. We examined bivariate associations among patient response variables, summarized in Figure 1. The figure shows that differences across response rates are significant for some variable groups (e.g. injection drug use, age groups).

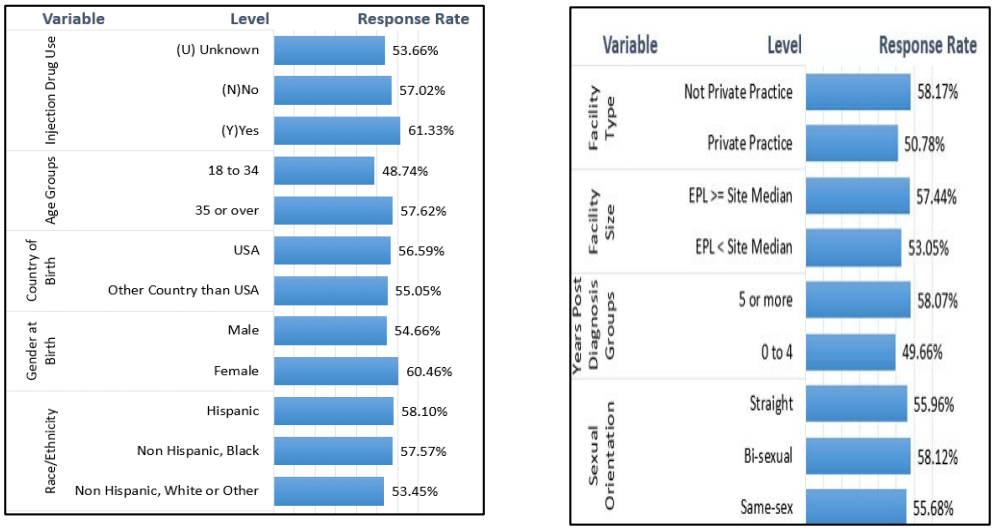


Figure 1: Bivariate associations

For example, among the age groups, one can see that patients 35 years and older and female patients have higher response rates than younger and male patients. The difference in response rates across subgroups defined by gender and age is much higher than the difference between subgroups defined by country of birth. Thus, the former variables are potentially better predictors of non-response.

For Step 3, after reviewing pairwise correlations among the response-associated predictors to minimize multicollinearity, at least seven predictors were retained as main effects in the multilevel model. These included (1) sexual orientation, (2) race/ethnicity, (3) injection drug use, (4) country of birth (USA vs. other country), (5) facility size (smaller vs. larger or equal to the region median), (6) age group (18 to 34 vs. 35 and over) and (7) gender at birth.

Step 4 involves running the multilevel logistic model. The patients are the level 1 units, which are nested within the level 2 clusters (facilities), which are furthered nested in level 3 clusters (project areas). While facility size varies only from facility to facility and is thus a level 2 covariate, we don't have any level 3 covariates measured on the project areas. Our model therefore has fixed effects at the first and second levels, and random intercepts and slopes at the second and third levels.

Figure 2 below is a Receiver Operating Characteristic (ROC) curve for the model, in which the area under the curve equals 0.7317. This curve shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). Also, the closer area under curve 1, the better the model performance.

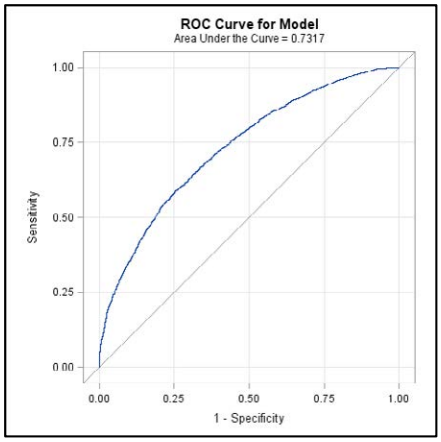


Figure 2: Multilevel model performance for patient survey

Step 5 involves running a multilevel model for response indicator with Proc Glimmix for level 1 (patient) and level 2 (facility) variables. Table 2 shows that all patient level variables are significant. A non-significant ($p > 0.05$) facility size variable is highlighted in red.

Table 2: Patient level variables

Solutions for Fixed Effects						
Effect	**Level**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept		0.07704	0.1381	22	0.56	0.5826
NREPL	EPL < Site Median	-0.169	0.09521	7343	-1.77	0.076
birth_country	Other Country than USA	-0.2877	0.07151	7343	-4.02	<.0001
NRAGE	18 to 34	-0.453	0.07014	7343	-6.46	<.0001
NRGENDER	Female	0.4691	0.07339	7343	6.39	<.0001
NRRACE	Hispanic	0.2137	0.08307	7343	2.57	0.0101
NRRACE	Non Hispanic, Black	0.3494	0.06789	7343	5.15	<.0001
idu	(N)No	-0.211	0.07847	7343	-2.69	0.0072
idu	(U)Unknown	-0.2746	0.07876	7343	-3.49	0.0005
sex_orientation	Bi-sexual	0.3356	0.08012	7343	4.19	<.0001
sex_orientation	Same-sex	0.3871	0.07645	7343	5.06	<.0001

Under the effects column, one can see the predictors for multivariate logistic model we've chosen based on bivariate analysis. When testing for random effects for the projected areas (site=level 3 and facility (level 2), we performed a test of covariance parameters. This approach tests whether the variances of the facility-level random intercept and the region-level random intercept are zero, which is also a test of the significance for these random effects. Effects for the project area and facility were both found to be significant (<.0001).

