USE OF WITHIN-PSU VARIANCES AND ERRORS-IN-VARIABLES REGRESSION TO ASSESS STABILITY OF A STANDARD DESIGN-BASED VARIANCE ESTIMATOR

D.S. Jang, Mathematica Policy Research and J.L. Eltinge, Texas A&M University
D.S. Jang, MPR, 600 Maryland Ave. SW, Suite 550, Washington, DC 20024-2512

Key words: Degrees of freedom; Diagnostic; Measurement error model; Satterthwaite approximation; Stratified multistage sample survey; Two-PSU-per-stratum design.

Abstract: Stability of a variance estimator is an important practical consideration in the analysis of sample survey data. For example, for a given point estimator $\hat{\theta}$ and design-based variance estimator $\hat{V}(\hat{\theta})$, the stability of $\hat{V}(\hat{\theta})$ is often quantified through a degrees-of-freedom term which is subsequently used in formal inference for the parameter θ. In addition, degrees-of-freedom terms are sometimes used as qualitative diagnostics for the amount of information conveyed by the variance estimator $\hat{V}(\hat{\theta})$. In analyses of stratified multistage sample data, degrees-of-freedom terms generally are calculated from the difference (number of primary sample units − number of strata), or from a standard Satterthwaite approximation. For surveys with large numbers of strata and small numbers of primary units per stratum, these approximations can be problematic under some conditions, especially in the analyses of subpopulations that are concentrated within a relatively small number of primary sample units. This paper examines the use of within-primary-sample unit variances and errors-in-variables regression to develop a modified degrees-of-freedom estimator. The proposed method is potentially applicable to cases for which there is a strong linear relationship between the overall stratum-level variances and within-primary-sample-unit variances.

1. Introduction

1.1 Large-sample design-based survey inference

Large-sample inference from sample survey data generally involves the following strategy. A fixed finite population U contains M elements with associated characteristics Y_k, $k = 1, \ldots, M$. Principal attention focuses on a parameter θ that is a function of $\{Y_k, k \in U\}$; customary examples include a population mean, total, ratio or regression coefficient. A sample design (possibly complex) is used to select a sample s of m out of the M elements in U. The sample observations $\{Y_k, k \in s\}$ are used to compute an estimator $\hat{\theta}$ of θ. Under relatively mild conditions (e.g., Cochran, 1977, Section 2.15), $m^{1/2}(\hat{\theta} - \theta)$ is distributed approximately as a normal random variable with mean zero and variance $mV(\hat{\theta})$, say. (We note in passing that with the exception of one condition considered in Section 2.2, all of the distributional work in this paper is evaluated with respect to the sample design; related ideas can be developed under a superpopulation model, but are beyond the scope of the present work.) The sample data $\{Y_j, j \in s\}$ are also used to compute an estimator $\hat{V}(\hat{\theta})$, say, of $V(\hat{\theta})$. See, e.g., Krewski and Rao (1981) for a detailed discussion of some specific design-based variance estimators and their asymptotic properties.

Under additional conditions, there exists a positive real number d such that $\{V(\hat{\theta})\}^{-1} d\hat{V}(\hat{\theta})$ is approximately distributed as a chi-square random variable on d degrees of freedom, and is approximately independent of $\hat{\theta}$. This suggests that

$$\{\hat{V}(\hat{\theta})\}^{-1/2}(\hat{\theta} - \theta) \tag{1.1}$$

is approximately distributed as a t random variable on d degrees of freedom. Then standard reasoning for parametric inference (e.g., Bickel and Doksum, 1977, Section 5.1) implies that t tests can be carried out using an approximate pivotal quantity of the form (1.1); and that approximate $(1 - \alpha)100\%$ confidence intervals for θ can be computed as,

$$\hat{\theta} \pm t_{d,1-\alpha/2}\{\hat{V}(\hat{\theta})\}^{1/2}. \tag{1.2}$$

where $t_{d,1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a t distribution on d degrees of freedom.

