Journal of Official Statistics
Vol. 10. No. 4, 1994, pp. 407–418

Reweighting and Variance Estimation for the Characteristics of Business Owners Survey

Phillip S. Kott [1]

Abstract: This paper explores the use of post-stratification to compensate for the unit nonresponse in the 1987 U.S. Characteristics of Business Owners (CBO) Survey. As is often the case, post-stratification leads to estimators with desirable properties under both a quasi-random (response) and a parametric (superpopulation) model. Some care is necessary in setting up these simple models because the CBO is a survey both of firm characteristics and of the characteristics of firm owners. Variance estimation methods are proposed that measure parametric model variance and quasi-design mean squared error simultaneously even when finite population correction cannot be ignored.

Key words: Unit nonresponse; quasi-randomization; parametric model; quasi-design unbiased.

1. Introduction

The 1987 Characteristics of Business Owners (CBO) Survey is used to estimate the proportion of U.S. firms and business owners with particular characteristics. Essentially, stratified simple random samples of roughly 25,000 firms were drawn independently from five mutually exclusive panels: a panel of Hispanic-owned firms, one of Asian-owned firms, one of women-owned firms, one of black-owned firms, and one of all other firms. Rules were established to assign firms with, say, both black and female owners to a single panel. The exact nature of these rules is beyond the scope of this paper.

CBO survey data are used to estimate two different types of proportions. The first is the proportion of owners with a certain survey characteristic (e.g., level of education). The second is the proportion of firms with a certain characteristic (e.g., number of women employees). The distinction arises when a sampled firm has more than one owner. The firm has women employees, but each owner has his (her) own level of education.

Although all owners within sampled firms were sent CBO questionnaires, only about 65% responded to the survey in each panel. To compensate for the nonresponse, two different but logically consistent owner reweighting schemes had to be developed – one for owner characteristics

[1] Chief Research Statistician, Research Division, National Agricultural Statistics Service, 3251 Old Lee Highway, Fairfax, VA 22030-1504, U.S.A. This paper reports on research undertaken at the U.S. Census Bureau. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau or the National Agricultural Statistics Service.

and one for firm characteristics. These schemes produce estimators that have desirable properties under both a quasi-randomization (response) and a parametric (superpopulation) model. In the former model, probabilities of responding to the survey are modelled, while in the latter, probabilities of having a particular characteristic are modelled. The paper develops new, but familiar-looking, variance estimators that simultaneously estimate conditional variance under the parametric model and design mean squared error under the quasi-randomization model. This variance methodology is also applied to certain domain estimators where the implied parametric model has not previously been discussed in the literature.

2. The Estimators

2.1. *The sample design*

The general method the Census Bureau has chosen to handle unit nonresponse in the 1987 CBO Survey is to post-stratify sampled owners into *response homogeneity groups* (Särndal, Swensson, and Wretman 1992, pp. 577–580) and then reweight them. The response or *quasi-randomization* model (Oh and Scheuren 1983) used in reweighting CBO data holds that the CBO respondent sample is effectively the product of a two-phase design. In the first phase, firms were selected using stratified simple random sampling. In the second phase, a simple random subsample of respondent owners was selected from among the owners of the originally sampled firms within each response homogeneity group. The second phase of sampling is, of course, a fiction; hence the prefix "quasi" on quasi-randomization model. Under this model, we can determine a *quasi-sampling* weight for each owner by taking the inverse of the product of his (her) firm's first-phase probability of selection and his (her) own second-phase quasi-probability of selection.

The original design strata within each panel of the 1987 CBO survey were determined by crossing state, industry, and receipt size classifications. There are between two and four sampled firms in a great number of the design strata. The CBO was deeply stratified in this manner to assure that the domain estimators for individual states, industries, and receipt size classes would be based on reasonably large samples.

Response homogeneity groups were much larger than design strata. They were defined by crossing industry and a combination of receipt size and legal form of organization: sole proprietor, partnership, or corporation (limited partnerships were out of scope). All owners in the same firm were allocated to the same group.

As is often the case in practice, necessity played a role in the development of the response homogeneity groups. Potential groups were collapsed together to assure that a minimum number of firms had respondent owners within each group. Nevertheless, the quasi-randomization assumption underpinning the treatment of nonresponse in the CBO is that all owners from the same response homogeneity group have the same probability of being "selected" for the sample.

2.2. *Owner characteristics*

Let y_{gi} equal 1 if owner i in response homogeneity group g has a particular survey characteristic (e.g., a level of education)

and zero otherwise. CBO data are used to estimate population proportions like

$$Y_O = \frac{\sum_{g=1}^{G} N_{Og}\left(\sum_{i=1}^{N_{Og}} y_{gi}/N_{Og}\right)}{N_O} = \sum_{g=1}^{G} N_{Og} Y_{Og}/N_O$$

where there are G groups in the population; N_{Og} is the number of owners in g, and Y_{Og} is the proportion of owners in group g with the survey characteristic.