Figure 3 presents the odds ratios for the multilevel logistic model for response indicator Y (0-1). Importantly, those born in the USA, people older than 35, women, and bisexual persons have higher response propensities.

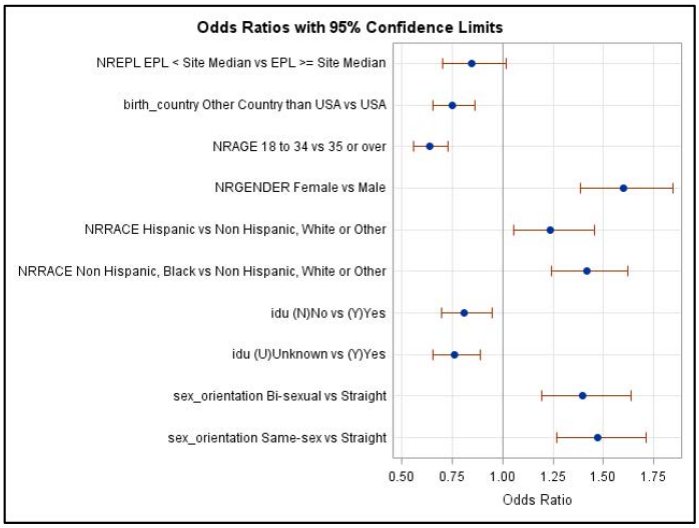


Figure 3: Odds ratios for multilevel logistic model

Finally, when we evaluated and compared results using a multilevel logistic model confusion matrix, the overall prediction accuracy was 68%. Figure 4 shows a cross tab of true and predicted respondents in health study.

	0-True NR	1-True R	Class. Error	
0-Pred NR	1697	830	33%	←False Omission Rate
1-Pred R	1696	3599	32%	←False Discovery Rate
			32%	←Overall Error Rate

Figure 4: HIV Multilevel model confusion matrix. Prediction accuracy of the HIV Multilevel model is 68%

4.2 Random Forest model for the patient health survey

Results of the Random Forest show that facility and site (region) played an important role as predictors, as well as sexual orientation and race.

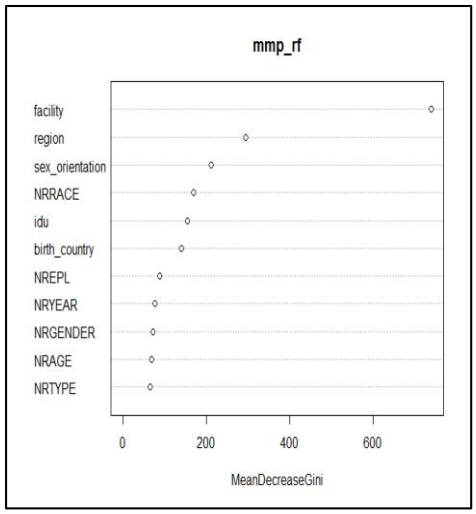


Figure 5: Measures of variable importance for Health Survey using Random Forest algorithm

We then produced a confusion matrix for the Random Forest model to see whether the model performs well in predicting respondents. The overall accuracy rate for this model is 60.

	0-True NR	1-True R	Class. Error	
0-Pred NR	1712	2213	56%	←False Omission Rate
1-Pred R	1377	3653	27%	←False Discovery Rate
			40%	←Overall Error Rate

Figure 6: Confusion matrix for the Random Forest method.

When compared with the multilevel logistic model confusion matrix, the overall error rate was higher for the Random Forest model. The prediction accuracy of the Random Forest method is 60%. Compared to the mixed effect logistic model, the Random Forest method was less sensitive but a bit more specific (fewer false positives)

4.3 Logistic regression model for school based study

In the first step, we selected a set of frame variables that were potentially related to school response propensities, displayed in Figure 7 below.