Within this inferential context, note that the degrees-of-freedom term d has two useful functions. First, given the choice of a given point estimator $\hat{\theta}$ and variance estimator $\hat{V}(\hat{\theta})$, the term d helps to ensure adequate performance of customary inferential methods, e.g., to ensure that the confidence interval (1.2) has a true coverage rate approximately equal to its nominal rate $1 - \alpha$. Second, assessment of d can help to identify cases in which a proposed variance estimator $\hat{V}(\hat{\theta})$ has unsatisfactory stability properties. For example, in some statistical agencies, it is customary to use a given general variance estimator $\hat{V}(\hat{\theta})$ for a wide variety of full-population and subpopulation-level analyses of a given survey

dataset. However, as one works through a sequence of subpopulations that are progressively more concentrated in a small number of primary sample units, at some point the customary variance estimator may begin to display marked instability. An appropriate estimator of d can provide a diagnostic that helps to identify the point at which this instability problem becomes severe enough to warrant consideration of alternative variance estimators, e.g., based on auxiliary information, or on generalized variance functions or average-design-effect approaches.

1.2 Degrees-of-freedom measures of the stability of a design-based variance estimator

Now consider estimation of the degrees-of-freedom term d. To simplify notation, the remainder of the paper will restrict attention to the case in which θ is a population total estimated from data collected through a stratified multistage sample design. Specifically, consider a population partitioned into L strata, with N_h primary sample units (PSUs) in stratum h. In addition, assume that in the first stage of sample selection, n_h primary units are selected with replacement from stratum h using per-draw selection probabilities p_{hi}, say, where $\sum_{i=1}^{N_h} p_{hi} = 1$ and $n = \sum_{h=1}^{L} n_h$. Also, within a selected PSU (h,i), n_{hi} secondary sample units (SSUs) are selected with per-draw selection probabilities p_{hij}, say, where $\sum_{j=1}^{N_{hi}} p_{hij} = 1$ and N_{hi} is the number of secondary units in primary unit (h,i). Finally, let N_{hij} be the number of population elements in secondary unit (h,i,j).

For a given element k in secondary unit (h,i,j), let Y_{hijk} be the survey item of interest; and define population totals $Y_{hij} = \sum_{k=1}^{N_{hij}} Y_{hijk}$, $Y_{hi} = \sum_{j=1}^{N_{hi}} Y_{hij}$, $Y_h = \sum_{i=1}^{N_h} Y_{hij}$, and $Y = \sum_{h=1}^{L} Y_h$. In addition, let $\hat{Y}_{hij}$ be a design-unbiased estimator of Y_{hij} based on observed elements within secondary unit (h,i,j); and define the related design-unbiased point estimators $\hat{Y}_{hi} = n_{hi}^{-1} \sum_{j=1}^{n_{hi}} p_{hij}^{-1} \hat{Y}_{hij}$, $\hat{Y}_h = n_h^{-1} \sum_{i=1}^{n_h} p_{hi}^{-1} \hat{Y}_{hi}$ and $\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h$. Under this design, a customary design-unbiased estimator of the design variance of $\hat{Y}$ is $\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \hat{V}_h$ where $\hat{V}_h = n_h^{-1}(n_h - 1)^{-1} \sum_{h=1}^{n_h} (p_{hi}^{-1} \hat{Y}_{hi} - \hat{Y}_h)^2$ (Wolter, 1985, p.44).

To evaluate the variance of $\hat{V}(\hat{Y})$, note that due to independence of sampling across strata, $V\{\hat{V}(\hat{Y})\} = \sum_{h=1}^{L} V(\hat{V}_h)$. In addition, we will use the following commonly employed assumption regarding the distribution of the stratum-level variance estimators $\hat{V}_h$.

