Methods for Reducing Bias for Estimation with Non-probability Samples: A Simulation Study

Clara Chen

April 2024

Abstract

Especially in the age of big data, non-probability samples are becoming increasingly common. Unlike the typical probability sampling framework, inclusion probabilities are unknown for non-probability samples, making it difficult to adjust for bias. While several methods have been developed to reduce bias of non-probability samples, it is not clear when each method performs best. Using data from the New York City Department of Education, this study investigates weighting and model-based approaches for reducing bias of non-probability samples via simulation. We find that bias reduction methods work well for estimating population parameters and a regression coefficient but not when estimating a difference between groups. This suggests that whether and which bias reduction method to use should be tailored to the estimation goal.

1 Introduction

A common statistical goal is to estimate the parameters of a large population from a sample. Typically a probability sampling design is used, where unbiased estimates can be made using known inclusion probabilities. However, in many practical scenarios, non-probability samples are conducted instead because of their cost-effectiveness and convenience. There are many methods that can adjust for the bias introduced by non-probability sampling, but it is not clear which methods are best suited for which conditions. This study aims to use simulation to compare the performance of post-stratification weighting and linear regression modeling as methods of reducing the bias of non-probability samples.

1.1 Probability Sampling

In the probability sampling framework, there is a finite population of size N, often denoted $\mathcal{U}$ for universe, from which a subset of size n, the sample $\mathcal{S}$, is chosen with a predetermined sampling design. Some examples of probability sampling designs include simple random sampling or stratified sampling. Importantly, in probability designs, each unit k in the population has a known, non-zero probability of being selected for the chosen sample, usually denoted π_k. The inverse of the inclusion probability, $w_k = \frac{1}{\pi_k}$, can be used as a survey weight, representing the number of people in the overall population that unit k represents.

The survey weights can be used to adjust for over or under representation of certain types of units in the sample. For example, the Horvitz-Thompson estimator shown for estimating the total and mean in equations (1) and (2) respectively uses inclusion probabilities to achieve unbiased estimates. Many other estimators such the Hayek estimator and the generalized regression estimator also incorporate inclusion probabilities to achieve asymptotic unbiasedness [Lohr, 2022] [Brewer and Hanif, 1983].

$$\hat{t}_y^{HT} = \sum_{k \in \mathcal{S}} y_k w_k = \sum_{k \in \mathcal{S}} \frac{y_k}{\pi_k} \tag{1}$$

$$\hat{\mu}_y^{HT} = \frac{1}{N} \sum_{k \in \mathcal{S}} y_k w_k = \frac{1}{N} \hat{t}_y^{HT} \tag{2}$$

1.2 Non-Probability Sampling

By contrast, in non-probability sampling, there is little to no control over which units are included in the sample, and, therefore, the inclusion probabilities are unknown. Most non-probability samples are a type of convenience sample where the main criteria for selection is the ease with which units can be recruited or located [Baker et al., 2013]. A sub-type of convenience samples are quota samples, where the population is divided into disjoint groups called quota classes and a set number of units are sampled within each quota. This is similar to stratification in probability sampling, but unlike stratification, within each quota class, a non-probabilistic convenience sample is taken.

In the big data age where digital data collection methods are becoming more widespread, non-probability sampling is becoming increasingly common. Unfortunately, selection bias is inherent in most non-probability samples because in convenience samples certain population units are more likely to participate in the survey than others, making the sample unrepresentative of the population. The larger the correlation between the study variable and the probability of participating in the sample, the larger the selection bias [Meng, 2018].

To derive meaningful estimates from nonprobability samples, we need to adjust for the bias. Because inclusion probabilities are unknown, we cannot directly adjust for bias as in probability sampling methods. However, we can often take advantage of auxiliary information, which is typically known for all population units from an external data source such as a census or administrative file, by using weighting or modeling methods.

1.3 Reducing Bias of Non-Probability Samples

1.3.1 Weighting

Weighting methods introduce pseudo survey weights by modeling participation probability. The participation indicator for unit k in the population is defined as:

$$R_k = \begin{cases} 1 & \text{if } k \in \mathcal{S} \\ 0 & \text{if } k \notin \mathcal{S} \end{cases} \tag{3}$$

Mirroring the probability sampling framework, we would like the survey weight to be $\frac{1}{P(R_k=1)}$, but since the true probability of participation is unknown, we will set the weight as:

$$\tilde{w_k} = \frac{1}{P(\widehat{R_k = 1})} \tag{4}$$

where $P(\widehat{R_k = 1})$ is an estimate of $P(R_k = 1)$. A simple method of estimating $P(\widehat{R_k = 1})$ is post-stratification [Valliant and Dever, 2011]. After sampling, the population is divided into H groups, or poststrata. The sample size for stratum h is n_h, and the population counts N_h for each stratum are also known. $P(R_k = 1)$ can be estimated with:

$$P(\widehat{R_k = 1}) = \frac{n_h}{N_h} \tag{5}$$

Weights introduced with post-stratification work best when $P(R_k = 1)$ is similar within each poststratum.

1.3.2 Model-Based Methods

Model-based methods attempt to estimate the unobserved data. A model is fit on the sample data to estimate the study variable from one or more auxiliary variables. Then, for the population units outside the sample, the model is used to estimate the study variable. For example, a model-based estimate of the total would be:

$$\hat{t}_y^{MB} = \sum_{k \in \mathcal{S}} y_k + \sum_{k \in \mathcal{U} - \mathcal{S}} \hat{y}_k \tag{6}$$

where $\hat{y}_k$ are the estimates produced from the model.