Variable	Level	Response Rate
Public vs. Non-Public School	Not-Public	14.30%
	Yes-Public	76.70%
Small vs. Large School	Large	73.40%
	Small	40.90%
Census Region	MW	73.70%
	NE	61.80%
	SO	73.10%
	WE	67.90%
Urban	No	81.40%
	Yes	58.50%
MDR Affluence	Low/Below Avg	74.10%
	Avg	68.80%
	Above Avg/High	63.10%
AP Classes Offered	No	63.80%
	Yes	72.10%
Before/After-School Programs Offered	No	73.30%
	Yes	29.40%
% Students College-Bound	Below median	70.50%
	Above median	79.60%
TechEd Courses Offered	No	55.20%
	Yes	75.80%
Enrollment Change Since Prior Year	Decrease	70.40%
	No Change	59.60%
	Increase	76.40%
% Students Enrolled in Free/Reduced-Price Lunch	Below median	74.00%
	Above median	80.80%

Variable	Level	Response Rate
NCES Locale	City	80.00%
	Suburb	75.00%
	Town	79.30%
	Rural	75.70%
Majority White School	No	100.00%
	Yes	70.40%
% Students below Poverty Line	Below median	66.70%
	Above median	83.00%
Per-Student Instructional Materials Expenditures	Below median	73.00%
	Above median	80.20%
Per-Student Curricular Materials Expenditures	Below median	74.00%
	Above median	80.20%
% Students in Special Education	Below median	74.60%
	Above median	80.80%
Student:Computer Ratio	Below median	70.90%
	Above median	70.60%
% Students Receiving Free/Reduced-Price Lunch and Title I Eligible	Below median	74.30%
	Above median	80.00%
Student:Teacher Ratio	Below median	77.60%
	Above median	75.30%
Per-Student Title I Spending	LT150	65.90%
	GE150	81.40%

Figure 7: Bivariate associations of potential variables related to school response rate

We observe more variability in response rates among subgroups for the school survey than in the health study. Of the 21 response-associated predictors, 16 predictors were retained for modeling, with the deletions designed to minimize multicollinearity. The selected variables included: (1) public vs. non-public school, (2) small vs. large school, (3) census region, (4) NCES locale, (5) MDR affluence locator, (6) AP classes offered (Y/N), (7) percent of students college-bound, (8) majority-white schools, (9) percent black students, (10) percent Hispanic students, (11) percent Asian students, (12) per-student instructional materials expenditures, (13) per-student curricular materials expenditures, (14) percent student in special education, (15) student/teacher ratio, and finally (16) per-student Title 1 spending. Figure 8 presents summary results of the logistic regression model.

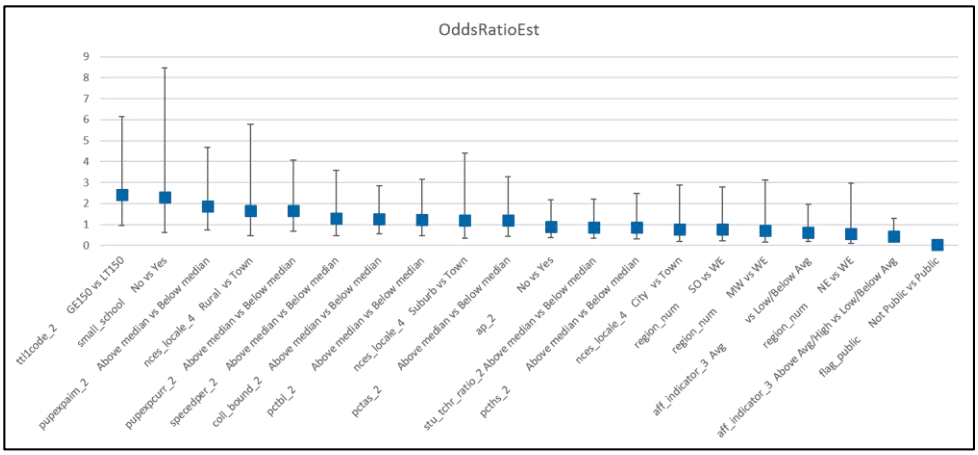


Figure 8: Odds Ratios for logistic regression for school response

The selected predictors were entered into a logistic regression model with school participant (Y/N) as the outcome. As shown in Figure 8, schools with higher Title I

spending are more likely to respond than schools with lower Title I spending (GE 150 versus LT150). Larger and public schools have higher response propensity. And schools with below average MDR affluence have higher response propensity than schools with above average MDR affluence.

Finally, when we evaluated and compared results summarized in the model confusion matrix, the overall accuracy rate was 79%. Figure 9 summarizes the error rates for the prediction.

	0-True NR	1-True R	Class. Error	
0-Pred NR	22	4	15%	←False Omission Rate
1-Pred R	33	121	21%	←False Discovery Rate
			21%	←Overall Error Rate

Figure 9: School logistic regression model confusion matrix

4.4 Random forest method for the school survey data

A model for school participation (Y/N) was built using the random forest algorithm. Figure 10 shows that the random forest algorithm suggests the use of the following predictors in non-response adjustments: school size, school type, region, and affluence (poverty).