(C.1) For $h = 1, 2, \ldots, L$, assume that $n_h \geq 2$, and that the terms $V_h^{-1}(n_h - 1)\hat{V}_h$ are distributed as independent chi-square random variables on $n_h - 1$ degrees of freedom, respectively, $h = 1, \ldots, L$.

Note that under condition (C.1), $V\{\hat{V}(\hat{Y})\} = \sum_{h=1}^{L} (n_h - 1)^{-1} 2V_h^2$. In addition, $\{V(\hat{Y})\}^{-1} d\hat{V}(\hat{Y})$ has the same first and second moments as a chi-square random variable on d degrees of freedom, where d is the solution to the equation,

$$2\{V(\hat{Y})\}^2 - V\{\hat{V}(\hat{Y})\}d = 0 \qquad (1.3)$$

Thus, under condition (C.1),

$$d = \{\sum_{h=1}^{L} (n_h - 1)^{-1} V_h^2\}^{-1} \{V(\hat{Y})\}^2. \qquad (1.4)$$

The sampling literature has proposed several estimators for d. The performance of these estimators tends to depend heavily on the degree of heterogeneity of the stratum-level variances V_h, and on the relative magnitudes of L and n_h. For cases in which the V_h are all equal, routine arguments show that $d = n - L$. Consequently, many applications use $n - L$ as the degrees-of-freedom term in the confidence interval (1.2) and related analyses. However, if the V_h terms are not equal, expression (1.4) can be substantially less than $n - L$, and it can be important to account explicitly for the heterogeneity of the V_h in estimation of d. If the V_h are moderately heterogeneous, this can be accomplished through direct use of the variance estimators $\hat{V}_h$. For example, for cases in which L is small and n_h is moderate or large for all h, it is customary to use the Satterthwaite (1946)-type estimator of d, (see, e.g., Cochran, 1977, p. 96; or Kendall et al., 1983, pp. 91-92),

$$\hat{d}_S = \{\sum_{h=1}^{L} (n_h - 1)^{-1} \hat{V}_h^2\}^{-1} \{\hat{V}(\hat{Y})\}^2. \qquad (1.5)$$

Also, for cases in which L is moderate or large, the n_h values are small and the V_h are moderately heterogeneous, Jang and Eltinge (1995) discussed possible modifications of the Satterthwaite estimator $\hat{d}_S$. For example, under the commonly encountered design in which $n_h = 2$ for all h, they considered the estimator,

$$\hat{d}_{mS} = (3L + 14)^{-1}(9L)\hat{d}_S. \qquad (1.6)$$

1.3 Use of stratum-level auxiliary information in assessment of variance estimator stability

For cases in which the V_h terms are expected to be relatively heterogeneous, it can be useful to supplement the $\hat{V}_h$ estimates with additional information regarding the relative magnitudes of the V_h terms. Depending on the application, one could consider several possible sources of such auxiliary information.

The remainder of this paper focuses on the use of within-PSU variances V_{Wh}, say, for this purpose. Section 2 reviews a detailed definition of V_{Wh}, discusses estimation of V_{Wh} and presents a simple errors-in-variables model for the relationship between V_h and V_{Wh}. Section 3 considers the use of this model to produce alternative estimators of d. Section 3.1 uses a model with an error in the equation, and Section 3.2 uses a related model with no equation error. Section 3.3 discusses the use of plots and other diagnostics in conjunction with the estimators in Sections 3.1 and 3.2. Finally, the use of auxiliary information to produce an improved estimator of d is somewhat analogous to previous sample survey work with the use of auxiliary data to produce improved point estimators and variance estimators. Section 4 explores this idea in concluding remarks.