Suppose we assume that each y_{gi} can be treated as a random variable with mean p_{Og}. Kott (1994) called this assumption a *parametric* model because in conventional design-based sampling theory the y_{gi} are parameters not random variables. It has also been called a superpopulation model. Some modifier on the term "model" is necessary to distinguish a parametric model from a quasi-randomization model.

Observe that the same G groups serve as both the response homogeneity groups under the quasi-randomization model and the *parametric model groups* under the parametric model. Unlike the quasi-randomization model, however, there is a separate parametric model for every survey variable. Like the quasi-randomization model, each parametric model is nothing more than an assumption.

The simple quasi-randomization and parametric models posited for the CBO are certainly subject to question. Nevertheless, they formally lay out the theory supporting the adjustments actually used in treating nonresponse. The quality of these adjustments is directly linked to the reasonability of the models. If they fail, the estimators discussed in this paper can be biased.

The post-stratified estimator for Y_O is

$$y_O = \frac{\sum_{g=1}^{G} N_{Og}\left(\sum_{i=1}^{r_{Og}} q_{gi} y_{gi} \Big/ \sum_{i=1}^{r_{Og}} q_{gi}\right)}{\sum_{g=1}^{G} N_{Og}} = \frac{\sum_{g=1}^{G}\sum_{i=1}^{r_{Og}} a_{Ogi} y_{gi}}{\sum_{g=1}^{G}\sum_{i=1}^{r_{Og}} a_{Ogi}} \quad (1)$$

where r_{Og} is the number of respondent owners in group g (which is assumed to be positive), $a_{Ogi} = q_{gi}[N_{Og}/\sum^{r_{Og}}(q_{gk})]$, and q_{gk} is the quasi-sampling weight of owner gk. That is, q_{gk} is equal to the product of w_{gk} and n_{Og}/r_{Og}, where w_{gk} is the original (CBO) sampling weight for the firm associated with gk, and n_{Og}/r_{Og} is the ratio of the number of owners in group g's original sample to the number of respondent owners in that group.

The value a_{Ogi} is the *adjusted owner weight* for gi. The estimator in equation (1) is unbiased under the parametric model in the sense that $E_M(y_O - Y_O) = 0$, where $E_M(\cdot)$ denotes expectation with respect to the parametric model.

The adjustment factor, $N_{Og}/\Sigma q_{gk}$, which when multiplied by q_{gi} yields a_{Ogi}, has an expectation under the quasi-randomization model of nearly 1. This is because $E_D[\sum^{r_{Og}}(q_{gk})] = N_{Og}$, where $E_D(\cdot)$ denotes expectation with respect to the quasi-design. As a result, y_O is approximately a design unbiased estimator for Y_O under the quasi-randomization model for sufficiently large r_{Og}.

Formally, if each $\min\{q_{gk}/q_{gk'}\}$ is bounded from below and each $\max\{q_{gk}/q_{gk'}\}$ is bounded from above, then each $N_{Og}/\Sigma q_{gk} = 1 + O_p(r_{Og}^{-1/2})$, which is approximately 1 for large r_{Og}.

Consequently, the quasi-design expectation of y_O/Y_O is approximately unity when all the r_{Og} are large.

2.3. Firm characteristics

Let us now turn our attention to firm characteristics. Let y_{gi} equal 1 if the firm that owner i is associated with has a certain survey characteristic and zero otherwise. We are interested in estimating firm population proportions like

$$Y_F = \frac{\sum\sum^{G \; N_{Og}} y_{gi}/n_{Ogi}}{N_F} \tag{2}$$

where n_{Ogi} is the number of owners in the firm associated with gi, and N_F is the total number of firms in the panel. Although equation (2) has an uncommon form, it is identical to the simple average of y_{gi} values among firms in the panel.

It is easy to see that $y'_F = \sum^{G}\sum^{r_{Og}}(q_{gi}y_{gi}/n_{Ogi})/\sum^{G}\sum^{r_{Og}}(q_{gi}/n_{gOi})$ is a nearly quasi-design unbiased estimator for y_F. An estimator that is both nearly quasi-design unbiased (for sufficiently large r_{Fg}) and unbiased under the parametric model is

$$y_F = \sum_{g=1}^{G}(N_{Fg}/N_F)\frac{\sum^{r_{Og}} q_{gi}y_{gi}/n_{Ogi}}{\sum q_{gi}/n_{Ogi}}$$

$$= \frac{\sum\sum^{G \; r_{Og}} a_{Fgi}y_{gi}}{\sum\sum a_{Fgi}}$$

where N_{Fg} is the number of firms in the population within group g and $a_{Fgi} = (q_{gi}/n_{Ogi})[N_{Fg}/\sum_{k=1}^{r_{Og}}(q_{gk}/n_{Ogk})]$ is the *adjusted firm weight* for owner gi. Observe that while the quasi-sampling weight for owner gi, q_{gi}, is the same for both owner and firm characteristics, the adjusted owner weight for gi, a_{Ogi}, may not equal the owner's adjusted firm weight, a_{Fgi}.

