Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis

Kenneth A. Bollen,[1,2] Paul P. Biemer,[3,4] Alan F. Karr,[4] Stephen Tueller,[4] and Marcus E. Berzofsky[4]

[1]Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, North Carolina 27599

[2]Department of Sociology, University of North Carolina, Chapel Hill, North Carolina 27599; email: bollen@unc.edu

[3]Odum Institute for Research in Social Science, University of North Carolina, Chapel Hill, North Carolina 27599

[4]RTI International, Research Triangle Park, North Carolina 27709

Annu. Rev. Stat. Appl. 2016. 3:375–92

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-011516-012958

Keywords

complex samples, survey weights, Hausman test, survey data, probability sampling

Abstract

Researchers apply sampling weights to take account of unequal sample selection probabilities and to frame coverage errors and nonresponses. If researchers do not weight when appropriate, they risk having biased estimates. Alternatively, when they unnecessarily apply weights, they can create an inefficient estimator without reducing bias. Yet in practice researchers rarely test the necessity of weighting and are sometimes guided more by the current practice in their field than by scientific evidence. In addition, statistical tests for weighting are not widely known or available. This article reviews empirical tests to determine whether weighted analyses are justified. We focus on regression models, though the review's implications extend beyond regression. We find that nearly all weighting tests fall into two categories: difference in coefficients tests and weight association tests. We describe the distinguishing features of each category, present their properties, and explain the close relationship between them. We review the simulation evidence on their sampling properties in finite samples. Finally, we highlight the unanswered theoretical and practical questions that surround these tests and that deserve further research.

1. INTRODUCTION

Social, health, and federal statistics scientists analyze survey data. Survey data typically arise from complex sample designs involving unequal probability sampling of population units, rather than simple random sampling and other equal probability of selection method (EPSEM) designs. Survey weights that are proportional to the inverse selection probabilities adjust for departures from EPSEM sampling. Survey statisticians may further adjust these selection weights to account for unit nonresponse, frame coverage errors, and other sources of sampling bias. Pfeffermann (1996), Biemer & Christ (2008), and Valliant et al. (2013) provide overviews of the methodologies and issues in survey weighting.

Survey weights are typically essential to avoid bias when estimating population means or proportions for variables as part of a descriptive analysis. However, whether researchers should use survey weights for modeling the relationships between explanatory and dependent variables has been debated in the literature since the early 1950s (e.g., Klein & Morgan 1951, Horvitz & Thompson 1952, Holt et al. 1980). A serious disadvantage of using survey weights unnecessarily is that they can substantially inflate the variance of the model parameter estimates, even when the unweighted analysis produces essentially the same estimates. In that case, efficiency of the estimators and statistical power are improved by not using weights.

Another concern about using weights is anchored more to privacy concerns than to statistical considerations (e.g., Fienberg 2009). Some have argued that releasing weights might provide sufficient information to permit deductive disclosure of respondents' identities, or sensitive information about them. In this situation, in deciding whether to release weights, data producers should consider whether weights are needed for unbiased inferences.

In our experience and in examining what is done in practice, we find that one group, mostly from biostatistics, public health, and survey methods traditions, generally uses weights. Another group, mostly from the social sciences (including econometrics), generally does not use weights. A third, smaller group estimates both weighted and unweighted analyses and performs informal ad hoc comparisons of the coefficients to reach a decision on weighting the analysis. The informal comparison consists of a subjective assessment of whether the coefficients differ significantly between the weighted and unweighted analyses. All these practices seem rooted in tradition rather than analytic results.

Although the question of whether to use weights applies more generally to the analysis of data from complex surveys, our focus here is on methods for testing whether survey weights are needed in a regression analysis, which is perhaps the most widely used statistical model in the social and behavioral sciences. Regression models are a useful point for the discussion in that they provide the arena in which most of the debate between weighted and unweighted analysis occurs. In the conclusion, we mention several references that consider weighting in regression with categorical dependent variables. Our article has four purposes: (*a*) to review the major diagnostic tests of the necessity for weighting, (*b*) to examine the assumptions and properties of these tests as documented in the literature, (*c*) to develop a framework to unite this seemingly diverse set of diagnostic tests under many fewer categories, and (*d*) to identify gaps in the literature where additional research is required. The article also lays the foundation for the study of other types of models, especially those that are closely related to regression models.

Interestingly, the literature on weighting tests often ignores possible effects of cluster sampling on the estimates, opting to focus primarily on unclustered, unequal probability sampling. We have done the same to simplify the exposition of the results and to remain consistent with the literature we have reviewed. Introducing clustering does little to illuminate the issues of weighting that are of primary interest in this article, but adds considerable complexity to the variance estimation

formulas. Thus, the above four purposes of the article can be addressed more clearly and succinctly by limiting the discussion to unclustered data. In addition, nearly all the material here applies more generally to the question of which of two or more sets of weights to use rather than whether to ignore the weights altogether. The latter question is just a special case in which one of the sets contains equal weights. This generalization is important. For instance, Asparouhov & Muthén (2007) compared two sets of weights: trimmed (or Winsorized) and untrimmed. Similarly, adjusting weights for nonresponse may be important for national estimates but not important in some modeling contexts. Comparing the two sets of weights addresses this question empirically.

Also, aspects of the "to weight or not to weight" decision may not be fully transparent to analysts. This is particularly true for survey datasets in which design variables are not released by agencies, which may be done for reasons such as protecting confidentiality (Fienberg 2009). In this case, analysts may treat weights in the modeling as surrogates for the unreleased design variables. For instance, in health data, if there is stratification by geography but geographic identifiers are not in, or are coarsened in, the released data, the geographic information may still be reflected in weights. The agency, with access to the full data, may determine that weights are not required to model an outcome of interest, whereas an analyst may conclude that they are. This is not a contradiction but simply a reflection of the fact that "the data" means different things to different people. The same phenomenon can arise if the released data have been subjected to statistical disclosure limitation, which, because it attenuates structure in the data, may reduce the role of weights (Cox et al. 2011, Fienberg 2009).

Sections 2 and 3 briefly describe the components of survey weights and the model-based and design-based perspectives on survey data that have shaped the views of weighting. Section 4 reviews and classifies many of the major tests of whether to weight. It also discusses the special case of testing when only a subset of coefficients is of interest. Section 5 provides a comparison of the tests and reviews studies that have investigated the statistical properties of the tests. Finally, Section 6 concludes with research questions, the answers to which will increase the practical usefulness of these diagnostic tests.

2. COMPONENTS OF SURVEY WEIGHTS

Before proceeding, we briefly consider the components of survey weights. A more complete discussion can be found in Biemer & Christ (2008). The construction of the survey weight begins with the selection weight, which is then adjusted to compensate for frame coverage errors and nonresponse. To see how this works, let P denote the target population, F denote the units in the target population that are on the sampling frame, S denote the units selected from the frame for the sample, and R denote the units in S that respond to the survey. Note that, in general, $R \subseteq S \subseteq F \subseteq P$. Suppose S constitutes a random sample of n persons selected with known selection probabilities, $\pi_{Si} = \Pr(i \in S)$, $i = 1, \ldots, n$. In the case of full response (i.e., $R = S$) and no frame coverage errors (i.e., $F = P$), the classic Horvitz-Thompson (HT) estimator of the population total, Y, is given as $\hat{Y} = \sum_{i \in S} w_{Si} y_i$, where $w_{Si} = \pi_{Si}^{-1}$ (Horvitz & Thompson 1952). Thus, the HT estimator gives rise to the survey weights not only for estimating totals and other descriptive statistics, but also for modeling.

