DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 721–752, http://dx.doi.org/10.2478/JOS-2018-0034

Accounting for Complex Sampling in Survey Estimation: A Review of Current Software Tools

Brady T. West[1], *Joseph W. Sakshaug*[2], *and Guy Alain S. Aurelien*[3]

In this article, we review current state-of-the art software enabling statisticians to apply design-based, model-based, and so-called "hybrid" approaches to the analysis of complex sample survey data. We present brief overviews of the similarities and differences between these alternative approaches, and then focus on software tools that are presently available for implementing each approach. We conclude with a summary of directions for future software development in this area.

Key words: Complex sample survey data; statistical software; design-based analysis; model-based analysis; multilevel modeling.

1. Introduction

Secondary analysis of survey data arising from complex sample designs is a ubiquitous research methodology in many applied fields. The "complex" terminology refers to features of sample designs that deviate from a design featuring simple random sampling with replacement, which in a finite population sampling framework is in accord with the theoretical notion of independent and identically distributed data. These complex design features, which generally include unequal probabilities of selection into the sample, cluster sampling, and stratification of the target population prior to sampling (Heeringa et al. 2017), need to be accounted for by secondary data analysts and applied statisticians who have many tools at their disposal for analyzing these types of data sets. A failure to account for these design features in analysis can lead to substantially biased inferences (e.g., Skinner et al. 1989; West et al. 2016; Heeringa et al. 2017). Over a period of more than 80 years, many different methods have been proposed by statisticians and survey methodologists for correctly accounting for these sample design features when performing survey data analysis.

The variety of approaches discussed and proposed in the survey statistics literature can generally be grouped into two main categories: *design-based analysis*, where the

[1] Survey Research Center, University of Michigan-Ann Arbor, 4118 Institute for Social Research, 426 Thompson Street Ann Arbor, MI, 48106, U.S.A. Email: bwest@umich.edu
[2] Institute for Employment Research, Regensburger Strasse 104, Nuremberg, 90478, Germany. Email: joe.sakshaug@iab.de
[3] Walter R. McDonald & Associates, 12300 Twinbrook Pkwy, Suite310, Rockville, MD 20852, U.S.A. Email: alainshamir@gmail.com

Acknowledgments: This work was supported by a grant from the U.S. National Science Foundation (NCSE-1424081).

randomized selection mechanism underlying the probability sampling governs all subsequent inference, and *model-based analysis*, where all inference depends on probability models posited by the analyst (Hansen et al. 1983). More recently (e.g., Little 2015), statisticians have advocated "hybrid" approaches that combine optimal properties of model-based and design-based approaches. A statistician responsible for analyzing survey data therefore needs to select one or more of these approaches to employ, depending on the objectives of a researcher's study and the parameters of scientific interest. And, once an approach has been selected, the statistician needs to identify software that implements the selected approach. In the present article, we aim to provide statisticians and survey researchers with an up-to-date review of state-of-the-art statistical software capable of implementing each of these different approaches, depending on the specific analysis of interest.

When thinking about these alternative approaches and the software tools implementing them, one needs to consider the objectives of a given analysis of survey data. Is one merely interested in generating descriptive inferences (means, proportions, totals, etc.), or is one also interested in more "analytic" objectives (regression coefficients, odds ratios, etc.)? The identification of appropriate software requires a cross-classification of "objective" (descriptive vs. analytic) and "approach" (design-based vs. model-based); see Table 1. We note that so-called "hybrid" approaches to analytic studies combine features of both design-based and model-based approaches. In the discussion moving forward, we assume that a formal probability sampling plan has been used to select a given sample from a finite population, and that the analyst is weighing different analysis approaches with this sample in hand. We do not consider software for analyzing data from non-probability samples, which are currently receiving a great deal of research attention (e.g., Baker et al. 2013; Elliott and Valliant 2017), in this article.

This article reviews state-of-the-art software tools in each of the five domains indicated in Table 1. Modern survey statisticians need to speak multiple computing languages in general, understanding the pros and cons of each, and effectively communicate software alternatives for clients who desire to analyze survey data. Not all software packages share the same capabilities for analyzing complex sample survey data, and we aim to review the state of the art in this regard. The article is structured as follows. In each of Sections 2 through 6, we first present a brief overview of one of the five approaches in Table 1, and then review current software tools that are available for implementing that particular approach. We then conclude in Section 7 with a summary of important directions for future software development in this area.

Table 1. Five possible combinations of research objectives and analysis approaches, to guide a review of current software for the analysis of complex sample survey data.

<table>
<tr><td></td><td>Design-based approaches</td><td>Model-based approaches</td></tr>
<tr><td>Descriptive objectives</td><td>1</td><td>2</td></tr>
<tr><td rowspan="2">Analytic objectives</td><td>3</td><td>4</td></tr>
<tr><td colspan="2">"Hybrid" approaches (5)</td></tr>
</table>

2. Descriptive Objectives: Design-Based Approaches

2.1. Weighted Estimation

When analysts employ design-based approaches to the descriptive analysis of survey data, their analytic objectives generally involve design-unbiased estimation (i.e., estimation that is unbiased with respect to the probability sample design used) of simple descriptive parameters characterizing a finite target population, such as means, proportions, totals, percentiles, and row or column percentages in contingency tables. These approaches generally feature weighted estimation of the parameters of interest, in addition to design-unbiased, nonparametric estimation of sampling variance for the weighted estimates and design-adjusted tests of associations between variables (e.g., Rao and Scott 1984). These approaches are quite popular among nonstatisticians because they are widely implemented in different statistical software packages, and they yield robust population inferences that do not require parametric assumptions regarding the variables of interest.

In general, the respondent weights computed by organizations collecting and producing survey data account for three key aspects of the sample design and the data collection: 1) unequal probabilities of selection into the sample for different population elements, 2) adjustment for nonresponse during data collection, and 3) calibration of the (possibly adjusted) respondent weights to known population totals (Kish 1965; Kalton and Flores-Cervantes 2003; Lohr 2009; Valliant et al. 2013; Lavallee and Beaumont 2016; Heeringa et al. 2017; Haziza and Beaumont 2017). The first element of a respondent weight is generally referred to as a *design weight*. The design weight for a given sampled unit is defined as the inverse of the probability of inclusion for that unit in a given sample, and these design weights can be computed for *all* sampled units in a probability sample (where every population element has a known nonzero probability of inclusion), including respondents and nonrespondents. Inference in design-based approaches is driven by these probabilities of selection, and these components of the weight ensure that estimates computed using the weights appropriately reflect the probability of selection for a given case from a specified target population. Under an extremely unusual scenario where 100 percent of the sampled population units respond to a survey request, one could compute population estimates of target parameters that are unbiased with respect to the sample design using this single design weight.

Unfortunately, not all sampled population units will respond to a survey request. If nonresponding units differ systematically from responding units in terms of key features of interest, nonresponse bias in estimates computed using design-based approaches may result. For this reason, the design weights are often adjusted to account for differential nonresponse among different population subgroups, treating the probability of responding as an additional stochastic stage of sample selection (Cassel et al. 1983; Särndal and Swensson 1987; Ekholm and Laaksonen 1991), and multiplying the design weights for responding units by the inverse of their response probability. Because these probabilities of response are not known in practice, they need to be estimated. Given auxiliary data for respondents and nonrespondents that are generally predictive of both the probability of responding and key survey variables (Lessler and Kalsbeek 1992; Bethlehem 2002; Kalton and Flores-Cervantes 2003; Little and Vartivarian 2005; Beaumont 2005; Groves 2006;

Kreuter et al. 2010), the literature provides extensive guidance on optimal methods for estimating these response probabilities and using them to adjust the design weights for nonresponse (Little 1986; Ekholm and Laaksonen 1991; Eltinge and Yansaneh 1997; Grau et al. 2006; Wun et al. 2007; Haziza and Beaumont 2007; West 2009; Kott 2012; Valliant et al. 2013; Brick 2013; Flores-Cervantes and Brick 2016).

The next (and generally final) step in computing adjusted design weights is to calibrate the (nonresponse adjusted) weights for responding units to sum to known population control totals, ensuring sound population representation in terms of the marginal distributions of (generally sociodemographic) population characteristics. There is a vast literature on this topic (Deville and Särndal 1992; Lundström and Särndal 1999; Rao 2005; Kott 2006; Kim and Park 2010; Kott 2011), and one can use a variety of approaches to perform calibration adjustments in practice, including poststratification (Holt and Smith 1979), raking (Oh and Scheuren 1983; Deville et al. 1993), and generalized regression estimation (Valliant et al. 2000), which can utilize population information for both continuous and categorical variables. Kott and Liao (2012) outline a calibration procedure implemented in the WTADJUST procedure of the SUDAAN software that builds on the developments in prior calibration literature to provide "double protection" against misspecification of either a substantive model or a response model (based on the auxiliary variables used in the calibration adjustment) when using calibration for nonresponse adjustment.

The WesVar software produced by Westat (https://www.westat.com/our-work/information-systems/wesvar®-support), the `calibrate()` function in the R `survey` package (Lumley 2010), the `ipfraking` user-written package in Stata (Kolenikov 2014), the `sreweight` user-written command in Stata (Pacifico 2014), and the CALMAR 2 software developed by Le Guennec and Sautory (2002) are also capable of computing calibration adjustments to design weights based on the methods described above, given population information on the chosen auxiliary variables (see http://vesselinov.com/CalmarEngDoc.pdf for more details on the various calibration options in the CALMAR 2 software). The final calibrated weights may then be trimmed to minimize the impact of weight variance on the precision of weighted survey estimates (Potter 1990; Elliott and Little 2000; Kalton and Flores-Cervantes 2003; Beaumont 2008; see also Asparouhov and Muthén 2007 for optimal weight trimming approaches using the Mplus software). The weights that result from this process then need to be input by analysts into software procedures enabling design-unbiased point estimation of population parameters (see Subsection 2.4).

We note that the final overall respondent weights that result from this three-step process are essentially adjusted versions of the design weights, but software procedures enabling design-based analysis treat these final respondent weights as if they were design weights that are "known" with certainty. Because *estimates* of response propensity are often used to adjust the design weights for unit nonresponse, this uncertainty in the final respondent weights should be accounted for in variance estimation. This is best handled using replication techniques, as outlined in Subsection 2.2 below, where the adjustment process (based on estimates) can be repeated for each replicate sample, and the variance in the adjustments across replicates is incorporated into the final variance estimates (Valliant 2004).

2.2. Variance Estimation

Taylor Series Linearization (TSL) is a design-based variance estimation technique that is widely implemented in different statistical software procedures and often serves as a default variance estimation procedure in these procedures when applying design-based approaches to complex samples. The basic idea behind TSL is to use a Taylor series expansion to approximate a non-linear estimator (e.g., a ratio mean, a ratio estimator of total, a regression parameter, a correlation coefficient) using a *linear* function of estimated sample totals. Once the nonlinear estimator is "linearized," then unbiased, design-based variance estimation formulae reflecting the complex sampling features (stratification, cluster sampling, weighting) can be applied to estimate the variance of the linear function of sample totals. The variance of the linearized estimator is estimated within each stratum (if applicable), and the stratum variances are combined to produce the total variance of the estimator. Wolter (2007, Chapter 6) reviews the TSL literature and provides technical details.

