Comparing Alternatives for Estimation from Nonprobability Samples

Richard Valliant
Research Prof. Emeritus
Universities of Michigan & Maryland

Abstract

Three approaches to estimation from nonprobability samples are quasi-randomization, superpopulation modeling, and doubly-robust estimation. In the first, the sample is treated as if it was obtained via a probability mechanism but, unlike in probability sampling, that mechanism is unknown. Pseudo selection probabilities of being in the sample are estimated by using the sample in combination with some external data set that covers the desired population. In the superpopulation approach, observed values of analysis variables are treated as if they had been generated by some model. The model is estimated from the sample and, along with external population control data, is used to project the sample to the population. The specific techniques are the same or similar to ones commonly employed for estimation from probability samples and include binary regression, regression trees, and calibration. When quasi-randomization and superpopulation modeling are combined, this is referred to as doubly-robust estimation. This paper reviews some of the estimation options and compares them in a series of simulation studies.

Keywords: doubly robust; multilevel regression and poststratification; pseudo-inclusion probabilities; quasi-randomization; superpopulation modeling

Word counts of text, excluding figures, tables, references, and appendices: approx. 7500

1 Introduction and Background

There has been a recent resurgence in interest in making inferences from nonprobability samples for several reasons. Response rates in probability surveys have been decreasing, particularly in telephone surveys. For example, Pew Research reported that their response rates (RRs) in typical

telephone surveys dropped from 36% in 1997 to 9% in 2012 (Kohut et al., 2012). With such low response rates, a sample initially selected randomly can hardly be called a probability sample from the desired population. Cost of follow-up to convert nonrespondents is expensive and often feckless, implying that poor response may remain so even with aggressive conversion attempts. Elliott and Valliant (2017) and Valliant et al. (2018, Ch. 18) review some of the problems that probability samples have encountered in the last decade.

There are also other data sources that are currently receiving attention and might be considered for finite population estimation (Couper, 2013). Social media and other data that can be scraped from the web might be used for gauging public opinion (Murphy et al., 2015) or measuring changes in consumer prices (Cavallo and Rigobon, 2016). These nonprobability sources may either replace probability samples or be combined with them for inference. Although the inferential issues raised subsequently apply to these "big data", we mainly concern ourselves with nonprobability samples that were directly collected for the purposes of making finite population estimates.

However, the appeal of collecting data quickly and cheaply does not mean that any nonprobability dataset is useful for inference. As an illustration, Figure 1, taken from Elliott and Valliant (2017), shows the general situation that is faced when using a sample of persons who have volunteered to be part of a panel recruited over the internet. The target universe is U; F_{pc} is the set of persons who visited webpages where the recruitment is done; F_c is the set that actually signed up for the panel; and s is the sample that is selected from the panel to participate in a particular survey. The inference task is to project s to the full population U. A few of the problems that can occur are:

(1) Only a small percentage of the population may visit the recruitment websites.

(2) Few persons who visit the recruitment sites may complete all of the steps to join the panel, e.g., click on a recruitment popup, respond to an email from the survey sponsor to verify that the person is not a robot, and complete a registration form to become part of the panel.

(3) The ones who join the panel are a poor cross-section of the population. For example, few persons between 18 and 24 years old may join and the racial composition may be much different from that of the population.

(4) The members of the panel who respond to a survey request may omit entire demographic groups, like elderly African American or Hispanic women, whose characteristics are unlike that of the rest of the population.

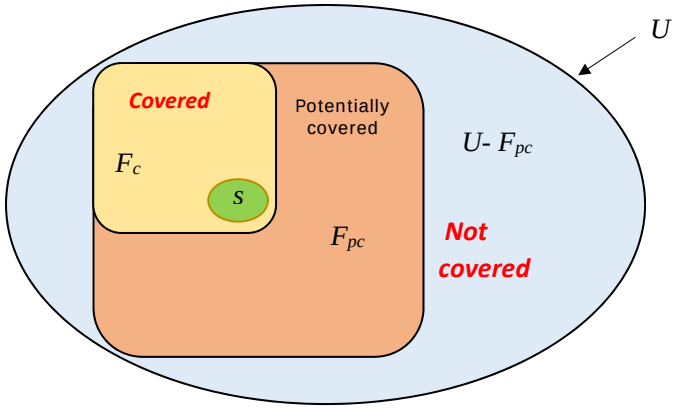

Figure 1: Illustration of potential and actual coverage of a target population when a non-probability sample is used

Projecting the data from s to U requires that all deficiencies in sample coverage and composition be corrected by estimating inclusion probabilities, poststratifying, raking, or some other means. The problem of having undercoverage in which the characteristics of the noncovered are different from those of the covered is essentially the same as that of having nonignorable unit nonresponse. Matei (2018) reviews several proposals for dealing with that type of unit nonresponse, including three reweighting schemes that are basically the same as the ones used in this paper. The reweighting methods are: (1) logistic regression to estimate the probability of being a response, (2) calibration directly to population totals, and (3) both (1) and (2) combined. Those methods correspond to the quasi-randomization, superpopulation, and doubly robust methods sketched in the following sections. More details are in Elliott and Valliant (2017) and Valliant et al. (2018, Ch. 18).

1.1 Quasi-Randomization

In the quasi-randomization approach, pseudo-inclusion probabilities are estimated and used to correct for selection bias. Given estimates of the pseudo-probabilities, their inverses are used as

weights just as done with the π-estimator. Design-based formulas are then used for point estimates and variances. The goal is to estimate the probability that we observe a unit in the sample (even though we did not control that probability). The inclusion probability of unit i, $\pi\left(i \in s|\mathbf{x}_i, \mathbf{y}_i; \mathbf{\Phi}\right)$, can depend on a vector of covariates, $\mathbf{x}_i$, the analysis variable(s), $\mathbf{y}_i$, and an unknown parameter, $\mathbf{\Phi}$, that must be estimated. This approach is also used in observational studies where it is known as the inverse probability-of-treatment weighted (IPTW) method (Robins et al., 2000). In the survey sampling literature, it is called propensity scoring or propensity score adjustment.

If the inclusion probabilities depend on the $\boldsymbol{y}$'s, this is a case of *not missing at random* (NMAR), which is very difficult to overcome. Since the nonsample $\boldsymbol{y}$'s are unknown, verifying that the sample data are not NMAR is impossible in most applications. There is some literature on estimation when nonsample data are NMAR (e.g., see Little, 2003, Matei, 2018), but the methods may require information on nonsample units that is available only in specialized applications. Thus, the only feasible approach is usually to estimate $\pi\left(i \in s|\mathbf{x}_i; \mathbf{\Phi}\right)$.

When the weights are $w_i = 1/\pi\left(i \in s|\mathbf{x}_i; \mathbf{\Phi}\right)$, estimated totals of the form $\hat{t}_y = \sum_{i \in s} w_i y_i$ are unbiased for target population totals in sense of repeated inclusion in the sample under the pseudo-probability distribution. As for design-based inference, every unit must have a non-zero chance of appearing in the sample. The difference from pure design-based inference is that we do not have control over the $\pi\left(i \in s|\mathbf{x}_i; \mathbf{\Phi}\right)$'s.

The technique used here is to estimate a pseudo-inclusion probability using a probability-based *reference survey*. The reference sample ($s_{\texttt{ref}}$) and the nonprobability sample are combined and the pseudo-inclusion probabilities for the nonprobability cases are estimated using a binary regression model. Other options for estimating the pseudo-inclusion probabilities would be to use machine learning methods, like CART, bagging, or random forests (James et al., 2014). The reference sample can be a probability sample that covers either the full target population—U in Figure 1—(which is the case for the empirical study here) or the potentially covered population, F_{pc}. In this discussion, we assume that the reference sample weights are scaled for estimating population totals, i.e., they are not normalized to sum to the sample size. An option would be to use an entire census file as the reference if it were available. The mechanics of estimation are:

(1) Code the cases in the reference sample as 0 and the cases in the nonprobability sample as 1.

(2) Reference sample cases receive their probability sample weight. Assign a weight of 1 to each nonprobability case.

(3) Fit a weighted binary regression to predict the probability of being in the nonprobability sample.