In the presence of nonresponse ($R \subset S$) and undercoverage ($F \subset P$), the HT estimator can be generalized (see, for example, Kalsbeek & Agans 2007) by replacing w_{Si} with $w_i = (\pi_i)^{-1}$, where $\pi_i = \Pr(i \in F, i \in S, i \in R)$. Note that π_i can be rewritten as

$$\pi_i = \Pr(i \in F) \times \Pr(i \in S | i \in F) \times \Pr(i \in R | i \in F, i \in S) = \pi_{Fi} \pi_{Si} \pi_{Ri},$$

where $\pi_{Fi} = \Pr(i \in F)$, $\pi_{Si} = \Pr(i \in S | i \in F)$, and $\pi_{Ri} = \Pr(i \in R | i \in F, i \in S)$ with inverses w_{Fi}, w_{Si}, and w_{Ri}, respectively. Thus, the survey weight is the product of three weights: $w_i = w_{Fi} w_{Si} w_{Ri}$. The component w_{Si} is referred to as the base or selection weight because it is the starting point for weight construction and derives from the survey design. The other weight components are essentially regarded as adjustments to this weight. One complication is that only w_{Si} is known; however, both w_{Fi} and w_{Ri} can be estimated. How exactly to do this is beyond the scope of this review, but numerous articles in survey statistics discuss their estimation (see, for example, Lohr 2010 or Valliant et al. 2013).

3. PERSPECTIVES ON POPULATIONS

One source of confusion that surrounds weighting is that there are at least two different perspectives on the nature of the population to which inferences are being made (e.g., see Sterba 2009). A common way to draw the distinction is to refer to one as the infinite population (or model-based) perspective and the other as the finite population (or design-based) perspective. In addition, within the model-based perspective there are two approaches: one based on a superpopulation of values of y from which the population of y-values is drawn and another based on Bayesian survey inference (see, for example, Little 2009), which adds the specification of a prior distribution for the values of y in the superpopulation. Although large-sample inferences for these two approaches are often similar, the Bayesian approach has some advantages in small samples (Little 2009). Nevertheless, in our review, we found no papers on weighting tests under the Bayesian survey inference approach and thus it is not considered here. Hence, in this section we describe the assumptions made about the regression equation when using the superpopulation (model-based) and finite population perspectives, starting with the former approach.

3.1. Model-Based Inference

Consider the regression equation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{Y}$ is the $n \times 1$ vector of values of a dependent variable, $\boldsymbol{X}$ is the $n \times k$ matrix of values of the explanatory variables with the first column of ones for the regression constant, $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of values of the disturbances or error variable. $\boldsymbol{\varepsilon}$ is the collection of all variables other than those in $\boldsymbol{X}$ that influence $\boldsymbol{Y}$.

In model-based inference, we make several assumptions about the error. We assume that it is a random variable with certain properties:

$$\text{A. } E(\boldsymbol{\varepsilon}) = 0;$$

$$\text{B. } E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I};$$

$$\text{C. } \boldsymbol{\varepsilon} \perp \boldsymbol{X}.$$

Assumption A is that all the omitted influences have a mean of zero when combined. Assumption B is that the variance of the error (σ^2) is the same for all cases and that the errors of different cases are uncorrelated. In other words, Assumption B represents homoscedasticity and no autocorrelation of the error variable. Assumption C is that the error variable ($\boldsymbol{\varepsilon}$) is distributed independently ($\perp$) of the explanatory variables ($\boldsymbol{X}$).

To these assumptions we add another,

$$\text{D. rank}(\boldsymbol{X}) = k,$$

which means that there are no linear dependencies (perfect collinearity) among the explanatory variables. Under these assumptions, the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, is unbiased [$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$] and is consistent [plim($\hat{\boldsymbol{\beta}}$) = $\boldsymbol{\beta}$], and the variance of the OLS estimator $\hat{\boldsymbol{\beta}}$ is $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Several points bear emphasis. One is that $\boldsymbol{\varepsilon}$ is a random variable, and this is one feature of the model that contributes to $\boldsymbol{Y}$ being a random variable. Note that, as presented above, the explanatory variables $\boldsymbol{X}$ are also random variables, though we make few assumptions about them other than that they are independent of the error and the matrix has full rank.

Variations on the above assumptions are sometimes used for model-based inferences; thus, the above should not be considered the only assumptions that might occur with this perspective. For instance, some researchers would replace Assumption C [$\boldsymbol{\varepsilon} \perp \boldsymbol{X}$] with the weaker assumption that they are uncorrelated (C′), Cov($\boldsymbol{\varepsilon}'$, $\boldsymbol{X}$) = 0.

The consistency of the OLS $\hat{\boldsymbol{\beta}}$ persists, but we can no longer establish that it is unbiased [$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$] in finite samples, though it would be asymptotically unbiased [$AE(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$]. In addition, we could relax the homoscedasticity assumptions, but we would need to make adjustments to our estimator of the variance of $\hat{\boldsymbol{\beta}}$.[1]

With these qualifications and alternative assumptions in mind, a key idea about the multiple regression equation from the model-based inference is that there is no assumption of a finite population to which inferences are being made. An infinite population is the target of inference. The sampling variability results from the use of a random error and, in many cases, the random explanatory variables that make the dependent variable a random variable. One or more of the preceding properties of OLS fail when one or more of the preceding assumptions fail.

3.2. Design-Based Inference

For design-based inference, we consider the finite population model defined for the population U, where there are N cases in the finite population. The finite population regression is

$$\boldsymbol{Y}_N = \boldsymbol{X}_N\boldsymbol{B} + \boldsymbol{E}_N, \tag{2}$$

where $\boldsymbol{Y}_N$ is an $N \times 1$ vector, $\boldsymbol{X}_N$ is a $N \times k$ matrix of explanatory variables with the first column of ones for the regression constants, $\boldsymbol{B}$ is the $k \times 1$ corresponding vector of coefficients, and $\boldsymbol{E}_N$ is the $N \times 1$ vector of errors.

The finite population formula for OLS regression coefficients is

$$\boldsymbol{B} = (\boldsymbol{X}_N'\boldsymbol{X}_N)^{-1}\boldsymbol{X}_N'\boldsymbol{Y}_N,$$

with residuals defined by

$$\boldsymbol{E}_N = \boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{B}.$$

[1]Another way to frame this problem is to consider maximum likelihood estimation of the multiple regression, which assumes a normal distribution of the disturbance term, and to consider the weights informative. Our estimation approach does not require the assumption of normality. In addition, if the weights are informative, then the key assumption that the error is independent of or uncorrelated with $\boldsymbol{X}$ is violated.

By construction, $\boldsymbol{E}_N$ has the following properties:

$$1'\boldsymbol{E}_N = 0$$

and

$$\boldsymbol{X}'_N\boldsymbol{E}_N = 0.$$

In other words, the population total of the regression residuals is zero and the population cross-product of the explanatory variables and residuals is also zero.

For estimability, we require that

$$\text{rank}(\boldsymbol{X}_N) = k.$$

If it were rank deficient, we could not apply OLS because $\boldsymbol{X}'_N\boldsymbol{X}_N$ would not have an inverse.