2. Within-PSU Variances

2.1 Estimation of within-PSU variances

Recall from Wolter (1985, p.41) that one may decompose $V_h = V(\hat{Y}_h)$ as,

$$V_h = V_{Bh} + V_{Wh}, \tag{2.1}$$

where $V_{Bh} = V\left\{\sum_{i=1}^{n_h}(n_h p_{hi})^{-1} Y_{hi}\right\}$ is a between-PSU variance term, $V_{Wh} = \sum_{i=1}^{N_h}(n_h p_{hi})^{-1}\sigma^2_{2hi}$ is a within-PSU variance term, and

$$\sigma^2_{2hi} = \mathrm{Var}(\hat{Y}_{hi}|\text{PSU } i, \text{ stratum } h)$$

reflects sampling variability within a specific primary unit (h,i).

Under the conditions stated in Section 1.2, routine arguments (e.g., Eltinge and Jang, 1996, Section 2.2) show that for any given (h,i),

$$\hat{\sigma}^2_{2hi} = n_{hi}^{-1}(n_{hi}-1)^{-1}\sum_{j=1}^{n_{hi}}(p_{hij}^{-1}\hat{Y}_{hij} - \hat{Y}_{hi})^2$$

is design unbiased for σ^2_{2hi}; and that

$$\hat{V}_{Wh} = n_h^{-2}\sum_{i=1}^{n_h} p_{hi}^{-2}\hat{\sigma}^2_{2hi}, \tag{2.2}$$

is design unbiased for V_{Wh}. In addition, a design unbiased estimator of the design variance of $\hat{V}_{Wh}$ is

$$\hat{V}(\hat{V}_{Wh}) = n_h^{-1}(n_h-1)^{-1}\sum_{i=1}^{n_h}(\hat{V}_{Whi} - \hat{V}_{Wh})^2. \tag{2.3}$$

Note that the variates $\hat{V}_h$ are functions of the sample means of the random variables $p_{hij}^{-1}\hat{Y}_{hij}$, taken over the selected primary units in stratum h. Conversely, the estimators $\hat{V}_{Wh}$ are functions of sample variances of the $p_{hij}^{-1}\hat{Y}_{hij}$ terms. Thus, if within each stratum h the terms $p_{hij}^{-1}\hat{Y}_{hij}$ are approximately distributed as normal random variables, then $\hat{V}_h$ and $\hat{V}_{Wh}$ are approximately independent.

2.2 Relationships between V_h and V_{Wh}

This work was motivated by some empirical cases in which the relationship between V_h and V_{Wh} appears to be well approximated by the linear regression equation,

$$V_h = \beta_0 + \beta_1 V_{Wh} + q_h, \tag{2.4}$$

where β_0 and β_1 are fixed regression coefficients and the equation error terms q_h are small relative to the overall variability of the V_h terms across $h = 1,\ldots,L$.

In addition, note that we may view $\hat{V}_h$ and $\hat{V}_{Wh}$ as sums

$$\hat{V}_h = V_h + e_h \qquad \text{and} \qquad \hat{V}_{Wh} = V_{Wh} + u_h \tag{2.5}$$

of true values (V_h, V_{Wh}) and errors (e_h, u_h).

Taken together, expressions (2.4) and (2.5) define an simple linear errors-in-variables model. In keeping with other errors-in-variables literature (e.g., Fuller, 1987, Chapter 1), we will assume that the errors (q_h, e_h, u_h) satisfy the following condition.

(C.2) Assume that q_h, e_h and u_h defined in expressions (2.4) and (2.5) are mutually independent random variables with common mean zero and variances σ_{qqh}, $V(\hat{V}_h)$ and $V(\hat{V}_{Wh})$, respectively, $h = 1,\ldots,L$.