This weighted regression will approximately estimate the census model that would be fit if the reference sample were the entire population excluding the nonprobability sample. If the size of the nonprobability sample is not a negligible fraction of the population, the weights for the reference sample should be adjusted so that the sum of the weights in the combined sample is an estimate of the population size N. The adjustment is

$$w_i^* = w_i \frac{\hat{N} - n_{\texttt{np}}}{\hat{N}} \tag{1}$$

where w_i is the weight of element i in the reference sample $s_{\texttt{ref}}$, $\hat{N} = \sum_{s_{\texttt{ref}}} w_i$, and $n_{\texttt{np}}$ is the sample size of the nonprobability sample. This adjustment results in the sum of the weights in the combined sample being $\sum_s 1 + \sum_{s_{\texttt{ref}}} w_i^* = n_{\texttt{np}} + \hat{N} - n_{\texttt{np}} = \hat{N}$. If $n_{\texttt{np}}$ is a small fraction of $\hat{N}$, this adjustment is unnecessary.

Two other technical requirements for the nonprobability-reference sample combination are called *common support* and *common covariates* and are described below.

Common support. This requirement says that for every value of $\mathbf{x}_i$ in the population, the probability of being in either the nonprobability or the reference sample must be positive. This is analogous to the requirement for a probability sample that all units have a positive probability of selection. Common support also implies the full range of values of each $\mathbf{x}_i$ is, at least potentially, covered by both the sample and the nonsample. In principle, this requirement ensures that there is sufficient overlap in the characteristics of the units in the nonprobability and the reference samples so that $\pi\left(i \in s | \mathbf{x}_i; \mathbf{\Phi}\right)$ can be estimated for every unit in the population. If this condition is violated, predicted probabilities for some units will be unreliable. For example, if the target population and

reference sample cover the adult population 18 years and older but s does not include any women who are 65+, then this would violate the common support requirement. The violation could occur either because 65+ women have zero probability of being in s (e.g., of volunteering for an opt-in survey) or that the sample does not allow that probability to be estimated even if it is positive because there are no sample cases in the 65+ group of women. Inferences would have to be limited to persons 18-65 years old, or the dubious assumption would have to be made that women over 65 can be represented by some subset of the people 18-65.

Another, usually minor, technical requirement is that the reference sample and the nonprobability sample do not include the same units. If the units do overlap, then the 0-1 coding for (in s) or (not in s) in step (1) is ambiguous. If overlap occurs, then any duplicates should be removed from one sample or the other.

The results of the binary regression are estimates of the probabilities of being in the nonprobability sample within whatever population the reference sample represents. For example, suppose that the nonprobability sample s is a panel of persons who volunteered through web advertisements to participate in surveys done over the Internet. If the reference sample is a U.S. national sample with complete coverage of the population 18 years and older, the (inverse pseudo-inclusion probability) weights for the reference sample will inflate it to the full U.S. adult population. Claiming that the volunteers are a sample of the full adult population—a requirement if common support is satisfied—is debatable at best.

On the other hand, if the reference sample is from only those adults who have an Internet subscription of some type and its weights inflate it to that population, then the inverses of the pseudo-inclusion probabilities will inflate s only to the adult population that have an Internet subscription. This situation more nearly satisfies the common support assumption, although estimates that refer to the Internet population only may not be what the survey designer desires.

Common covariates. Another important requirement is that the reference survey and the nonprobability sample both collect the same set of covariates that will be used to model the inclusion probabilities. Otherwise, the binary regression cannot be fit. This common-covariate requirement

will necessitate comparability between how data are collected in the reference and nonprobability samples. For example, if categorical ethnicity is a covariate, the wording of the question should be exactly the same in both surveys. If the reference sample is an existing survey, like the U.S. American Community Survey (ACS), the nonprobability sample should use the same ethnicity question as does ACS.

Using existing surveys like the ACS is economical as long as the items in the existing survey are good predictors of the items to be collected in the nonprobability survey. On the other hand, using a specially designed reference survey opens up the possibility of collecting new items that may be better predictors. Schonlau et al. (2007) called these "webographic" variables since the authors were concerned with surveys collected over the Internet. Those authors found, in an example using a telephone reference survey, that phone and propensity-adjusted Web survey estimates were significantly different for a number of characteristics, but this difference was largely eliminated if both demographic and webographic covariates were used to estimate propensities. On the other hand, Lee (2006) found that adding non-demographic webographics was ineffective in either reducing biases or variances in her application.

1.2 Superpopulation Models

In the superpopulation modeling approach, models are fit to analytic survey variables and used to project to the full population. Each analysis variable y can potentially follow a different model making this approach seem less flexible than the quasi-randomization approach in Section 1.1 where the same pseudo-inclusion probability can be used to make estimates for any y variable. The practically expedient approach to achieving a similar level of generality in superpopulation modeling is to identify a form of model and set of covariates that produce reasonably good results for many y's. In that case, a single set of model-based weights can be used for all y's.

The general idea in model-based estimation when estimating a population total is to sum the responses for the sample cases and add to them the sum of predictions for nonsample cases. As before, s is the set of nonprobability cases and, in addition, let $\bar{s}$ denote the set of nonsample cases.

In order for inferences to be for the desired target population U, we must have $s \cup \bar{s} = U$. That is, "nonsample" means all units that are in the target population but not in the sample.

The key to forming unbiased estimates is that the sample and nonsample cases follow a common model and that this model can be discovered by analyzing the sample responses. An appropriate model usually includes covariates, and these must be known for each individual, nonprobability sample case. The covariates may or may not be known for individual nonsample cases, but, at a minimum, population totals of the covariates are required to construct the estimator. Requiring that population totals be available makes the requirements for superpopulation modeling akin to those for calibration estimators used in probability samples. Suppose that a linear model for a variable y is

$$E_M(y_i) = \mathbf{x}_i^T \beta$$

where the subscript M means that the expectation is with respect to the model, $\mathbf{x}_i$ is a vector of p covariates for unit i and β is a parameter vector. Given a sample s, the ordinary least squares estimator of the slope parameter is $\hat{\beta} = \mathbf{A}_s^{-1}\mathbf{X}_s^T\mathbf{y}_s$ where $\mathbf{A}_s = \mathbf{X}_s^T\mathbf{X}_s$, $\mathbf{X}_s$ is the $n \times p$ matrix of covariates for the sample units, and $\mathbf{y}_s$ is the n-vector of sample y's. (If $var_M(\mathbf{y}) = \mathbf{V}$, a diagonal or non-diagonal covariance matrix, generalized least squares can be used to estimate β.) A prediction of the value of a unit in the set of nonsample units is $\hat{y}_i = \mathbf{x}_i^T\hat{\beta}$, $i \in \bar{s}$. A predictor of the population total, t_y, is

$$\begin{aligned} \hat{t}_{y1} &= \sum_{i \in s} y_i + \sum_{i \in \bar{s}} \hat{y}_i \\ &= \sum_{i \in s} w_{1i} y_i \end{aligned} \tag{2}$$

where $w_{1i} = 1 + \mathbf{t}_{\bar{s}x}^T \mathbf{A}_s^{-1}\mathbf{x}_i$ with $\mathbf{t}_{\bar{s}x} = \mathbf{t}_{Ux} - \mathbf{t}_{sx}$. The theory for this *prediction approach* is extensively covered in Valliant et al. (2000). Other options for predicting nonsample values would be to use machine learning methods described in James et al. (2014) or the LASSO (Tibshirani, 1996, Chen et al., 2018).

If the sample is a small fraction of the population, as would be the case for applications like opt-in web surveys, the prediction estimator is approximately the same as predicting the value for every unit in the population and summing the predictions:

$$\begin{aligned}\hat{t}_{y2} &= \sum_{i \in U} \hat{y}_i = \mathbf{t}_{Ux}^T \hat{\beta} \\ &= \sum_{i \in s} w_{2i} y_i \end{aligned} \tag{3}$$

where $w_{2i} = \mathbf{t}_{Ux}^T \mathbf{A}_s^{-1} \mathbf{x}_i$. As is clear from the form of the weights, w_{1i} and w_{2i}, which population is being represented is then governed, in part, by the set of control totals used in estimation. (The parallel consideration for quasi-randomization is which population a reference sample represents.)

A mean or proportion can be estimated using the standard approach of dividing an estimate of a total by the sum of the weights: $\hat{\bar{y}} = \hat{t}_y / \hat{N}$, where $\hat{N} = \sum_s w_i$ and w_i is again either of the model-based weights defined above. The denominator is denoted as $\hat{N}$ because the sum of these weights will be near N in all situations as long as an intercept is included among the model covariates.

The estimators in (2) or (3) are quite flexible in what covariates can be included. For example, we might predict the amount that people have saved for retirement based on their occupation, years of education, marital status, age, number of children they have, and region of the country in which they live. Interactions can also be used. Constructing the estimator would require that census counts be available for each of those covariates. Another possibility is to use estimates from some other larger or more accurate survey (e.g., Dever and Valliant, 2010, 2016). The reference surveys mentioned earlier could be a source of estimated control totals in which webographic covariates might be used.