Suppose a sample, S, of n units is drawn from U and let P_i denote the probability that the ith population unit is selected, for $i = 1, \ldots, N$. Further assume that $\Pr(i \in S, i' \in S \mid i, i' \in U) = \Pr(i \in S \mid i \in U)\Pr(i' \in S \mid i' \in U) = P_i P_{i'}$ for $i \neq i'$; that is, sampling proceeds independently at each draw. A design-consistent estimator of $\boldsymbol{B}$ is

$$\hat{\boldsymbol{B}}_W = (\boldsymbol{X}'_n\boldsymbol{W}\boldsymbol{X}_n)^{-1}\boldsymbol{X}'_n\boldsymbol{W}\boldsymbol{Y}_n,$$

where $\boldsymbol{Y}_n$ is an $n \times 1$ vector of sample Y values, $\boldsymbol{X}_n$ is a $n \times k$ matrix of explanatory variables with the first column of ones for the regression constants, and $\hat{\boldsymbol{B}}_W$ is the $k \times 1$ vector of estimated regression coefficients. $\boldsymbol{W}$ is the $n \times n$ diagonal matrix with w_j on the jth diagonal and

$$w_j = P_i^{-1} \quad \text{for} \quad j = i, j \in S, i \in U.$$

Pfeffermann & Sverchkov (2009) show that $\hat{\boldsymbol{B}} = (\mathbf{X}'_n\boldsymbol{X}_n)^{-1}\boldsymbol{X}'_n\boldsymbol{Y}_n$ will also be a consistent estimator of $\boldsymbol{B}$ if and only if

$$\text{Cov}_S(w_j, y_j \mid \boldsymbol{x}_j) = 0 \tag{3}$$

for all $j \in S$ or, equivalently, if

$$E_S(w_j \mid y_j, \boldsymbol{x}_j) = E_S(w_j \mid \boldsymbol{x}_j),$$

where $\boldsymbol{x}_j$ is the vector of explanatory variables for the jth sample unit and $E_S(\bullet)$ is the expectation over all possible samples of size n selected with probabilities P_i for $i = 1, \ldots, N$.

3.3. Model-Based Inference and Weighting

We can combine the model-based and design-based perspectives if we regard the finite population of the design-based perspective as a simple random sample of size N from the infinite population described in the model-based perspective. Let $E_P(\bullet)$ denote the joint expectation of this hypothetical sampling mechanism as well as the distribution of the error term, $\boldsymbol{\varepsilon}$ in Equation 1, and let $E_S(\bullet)$ denote the expectation with respect to the sampling from the finite population. Pfeffermann & Sverchkov (2009) show that the result in Equation 3 can be extended to include this perspective in that $\hat{\boldsymbol{B}} = (\boldsymbol{X}'_n\boldsymbol{X}_n)^{-1}\boldsymbol{X}'_n\boldsymbol{Y}_n$ is now a consistent estimator of $\boldsymbol{\beta}$ in Equation 1 if

$$\text{Cov}(w_j, y_j \mid \boldsymbol{x}_j) = 0 \tag{4}$$

or, equivalently, if

$$E_P(y_j \mid \boldsymbol{x}_j) = E_S(w_j y_j \mid \boldsymbol{x}_j). \tag{5}$$

These results are key to showing the equivalence between the difference in coefficients (DC) tests and the weight association (WA) tests described in the next section.

The remainder of this article focuses primarily on this combined model-based and design-based perspective because most of the weighting tests in the literature either explicitly or tacitly assume this perspective. In fact, much of the confusion in the literature stems from a lack of specificity about which perspective is assumed for a test.[2]

4. TESTS OF WHETHER TO WEIGHT

Suppose that a researcher wishes to test whether weights are needed in a regression analysis. A search of the literature reveals a variety of diagnostic tests. Using data from Pfeffermann & Sverchkov (1999), this researcher might form the residuals from the OLS regression of the dependent variable on the original covariates and then correlate these residuals and various moments of the residuals with the weights. A significance test of whether the correlation is zero in the population and hence whether weights are required could be conducted using Fisher's Z transformation of the correlation (Fisher 1915, 1921) or a bootstrap estimate of the variance of the correlation (Efron 1979). Alternatively, still following Pfeffermann & Sverchkov, the analyst might decide to regress the residuals on the weight variable (w_j) and to perform a t-test of whether the coefficient of w_j is statistically significant. Another researcher might turn to DuMouchel & Duncan (1983) and regress the dependent variable on both the unweighted and weighted covariates. An F-test of whether all coefficients of the weighted covariates are zero would indicate whether weights are required. Finally, a different researcher might turn to a Hausman (1978) chi-square test as modified by Pfeffermann (1993) to determine whether the differences in the coefficients of the weighted and unweighted estimates are significant.

These are just a few of the diagnostic tests that the literature provides to decide the necessity of weights. The number of tests and the scarcity of comparisons of their properties create confusion about which, if any, to use. Do these and other diagnostics available in the literature test equivalent hypotheses? What assumptions are required for each test? Should we expect similar or different conclusions depending on the test employed? More generally, which test should be used and under what conditions? Though our review cannot answer all such questions, we can simplify the matter by demonstrating the relationships between most of these tests and clarifying the assumptions that underlie them. To begin, we classify nearly all these tests into two groups: DC tests and WA tests. Later in the article we also show that most of the tests in these two groups are closely related. But first we give an overview of these two groups of tests.

4.1. Difference in Coefficients Tests

The DC tests compare the coefficients of the weighted and unweighted analyses and assess whether these differences are statistically significantly different from zero. Start with the regression equation

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{6}$$

assuming

$$E(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = 0$$

[2]Bayesian statistics provides another perspective on weighting (e.g., Sugden & Smith 1984; Smith 1988; Smith & Sugden 1988; Fienberg 2009, 2011; Little 2009). However, we are unaware of any papers that propose tests for weights from a Bayesian perspective despite its potential usefulness.

and

$$V(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I},$$

where $\boldsymbol{Y}$ is the vector of values for a dependent variable, $\boldsymbol{X}$ is the matrix of values of the covariates (including a vector of ones in the first column) with $\boldsymbol{\beta}$ a vector of their corresponding coefficients, and $\boldsymbol{\varepsilon}$ is the vector of disturbances or errors. As Hausman (1978, p. 1,251) notes, we can replace the first assumption with $\text{plim}\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon} = 0$ in large samples. These tests originate from and are justified by Hausman (1978). He describes a general model misspecification test to detect omitted variables, incorrect functional form, and other model misspecifications. This test is based on the idea that under correct model specification, two different consistent estimators of the same parameters converge to the same parameter values as the sample size increases, but diverge when there is a misspecification.

Call the vector of regression coefficient estimates from the first estimator $\hat{\boldsymbol{\beta}}_1$ and the estimates of the same vector of coefficients from the second estimator $\hat{\boldsymbol{\beta}}_2$. In a correctly specified model the asymptotic expected value of $[\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2]$ should be zero, whereas in a misspecified model the asymptotic mean of $[\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2]$ need not be zero. The other assumptions underlying the Hausman (1978) test are (*a*) $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are consistent estimators of $\boldsymbol{\beta}$, (*b*) $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ have asymptotic normal distributions, and (*c*) $\hat{\boldsymbol{\beta}}_2$ is asymptotically efficient in that it attains the asymptotic Cramer-Rao bound. The null hypothesis is that of no misspecification.

We can test whether the asymptotic mean of $[\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2]$ is zero by having the asymptotic covariance matrix ($\boldsymbol{V}$) for the difference vector of $[\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2]$. The general form of the Hausman test statistic is $T_H = [\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2]'\hat{\boldsymbol{V}}^{-1}[\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2]$, where $\hat{\boldsymbol{V}}$ is an estimator of the variance of $[\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2]$. Hausman (1978) proves that under the null hypothesis of no misspecification and assumptions *a–c* above, the estimator of the asymptotic covariance matrix of $[\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2]$ is $\hat{\boldsymbol{V}}_H = [\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_1) - \hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_2)]$. The T_H asymptotically follows a chi-square distribution with degrees of freedom equal to the number of coefficients in $\hat{\boldsymbol{\beta}}_1$ (or $\hat{\boldsymbol{\beta}}_2$). The null hypothesis postulates that the model is correct and hence that each estimator converges in expectation to $\boldsymbol{\beta}$.