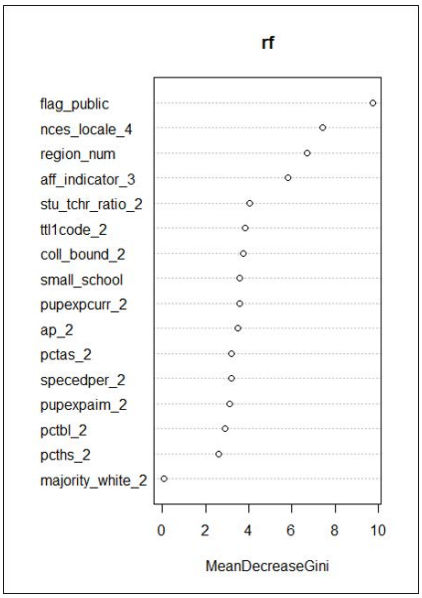
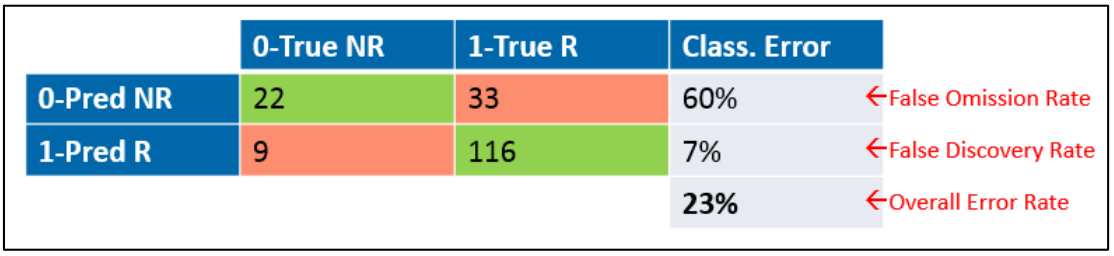


Figure 10: Measures of variable importance for the school survey using the Random Forest algorithm

To assess the accuracy of the Random Forest method, Figure 11 presents a confusion matrix for the method.



	0-True NR	1-True R	Class. Error	
0-Pred NR	22	33	60%	←False Omission Rate
1-Pred R	9	116	7%	←False Discovery Rate
			23%	←Overall Error Rate

Figure 11: Random forest confusion matrix.

The prediction accuracy of the Random Forest method is 77%. The multilevel logistic regression model had slightly better overall accuracy (79% correct classification) than the random forest method (77% correct classification). The random forest model was more specific but less sensitive (i.e., more false negatives) compared to the logistic model.

5. Discussion

For the health study, both methods – mixed effects models and Random Forest – agreed on the following variables as top predictors of patient's response propensity: facility, project area, sexual orientation, country of birth. The Random Forest model also identified race/ethnicity and injection drug use as important variables; the multilevel model, on the other hand, yielded gender and race group.

For the school survey, both methods identified school affluence and school type (public/nonpublic) as important predictors. The Random Forest method additionally selected NCES locale and Census region, whereas using the logistic model, school size and Title I spending were also significant.

In both studies, the logistic model had a slightly better overall performance, but for the Random Forest model, both cases were more specific (fewer false positives) for both studies. Further research will assess bias and variances for estimates based on survey weights adjusted with the different approaches.

References

Borgoni, R., & Berrington, A. (2011). Evaluating a sequential tree-based procedure for multivariate imputation of complex missing data structures. *Qualitative and Quantities Journal*, 47, 1991-2008.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29(3), 329-353.

Brick, J.M., & Montaquila, J.M. (2009). Nonresponse and weighting. In D. Pfeffermann and C.R. Rao (Eds.), *Sample surveys: Design, methods and applications* (vol. 29A, pp. 163-186). Amsterdam, The Netherlands: North-Holland.

Iachan, R., Harding, P., & Harding, L. (2014). *Non-response adjustments*. Presented at the Survey Research Methods Section of the American Statistical Association, Boston, MA.

Khan, M.H., & Shaw, J.E.H. (2011). Multilevel logistic regression analysis applied to binary contraceptive prevalence data. *Journal of Data Science*, 9, 93-110.

Larsen, K., & Merlo, J. (2005). Appropriate assessment of neighborhood effects on individual health: Integrating random and fixed effects in multilevel logistic regression. *American Journal of Epidemiology*, 161(1), 81-88.

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel structural equation modeling. In S.Y. Lee (Ed.), *Handbook of latent variable and related models* (pp.209-227). Amsterdam: Elsevier.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forest. *Psychology Methods*, 14(4), 323-348.

Zhu, M. (2014). *Analyzing multilevel models with the GLIMMIX procedure*. Paper SAS026-2014, SAS Institute. Retrieved from http://support.sas.com/resources/papers/proceedings14/SAS026-2014.pdf.