In what follows, all r_{Fg} – and thus all r_{Og} – are assumed to be sufficiently large that each $a_{Fgi}/(q_{gi}/n_{Ogi})$ (and a_{Ogi}/q_{gi}) is approximately 1. Formally, $a_{Fgi}/(q_{gi}/n_{Ogi}) = 1 + O_p(r_{Fg}^{-1/2})$. Observe that when this value is approximately 1, so is $r_{Fg}/(r_{Fg}-1)$.

3. Variance Estimation

3.1. A general form

Let $d_{Ogi} = y_{gi} - \sum_{k=1}^{N_{Og}} y_{gk}/N_{Og}$. The quasi-design mean squared error for y_O is identical to that for $\sum^{G}\sum^{r_{Og}} a_{Ogi}d_{Ogi}/N_O$ and approximately equal to that of $\sum^{G}\sum^{r_{Og}} q_{gi}d_{Ogi}/N_O$ for large r_{Fg}. Equation (8) in Kott (1990) contains an estimator for the design variance for an estimator of the form $\sum^{G}\sum^{r_{Og}} q_{gi}d_{Ogi}/N_O$ based on a two-phase sample (actually the estimator in Kott is not divided by N_O). Nothing is lost by replacing the q_{gi} in that estimator by a_{Ogi} and the d_{Ogi} by $y_{gi} - \sum^{r_{Og}} a_{Ogk}y_{gk}/N_{Og}$; that is to say, it remains an approximately quasi-design unbiased estimator for the quasi-design mean squared error of y_O. Formally, its relative design bias is at most $O(\max\{r_{Og}^{-1/2}\})$ under the quasi-randomization model.

It is necessary to replace the d_{Ogi} by $y_{gi} - \Sigma a_{Ogk}y_{gk}/N_{Og}$ because they are not known. Replacing q_{gi} by a_{Ogi} parallels the technique used in the weighted residual variance estimator of Särndal, Swensson, and Wretman (1989) to estimate (parametric) model variance simultaneously with (quasi-) design mean squared error. We will return to the topic of parametric model variance in the next subsection.

A similar derivation is possible for y_F. Let $d_{Fgi} = y_{gi} - \sum_{k=1}^{N_{Og}}(y_{gk}/n_{Ogk})/N_{Fg}$. The quasi-design mean squared error of y_F is identical to that of $\sum^{G}\sum^{r_{Og}} a_{Fgi}d_{Fgi}/N_F$ and approximately equal to that of $\sum^{G}\sum^{r_{Og}}(q_{gi}/n_{Ogi})d_{Fgi}/N_F$ for large r_{Fg}. Nothing is lost by replacing the q_{gi}/n_{Ogi} in

that estimator by a_{Fgi} and the d_{Fgi} by $y_{gi} - \sum^{r_{Og}} a_{Fgk} y_{gk}/N_{Fg}$; that is to say, it remains an approximately quasi-design unbiased estimator for the quasi-design mean squared error of y_F.

The quasi-design mean squared error estimates for y_O and y_F discussed above can be derived with a single set of notation. To this end, let a_{gi} denote either a_{Ogi} or a_{Fgi} and N either N_O or N_F as appropriate, and let $e_{gi} = a_{gi}[y_{gi} - \sum^{r_{Og}} (a_{gk} y_{gk})/\Sigma a_{gk}]/N$ be the *weighted residual* for owner i (Särndal et al. 1989). In addition, let S_h be the set of originally sampled firms in design stratum h, $h = 1, \ldots, H$, and S_h' be the set of respondent owners in h. Finally, let f_h be the original sampling fraction for stratum h, and n_{Fh} be the number of originally sampled firms in stratum h. Note that f_h equals $1/w_{gi}$ for an owner gi within a firm in sampling stratum h.

After much manipulation, the quasi-design mean squared error estimator for both y_O and y_F discussed above can be expressed as

$$v' = A' - B' - C' - D' \tag{3}$$

where

$$A' = \sum_{g=1}^{G} (r_{Og}/[r_{Og} - 1]) \times ([n_{Og} - 1]/n_{Og}) \sum_{h=1}^{H} \sum_{j \in S_h} u_{jg}^2$$

$$B' = \sum_{h=1}^{H} (1 - f_h) \left[\left(\sum_{g=1}^{G} \sum_{j \in S_h} u_{jg} \right)^2 - \sum_{g=1}^{G} \sum_{j \in S_h} u_{jg}^2 \right] \Big/ (n_{Fh} - 1)$$

$$C' = \sum_{h=1}^{H} f_h \sum_{g=1}^{G} (r_{Og}/[r_{Og} - 1]) \times \left\{ \sum_{j \in S_h} (1 - 1/n_{Og}) u_{jg}^2 - \sum_{mi \in S_h'} (1 - r_{Og}/n_{Og}) e_{mig}^2 \right\}$$