Hausman (1978) proposed this test for misspecifications in general. Pfeffermann (1993) proposed using the Hausman test to compare the coefficients from the weighted and unweighted regressions, so that $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_w$, the estimates from the weighted analysis (with lower asymptotic efficiency), and $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_u$, the estimates from the unweighted analysis (with higher asymptotic efficiency). Following Hausman's (1978) results, $\hat{\boldsymbol{V}}$ is set to $[\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_w) - \hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_u)]$, where $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_w)$ is the estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_w$ and $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_u)$ is the analogous quantity for $\hat{\boldsymbol{\beta}}_u$.

Asparouhov & Muthén (2007) use a Hausman test to compare weighted and unweighted estimates and even two different types of weighted estimates (e.g., one for trimmed and another for untrimmed weights). They also propose a modified estimate of $\hat{\boldsymbol{V}}$ that has superior finite sample properties. To see how Asparouhov & Muthén's (2007) application of the Hausman test differs from Pfeffermann's (1993), we consider the Hausman test statistic $T_H = [\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_u]'\hat{\boldsymbol{V}}^{-1}[\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_u]$. The middle term $\hat{\boldsymbol{V}}$ is an estimate of the asymptotic covariance matrix of the difference in the coefficients across the weighted and unweighted data, which are estimable in more than one way. The estimate used by Hausman (1978) is $\hat{\boldsymbol{V}}_H = [\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_1) - \hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_2)]$, where sample estimates of the asymptotic covariance matrices replace their population counterparts. Hausman's (1978) and Pfeffermann's (1993) use of this simple difference is justified by the asymptotic efficiency of the unweighted OLS estimator $\hat{\boldsymbol{\beta}}_2$ under the assumptions of their models. Indeed, Hausman's (1978, pp. 1253–54) proof that $\boldsymbol{V} = [V(\hat{\boldsymbol{\beta}}_1) - V(\hat{\boldsymbol{\beta}}_2)]$ depends on this assumption. If neither estimator is asymptotically efficient, we cannot rely on this simplification.

Table 1 Difference in coefficients tests for weighting

Model(s)	Test	Reference
$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ H_0: no misspecification $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2$ consistent $\hat{\boldsymbol{\beta}}_2$ asymptotically efficient	$T_H = (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)' \hat{\boldsymbol{V}}_H^{-1} (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)$ $\hat{\boldsymbol{V}}_H = [\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_1) - \hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_2)]$, $k = \dim(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)$ T_H asymptotically follows χ_k^2	Hausman (1978)
Hausman test with $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{\boldsymbol{w}}, \hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}$	Hausman test with $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{\boldsymbol{w}}, \hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}$	Pfeffermann (1993)
Hausman test with $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{\boldsymbol{w}_1}, \hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_{\boldsymbol{w}_2}$	Hausman test with $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{\boldsymbol{w}_1}, \hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_{\boldsymbol{w}_2}$ and $\hat{\boldsymbol{V}}_{AM} = [\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{w}_1}) - \hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{w}_2}) - 2\mathbf{C}]$, where $\mathbf{C} = \left(\frac{\partial^2 L_1(\hat{\boldsymbol{\beta}}_{\boldsymbol{w}_1})}{(\partial \boldsymbol{\beta})^2}\right)^{-1} M \left(\frac{\partial^2 L_2(\hat{\boldsymbol{\beta}}_{\boldsymbol{w}_2})}{(\partial \boldsymbol{\beta})^2}\right)^{-1'}$ $M = \sum_i w_{1i} w_{2i} \frac{\partial l_i(\hat{\boldsymbol{\beta}}_{\boldsymbol{w}_1})}{\partial \boldsymbol{\beta}} \left[\frac{\partial l_i(\hat{\boldsymbol{\beta}}_{\boldsymbol{w}_2})}{\partial \boldsymbol{\beta}}\right]'$	Asparouhov & Muthén (2007)

An alternative estimate for $\hat{\boldsymbol{V}}$ in the context of testing for weights comes from Asparouhov & Muthén (2007). They still use the DC test proposed by Hausman (1978) and applied to weights as in Pfeffermann (1993), but they suggest a different estimator of $\hat{\boldsymbol{V}}$. Specifically, they suggest $\hat{\boldsymbol{V}}_{AM} = [\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{w}_1}) - \hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{w}_2}) - 2\mathbf{C}]$, where $\mathbf{C}$ is the covariance matrix of the two estimators.[3] Their asymptotic covariance matrix allows for covariances between the two different estimators.

Table 1 summarizes the DC tests for weighting that we have reviewed. Because these tests of weights are based on asymptotic theory, it would be useful to know their performance in finite samples in the range of values that are typical in survey research. Unfortunately, there is scarce evidence on these test statistics in the context of testing for weighting. We are not aware of any analytic finite sample results. Asparouhov & Muthén (2007) conducted a small simulation study and found that the Type I error rates for the classic Hausman test described above were too large (i.e., rejected too frequently). Asparouhov & Muthén's (2007) modification performed better but still had some inaccuracies at small to moderate samples. Much more systematic study is required before drawing conclusions on how to best proceed with these tests.

In addition, the tests are implemented by comparing OLS (the unweighted model) with probability weighted least squares (PWLS; the weighted model), with the exception of Asparouhov & Muthén (2007), who use Skinner's (1989) pseudomaximum likelihood estimator. Finally, several authors (see, for example, Pfeffermann 1993) mention that one advantage of these tests is that they can compare subsets of coefficients (for example, slope coefficients only omitting the intercept) for differences. As we discuss below, targeted testing is helpful when one or two coefficients are of central interest and it would be practical and valuable to know which tests work best for such evaluations.

Researchers have suggested other less formal DC tests. For instance, Hahs-Vaughn & Lomax (2006) propose fitting models with and without weights and assessing whether the associated confidence intervals (CIs) overlap, concluding that the weights make a significant difference if the CIs do not overlap.[4] The authors base their argument on Schenker & Gentleman (2001), even

[3]Note that their derivation applies to structural equation models in general. We have adapted their formula to apply to regression coefficients in a multiple regression.

[4]The CI for the weighted estimates should take account of the heteroscedasticity that might be introduced by weighting.

Table 2 Weight association tests

Model(s)	Test	Reference(s)
$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{X_M\beta_M} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X_M}$ is a transformed version of $\boldsymbol{X}$	F-test of H_0: $\boldsymbol{\beta_M} = 0$	Hausman (1978)
$\boldsymbol{Y} = \boldsymbol{X\beta} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\boldsymbol{X}} = \boldsymbol{WX}$	F-test of H_0: $\tilde{\boldsymbol{\beta}} = 0$	DuMouchel & Duncan (1983) [Fuller (1984)]
$\boldsymbol{Y} = \boldsymbol{X\beta} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\boldsymbol{X}} = \begin{cases} \boldsymbol{WX} \\ \text{subset of } \boldsymbol{WX} \\ \text{or } \boldsymbol{W} \end{cases}$	F-test of H_0: $\tilde{\boldsymbol{\beta}} = 0$	Fuller (2009)
$\boldsymbol{Y} = \boldsymbol{X\beta} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\boldsymbol{X}} = \boldsymbol{QX}$ with $\boldsymbol{Q} = \text{diag}(q_1, q_2, \ldots, q_n)$ and $q_i = w_i b(\boldsymbol{x}_i) = w_i \hat{w}^{-1}(\boldsymbol{x}_i)$, with $\hat{w}(\boldsymbol{x}_i)$ from regression of w_i on $f(\boldsymbol{x}_i)$	F-test of H_0: $\tilde{\boldsymbol{\beta}} = 0$	Wu & Fuller (2005)
$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ 1. $\text{CORR}_s(\hat{\varepsilon}_i^k, W_i) = 0, k = 1, 2, 3$ 2. $\boldsymbol{W} = \boldsymbol{\alpha}_w + B_{W\hat{\boldsymbol{\varepsilon}}}\hat{\boldsymbol{\varepsilon}}^k + \boldsymbol{e}_W$	1. *a*. Fisher's Z transformation of H_0: $\text{CORR}_s(\varepsilon_i^k, w_i) = 0$ variance $\approx \frac{1}{(n-3)}$ *b*. Bootstrap standard deviation for Z-test of H_0: $\text{CORR}_s(\varepsilon_i^k, w_i) = 0$ 2. *t*-test of $\hat{B}_{W\hat{\boldsymbol{\varepsilon}}}$ for H_0: $B_{W\hat{\boldsymbol{\varepsilon}}} = 0$	Pfeffermann & Sverchkov (1999)
$\boldsymbol{W} = \boldsymbol{X\gamma}_1 + \gamma_2\boldsymbol{Y} + \boldsymbol{\delta_W}$	F-test of H_0: $\gamma_2 = 0$ normality of disturbance Note: test proposed for small area estimation	Pfeffermann & Sverchkov (2007)

though the latter suggest CI comparison should not be used when a test is available. Conceptually, comparing CIs in this way is an informal analog of the other DC tests in **Table 1**.