Variance estimation. There are several choices, described in Valliant et al. (2000, chaps. 5 and 9) and Elliott and Valliant (2017), for variance estimators when model-based weighting is used. To fully define the model, we need to add a variance specification. The variance estimator we sketch here is appropriate for models in which units are mutually independent. Although model-based

estimators have been extended to cases where units are correlated within clusters (see Valliant et al., 2000, chap. 9), these clustered structures are typically unnecessary for opt-in web surveys and similar cases. Suppose that the full model is

$$E_M(y_i) = \mathbf{x}_i^T \beta \tag{4}$$
$$V_M(y_i) = v_i$$

where v_i is a variance parameter that does not have to be specifically defined. The variance estimator below will work regardless of the form of v_i (as long as the value is finite).

When the size of s is negligible compared to the size of the population, a variance estimator that is approximately unbiased under (4) is

$$v\left(\hat{t}_y\right) = \sum_s a_i^2 \hat{v}_i \tag{5}$$

where $\hat{v}_i = y_i - \mathbf{x}_i^T \hat{\beta}$ with $\hat{\beta}$ being the ordinary least squares estimator of β, and $a_i = w_i - 1$ with w_i being either w_{1i} or w_{2i}. This estimator is robust in the sense of being approximately model-unbiased regardless of the form of v_i (which is unknown) as long as the sampling fraction is small. The estimator in (5) also has the convenience of being approximately equal to the default estimator computed in the R `survey` package (Lumley, 2017), which will be used in the simulation study.

If the population totals for some of the covariates are estimated from an independent survey, then the variance in (5) should be modified by adding a term to reflect that additional uncertainty (see Dever and Valliant, 2010, 2016).

1.3 Multilevel Regression and Poststratification

Multilevel regression and poststratification (MRP) (Gelman and Little, 1997, Gelman, 2007) is a variation on superpopulation modeling. An elaborate set of poststrata is formed by crossing the covariates that are predictors of the survey analysis variables. Any continuous predictors are broken

into categories. The intuition behind this method is that if enough finely defined cells are formed, then almost any underlying model can be approximated, regardless of how complicated. A mean or proportion is estimated as

$$\hat{\bar{y}} = \sum_{\gamma=1}^{G} P_\gamma \hat{\mu}_\gamma \tag{6}$$

where P_γ is the proportion of the population in poststratum (PS) γ, $\hat{\mu}_\gamma$ is the estimated mean per element in poststratum γ. If the P_γ are unknown, then estimates, $\hat{P}_\gamma$, are used (e.g., see Dong et al., 2014, Zhou et al., 2016). The PS mean is estimated by a random (or mixed) effects model, which can be Bayesian or not. The general approach is to begin with the cross-classification of many covariates and dynamically decide which crosses to retain. Bayesian methods are especially useful in this regard since they can adapt the PS estimates to cases where cells have little, if any, sample. However, if sample sizes are small or zero in some poststrata, and the $\hat{\mu}_\gamma$ for such PS are shrunk toward an overall mean, MRP can be biased if the population means in those PS are substantially different from those of the PS that are well-covered. This can happen when, for example, the PS are defined by demographic variables, and the sample frame poorly covers some combinations of the demographics.

In cases where a linear model is used to predict y, MRP does produce unit-level weights as shown in Gelman (2007). However, when y is binary and a logistic or similar model is used at the unit level, $\hat{\bar{y}}$ cannot be written as a weighted sum of the y's. This can be a practical disadvantage if many estimates are to be made from the nonprobability sample, and a separate model must be fitted for every y.

MRP has been particularly interesting to political scientists who want to make state-level estimates based on relatively small national samples of 1,500 to 2,000 voters (Park et al., 2004, 2006, Selb and Munzert, 2011, Shapiro, 2011). An interesting application is Wang et al. (2015) who were able to predict the outcome of the 2012 U.S. presidential election using MRP with a sample that had gaping holes in coverage. In contrast, Buttice and Highton (2013) illustrated that strong covariates in the multilevel model and substantial variation across geographic units were necessary for MRP to perform well for both national and state-level estimates; otherwise MRP can have mediocre performance.

2 Simulation Study

A simulation study was conducted to compare the performances of the quasi-randomization, model-based, doubly robust, and MRP approaches. To simulate the situation that might obtain in a nonprobability sample of persons, probability samples were selected that were disproportionately allocated to age groups. At the estimation stage, it was assumed that an analyst had no knowledge of the selection probabilities but had to construct a set of weights to estimate population quantities. The simulation was also done in such a way that each sample had a considerable amount of undercoverage that needed to be corrected through estimation.

The population in the simulation was derived from the dataset `mibrfss` in the R PracTools package (Valliant et al., 2017). The Michigan Behavioral Risk Factor Surveillance Survey (MIBRFSS) is part of a national state-by-state system of surveys used to monitor health conditions in the U.S. Data are collected through telephone household interviews. Demographic variables and a few health related variables are included in the `mibrfss` subset. The `mibrfss` data set contains observations on 2,845 persons and is extracted from the 2003 U.S. survey. The file contains only persons 18 years and older.

To create a larger population (U in Figure 1) to use for the simulation, `mibrfss` was bootstrapped to a larger artificial dataset by selecting a simple random sample of $N = 50,000$ with replacement from the 2,845 persons. The variables used in the simulation were:

- Demographics
  - `AGECAT`: Age (1 = 18-24 years; 2 = 25-34 years; 3 = 35-44 years; 4 = 45-54 years; 5 = 55-64 years; 6 = 65+);
  - `RACECAT`: Race (1 = White; 2 = African American; 3 = Other);
  - `EDCAT`: Education level (1 = Did not graduate high school; 2 = Graduated high school; 3 = Attended college or technical school; 4 = Graduated from college or technical school);
  - `INCOMC3`: Income (1 = Less than $15,000; 2 = $15,000 to less than $25,000; 3 = $25,000 to less than $35,000; 4 = $35,000 to less than $50,000; 5 = $50,000 or more)

- Analysis variables
  - `GENHLTH`: General health (self-reported) (1 = Excellent; 2 = Very good; 3 = Good; 4 = Fair; 5 = Poor). For the simulation analysis, the Fair and Poor categories were combined.
  - `SMOKE100`: Smoked 100 or more cigarettes in lifetime (1 = Yes; 2 = No)

The analysis variables are ones for which population estimates will be computed while the demographics are used in covariates for the quasi-randomization and superpopulation models. In particular, the quantities estimated are:

(1) `SMOKE100`, proportion of persons who have smoked 100 or more cigarettes

(2) `GENHLTH`, proportion of persons in each category

(3) `FairPoor`, proportion of persons reporting fair or poor health

(4) `GoB`, proportion of persons reporting good or better health

(5) `mGenHlth`, mean reported value of general health

2.1 Population Statistics and Sampling Methods

The population from which samples were drawn consisted of the 65% of persons who reported having access to the Internet at home (F_c in Figure 1). Thus, the sampling frame has a substantial amount of undercoverage that must be corrected for in estimation. Stratified samples (s in Figure 1) were selected using age groups as strata with the allocation in the last column of Table 1. The age distributions of the population and the subgroup with Internet access at home are also shown in Table 1. The Internet subgroup skews younger than the full population. For example, 61% of the population is less than 55 years old while 70% of the Internet subgroup is. The sample is even younger with 82% being under 55.

As illustrated in Tables 2 – 4, the proportions and means of the analysis variables do differ noticeably between the full population U and the Internet subgroup F_c, which constitutes the

Table 1: Distribution by age of the population of persons, the subgroup that has Internet access at home, and the samples.

	Proportions		
Age	Population	Internet	Sample
1 = 18-24	0.057	0.061	0.120
2 = 25-34	0.135	0.156	0.310
3 = 35-44	0.197	0.234	0.190
4 = 45-54	0.224	0.255	0.200
5 = 55-64	0.170	0.175	0.130
6 = 65+	0.218	0.119	0.050
Total	1.000	1.000	1.000

sampling frame. In the Internet frame, 50.4% have smoked 100 cigarettes but 53.5% have in the population. People in the frame population rate themselves as healthier than the full population: 89.4% of the Internet frame rate their themselves as having good or better health but 84.2% of the full population give themselves that rating. Persons in the population generally rate themselves as being unhealthier than the persons with Internet access; this is a result of the Internet subgroup being younger on the whole than the full population. Consistent with the distribution of health rating, the mean of the general health rating codes is higher at 2.443 for the full population than the 2.278 for the Internet persons.