$$D' = \sum_{g=1}^{G} (1/[r_{Og} - 1])([n_{Og} - r_{Og}]/n_{Og}) \times \sum_{h=1}^{H} (1 - f_h)/(n_{Fh} - 1) \times \left[\left(\sum_{j \in S_h} u_{jg} \right)^2 - \sum_{j \in S_h} u_{jg}^2 \right]$$

and $e_{mig} = e_{mi}$ when $m = g$ and zero otherwise, and u_{jg} is the sum of e_{mig} taken over all respondent owners in sampled firm j (if that set is empty, u_{jg} is zero).

3.2. *A better variance estimator for y_O*

When claiming that y_O is an unbiased estimator of Y_O under the parametric model in Section 2.2, we made no assumptions about the distributions of the y_{gi} apart from their having a common mean within groups. The estimation of the parametric model variance of y_O, however, requires additional assumptions. Assume, for now, that the y_{gi} values for owners from different firms are uncorrelated under the parametric model. This means that the model expectations of B' and D' in equation (3) are approximately zero. They would be exactly zero if each e_{gi} were equal to $a_{gi}[y_{gi} - E(y_{gi})]/N$, but this equality is itself only approximate.

The quality of v' in equation (3) as an estimator for the parametric model variance of y_O is suspect when the f_h are not negligible (see Särndal et al. 1989, p. 535). This is the case for some design strata in many business surveys including the 1987 CBO. As a result, the following alternative variance estimator is proposed

$$v_O = A_O - B_O - C_O \tag{4}$$

where

$$A_O = \sum_{g=1}^{G}(r_{Fg}/[r_{Fg}-1])\sum_{h=1}^{H}\sum_{j\in S_h} u_{Ojg}^2$$

$$C_O = \sum_{g=1}^{G}(r_{Fg}/[r_{Fg}-1])\sum_{h=1}^{H} f_h \sum_{j\in S_h} u_{Ojg}^2$$

$$-\sum_{g=1}^{G}(r_{Og}/[r_{Og}-1])$$

$$\times \sum_{i=1}^{r_{Og}}(1/w_{gi}-1/a_{Ogi})e_{Ogi}^2.$$

Here, r_{Fg} is the number of firms represented by respondent owners in group g, $e_{Ogi} = a_{Ogi}[y_{gi} - \sum^{r_{Og}}(a_{Ogk}y_{gk})/\Sigma a_{Ogk}]/N_O$, and u_{Ojg} is the sum of e_{Ogi} taken over all respondent owners in both sampled firm j and group g; $B_O = B'$ in equation (3), where u_{jg} denotes u_{Ojg}. The term D' in equation (3) has been dropped because it is small compared to the others (since all the r_{Og} are large).

The justification for v_O follows. Let us assume a parametric model for owner characteristics in which the y_{gi} are uncorrelated random variables with mean p_{Og} and variance $v_{Og} = p_{Og}(1-p_{Og})$. This is a more restrictive model than that discussed above since the y_{gi} are now uncorrelated within firms. The additional assumption is needed to support a finite population correction term, in this case, C_O.

Let $\delta_{gi} = y_{gi} - p_{Og}$, so that $\mathrm{Var}_M(\delta_{gi}) = v_{Og}$. The model variance of y_O as an estimator for Y_O can be expressed as

$$E_M[(y_O - Y_O)^2] = N_O^{-2}$$

$$\times \sum_{g=1}^{G} E_M\left[\left(\sum_{i=1}^{r_{Og}} a_{Ogi}y_{gi} - \sum_{i=1}^{N_{Og}} y_{gi}\right)^2\right]$$

$$= N_O^{-2}$$

$$\times \sum^{G}\left[\left(\sum^{r_{Og}} a_{Ogi}\delta_{gi} - \sum^{N_{Og}}\delta_{gi}\right)^2\right]$$

$$= N_O^{-2}$$

$$\times \sum^{G} v_{Og}\left(\sum^{r_{Og}} a_{Ogi}^2 - 2\sum^{r_{Og}} a_{Ogi} + N_{Og}\right)$$

$$= N_O^{-2}\sum^{G} v_{Og}\left(\sum^{r_{Og}} a_{Ogi}^2 - N_{Og}\right). \quad (5)$$

Consider the expression

$$e_{Ogi} = (a_{Ogi}/N_O)\left[y_{gi} - \sum^{r_{Og}}(a_{Ogk}y_{gk})/\sum a_{Ogk}\right]$$

$$= (a_{Ogi}/N_O)\left[\delta_{gi} - \sum a_{Ogk}\delta_{gk}/\sum a_{Ogk}\right]$$

for owner i in group g. It is not difficult to show that $E_M(e_{Ogi}^2) \approx a_{Ogi}^2 v_{Og}/N_O^2$. It is now a straightforward, if tedious, exercise to show that v_O in equation (4) is a nearly model unbiased estimator of the right hand side of (5); that is, when terms like $1 + O(1/r_{Fg})$ are approximately unity.

