ARE SURVEY WEIGHTS NEEDED?
A REVIEW OF DIAGNOSTIC TESTS IN REGRESSION ANALYSIS

Kenneth A. Bollen
University of North Carolina at Chapel Hill

Paul Biemer
Alan Karr
Stephen Tueller
Marcus Berzofsky
RTI International

ABSTRACT

When analyzing survey data the decision to weight variables or not can have serious implications. Sample selection probabilities might be unequal by design or as a result of frame errors, nonresponse or other data collection problems. If we do not weight when appropriate, we run the risk of having biased coefficient estimates and poor inferences. Alternatively, when we unnecessarily apply weights, we can create an inefficient estimator with reduced statistical power and no gain in accuracy. Yet in practice it is rare to see researchers testing whether weighting is necessary. One reason for this is that researchers are sometimes guided more by the current practice in their field than by scientific evidence. Another reason is that the statistical tests for weighting are neither widely known nor widely available in major statistical software packages. This is further complicated by the broad array of tests with little guidance on which has optimal properties under which conditions.

This paper reviews a wide variety of empirical tests to determine whether weighted analyses are justified. Our focus is on regression models, though the implications of our review extend beyond regression. Even with this focus, there are a half-dozen or more diagnostic tests that are a source of confusion for analysts. Our review demonstrates that nearly all such tests are classifiable into two categories that we refer to as Hausman-type Tests and Weight Association Tests. We describe the distinguishing features of each category, present the known properties of the tests, and how they are related to each other. We also review the limited simulation evidence that has investigated the sampling properties of these tests in finite samples. Finally, we highlight the unanswered theoretical and practical questions that surround these tests and that deserve further statistical research.
INTRODUCTION

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

Survey weights are 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 statistical models of the relationships between explanatory and dependent variables has been the topic of debate in the literature dating back to the early 1950s (see, e.g., Klein & Morgan, 1951; Holt, Smith, & Winter, 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 in privacy concerns than in statistical considerations. Some have argued that weights might provide sufficient information to permit deductive disclosure of respondents (e.g., Feinberg, 2??). In this situation, researchers might want to determine whether the omission of weights to maintain confidentiality leads to bias coefficient estimates.

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

The focus of our paper is to review and to compare empirical scientific tests of whether to use weights. Our paper has several purposes. One is to review the major diagnostic tests of the necessity for weighting. Second, we examine the assumptions and properties of these tests as documented in the literature. Third, we develop a framework to unite this seemingly diverse set of diagnostic tests under many fewer categories. Finally, we identify gaps in the literature where additional research is required. Furthermore we conduct this review in the context of regression analysis; 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. It also lays the foundation for the study of other types of models that are closely related to or are generalizations of
regression. To simplify the presentation we first present the diagnostic tests ignoring clustering of data and then provide a separate section to describe modifications to diagnostic tests that are required to take account of clusters that are common in complex samples.

The next section reviews and classifies many of the major tests of whether to weight. Following this is a section on other issues surrounding these tests including a comparison of what is being assessed with each type of test and a discussion of which of these tests apply to subsets of coefficients. Modifications to these tests that permit us to take account of clusters are covered in the next section. We end this review with our conclusion on what research questions remain for these diagnostic tests to increase their practical usefulness.

TESTS OF WHETHER TO WEIGHT

Suppose that a researcher decides to test whether weights are needed in a regression analysis. A search of the literature reveals a variety of diagnostic tests. Using Pfeffermann & Sverchkov (1999) this researcher might form the residuals from the Ordinary Least Squares (OLS) regression of the dependent variable on the original covariates and then correlate these residuals with weights; and correlate the square and cube of these residuals with the weights. A Fisher's Z transformation (Fisher, 1915; 1921) or a bootstrap estimate (Efron, 1979) of the standard deviations of these correlations would enable significance tests of whether these correlations are zero in the population and hence whether weights are required. Or using the same article, the analyst might decide to use an OLS regression of these residuals on the weight variable (W) and to perform a t-test of whether the coefficient of W is statistically significantly different from zero. Another researcher might turn to DuMouchel & Duncan (1983) and use an OLS regression of the dependent variable on the original covariates and on the weighted
covariates. An F-test of whether all coefficients of the weighted covariates are zero would be a test of whether weights are required. Finally, a third researcher might turn to a Hausman (1978) test as described by Pfefferman (1993) where the differences in the coefficients of the weighted and unweighted estimates pre- and post-multiply an estimate of the inverse of the covariance matrix of these coefficient differences to form a chi-square test statistic.