1.3.3 Comparing Methods

[Buelens et al., 2015] and [Buelens et al., 2018] compare weighting methods and model-based methods with a variety of machine learning models on simulated samples from an automobile dataset. In their study, they find that model-based methods outperformed weighting methods for reducing bias in a predictive inference setting. The methods and simulation structure in this study are inspired by the work of Buelens et. al. but extend beyond to statistical goals other than predictive inference. Additionally, by using a smaller population dataset and different approaches to mimicking non-probability sampling, this study investigates whether similar results hold in alternate conditions. In general, we aim to address the unanswered question of when bias reduction methods for non-probability samples should be used and which ones work best.

2 Methods

2.1 Data Overview

Variable Name	Description
dbn	School unique ID number
school_name	Name of school
total_enrollment	Total enrollment for all grades
grade_1, grade_2, ..., grade_12	Enrollment for each grade 1-12
high_school_enrollment	Sum of enrollment for grades 9-12
number_female, number_male	Enrollment by gender group
percent_female, percent_male	Percent of study body by gender group
number_asian, number_black, number_hispanic, number_multi_racial, number_native_american number_white	Enrollment by race/ethnicity group
percent_asian, percent_black, percent_hispanic, percent_multi_racial, percent_native_american percent_white	Percent of student body by race/ethnicity group
number_students_with_disabilities	Number of students with disabilities
percent_students_with_disabilities	Percent of students with disabilities
number_english_language_learners	Number of ELL students
percent_english_language_learners	Percent of students who are ELL
number_poverty	Number of students in poverty
percent_poverty	Percent of students in poverty
number_grads	Number of graudates
percent_grads	Graduation rate
is_charter	Indicator of being a charter school

Table 1: Non-exhaustive list of NYC Department of Education dataset variables

Data from the New York City Department of Education will be used as the reference population in this simulation. The first dataset used provides demographic information on all public NYC schools with enrollment by grade, gender, race, disability, english language learner (ELL), and poverty status. A second dataset provides graduation numbers and rates. After merging the two datasets and filtering to schools with non-zero high school enrollment in the 2022-2023 school year, there were a total of $N = 538$ schools. For the simulation study, demographic information is treated as auxiliary data known for all schools, and the graduation numbers and percentages are treated as only known for sampled schools.

Table 1 lists the relevant variables and their descriptions. The charter status of a school (`is_charter`) was determined by whether the school ID was present in a third dataset with test scores for charter schools only. For schools with poverty rate above 95%, the raw data encodes `percent_poverty` and `number_poverty` as "Above 95%". These entries were imputed with 0.95 and 0.95×`total_enrollment` respectively.

2.2 Simulation Setup

2.2.1 Drawing Samples

We will start by drawing samples of size $n = 100$ from the population data using three methods.

1. A simple random sample, intended to serve as a control.

2. A "convenience sample" over-sampling schools with larger enrollment and lower poverty rate.

   This sampling scheme could plausibly represent a non-probability sample since in an opt-in survey, schools with more resources would likely be more willing to participate. Specifically, the inclusion probabilities are proportional to the output of the equation 7, but scaled so that the weights of all schools sum to n.

   $$[\text{Normalized Total Enrollment}]^2 + [1 - (\text{Poverty Rate})]^2 \tag{7}$$

   Figure 1 shows the distribution of inclusion probabilities over enrollment and poverty. A subset of schools have extremely high inclusion probability close to or equal to 1, but the inclusion probabilities are generally low enough to ensure samples are sufficiently different.

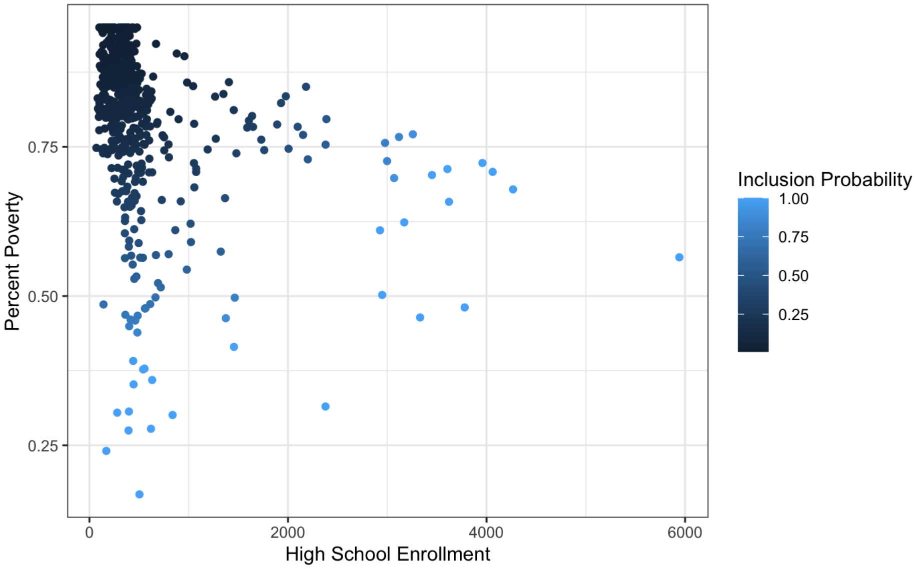


Figure 1: Inclusion probabilities for the "convenience sample" by poverty and enrollment

3. A "quota sample" by charter status.

   The number of charter and non-charter schools included in the sample are set to 13 and 87 respectively, proportional to their numbers in the overall population. Within each school type, schools with larger enrollment and lower poverty rates are over-sampled with the same process as sampling method 2.