Strictly speaking, the expressions of the forms $r_{Og}/(r_{Og}-1)$ and $r_{Fg}/(r_{Fg}-1)$ within v_O are themselves approximately 1. They have not been rounded as ad hoc compensations for the ignorable negative bias in the e_{Ogi}^2 and u_{Ojg}^2 as estimators for the $a_{Ogi}^2 v_{Og}/N_O^2$ and $\Sigma_{gi\in j} a_{Ogi}v_{Og}/N_O^2$, respectively.

The near parametric model unbiasedness of v_O in equation (4) depends on terms like $1 + O(1/r_{Fg})$ being approximately 1. By contrast, the approximate quasi-design unbiasedness of v' in equation (3) depends on terms like $1 + O(r_{Fg}^{-1/2})$ being approximately 1, a tighter requirement on the size of r_{Fg}.

It is easy to see that the ratio of v_O and v' is itself approximately 1 when each $1 + O_p(r_{Fg}^{-1/2})$ is approximately unity (the

subscript p as before refers to the quasi-design probability space). Thus, v_O is also an approximately quasi-design unbiased estimator of the quasi-design mean squared error of y_O.

It is interesting to note that if all the f_h in v_O were set equal to zero, then v_O would be virtually identical to the post-stratified estimator computed by SUDAAN (Shah, Barnwell, Hunt, and Lavange 1991) when the first-phase sample is selected with replacement (the only distinction comes from the ad hoc $r_{Fg}/(r_{Fg}-1)$ terms in A_O).

3.3. *A better variance estimator for y_F*

In the parametric model for owner characteristics given above, we assumed that the y_{gi} were independent. Obviously, the y_{gi} will not be independent across owners when y_{gi} denotes a firm characteristic. In fact, the y_{gi} will be perfectly correlated within firms (if we ignore measurement error). For that reason, the following variance estimator for y_F is proposed

$$v_F = A_F - B_F - C_F \tag{6}$$

where

$$A_F = \sum_{g=1}^{G}(r_{Fg}/[r_{Fg}-1])\sum_{h=1}^{H}\sum_{j\in S_h} u_{Fjg}^2$$

$$C_F = \sum_{g=1}^{G}(r_{Fg}/[r_{Fg}-1])\sum_{h=1}^{H} f_h \sum_{j\in S_h} u_{Fjg}^2$$

$$-\sum_{g=1}^{G}(r_{Og}/[r_{Og}-1])$$

$$\times \sum_{i=1}^{r_{Og}}(r_{Ogi}/w_{gi} - 1/a_{Fgi})e_{Fig}^2$$

and $e_{Fig} = a_{Fgi}[y_{gi} - \sum^{r_{Og}}(a_{Fgk}y_{gk})/\Sigma a_{Fgk}]/N_F$. Here, u_{Fjg} is the sum of e_{Fig} taken over all respondent owners in both sampled firm j and group g, and r_{Ogi} is the number of respondent owners in the same firm as gi; $B_F = B'$ in equation (3), where now u_{jg} denotes u_{Fjg}.

The justification for v_F follows. Let us assume that a model for characteristics in which the y_{gi} are random variables with mean p_{Fg} and variance $v_{Fg} = p_{Fg}(1-p_{Fg})$. By definition, the value of y_{gi} is the same for all the owners of the same firm. We add the assumption that the y_{gi} are independent across firms.

Let $\theta_{gi} = y_{gi} - p_{Fg}$, so that $\mathrm{Var}_M(\theta_i) = v_{Fg}$. The model variance of y_F as an estimator for Y_F can be expressed as

$$E_M[(y_F - Y_F)^2] = \sum_{g=1}^{G} E_M\left[\left(\sum_{i=1}^{r_{Og}} a_{Fgi}y_{gi} - \sum_{i=1}^{N_{Og}}\{y_{gi}/n_{Ogi}\}\right)^2\right] \Big/ N_F^2$$

$$= \sum^{G} E_M\left[\left(\sum^{r_{Og}} a_{Fgi}\theta_{gi} - \sum^{N_{Og}}\{\theta_{gi}/n_{Ogi}\}\right)^2\right] \Big/ N_F^2$$

$$= \sum^{G} v_{Fg}\left(\sum^{r_{Og}} r_{Ogi}a_{Fgi}^2 - 2\sum^{r_{Og}} a_{Fgi} + N_{Fg}\right) \Big/ N_F^2$$

$$= \sum v_{Fg}\left(\sum^{r_{Og}} r_{Ogi}a_{Fgi}^2 - N_{Fg}\right) \Big/ N_F^2. \tag{7}$$

It is again a straightforward, and tedious, exercise to show that v_F in equation (6) is a nearly model unbiased estimator of the right hand side of (7); that is, when terms of order $1/r_{Fg}$ are ignored. There are also analogous ad hoc compensations for the ignorable negative biases of the e_{Fgi}^2 and u_{Fjg}^2.