The goal of inference is to make estimates for the entire population U based on a sample from F_c even though frame quantities may be different than those of the full population. Regardless of the method of estimation, the problems that must be overcome are (1) the analytic variables have different distributions in the Internet frame and the population and (2) the actual probabilities of being observed are unknown. Although the degree of undercoverage (35%) may seem large, it is not unrealistic for many surveys—volunteer or not—where cooperation is both low and quite selective.

Figure 2 gives plots of the proportions of persons in the target population who have smoked at least 100 cigarettes in their lifetime. Separate plots are shown for age-group $\times$ race, age-group $\times$ education, age-group $\times$ income, race $\times$ education, race $\times$ income, and education $\times$ income. The fact that the lines in each plot are not parallel implies that there is some degree of interaction that might be accounted for in estimation. To test this further, we fit two linear models using the

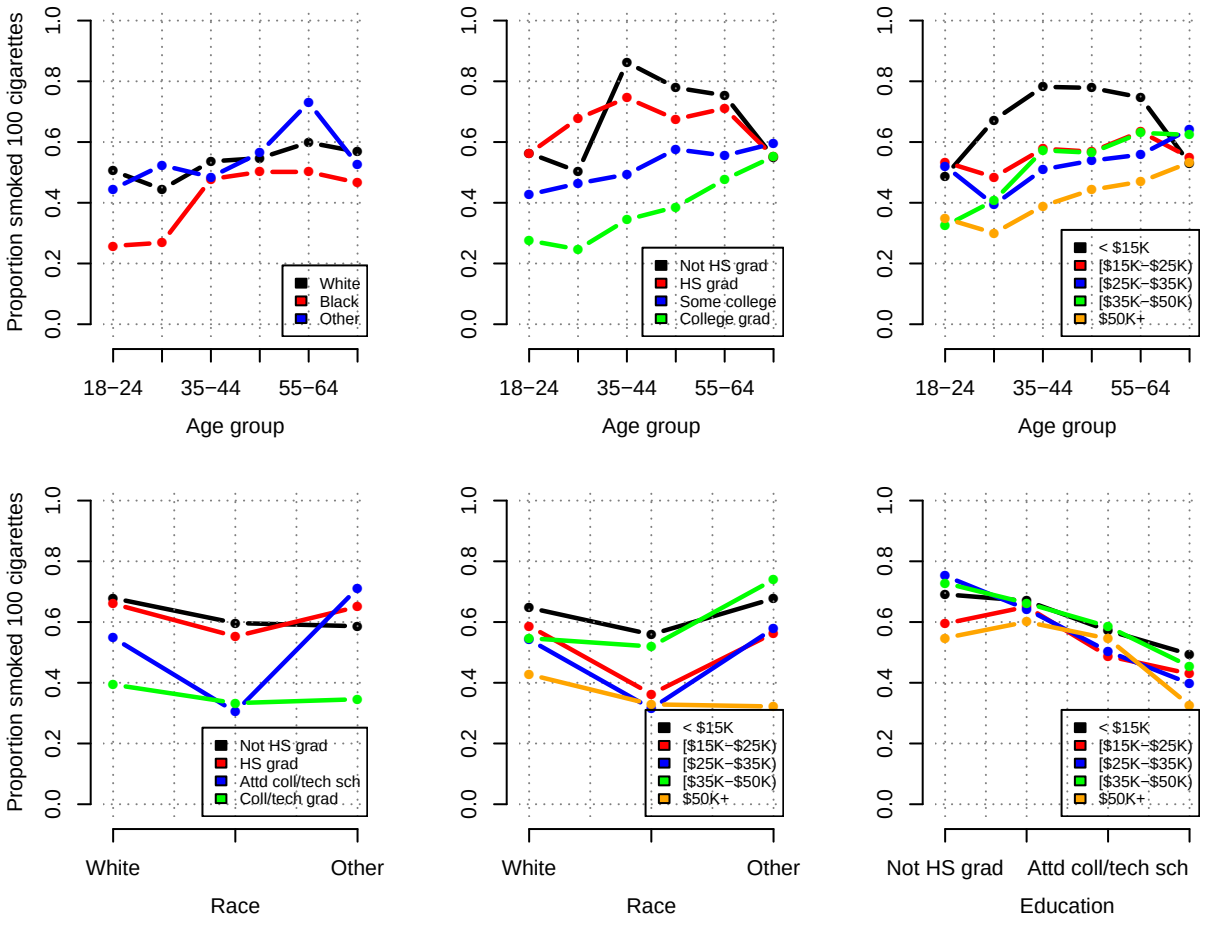

Figure 2: Plot of proportions of persons in the target population who have smoked at least 100 cigarettes in lifetime by 2-way crosses of age, race, education, and income groups.

Table 2: Proportions and means of analysis variables in the full population and subpopulation of persons with access to the Internet at home.

	Smoked 100 cigarettes		Good or better health	
	Population	Internet	Population	Internet
Yes	0.535	0.504	0.842	0.894
No	0.465	0.496	0.158	0.106

Table 3: Proportions and means of analysis variables in the full population and subpopulation of persons with access to the Internet at home.

	General health		Mean general health	
	Population	Internet	Population	Internet
Excellent	0.178	0.215	2.443	2.278
Very good	0.358	0.398		
Good	0.306	0.281		
Fair or poor	0.158	0.106		

Internet subset of the full population to predict analysis variables:

- M1, Main effects only: `AGECAT`, `RACECAT`, `EDCAT`, `INCOMC3`
- M2, Main effects + 2-way interactions: `AGECAT*RACECAT`, `AGECAT*INCOMC3`, `RACECAT*EDCAT`, `RACECAT*INCOMC3`, `EDCAT*INCOMC3` (`AGECAT*EDCAT` was not significant)

Each of these fitted models was then used to predict means of `GENHLTH` and `SMOKE100` (recoded to 0=No, 1=Yes) for the *full* population by age groups. Although fitting linear models to these variables would not be done in conventional data analysis, the implied models are linear for estimators of the form $\hat{t}_y = \sum_s w_i y_i$ of totals and $\hat{\bar{y}} = \sum_s w_i y_i / \hat{N}$ of means. Thus, examining the performance of linear models is consistent with standard survey practice.

Table 4 gives the means of `GENHLTH` by age group for the full population, the subset of persons with Internet access, the mean M1 predictions, and the mean M2 predictions. The Internet means by age are all less than those of the full population reflecting the fact that the Internet persons are younger and healthier. The means of the predictions from the two models are almost the same in each group and are intermediate between the Internet and full population means. That

is, using either model fitted on the Internet subset improves the predicted means compared using unadjusted Internet means by age group. The closeness of the two models also implies that the degree of improvement to be had in this population by calibrating with a model that has interactions is likely to be small.

Table 4: Means by age of general health for the population, the subset with the Internet at home, the main effects model (M1), and the model with 2-way interactions (M2).

Age	Population	Internet	M1	M2
18-24	2.27	2.19	2.26	2.23
25-34	2.18	2.14	2.18	2.17
35-44	2.26	2.18	2.23	2.23
45-54	2.36	2.21	2.33	2.33
55-64	2.55	2.42	2.50	2.51
65+	2.82	2.63	2.76	2.76

Figure 3 gives plots of the mean prediction by age from M1 and M2 versus the target population means for `GENHLTH` and `SMOKE100`. Most of the points being below the 45^o line is consistent with the means in the Internet subset being less than the full population means in Table 4. Although Figure 2 suggested that there might be interactions among the covariates that might be worth accounting for, the mean predictions for M1 and M2 in Figure 3 are hardly different.

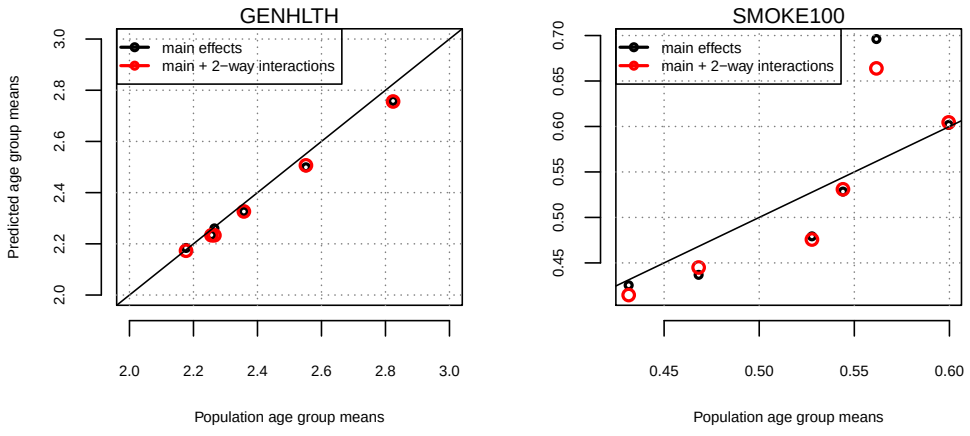

Figure 3: Predicted means by age group from main effects-only (M1) and main effects + 2-way interaction (M2) models plotted vs. population means for `GENHLTH` and `SMOKE100`. Reference lines are drawn at 45^o.