These are just a few of the diagnostic tests that the literature provides to test 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 diagnostics test the same or different hypotheses? What assumptions are required for each of the tests? Should we expect similar or different results 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 are able to simplify the matter by demonstrating the relationships between most of these tests and clarifying the assumptions that underlie them. To begin with we classify nearly all of these tests into two types: Difference in Coefficient Tests and Weight Association Tests. Later in the paper we also will show a tight connection between most of the tests in these two groups. But first we give an overview of these two types of tests.

**Difference in Coefficients Tests**

The first group of tests have in common that they compare the coefficients of the weighted and unweighted analyses and provide an assessment of whether these differences are statistically significantly different from zero. Start with the regression equation,

\[ Y = X\beta + \epsilon \]
assuming
\[ E(\epsilon|X) = 0, \]
\[ V(\epsilon|X) = \sigma^2 I. \]
where \( Y \) is the vector of values for a dependent variable, \( X \) is the matrix of values of the covariates (including a vector of ones in the first column) with \( \beta \) a vector of their corresponding coefficients, and \( \epsilon \) the vector of disturbances or errors.\(^1\) These tests have their origin and justification in Hausman (1978). His paper described a general model misspecification test to detect omitted variables, incorrect functional form, and other model misspecifications. It 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, \( \hat{\beta}_1 \), and the estimates of the same vector of coefficients from the second estimator, \( \hat{\beta}_2 \). In a correctly specified model, the asymptotic expected value of \( [\hat{\beta}_1 - \hat{\beta}_2] \) should be zero, while in a misspecified model the asymptotic mean of \( [\hat{\beta}_1 - \hat{\beta}_2] \) need not be zero. The other assumptions underlying the Hausman (1978) test are: (1) \( \hat{\beta}_1, \hat{\beta}_2 \) are consistent estimators, (2) \( \hat{\beta}_1, \hat{\beta}_2 \) each have asymptotic normal distributions, and (3) \( \hat{\beta}_2 \) 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 \( [\hat{\beta}_1 - \hat{\beta}_2] \) is zero by having the asymptotic covariance matrix \( (V) \) for the difference vector of \( [\hat{\beta}_1 - \hat{\beta}_2] \). The general form of the Hausman test statistic is: \( T_H = [\hat{\beta}_1 - \hat{\beta}_2]' \hat{V}^{-1}[\hat{\beta}_1 - \hat{\beta}_2] \), where \( \hat{V} \) is an estimator of the variance of

\(^1\) As Hausman (1978, p. 1251) notes, we can replace the first assumption with \( \text{plim} \frac{1}{n} X' \epsilon = 0 \) in large samples.
$[\hat{\beta}_1 - \hat{\beta}_2]$. Hausman (1978) proves that under the assumptions of his test the estimator of the asymptotic covariance matrix of $[\hat{\beta}_1 - \hat{\beta}_2]$ is $\tilde{V} = [\tilde{V}(\hat{\beta}_1) - \tilde{V}(\hat{\beta}_2)]$. The $T_H$ asymptotically follows a chi-square distribution with degrees of freedom equal to the number of coefficients in $\hat{\beta}_1$ (or $\hat{\beta}_2$). The null hypothesis postulates that the model is correct and hence that the estimators each converge to $\beta$.

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

Asparouhov & Muthén (2007) use a Hausman test for comparing weighted and unweighted estimates, and even two different types of weighted estimates. They also propose a modified estimate of $\tilde{V}$ that they suggest 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 $T_H = [\hat{\beta}_w - \hat{\beta}_u] \tilde{V}^{-1} [\hat{\beta}_w - \hat{\beta}_u]$. The middle term $\tilde{V}$ is an estimate of the asymptotic covariance matrix of the difference in the coefficients across the weighted and unweighted data, which is estimable in more than one way. The estimate used by Hausman (1978) is $\tilde{V} = [\tilde{V}(\hat{\beta}_1) - \tilde{V}(\hat{\beta}_2)]$, where sample estimates of these asymptotic covariance matrices replace their population counterparts. Hausman (1978) and Pfeffermann (1993) use of this simple difference is justified by the asymptotic efficiency of the unweighted OLS estimator $\hat{\beta}_2$ under the assumptions of their models. Indeed Hausman’s (1978, pages 1253-54) proof that $\tilde{V} =
\[ V(\hat{\beta}_1) - V(\hat{\beta}_2) \] depends on this assumption. If neither estimator is asymptotically efficient, we cannot rely on this simplification.