The approximate quasi-design unbiasedness of v_F is harder to establish. It relies

on the fact that each

$$Q_g = \sum_{i=1}^{r_{Og}} (r_{Ogi}/w_{gi} - 1/a_{Fgi}) / \sum_{i=1}^{r_{Og}} (1/w_{gi})(1 - r_{Og}/n_{Og}) = 1 + O_p(r_{Fg}^{-1/2}),$$

where the numerator of Q_g comes from C_F in equation (6) and the denominator from C' in (3). As a result, the v_F has the same approximate quasi-design bias as v'.

To see that $Q_g = 1 + O_p(r_{Fg}^{-1/2})$, rewrite Q_q as

$$Q_q = \sum^{r_{Fg}} (\tilde{r}_{Oj}^2/\tilde{w}_j - \tilde{r}_{Oj}/\tilde{a}_{Fj}) / \sum (\tilde{r}_{Oj}/\tilde{w}_j)(1 - r_{Og}/n_{Og}),$$

where $\tilde{r}_{Oj}$, $\tilde{w}_j$, and $\tilde{a}_{Fj}$ have been defined in the obvious way (for example, $\tilde{w}_j = w_{gi}$, where gi is any owner in firm j). Let $m_g = r_{Og}/n_{Og}$. The quasi-design expectations of $\tilde{r}_{Oj}$ and $\tilde{r}_{Oj}^2$ are $\tilde{n}_{Oj} m_g$ and $(\tilde{n}_{Oj} m_g)^2 + \tilde{n}_{Oj} m_g (1 - m_g)$, respectively, while $1/\tilde{a}_{Fj}$ is approximately $\tilde{n}_{Oj}/\tilde{q}_j = (1/\tilde{w}_j)\tilde{n}_{Oj} m_g$. As a consequence, both the numerator and denominator of Q_g have the same quasi-design expectation. Both also have relative quasi-design variances of order $O_p(r_{Fg}^{-1/2})$. Thus, $Q_g = 1 + O_p(r_{Fg}^{-1/2})$.

It is again interesting to note that if all the f_h in v_F were set equal to zero, then v_F would be virtually identical to the post-stratified estimator computed by SUDAAN (Shah et al. 1991) when the first-phase sample is selected with replacement.

4. Domain Estimation

The U.S. Census Bureau is interested in the proportion of owners (or firms) with a particular survey characteristic within various domains. Mathematically, a domain proportion for an owner characteristic has the form

$$Y_{O(d)} = \frac{\sum^{G} \sum^{N_{Og}} y_{gi} d_{gi}}{\sum \sum d_{gi}}$$

where $d_{gi} = 1$ if owner gi is in domain d and zero otherwise.

The estimator for $Y_{O(d)}$ used by the U.S. Census Bureau is

$$y_{O(d)} = \frac{\sum^{G} \sum^{r_{Og}} a_{Ogi} y_{gi} d_{gi}}{\sum \sum a_{Ogi} d_{gi}}. \quad (8)$$

The population size of each response homogeneity group g within certain domains of interest, like industries, is either N_{Og} or zero. For other domains of interest, like employment classes, this size can be variable. This estimator in equation (8) does not require knowledge – unavailable to the U.S. Census Bureau – about the population sizes of the response homogeneity groups within domain d.

It is easy to see that the numerator and denominator of $y_{O(d)}$ are, respectively, nearly design unbiased estimators for the numerator and denominator of $Y_{O(d)}$ under the quasi-randomization model. For certain domains, like industries, the denominator of equation (8) will be a constant; for others, like employment classes, a random variable.

Similarly, if we assume that d_{gi} and $y_{gi(d)} = y_{gi} d_{gi}$, respectively, can be treated as random variables with a constant mean within groups, then the numerator and denominator of $y_{O(d)}$ are unbiased estimators for the numerator and denominator of $Y_{O(d)}$ under the parametric model. Strictly speaking, $y_{O(d)}$ is not exactly an unbiased estimator for $Y_{O(d)}$ under the parametric model, but in the context of the CBO where all domain estimates of interest are based on relatively large respondent samples, the potential for bias can be ignored for all practical purposes.