There are two important issues that analysts need to handle carefully when employing TSL for design-based variance estimation: subpopulation analysis, and "singleton" sampling clusters. First, considering subpopulation analysis, complex sample designs often employ the sampling of clusters of population elements within sampling strata for reasons of cost efficiency. The clusters sampled at the first stage of random selection (possibly within strata) are often referred to as primary sampling units, or PSUs, and these could be geographic areas in area probability samples, naturally occurring groups of population elements (e.g., colleges), or individual sampled elements if no cluster sampling is employed (software enabling design-based analysis will estimate variances under this assumption if no cluster ID variables are indicated). When analyzing subpopulations (e.g., elderly males) and using TSL for variance estimation, analysts need to explicitly form binary variables indicating which sampled cases fall into the subpopulation of interest, and use these indicators for variance estimation (which is often facilitated by "subpopulation" options in the different software procedures, e.g., the `subpop()` option in Stata). This approach enables PSUs with no sample from the subpopulation to still be accounted for in the variance estimation (in that they contribute totals of zero for the variables of interest), rather than being removed entirely. The physical removal of entire PSUs due to the deletion of cases that do not belong to a subpopulation can lead to scenarios where sampling strata only have a single PSU present, preventing variance estimation within that stratum when using TSL (more on this below). See West et al. (2008) and Heeringa et al. (2017) for more on this TSL-specific issue, which becomes irrelevant when using *replication methods* for variance estimation (as clusters with no subpopulation sample simply do not contribute to replicate estimates).

Second, considering the "singleton" sampling cluster issue, some PSUs may also be selected with *certainty*, meaning (in a design-based setting) that they would be included in every possible hypothetical sample that might be selected; that is, they have a probability of inclusion of one. When employing TSL for variance estimation, there need to be at least two PSUs present within a sampling stratum to estimate the contribution of that stratum to the overall sampling variance, and certainty PSUs often define their own stratum (e.g., the city of New York in the United States). Data producers can facilitate variance estimation using TSL by dividing the sampled elements in a certainty PSU into multiple *random* *groups* (Wolter 2007), and providing codes for these “pseudo clusters” in a public-use data set. If a data user for some reason encounters a stratum with only a single PSU code present in such a data set, most design-based software will provide some form of ad-hoc solution for estimating the contribution of that stratum to the overall sampling variance. For example, Stata provides users with several choices via the `singleunit()` option in the `svyset` command, which is used to identify PSUs for variance estimation; SUDAAN provides the MISSUNIT option (see http://sudaansupport.rti.org/sudaan/page.cfm/Theory); and the `survey` package in R provides the user with a variety of global options (see http://faculty.washington.edu/tlumley/old-survey/exmample-lonely.html for examples).

Replication methods represent a second nonparametric design-based approach to estimating the variance of a weighted estimate. In general, these methods involve dividing the full sample into various subsamples, calculating an estimate of the parameter of interest within each subsample, and calculating the variation among the subsample estimates to estimate the variance of the full sample estimate. These methods can be implemented in various forms, including the random groups method (RGM), Jackknife repeated replication (JRR), balanced repeated replication (BRR), bootstrapping, and various modifications of these methods (Wolter 2007; Shao and Tu 1995). A key advantage of these replication methods is that they do not require the linearization of a nonlinear estimator (Krewski and Rao 1981), and can generally be applied to many different forms of estimators. These methods also enable survey organizations to disseminate public-use survey data sets including (adjusted) weights for each of the replicate samples in lieu of stratum and PSU codes, minimizing the risk of identifying survey respondents within small PSUs. This requires the data user to employ variance estimation software that supports the specific type of replication weighting scheme used by the survey organization, and nearly all major statistical software packages with procedures enabling variance estimation for complex samples currently enable the use of these “replicate weights” (e.g., SAS, Stata, R). At present, the bootstrapping approach can be applied to complex samples in Stata (Kolenikov 2010), R (Lumley 2010), Mplus (Asparouhov and Muthén 2010), WesVar, SAS, and SUDAAN (Gagne et al. 2014).

So how does a survey statistician choose which variance estimation procedure to use when employing a particular software procedure for design-based descriptive analysis? Numerous studies have compared the performance of these alternative variance estimation methods under different complex sample designs. These include Kish and Frankel (1968, 1970, 1974), Frankel (1971), Bean (1975), Campbell and Meyer (1978), Lemeshow and Levy (1978), Shah et al. (1977), Rao and Wu (1985, 1987, 1988), Kovar et al. (1988), Judkins (1990), Shao and Sitter (1996), Korn and Graubard (1999), Canty and Davison (1999), Rao and Shao (1999), Shao (2003), and Heeringa et al. (2017). These studies have consistently demonstrated that for many common types of survey estimates (e.g., means, proportions, regression coefficients), all methods perform well and differences between the methods are *negligible*. Exceptions include small samples, where linearization can be unstable and perform worse than replication methods, and quantiles, where alternative forms of linearization are needed given that quantiles cannot generally be approximated using smooth functions of population totals or means (Woodruff 1952; Francisco and Fuller 1991; Sitter and Wu 2001). Many of the studies above demonstrate that BRR and the bootstrap perform well for medians and functions of quantiles. In addition, linearization methods covering all possible nonlinear estimators (e.g., correlation coefficients) and complex sample designs may not be readily programmed in all software packages.

2.3. Calculation of Degrees of Freedom for Confidence Intervals

Analysts often desire to form confidence intervals for population parameters when applying design-based methods to complex samples. These intervals, which under an assumption of large-sample normality of the sampling distribution for the weighted estimator rely on a critical t-value, also require specification of the appropriate degrees of freedom for the critical t-value. At present, most statistical software computes these degrees of freedom based on the aforementioned assumption of large-sample normality, setting the degrees of freedom equal to the number of clusters used for variance estimation minus the number of strata (Heeringa et al. 2017). While this approach makes intuitive sense, given that design-based variance estimates are driven by between-cluster variance within strata and the standard deviation of the sampling distribution is estimated rather than known, it is heavily dependent on the aforementioned assumption and can be severely limited in certain cases (Valliant and Rust 2010). Valliant and Rust (2010) propose an alternative estimator of the degrees of freedom for the critical t-value and show that it leads to improved coverage in some cases, but more work in this area, including sensitivity analyses, is certainly needed. Furthermore, the alternative estimator proposed by Valliant and Rust has yet to make its way into any statistical software.

Dean and Pagano (2015) provide a recent review of several different methods for computing confidence intervals for estimated proportions in the descriptive context, with and without adjustment for the degrees of freedom according to a complex sample design. Via simulation, these authors found support for use of the logit, Wilson, Jeffreys, and Agresti-Coull intervals (Agresti and Coull 1998) in complex samples, especially when proportions are very small or very large. Some of these methods (e.g., the logit approach) are readily implemented in existing software (e.g., the `svy: tab` command in Stata). While the other methods may not be as widely implemented, these authors provide clear guidance on their computation in practice.

2.4. Software

We now consider state-of-the-art statistical software that is currently available for implementing the design-based descriptive estimation and inference approaches outlined above when analyzing complex sample survey data. Table 2 provides a list of presently available software procedures and profiles their capabilities, in particular considering 1) percentile estimation, 2) variance estimation options, and 3) subpopulation analysis. All procedures in Table 2 enable appropriate weighted estimation of various descriptive parameters. A key take-away message from Table 2 is that weighted estimation of percentiles, combined with design-based variance estimation for the weighted estimates based on the aforementioned approaches, is not yet widely implemented across the different software packages. Aside from this, most software packages offer similar capabilities for design-based descriptive analyses of complex sample survey data. Examples of the use of syntax for many of these procedures can be found at http://www.isr.umich.edu/src/smp/asda.

Table 2. Capability profile of current statistical software enabling design-based descriptive estimation based on complex sample survey data (all procedures enable weighted estimation).

		Variance estimation options							Subpopulation analysis	
Software	Percentile estimation?	User-specified FPCs?	Handles replicate weights?	TSL?	JRR?	BRR?	Bootstrapping?	Appropriate variance estimation for post-stratification?	Subpopulation estimation?	Subpopulation comparisons?
SAS/STAT (V9.4)										
SURVEYMEANS	Y	Y	Y	Y	Y	Y	Y[1]	Y	Y	Y[2]
SURVEYFREQ	N	Y	Y	Y	Y	Y	Y[1]	Y	Y	Y
IBM SPSS Statistics: Complex Samples Module (V25)										
CSDESCRIPTIVES	N	Y	N	Y	N	N	N	N	Y	N
CSTABULATE	N	Y	N	Y	N	N	N	N	Y	Y
Stata (V15+)										
`svy: mean`	N	Y	Y	Y	Y	Y	Y	Y[3]	Y	Y[4]
`svy: prop`	N	Y	Y	Y	Y	Y	Y	Y[3]	Y	Y[4]
`svy: tab`	N	Y	Y	Y	Y	Y	Y	Y[3]	Y	Y
R: survey package[5]										
`svymean()`	N	Y	Y	Y	Y	Y	Y	Y[7]	N	N
`svyby()`	Y[6]	Y	Y	Y	Y	Y	Y	Y[7]	Y	N
`svytable()`	N	Y	Y	Y	Y	Y	Y	Y[7]	Y	Y
`svyquantile()`	Y	Y	Y	Y	Y	Y	Y	Y[7]	N	N
`svycontrast()`	Y	Y	Y	Y	Y	Y	Y	Y[7]	N	Y[8]
WesVar										
Mean	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Median	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Quantile	Y	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Ratio	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Totals	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y

Table 2. Continued.

		Variance estimation options							Subpopulation analysis	
Software	Percentile estimation?	User-specified FPCs?	Handles replicate weights?	TSL?	JRR?	BRR?	Bootstrapping?	Appropriate variance estimation for post-stratification?	Subpopulation estimation?	Subpopulation comparisons?
Variance	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
CV	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Skewness	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
Kurtosis	N	Y	Y	N	Y	Y	Y[9]	Y	Y	Y
SUDAAN										
DESCRIPT	Y	Y	Y	Y	Y	Y	Y[9]	Y	Y	Y
TABULATE	N	Y	Y	Y	Y	Y	Y[9]	Y	Y	Y
RATIOS	N	Y	Y	Y	Y	Y	Y[9]	Y	Y	Y
PROC CROSSTAB	N	Y	Y	Y	Y	Y	Y[9]	Y	Y	Y
IVEware[10]										
%DESCRIBE	N	N	N	Y	Y	N	N	N	Y	Y
VPLX[11]										
Summary statistics	N	N	Y	Y	Y	Y	N	Y	Y	Y
Epi Info CSAMPLE[12]										
Summary statistics	N	N	N	Y	N	N	N	N	Y	Y
AM Software[13]										
Summary statistics	Y	N	N	Y	Y	Y	Y	N	N	N

Table 2. Continued.