2.2.2 Estimating Quantities

For each sample, we will estimate the following quantities:

1. Mean graduation rate
2. Total number of graduates
3. Difference in the mean graduation rate between charter schools and non-charter schools
4. Linear regression coefficient for number of graduates regressed on number of students in poverty

Study Variable	**True Population Value**
Mean `percent_grads`	82.87%
Total `number_grads`	62697
Difference in mean `number_grads` in charters and non-charters	-4.356%
Linear regression coefficient for `number_grads`∼`number_poverty`	0.305

Table 2: True population quantities

Together, these study variables cover three statistical goals of estimating a population parameter, a difference between groups or treatment effect, and a regression coefficient. The true population values are shown in Table 2. Both the difference and regression coefficient are significant at the 95% confidence level in the population.

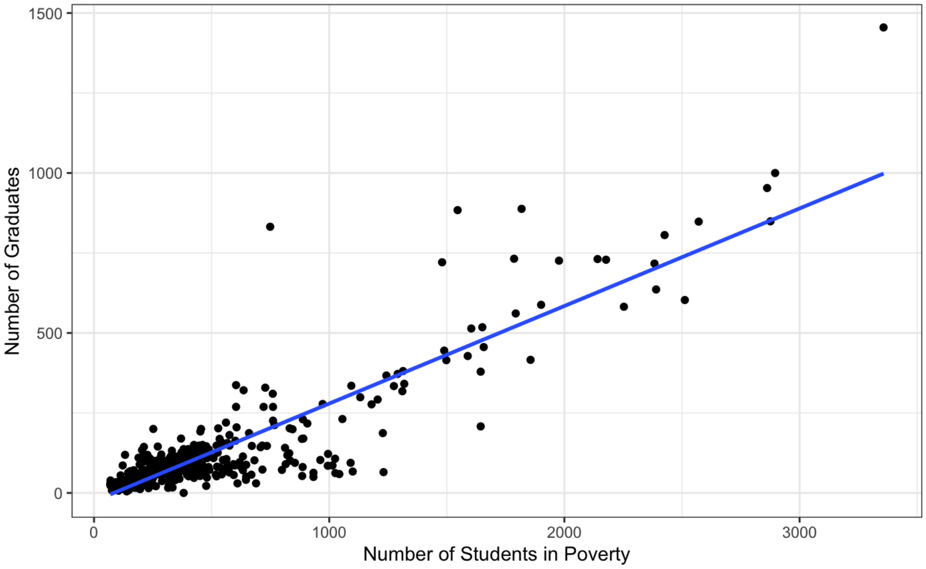


Figure 2: Number of graduates vs. number of students in poverty for the population ($R^2 = 0.79$)

2.2.3 Applying Bias Reduction Methods

For each sample, we will apply the following bias reduction methods.

1. Post-stratification weighting based on enrollment and poverty.

   Based on `high_school_enrollment`, each school is categorized as having "low", "medium", or "high" enrollment using thresholds of 372 (the population median) and 1000. Based on `percent_poverty` each school is categorized a having "low" or "high" poverty using a threshold of 0.75 (the 25% percentile). Post-stratification groups are defined as the interaction between these two categorical variables.

2. Post-stratification weighting based on race, disability, and ELL percentages.

   Each school is categorized as having "low" or "high" minority level, disability level, and ELL level using the population median for each respective variable as the threshold. Post-stratification groups are determined by the interaction of the three resulting categorical variables.

3. A model-based approach.

   Two linear regression models are fit on the sample data. The first model predicts `percent_grads` from the percentages for female, Asian, Black, Hispanic, white, disability, and ELL students. We obtain estimated graduation rates for schools outside the sample with this model in order to estimate the mean graduation rate and the difference between charter and non-charter schools.

   The second model predicts `number_grads` from enrollment numbers for female, Asian, Black, Hispanic, white, disability, and ELL students. The estimated number of graduates from this model are used for estimating the total number of graduates and the regression coefficient.

The study variables from section 2.2.2 will be re-estimated using each method. For the weighting methods, the Horvitz-Thompson estimator will be used for means and totals, and a survey-weighted model will be used to estimate the regression coefficient. We will compare the bias and variance of the estimates under each bias reduction method and sampling condition based on repeating this process for 1000 iterations.

3 Results

3.1 Bias

The distribution of estimated quantities for each sampling procedure and bias reduction method is shown in Figure 3. The bias for each condition can be found in Figure 8 and Table 3.

For simple random sampling, there was close to no bias across all conditions. When estimating the mean graduation rate and total number of graduates, the control case estimate was unbiased and remained unbiased when weighting or model-based methods were applied. For the regression coefficient, simple random sampling had slight negative bias in the control case, and applying the bias reduction methods decreased the bias. The only exception was when estimating the difference in mean graduation rate between charter and non-charter schools. In this case, applying the bias reduction methods introduced negative bias, particularly the model-based method.

The results for the convenience and quota samples were similar to each other. Without bias adjustment, estimates produced by both samples were very biased, particularly for the total number of graduates. For estimating the mean, total, and regression coefficient, the bias reduction methods all successfully reduced the bias. The first weighting method and the model-based approach performed the best. The second weighting method also moderately reduced bias, suggesting that some weighting is still beneficial even when the post-stratification groups do not explicitly relate to the variables underlying true inclusion probability.