Observe that

$$y_{O(d)} - Y_{O(d)} = \sum\sum^{G}\sum^{r_{Og}}\{a_{Ogi}(y_{gi(d)} - Y_{O(d)}d_{gi})\}/\sum\sum a_{Ogi}d_{gi}$$
$$= \sum\sum a_{Ogi}y'_{gi}/\sum\sum a_{Ogi}d_{gi},$$

where

$$y'_{gi} = y_{gi(d)} - Y_{O(d)}d_{gi} \approx y_{gi(d)} - y_{O(d)}d_{gi}.$$

This suggests that v_O in equation (4) can serve as an estimator for parametric model variance and quasi-design mean squared error of $y_{O(d)}$ as an estimator for $Y_{O(d)}$, if we redefine e_{Ogi} as

$$e_{Ogi} = \left(\sum_{g'=1}^{G}\sum_{k=1}^{r_{Og'}} a_{Og'k}d_{g'k}\right)^{-1} a_{Ogi}$$
$$\times\left[\left(y_{gi(d)} - \sum_{k=1}^{r_{Og}} a_{Ogk}y_{gk(d)}/\sum_{k=1}^{r_{Og}} a_{Ogk}\right)\right.$$
$$\left. - y_{O(d)}\left(d_{gi} - \sum_{k=1}^{r_{Og}} a_{Ogk}d_k/\sum_{k=1}^{r_{Og}} a_{Ogk}\right)\right]. \quad (9)$$

It is a trivial matter to extend the argument made in this section to domain estimators of firm characteristics. For example, we can redefine the e_{Fgi} using equation (9) by replacing all the a_{Ogk} with a_{Fgk} and $y_{O(d)}$ with $y_{F(d)}$. The r_{Og} are (as always) unchanged.

5. Empirical Results

The U.S. Census Bureau was interested in determining whether there were appreciable finite population and design stratum effects (defined below) in the CBO data. If there were not, a simplified variance estimator could be used operationally.

To study this matter while containing costs, eight test variables were created based on meaningful combinations of actual CBO survey variables (e.g., the seven levels of education variables were combined into a single 0/1 variable). The research data set was composed of one randomly selected record per firm, and so the analysis was restricted to the estimated components of variance in equation (6). The term A_F in that equation is an estimator for the parametric variance assuming that the population of interest is infinite, while C_F is a measure of the finite population effect on the variance.

The term B_F is a measure of the *design stratum effect*; i.e, it captures the effect on variance of any tendency for the e_{Fgi} values of firms within the same design stratum to be similar. Remember, "group effects," have already been removed from the e_{Fgi}.

To compare alternative variance estimators, z and w (say), we use the measure $(1/2)\log(z/w)$. This measure is symmetric; i.e., $(1/2)\log(z/w) = -(1/2)\log(w/z)$. Moreover, it directly relates the implied standard error estimates, $z^{1/2}$ and $w^{1/2}$, since $(1/2)\log(z/w) = \log[(z/w)^{1/2}]$. Observe that $(1/2)\log(z/w)$ is approximately the percentage difference between the two implied standard error estimators (since $(1/2)\log(z/w) \approx (z^{1/2} - w^{1/2})/w^{1/2}$).

The two groups of numbers in Table 1 demonstrate how small the finite population and design stratum effects are on the variance estimators for the panel. The lack of an appreciable design stratum effect carries over to all domain estimators of interest to the Census Bureau.

There are also no appreciable finite population effects in most of the domains investigated. The only exceptions occur in the three (of ten) largest receipt size classes and three (of nine) largest employment size classes for a number of panels. This makes sense because the sampling fractions were often noticeably large in these domains (1/2 or greater).

Table 2 displays the finite population effects on the three largest receipt size

Table 1. The relative effects on standard errors of design stratification and finite population correction

Test variable	The design stratum effect[1]				
	Panel				
	Asian	Black	Hispanic	Women	Other
1	0.0066	0.0019	0.0001	0.0118	0.0060
2	–0.0008	–0.0020	–0.0013	–0.0056	0.0018
3	–0.0070	0.0004	0.0005	–0.0017	0.0031
4	–0.0027	–0.0006	0.0017	–0.0029	0.0006
5	–0.0075	–0.0045	–0.0023	0.0012	–0.0049
6	–0.0075	0.0007	0.0013	0.0016	0.0042
7	0.0047	0.0006	0.0012	0.0031	–0.0005
8	0.0057	–0.0003	0.0003	–0.0007	–0.0061
Average	–0.0010	–0.0005	0.0002	0.0009	0.0005
	The combined design stratum and finite population effect[2]				
Test variable	Panel				
	Asian	Black	Hispanic	Women	Other
1	0.0041	–0.0087	–0.0078	0.0114	0.0059
2	–0.0032	–0.0114	–0.0080	–0.0059	0.0016
3	–0.0093	–0.0085	–0.0057	–0.0020	0.0029
4	–0.0050	–0.0107	–0.0057	–0.0032	0.0004
5	–0.0096	–0.0133	–0.0085	0.0010	–0.0050
6	–0.0097	–0.0082	–0.0049	0.0012	0.0041
7	0.0020	–0.0111	–0.0078	0.0027	–0.0008
8	0.0036	–0.0091	–0.0059	–0.0010	–0.0063
Average	–0.0034	–0.0101	–0.0068	0.0005	0.0003

[1] $(1/2)\log\{(A_F - B_F)/A_F\}$.
[2] $(1/2)\log\{(A_F - B_F - C_F)/A_F\}$.

classes for four of the panels. The effect never exceeds 0.005 (in absolute value) for the sparsely sampled Other panel, which is not displayed in the table. Notice that for a particular domain and panel the finite population effect is fairly stable across variables.

It is important to realize that the properties of the 1987 CBO survey revealed in this section need not be shared by other surveys. That is to say, it is not universally the case that stratum effects do not matter and that finite population corrections rarely do. Nevertheless, survey statisticians facing similar questions about their surveys may want to use some of the methods discussed here in determining their approach to variance estimation.

6. Concluding Remarks

The principle contribution of this paper is in variance estimation for certain post-stratified estimators. The weighted residual method of Särndal et al. (1989) was the inspiration. That method can produce an estimator for parametric model variance and quasi-design mean squared error simultaneously. Its parametric model property, however, relies on having small sampling fractions in all design strata.

*Table 2. The relative effects on standard errors of finite population correction in the three highest receipt classes**

Test variable	Receipt class	Panel			
		Asian	Black	Hispanic	Women
1	250 K–499 K	–0.0217	–0.1016	–0.0726	–0.0033
1	500 K–999 K	–0.0454	–0.2343	–0.1685	–0.0072
1	1,000 K+	–0.1125	–0.5464	–0.2817	–0.0153
2	250 K–499 K	–0.0215	–0.1022	–0.0736	–0.0032
2	500 K–999 K	–0.0470	–0.2324	–0.1671	–0.0070
2	1,000 K+	–0.1134	–0.5419	–0.2714	–0.0164
3	250 K–499 K	–0.0213	–0.1017	–0.0717	–0.0030
3	500 K–999 K	–0.0470	–0.2288	–0.1654	–0.0067
3	1,000 K+	–0.1156	–0.5410	–0.2776	–0.0167
4	250 K–499 K	–0.0221	–0.1023	–0.0725	–0.0033
4	500 K–999 K	–0.0509	–0.2341	–0.1660	–0.0069
4	1,000 K+	–0.1198	–0.5454	–0.2766	–0.0163
5	250 K–499 K	–0.0215	–0.1011	–0.0723	–0.0034
5	500 K–999 K	–0.0454	–0.2299	–0.1652	–0.0071
5	1,000 K+	–0.1183	–0.5439	–0.2701	–0.0170
6	250 K–499 K	–0.0210	–0.1017	–0.0718	–0.0032
6	500 K–999 K	–0.0475	–0.2334	–0.1653	–0.0069
6	1,000 K+	–0.1107	–0.5488	–0.2742	–0.0165
7	250 K–499 K	–0.0206	–0.1029	–0.0728	–0.0034
7	500 K–999 K	–0.0471	–0.2322	–0.1647	–0.0067
7	1,000 K+	–0.1097	–0.5326	–0.2704	–0.0171
8	250 K–499 K	–0.0217	–0.1016	–0.0720	–0.0033
8	500 K–999 K	–0.0454	–0.2345	–0.1653	–0.0069
8	1,000 K+	–0.1160	–0.5488	–0.2741	–0.0161
Average	250 K–499 K	–0.0214	–0.1019	–0.0724	–0.0033
Average	500 K–999 K	–0.0470	–0.2325	–0.1659	–0.0069
Average	1,000 K+	–0.1145	–0.5436	–0.2745	–0.0164

*$(1/2)\log\{(A_F - C_F)/A_F\}$.

Adjustments of weighted residual variance estimators were proposed to compensate for potentially large sampling fractions in some design strata. Moreover, the implicit parametric model for a commonly used domain estimator was given explicit expression.

One interesting observation was made in the text: In the absence of appreciable first-phase finite population corrections, the variance estimators introduced here are virtually identical to the post-stratified variance estimators computed by SUDAAN, which does not normally handle two-phase designs.

7. References

Kott, P.S. (1990). Variance Estimation When a First Phase Area Sample Is Restratified. Survey Methodology, 16, 99–103.

Kott, P.S. (1994). A Note on Handling Nonresponse in Sample Surveys. Journal of the American Statistical Association, 89, 693–696.

Oh, H.L. and Scheuren, F.J. (1983). Weighting Adjustment for Unit Nonresponse. In Incomplete Data and Sample Surveys, Volume 2: Theory and Bibliographies, W.G. Madow, I. Olkin, and D.B. Rubin (eds.). New York: Academic Press.

Särndal, C.E., Swensson, B., and Wretman, J.H. (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of a Finite Population Total. Biometrika, 76, 527–538.

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

Shah, B.V., Barnwell, B.G., Hunt, P.N., and Lavange, L.M. (1991). SUDAAN™ User's Manual, Technical Appendix. Research Triangle Park, NC: Research Triangle Institute.

Received January 1993
Revised July 1994