		Variance estimation options							Subpopulation analysis	
Software	Percentile estimation?	User-specified FPCs?	Handles replicate weights?	TSL?	JRR?	BRR?	Bootstrapping?	Appropriate variance estimation for post-stratification?	Subpopulation estimation?	Subpopulation comparisons?
Bascula 4[14]										
Summary statistics	N	N	Y	Y	N	Y	N	Y	N	N
CLUSTERS[15]										
Summary statistics	N	N	N	Y	N	N	N	N	Y	Y
Generalized estimation system[16]										
Summary statistics	N	N	N	Y	Y	N	N	Y	N	N
PCCARP[17]										
Summary statistics	Y	Y	N	Y	N	N	N	Y	Y	Y

[1] See https://support.sas.com/documentation/onlinedoc/stat/143/surveymeans.pdf for general discussion.

[2] Available in the newest version of SURVEYMEANS in SAS/STAT 14.2: see http://support.sas.com/kb/34/607.html.

[3] Via the `poststrata()` and `postweight()` options in the `svyset` command.

[4] Via the `lincom` post-estimation command.

[5] See http://r-survey.r-forge.r-project.org/survey/ for extensive annotated examples of the use of these procedures in R.

[6] Via inclusion of the `svyquantile()` function.

[7] When the `postStratify()` function has been used to update the `svydesign()` object.

[8] Given a `svyby()` object; covariances of subgroup estimates are not computed and accounted for in comparisons when using linearization.

[9] See https://www.researchgate.net/publication/255643504_Using_bootstrap_weights_with_Wes_Var_and_SUDAAN for discussion.

[10] http://www.iveware.org

[11] https://www.census.gov/sdms/www/vwelcome.html (still active)

[12] http://www.cdc.gov/epiinfo/index.html

[13] http://am.air.org/

[14] http://www.hcp.med.harvard.edu/statistics/survey-soft/bascula.html (A European product, no longer active)

[15] http://www.hcp.med.harvard.edu/statistics/survey-soft/clusters.html (A European product, still active)

[16] http://www.hcp.med.harvard.edu/statistics/survey-soft/genest.html (A product of Statistics Canada, still active)

[17] http://www.hcp.med.harvard.edu/statistics/survey-soft/pccarp.html (no longer active)

3. Descriptive Objectives: Model-Based Approaches

3.1. Overview

While descriptive inferences based on complex sample survey data sets generally tend to arise from design-based approaches (Little 2004), model-based approaches to descriptive inference have their relative merits as well. Design-based approaches can be heavily affected by *non-sampling errors*, such as unit nonresponse, given that they are governed by knowledge of sampling probabilities for all cases included in a sample. Unlike design-based approaches, which are based on the notion of random sampling from a finite population, strictly model-based approaches assume that some superpopulation model exists, from which the finite populations in the design-based setting are actually sampled. Interest lies in unbiased estimation of the parameters of that superpopulation model. Model-based approaches to making descriptive inference generally involve the specification of a probability model for a variable (or variables) of interest (where the variable for which descriptive inference is desired is a dependent variable), estimation of the descriptive parameters of interest (e.g., means) defined by the model, and estimation of the variance of that estimate with respect to the specified model (Binder and Roberts 2003).

One can also employ model-based prediction approaches when making descriptive inferences about finite populations. In this case, various auxiliary predictors available for the larger population (usually in aggregate form) may be included in the specification of the probability model for the variable of interest. In the case of complex sampling, these auxiliary predictors can and should generally include some function of the probability of selection, in addition to stratum identifiers (if these design features are relevant and informative about the variable of interest; Hansen et al. 1983; Little 2004). In these cases, predictions are computed on the variable of interest for nonsampled cases or nonrespondents, using the auxiliary information and parameter estimates in the specified model, and estimates are computed by combining the observed sample data on the dependent variable and the model-based predictions for nonsampled or nonresponding cases (Valliant et al. 2000). Variances of the resulting estimates are then computed with respect to the properties of the model used. Predictions for the nonsampled cases and measures of uncertainty for the descriptive parameter of interest may also be computed based on Bayesian methods (Little 2003), where informative design features should again be included in the specification of the model (likelihood) for the available data.

Särndal et al. (1992) describe an alternative approach that combines elements of design-based inference and model-based inference known as *model-assisted* inference, where design-based estimates of descriptive parameters (e.g., totals) are adjusted given known auxiliary variables for the entire population and their relationships with the variable of interest, and variances of the adjusted estimates are computed with respect to the randomization distribution (as in design-based inference). The generalized regression (GREG) estimator is a popular example of the model-assisted approach to making descriptive inferences from complex sample survey data. Valliant et al. (2000) provide a comprehensive theoretical overview of related model-based prediction approaches to the descriptive analysis of survey data.

Best practices in this area generally focus on how probabilities of selection/weights should be accounted for in the models, how to make the most efficient use of auxiliary information available for a finite population, and how to handle nonresponse. For example, Elliott and Little (2000) discuss efficient model-based Bayesian approaches to accommodating sampling weights in descriptive inference when large or highly-variable survey weights may cause design-based approaches to become very inefficient. Little (2004) presented a model-based approach to descriptive inference combining precision weighting and probability weighting in a Bayesian framework. This was further expanded on in Little (2012), who advocated "Calibrated Bayes" (CB) as a framework for survey inference. The basic idea behind CB is to use a Bayesian model-based approach to produce inferences that have good design-based properties. The CB approach is intended to combine the strengths of both design-based and model-based perspectives by explicitly accounting for survey design information in the model and using only weak prior distributions that allow the observed data to dominate the inference. Inferences are calibrated in the sense that they produce posterior credibility intervals that correspond to their nominal design-based coverage in repeated sampling (Little 2006, 2011, 2012, 2015). The incorporation of all key survey design features in the model is paramount to this approach to minimize the effect of model misspecification.

Peress (2010) discussed the use of selection models in a model-based approach to account for nonignorable nonresponse as a part of the modeling process in estimating a proportion. Using a related approach, Barnighausen et al. (2011) applied a Heckman-type bivariate probit selection model in estimating HIV prevalence estimates that adjusted for nonignorable nonresponse based on a set of selection variables correlated with survey participation. More recently, West and McCabe (2017) demonstrated how this approach can be implemented using the Stata software to make descriptive inferences in a longitudinal context, where nonignorable attrition may be occurring in the future waves of a panel survey.

3.2. Software

Regarding available software for implementing these model-based approaches to descriptive inference in surveys, there are not nearly as many "canned" software procedures implementing these approaches as there are for design-based approaches, meaning that statisticians would generally need to write code implementing these approaches for nonstatistical clients. For example, Zheng and Little (2003), Little and Zheng (2007), and Zangeneh and Little (2015) demonstrate the improvements in estimates of population totals when using a penalized spline regression model over a design-based Horvitz-Thompson approach when the sizes of nonsampled units are either known or unknown, and error variances in the model of interest may be heteroscedastic. Zangeneh and Little (2015) have developed R code implementing their proposed approach (available from the authors upon request).

Chapter 8 of Lunn et al. (2012) discusses how the BUGS software (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml) can be used to generate predictions for nonsampled cases using a Bayesian approach, where again any complex sampling features would need to be accounted for in the model specification (Little 2004). More recently, the Stan software (http://mc-stan.org/) has become a popular alternative for similar types of Bayesian approaches, and this software can be readily used in R and Stata (among other platforms). Interested readers can see http://rpubs.com/corey_sparks/157901 for an example of a model-based descriptive analysis of survey data using a Bayesian approach in R (calling the Stan software). Valliant et al. (2000) provided a comprehensive library of S-plus code for implementing various model-based and model-assisted approaches, and these functions can generally be adapted in the R software with ease (examples are available upon request from the first author; see also Valliant et al. (2013) for additional examples). In general, we recommend that statisticians compare standard errors for descriptive estimates computed using design-based and model-based approaches, and determine whether efficiency gains are possible when employing the model-based approaches discussed in this section.

In more recent work, Si et al. (2015) presented model-based Bayesian methodology for making robust finite population inferences about means or proportions of interest when only final survey weights (and no stratum or cluster codes) are available for survey respondents. This model-based approach simultaneously predicts the distribution of the final survey weights among nonsampled cases in the population of interest and the values of the survey variable of interest for these cases (as a function of the weights), enabling simulations of the full population means or proportions based on posterior distributions for these descriptive parameters (given the sampled cases and their data). These authors demonstrated the advantages of this approach for the efficiency of descriptive finite population estimates (for both full populations and subpopulations), and implemented this approach in the Stan software (see http://www.isr.umich.edu/src/smp/asda for example Stan code).

4. Analytic Objectives: Design-Based Approaches

4.1. Overview

Design-based approaches that utilize (adjusted) design weights to fit regression models to complex sample survey data are in common use (e.g., DuMouchel and Duncan 1983; Pfefferman 1993; Pfeffermann and Sverchkov 2009; Pfefferman 2011; Lumley and Scott 2017). In the simple case of estimating the parameters of a specified linear regression model, the standard ordinary least squares (OLS) approach can be modified by incorporating the final respondent weight into the objective function that minimizes the finite population residual sum of squares. This weighted least squares (WLS) approach provides a closed-form, model-unbiased estimator for the regression parameters that also serves as a pseudo-maximum likelihood estimator for the regression parameters in the finite population (Binder 1981, 1983; Pfeffermann 2011, Section 3.4). Lohr (2014) describes how to estimate design effects in this context reflecting complex sampling features.

For generalized linear models featuring nonlinear relationships between the predictors and the expectation of the dependent variable of interest (e.g., logistic regression models), closed-form solutions do not exist for estimation of model parameters. Furthermore, "standard" maximum likelihood estimation is not possible with complex sample designs because the assumption of independent observations is violated by the stratification and cluster sampling inherent to complex samples (Archer et al. 2007). Binder (1981, 1983) proposed a pseudo-maximum likelihood estimation (PMLE) framework for fitting generalized linear models to complex sample survey data. The basic idea of the PMLE method is to estimate model parameters by replacing finite population likelihood estimating equations with design-unbiased, weighted estimating equations for the responding units. Positive evaluations of the PMLE method and its properties can be found in several studies (Binder 1983; Chambless and Boyle 1985; Roberts et al. 1987; Morel 1989; Skinner et al. 1989; Nordberg 1989; Pfeffermann 1993; Godambe and Thompson 2009). PMLE is now the standard method implemented in many software procedures for fitting generalized linear models to complex sample survey data.