An alternative estimate for \( \hat{V} \) in the context of testing for weights comes from Asparouhov & Muthén (2007). They still use the difference in coefficient test proposed by Hausman (1978) and applied to weights as in Pfeffermann (1993), but they suggest a different estimator of \( \hat{V} \).

Specifically, they suggest \( \hat{V} = [\hat{V}(\hat{\beta}_{w_1}) - \hat{V}(\hat{\beta}_{w_2}) - 2C] \) where \( C \) is the covariance matrix of the two estimators. Their asymptotic covariance matrix allows for covariances between the two different estimators.

Table 1 summarizes these Difference in Coefficient 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. They found that the Type I errors 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 the best way to proceed with these tests.

In addition, the tests are implemented comparing OLS (the unweighted model) with probability weighted least squares (the weighted model) with the exception of Asparouhov & Muthén (2007), who use Skinner’s (1989) pseudo-maximum likelihood estimator. Finally, several authors who have presented these tests mention that an advantage of them is that it is possible to compare subsets of coefficients for differences rather than just perform a global test.
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 the best for such evaluations.

(Note the C in the Asparouhov & Muthen (2007) test needs modification to correspond to betas only. Check what I have done below.)

Table 1 Coefficient Difference Tests for Weighting

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>MODEL(S)</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausman (1978)</td>
<td>$Y = X\beta + \varepsilon$ $H_0$: No misspecification $\hat{\beta}_1, \hat{\beta}_2$ consistent $\hat{\beta}_2$ asym. efficient</td>
<td>$T_H = (\hat{\beta}_1 - \hat{\beta}_2)'\hat{\vartheta}^{-1}(\hat{\beta}_1 - \hat{\beta}_2)$ $T_H$ asymp. $\chi_k^2$ $\hat{\vartheta} = [\hat{\vartheta} (\hat{\beta}_1) - \hat{\vartheta} (\hat{\beta}_2)]$, $k = \dim(\beta_1 - \beta_2)$</td>
</tr>
<tr>
<td>Pfeffermann (1993)</td>
<td>Hausman Test with $\hat{\beta}_1 = \hat{\beta}_w, \hat{\beta}_2 = \hat{\beta}_u$</td>
<td>Hausman Test with $\hat{\beta}_1 = \hat{\beta}_w, \hat{\beta}_2 = \hat{\beta}_u$</td>
</tr>
<tr>
<td>Asparouhov &amp; Muthen (2007)</td>
<td>Hausman Test with $\hat{\beta}<em>1 = \hat{\beta}</em>{w_1}, \hat{\beta}<em>2 = \hat{\beta}</em>{w_2}$</td>
<td>Hausman Test with $\hat{\beta}<em>1 = \hat{\beta}</em>{w_1}, \hat{\beta}<em>2 = \hat{\beta}</em>{w_2}$ and $\hat{\vartheta} = [\hat{\vartheta} (\hat{\beta}<em>{w_1}) - \hat{\vartheta} (\hat{\beta}</em>{w_2}) - 2C]$ where $C = \left(\frac{\partial^2 L_1(\hat{\beta}<em>w)}{\partial \beta} \right)^{-1} M \left(\frac{\partial^2 L_2(\hat{\beta}<em>w)}{\partial \beta} \right)^{\chi_1}$ $M = \sum_i w</em>{1i} w</em>{2i} \frac{\partial l_i(\hat{\beta}<em>{w_1})}{\partial \beta} \left[\frac{\partial l_i(\hat{\beta}</em>{w_1})}{\partial \beta}\right]^{\chi_1}$</td>
</tr>
</tbody>
</table>
## Table 2 Weight Association Tests