2.2 Estimation Methods

Sets of stratified samples of sizes 500 and 1,000 were selected from the Internet subset of 32,613 in the full population of 50,000. Consequently, each sample has built-in undercoverage. Strata were defined by age group with the allocation given in the last column of Table 1. Separate sets of 10,000 samples were selected for quasi-randomization, superpopulation linear models, and doubly robust estimation. Sets of 500 samples were selected for MRP estimation since that method is much more computationally demanding.

Quasi-randomization. For quasi-randomization, a simple random reference sample was selected without replacement from the full population in each iteration of the simulation. The reference sample size was the same size as the nonprobability sample (500 or 1,000). The two samples were combined, as described in Section 1, and a main-effects logistic model was fitted to predict the probability of being in the nonprobability sample. The covariates in the model were `AGECAT`, `RACECAT`, `EDCAT`, and `INCOMC3` described in Section 2.1. The weights for the reference sample were those in equation (1). The analysis weights for the persons in the nonprobability sample were then the inverses of their estimated observation probabilities. Variances were estimated in two ways. The first was with the linearization, ultimate cluster variance estimator, which is appropriate for with-replacement, varying probability sampling (Valliant et al., 2018, sec. 15.3). Using $\hat{\theta}$ to denote either an estimated proportion or mean, the formula is

$$v_{wr}\left(\hat{\theta}\right) = \hat{N}^{-2}\frac{n}{(n-1)}\sum_{i\in s}\left(\hat{z}_i - \hat{\bar{z}}\right)^2, \tag{7}$$

where $\hat{z}_i = w_i z_i$, z_i is a linearized deviate (or score) associated with $\hat{\theta}$, $\hat{\bar{z}} = n^{-1}\sum_{i\in s}\hat{z}_i$ with w_i being the quasi-randomization weight, and $\hat{N} = \sum_{i\in s} w_i$. The second variance estimator was a grouped jackknife. Each sample/(reference sample) combination was divided into $G = 50$ equal-sized random groups and the logistic model refitted in every group. The proportion or mean was

estimated from each. The grouped jackknife estimator is then

$$v_J\left(\hat{\theta}\right) = \frac{G-1}{G}\sum_{g=1}^{G}\left(\hat{\theta}_g - \hat{\theta}\right)^2 \tag{8}$$

where $\hat{\theta}_g$ is the estimated mean or proportion from all units not in group g.

Linear Model-based. Two sets of weights were computed in each sample. Both were calibrated estimates using the same covariates as the quasi-randomization estimator: `AGECAT`, `RACECAT`, `EDCAT`, and `INCOMC3`. These were entered into a calibration model as main effects with no interactions. The first set of weights were those for a linear model estimator (w_{1i} in equation (2)). The second set was from raking. In some applications, raking can preserve interactions among the main effects that are implicit in the data (Han, 2017). In both cases, the population control totals were from the full target population U which has $N = 50,000$.

Both model-based estimators were computed using `calibrate` in the R `survey` package (Lumley, 2017). The default variance estimator (labeled "wr" in the tables below) in the `survey` package is approximately equal to the right-hand side of equation (7). The grouped jackknife in (8) was also calculated; all calibration steps were repeated separately for each group.

Doubly robust. These estimators were computed by first computing quasi-randomization weights and then using linear model calibration to full population totals of covariates based on a main effects model. Two variance estimators were computed: (1) one that would be appropriate for single-stage with-replacement sampling that treats the weights as if they were inverse selection probabilities, and (2) the grouped jackknife in (8) that repeated the quasi-randomization and calibration steps separately for each jackknife group. The former is expected to be an underestimate because it does not reflect the variation due to estimating pseudo-probabilities or to calibrating.

Multilevel Regression and Poststratification. Bayesian models were used for MRP using the R package `rstanarm` (Stan Development Team, 2016). For the MRP estimates of the proportions of `SMOKE100`, the individual categories of `GENHLTH`, `FairPoor`, and `GoB` a logistic regression

model was fitted to the binary, person-level indicator variables using `stan_glmer` in `rstanarm`. The models included a fixed-effect intercept and random effects for all main effects and two-way interactions for `AGECAT`, `RACECAT`, `EDCAT`, `INCOMC3`. Mean general health was fitted with a linear regression model that included the same four main effects and two-way interactions. The prior on the intercept was normal with a variance of 100; several specifications for the prior mean were tried, including the overall sample proportion, its log-odds, and zero. All choices led to essentially the same simulation results. Noninformative, flat priors were used for the regression parameters.

For comparison the following alternative way of model-fitting was also tested, but detailed results are not reported here. The unweighted sample proportion or mean was computed for each PS that was present in a sample. The same random effects model as above was fitted, treating the PS proportions or means as normally distributed. We used the structured prior for variances of regression coefficients described in (Si et al., 2017). Their formulation is designed to de-emphasize unstable, direct PS estimates by shrinking them toward an overall estimate. Simulations of $n = 500, S = 500$ were run with the structured prior method for excellent health and mean general health with biases and coverage of credible intervals being almost the same as the ones reported below for the models on individual y values.

For each sample, Markov chain Monte Carlo (MCMC) estimation was used with two chains. Each chain had 500 iterations with the last 250 in each chain used for estimation, i.e., a total of 500 MCMC estimates for each sample. If one were analyzing a single sample, more chains and iterations would be advisable. As a partial check on whether two chains and 500 iterations were adequate for assessing properties, we ran a simulation of 200 samples of $n = 500$ with each using four chains of 1,000 iterations. This was done for the proportion of persons who rated themselves as being in excellent health and produced summary results that were virtually the same as for two chains and 500 iterations.

The poststrata were the set of 273 combinations of the levels of `AGECAT`, `RACECAT`, `EDCAT`, `INCOMC3` that occurred in the population. Samples of 500 contained an average of 133 of the PS while samples of 1,000 included an average of 165 of the poststrata. For the proportions of the binary variables, posterior predictions (which were 0's or 1's) were made from each of the 500 saved iterations in

each sample and the average prediction calculated across the 500. These average model predictions were made for all 50,000 persons in the population. The means of the posterior predictions were computed for each of the 273 PS in the population. Note that this produces a value of $\hat{\mu}_\gamma$ in all PS, including those not represented in the sample. The poststratified estimate was then computed using expression (6).

2.3 Results

Results were summarized over the sets of 10,000 samples using the statistics described below.

- Empirical percent relative bias of an estimated proportion or mean

$$relbias(\hat{\theta}) = \frac{100}{\theta} \times S^{-1} \sum_{s=1}^{S} \left(\hat{\theta}_s - \theta\right)$$

  where s is a simulated sample, $S = 10,000$ is the total number of samples selected ($S = 500$ for MRP), $\hat{\theta}_s$ is the estimate from sample s and θ is the full population value.

- Empirical root mean square error ($rmse$)

$$rmse(\hat{\theta}) = \sqrt{S^{-1} \sum_{s=1}^{S} \left(\hat{\theta}_s - \theta\right)^2}$$

- Empirical percent relbias of standard error estimators for quasi-randomization, linear model, and doubly robust estimates

$$relbias\left(\sqrt{v(\hat{\theta})}\right) = 100 \times \left(S^{-1} \sum_{s=1}^{S} \sqrt{v\left(\hat{\theta}_s\right)} - \sqrt{Var(\hat{\theta})}\right) \Big/ \sqrt{Var(\hat{\theta})}$$

  where $v(\hat{\theta}_s)$ is a variance estimator in sample s and $Var(\hat{\theta})$ is the empirical variance of $\hat{\theta}$ across all samples. For MRP no variance estimates were computed.

- 95% confidence interval coverage for quasi-randomization, linear model, and doubly robust estimates. t-statistics were computed for each sample as $t\left(\hat{\theta}_s\right) = \left(\hat{\theta}_s - \theta\right) \Big/ \sqrt{v\left(\hat{\theta}_s\right)}$. Con-

fidence interval coverage was calculated as

$$CI.cov\left(\hat{\theta}\right) = 100 \times S^{-1} \sum_{s=1}^{S} I\left(-1.96 \leq t\left(\hat{\theta}_s\right) \leq 1.96\right)$$

where $I(\cdot)$ is the indicator of the event in parentheses.