When estimating the difference in mean graduation rate between charter and non-charter schools in the convenience and quota samples, bias reduction methods do not seem to perform well. The model-based approach increased bias for the convenience sample, and both the second weighting method and model-based approach increased bias for quota sample. Although the first weighting method successfully reduced the magnitude of bias for both non-probability samples, this method relies on knowing the variables underlying inclusion probability, which would likely not be known in non-simulation settings. Moreover, based on Figure 3, many of the estimated differences were positive when using weighting, when the true difference was negative. Incorrectly estimating the sign is particularly problematic for a difference, which further indicates that these weighting methods should not be used in this case.

Mean Graduation Rate			
Bias Reduction Method	Convenience Sample	Quota Sample	Random Sample
None	7.273 (8.8%)	7.071 (8.5%)	0.003 (0.004%)
Weighting 1	1.599 (1.9%)	1.381 (1.7%)	-0.001 (0.001%)
Weighting 2	2.169 (2.6%)	1.890 (2.3%)	0.011 (0.013%)
Model-Based	1.181 (1.4%)	1.043 (1.3%)	–0.002 (0.002%)

Total Number of Graduates			
Bias Reduction Method	Convenience Sample	Quota Sample	Random Sample
None	74187 (118%)	71874 (115%)	-176 (0.3%)
Weighting 1	6123 (9.8%)	7746 (12%)	32 (0.05%)
Weighting 2	32840 (52%)	35143 (56%)	-200 (0.32%)
Model-Based	-1750 (2.8%)	-4364 (6.7%)	–114 (0.18%)

Charter Difference in Mean Graduation Rate			
Bias Reduction Method	Convenience Sample	Quota Sample	Random Sample
None	-1.285 (29%)	-1.913 (44%)	-0.021 (0.5%)
Weighting 1	0.308 (7.1%)	1.167 (27%)	0.011 (0.3%)
Weighting 2	-1.142 (26%)	-2.237 (51%)	0.132 (3.0%)
Model-Based	2.672 (61%)	2.782 (64%)	0.397 (9.1%)

Regression Coefficient of Graduates on Poverty			
Bias Reduction Method	Convenience Sample	Quota Sample	Random Sample
None	0.0318 (10%)	0.0274 (9.0%)	-0.0057 (1.9%)
Weighting 1	0.0096 (3.1%)	-0.0199 (6.4%)	-0.0030 (1.0%)
Weighting 2	0.0249 (8.2%)	0.0154 (5.1%)	-0.0058 (1.9%)
Model-Based	0.0205 (6.7%)	0.0106 (3.5%)	-0.0031 (1.0%)

Table 3: Bias for each estimated quantity, absolute relative percent bias in parentheses; a visual with standard errors can be found in the appendix