4.2. Weight Association Tests

Table 2 summarizes the WA tests in our review. Several of these tests take the form of a regression of the dependent variable on the unweighted and weighted covariates, while other WA tests assess the association of the weights $\boldsymbol{W}$ to $\boldsymbol{Y}$ conditional on the explanatory variables $\boldsymbol{X}$. The direct comparison of the coefficients from the weighted and unweighted regressions is not explicit in WA tests. Later we describe how some WA tests are more closely related to the DC tests than they first appear.

As with the tests in **Table 1**, we find a linkage between the Hausman (1978) test and several of the WA tests. Hausman (1978) suggested that another form of his misspecification test is to assess the statistical significance of $\boldsymbol{\beta_M}$ in the equation $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{X_M\beta_M} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X_M}$ is a suitably transformed version of $\boldsymbol{X}$. An F-test of $H_0 : \boldsymbol{\beta}_M = 0$ is a test of misspecification. In addition to the usual multiple regression assumptions, use of the F-test requires us to assume that $\boldsymbol{\varepsilon}$ comes from a normal distribution.

Although Hausman suggested this form for a variety of misspecifications, he did not consider it for tests of weighting. However, DuMouchel & Duncan (1983) and Fuller (1984) take this Hausman (1978) regression approach and apply it to the decision of whether to weight. In this context, consider the equation $\boldsymbol{Y} = \boldsymbol{X\beta}_u + \boldsymbol{X}_w\boldsymbol{\beta}_w + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y}$ is the vector of values for the

dependent variable, $\boldsymbol{X}$ is the matrix of unweighted values of the explanatory variables with $\boldsymbol{\beta}_u$ the corresponding coefficients, $\boldsymbol{X}_w$ is the matrix of the same explanatory variables but weighted with $\boldsymbol{\beta}_w$ their corresponding coefficients, and $\boldsymbol{\varepsilon}$ is the vector of errors. DuMouchel & Duncan (1983) recommend estimating this regression model with OLS and then applying an F-test of $H_0 : \boldsymbol{\beta}_w = 0$ to determine whether weights are needed. Rejection of this null hypothesis implies that weights are required, while failure to reject supports an unweighted analysis. Fuller (1984) makes a similar argument. We consider the F-test regression approaches of DuMouchel & Duncan (1983) and Fuller (1984) as WA tests that follow Hausman's (1978) alternative regression-based misspecification test, even though neither makes the link to Hausman (1978). Fuller (2009) suggests a variant on this approach in that he recommends the regression $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_u + \boldsymbol{W}\alpha + \boldsymbol{\varepsilon}$, where $\boldsymbol{W}$ is the weight variable, and testing whether its coefficient α is significantly different from zero. He also suggests testing just a subset of the variables in $\boldsymbol{WX}$ when interest lies in a few but not all of the covariates.

Wu & Fuller (2005) take the same WA test approach with $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\boldsymbol{X}} = \boldsymbol{QX}$ and $\boldsymbol{Q} = \text{diag}(q_1, q_2, \ldots, q_n)$. The q values are $q_i = w_i \hat{w}^{-1}(\boldsymbol{x}_i)$, where $\hat{w}(\boldsymbol{x}_i)$ are predicted values from the regression of w_i on $f(\boldsymbol{x}_i)$ and where $f(\boldsymbol{x}_i)$ is a function of the covariates determined by the researcher. In a sense, the part of the original weights that is predictable by the covariates is factored out of the original weight, and these unpredictable parts of weights transform $\boldsymbol{X}$. As with the previous regression models, an F-test of $H_0 : \tilde{\boldsymbol{\beta}} = 0$ determines the need for weights.

Clearly, the tests of DuMouchel & Duncan (1983), Fuller (1984, 2009), and Wu & Fuller (2005) are special cases of the Hausman (1978) regression test formulated as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}$. When they differ, it is in their definition of $\tilde{\boldsymbol{X}}$, where, for example, DuMouchel & Duncan (1983) use $\tilde{\boldsymbol{X}} = \boldsymbol{X}_w$ and Fuller (2009) uses a subset of the variables in $\boldsymbol{X}_w$ for $\tilde{\boldsymbol{X}}$. They make the usual regression assumptions of

$$\text{A. } E(\boldsymbol{\varepsilon}) = 0,$$

$$\text{B. } E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \boldsymbol{I},$$

$$\text{C. } \text{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{X}) = \text{Cov}(\boldsymbol{\varepsilon}, \tilde{\boldsymbol{X}}) = 0,$$

$$\text{D. } [\boldsymbol{X}\ \tilde{\boldsymbol{X}}] \text{ is nonsingular.}$$

To justify the application of the F-test of $H_0 : \tilde{\boldsymbol{\beta}} = 0$, they also assume

$$\text{E. } \boldsymbol{\varepsilon} \sim N(0, \sigma^2).$$

Fuller (2009) and others hold open the option of having the second set of covariates consist solely of the weight ($\boldsymbol{W}$) or a subset of $\boldsymbol{WX}$ rather than the full set. This strategy makes sense if the analyst hypothesizes that the weight alone or a subset of the transformed $\boldsymbol{WX}$ is what matters rather than the full set. If in reality these are the only variables that require weighting, then this subset strategy could increase the statistical power of the F-test and be defended on these grounds.

Other WA tests approach this issue differently than the Hausman (1978) regression method. Pfeffermann & Sverchkov (1999) derive relationships between sample residuals and weights. They begin with a distinction between the population and sample distributions of the errors in a regression. Their argument is that if the sample distribution of the residuals is the same as the population distribution of the errors, then one can ignore the sampling scheme and apply the unweighted regression (Pfeffermann & Sverchkov 1999, p. 171). Their test is grounded in the

idea that by comparing the moments of the sample and population residuals, one can compare whether their distributions are the same. They suggest that $E_p(\varepsilon_i^k) = E_s(\varepsilon_i^k)$, $k = 1, 2, \ldots$, where E_p and E_s are the expectations under the population and the sample probability density functions, respectively. Their paper also derives that $E_p(\varepsilon_i^k) = E_s(\varepsilon_i^k w_i)/E_s(w_i)$ and that an equivalent set of hypotheses is

$$\mathrm{CORR}_s(\varepsilon_i^k, w_i) = 0, \quad k = 1, 2, \ldots,$$

where CORR_s is the correlation under the sample distribution.