<table>
<thead>
<tr>
<th>Source</th>
<th>Equation</th>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausman (1978)</td>
<td>$Y = X\beta + \tilde{X}\tilde{\beta} + \tilde{\epsilon}$ where $\tilde{X}$ transformed version of $X$</td>
<td>$F$ Test of $H_0: \tilde{\beta} = 0$</td>
</tr>
<tr>
<td>DuMouchel &amp; Duncan (1983) [Fuller (1984)]</td>
<td>$Y = X\beta + \tilde{X}\tilde{\beta} + \tilde{\epsilon}$ where $\tilde{X} = WX$</td>
<td>$F$ Test of $H_0: \tilde{\beta} = 0$</td>
</tr>
</tbody>
</table>
| Fuller (2009) | $Y = X\beta + \tilde{X}\tilde{\beta} + \tilde{\epsilon}$ where $\tilde{X} = \left\{ \begin{array}{ll} WX \\
subset of WX \\
or WX \end{array} \right.$ | $F$ Test of $H_0: \tilde{\beta} = 0$ |
| Wu & Fuller (2005) | $Y = X\beta + \tilde{X}\tilde{\beta} + \tilde{\epsilon}$ where $\tilde{X} = QX$ with $Q = diag(q_1, q_2, \ldots, q_n)$ and $q_i = w_i h(x_i) = w_i \hat{w}^{-1}(x_i)$ with $\hat{w}(x_i)$ from regression of $w_i$ on $f(x_i)$ | $F$ Test of $H_0: \tilde{\beta} = 0$ |
| Pfeffermann & Sverchkov (1999) | $\hat{\epsilon} = Y - X\hat{\beta}$ 1) CORR($\hat{\epsilon}_k^k, W_i$) $k = 1, 2, 3$ 2) $W = \alpha_W + B_{W\hat{\epsilon}}^k + \epsilon_W$ | 1) a) Fisher’s transformation, Z-test of $H_0: \text{CORR}(\hat{\epsilon}_1^k, W_i) = 0$ with variance approximately $1/(N-3)$ 1) b) Bootstrap s.d. for Z-test of $H_0: \text{CORR}(\hat{\epsilon}_1^k, W_i) = 0$ 2) t-test of $B_{W\hat{\epsilon}}$ for $H_0: B_{W\hat{\epsilon}} = 0$ |
| Pfeffermann & Sverchkov (2007) | $W = XY_1 + \gamma_2 Y + \delta_W$ | $F$ Test of $H_0: \gamma_2 = 0$ normality of disturbance Note: test proposed for small area estimation. |

### Weight Association Tests

Table 2 summarizes the weight association tests. Several of these take the form of a regression of the dependent variable on the raw and transformed covariates while the other tests assess the association of the weights $W$ to $Y$ conditional on the explanatory variables $X$. The direct comparison of the coefficients from the weighted and unweighted regressions is not explicit in the tests. Though later we describe how some of these tests are more closely tied to the differences in coefficients than they first appear.
As with the tests in Table 1, we find a linkage between Hausman (1978) and several of these Weight Association Tests. Hausman (1978) suggested that another form of his misspecification test is to assess the statistical significance of $\beta_M$ in the equation

$$Y = X\beta + X_M\beta_M + \varepsilon$$

where $X_M$ is a suitably transformed version of $X$. An F-test of $Ho: \beta_M = 0$ is a test of misspecification. In addition to the usual multiple regression assumptions, use of the F-test requires that we assume that $\varepsilon$ comes from a normal distribution.

Though Hausman suggested this 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 problem of whether to weight or not. In this context, consider the equation $Y = X\beta_u + X_w\beta_w + \varepsilon$, where $Y$ is the vector of values for the dependent variable, $X$ is the matrix of unweighted values of the explanatory variables with $\beta_u$ the corresponding coefficients, $X_w$ is the matrix of the same explanatory variables but weighted with $\beta_w$ their corresponding coefficients, and $\varepsilon$ is the vector of errors. DuMouchel & Duncan (1983) recommend estimating this regression model with OLS and then applying an F-test of $Ho: \beta_w = 0$ 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 DuMouchel & Duncan (1983) and Fuller (1984) F-test regression approaches as Weight Association Tests that follow Hausman’s (1978) alternative regression-based misspecification test, even though neither make the link to Hausman (1978). Fuller (2009) suggests a variant on this approach in that he recommends the regression $Y = X\beta_u + W\alpha + \varepsilon$, where $W$ is the weight variable and that we test its coefficient $\alpha$ to determine whether it is significantly different from zero. He also suggests
testing just a subset of the variables in $WX$ when interest lies in a few but not all of the covariates.

Wu & Fuller (2005) take the same Weight Association Regression Test approach with $Y = X\beta + \bar{X}\bar{\beta} + \bar{\varepsilon}$ where $\bar{X} = QX$ and $Q = \text{diag}(q_1, q_2, \cdots, q_n)$. The $qs$ are $q_i = w_i h(x_i) = w_i \hat{w}^{-1}(x_i)$ with $\hat{w}(x_i)$ from the regression of $w_i$ on $f(x_i)$ where $f(x_i)$ is some function of the covariates. In a sense, the original weights are adjusted by that part of the weights that are predictable by the covariates and these adjusted weights are what transform the $X$. As with the previous regression models, an $F$-test of $H_0: \bar{\beta} = 0$ determine the need for weights.