- 95% credible interval coverage for MRP. In each simulated sample, 500 poststratified estimates were computed using the MCMC results. A credible interval was formed with endpoints equal to the 2.5 and 97.5 percentiles of the 500 PS estimates. Coverage was the percentage across the $S = 500$ samples that the intervals contained the population value.

Table 5 shows the percent relbiases for the estimated proportions and means for the seven estimands, four estimators, and two sample sizes. Figure 4 plots the relbiases of the different estimators. MRP is the most biased in six of seven cases when $n = 500$ and 3 of 7 cases when $n = 1,000$. Quasi-randomization estimates are typically more biased than the linear model and raking estimates. Of the two M1 model-based estimators, raking generally has somewhat smaller absolute biases than M1(linear). Overall, the doubly robust estimates have near the smallest absolute bias for all variables and both sample sizes.

Table 6 gives the *rmse*'s of each estimator. For $n = 500$ and 1,000 quasi-randomization and MRP have somewhat larger *rmse*'s for five of seven variables. For example when $n = 500$, the quasi-randomization *rmse* for mean health is 24% larger than that of the doubly robust estimator (0.082/0.066) while MRP's *rmse* is 38% larger (0.091/0.066); when $n = 1,000$, the quasi-randomization *rmse* is 35% larger than that of the doubly robust estimator (0.70/0.052); while MRP's *rmse* is 38% larger (0.072/0.052).

Table 7 presents the percent relbiases of the standard error estimators for quasi-randomization, M1(linear), M1(raking), and doubly robust. (As noted above, variance estimates were not computed for MRP.) The with-replacement SE estimators are nearly unbiased for the quasi-randomization estimators when $n = 500$ and 1 to 2 percent overestimates when $n = 1,000$. The model-based wr estimators tend to be underestimates at both sample sizes for both the linear and raking estimators. For the doubly robust estimator the wr is generally an overestimate for both sample sizes with the

Table 5: Percent relbiases of estimators in a simulation of 10,000 samples; 500 samples for MRP.

Variable	Quasi-rand	M1 (linear)	M1 (raking)	Doubly robust	MRP
		n = 500			
Smoke100	-0.2	3.2	1.6	1.7	-1.3
Health excellent	6.9	4.9	5.2	4.9	12.9
Health very good	2.9	1.0	0.3	0.5	5.2
Health Good	-0.7	0.6	0.9	0.8	-1.6
Health fair or poor	-13.1	-8.9	-8.3	-8.2	-13.7
GoB health	2.5	1.7	1.6	1.5	2.6
Mean health	-2.3	-1.4	-1.3	-1.3	-3.0
		n = 1,000			
Smoke100	-0.3	3.3	1.3	1.5	0.0
Health excellent	7.1	4.7	3.9	4.9	11.9
Health very good	3.0	1.1	0.0	0.6	4.1
Health Good	-0.7	0.5	2.5	0.8	-0.9
Health fair or poor	-13.4	-8.8	-9.4	-8.2	-10.7
GoB health	2.5	1.7	1.8	1.5	2.0
Mean health	-2.3	-1.4	-1.2	-1.3	-2.4

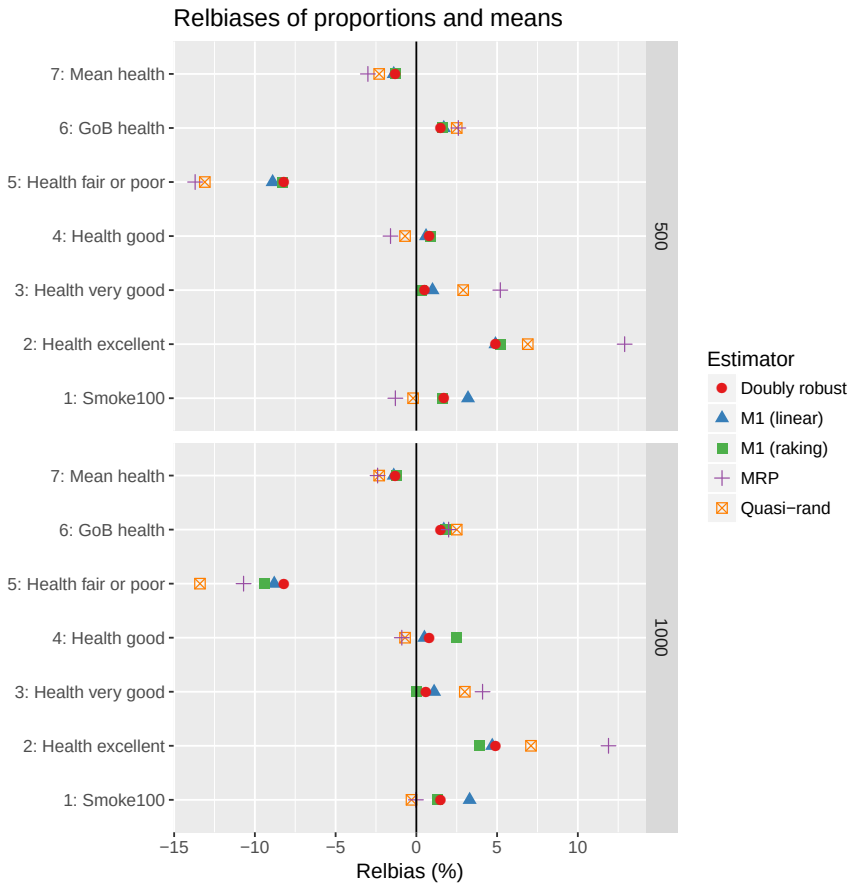

Figure 4: Relbiases of five estimators for seven estimands in 10,000 samples; 500 sample for MRP

Table 6: Root mean squared errors of estimators in a simulation of 10,000 samples; 500 samples for MRP.

Variable	Quasi-rand	M1 (linear)	M1 (raking)	Doubly robust	MRP
		$n = 500$			
Smoke100	0.031	0.035	0.032	0.032	0.029
Health excellent	0.025	0.023	0.024	0.023	0.030
Health very good	0.030	0.028	0.028	0.028	0.030
Health Good	0.029	0.029	0.031	0.030	0.025
Health fair or poor	0.032	0.028	0.029	0.028	0.032
GoB health	0.032	0.028	0.029	0.028	0.032
Mean health	0.082	0.067	0.068	0.066	0.091
		$n = 1000$			
Smoke100	0.021	0.027	0.023	0.023	0.019
Health excellent	0.019	0.017	0.017	0.017	0.025
Health very good	0.022	0.020	0.019	0.020	0.022
Health Good	0.020	0.020	0.022	0.021	0.017
Health fair or poor	0.027	0.021	0.023	0.022	0.025
GoB health	0.027	0.021	0.023	0.022	0.025
Mean health	0.070	0.053	0.050	0.052	0.072

relbias being about 10% for mean health. The grouped jackknife is a 3 to 5% overestimate for quasi-randomization for $n = 500$ and somewhat less of an overestimate for the larger sample size. The jackknife is nearly unbiased for M1(linear) but has a slight positive bias for the raking estimator. For the doubly robust estimator, the jackknife is always positively biased, although the bias decreases for $n = 1,000$.