Note that the distributional assumptions on e_h and u_h are consistent with the general randomization-based ideas in Section 2.1. However, the assumption of randomness of q_h goes somewhat beyond traditional randomization approaches, which generally view V_h and V_{Wh} as fixed quantities. If one prefers to consider a pure randomization approach to survey inference, one could instead define (β_0, β_1) to be the coefficients from the least-squares regression of

the fixed true V_h on the fixed true V_{Wh} and an intercept term; define $q_h = V_h - \beta_0 - \beta_1 V_{Wh}$; and then assume that the finite-population mean of q_h, and related covariances of q_h with other stratum-level quantities, are all negligible. In an infomal sense, these finite-population assumptions amount to assuming that the equation errors q_h are not strongly associated with the other essential features of expressions (2.4) and (2.5). See Eltinge (1994) for related comments, and for formal details of an associated asymptotic framework. To simplify notation, the remainder of the paper will not consider further this finite-population approach, and will instead make direct use of condition (C.2).

3. Errors-in-Variables Estimation of d

Under conditions (C.1) and (C.2) and additional asymptotic regularity conditions, the limiting first and second moments of $L^{-1}\hat{V}$, conditional on the true $(V_{W1}, \ldots, V_{WL})'$, are

$$\lim_{L\to\infty} L^{-1} \sum_{h=1}^{L} (\beta_0 + \beta_1 V_{Wh}) \qquad \text{and}$$

$$\lim L^{-1} 2 \sum_{h=1}^{L} (n_h - 1)^{-1} \{(\beta_0 + \beta_1 V_{Wh})^2 + \sigma_{qqh}\}.$$

Application of these results to expression (1.4) suggests the definition,

$$L^{-1} d_{EIV} = D^{-1} N, \tag{3.1}$$

where

$$N = \beta_0^2 + 2\beta_0\beta_1 \bar{V}_W + \beta_1^2 (\bar{V}_W)^2,$$

$$D = \beta_0^2 L^{-1} \sum_{h=1}^{L} (n_h - 1)^{-1} + 2\beta_0\beta_1 \bar{V}_W(n_h) + \beta_1^2 \bar{V}_{W2}(n_h) + \bar{\sigma}_{qq\cdot},$$

$$\bar{V}_W = L^{-1} \sum_{h=1}^{L} V_{Wh},$$

$$\bar{V}_W(n_h) = L^{-1} \sum_{h=1}^{L} (n_h - 1)^{-1} V_{Wh},$$

$$\bar{V}_{W2}(n_h) = L^{-1} \sum_{h=1}^{L} (n_h - 1)^{-1} V_{Wh}^2 \qquad \text{and}$$

$$\bar{\sigma}_{qq\cdot} = L^{-1} \sum_{h=1}^{L} (n_h - 1)^{-1} \sigma_{qqh}.$$

Note that the limit of $L^{-1} d_{EIV}$ is a function of the limit of

$$\theta = (\beta_0, \beta_1, \bar{V}_W, (\bar{V}_W)^2, \bar{V}_W(n_h), \bar{V}_{W2}(n_h), \bar{\sigma}_{qq\cdot})',$$

say. Consequently, one can construct an estimator of d_{EIV} by substituting consistent estimates of the components of θ into expression (3.1). Under moderate regularity conditions that are not specified here for reasons of space,

$$\hat{\bar{V}}_W = L^{-1} \sum_{h=1}^{L} \hat{V}_{Wh}$$

$$\widehat{\bar{V}_W^2} = \left(L^{-1} \sum_{h=1}^{L} \hat{V}_{Wh} \right)^2 - \widehat{\text{Var}} \left(L^{-1} \sum_{h=1}^{L} \hat{V}_{Wh} \right),$$

$$\widehat{\bar{V}_W}(n_h) = L^{-1} \sum_{h=1}^{L} (n_h - 1)^{-1} \hat{V}_{Wh} \qquad \text{and}$$

$$\widehat{\bar{V}_{W2}}(n_h) = L^{-1} \sum_{h=1}^{L} (n_h - 1)^{-1} [\hat{V}_{Wh}^2 - \hat{V}(\hat{V}_{Wh})].$$

are consistent estimators of the limits of $\bar{V}_W$, $(\bar{V}_W)^2$, $\bar{V}_W(n_h)$ and $\bar{V}_{W2}(n_h)$, respectively, where $\hat{V}(\hat{V}_{Wh})$ is defined in expression (2.3). Thus, it remains to construct consistent estimators of β_0, β_1 and the limiting value of $\sigma_{qq\cdot}^2$. In developing such estimators, the errors-in-variables literature generally gives separate consideration to the cases in which $\sigma_{qq\cdot}^2$ is greater to zero and equal to zero. These cases are addressed in Sections 3.1 and 3.2, respectively.