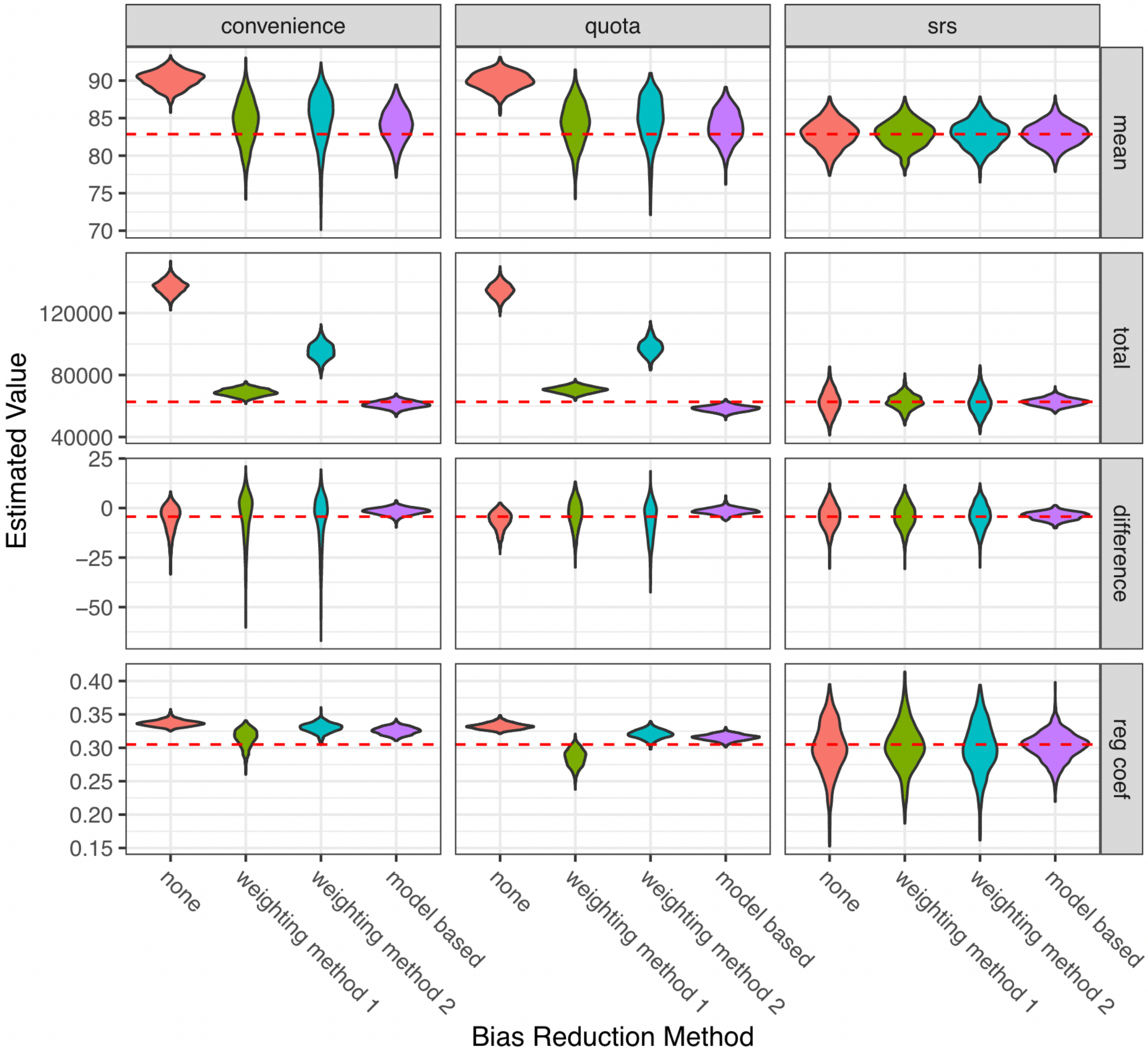


Figure 3: Violin plot of estimates for each bias reduction method faceted by sampling method and estimated quantity; true population value indicated by dashed red line

3.2 Variance

Overall, the variances for the non-probability samples are lower than the simple random samples. The convenience and quota sampling procedure inherently produce samples more similar to each other, thereby decreasing variance. In terms of variance, the model-based approach performed best. It had the lowest variance compared to other methods under all conditions, and even reduced variance for all quantities in simple random sampling.

The effect of the weighting methods on the variance of the estimates varied based on the study variable. For the mean and difference between groups, both weighting methods greatly increased variance. For the regression coefficient, the weighting methods both slightly increased variance. For estimating the total number of graduates, the first weighting method had much smaller variance while the second weighting method had similar variance to the control.

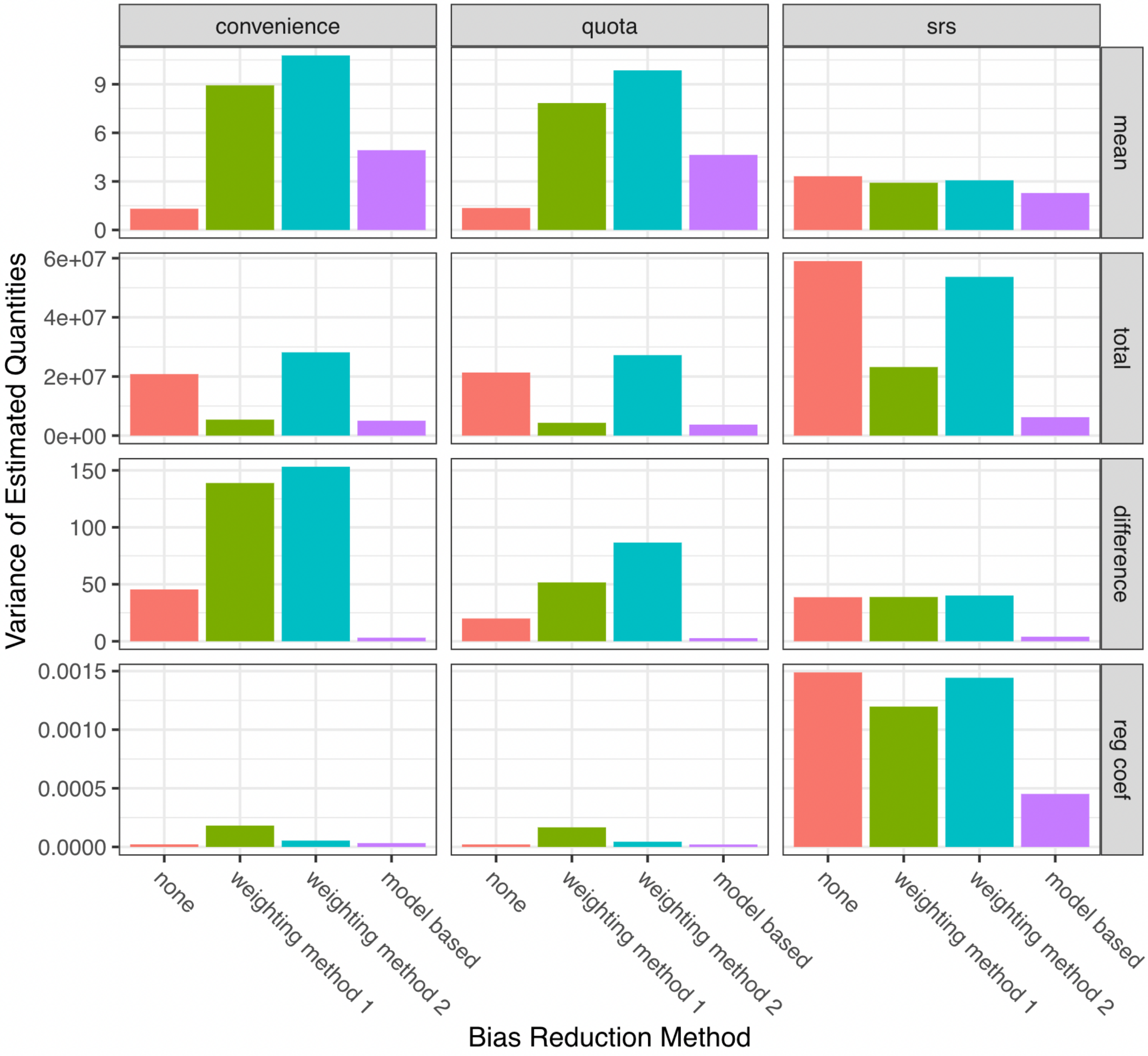
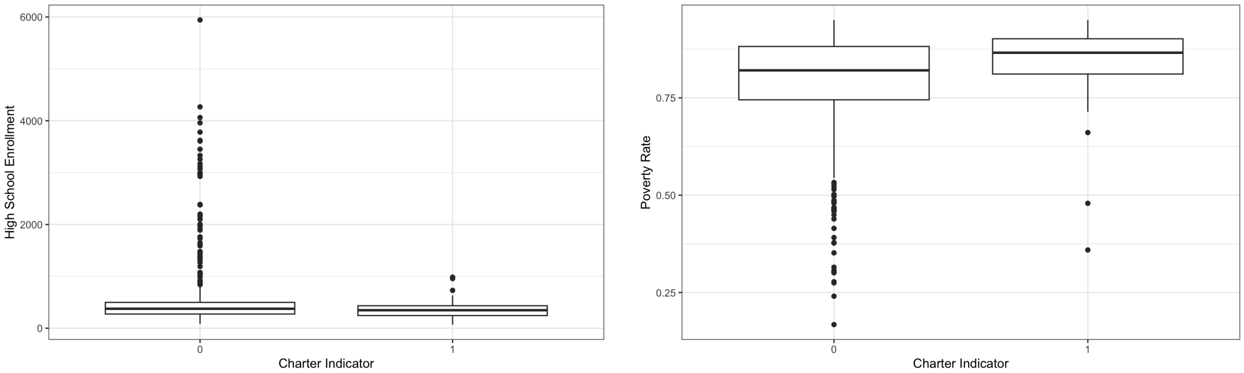