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

$E(\varepsilon) = 0$

$E(\varepsilon\varepsilon') = \sigma^2 I$

$\text{COV}(\varepsilon, X) = \text{COV}(\varepsilon, \bar{X}) = 0$

No perfect collinearity for $[X \bar{X}]$

$\varepsilon \sim N(0, \sigma^2)$

where the last normality assumption justifies the $F$-test of $H_0: \bar{\beta} = 0$.

Other Weight Association Tests approach this issue differently than this. Pfeffermann & Sverchkov (1999) derived relationships between sample residuals and weights. The first step is
the OLS regression of the dependent variable on the raw covariates. Then the residuals are formed as \( \hat{\varepsilon} = Y - X\hat{\beta} \). These residuals have the linear effect of the covariates removed and the question is whether the remaining part is associated with the weights. They suggest tests of the null hypotheses that the correlations between the weights and the sample residuals, squared residuals, cubed residuals, etc., are zero; i.e., \( \text{CORR}(\varepsilon_i^k, w_l) = 0 \) for \( k = 1, 2, \ldots \). The null hypothesis is that the correlation is zero, and when the alternative hypothesis is retained, sample weights are informative. They evaluated the performance of these tests using five estimators, OLS, probability weighted least squares (PWLS), Skinner’s pseudo-maximum likelihood (PML), parametric (maximum likelihood, or ML), and semiparametric estimators. To test the correlations they examined both the Fisher’s Z transformation of the correlations and using the bootstrap 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), but the two methods performed similarly. The approach using the squared residuals correlations performed very poorly, while the approach using the original residuals performed well. No explanation of these differences is provided, but we need to keep in mind that the scope of the simulations was very limited.

The last Weight Association Test that we discuss comes from Pfeffermann & Sverchkov (2007). They suggest regressing the weight on the covariates and the dependent variable \( (W = X\gamma_1 + \gamma_2 Y + \delta_w) \). A test of \( H_0: \gamma_2 = 0 \) is a test of whether weights make a difference. We note that Pfeffermann & Sverchkov (2007) suggest this approach in the context of small area estimation and we are presenting it as a more general test for weights.
Other Tests of Weight Informativeness

We describe these other tests for completeness. Hahs-Vaughn and 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. The authors base their argument on Schenker and Gentleman (2001) even though the latter suggest CI comparison should not be used when a test is available. Conceptually, comparing CI in this way is an informal analog of a Hausman-type Test of the differences in coefficients.

Bertolet (2008) and Faiella (2010) cast the problem of weight ignorability as a design-based versus model-based issue. Other authors go into theoretical detail on informativeness of the sampling weights, but do not propose tests (e.g., Sugden & Smith, 1984; Smith, 1988). Sugden & Smith (1984) note Rubin’s (1976) treatment of sampling as a special case of missing values. This raises the possibility that the design may be noninformative to the data owners (who have all the design variables), but informative to secondary data analysts who do not have access to all of the design variables. Since this is often the case, it underscores the need for research to provide guidance on assessing design informativeness via the weights that are available to secondary data analysts. [SHOULD WE MODIFY THIS LAST PARAGRAPH TO MAKE IT BLEND BETTER WITH THE REST OF THE MATERIAL?]

OTHER ISSUES IN TESTING

2 The CI for the weighted estimates should take account of the heteroscedasticity that might be introduced by weighting.
Beyond the diversity of tests and the different types, there are other issues that should be considered. In this section we discuss comparisons of these tests, what to do when testing only a subset of coefficients, and what we know about the finite sample properties of the tests.

**Comparison of Tests**

One difference that deserves attention are the properties of Hausman-type Test statistics based on the $T_H = [\hat{\beta}_w - \hat{\beta}_u] \prime \tilde{V}^{-1}[\hat{\beta}_w - \hat{\beta}_u]$ versus Weight Association Tests formulated as regressions (e.g., $Y = X\beta_u + X_w\beta_w + \epsilon$, with a test on $\hat{\beta}_w$). The Hausman-type Tests of differences in coefficients as described above (see Hausman, 1978; Pfeffermann, 1993; Asparouhov & Muthén, 2007) are asymptotically distributed chi-square, while the regression misspecification form of the test (see Hausman, 1978; DuMouchel & Duncan, 1983; Fuller, 1984) recommend application of the usual regression F-test for $H_0: \beta_w = 0$. Kott (1990) proposes a Hausman-type Test for comparing ordinary least squares (OLS) to probability weighted least squares regression that is chi-square distributed, but shows an alternative derivation using an F-test to use when the degrees of freedom is not large (echoing DuMouchel & Duncan, 1983). Whether the Hausman-type Tests of coefficient differences versus the regression based formulation have important impact on power is unknown, and we do not know which forms of the tests have superior finite sample properties.