Binder (1981, 1983) also proposed a general method for linearized variance estimation for pseudo-maximum likelihood estimates of regression coefficients that is implemented as the default method in many statistical software packages, and replication methods generally work equally well in many regression settings, as noted earlier. Hypothesis tests for regression parameters based on complex sample survey data are carried out by using the design-based estimates of the variances and covariances and applying commonly used test statistics, such as Student's t and the Wald chi-square or Wald F-test. Rao and Scott (1981, 1984, 1987) proposed a modified Wald chi-square statistic for survey data that accounts for complex sample design features. This procedure is implemented in many statistical software packages that support the analysis of complex sample survey data (e.g., the `svy: tab` command and the `test` post-estimation command in Stata).

4.2. *Should Survey Weights Even Be Used to Fit Models?*

The use of (adjusted) design weights to fit regression models has some limitations which have provoked controversy among statisticians (see Pfeffermann 1993; Gelman 2007; Pfefferman 2011; or Bollen et al. 2016 for reviews of the general issues). Design-based estimation strategies utilizing probability-weighted estimators generally yield larger variances than model-based estimation strategies (Korn and Graubard 1999). This loss in efficiency is more notable for small sample sizes and cases where there is large variation in the survey weights. For this reason, one best practice in this area is to examine the sensitivity of the regression results by comparing weighted and unweighted analyses, which is quite easy to do using current software. If these analyses yield notable differences, then this may indicate model misspecification and the weighted estimates should be reported to ensure that they are unbiased with respect to the sample design used. A review of formal tests for differences between weighted and unweighted regression analyses can be found in Bollen et al. (2016). It is also worth noting that the use of probability weighting in the analysis of complex sample survey data is not customary in some disciplines (e.g., economics) which favor the flexibility of explicitly featuring the relevant design variables as part of the model-building process.

4.3. *Software for Model Fitting*

Table 3 presents a summary of available software procedures for fitting regression models to complex sample survey data using design-based approaches. We emphasize software

Table 3. Current software procedures enabling design-based estimation of various regression models.

	Regression modeling options										Goodness of fit tests	
Software	Linear	Binary logistic	Ordinal logistic	Multinomial logistic	Poisson regression	Negative binomial regression	Probit	Cloglog	Survival (Cox) models	Quantile regression	Archer and Lemeshow test	Design-based model diagnostics?
Stata (V15+)												
`svy: regress`[2]	Y	N	N	N	N	N	N	N	N	N[1]	N	N
`svy: logit`	N	Y	N	N	N	N	N	N	N	N	Y[4]	N
`svy: mlogit`	N	N	N	Y	N	N	N	N	N	N	N	N
`svy: ologit`	N	N	Y	N	N	N	N	N	N	N	N	N
`svy: poisson`	N	N	N	N	Y[3]	N	N	N	N	N	N	N
`svy: nbreg`	N	N	N	N	N	Y[3]	N	N	N	N	N	N
`svy: stcox`	N	N	N	N	N	N	N	N	Y	N	N	N
`svy: probit`	N	N	N	N	N	N	Y	N	N	N	N	N
`svy: cloglog`	N	N	N	N	N	N	N	Y	N	N	N	N
SAS (V9.4)												
SURVEYREG	Y	N	N	N	N	N	N	N	N	N	N	N
SURVEYLOGISTIC	N	Y	Y	Y	N	N	Y	Y	N	N	N	N
SURVEYPHREG	N	N	N	N	N	N	N	N	Y	N	N	N
IBM SPSS Statistics: complex camples module (V25)												
CSLOGISTIC	N	Y	Y	Y	N	N	Y	Y	N	N	N	N
CSORDINAL	N	N	Y	N	N	N	Y	Y	N	N	N	Y
CSGLM	Y	N	N	N	N	N	N	N	N	N	N	N
CSCOXREG	N	N	N	N	N	N	N	N	Y	N	N	N
R survey package												
`svyglm()`	Y	Y	Y	Y	Y	Y	N	N	N	N	N	Y[5]
`rq()`	N	N	N	N	N	N	N	N	N	Y[6]	N	N
`svycoxph()`	N	N	N	N	N	N	N	N	Y	N	N	N

Table 3. Continued.

	Regression modeling options										Goodness of fit tests	
Software	Linear	Binary logistic	Ordinal logistic	Multinomial logistic	Poisson regression	Negative binomial regression	Probit	Cloglog	Survival (Cox) models	Quantile regression	Archer and Lemeshow test	Design-based model diagnostics?
Wesvar												
Linear regression	Y	N	N	N	N	N	N	N	N	N	N	N
Logistic regression	N	Y	N	N	N	N	N	N	N	N	N	N
Multinomial regression	N	N	N	Y	N	N	N	N	N	N	N	N
SUDAAN												
REGRESS	Y	N	N	N	N	N	N	N	N	N	N	N
LOGISTIC	N	Y	Y	Y	N	N	N	N	N	N	N	N
MULTILOG	N	N	Y	Y	N	N	N	N	N	N	N	N
LOGLINK	N	N	N	N	Y	N	N	N	N	N	N	N
SURVIVAL	N	N	N	N	N	N	N	N	Y	N	N	N
IVEware												
%REGRESS	Y	Y	Y	Y	Y	N	N	N	Y	N	N	N
Epi info CSAMPLE												
Regress	Y	N	N	N	N	N	N	N	N	N	N	N
Logistic	N	Y	N	N	N	N	N	N	N	N	N	N
AM software												
Regression	Y	Y	Y	Y	N	N	Y	N	N	N	N	N

[1] A full list of possible models that can be fitted in the design-based framework using Stata 15+ (including specialized modeling options, such as structural equation models, instrumental variables regression, finite mixture models, among others) can be found at https://www.stata.com/manuals/svy.pdf.

[2] Quantile regression models can be fitted using design-based approaches in Stata, given replicate weights and the user-written `bs4rw` command.

[3] Procedures for fitting zero-inflated versions of these models, including `svy: zip` and `svy: zinb`, are also available.

[4] Implemented in the post-estimation command `estat gof`, which can be executed after running `svy: logit`.

[5] Currently available in a working R package `svydiags`; see Heeringa et al. (2017).

[6] Possible when using the `rq()` function from the `quantreg` package in combination with replicate weights in the `survey` package; see http://www.isr.umich.edu/src/smp/asda/Additional%20R%20Examples%20bootstrapping%20with%20quantile%20regression.pdf for details.

procedures in general-purpose statistical software packages, but other stand-alone software tools primarily focused on modeling, such as Mplus (see http://www.statmodel.com/resrchpap.shtml for examples) and Latent GOLD (see the *Advanced/Syntax add-on* at https://www.statisticalinnovations.com/latent-gold-5-1/), can also easily fit common regression models using design-based approaches.

Readily apparent from Table 3 are the following take-away points: 1) different packages currently vary in terms of the different types of regression models that can be fitted using design-based methods; 2) design-based post-estimation goodness-of-fit tests are currently only implemented for logistic regression modeling in Stata; 3) design-based quantile regression is only implemented in the R `survey` package and Stata add-on commands at present; and 4) model diagnostics using design-based approaches are currently only available for linear regression models in a working package in R (see http://www.isr.umich.edu/src/smp/asda/svydiags-manual.pdf for details). Examples of the use of syntax for many of these procedures can be found at http://www.isr.umich.edu/src/smp/asda.

4.4. Software for Model Evaluation and Selection

Numerous design-adjusted model evaluation tools have been developed to evaluate the fits of regression models based on complex sample survey data. However, the implementation of some of these tools in popular statistical software packages is not yet widespread. A modified version of the R^2 statistic is often available for linear regression models, which estimates the "weighted" proportion of explained variance in the dependent variable after controlling for the independent variables. Residual diagnostics for complex samples more generally is an active area of research. Li and Valliant (2015) document the latest advances in this area and review their implementation in R; a working package for R entitled `svydiags` is available from these authors upon request, and examples of the use of this package are provided in Heeringa et al. (2017). Liao and Valliant (2012a, 2012b) developed collinearity diagnostics for identifying excessively high correlations between independent variables that explicitly account for complex sampling features; however, these diagnostics have not yet made their way into popular statistical software packages. Li and Valliant (2009, 2011a, 2011b) and Ryan et al. (2015) have proposed methods for identifying influential data points in linear and logistic regression analyses based on complex sample survey data, and these also need software development.

Model selection methods for complex samples have also seen recent development. For instance, Lumley and Scott (2015) developed survey analogues of the popular AIC and BIC information criteria for regression models fitted using pseudo-maximum likelihood estimation methods. These methods have been implemented in the R `survey` package. Archer et al. (2007) demonstrate that standard goodness-of-fit tests are not suitable for complex sample survey data and propose alternative tests that account for complex design features, including an F-test which is a survey analogue to the Hosmer-Lemeshow chi-square test for logistic regression. Heeringa et al. (2017) provide Wald tests for comparing nested regression models, following from Hosmer et al. (2013), who note that the standard likelihood ratio chi-square test is inappropriate for complex sample survey data due to the violation of key assumptions about the likelihood function that underlie the test. This issue is addressed further by Lumley and Scott (2013, 2014), who developed partial likelihood ratio tests for Cox regression models in the survival analysis context and adapted the Rao and Scott (1984) chi-square test to the case of design-based likelihood ratio tests in arbitrary regression models fitted to survey data. These approaches are also currently implemented in the `survey` package in R.

4.5. Software for Structural Equation Modeling and Classification Trees

There has also been some work on accounting for complex sample design features in structural equation and latent variable models (e.g., Muthén and Satorra 1995; Kaplan and Ferguson 1999; Stapleton 2002, 2006). These design-based approaches to fitting structural equation models are currently implemented in the Mplus software, the `lavaan.survey` package of R (Oberski 2014), the LISREL software, the `svy: sem` and `svy: gsem` commands of Stata, the Latent GOLD software, and PROC LCA (a user-written add-on for SAS). Some analysts of survey data may also be interested in building classification or regression trees for generating finite population predictions, and the recently-developed `rpms` package in R (Toth 2017) enables analysts to apply regression trees to complex sample survey data. A more general summary of additional specialized procedures for fitting models to survey data in R using design-based approaches can be found at https://cran.r-project.org/web/views/OfficialStatistics.html.

5. Analytic Objectives: Model-Based Approaches

5.1. Overview

Model-based approaches for analyzing survey data given analytic objectives vary. These approaches are typically implemented under a population- or sample-based modeling perspective. Under the population modeling perspective, all population units (including nonsampled units) are included in the analysis model, whereas under the sample-based perspective, only the sampled (or responding) units are analyzed. Under the population modeling perspective, one possible approach is to include all design variables and relevant interaction terms as covariates in the analysis model and effectively integrate these variables out (Pfeffermann 2011), leaving only the covariates of substantive interest. Implementing such an approach can be difficult for secondary analysts, because design variables for the entire population are typically not made available to secondary data users.