Figure 4: Variance of estimated quantities for each sampling method and bias reduction method



(a) Charter vs. Enrollment

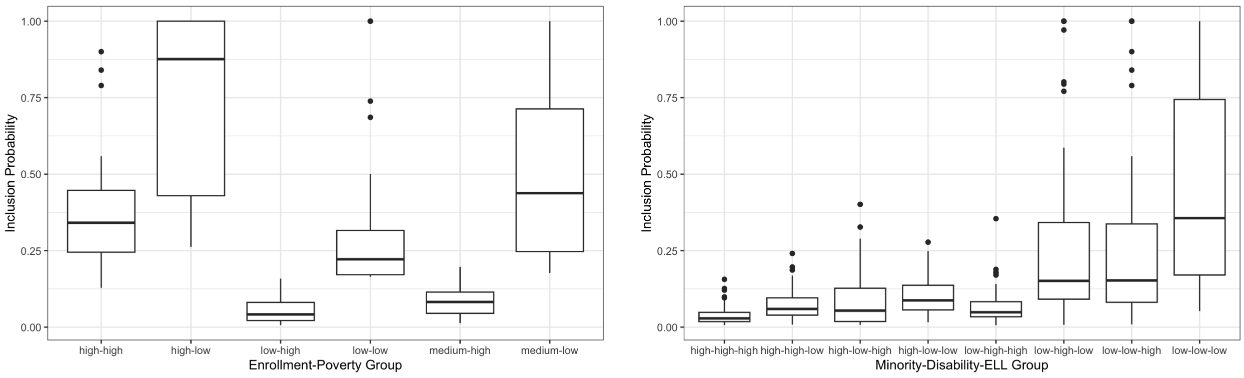
(b) Charter vs. Poverty Rate

Figure 5: Distribution of variables determining non-probability sample inclusion by quota

4 Discussion

Based on the results of this study, whether bias reduction methods for non-probability samples should be used depends on the estimation goal. When estimating a difference between groups, it seems the bias reduction methods should not be applied; the weighting methods slightly reduced bias but greatly increased variance, while the model-based method improved variance at the expense of increased bias. For estimating the population parameters and the regression coefficient, the bias reduction methods performed much better. Weighting based on true inclusion groups or the model-based approach performed best, but weighting based on other variables was moderately successful as well.

The conditions tested in this study were relatively narrow, and some of these results may be specific to the New York City schools data, leaving many possible directions for future work. The convenience and quota sampling results were likely so similar because the distribution of enrollment and poverty rate, the two variables determining inclusion probability, are not noticeably different between charter and non-charter schools (Figure 5). If another category with more heterogeneity between categories was available, quota sampling may have reduced bias even without bias adjustment methods. Other ways of simulating the non-probability samples with more extreme levels of participation imbalance could also be considered to investigate how effective weighting and model-based fixes are in the most extreme cases. Due to data limitations, both the population and sample size were quite small in this simulation. With a larger dataset, the impact of sample size could be tested as well.