A practical difference between the Hausman-type tests of differences in coefficients and the regression based Weight Association Tests also arises. As noted above, Hausman recommended that $\tilde{V} = [\tilde{V}(\hat{\beta}_w) - \tilde{V}(\hat{\beta}_u)]$. In a sample it is possible for this difference in estimates of asymptotic covariance matrices not to be positive definite. This same computational issue is not present for the regression forms of the tests.
Fuller (1984), Winship & Radbill (1994), Wu & Fuller (2005) use and evaluate the regression Weight Association Tests, but focus only on whether one correctly concludes whether ignorability holds, not on the relative performance of various tests. In a simulation study, Wu & Fuller (2005) use the DuMouchel & Duncan (1983)/Fuller (1984) regression tests of weighting to decide whether weights were needed. Then they compared multiple estimators, including the general least squares (GLS) approach of Pfeffermann & Sverchkov (1999), which was informed by the Weight Association Test the latter derived. However, Wu & Fuller (2005) did not compare the performance of their Hausman-type Test comparing coefficients with the Weight Association Tests.

**Tests for Subsets of Coefficients**

Thus far, our discussion has focused on tests of whether all coefficients are equal in weighted versus unweighted analyses. Sometimes there are coefficients of particular variables in a regression that 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 find a significant difference for at least one of the variables. This leads the researcher to a weighted analysis and the reduced efficiency that often accompanies using weights. However, it is possible that the discrepancy in coefficients is not because of the coefficient for minority status, but to some other explanatory variable in the analysis. 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 of whether one or a subset of coefficients are not equal with and without weights. If the key coefficient(s) are the same, the more efficient unweighted analysis could be done with attention directed to these key variables.¹

One form of the Hausman test 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 this to a chi-square distribution with 1 degree of freedom. A significant test statistic suggests that weighting is required, whereas an insignificant one suggests it is not. Among the Weight Association Tests, we could use the regression based approaches that include the weighted and unweighted variables and only test those coefficients of greatest interest for the weighted variables. However, we have found little empirical or analytic evidence on which tests would perform the best when testing subsets of coefficients.

**Simulation Studies of Finite Sample Properties**

For the most part, the properties of the two major types of tests are justified asymptotically. The classic Hausman-type 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 weights are not needed. Tests based on regressing the sample residual on the weights assume that the sample residuals and sample weights come from large

---

¹ 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.
enough samples 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.

This state of affairs is unsatisfactory. Researchers do not have evidence on 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 would be available that would inform us of the test properties under different conditions. But these are not available. Simulation studies under a variety of design conditions would be helpful, but even simulation work on these different tests is sparse.

**CONCLUSION**

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 option. But unnecessarily weighting means lower efficiency and lower statistical power. Tests of whether weights are required are available but rarely applied. They are not utilized for several reasons. One is the lack of awareness among researchers. Another reason is the power of tradition in different fields where some always weight and others never weight. An additional issue is that not all of these tests are readily available in software packages. Furthermore, even when easy to implement there is little guidance on which of the many tests to choose.
Our review of the tests of whether to weight highlights several areas in which research is required. Our review accomplishes one aim in that we classify tests into two major groups, Hausman-type Tests of coefficient differences and Weight Association Tests. This makes the problem more manageable and highlights the close relation among seemingly different tests. But there are a number of questions to which we still have no answers. Which of these tests have the best finite sample properties? To what degree are the Hausman-type Tests and the Weight Association Tests testing the same or different things? A related follow-up question is what should a researcher do if one test is passed and another fails? Another question comes into play when a researcher’s interest focuses on just one coefficient and wants to test whether weighting makes a difference or not. Which testing approach makes the most sense in this situation? Finally, a practical question is which of these tests require custom programming and which of these can a researcher obtain by “tricking” existing software into providing the necessary results?

Considering the dominance of complex samples with unequal selection probabilities, it is surprising that we have not seen further development of diagnostic tests to determine whether to weight or not. Hopefully, this review adds to the pressure to pursue this question more seriously than has been done up to the present.

REFERENCES