Table 7: Percent relbiases of with-replacement and grouped jackknife standard error estimators in simulations of 10,000 samples; 500 samples for MRP. wr = with-replacement; JK = grouped jackknife

	Quasi-rand		M1 (linear)		M1 (raking)		Doubly robust	
Variable	wr	JK	wr	JK	wr	JK	wr	JK
				$n = 500$				
Smoke100	-0.7	3.8	-1.3	1.0	0.1	2.5	1.2	3.3
Health excellent	0.2	3.3	-0.7	1.1	-1.1	1.6	3.4	2.7
Health very good	0.3	4.3	-1.0	1.4	-1.1	2.6	4.1	3.5
Health Good	-1.3	3.1	-2.8	-0.4	-3.2	1.1	0.5	3.3
Health fair or poor	-0.8	3.8	-4.2	-1.0	-4.9	0.8	4.9	3.5
GoB health	-0.8	3.8	-4.2	-1.0	-4.9	0.8	4.9	3.5
Mean health	1.5	5.1	-2.5	0.0	-3.0	1.9	10.5	4.6
				$n = 1000$				
Smoke100	1.1	2.6	0.4	0.8	3.2	2.4	2.6	2.5
Health excellent	2.2	2.9	0.0	0.0	1.1	1.7	4.3	1.8
Health very good	1.7	3.2	-0.3	0.2	1.6	2.3	4.2	1.9
Health Good	0.8	2.3	-1.1	-0.9	-0.2	0.9	1.4	1.7
Health fair or poor	1.5	2.8	-1.0	-0.4	-0.1	1.9	7.6	3.2
GoB health	1.5	2.8	-1.0	-0.4	-0.1	1.9	7.6	3.2
Mean health	2.3	2.7	-1.2	-0.9	0.6	2.5	9.9	1.9

Table 8 shows the coverage rates of 95% normal approximation CIs and 95% credible intervals for MRP. Figure 5 plots the coverage rates for the estimators, excluding MRP whose coverage for mean health would distort the scale. Since MRP has noticeably different properties for two of the variables, we will discuss it separately below. Having an estimated SE that is positively biased is often an advantage when constructing normal approximation confidence intervals (CIs); this is true here for quasi-randomization, M1(linear), M1(raking), and doubly robust. For `SMOKE100`,

excellent health, very good health, and good health, CI coverage is within about 2 percentage points of 95%. For fair or poor health, good or better health, and mean health, coverage rates are often poor with the worst being 71.6% for mean health for (quasi-randomization, wr) when $n = 1,000$. CI coverage is often worse for the larger sample size because the biases of the point estimators of means and proportions result in the CIs not being centered on the population values, and, as the SEs decrease with the larger sample size, the length of the CIs becomes shorter. The jackknife typically gives better coverage rates than the wr estimator, but the advantage is slight. The model-based linear, raking, and the doubly robust estimators give noticeably better coverage with either wr or JK variance estimators for fair or poor health, GoB, and mean health than does the quasi-randomization approach with either type of variance estimator.

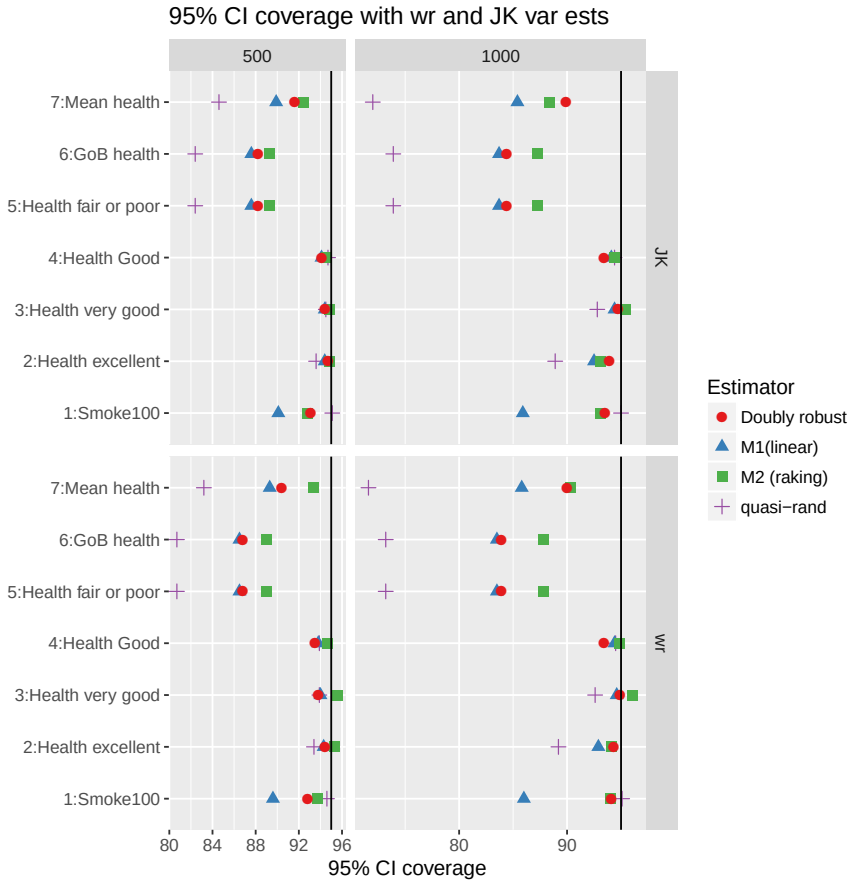

Figure 5: Coverage of 95% confidence intervals for four estimators and seven estimands in 10,000 samples

Coverage of the MRP credible intervals is over 90% for `SMOKE100`, very good health, and good health. MRP coverage is notably poor, and worse than for the other estimators, for excellent health

(79.4% when $n = 500$; 71% for $n = 1,000$) and mean health (68.4% when $n = 500$; 64.8% when $n = 1,000$).

To explore how well MRP did at predicting individual PS means, a separate set of 100 samples each of size $n = 500$ was selected using the same stratified design as in Table 1. The average posterior prediction of mean health was calculated for each poststratum across the 100 samples. These are plotted versus population PS means in Figure 6. Separate panels are shown based on how often the PS were included in the 100 samples—0 to 0.2 of the samples in the first panel, 0.2 to 0.5 in the second, 0.5 to 0.9 in the third, and over 0.9 in the fourth. The posterior predictions on average are too large when the population PS means are small and are too small when the population means are large. The same pattern holds regardless of the rate at which poststrata are covered by the samples. This is evidence that the more extreme posterior means are being inefficiently shrunk toward the overall average.

It might be argued that inclusion of more covariates would make MRP more competitive with the other estimates. However, in this population no other covariates were available that could be reasonably used in modeling. Even if they had been available, additional covariates would also improve the performance of the other four estimators in this study making it difficult for MRP to catch or surpass them.

3 Conclusion

This article compared the performance of five types of estimators of population means and proportions using samples whose probabilities of observation were unknown to an analyst. The samples were constructed to substantially undercover the desired target population—a situation that likely applies to many nonprobability samples (and probability samples with very low response rates). Using covariates whose totals were known for the target population, estimators were constructed based on quasi-randomization and superpopulation models.

The success of these approaches was mixed at best. The doubly robust estimators were generally

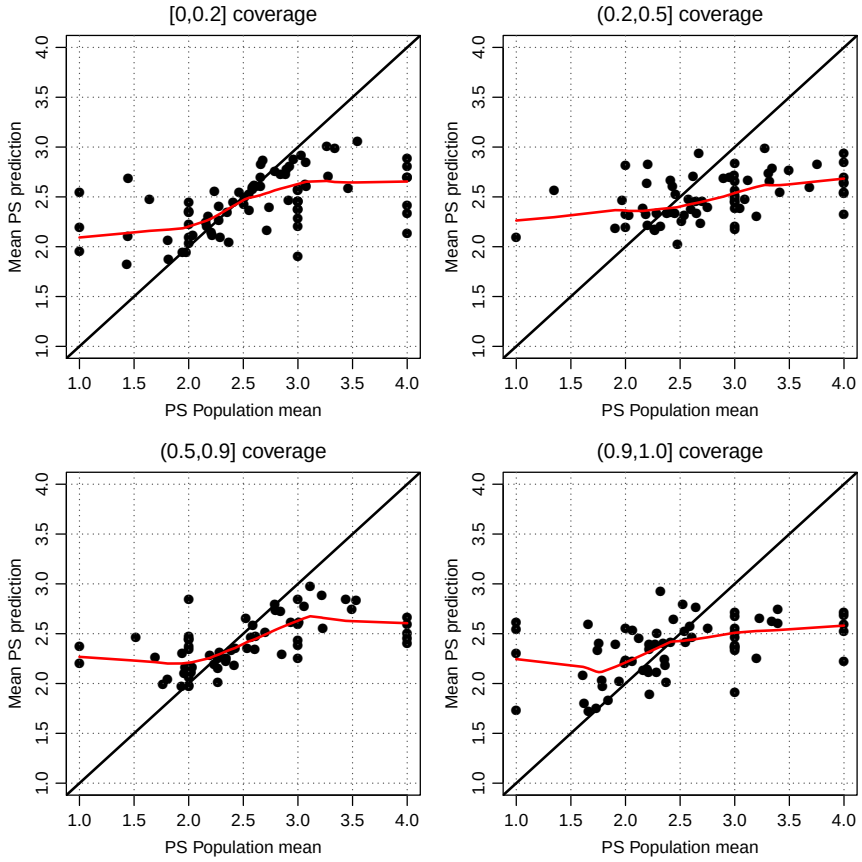

Figure 6: Plot of average sample MRP PS means of mean health vs. population PS means for 100 samples. Separate panels for proportion of samples in which poststrata had nonzero sample sizes. Red line is a nonparametric smoother. A 45^o line is drawn in each plot.

Table 8: Coverage of 95% confidence intervals constructed with with-replacement and grouped jackknife standard error estimators in a simulation of 10,000 samples; 500 samples for MRP credible intervals. wr = with-replacement; JK = grouped jackknife

	Quasi-rand		M1 (linear)		M1 raking		Doubly robust		MRP
Variable	wr	JK	wr	JK	wr	JK	wr	JK	
				$n = 500$					
Smoke100	94.6	95.1	89.6	90.1	92.8	93.1	93.7	92.8	93.0
Health excellent	93.4	93.6	94.3	94.4	94.4	94.7	95.3	94.8	79.4
Health very good	93.9	94.5	94.0	94.4	93.8	94.4	95.6	94.8	90.2
Health Good	93.9	94.7	93.8	94.1	93.5	94.1	94.7	94.6	94.4
Health fair or poor	80.7	82.4	86.5	87.6	86.8	88.2	89.0	89.3	81.2
GoB health	80.7	82.4	86.5	87.6	86.8	88.2	89.0	89.3	81.2
Mean health	83.2	84.6	89.3	89.9	90.4	91.6	93.4	92.4	68.4
				$n = 1000$					
Smoke100	95.1	95.0	86.0	85.9	94.1	93.5	94.0	93.1	94.4
Health excellent	89.2	88.9	92.9	92.5	94.3	93.9	94.1	93.1	71.0
Health very good	92.6	92.8	94.6	94.4	94.9	94.7	96.1	95.4	90.0
Health Good	94.5	94.4	94.4	94.1	93.4	93.4	94.9	94.4	98.2
Health fair or poor	73.2	73.9	83.5	83.7	83.9	84.4	87.8	87.3	81.2
GoB health	73.2	73.9	83.5	83.7	83.9	84.4	87.8	87.3	80.4
Mean health	71.6	72.0	85.8	85.4	90.0	89.9	90.3	88.4	64.8

least biased; *rmse*s of the doubly robust and model-based estimators were usually smaller than those of the quasi-randomization or MRP estimators. However, for some variables, estimates had substantial biases that did not decrease with increasing sample size. Thus, use of covariates for estimation could not always overcome the population undercoverage problem, even in large samples.

Two methods of variance estimation were used for estimators other than MRP. The first was appropriate under an independence randomization model or an independence superpopulation model. The second was a grouped jackknife estimator in which all estimation steps were repeated for every jackknife group. The jackknife is generally preferable for variables where point estimators are nearly unbiased since its confidence interval coverage was nearer the nominal level. However, when the relative biases of point estimators are more than 2 or 3%, neither variance estimator provides good CI coverage. Credible intervals based on the MRP estimators covered at less than the desired rate with coverage being extremely poor for some variables.

Overall, a doubly robust estimator in combination with the jackknife variance estimator was the best combination in this study in terms of bias of the point estimator, its root mean square error, and the coverage of its associated confidence intervals. Additionally, weights can be computed for the doubly robust estimator, which analysts can handle as they normally do for survey estimation. Multilevel regression and poststratification may have advantages in applications, like election polling, where very strong covariates are available, but it did not do well in this study.

When making finite population estimates from a nonprobability sample, the worry is how "representative" the sample is of the full population. Does the sample contain all important subgroups whose characteristics differ from each other? If the sample has coverage problems, the unrepresented groups must be enough like sample groups that their values can be predicted using models fitted from the sample. There are cases where this can be done despite what appear to be insurmountable problems. For example, using multilevel regression and poststratification (MRP), Wang et al. (2015) were able to predict the outcome of the 2012 U.S. presidential election using a sample whose demographic distribution was vastly different from the full electorate. However, an election poll is a specialized application, and they had access to powerful covariates such as party affiliation and how a person had voted in the previous presidential election. The results here are consistent

with those in Mercer et al. (2018) who found that even with their most judiciously chosen set of covariates, they were unable to remove biases from all estimates.

A difficulty with nonprobability samples (and even with probability samples) is that in a particular survey, we never really know how close to the truth estimates are. In probability samples where the frame correctly covers the full population, practitioners can fall back on the argument that, in the long run, the method will produce results that are correct on average. No such repeated, controlled sampling justification exists for nonprobability samples. If the models are wrong, then the estimates are wrong. In an election poll, survey designers do get a report card on how their polls did—polling results either predict the eventual winner or not. In most applications the population values will never be known. Whether the models for inclusion probabilities or those for analysis variables accurately predict nonsample values will never be known.

Some means of validating results is needed, although devising tests is difficult. Including validation questions (e.g., employment status or self-reported health status) from large, well-designed probability surveys, like the American Community Survey or National Health Interview Survey, would seem to be one way to gauge the accuracy of a nonprobability sample. Of course, a shortcoming of such tests is that the validation questions may have different potential for bias than the items that are really of interest in the survey.

References

Buttice, M. and Highton, B. (2013). How does multilevel regression and poststratification perform with conventional national surveys? *Political Analysis*, 21:449–467.

Cavallo, A. and Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *The Journal of Economic Perspectives*, 30(2):151–178.

Chen, J. K. T., Valliant, R., and Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive lasso. *Survey Methodology*. under review.

Couper, M. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7:145–156.

Dever, J. A. and Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*, 36:45–56.

Dever, J. A. and Valliant, R. (2016). General regression estimation adjusted for undercoverage and estimated control totals. *Journal of Survey Statistics and Methodology*, 4:289–318.

Dong, Q., Elliott, M. R., and Raghunathan, T. (2014). A nonparametric method to generate synthetic populations to adjust for complex sample designs. *Survey Methodology*, 40:29–46.

Elliott, M. R. and Valliant (2017). Inference for nonprobability samples. *Statistical Science*, 32:249–264.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164.

Gelman, A. and Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23:127–135.

Han, D. (2017). *Investigation of Alternative Calibration Estimators in the Presence of Nonresponse*. PhD thesis, University of Maryland. https://drum.lib.umd.edu/handle/1903/19939.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning*. Springer, New York.

Kohut, A., Keeter, S., Doherty, C., Dimock, M., and Christian, L. (2012). Assessing the representativeness of public opinion surveys. http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22:329–349.

Little, R. J. A. (2003). Bayesian methods for unit and item nonresponse. In Chambers, R. and Skinner, C., editors, *Analysis of Survey Data*, chapter 18. John Wiley, Chichester.

Lumley, T. (2017). *survey: analysis of complex survey samples R package v. 3.32*.

Matei, A. (2018). On some reweighting schemes for nonignorable unit nonresponse. *Survey Statistician*, (77):21–33.

Mercer, A., Lau, A., and Kennedy, C. (2018). For weighting online opt-in samples, what matters most? Technical report, Pew Research Center, Washington DC. http://www.pewresearch.org/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/.

Murphy, J., Link, M., Childs, J., Tesfaye, C., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M., and Harwood, P. (2015). Social media in public opinion research. *Public Opinion Quarterly*, 78:788–794.

Park, D., Gelman, A., and Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: Statelevel estimates from national polls. *Political Analysis*, 12:375–385.

Park, D. K., Gelman, A., and Bafumi, J. (2006). State-level opinions from national surveys: Poststratification using multilevel logistic regression. In Cohen, J. E., editor, *Public Opinion in State Politics*. Stanford University Press.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Schonlau, M., van Soest, A., and Kapteyn, A. (2007). Are "Webographic" or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods*, 1(3):155–163.

Selb, P. and Munzert, S. (2011). Estimating constituency preferences from sparse survey data using auxiliary geographic information. *Political Analysis*, 19:455–470.

Shapiro, R. Y. (2011). Public opinion and American democracy. *Public Opinion Quarterly*, 759:982–1017.

Si, Y., Trangucci, R., Gabry, J., and Gelman, A. (2017). Bayesian hierarchical weighting adjustment and survey inference. https://arxiv.org/abs/1707.08220.

Stan Development Team (2016). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.15.3.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288.

Valliant, R., Dever, J., and Kreuter, F. (2017). *PracTools: Tools for Designing and Weighting Survey Samples, R package version 0.8*. https://CRAN.R-project.org/package=PracTools.

Valliant, R., Dever, J. A., and Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York, 2nd edition.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley & Sons, Inc., New York.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31:980–991.

Zhou, H., Elliott, M. R., and Raghunathan, T. (2016). A two-step semiparametric method to accommodate sampling weights in multiple imputation. *Biometrics*, 72:242–252.