Model-based approaches for imputing both the design and substantive variables for the non-observed portion of the population have been proposed (Feder 2011; Si et al. 2015), though issues arise when the sample selection is dependent on the substantive variables of interest – a situation realized in a non-ignorable sampling setting (Pfeffermann and Sikov 2011). A further complication, noted by Pfefferman (2011), is that modeling the relationship between the design and substantive variables can be quite cumbersome, and integrating the design variables out of the model may result in an analysis model that does not reproduce the target model of substantive interest. Pfeffermann (2011) addresses this issue by demonstrating that the analysis model can be estimated without integrating the design variables out of the model. When not all design variables are available to the analyst for the entire population, then the sample weight is sometimes used as a proxy for the design variables (DuMouchel and Duncan 1983; Rubin 1985; Chambers et al. 1998; Wu and Fuller 2006). However, this still requires that the sample inclusion probabilities be made available to the secondary data analyst for the entire population, which may not be possible due to confidentiality or other data restrictions.

In contrast, approaches based on the sample modeling perspective need only make use of the design variables known for the sampled (or responding) units. Model-based methods employing maximum likelihood techniques estimate the unknown population parameters using the likelihood of the joint distribution of the design variables and sample covariates (Gelman et al. 2003; Little 2004). Alternative full-likelihood methods, which utilize the Missing Information Principle (Orchard and Woodbury 1972), have been explored in different contexts (Breckling et al. 1994; Chambers et al. 1998; Chambers and Skinner 2003, Chapter 2). Empirical likelihood methods have also been considered for complex samples (Hartley and Rao 1968; Owen 2001). These methods, while generally more computationally intensive than design-based approaches, can produce much more efficient estimates of regression parameters with improved coverage properties (see Pfeffermann et al. 2006 for an illustration).

5.2. *Software*

Given the considerations outlined in Subsection 5.1, model-based approaches to analytic objectives can make use of existing software procedures for fitting regression models. There is no need to use specialized software for design-based survey analysis to fit these models. The important aspect of implementing these procedures is making sure that the design features have been carefully accounted for in the design matrices of the specified models.

6. Analytic Objectives: “Hybrid” Approaches

6.1. *Overview*

So-called “hybrid” approaches to regression modeling of complex sample survey data employ multilevel models, and are distinguished by the explicit desire of the researcher to make finite population inferences about the components of variance in dependent variables of interest attributable to the different stages of a multi-stage sample design. The theory and methods for incorporating survey weights into pseudo-maximum likelihood estimation of the fixed effect and covariance parameters defining a multilevel model were initially described by Pfeffermann et al. (1998). These methods were later expanded on and evaluated via simulation by Kovacevic and Rai (2003), Grilli and Pratesi (2004), Asparouhov (2006), Rabe-Hesketh and Skrondal (2006), Carle (2009), and Pfeffermann (2011). Skinner and Holmes (2003) and Heeringa et al. (2017) have elaborated on the appropriate use of survey weights when fitting multilevel models to *longitudinal* survey data.

These methods for computing weighted estimates of the parameters in multilevel regression models all require the following: 1) *conditional* weights at lower levels of the data hierarchy (e.g., students within schools), which indicate inverses of the probability of selection *conditional* on a given higher-level unit (e.g., school) being sampled, and 2) unit-level weights at the highest level of the data hierarchy (e.g., counties), representing inverses of the probabilities of selection for the highest-level sampling unit. Pfeffermann et al. (1998) and Rabe- Hesketh and Skrondal (2006) clearly describe how the likelihood functions used to estimate these models are partitioned in a way that requires this combination of conditional and unconditional weights for unbiased estimation of the model parameters. More recently, Stapleton and Kang (2016) have described how to estimate design effects in this context, representing the effects of complex sampling features on the variance of estimated parameters in multilevel models.

This requirement that the conditional lower-level weights and unconditional higher-level weights be available for estimation has limited these approaches from gaining traction outside of the survey statistics literature (see West et al. 2015 for a recent case study), given the need for public-use data files to include these "specialized" weights for users, which could introduce disclosure risk concerns. The final respondent weights provided in a public-use survey data set typically represent inverses of the products of the probabilities of selection at *all* stages of a complex sample design; computation of the conditional weights at lower levels requires dividing the final weights by the higher-level sampling weights to determine the inverse of the conditional sampling probability required for estimation. The computation of these weights therefore represents an additional burden that survey organizations would need to take on for users interested in these "hybrid" approaches. Chantala et al. (2011) provide important practical guidance and software tools to assist with this process.

The conditional weights that are specific to each lower-level unit also need to be *scaled* or *normalized* across all higher-level units, to reduce the varying magnitudes of these weights across the higher-level units. This weight scaling is important because it minimizes the bias in parameter estimates based on the models (Pfeffermann et al. 1998). Pfeffermann et al. (1998), Rabe-Hesketh and Skrondal (2006), and Carle (2009) describe alternative methods for performing weight scaling (e.g., normalizing the lower-level weights by dividing all of the weights in a higher-level unit by their average, so that they sum to the sample size within that unit). The literature to date has not demonstrated that one weight scaling method is superior over another; there has, however, been consistent agreement that weight scaling needs to be done to minimize bias, especially in the case of generalized linear regression models (e.g., multilevel logistic regression models; Rabe-Hesketh and Skrondal 2006). Weight scaling represents an additional data processing step that may not be "automatic" in the software that is presently available for these "hybrid" approaches (e.g., the `mixed` command in Stata); see Rabe-Hesketh and Skrondal (2006) for worked examples.

6.2. Software

At present, these approaches for weighted estimation of multilevel models are not widely implemented across statistical software packages. This kind of implementation will be especially important for these "hybrid" model-based approaches to gain traction among nonstatisticians. Software packages and specific procedures capable of implementing these "hybrid" approaches for both linear and generalized linear regression models include Stata (Version 15.1+), SAS (PROC GLIMMIX, SAS/STAT Version 13.1+; Zhu 2014), HLM (Version 7.01+), MLwiN (Version 2.35+; see http://www.bristol.ac.uk/cmm/software/

mlwin), Mplus (Version 7.4+; see http://www.statmodel.com), and the gllamm command for Stata (www.gllamm.org). The online documentation for each of these packages provides worked examples of implementing these "hybrid" approaches (e.g., type "help mixed#sampling" in the Stata Viewer). Importantly, all of these tools are capable of implementing either model-based approaches or "hybrid" approaches, depending on how the survey weights are used.

7. Directions for Future Software Development

First, considering design-based approaches, additional software options enabling variance estimation for quantile estimates are still needed, where BRR and bootstrap methods have been shown to produce the best confidence interval coverage (Kovar et al. 1988). Techniques for accounting for complex sample design features when evaluating the goodness-of-fit of various regression models in the design-based framework (e.g., Archer et al. 2007) also need theoretical and computational development. Furthermore, methods for assessing regression diagnostics need further theoretical development (especially for generalized linear models), and state-of-the-art diagnostic methods for linear regression models need to be more widely incorporated in survey analysis software. Finally, more research needs to consider whether there are better approaches to estimating the design-based degrees of freedom associated with a given variance estimate when forming confidence intervals, and implementation of alternative approaches (e.g., Valliant and Rust 2010) in existing software is still needed.

Second, considering model-based and "hybrid" approaches, the literature currently lacks a coherent theoretical framework enabling hypothesis testing for the variance components in a multilevel model estimated using pseudo-maximum likelihood estimation (see Zhang and Lin 2008 for a review of these methods). The recent work by Lumley and Scott (2015) needs to be adapted to these types of tests based on multilevel models estimated using sampling weights. Also important in this area will be the development of diagnostics for fitted multilevel models (Claeskens 2013) that recognize complex sampling features. Finally, there is still work to be done in assessing optimal approaches for fitting multilevel models to longitudinal survey data (Thompson 2015); for example, should time-varying weights be computed to adjust for differential attrition at different waves? Or should only cases with complete data be analyzed when fitting the multilevel models (Heeringa et al. 2017, Chapter 11)? Empirical and theoretical developments in this area will be important moving forward.

Finally, this review has not touched on statistical analysis approaches involving item-missing survey data, and how complex sampling features should be accounted for in this context. Briefly, initial work in this area suggested that models for imputing item-missing values should include the complex sample design features as covariates, similar to some of the model-based approaches discussed above (Reiter et al. 2006). More recently, methods have been developed for simulating synthetic populations, given the complex sampling features available for a sample, and then imputing missing values using straightforward methods in these simulated populations prior to making population inferences (Zhou et al. 2016b; Zhou et al. 2016c; see also Zhou et al. 2016a, for example R code). Alongside these developments using model-based imputation methods, Kim and Fuller (2004) and Kim and Shao (2014) have developed fractional hot-deck imputation techniques for complex sample survey data sets that may offer efficiency advantages over other competing imputation approaches. These approaches have been implemented in the SURVEYIMPUTE procedure of the SAS software (Version 9.4). Future research should consider the competing benefits and costs of the simulation-based imputation approaches and the fractional imputation approaches in terms of computational costs and the efficiency of the finite population estimates produced.

8. References

Agresti, A. and B. A. Coull. 1998. “Approximate Is Better than ‘Exact’ for Interval Estimation of Binomial Proportions.” *American Statistician* 52: 119–126. Doi: https://doi.org/10.1080/00031305.1998.10480550.

Archer, K.J., S. Lemeshow, and D.W. Hosmer. 2007. “Goodness-of-fit Tests for Logistic Regression Models When Data are Collected using a Complex Sampling Design.” *Computational Statistics and Data Analysis* 51: 4450–4464. Doi: https://doi.org/10.1016/j.csda.2006.07.006.

Asparouhov, T. 2006. “General Multi-level Modeling with Sampling Weights.” *Communications in Statistics—Theory and Methods* 35: 439–460. Doi: https://doi.org/10.1080/03610920500476598.

Asparouhov, T. and B. Muthén. 2007. “Testing for Informative Weights and Weights Trimming in Multivariate Modelling with Survey Data.” In Proceedings of the Survey Research Methods Section of the American Statistical Association, 2007, Salt Lake City, Utah, 3394–3399. Available at: https://www.statmodel.com/download/JSM2007000745.pdf (accessed April 14, 2017).

Asparouhov, T. and B. Muthén. 2010. “Resampling Methods in Mplus for Complex Survey Data.” *Mplus Technical Report, May 4, 2010*. Available at: https://www.statmodel.com/download/Resampling_Methods5.pdf (Accessed October 10, 2016).

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, and R. Tourangeau. 2013. “Summary Report of the AAPOR Task Force on Non-probability Sampling.” *Journal of Survey Statistics and Methodology* 1: 90–143. Doi: https://doi.org/10.1093/jssam/smt008.

Barnighausen, T., J. Bor, S. Wandira-Kazibwe, and D. Canning. 2011. “Correcting HIV Prevalence Estimates for Survey Nonparticipation using Heckman-type Selection Models.” *Epidemiology* 22: 27–35. Doi: 10.1097/EDE.0b013e3181ffa201.