The first step in implementing this is the OLS regression of the dependent variable on the unweighted covariates. Then the residuals are formed as $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. These residuals have the linear effect of the covariates removed and the question is whether the remaining part is associated with the weights. On the basis of the preceding arguments, Pfeffermann & Sverchkov (1999) suggest tests of the null hypotheses that the correlations between the weights and the sample residuals, squared residuals, cubed residuals, etc., are zero. In other words, $\mathrm{CORR}_s(\varepsilon_i^k, w_i) = 0$, for $k = 1, 2, \ldots$. The null hypothesis is that all these correlations are zero, and when the null hypothesis is rejected, sample weights are informative. The authors evaluated the performance of these tests using OLS, PWLS, Skinner's pseudomaximum likelihood, parametric (maximum likelihood), and semiparametric estimators. To test the correlations, they both examined the Fisher's Z transformation of the correlations and used bootstrap resampling to develop estimates of the sampling variability of the correlations to use in the significance tests. They found the bootstrap method to work better in their simulations than the Fisher's Z transformation approach.

As an alternative to correlations, Pfeffermann & Sverchkov (1999) also examined regressing the weights on the residuals (see **Table 2**) and found that the two methods performed similarly. The approach using the squared residuals correlations performed very poorly, whereas the approach using the original residuals performed well. No explanation of these differences is provided, but the scope of the simulations was very limited.

The last WA test that we discuss comes from Pfeffermann & Sverchkov (2007). They suggest regressing the weight on the covariates and the dependent variable ($\boldsymbol{W} = \boldsymbol{X}\boldsymbol{\gamma}_1 + \gamma_2 \boldsymbol{Y} + \boldsymbol{\delta}_W$). A test of $H_0 : \gamma_2 = 0$ determines whether weights make a difference. We note that Pfeffermann & Sverchkov (2007) suggest this approach in the context of small area estimation and we present it as a more general test for weights.

In our review, we uncovered a few papers that cannot be neatly categorized as presenting DC or WA tests. Bertolet (2008) and Faiella (2010) cast the problem of weight ignorability as a design-based versus model-based issue. In addition, a few authors go into theoretical detail on the informativeness of the sampling weights but do not propose tests (e.g., Smith 1988, Sugden & Smith 1984). We refer readers to these papers for more details.

4.3. Testing Subsets of Coefficients

Thus far, our discussion has focused primarily on tests of whether all coefficients are equal in weighted versus unweighted analyses. Sometimes coefficients of particular variables in a regression are of special interest, whereas the coefficients of the other explanatory variables are of less concern. For instance, a study of discrimination would focus on the coefficient of the variable measuring minority status and might be less interested in the coefficients of the other explanatory variables. A global test of all coefficients might be statistically significant. This leads the researcher to a weighted analysis and the reduced efficiency that often accompanies using weights. However, it

is possible that the significant test statistic is due not to the coefficient for minority status, but to some other explanatory variable(s). In this situation, the efficiency of all coefficient estimates might be impacted when the coefficient estimates for minority status were essentially the same in the weighted and unweighted analyses.

Similarly, many researchers are not interested in differences of the intercepts between weighted and unweighted analyses and would like to test equality of all coefficients except the intercepts (e.g., Scott 2006). Under these circumstances it would be helpful to have a test to determine whether one coefficient or a subset of coefficients with and without weights is equal. If the key coefficients are the same, unweighted analysis with attention directed to these key variables could be performed more efficiently.[5]

Hausman's (1978) and Pfeffermann's (1993) DC tests would calculate $(\hat{\beta}_{w1} - \hat{\beta}_{u1})[\hat{V}(\hat{\beta}_{w1}) - \hat{V}(\hat{\beta}_{u1})]^{-1}(\hat{\beta}_{w1} - \hat{\beta}_{u1})$ and compare it to a chi-square distribution with one degree of freedom. A significant test statistic suggests that weighting is required, whereas an insignificant one suggests it is not. Among the WA tests, one could use the regression-based approaches that include the weighted and unweighted variables and test only those coefficients of greatest interest for the weighted variables (e.g., Winship & Radbill 1994). However, we have found little empirical or analytic evidence on which tests would perform the best when testing subsets of coefficients.

5. COMPARING DIFFERENCE IN COEFFICIENTS AND WEIGHT ASSOCIATION TESTS

5.1. Equivalence of the Tests

Beyond the diversity and different types of tests, there are other issues to consider. In this section we discuss comparisons of these tests and what is known from the literature about their finite sample properties.

An important question when determining the necessity of weights is whether a DC or WA test should be chosen. Should both be tested or are both approaches equivalent? As noted in Section 2, Pfeffermann & Sverchkov (2009) provide a key result here. They showed the equivalence of Equation 4 and Equation 5, which proves that the DC and WA tests are equivalent in that they essentially test the same thing; that is, testing the difference in weighted and unweighted regression coefficients is equivalent to testing whether, conditional on $\boldsymbol{X}$, y and w are correlated. That is not to say that a particular DC test will provide the same conclusion as a WA test. Indeed, even within the same category of test (i.e., DC or WA), results may differ; so the question of which test to use is not resolved by simply noting the equivalency of WA and DC tests.

The properties of DC tests based on $T_H = [\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_u]' \hat{\boldsymbol{V}}^{-1} [\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_u]$ versus WA tests formulated as regressions (e.g., $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_u + \boldsymbol{X}_w\boldsymbol{\beta}_w + \boldsymbol{\varepsilon}$, with a test on $\hat{\boldsymbol{\beta}}_w$) deserve attention. The DC tests, as described above (see Asparouhov & Muthén 2007, Hausman 1978, Pfeffermann 1993), are asymptotically distributed chi-square, whereas the regression misspecification form of the WA test (see DuMouchel & Duncan 1983, Fuller 1984, Hausman 1978) recommends the usual regression F-test for $H_0 : \boldsymbol{\beta}_w = 0$. Kott (1990) proposes a DC test for comparing OLS to PWLS regression that is chi-square distributed, but shows an alternative derivation using an F-test when the degrees of freedom are not large (echoing DuMouchel & Duncan 1983). Any power advantages of the DC

[5]Here, too, we need to consider the differences between statistical significance and practical significance and the manner in which a large sample increases the probability of detecting even small differences.

tests compared to the WA tests are unknown. Likewise, which forms of the tests have superior finite sample properties are also unknown.

Closer examination of these tests reveals a fundamental similarity between them. First, consider the DC test of $T_H = [\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_u]' \hat{\boldsymbol{V}}^{-1} [\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_u]$. If we write the asymptotic expectation as AE, the null hypothesis of this test is

$$H_0 : AE[\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_u] = 0,$$

or that the large sample mean of the difference (i.e., as both N and n approach infinity, holding n/N constant) in coefficients is zero for all coefficients.

Now consider the WA (regression) tests of **Table 2** [$\boldsymbol{Y} = \boldsymbol{X\beta} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}}$]. An F-test of $H_0 : \tilde{\boldsymbol{\beta}} = 0$ appears to differ from the test that determines whether the coefficients of the weighted and unweighted coefficients are equal. However, DuMouchel & Duncan (1983) prove that a test of $H_0 : \tilde{\boldsymbol{\beta}} = 0$ is the same as a test of $H_0 : AE[\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_u] = 0$ (see DuMouchel & Duncan 1983, p. 539). In other words, the preceding DC tests and the WA regression tests are tests of the same null hypothesis $H_0 : AE[\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_u] = 0$ of whether the coefficients differ from each other. The main difference is that the DC test statistic has a chi-square distribution and the WA test statistic has an F-distribution.

The differences in the distributions of the test statistics mean that the results of the tests will not always be the same. Furthermore, we do not know which of these test statistics has greater power. Finally, we do not know the frequency of their Type I errors. These issues would be valuable to pursue.