3.1 A model with equation errors

Consider first the case in which $\bar{\sigma}_{qq\cdot} > 0$. Then following Fuller (1987, pp. 187-189), define the estimators,

$$\hat{\beta}_1 = \left[\sum_{h=1}^{L} \{(\hat{V}_{Wh} - \hat{\bar{V}}_W)^2 - \hat{V}(\hat{V}_{Wh})\} \right]^{-1} \times \sum_{h=1}^{L} (\hat{V}_{Wh} - \hat{\bar{V}}_W) \hat{V}_h,$$

$$\hat{\beta}_0 = L^{-1} \sum_{h=1}^{L} \hat{V}_h - \hat{\beta}_1 \hat{\bar{V}}_W$$

and

$$\hat{\bar{\sigma}}_{qq\cdot} = L^{-1} \sum_{h=1}^{L} (n_h - 1)^{-1} \{(L-2)^{-1} L (\hat{V}_h - \hat{\beta}_0 - \hat{\beta}_1 \hat{V}_{Wh})^2 - \hat{\sigma}_{eeh} - \hat{\beta}_1^2 \hat{V}(\hat{V}_{Wh})\},$$

where $\hat{\sigma}_{eeh} = 2(n_h + 1)^{-1} \hat{V}_h^2$ from condition (C.1). Naturally, the estimator of $\bar{\sigma}_{qq\cdot}$ is taken to be the maximum of (3) and zero. Then an estimator $L^{-1}\hat{d}_{EIV}$, say, follows from substitution of the abovementioned estimators into expression (3.1).

3.2 A model without equation errors

In some cases, empirical evidence suggests that $\sigma_{qq\cdot} = 0$. For such a case, called an errors-in-variables model with no equation error, one generally uses alternative point estimators $(\tilde{\beta}_0, \tilde{\beta}_1)$, say. See Fuller (1987, pp. 190-191) for details. Substitution of $(\tilde{\beta}_0, \tilde{\beta}_1)$ and $\sigma_{qq\cdot} \doteq 0$ into the previous expression for $L^{-1}\hat{d}_{EIV}$ leads to an alternative estimator $\tilde{d}_{EIV}$, say.

3.3 Related diagnostics

In general, the performance of $\hat{d}_{EIV}$ or $\tilde{d}_{EIV}$ will depend heavily on the extent to which the within-PSU variance estimators $\hat{V}_{Wh}$ provide useful supplementary information regarding the relative magnitudes of the true overall stratum-level variances V_h. To assess this, the following diagnostics are potentially useful. First, a scatterplot of $\hat{V}_h$ against $\hat{V}_{Wh}$ provides some qualitative information regarding the adequacy of the linear approximation (2.4). Of special interest are the general dispersion of the scatterplot around a straight-line errors-in-variables fit and the possible presence of nontrivial curvature, a nonzero intercept β_0, or outliers.

Second, to some degree, interpretation of such scatterplots can be complicated by the presence of error in the $\hat{V}_{Wh}$. Consequently, it is useful to consider some other diagnostics that account explicitly for these errors. For a detailed discussion of errors-in-variables diagnostics, see, e.g., Fuller (1987), Carroll, Ruppert and Stefanski (1995) and references cited therein. Here, we will restrict attention to some simple moment-based tools. For example, consider $\hat{\beta}_0 + \hat{\beta}_1\hat{V}_{Wh}$ as an estimator of V_h, and note that under the linear approximation (2.4),

$$\begin{aligned}(\hat{\beta}_0 + \hat{\beta}_1\hat{V}_{Wh}) - V_h &= \beta_1(\hat{V}_{Wh} - V_{Wh}) - q_h \\ &+ (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)V_{Wh} \\ &+ (\hat{V}_{Wh} - V_{Wh})(\hat{\beta}_1 - \beta_1)\end{aligned} \tag{3.2}$$