Bean, J.A. 1975. “Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples: An Empirical Distribution.” In *Vital and Health Statistics: Series 2, Data Evaluation and Methods Research* 65: i–iv.

Beaumont, J.F. 2005. “On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment.” *Survey Methodology* 31: 227–231.

Beaumont, J.F. 2008. “A New Approach to Weighting and Inference in Sample Surveys.” *Biometrika* 95: 539–553. Doi: https://doi.org/10.1093/biomet/asn028.

Bethlehem, J.G. 2002. “Weighting Nonresponse Adjustments Based on Auxiliary Information.” In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 275–288. New York: Wiley.

Binder, D.A. 1981. “On the Variances of Asymptotically Normal Estimators for Complex Surveys.” *Survey Methodology* 7: 157–170.

Binder, D.A. 1983. “On the Variances of Asymptotically Normal Estimators from Complex Surveys.” *International Statistical Review* 51: 279–292. Doi: 10.2307/1402588.

Binder, D.A. and G.R. Roberts. 2003. “Design-based and Model-based Methods for Estimating Model Parameters.” In *Analysis of Survey Data*, edited by R.L. Chambers and C.J. Skinner, 29–48. Chichester, West Sussex: Wiley.

Bollen, K.A., P.P. Biemer, A.F. Karr, S. Tueller, and M.E. Berzofsky. 2016. “Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis.” *Annual Review of Statistics and Its Application* 3: 375–392. Doi: https://doi.org/10.1146/annurev-statistics-011516-012958.

Breckling, J.U., R.L. Chambers, A.H. Dorfman, S.M. Tam, and A.H. Welsh. 1994. “Maximum Likelihood Inference from Sample Survey Data.” *International Statistical Review* 62: 349–363. Doi: 10.2307/1403766.

Brick, J.M. 2013. “Unit Nonresponse and Weighting Adjustments: A Critical Review.” *Journal of Official Statistics* 29: 329–353. Doi: https://doi.org/10.2478/jos-2013-0026.

Campbell, C. and M. Meyer. 1978. “Some Properties of T Confidence Intervals for Survey Data.” In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 437–442. Available at: https://ww2.amstat.org/sections/srms/Proceedings/papers/1978_089.pdf (accessed April 14, 2017).

Canty, A.J. and A.C. Davison. 1999. “Resampling-based Variance Estimation for Labour Force Surveys.” *The Statistician* 48: 379–391. Doi: 10.1111/1467-9884.00196.

Carle, A.C. 2009. “Fitting Multilevel Models in Complex Survey Data with Design Weights: Recommendations.” *BMC Medical Research Methodology* 9(49). Doi: https://doi.org/10.1186/1471-2288-9-49.

Cassel, C., C.-E. Särndal, and J. Wretman. 1983. “Some Uses of Statistical Models in Connection with the Nonresponse Problem.” In *Incomplete Data in Sample Surveys*, edited by W.G. Madow and I. Olkin, 143–160. New York: Academic Press.

Chambers, R.L., A.H. Dorfman, and S. Wang. 1998. “Limited Information Likelihood Analysis of Survey Data.” *Journal of the Royal Statistical Society (Series B)* 60: 397–411. Doi: 10.1111/1467-9868.00132.

Chambers, R.L. and C.J. Skinner (Editors). 2003. *Analysis of Survey Data*. New York: John Wiley and Sons.

Chambless, L.E. and K.E. Boyle. 1985. “Maximum Likelihood Methods for Complex Sample Data: Logistic Regression and Discrete Proportional Hazards Models.” *Communications in Statistics-Theory and Methods* 14: 1377–1392. Doi: https://doi.org/10.1080/03610928508828982.

Chantala, K., D. Blanchette, and C.M. Suchindran. 2011. “Software to Compute Sampling Weights for Multilevel Analysis.” Technical Report, Carolina Population Center, UNC at Chapter Hill. Available at http://www.cpc.unc.edu/research/tools/data_analysis/ml_sampling_weights (accessed January 30, 2018).

Claeskens, G. 2013. “Lack of Fit, Graphics, and Multilevel Model Diagnostics.” In *The SAGE Handbook of Multilevel Modeling*, edited by M.A. Scott, J.S. Simonoff, and B.D. Marx, 425–444. Los Angeles: SAGE Publications.

Dean, N. and M. Pagano. 2015. “Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys.” *Journal of Survey Statistics and Methodology* 3: 484–503. Doi: https://doi.org/10.1093/jssam/smv024.

Deville, J.C. and C.-E. Särndal. 1992. “Calibration Estimators in Survey Sampling.” *Journal of the American Statistical Association* 87: 376–382. Doi: https://doi.org/10.1080/01621459.1992.10475217.

Deville, J.C., C.-E. Särndal, and O. Sautory. 1993. “Generalized Raking Procedures in Survey Sampling.” *Journal of the American Statistical Association* 88: 1013–1020. Doi: https://doi.org/10.1080/01621459.1993.10476369.

DuMouchel, W.H. and G.J. Duncan. 1983. “Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples.” *Journal of the American Statistical Association* 78: 535–543. Doi: https://doi.org/10.1080/01621459.1983.10478006.

Ekholm, A. and S. Laaksonen. 1991. “Weighting Via Response Modeling in the Finnish Household Budget Survey.” *Journal of Official Statistics* 7: 325–337.

Elliott, M.R. and R.J. Little. 2000. “Model-based Alternatives to Trimming Survey Weights.” *Journal of Official Statistics* 16: 191–210.

Elliott, M.R. and R. Valliant. 2017. “Inference for Nonprobability Samples.” *Statistical Science* 32: 249–264. Doi: 10.1214/16-STS598.

Eltinge, J.L. and I.S. Yansaneh. 1997. “Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey.” *Survey Methodology* 23: 33–40.

Feder, M. 2011. “Fitting Regression Models to Complex Survey Data – Gelman’s Estimator Revisited.” In *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*, Dublin, Ireland, August 2011. Available at: http://2011.isiproceedings.org/papers/950551.pdf (accessed January 30, 2018).

Flores-Cervantes, I. and J.M. Brick. 2016. “Nonresponse Adjustments with Misspecified Models in Stratified Designs.” *Survey Methodology* 42: 161–177.

Francisco, C.A. and W.A. Fuller. 1991. “Quantile Estimation with a Complex Survey Design.” *The Annals of Statistics* 19: 454–469. Doi: http://www.jstor.org/stable/2241867.

Frankel, M.R. 1971. *Inference from Survey Samples: an Empirical Investigation*. Institute for Social Research, University of Michigan, Ann Arbor, MI, USA.

Gagne, C., G. Roberts, and L.-A. Keown. 2014. “Weighted Estimation and Bootstrap Variance Estimation for Analyzing Survey Data: How to Implement in Selected Software.” *Statistics Canada: The Research Data Centres Information and Technical Bulletin, August 7, 2014*. Available at: http://www.statcan.gc.ca/pub/12-002-x/2014001/article/11901-eng.htm (accessed January 30, 2018).

Gelman, A. 2007. “Struggles with Survey Weighting and Regression Modeling.” *Statistical Science* 22: 153–164. Doi: 10.1214/088342306000000691.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2003. *Bayesian Data Analysis* (*2nd Edition*). Boca Raton, FL: Chapman & Hall/CRC.

Godambe, V.P. and M.E. Thompson. 2009. "Estimating Functions and Survey Sampling." *Handbook of Statistics Vol 29B (Sample Surveys: Inference and Analysis)*: 83–101. Doi: https://doi.org/10.1016/S0169-7161(09)00226-0.

Grau, E., F. Potter, S. Williams, and N. Diaz-Tena. 2006. "Nonresponse Adjustment Using Logistic Regression: To Weight or Not to Weight?" In *Proceedings of the Survey Research Methods Section of the American Statistical Association, Alexandria, VA, 2006*, 3073–3080. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.586.3263&rep=rep1&type=pdf (accessed January 30, 2018).

Grilli, L. and M. Pratesi. 2004. "Weighted Estimation in Multilevel Ordinal and Binary Models in the Presence of Informative Sampling Designs." *Survey Methodology* 30: 93–104.

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70: 646–675. Doi: https://doi.org/10.1093/poq/nfl033.

Hansen, M.H., W.G. Madow, and B.J. Tepping. 1983. "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys." *Journal of the American Statistical Association* 78: 776–793. Doi: https://doi.org/10.1080/01621459.1983.10477018.

Hartley, H.O. and J.N.K. Rao. 1968. "A New Estimation Theory for Sample Surveys." *Biometrika* 55: 547–557. Doi: https://doi.org/10.1093/biomet/55.3.547.

Haziza, D. and J.F. Beaumont. 2007. "On the Construction of Imputation Classes in Surveys." *International Statistical Review* 75: 25–43. Doi: 10.1111/j.1751-5823.2006.00002.x.

Haziza, D. and J.F. Beaumont. 2017. "Construction of Weights in Surveys: A Review." *Statistical Science* 32: 206–226. Doi: 10.1214/16-STS608.

Heeringa, S.G., B.T. West, and P.A. Berglund. 2017. *Applied Survey Data Analysis, Second Edition*. Boca Raton, FL: Chapman & Hall/CRC Press.

Holt, D. and T.M.F. Smith. 1979. "Post Stratification." *Journal of the Royal Statistical Society, Series A (General)* 142: 33–46. Doi: http://www.jstor.org/stable/2344652.

Hosmer, D.W., S. Lemeshow, and X. Sturdivant. 2013. *Applied Logistic Regression, Third Edition*. New York, NY: Wiley.

Judkins, D.R. 1990. "Fay's Method for Variance Estimation." *Journal of Official Statistics* 6: 223–239.

Kalton, G. and I. Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19: 81–97.

Kaplan, D. and A.J. Ferguson. 1999. "On the Utilization of Sample Weights in Latent Variable Models." *Structural Equation Modeling: A Multidisciplinary Journal* 6: 305–321. Doi: https://doi.org/10.1080/10705519909540138.

Kim, J.K. and W.A. Fuller. 2004. "Fractional Hot Deck Imputation." *Biometrika* 91: 559–578. Doi: https://doi.org/10.1093/biomet/91.3.559.

Kim, J.K. and J. Shao. 2014. *Statistical Methods for Handling Incomplete Data*. Boca Raton, FL: CRC Press.

Kim, J.K. and M. Park. 2010. "Calibration Estimation in Survey Sampling." *International Statistical Review* 78: 21–39. Doi: 10.1111/j.1751-5823.2010.00099.x.

Kish, L. 1965. *Survey Sampling*. New York, NY: Wiley.

Kish, L. and M.R. Frankel. 1968. "Balanced Repeated Replication for Analytical Statistics." In *Proceedings of the Social Statistics Section of the American Statistical Association,* 1968, 2–10. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y1968/Balanced%20Repeated%20Replications%20For%20Analytical%20Statistics.pdf (accessed January 30, 2018).

Kish, L. and M.R. Frankel. 1970. "Balanced Repeated Replications for Standard Errors." *Journal of the American Statistical Association* 65: 1071–1094. Doi: https://doi.org/10.1080/01621459.1970.10481145.

Kish, L. and M.R. Frankel. 1974. "Inference from Complex Samples." *Journal of the Royal Statistical Society. Series B (Methodological)* 36: 1–37. Doi: http://www.jstor.org/stable/2984767.

Kolenikov, S. 2014. "Calibrating Survey Data using Iterative Proportional Fitting (Raking)." *The Stata Journal* 14: 22–59.

Kolenikov, S. 2010. "Resampling Variance Estimation for Complex Survey Data." *Stata Journal* 10: 165–199.

Kott, P.S. 2006. "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors." *Survey Methodology* 32: 133.

Kott, P.S. 2011. "A Nearly Pseudo-Optimal Method for Keeping Calibration Weights From Falling Below Unity In The Absence Of Nonresponse Or Frame Errors." *Pakistan Journal of Statistics* 27: 391–396.

Kott, P.S. 2012. "Why One Should Incorporate the Design Weights When Adjusting for Unit Nonresponse Using Response Homogeneity Groups." *Survey Methodology* 38: 95–99.

Kott, P.S. and D. Liao. 2012. "Providing Double Protection for Unit Nonresponse with a Nonlinear Calibration-Weighting Routine." *Survey Research Methods* 6: 105–111. Doi: http://dx.doi.org/10.18148/srm/2012.v6i2.5076.

Korn, E.L. and B.I. Graubard. 1999. *Analysis of Health Surveys*. New York, NY: Wiley.

Kovačević, M.S. and S.N. Rai. 2003. "A Pseudo Maximum Likelihood Approach to Multilevel Modelling of Survey Data." *Communications in Statistics-Theory and Methods* 32: 103–121. Doi: https://doi.org/10.1081/STA-120017802.

Kovar, J.G., J.N.K. Rao, and C.F.J. Wu. 1988. "Bootstrap and Other Methods to Measure Errors in Survey Estimates." *Canadian Journal of Statistics* 16: 25–45. Doi: 10.2307/3315214.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T.M. Ezzati-Rice, C. Casas-Cordero, M. Lemay, A. Peytchev, R.M. Groves, and T.E. Raghunathan. 2010. "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173: 389–407. Doi: 10.1111/j.1467-985X.2009.00621.x.

Krewski, D. and J.N.K. Rao. 1981. "Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods." *The Annals of Statistics* 9: 1010–1019. Doi: http://www.jstor.org/stable/2240615.

Lavallée, P. and J.F. Beaumont. 2016. "Weighting: Principles and Practicalities." In *The SAGE Handbook of Survey Methodology*, edited by C. Wolf, D. Joye, T.W. Smith, and Y. Fu, 460–476. London: Sage.

Le Guennec, J., and O. Sautory. 2002. "CALMAR 2: Une nouvelle version de la macro Calmar de redressment d'echantillon par calage." Actes des Journeés de Méthodologie Statistique, INSEE, Paris. Available in French at: http://jms.insee.fr/files/documents/2002/327_1-JMS2002_SESSION1_LE-GUENNEC-SAUTORY_CALMAR-2_ACTES.PDF (accessed January 30, 2018).

Lemeshow, S. and P. Levy. 1978. "Estimating the Variance of Ratio Estimates in Complex Sample Surveys with Two Primary Units per Stratum—A Comparison of Balanced Replication and Jackknife Techniques." *Journal of Statistical Computation and Simulation* 8: 191–205. Doi: https://doi.org/10.1080/00949657908810266.

Lessler, J.T. and W.D. Kalsbeek. 1992. *Nonsampling Error in Surveys*. Wiley.

Li, J. and R. Valliant. 2009. "Survey Weighted Hat Matrix and Leverages." *Survey Methodology* 35: 15–24.

Li, J. and R. Valliant. 2011a. "Linear Regression Influence Diagnostics for Unclustered Survey Data." *Journal of Official Statistics* 27: 99–119.

Li, J. and R. Valliant. 2011b. "Detecting Groups of Influential Observations in Linear Regression using Survey Data: Adapting the Forward Search Method." *Pakistan Journal of Statistics* 27: 507–528.

Li, J. and R. Valliant. 2015. "Linear Regression Diagnostics in Cluster Samples." *Journal of Official Statistics* 31: 61–75. https://doi.org/10.1515/jos-2015-0003.

Liao, D. and R. Valliant. 2012a. "Variance Inflation Factors in the Analysis of Complex Survey Data." *Survey Methodology* 38: 53–62.

Liao, D. and R. Valliant. 2012b. "Condition Indexes and Variance Decompositions for Diagnosing Collinearity in Linear Model Analysis of Survey Data." *Survey Methodology* 38: 189–202.

Little, R.J.A. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review* 54: 139–157. Doi: http://www.jstor.org/stable/1403140.

Little, R.J.A. 2003. "The Bayesian Approach to Sample Survey Inference." In *Analysis of Survey Data*, edited by R.L. Chambers and C.J. Skinner, 49–57. Chichester, West Sussex: Wiley.

Little, R.J.A. 2004. "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling." *Journal of the American Statistical Association* 99: 546–556. Doi: https://doi.org/10.1198/016214504000000467.

Little, R.J.A. 2006. "Calibrated Bayes: A Bayes/Frequentist Roadmap." *The American Statistician* 60: 213–223. Doi: https://doi.org/10.1198/000313006X117837.

Little, R.J.A. 2011. "Calibrated Bayes, for Statistics in General, and Missing Data in Particular." *Statistical Science* 26: 162–186. Doi: 10.1214/10-STS318.

Little, R.J.A. 2012. "Calibrated Bayes: An Alternative Inferential Paradigm for Official Statistics (with discussion and rejoinder)." *Journal of Official Statistics* 28: 309–372.

Little, R.J.A. 2015. "Calibrated Bayes, An Inferential Paradigm for Official Statistics in the Era of Big Data." *Statistical Journal of the IAOS* 31: 555–563. Doi: https://doi.org/10.3233/SJI-150944.

Little, R.J.A. and S. Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31: 161–168.

Little, R.J.A. and H. Zheng. 2007. “The Bayesian Approach to the Analysis of Finite Population Surveys.” *Bayesian Statistics* 8: 1–20.

Lohr, S. 2009. *Sampling: Design and Analysis, Second Edition*. Boston, MA: Cengage Learning.

Lohr, S. 2014. “Design Effects for a Regression Slope in a Cluster Sample.” *Journal of Survey Statistics and Methodology* 2: 97–125. Doi: https://doi.org/10.1093/jssam/smu003.

Lumley, T. 2010. *Complex Surveys: A Guide to Analysis Using R*. New York, NY: Wiley.

Lumley, T. and A. Scott. 2013. “Partial Likelihood Ratio Tests for the Cox Model under Complex Sampling.” *Statistics in Medicine* 32: 110–123. Doi: https://doi.org/10.1002/sim.5492.

Lumley, T. and A. Scott. 2014. “Tests for Regression Models Fitted to Survey Data.” *Australian & New Zealand Journal of Statistics* 56: 1–14. Doi: https://doi.org/10.1111/anzs.12065.

Lumley, T. and A. Scott. 2015. “AIC and BIC for Modeling with Complex Survey Data.” *Journal of Survey Statistics and Methodology* 3: 1–18. Doi: https://doi.org/10.1093/jssam/smu021.

Lumley, T. and A. Scott. 2017. “Fitting Regression Models to Survey Data.” *Statistical Science* 32: 265–278. Doi: https://10.1214/16-STS605.

Lundström, S. and C.E. Särndal. 1999. “Calibration as a Standard Method for Treatment of Nonresponse.” *Journal of Official Statistics* 15: 305–327.

Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. 2012. *The BUGS book: A Practical Introduction to Bayesian Analysis*. CRC press.

Morel, G. 1989. “Logistic Regression under Complex Survey Designs.” *Survey Methodology* 15: 203–223.

Muthén, B.O. and A. Satorra. 1995. “Complex Sample Data in Structural Equation Modeling.” *Sociological Methodology* 25: 267–316. Doi: https://doi.org/10.2307/271070.

Nordberg, L. 1989. “Generalized Linear Modeling of Sample Survey Data.” *Journal of Official Statistics* 5: 223–239.

Oberski, D.L. 2014. “lavaan.survey: An R package for Complex Survey Analysis of Structural Equation Models.” *Journal of Statistical Software* 57: 1–27. Doi: https://doi.org/10.18637/jss.v057.i01.

Oh, H.L. and F.J. Scheuren. 1983. “Weighting Adjustment for Unit Nonresponse.” In *Incomplete Data in Sample Surveys*, edited by W.G. Madow, I. Olkin, and D.B. Rubin, 143–184. New York: Academic Press.

Orchard, T. and M.A. Woodbury. 1972. “A Missing Information Principle: Theory and Applications.” Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics, 697–715. University of California Press: Berkeley, CA. Available at: https://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200514117 (accessed January 30, 2018).

Owen, A.B. 2001. *Empirical Likelihood*. New York: Chapman & Hall.

Pacifico, D. 2014. “sreweight: A Stata Command to Reweight Survey Data to External Totals.” *The Stata Journal* 14: 4–21.

Peress, M. 2010. "Correcting for Survey Nonresponse using Variable Response Propensity." *Journal of the American Statistical Association* 105: 1418–1430. Doi: https://doi.org/10.1198/jasa.2010.ap09485.

Pfeffermann, D. 1993. "The Role of Sampling Weights When Modeling Survey Data." *International Statistical Review* 61: 317–337. Doi: https://doi.org/10.2307/1403631.

Pfeffermann, D. 2011. "Modelling of Complex Survey Data: Why Model? Why Is It a Problem? How Can We Approach It?" *Survey Methodology* 37: 115–136.

Pfeffermann, D., F.A.D.S. Moura, and P.L.D.N. Silva. 2006. "Multi-level Modelling Under Informative Sampling." *Biometrika* 93: 943–959. Doi: https://doi.org/10.1093/biomet/93.4.943.

Pfeffermann, D. and A. Sikov. 2011. "Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information." *Journal of Official Statistics* 27: 181–209.

Pfeffermann, D., C.J. Skinner, D.J. Holmes, H. Goldstein, and J. Rasbash. 1998. "Weighting for Unequal Selection Probabilities in Multilevel Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60: 23–40. Doi: https://doi.org/10.1111/1467-9868.00106.

Pfeffermann, D. and M. Sverchkov. 2009. "Inference Under Informative Sampling." In *Handbook of Statistics – Sample Surveys: Inference and Analysis (Volume 29, Part B)*, edited by V.N. Gudivada, V.V. Raghavan, V. Govindaraju, and C.R. Rao, 455–487.

Potter, F.J. 1990. "A Study of Procedures to Identify and Trim Extreme Sampling Weights." In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 225–230. Available at: http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1990_034.pdf (accessed January 30, 2018).

Rabe-Hesketh, S. and A. Skrondal. 2006. "Multilevel Modelling of Complex Survey Data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169: 805–827. Doi: https://doi.org/10.1111/j.1467-985X.2006.00426.x.

Rao, J.N.K. 2005. "Interplay Between Sample Survey Theory and Practice: An Appraisal." *Survey Methodology* 31: 117–138.

Rao, J.N.K. and J. Shao. 1999. "Modified Balanced Repeated Replication for Complex Survey Data." *Biometrika* 86: 403–415. Doi: https://doi.org/10.1093/biomet/86.2.403.

Rao, J.N.K. and A.J. Scott. 1981. "The Analysis of Categorical Data from Complex Sample Surveys: Chi-squared Tests for Goodness of Fit and Independence in Two-way Tables." *Journal of the American Statistical Association* 76: 221–230. Doi: https://doi.org/10.2307/2287815.

Rao, J.N.K. and A.J. Scott. 1984. "On Chi-squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data." *The Annals of Statistics* 12: 46–60. Doi: http://dx.doi.org/10.1214/aos/1176346391.

Rao, J.N.K. and A.J. Scott. 1987. "On Simple Adjustments to Chi-square Tests with Sample Survey Data." *The Annals of Statistics* 15: 385–397. Doi: https://doi.org/10.1214/aos/1176350273.

Rao, J.N.K. and C.F.J. Wu. 1985. "Inference from Stratified Samples: Second-order Analysis of Three Methods for Nonlinear Statistics." *Journal of the American Statistical Association* 80: 620–630. Doi: https://doi.org/10.2307/2288478.

Rao, J.N.K. and C.F.J. Wu. 1987. “Methods for Standard Errors and Confidence Intervals from Sample Survey Data: Some Recent Work.” *Bulletin of the International Statistical Institute* 3: 5–21.

Rao, J.N.K. and C.F.J. Wu. 1988. “Resampling Inference with Complex Survey Data.” *Journal of the American Statistical Association* 83: 231–241. Doi: https://doi.org/10.2307/2288945.

Reiter, J.P., T.E. Raghunathan, and S.K. Kinney. 2006. “The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data.” *Survey Methodology* 32: 143–149.

Roberts, G., J.N.K. Rao, and S. Kumar. 1987. “Logistic Regression Analysis of Sample Survey Data.” *Biometrika* 74: 1–12. Doi: https://doi.org/10.2307/2336016.

Rubin, D.B. 1985. “The Use of Propensity Scores in Applied Bayesian Inference.” In *Bayesian Statistics 2*, edited by J.M. Bernardo, M.H. Degroot, D.V. Lindley, and A.F.M. Smith, 463–472. Elsevier Science Publishers B.V.

Ryan, B.L., J. Koval, B. Corbett, A. Thind, M.K. Campbell, and M. Stewart. 2015. “Assessing the Impact of Potentially Influential Observations in Weighted Logistic Regression.” *The Research and Data Centres Information and Technical Bulletin (Statistics Canada)* 7. Available at: http://www.statcan.gc.ca/pub/12-002-x/2015001/article/14147-eng.htm (accessed January 30, 2018).

Särndal, C.E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag Inc.

Särndal, C.E. and B. Swensson. 1987. “A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse.” *International Statistical Review* 55: 279–294. Doi: https://doi.org/10.2307/1403406.

Shah, B.V., M.M. Holt, and R.E. Folsom. 1977. “Inference about Regression Models from Sample Survey Data.” *Bulletin of the International Statistical Institute* 47: 43–57.

Shao, J. 2003. “Impact of the Bootstrap on Sample Surveys.” *Statistical Science* 18: 191–198.

Shao, J. and R.R. Sitter. 1996. “Bootstrap for Imputed Survey Data.” *Journal of the American Statistical Association* 91: 1278–1288. Doi: https://doi.org/10.2307/2291746.

Shao, J. and D. Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer.

Si, Y., N.S. Pillai, and A. Gelman. 2015. “Bayesian Nonparametric Weighted Sampling Inference.” *Bayesian Analysis* 10: 605–625. Doi: http://dx.doi.org/10.1214/14-BA924.

Sitter, R.R. and C. Wu. 2001. “A Note on Woodruff Confidence Intervals for Quantiles.” *Statistics & Probability Letters* 52: 353–358. Doi: https://doi.org/10.1016/S0167-7152(00)00207-8.

Skinner, C.J. and D.J. Holmes. 2003. “Random Effects Models for Longitudinal Survey Data.” Chapter 14 in *Analysis of Survey Data*, edited by R.L. Chambers and C.J. Skinner. John Wiley and Sons.

Skinner, C.J., D. Holt, and T.F. Smith. 1989. *Analysis of Complex Surveys*. John Wiley & Sons.

Stapleton, L.M. 2002. “The Incorporation of Sample Weights into Multilevel Structural Equation Models.” *Structural Equation Modeling* 9: 475–502. Doi: https://doi.org/10.1207/S15328007SEM0904_2.

Stapleton, L.M. 2006. “An Assessment of Practical Solutions for Structural Equation Modeling with Complex Sample Data.” *Structural Equation Modeling* 13: 28–58. Doi: https://doi.org/10.1207/s15328007sem1301_2.

Stapleton, L.M. and Y. Kang. 2016. “Design Effects of Multilevel Estimates From National Probability Samples.” *Sociological Methods & Research*, available at http://journals.sagepub.com/doi/abs/10.1177/0049124116630563 (accessed January 30, 2018). Doi: https://doi.org/10.1177/0049124116630563.

Thompson, M.E. 2015. “Using Longitudinal Complex Survey Data.” *Annual Review of Statistics and Its Application* 2: 305–320. Doi: https://doi.org/10.1146/annurev-statistics-010814-020403.

Toth, D. 2017. “rpms: An R Package for Modeling Survey Data with Regression Trees.” Available at https://cran.r-project.org/web/packages/rpms/vignettes/rpms_2017_02_10.pdf (accessed January 1, 2018).

Valliant, R. 2004. “The Effect of Multiple Weighting Steps on Variance Estimation.” *Journal of Official Statistics* 20(1): 1–18.

Valliant, R., J.A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite Population Sampling and Inference: a Prediction Approach*. New York: Wiley.

Valliant, R. and K.F. Rust. 2010. “Degrees of Freedom Approximations and Rules-of-Thumb.” *Journal of Official Statistics* 26: 585–602.

West, B.T. 2009. “A Simulation Study of Alternative Weighting Class Adjustments for Nonresponse when Estimating a Population Mean from Complex Sample Survey Data.” In *Proceedings of the section on Survey Research Methods: Joint Statistical Meetings*, 4920–4933. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2009/Files/305394.pdf (accessed January 30, 2018).

West, B.T., L. Beer, W. Gremel, J. Weiser, C. Johnson, S. Garg, and J. Skarbinski. 2015. “Weighted Multilevel Models: A Case Study.” *American Journal of Public Health* 105: 2214–2215. Doi: https://dx.doi.org/10.2105%2FAJPH.2015.302842.

West, B.T., P.A. Berglund, and S.G. Heeringa. 2008. “A Closer Examination of Subpopulation Analysis of Complex-Sample Survey Data.” *The Stata Journal* 8: 520–531.

West, B.T. and S.E. McCabe. 2017. “Alternative Approaches to Assessing Nonresponse Bias in Longitudinal Survey Estimates: An Application to Substance Use Outcomes among Young Adults in the U.S.” *American Journal of Epidemiology* 185: 591–600. Doi: https://doi.org/10.1093/aje/kww115.

West, B.T., J.W. Sakshaug, and G.A.S. Aurelien. 2016. “How Big of a Problem is Analytic Error in Secondary Analyses of Survey Data?” *PLoS ONE* 11. Doi: https://doi.org/10.1371/journal.pone.0158120.

Wolter, K.M. 2007. *Introduction to Variance Estimation*, *Second Edition*. New York: Springer-Verlag.

Woodruff, R.S. 1952. “Confidence Intervals for Medians and other Position Measures.” *Journal of the American Statistical Association* 47: 635–646. Doi: https://doi.org/10.2307/2280781.

Wu, Y.Y. and W.A. Fuller. 2006. "Estimation of Regression Coefficients with Unequal Probability Samples." In *Proceedings of the Survey Research Methods Section: American Statistical Association*, 3892–3899. Available at: https://ww2.amstat.org/sections/srms/Proceedings/y2006/Files/JSM2006-000807.pdf (accessed January 30, 2018).

Wun, L.M., T.M. Ezzati-Rice, N. Diaz-Tena, and J. Greenblatt. 2007. "On Modeling Response Propensity for Dwelling Unit (DU) Level Non-response Adjustment in the Medical Expenditure Panel Survey (MEPS)." *Statistics in Medicine* 26: 1875–1884. Doi: https://doi.org/10.1002/sim.2809.

Zangeneh, S.Z. and R.J. Little. 2015. "Bayesian Inference for the Finite Population Total from a Heteroscedastic Probability Proportional to Size Sample." *Journal of Survey Statistics and Methodology* 3: 162–192. Doi: https://doi.org/10.1093/jssam/smv002.

Zhang, D. and X. Lin. 2008. "Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and Other Related Topics." In *Random Effect and Latent Variable Model Selection*, edited by D.B. Dunson. Springer Lecture Notes in Statistics, 192.

Zheng, H. and R.J. Little. 2003. "Penalized Spline Model-Based Estimation of the Finite Populations Total from Probability-Proportional-to-Size Samples." *Journal of Official Statistics* 19: 99–117.

Zhou, H., M.R. Elliott, and T.E. Raghunathan. 2016a. "Synthetic Multiple-Imputation Procedure for Multistage Complex Samples." *Journal of Official Statistics* 32: 231–256. Doi: https://doi.org/10.1515/JOS-2016-0011.

Zhou, H., M.R. Elliott, and T.E. Raghunathan. 2016b. "Multiple Imputation in Two-Stage Cluster Samples Using the Weighted Finite Population Bayesian Boostrap." *Journal of Survey Statistics and Methodology* 4: 139–170. Doi: https://doi.org/10.1093/jssam/smv031.

Zhou, H., M.R. Elliott, and T.E. Raghunathan. 2016c. "A Two-Step Semiparametric Method to Accommodate Sampling Weights in Multiple Imputation." *Biometrics* 72: 242–252. Doi: https://10.1111/biom.12413.

Zhu, M. 2014. "Analyzing Multilevel Models with the GLIMMIX Procedure." Paper SAS026-2014. Cary, NC: SAS Institute, Inc.

Received July 2017
Revised December 2017
Accepted February 2018