Pfeffermann & Sverchkov (1999, 2007) proposed WA tests different than those that we have so far compared to the DC tests. For instance, they recommend tests that relate the residual from an unweighted regression to the weights ($\boldsymbol{W}$) or regress $\boldsymbol{W}$ on $\boldsymbol{Y}$ and $\boldsymbol{X}$. The relationship of these tests to the others is less direct than the comparisons we just drew, but there are connections. Perhaps the closest relation comes from an examination of Fuller's (2009) suggestion to use the regression equation $\boldsymbol{Y} = \boldsymbol{X\beta}_u + \boldsymbol{W\alpha} + \boldsymbol{\varepsilon}$. If we subtract the explanatory variables and their coefficients from the left-hand side, we get $\boldsymbol{Y} = \boldsymbol{X\beta}_u + \boldsymbol{W\alpha} + \boldsymbol{\varepsilon}$. This new left-hand side is similar to Pfeffermann & Sverchkov's (1999) residual, and if we regress it on $\boldsymbol{W}$, we have a test similar to that in **Table 2**. The main difference is that Pfeffermann & Sverchkov (1999) suggest the OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}$ and then form the residual and regress it on $\boldsymbol{W}$. In Fuller (2009), $\boldsymbol{Y}$ is regressed on $\boldsymbol{X}$ and $\boldsymbol{W}$ at the same time. Both approaches assume that most interest lies in the $\boldsymbol{W}$ variable and not the other weighted covariates. In practice, researchers could be more concerned with whether other variables differ in the weighted and unweighted samples, so that they might want to include more weighted covariates for the test.

Returning to the contrast of the DC and WA tests, a practical difference between the DC and the WA (regression) tests also arises. As noted above, Hausman recommended that $\hat{\boldsymbol{V}}_H = [\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_w) - \hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_u)]$. In a sample it is possible for this difference in estimates of asymptotic covariance matrices to not be positive definite. This same computational issue is not present for the regression forms of the tests.

Fuller (1984), Winship & Radbill (1994), and Wu & Fuller (2005) used and evaluated the regression form of the WA test. Wu & Fuller (2005) used the DuMouchel & Duncan (1983)/Fuller (1984) regression tests of weighting to decide whether weights were needed. In a simulation they then compared multiple estimators, including the general least squares approach of Pfeffermann & Sverchkov (1999), which was informed by the WA test the latter derived, concluding that when the correlation between the weights and the error term was low, OLS performed better, but when the correlation was high, the Pfeffermann & Sverchkov (1999) estimator performed better.

However, Wu & Fuller (2005) did not evaluate the performance of their DC test in the simulation or compare it with any of the WA tests.

5.2. Studies of the Statistical Properties of the Tests

The validity of the two major types of tests is mostly justified asymptotically. The classic Hausman DC test statistic of differences in coefficients, for instance, follows a chi-square distribution in large samples. Asparouhov & Muthén (2007) find that in their small simulation study, this test, which they refer to as the Hausman-Pfeffermann test, rejects too frequently when the null hypothesis of "no weights needed" is true. Tests based on regressing the sample residual on the weights assume that the sample residuals and sample weights come from samples large enough to be good approximations to their population counterparts. We have found only limited analytic material on the finite sample properties of these tests that could guide empirical research.

Even the Monte Carlo simulation research on the finite sample properties of these tests is quite limited and largely designed to illustrate a new test with a pilot simulation analysis to demonstrate its potential. Among all the literature we reviewed, we found five simulation studies on tests of weight informativeness. They typically appear as appendages to estimation-focused studies. Two studies focused only on DC tests (Asparouhov & Muthén 2007, Pfeffermann & Sverchkov 2003), two focused only on WA tests (Eideh & Nathan 2006, Pfeffermann & Sverchkov 1999), and one compared the two types of tests to each other (Chambers et al. 2003).

Four of these five studies focused on bivariate regression models, though Asparouhov & Muthén (2007) also explored tests of weights in the context of factor analysis. Eideh & Nathan (2006) explored weights in the context of short time series (3 or 10 repeated observations). Asparouhov & Muthén (2007) looked at sample sizes ranging from 200 to 10,000 taken from an infinite population, while Pfeffermann & Sverchkov (1999) simulated finite populations of 1,000 and 3,000 and then drew samples of 100 and 300 from the simulated populations. Pfeffermann & Sverchkov (2003) simulated a finite sample of 3,000 and drew from it samples of 300. Eideh & Nathan (2006) used a sample size of 500. Chambers et al. (2003) simulated a finite sample of 1,000 and drew samples of 100 from it.

Methods of simulating informative sampling designs are diverse. Asparouhov & Muthén (2007) sampled data as exponential functions of Y for bivariate regression and as exponential functions of the latent variable for the factor model. Pfeffermann & Sverchkov (1999) sampled data as exponential and polynomial functions of Y. Pfeffermann & Sverchkov (2003) sampled data as a Poisson function of Y as well as stratified sampling. Eideh & Nathan (2006) sampled cases as exponential and linear functions of Y. Chambers et al. (2003) sampled data proportional to Z where Y was a function of the product of Y and Z.

Pfeffermann & Sverchkov (1999) proposed a correlation test based on moments of the correlation between the predictor and residuals and reported near-nominal detection of informative weights for first-order moments, though higher-order moments did not perform well. Asparouhov & Muthén (2007) modified Pfeffermann's (1993) Hausman test by changing the definition of the variance to address small sample sizes. The modified and original tests performed as expected for large sample sizes. For small sample sizes, both tests incorrectly rejected the null hypothesis more frequently than the nominal 5% rate, though the modified test performed better. Eideh & Nathan (2006) reported that their use of a t-statistic to implement a correlation test leads to near-nominal expected performance for simulated time series data. Chambers et al. (2003) found that the correlation test of Pfeffermann & Sverchkov (1999) performed well for both homogenous and heterogeneous error conditions, whereas a version of the Hausman test underperformed when weights were informative under heterogeneous data. Pfeffermann & Sverchkov (2003) developed

a Hotelling test generalized to the Hausman test and report that when weights were informative, the null hypothesis of noninformative weights was rejected at the 1% level, though the target alpha was 0.05.

This state of affairs is unsatisfactory. Researchers do not have evidence of how well the tests perform in small to moderate samples nor do they know how large a sample must be to safely rely on the asymptotic properties. Ideally, analytic results that inform us of the test properties under different conditions would be available, but they are not. Simulation studies under a variety of design conditions would be helpful, but even simulation work on these different tests is sparse.

6. CONCLUSIONS

At a time when most surveys have unequal probabilities of selection either by design or by other practical constraints, the question of whether to weight variables during the analysis takes on added importance. If weighting data were a cost-free option, then always weighting would be a reasonable strategy. But unnecessarily weighting means lower efficiency and lower statistical power. Tests that determine whether weights are required do exist, but they are rarely applied for several reasons. One is the lack of awareness among researchers. Another is the influence of tradition in different fields—some always weight and others never do. An additional reason is that some of these tests are not readily available in software packages. Furthermore, even when these tests are easy to implement, there is little guidance on which of the many tests to choose.

Though our focus is on the most widely studied case of weighting for linear regression models, a few papers concentrate on using weights in generalized linear models (e.g., Binder 1983, Chambless & Boyle 1985, Morel 1989, Roberts et al. 1987) and the software to implement (e.g., An 2008). However, these papers do not provide any new information on weighting tests for the models. In addition, as previously noted, most unequal probability samples are also clustered, yet for the most part the extant literature has not considered the implications of clustering on the weighting decision. This is another area where more research is needed.

This review accomplishes our aim to classify tests into two major groups: DC tests and WA tests. We also find that these two groups are more closely related than at first appearance. This makes the problem more manageable and highlights the close relation among seemingly different tests, but numerous questions remain. Which of these tests has the best finite sample properties? Although they are theoretically equivalent, are DC and WA tests interchangeable in any given situation or is one favored over the other in certain contexts? Further, what should a researcher do if one test is passed and another is failed? When a researcher's interest focuses on just one coefficient and she wants to determine whether weighting makes a difference, which testing approach makes the most sense? Finally, which of these tests requires custom programming and which can a researcher obtain by tricking existing software into providing the necessary results? Many fields of research would benefit from answers to these questions.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

K.A. Bollen gratefully acknowledges RTI International and their CoDA Visiting Scholar program for the support they provided in preparing this research and paper. All authors appreciate the helpful comments of Stephen Fienberg.