Under regularity conditions, expression (3.2) is dominated by its leading term, $\beta_1(\hat{V}_{Wh} - V_{Wh}) - q_h$, which has a variance estimated by $\hat{\beta}_1^2\hat{V}(\hat{V}_{Wh}) + \hat{\bar{\sigma}}_{qq\cdot}$ under the assumption of homogeneous $\sigma_{qqh\cdot}$. Also, $\hat{V}(\hat{V}_h)$ reflects the precision of $\hat{V}_h$ as an estimator of V_h. Consequently, define the ratio,

$$r = \{\sum_{h=1}^{L}\hat{V}(\hat{V}_h)\}^{-1}\sum_{h=1}^{L}\{\hat{\beta}_1^2\hat{V}(\hat{V}_{Wh}) + \hat{\bar{\sigma}}_{qq\cdot}\}$$

One generally would consider use of $\hat{d}_{EIV}$ or $\tilde{d}_{EIV}$ only if r is substantially less than unity.

Another quantity that assesses the magnitude of errors in predictor variables is the *reliability ratio*, κ_{xx}, defined to be the ratio of the population variances of true values, divided by the population variance of observed values; see, e.g., Fuller (1987, p. 3). For the present case, an estimator of κ_{xx} is,

$$\begin{aligned}\hat{\kappa}_{xx} = &\left[\sum_{h=1}^{L}(\hat{V}_{Wh} - \hat{\bar{V}}_W)^2\right]^{-1} \\ &\times \left[\sum_{h=1}^{L}(\hat{V}_{Wh} - \hat{\bar{V}}_W)^2 - \hat{\sigma}_{uu\cdot}\right],\end{aligned}$$

Given independence of the errors from the true values, κ_{xx} falls between 0 and 1, with κ_{xx} close to 1 indicating that measurement error makes a relatively small contribution to the overall variability of the observed values. In application and simulation work not detailed here, we used $\hat{\kappa}_{xx}$ as a preliminary diagnostic to identify cases for which the $\hat{V}_{Wh}$ could potentially provide useful supplementary information. In particular, we found that use of $\hat{d}_{EIV}$ or $\tilde{d}_{EIV}$ was problematic for applications involving $\hat{\kappa}_{xx}$ less than 0.7 or 0.8.

Finally, the estimator $\hat{\bar{\sigma}}_{qq\cdot}$ gives a pooled indication of the amount of dispersion in the approximate linear relationship (2.4). For cases in which r is not small, the two component ratios

$$\{\sum_{h=1}^{L}\hat{V}(\hat{V}_h)\}^{-1}\sum_{h=1}^{L}\hat{\beta}_1^2\hat{V}(\hat{V}_{Wh})$$

and $\{\sum_{h=1}^{L}\hat{V}(\hat{V}_h)\}^{-1}\hat{\bar{\sigma}}_{qq\cdot}$ gives some indication of whether the limitations of the errors-in-variables approach are attributable to a large estimation error variance $V(\hat{V}_{Wh})$, a large average equation error variance $\sigma_{qq\cdot}$, or both. Also, for cases in which an errors-in-variables approach may be appropriate, a formal test of the null hypothesis $\mathrm{H}_0 : \sum_{h=1}^{L}\sigma_{qqh\cdot} = 0$ is useful in deciding whether to use $\hat{d}_{EIV}$ or $\tilde{d}_{EIV}$; see, e.g., Fuller (1987) for examples of formal tests for equation error.

4. Discussion

Recall from Section 1.1 that customary design-based large-sample inference with survey data involves a point estimator $\hat{\theta}$, a variance estimator $\hat{V}(\hat{\theta})$ and a degrees-of-freedom term d. In parallel with this, one may consider use of auxiliary data to develop improved estimators for one or more of these three parts of an analysis. For point estimation, use of auxiliary data has received extensive attention. See,

e.g, the classical discussion of ratio- and regression-based point estimators in Cochran (1977, Chapters 6 and 7); and other work with regression estimators and related approaches in Fuller (1975), Isaki and Fuller (1982), Särndal et al. (1992) and references cited therein. In that work, a common theme is that efficiency improvements, if any, obtained through the use of auxiliary data will depend on the adequacy of approximations (generally involving regression equations) for the relationship between the auxiliary data and the principal variables of interest. Also, to reflect these potential efficiency gains, the use of auxiliary data for point estimation generally also leads to modified variance estimators and may also require modified degrees-of-freedom terms.

Also, in some cases auxiliary data is used to produce a variance estimator $\tilde{V}(\hat{\theta})$, say, that is expected to be more stable than the customary design-based variance estimator; see, e.g., Isaki (1983). Again here, the magnitude of improvements in variance estimator stability will depend on the strength of the relationship between the available auxiliary data and the variables of principal interest.

In this paper, we have considered the use of auxiliary data to produce a degrees-of-freedom estimator that may be more stable than customary Satterthwaite-type estimators such as $\hat{d}_S$ and $\hat{d}_{mS}$. The resulting errors-in-variables estimators are intended primarily for cases in which L is relatively large, n_h values are small and the true stratum variances V_h are heterogeneous. Also, as emphasized in Sections 2.2 and 3.3, performance of the proposed estimator will depend heavily on the adequacy of the approximation (2.4) for the relationship of V_h with V_{Wh} across $h = 1, \ldots, L$.

Acknowledgements

This research was supported in part by the U.S. National Center for Health Statistics. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. National Center for Health Statistics.

References

Bickel, P.J. and Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco: Holden-Day.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models.* New York: Chapman and Hall.

Cochran, W.G. (1977). *Sampling Techniques* (Third Edition). New York: Wiley.

Eltinge, J.L. (1994). "A Finite Population Sampling Approach to Approximating Small Measurement Error Effects in Regression." *Sankhyā, Ser. B* **56**, 234-250.

Eltinge, J.L. and Jang, D.S. (1996). "Stability Measures for Variance Component Estimators Under a Stratified Multistage Design." *Survey Methodology* **22**, 157-165.

Fuller, W.A. (1975). "Regression Analysis for Sample Survey." *Sankhyā, Series C* **37**, 117-132.

Fuller, W.A. (1987). *Measurement Error Models.* New York: John Wiley.

Isaki, C.T. (1983). "Variance Estimation Using Auxiliary Information." *J. Amer. Statist. Assoc.* **78** 117-123.

Isaki, C.T. and Fuller, W.A. (1982). "Survey Design Under the Regression Superpopulation Model." *J. Amer. Statist. Assoc.* **77**, 89-96.

Jang, D.S. and Eltinge, J.L. (1995). "Empirical Assessment of The Stability of Variance Estimators Based on a Two-Clusters-Per-Stratum Design." Technical Report # 225, Department of Statistics, Texas A&M University, Submitted for publication.

Kendall, M., Stuart, A. and Ord, J.K. (1983). *The Advanced Theory of Statistics, Volume 3: Design and Analysis, and Time-Series.* New York: Macmillan.

Krewski, D. and Rao, J.N.K. (1981). "Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods." *Ann. Statist.* **9**, 1010-1019.

Rao, C.R. (1973). *Linear Statistical Inference and Its Application*, New York: John Wiley.

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

Satterthwaite, F.E. (1946). "An Approximate Distribution of Estimates of Variance Components." *Biometrics* **2**, 110-114.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.