(a) Weighting method 1

(b) Weighting method 2

Figure 6: Inclusion probability distribution by post-stratification groups

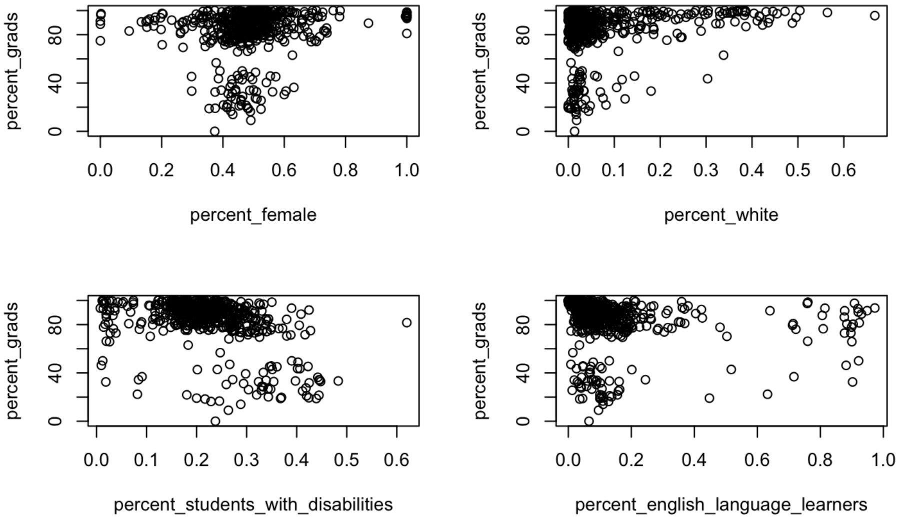


Figure 7: Scatterplots of graduation percentage with some of the predictors used in the model-based approach

Additional variations on the implementation of bias reduction methods could also be explored. For the weighting methods, each post-strata had relatively different distributions of inclusion probability, which contributed to the weighting methods working well (Figure 6), but this may not be true in other datasets. Therefore, other calibration methods or models for predicting participation probability would be worth considering. For the model-based approach, extending beyond the linear regression framework, more complex modeling methods such as regression trees and neural networks could be explored. The predictors for the model predicting the graduation rate demonstrated non-linear relationships with the outcome variable, which alternative models could better capture (Figure 7).

Importantly, while the methods tested in this study reduced bias, with varying degrees of success depending on the conditions, none of the methods were able to completely remove bias to the level of simple random sampling. Therefore, both weighting and model-based approaches require further refinement before we can fully leverage the potential insights from non-probability samples.

References

[Baker et al., 2013] Baker, R., Brick, M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., and Tourangeau, R. (2013). Summary report of the aapor task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1:90–143.

[Brewer and Hanif, 1983] Brewer, K. R. W. and Hanif, M. (1983). *Sampling with unequal probabilities.* New York: Springer-Verlag.

[Buelens et al., 2015] Buelens, B., Burger, J., and van den Brakel, J. A. (2015). Predictive inference for non-probability samples: a simulation study. *Statistics Netherlands.*

[Buelens et al., 2018] Buelens, B., Burger, J., and van den Brakel, J. A. (2018). Comparing inference methods for non-probability samples. *International Statistical Review*, 86.

[Lohr, 2022] Lohr, S. L. (2022). *Sampling: Design and Analysis.* CRC Press, third edition.

[Meng, 2018] Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2):685–726.

[Valliant and Dever, 2011] Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods Research*, 40(1):105–137.

Appendix

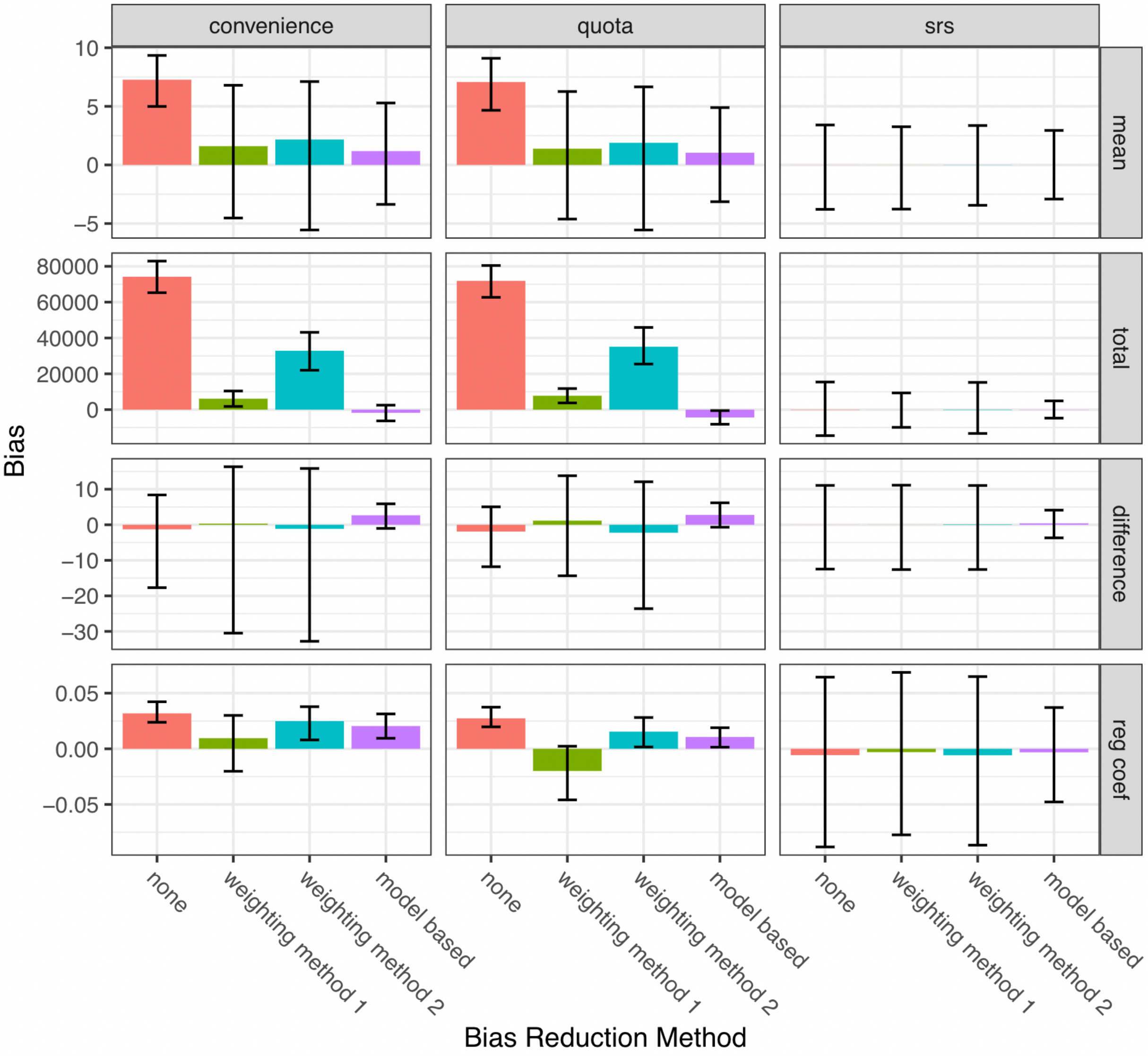
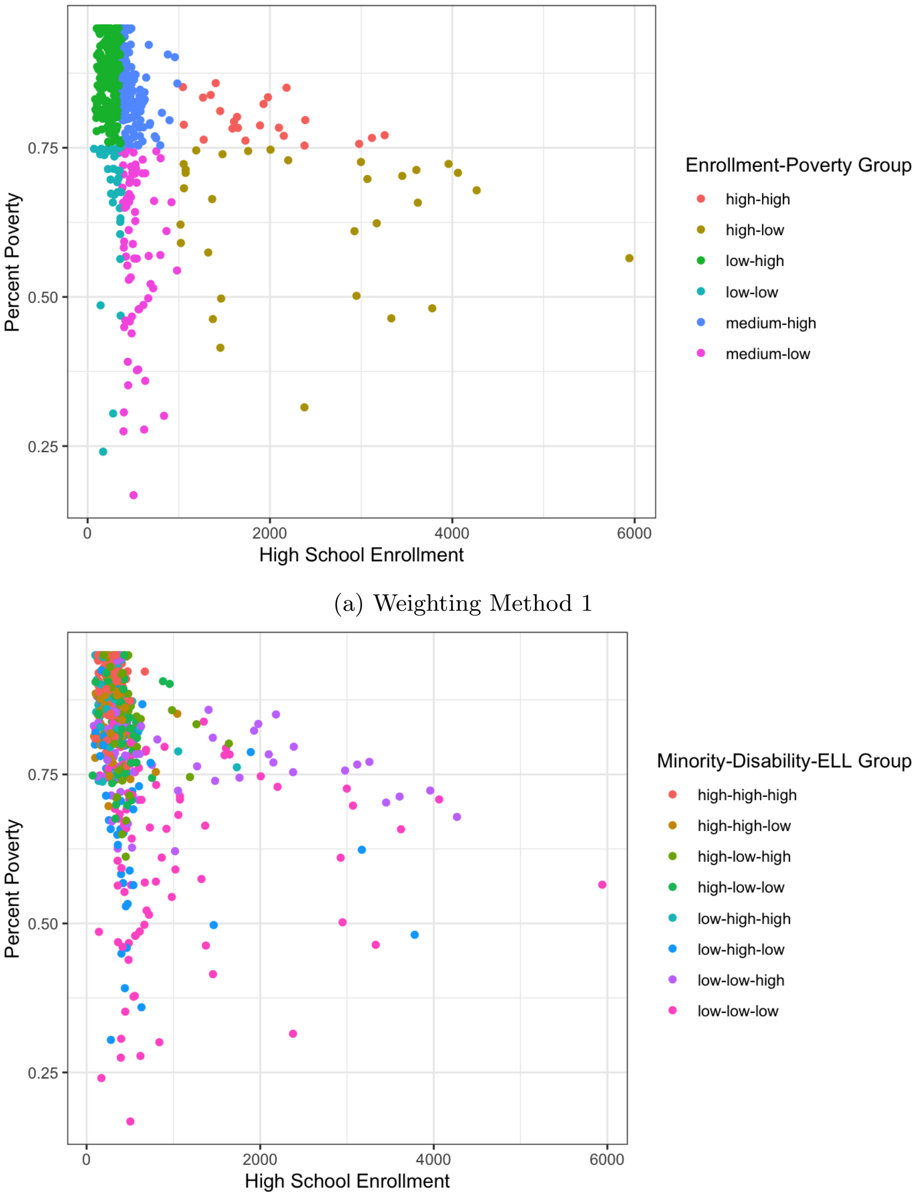


Figure 8: Bias of estimated quantities for each sampling method and bias reduction method, 0.025 and 0.975 quantiles indicated by error bars



(a) Weighting Method 1

(b) Weighting Method 2

Figure 9: Enrollment and poverty rate by post-stratification groups