LITERATURE CITED

An AB. 2008. Performing logistic regression on survey data with the new SURVEYLOGISTIC procedure. In *Proceedings of the Statistics and Data Analysis Section, SUGI, 27th*, Orlando, FL. Cary, NC: SAS. **http://www2.sas.com/proceedings/sugi27/p258-27.pdf**

Asparouhov T, Muthén B. 2007. Testing for informative weights and weights trimming in multivariate modeling with survey data. *Proc. Jt. Stat. Meet., Surv. Res. Methods Sect.*, Salt Lake City, UT, July 29–Aug. 2, pp. 3394–99. Alexandria, VA: Am. Stat. Assoc. **http://www.amstat.org/sections/srms/Proceedings/y2007/Files/JSM2007-000745.pdf**

Bertolet M. 2008. *To weight or not to weight? Incorporating sampling designs into model-based analyses*. PhD thesis, Carnegie Mellon Univ.

Biemer PP, Christ SL. 2008. Weighting survey data. In *International Handbook of Survey Methodology*, ed. ED de Leeuw, JJ Hox, DA Dillman, pp. 317–41. London: Routledge

Binder DA. 1983. On the variance of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* 51:279–91

Chambers RL, Dorfman AH, Sverchkov MY. 2003. Nonparametric regression with complex survey data. In *Analysis of Survey Data*, ed. RL Chambers, CJ Skinner, pp. 151–74. Chichester, UK: Wiley

Chambless LE, Boyle KE. 1985. Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Commun. Stat. Theory Methods* 14(6):1377–92

Cox LH, Karr AF, Kinney SK. 2011. Risk-utility paradigms for statistical disclosure limitation: how to think, but not how to act (with discussion). *Int. Stat. Rev.* 79(2):160–99

DuMouchel WH, Duncan GJ. 1983. Using sample survey weights in multiple regression analysis. *J. Am. Stat. Assoc.* 78(383):535–43

Efron B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7(1):1–26

Eideh AAH, Nathan G. 2006. Fitting time series models for longitudinal survey data under informative sampling. *J. Stat. Plann. Inference* 136:3052–69

Faiella I. 2010. *The use of survey weights in regression analysis*. Work. Pap. 739, Bank of Italy, Rome

Fienberg SE. 2009. The relevance or irrelevance of weights for confidentiality and statistical analyses. *J. Priv. Confid.* 1(2):183–95

Fienberg SE. 2011. Bayesian models and methods in public policy and government settings. *Stat. Sci.* 26(2):212–26

Fisher RA. 1915. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika* 10(4):507–21

Fisher RA. 1921. On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron* 1:3–32

Fuller WA. 1984. Least squares and related analyses for complex survey designs. *Surv. Methodol.* 10(1):97–118

Fuller WA. 2009. *Sampling Statistics*. Hoboken, NJ: Wiley

Hahs-Vaughn DL, Lomax RG. 2006. Utilization of sample weights in single-level structural equation modeling. *J. Exp. Educ.* 74(2):161–90

Hausman JA. 1978. Specification tests in econometrics. *Econometrica* 46(6):1251–71

Holt D, Smith TMF, Winter PD. 1980. Regression analysis of data from complex surveys. *J. R. Stat. Soc. A* 143(4):474–87

Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47:663–85

Kalsbeek WD, Agans RP. 2007. Sampling and weighting in household telephone surveys. In *Advances in Telephone Survey Methodology*, ed. JM Lepkowski, C Tucker, JM Brick, ED de Leeuw, L Japec, et al. Hoboken, NJ: Wiley

Klein LR, Morgan JN. 1951. Results of alternative statistical treatments of sample survey data. *J. Am. Stat. Assoc.* 46:442–60

Kott PS. 1990. What does performing a regression on survey data mean? *Proc. Jt. Stat. Meet., Surv. Res. Methods Sect.*, Anaheim, CA, Aug. 6–9, pp. 337–41. Alexandria, VA: Am. Stat. Assoc. **http://www.amstat.org/sections/srms/Proceedings/papers/1990_053.pdf**

Little R. 2009. *Weighting and prediction in sample surveys*. Work. Pap. 81, Dep. Biostat., Univ. Michigan. **http://www.bepress.com/umichbiostat/paper81**

Lohr SL. 2010. *Sampling: Design and Analysis.* Pacific Grove, CA: Duxbury. 2nd ed.

Morel G. 1989. Logistic regression under complex survey designs. *Surv. Methodol.* 15:203–23

Pfeffermann D. 1993. The role of sampling weights when modeling survey data. *Int. Stat. Rev.* 61(2):317–37

Pfeffermann D. 1996. The use of sampling weights for survey data analysis. *Stat. Methods Med. Res.* 5:239–61

Pfeffermann D, Sverchkov M. 1999. Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā Indian J. Stat. B* 61:166–86

Pfeffermann D, Sverchkov M. 2003. Fitting generalized linear models under informative sampling. In *Analysis of Survey Data*, ed. RL Chambers, CJ Skinner, pp. 175–94. Chichester, UK: Wiley. doi: 10.1002/0470867205.ch12

Pfeffermann D, Sverchkov M. 2007. Small area estimation under informative probability sampling of areas and within the selected areas. *J. Am. Stat. Assoc.* 102(480):1427–38

Pfeffermann D, Sverchkov M. 2009. Inference under informative sampling. In *Handbook of Statistics*, Vol. 29, Pt. B: *Sample Surveys: Inference and Analysis*, ed. D Pfeffermann, CR Rao, pp. 455–87. Amsterdam: Elsevier

Roberts G, Rao JNK, Kumar S. 1987. Logistic regression analysis of sample survey data. *Biometrika* 74(1):1–12

Schenker N, Gentleman JF. 2001. On judging the significance of differences by examining the overlap between confidence intervals. *Am. Stat.* 55(3):182–86

Scott A. 2006. Population-based case control studies. *Surv. Methodol.* 32(2):123–32

Skinner CJ. 1989. Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, ed. CJ Skinner, D Holt, TMF Smith, pp. 59–87. New York: Wiley

Smith TMF. 1988. To weight or not to weight, that is the question. *Bayesian Stat.* 3:437–51

Smith TMF, Sugden RA. 1988. Sampling and assignment mechanisms in experiments, surveys and observational studies. *Rev. Int. Stat.* 56(2):165–80

Sterba SK. 2009. Alternative model-based and design-based frameworks for inference from samples to populations: from polarization to integration. *Multivar. Behav. Res.* 44(6):711–40

Sugden RA, Smith TMF. 1984. Ignorable and informative designs in survey sampling inference. *Biometrika* 71(3):495–506

Valliant R, Dever JA, Kreuter F. 2013. *Practical Tools for Designing and Weighting Sample Surveys*. New York: Springer-Verlag

Winship C, Radbill L. 1994. Sampling weights and regression analysis. *Sociol. Methods Res.* 23(2):230–57

Wu Y, Fuller W. 2005. Preliminary testing procedures for regression with survey samples. *Proc. Jt. Stat. Meet., Surv. Res. Methods Sect.*, Minneapolis, MN, Aug. 7–11, pp. 3683–88. Alexandria, VA: Am. Stat. Assoc. **http://www.amstat.org/sections/srms/Proceedings/y2005/Files/JSM2005-000099.pdf**

Annual Review of Statistics and Its Application

Volume 3, 2016

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics