Survey Methodology, December 1994
Vol. 20, No. 2, pp. 97–105
Statistics Canada

Generalized Regression Estimation for a Two-Phase Sample of Tax Records

JOHN ARMSTRONG and HÉLÈNE ST-JEAN[1]

ABSTRACT

A generalized regression estimator for domains and an approximate estimator of its variance are derived under two-phase sampling for stratification with Poisson selection at each phase. The derivations represent an application of the general framework for regression estimation for two-phase sampling developed by Särndal and Swensson (1987) and Särndal, Swensson and Wretman (1992). The empirical efficiency of the generalized regression estimator is examined using data from Statistics Canada's annual two-phase sample of tax records. Three particular cases of the generalized regression estimator – two regression estimators and a poststratified estimator – are compared to the Horvitz-Thompson estimator.

KEY WORDS: Model assisted estimation; Domain estimation; Poisson sampling.

1. INTRODUCTION

In this paper the problem of domain estimation under two-phase sampling for stratification is examined in a case in which Poisson sampling is used at both phases of selection. Consider a population of N units and suppose that it is necessary to estimate the total of a characteristic of interest, y, for L disjoint domains. Domain membership can be well, but not exactly, predicted using an auxiliary variable, Θ, that is not observed before sampling. The cost of obtaining information on Θ is lower than the cost of obtaining information on y and lower than the cost of obtaining exact domain membership data. At the first phase of sampling, a Poisson sample is drawn from the population and the value of Θ is observed for each sampled unit. The units in the first-phase sample are stratified using Θ-values. This stratification is an approximation to stratification by domain. At the second phase of sampling, a Poisson sample is drawn from each stratum. The value of y, as well as exact domain membership data, is observed for each unit in the second-phase sample.

The Horvitz-Thompson estimator of the total of y for domain d is $\hat{Y}_{H-T}(d) = \sum_{i \in s2} y_i(d)/(p_{1i}p_{2i})$, where $y_i(d)$ takes the value of y_i if unit i falls in domain d and otherwise takes the value zero, $s2$ denotes the second-phase sample and p_{1i} and p_{2i} are first- and second-phase selection probabilities, respectively, for unit i. Since the sample sizes obtained using Poisson sampling are random variables, this estimator may be inefficient. (See Sunter 1986 or Särndal, Swensson and Wretman 1992, p. 63.) Generalized regression estimation is an alternative to the Horvitz-Thompson estimator that can be employed when auxiliary information is available. A generalized regression estimator for two-phase Poisson sampling and an approximate estimator of its variance are derived in this paper.

Section 2 contains the derivation of the generalized regression estimator and approximate variance estimator. Section 3 includes a description of the application that motivated the estimation problem – Statistics Canada's annual two-phase sample of tax records. The results of an empirical study comparing the Horvitz-Thompson estimator with three particular cases of the generalized regression estimator – the poststratified estimator currently used in production and two regression estimators – are described in Section 4.

2. GENERALIZED REGRESSION ESTIMATION

Generalized regression estimation is not a new technique. A generalized regression estimator for a one-phase sample design is described by Deming and Stephan (1940). Recent applications of generalized regression estimation at Statistics Canada include the work of Lemaître and Dufour (1987) and Bankier, Rathwell and Majkowski (1992). Hidiroglou, Särndal and Binder (1993) provide an extensive discussion of the use of generalized regression estimators for business surveys.

Derivation of generalized regression estimators can be approached from the perspective of model assisted survey sampling (Särndal, Swensson and Wretman 1992) or from the perspective of calibration (Deville and Särndal 1992). Let $U = \{u\}$ and $V = \{v\}$ denote sets of first-phase poststrata and second-phase poststrata, respectively. During generalized regression weighting of the first-phase sample, the design weights $1/p_{1i}$ are adjusted to yield weights $w_{1i} = g_{1i}/p_{1i}$ that respect the calibration equations

[1] John Armstrong, Social and Economic Studies Division, 24 – R.H. Coats Bldg., and Hélène St-Jean, Business Survey Methods Division, 11 – R.H. Coats Bldg., Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

$$\sum_{i \in s1 \cap v} w_{1i} x_i = X_u,$$

for each first-phase poststratum u, where x_i is an $L_1 \times 1$ vector of auxiliary variables known for all units in the population and X_u is the vector of auxiliary variable totals for poststratum u. The adjusted weights minimize the distance measure $\sum_{i \in s1} (g_{1i} - 1)^2 / p_{1i}$. The same weights can be obtained from a model assisted perspective using

$$E_\xi(y_i) = x_i' \beta_u, \; i \in u$$
$$V_\xi(y_i) = \sigma^2,$$

where y_i is the value of the variable of interest for unit i, and $E_\xi(\cdot)$ and $V_\xi(\cdot)$ denote expectation and variance, respectively, with respect to the model.

For the generalized regression estimators of interest, weighting of the second-phase sample involves a calibration procedure that is conditional on the results of first-phase weighting. The initial weights, w_{1i}/p_{2i}, are adjusted to give final weights, $w_i = g_{2i} w_{1i} / p_{2i}$, that satisfy the calibration equations

$$\sum_{i \in s2 \cap v} w_i z_i = \tilde{Z}_v,$$

for each second-phase poststratum v, where z_i is an $L_2 \times 1$ vector of auxiliary variables known for all units in the first-phase sample and $\tilde{Z}_v = \sum_{i \in s1 \cap v} w_{1i} z_i$ is an estimate of the vector of auxiliary variable totals for post-stratum v, computed using the adjusted first-phase weights w_{1i}. Note that these calibration equations differ in an important way from the examples given by Särndal and Swensson (1987, pp. 284-288) and Särndal, Swensson and Wretman (1992, pp. 359-366) because they involve adjusted first-phase weights rather than first-phase design weights. The final weights minimize the distance measure $\sum_{i \in s2} w_{1i} (g_{2i} - 1)^2 / p_{2i}$. The model needed to obtain the same weights from a model assisted perspective is

$$E_\xi(w_{1i} y_i) = w_{1i} z_i' \beta_v, \; i \in v$$
$$V_\xi(w_{1i} y_i) = w_{1i} \sigma^2.$$

Use of adjusted first-phase weights rather than first-phase design weights in the second-phase calibration equations has two important advantages. First, the generalized regression estimator for domain d can be written as

$$\hat{Y}_{\mathrm{GREG}}(d) = \sum_{i \in s2} y_i(d) g_{1i} g_{2i} / p_{1i} p_{2i},$$

using first-phase and second-phase g-weights. Second, suppose that some auxiliary variables are used for calibration at both phases of weighting. Estimates of population totals for such variables that are equal to actual totals can be constructed using final weights.

Let $\check{X}_u = \sum_{i \in s1 \cap u} x_i / p_{1i}$ denote the $L_1 \times 1$ vector of Horvitz-Thompson estimates of auxiliary variable totals for first-phase poststratum u. The first-phase g-weight is

$$g_{1i} = 1 + \lambda_u' x_i,$$

where $\lambda_u' = (X_u - \check{X}_u)' M_u^{-1}$ and $M_u^{-1} = (\sum_{i \in s1 \cap u} x_i x_i' / p_{1i})^{-1}$. For second-phase poststratum v, denote the estimate of $\tilde{Z}_v$ based on initial second-phase weights by $\dot{Z}_v = \sum_{i \in s2 \cap v} w_{1i} z_i / p_{2i}$. The second-phase g-weight is

$$g_{2i} = 1 + \lambda_v' z_i,$$

where $\lambda_v' = (\tilde{Z}_v - \dot{Z}_v)' M_v^{-1}$ and $M_v^{-1} = (\sum_{i \in s2 \cap v} w_{1i} z_i z_i' / p_{2i})^{-1}$.

The approximate variance of $\hat{Y}_{\mathrm{GREG}}(d)$ is given by

$$V(\hat{Y}_{\mathrm{GREG}}(d)) \approx \sum_i \frac{1 - p_{1i}}{p_{1i}} Q_{1i}^2 + E_1 \left[\sum_{i \in s2} \frac{1 - p_{2i}}{p_{2i}} (w_{1i} Q_{2i})^2 \right],$$

where $E_1(\cdot)$ denotes expectation with respect to the first phase of sampling, $Q_{1i} = y_i(d) - x_i' B_u$ for each unit in first-phase poststratum u, and B_u, the vector of estimated coefficients from the regression of $y(d)$ on x that would be obtained if $y(d)$ was available for all units in first-phase poststratum u, is given by

$$B_u = \left(\sum_{i \in u} x_i x_i' \right)^{-1} \left(\sum_{i \in u} x_i y_i(d) \right).$$

Similarly, $Q_{2i} = y_i(d) - z_i' B_v$ for each unit in second-phase poststratum v and B_v, the vector of estimated coefficients from the regression of $y(d)$ on z that would be obtained, conditional on the first-phase calibration, if $y(d)$ was available for all units in the component of the first-phase sample falling in second-phase poststratum v, is given by

$$B_v = \left(\sum_{i \in s1 \cap v} w_{1i} z_i z_i' \right)^{-1} \left(\sum_{i \in s1 \cap v} w_{1i} z_i y_i(d) \right).$$

An estimator of the approximate variance of $\hat{Y}_{\mathrm{GREG}}(d)$ is

$$\hat{V}(\hat{Y}_{\mathrm{GREG}}(d)) = \sum_i \frac{1 - p_{1i}}{p_{1i}^2 p_{2i}} (g_{1i} q_{1i})^2 + \sum_i \frac{1 - p_{2i}}{(p_{1i} p_{2i})^2} (g_{1i} g_{2i} q_{2i})^2.$$

Since $y(d)$ is available only for units in $s2$, estimates of B_u and B_v are

$$\hat{B}_u = \left(\sum_{i \in s2 \cap u} w_i x_i x_i' \right)^{-1} \left(\sum_{i \in s2 \cap u} w_i x_i y_i(d) \right),$$

$$\hat{B}_v = \left(\sum_{i \in s2 \cap v} w_i z_i z_i' \right)^{-1} \left(\sum_{i \in s2 \cap v} w_i z_i y_i(d) \right).$$

The sample residuals needed to compute the variance estimator are $q_{1i} = y_i(d) - x_i' \hat{B}_u$ and $q_{2i} = y_i(d) - z_i' \hat{B}_v$. More details of the derivation of the approximate variance of $\hat{Y}_{\text{GREG}}(d)$ and the estimator of the approximate variance are given in Appendix A.

If y is strongly correlated with x and z, the variance of the generalized regression estimator of the population total of y will be relatively small. However, it is important to note that strong correlations between y and x and z will not necessarily lead to a relatively small variance for the estimate of the total of y for a particular domain, since $y(d)$ may be poorly correlated with x and z within poststrata that include at least one sampled unit falling in domain d.

The correlation between $y(d)$ and x and z within a poststratum that includes at least one sampled unit falling in domain d may be low if some sampled units in the poststratum do not fall in domain d. This situation may arise often if domain totals of auxiliary variables and/or exact domain membership information for units in the first-phase sample are unavailable. In the context of two-phase sampling for stratification, there is no domain membership information available before selection of the first-phase sample. If each first-phase poststratum is formed by combining one or more first-phase sampling strata, for example, most first-phase poststrata will include more than one domain. The variable Θ used to predict domain membership during stratification of the first-phase sample is not an exact predictor. If second-phase poststrata are formed by combining second-phase sampling strata, each domain may be divided between a number of second-phase poststrata.

Depending on the type of auxiliary information used, the g-weights associated with the generalized regression estimator and, consequently, generalized regression estimates, may be negative.

3. APPLICATION: TWO-PHASE SAMPLING OF TAX RECORDS

The two-phase tax sample is part of a general strategy at Statistics Canada for production of annual estimates of Canadian economic activity. Annual economic data for large businesses are collected through mail-out sample surveys. Data for small businesses are obtained from the tax sample. Estimates of financial variables for the business population are obtained by combining tax and survey estimates. Tax data rather than survey data are used to obtain small business estimates in order to reduce costs and response burden.

The two-phase sample design was introduced in response to a requirement for estimates for domains defined using the four-digit Standard Industrial Classification (SIC) code (Statistics Canada 1980). The first two digits of SIC (SIC2) provides a classification of businesses activity into 76 groups. Within each group, four-digit SIC (SIC4) codes provide classification into finer categories. For example, the SIC2 code of a business might classify it in the transportation industry while the SIC4 code describes the activity of the business as bulk liquids trucking.

There are two types of taxfilers - T1s and T2s. A T1 taxfiler is an individual, who may own all or part of one or more unincorporated businesses, while a T2 taxfiler is an incorporated business. Administrative files that contain limited information for all taxfilers that are associated with businesses are provided to Statistics Canada by Revenue Canada, the Canadian government department responsible for tax collection. These files are used to construct a sampling frame. Information concerning numbers of businesses owned by T1 taxfilers and ownership shares is not available on the sampling frame. Frame data does include geographical information, as well as gross business income and net profit for both T1 and T2 taxfilers. A few other major financial variables, including salary and inventory data, are generally available for T2 taxfilers. Estimates are required for about 35 financial variables that can be obtained from tax returns and associated financial statements but are not available on administrative files supplied by Revenue Canada.

Taxfilers that are associated with businesses are classified by Revenue Canada using the SIC system. In most cases, descriptions of business activity reported on tax returns are sufficient to accurately determine SIC2 codes. Revenue Canada assigns additional digits of SIC to most taxfilers. However, not all taxfilers are classified to the four-digit level and the third and fourth digits of SIC4 codes assigned by Revenue Canada are relatively inaccurate. A two-phase approach to sampling of tax records was adopted to facilitate accurate estimation of economic production at the SIC4 level.

Section 3.1 includes a brief description of the two-phase sampling design. More information about the two-phase design is provided in Armstrong, Block and Srinath (1993). Sections 3.2 and 3.3 contain information concerning estimation for the two-phase design. The Horvitz-Thompson estimator is described in Section 3.2 and a poststratified estimator is discussed in Section 3.3.

3.1 Sampling Design

The administrative information used to construct the sampling frame for a particular tax year is accumulated by Revenue Canada over a period of two calendar years as tax returns are received and processed. The use of Poisson sampling offers substantial operational advantages because sampling operations can begin before a complete sampling frame is available.

The target (in-scope) population for tax sampling is the population of businesses with gross income over $25,000, excluding large businesses covered by mail-out sample surveys. The first-phase sample is a longitudinal sample of taxfilers. Strata are defined by SIC2, province and size (gross business income). All taxfilers that are included in the first-phase sample for tax year T and are still in-scope for tax sampling in tax year $T + 1$ remain in the first-phase sample for tax year $T + 1$. Taxfilers may be added to the first-phase sample each year to improve the precision of certain estimates and to replace taxfilers sampled in previous years that are no longer in-scope.

To implement Poisson sampling for first-phase sample selection, each taxfiler is assigned a pseudo-random number (hash number) in the interval (0,1) generated by a hashing function that uses the unique taxfiler identifier as input. The hash number for each taxfiler is compared to the sampling interval for the corresponding stratum. If the hash number for a particular taxfiler falls in the corresponding sampling interval and the taxfiler is not already in the first-phase sample, then the taxfiler is added to the first-phase sample. Since taxfiler identifiers do not change over time, Poisson sampling facilitates selection of a longitudinal first-phase sample.

First-phase selection probabilities for taxfilers that are already included in the first-phase sample are updated each year. Longitudinal updating is necessary because: (i) a taxfiler may fall in different first-phase sampling strata in consecutive tax years; and (ii) first-phase sampling fractions for a given stratum may vary from one year to the next.

Copies of tax returns and associated financial statements for taxfilers in the first-phase sample are sent to Statistics Canada from Revenue Canada. In order to select the second-phase sample, statistical entities are created using information about businesses corresponding to taxfilers in the first-phase sample. Let $J = \{j\}$ denote the population of businesses that is the target population for tax sampling. A statistical entity, denoted by (i,j), is created for every taxfiler-business combination in the first-phase sample. For each T1 taxfiler in the first-phase sample, data for all businesses wholly or partially owned by the taxfiler (including ownership shares) that are needed to create statistical entities are available from tax returns and associated financial statements. Since there is a one-to-one correspondence between businesses and T2 taxfilers, a single statistical entity is created for each T2 taxfiler in the first-phase sample.

For each tax year, statistical entities that have not appeared in previous tax samples are assigned SIC4 codes by Statistics Canada. These codes are determined using information supplementary to business activity descriptions reported on tax returns and are more accurate in digits three and four than codes assigned by Revenue Canada. For statistical entities that have appeared in previous tax samples, the SIC4 assigned earlier is carried forward.

Conceptually, the second-phase sample is a sample of businesses. Operationally, it is a sample of taxfilers selected using statistical entities. Statistical entities are stratified using SIC4 codes assigned by Statistics Canada, as well as province and size. The total revenue of business j is used as the size variable for statistical entity (i,j). If one statistical entity corresponding to a T1 taxfiler is selected for the second-phase sample, then all statistical entities corresponding to the taxfiler are selected. Consequently, the second-phase selection probability for statistical entity (i,j) depends only on i.

Second-phase sample selection is done by the Poisson sampling method using hash numbers generated from taxfiler identifiers. The hashing function used for second-phase sample selection is independent of the first-phase hashing function.

Data for about 35 financial variables are transcribed from tax returns and associated financial statements for taxfilers selected in the second-phase sample. SIC4 codes assigned by Statistics Canada are updated, if necessary, to ensure that all SIC4 codes used during tabulation of estimates correspond to the current tax year.

3.2 Horvitz-Thompson Estimator

The second-phase sample is a sample of businesses selected using statistical entities. Since some businesses are partnerships, more than one statistical entity may correspond to the same business. To construct estimates for the population of businesses, an adjustment for the effects of partnerships is required. If business j is a partnership, it will be included in the second-phase sample if any of the corresponding taxfilers are selected. The usual Horvitz-Thompson estimator must be adjusted for partnerships to avoid over-estimation. Let δ_{ij} denote the proportion of business j owned by taxfiler i and suppose that statistical entity (i,j) is selected for the second-phase sample. The data for business j is adjusted by multiplying it by δ_{ij} so that only the component of income and expense items corresponding to taxfiler i is included in estimates. Rao (1968a) describes a similar adjustment in a slightly different context.

Let y_j denote the value of the variable y for business j. The Horvitz-Thompson estimate of the total of y over domain d, incorporating adjustment for partnerships, is given by

$$\hat{Y}_{H-T}(d) = \sum_{i \in s2} \sum_{j \in J_i} \delta_{ij} y_j(d) / (p_{1i} p_{2i}),$$

where J_i is a set containing the indices of the businesses wholly or partially owned by taxfiler i. Since selection probabilities depend only on the taxfiler index i, $\hat{Y}_{H-T}(d)$ can be written as

$$\hat{Y}_{H-T}(d) = \sum_{i \in s2} y_i(d) / (p_{1i} p_{2i}),$$

where

$$y_i(d) = \sum_{j \in J_i} \delta_{ij} y_j(d).$$

$\hat{Y}_{H-T}(d)$ is an unbiased estimator of the population total of y for businesses in domain d. Refer to Rao (1968a).

The second-phase sample is obtained by Poisson subsampling of the first-phase Poisson sample. Consequently, the second-phase sample is also a Poisson sample and the variance of $\hat{Y}_{H-T}(d)$ is

$$V(\hat{Y}_{H-T}(d)) = \sum_i \left[(1 - p_{1i} p_{2i}) / (p_{1i} p_{2i})\right] y_i(d)^2.$$

An unbaised estimator of this variance is

$$\hat{V}(\hat{Y}_{H-T}(d)) = \sum_{i \in s2} \left[(1 - p_{1i} p_{2i}) / (p_{1i} p_{2i})^2\right] y_i(d)^2.$$

3.3 Poststratified Horvitz-Thompson Estimator

Adjustment of the Horvitz-Thompson estimator to account for differences between actual and expected sample sizes under Poisson sampling was suggested by Brewer, Early and Joyce (1972). The methodology currently used to produce estimates based on the two-phase tax sample incorporates such adjustments.

Ratio adjustments are applied within poststrata during weighting of both the first- and second-phase samples. Choudhry, Lavallée and Hidiroglou (1989) provide a general discussion of weighting for a two-phase Poisson sample using poststratified ratio adjustments. Suppose that first-phase poststratum u contains N_u taxfilers. An estimate of the number of taxfilers in the population that fall in first-phase poststratum u, based on the first-phase sample, is

$$\check{N}_u = \sum_{i \in s1 \cap u} (1/p_{1i}).$$

The poststratified first-phase weight for taxfiler i, $i \in u$ is

$$w_{1i} = (1/p_{1i})(N_u/\check{N}_u).$$

An estimate of the number of taxfilers in second-phase poststratum v, based on the first-phase sample, is

$$\tilde{N}_v = \sum_{i \in s1 \cap v} w_{1i}.$$

An alternative estimate, using only units in the second-phase sample, is

$$\dot{N}_v = \sum_{i \in s2 \cap v} w_{1i}/p_{2i}.$$

The poststratified second-phase weight for statistical entity (i,j) in poststratum v is

$$w_{2i} = (1/p_{2i})(\tilde{N}_v/\dot{N}_v)$$

and the final weight is

$$w_i = w_{1i} w_{2i}.$$

The poststratified estimate of the total of y over domain d is

$$\hat{Y}(d) = \sum_{i \in s2} w_i y_i(d).$$

Choudhry, Lavallée and Hidiroglou (1989) note that the variance of $\hat{Y}(d)$ is approximately given by

$$V(\hat{Y}(d)) \approx \sum_u \sum_{i \in u} \frac{(1 - p_{1i})}{p_{1i}} \left(y_i(d) - \frac{Y_u(d)}{N_u}\right)^2 + \sum_v \sum_{i \in v} \frac{(1 - p_{2i})}{p_{1i} p_{2i}} \left(y_i(d) - \frac{Y_v(d)}{N_v}\right)^2,$$

where $Y_u(d)$ and $Y_v(d)$ are population totals for the variable y over the portions of the domain d belonging to poststrata u and v respectively.

This variance is estimated by

$$\hat{V}(\hat{Y}(d)) = \sum_u \sum_v \left(\frac{N_u}{\check{N}_u}\right)^2 \left(\frac{\tilde{N}_v}{\dot{N}_v}\right)^2 \sum_{i \in s2 \cap u \cap v} \frac{(1 - p_{1i})}{p_{1i}^2 p_{2i}} \left(y_i(d) - \frac{\hat{Y}_u(d)}{\hat{N}_u}\right)^2 + \sum_u \sum_v \left(\frac{N_u}{\check{N}_u}\right)^2 \left(\frac{\tilde{N}_v}{\dot{N}_v}\right)^2 \sum_{i \in s2 \cap u \cap v} \frac{(1 - p_{2i})}{(p_{1i} p_{2i})^2} \left(y_i(d) - \frac{\hat{Y}_v(d)}{\hat{N}_v}\right)^2,$$

where the estimates $\hat{N}_u$ and $\hat{N}_v$ are calculated using final weights.

The inclusion of the factor $(N_u/\check{N}_u)^2(\tilde{N}_v/\dot{N}_v)^2$ can be motivated by an improvement in the conditional properties of the estimator (Royall and Eberhardt 1975). A variance estimator for the ratio estimator for a one-phase sample design including an analogous adjustment factor has also been studied by Wu (1982). Empirical work reported by Wu and Deng (1983) indicates that the coverage properties of confidence intervals based on the normal approximation are improved using the adjustment factor.

$\hat{Y}(d)$ is a particular case of $\hat{Y}_{\text{GREG}}(d)$ that can be obtained if a single auxiliary variable with value one for all taxfilers is employed during both first-and second-phase weighting. In this case, we have $g_{1i} = N_u/\check{N}_u$ for all taxfilers in first-phase poststratum u and $g_{2i} = \tilde{N}_v/\dot{N}_v$ for all taxfilers in second-phase poststratum v. Note that negative g-weights are precluded by this choice of auxiliary variables. The variance estimator $\hat{V}(\hat{Y}(d))$ differs in a minor way from the estimator $\hat{V}(\hat{Y}_{\text{GREG}}(d))$ for this particular case of $\hat{Y}_{\text{GREG}}(d)$. The second-phase g-weight appears in the leading term of $\hat{V}(\hat{Y}(d))$ but does not appear in $\hat{V}(\hat{Y}_{\text{GREG}}(d))$.

4. EMPIRICAL STUDY

In order to compare the performance of $\hat{Y}_{H-T}(d)$, $\hat{Y}(d)$ and $\hat{Y}_{\text{GREG}}(d)$, an empirical study was conducted using data from the province of Quebec for tax year 1989. Since the estimator $\hat{Y}(d)$ is a special case of $\hat{Y}_{\text{GREG}}(d)$, it will be called $\hat{Y}_{\text{GREG-TPH}}(d)$ in subsequent discussion. (TPH is an abbreviation for two-phase Hájek.) Two other generalized regression estimators were considered. In both cases, x and z contains a variable with value one for all taxfilers. One generalized regression estimator involves calibration on taxfiler revenue during second-phase weighting. (Taxfiler revenue is included as a second auxiliary variable in z.) The second estimator involves calibration on taxfiler revenue at both phases of weighting. (Taxfiler revenue is included as a second auxiliary variable in both x and z.) Estimates of domain totals computed using these two estimators are denoted by $\hat{Y}_{\text{GREG-R2}}(d)$ and $\hat{Y}_{\text{GREG-R1R2}}(d)$, respectively, in subsequent discussion.

Estimates were produced for two variables of interest – transcribed revenue and total expenses. There are some conceptual differences between transcribed revenue and taxfiler revenue. For example, capital gains and extraordinary items are included in taxfiler revenue in many industries while they are excluded from transcribed revenue. In addition, taxfiler revenue contains more data capture errors than transcribed revenue since it is not subject to the same level of quality control.

The population used for the study included about 140,000 T2 taxfilers who reported over $25,000 in revenue for tax year 1989. The first- and second-phase selection probabilities used during sampling for production for tax year 1989 were employed. The first-phase sample included approximately 31,000 taxfilers and there were about 23,000 businesses in the second-phase sample. The correlation between taxfiler revenue and transcribed revenue for businesses in the second-phase sample was 0.969, while the correlation between taxfiler revenue and total expenses was 0.960.

Large proportions of units in the first- and second-phase samples were selected with certainty. All units with first-phase selection probability one were excluded from first-phase weighting and the corresponding g-weights were set to one. Units with second-phase selection probability one were treated analogously during second-phase weighting. There were 9,884 units in the first-phase sample with first-phase selection probabilities different from one and 910 units in the second-phase sample with second-phase selection probabilities different from one. Each first-phase poststratum consisted of one or more of the first-phase sampling strata used during sampling for 1989 production. These strata were defined using five revenue classes. All the sampling strata included in any particular first-phase poststratum corresponded to the same revenue class. Each first-phase poststratum contained a minimum of twenty sampled units. The use of a minimum sample size was motivated by concerns about the bias in $\hat{V}(\hat{Y}_{\text{GREG}}(d))$ when the number of sampled units used for estimation of regression coefficients is very small (Rao 1968b). If a first-phase sampling stratum included fewer than twenty sampled units, it was combined with sampling strata for similar SIC2 codes and the same revenue class until a poststratum containing at least twenty sampled units was obtained. Application of this procedure led to 166 first-phase poststrata. Second-phase poststrata were formed analogously, combining sampling strata for similar SIC4 codes to obtain a minimum sample size of twenty for each poststrata. There were 30 second-phase poststrata.

First and second-phase weights for $\hat{Y}_{\text{GREG-TPH}}(d)$, $\hat{Y}_{\text{GREG-R2}}(d)$ and $\hat{Y}_{\text{GREG-R1R2}}(d)$ were calculated using a modified version of the SAS macro CALMAR (Sautory 1991). The set of first-phase sampling weights calculated for the GREG-R1R2 estimator included twelve negative weights. There were no negative second-phase weights calculated for either GREG-R2 or GREG-R1R2. (Negative weights are not possible for the GREG-TPH estimator.) Estimates of transcribed revenue and total expenses were produced for 77 SIC2 domains, 256 SIC3 domains and 587 SIC4 domains using the three GREG estimators, as well as $\hat{Y}_{\text{H-T}}(d)$. Since GREG-R1R2 did not produce any negative estimates, no measures were taken to modify the negative weights associated with the estimator.

Results of comparisons of the GREG-TPH and H-T estimators are presented in Table 1 and Table 2. The mean gains and mean losses reported in the tables are averages of ratios of coefficients of variation. The GREG-TPH estimator performs better than the H-T estimator for the

Table 1

Comparison of GREG-TPH and H-T Estimators for Transcribed Revenue, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-TPH		Losses Using GREG-TPH	
	Number	Mean	Number	Mean
SIC2	57	0.768	20	1.113
SIC3	175	0.909	81	1.082
SIC4	359	0.945	228	1.079

Table 2

Comparison of GREG-TPH and H-T Estimators for Total Expenses, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-TPH		Losses Using GREG-TPH	
	Number	Mean	Number	Mean
SIC2	57	0.773	20	1.100
SIC3	175	0.910	81	1.082
SIC4	355	0.945	232	1.079

majority of domains. The gains obtained using GREG-TPH are particularly large for SIC2 domains. At the SIC4 level, the estimated coefficient of variation (CV) for the GREG-TPH estimate of total expenses is lower than the estimated CV for the H-T estimate for 60.5% of domains. In cases in which the estimated CV for GREG-TPH is lower it is 5.5% smaller, on average, than the estimated CV for H-T. When the estimated CV for GREG-TPH is higher it is 7.9% larger than the estimated CV for H-T, on average. In addition to the information in Tables 1 and 2, there is another reason to prefer GREG-TPH to H-T. Each year, tax return information for some sampled taxfilers is not received by Statistics Canada or is unusable because it does not include the necessary financial statements. Assuming that such cases of nonresponse are ignorable, the GREG-TPH estimator provides an automatic nonresponse adjustment.

The results in Tables 1 and 2 indicate that the relative performance of the GREG-TPH and H-T estimators are very similar for both variables of interest. The results of the other comparisons of estimators done as part of this empirical study did not depend on the variable of interest in any important way. Consequently, only results for total expenses are reported in subsequent tables.

The GREG-TPH estimator is compared to GREG-R2 and GREG-R1R2 in Tables 3 and 4. Based on estimated coefficients of variation, GREG-R2 performs slightly better than GREG-TPH. Since a large proportion of units in the second-phase tax sample have second-phase selection probability one and both GREG-R2 and GREG-TPH use the same auxiliary variables during first-phase weighting, the marginal differences between GREG-R2 and GREG-TPH are not surprising. Estimated CVs for GREG-R1R2 are generally smaller than estimated CVs for GREG-TPH and the relative performance of GREG-R1R2 improves as domain size increases. Nevertheless, GREG-R1R2 is superior to GREG-TPH for only 64% of SIC4 domains, and the average increase in estimated CVs for those domains in which GREG-R1R2 did worse than GREG-TPH is larger than the average decrease in estimated CVs for domains in which GREG-R1R2 performed better.

Table 3

Comparison of GREG-R2 and GREG-TPH Estimators for Total Expenses, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-R2		No Difference	Losses Using GREG-R2	
	Number	Mean	Number	Number	Mean
SIC2	38	0.993	26	13	1.001
SIC3	58	0.991	158	40	1.002
SIC4	88	0.988	439	60	1.009

Table 4

Comparison of GREG-R1R2 and GREG-TPH Estimators for Total Expenses, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-R1R2		Losses Using GREG-R1R2	
	Number	Mean	Number	Mean
SIC2	51	0.867	26	1.170
SIC3	160	0.934	96	1.093
SIC4	377	0.954	210	1.074

The results in Tables 3 and 4 indicate that, although the GREG-R1R2 estimator shows some promise, it would be inappropriate to completely replace the GREG-TPH estimator currently used in production by GREG-R1R2. The improvements obtained using GREG-R1R2 are relatively marginal, given the strong correlation between taxfiler revenue and total expenses. Larger improvements could be obtained if: (i) SIC codes used for first-and second-phase stratification were always consistent with SIC codes used to determine the domain membership of sampled units; and (ii) formation of first-and second-phase poststrata did not require combination of sampling strata to obtain a minimum sample size in each poststratum.

The results reported in Table 5 were obtained after SIC codes assigned to taxfilers by Revenue Canada and SIC codes used for stratification of the second-phase sample were changed for sampled units, where necessary, to eliminate inconsistencies between these codes and those

Table 5
Comparison of GREG-R1R2 and GREG-TPH Estimators for Total Expenses, Estimated Coefficients of Variation, No Misclassification

Type of Domain	Gains Using GREG-R1R2		Losses Using GREG-R1R2	
	Number	Mean	Number	Mean
SIC2	66	0.778	11	1.057
SIC3	184	0.916	72	1.047
SIC4	402	0.944	185	1.034

used to determine domain membership. A comparison of Tables 4 and 5 indicates that the relative performance of GREG-R1R2 is considerably better when there are no classification errors. GREG-R1R2 reduces estimated CVs by over 22% (on average) for over 85% of SIC2 domains.

Throughout the empirical results reported here, performance improvements obtained through the use of additional auxiliary information increase as domain size increases. This result is consistent with the observations in Section 2 concerning the conditions under which correlations between $y(d)$ and the vectors of auxiliary variables, x and z, will be high. Provided that the variable of interest and the auxiliary variables are highly correlated, correlations involving $y(d)$ will be strong if each poststratum containing at least one sampled unit falling in domain d also contains relatively few sampled units that do not fall in domain d.

5. CONCLUSIONS

Generalized regression estimation provides a convenient framework for the use of auxiliary information. A generalized regression estimator for a two-phase sample design with Poisson sampling at both phases of selection is derived in this paper. The efficiency of the estimator is investigated through application to the two-phase tax sample selected by Statistics Canada to obtain annual estimates of the economic activity of small businesses. The estimation method currently used in production for this survey incorporates poststratified ratio adjustments during both first-and second-phase weighting to compensate for differences between actual and expected sample sizes. This poststratified estimator is a particular case of the generalized regression estimator.

In an empirical study, the generalized regression estimator currently used in production (GREG-TPH) performs much better then the Horvitz-Thompson estimator. Two other generalized regression estimators are also compared to GREG-TPH. The alternative estimators produce improvements for large domains. However, their performance for the smaller domains that are of particular interest to users of estimates based on the two-phase tax sample does not justify complete replacement of the current production methodology.

ACKNOWLEDGEMENTS

The authors would like to thank René Boyer for providing a modified version of the SAS macro CALMAR suitable for the empirical study, as well as K.P. Srinath and Michael Hidiroglou for helpful discussions. Thanks are also due to Michael Bankier and Jean Leduc for helpful comments on a earlier draft of this paper.

APPENDIX A: DERIVATION OF VARIANCE OF $\hat{Y}_{GREG}(d)$ AND VARIANCE ESTIMATOR

The variance of $\hat{Y}_{\mathrm{GREG}}(d)$ can be derived using the identity

$$V(\hat{Y}_{\mathrm{GREG}}(d)) = E_1V_2(\hat{Y}_{\mathrm{GREG}}(d)) + V_1E_2(\hat{Y}_{\mathrm{GREG}}(d)).$$

First, consider the variance of the estimator with respect to the second phase of sampling, conditional on the results of first-phase calibration. If the vector of auxiliary variables for second-phase weighting, z, includes a variable with value one for all taxfilers (or a linear combination of auxiliary variables that is equal to one for all taxfilers can be constructed), the generalized regression estimator can be written as

$$\hat{Y}_{\mathrm{GREG}}(d) = \sum_{i\in s2} w_{1i}w_{2i}y_i(d)$$

$$= \sum_v \sum_{i\in s2\cap v} w_{1i}(y_i(d) - z_i'\hat{B}_v)/p_{2i} + \sum_v \bar{Z}_v\hat{B}_v.$$

Ignoring the variability due to the estimation of regression coefficients during second-phase weighting, we have

$$E_1V_2(\hat{Y}_{\mathrm{GREG}}) \approx E_1V_2\left(\sum_{i\in s2} w_{1i}Q_{2i}/p_{2i}\right)$$

$$= E_1\left(\sum_{i\in s1} \frac{(1-p_{2i})}{p_{2i}} w_{1i}^2Q_{2i}^2\right).$$

The estimator of $E_1V_2(\hat{Y}_{\mathrm{GREG}}(d))$ based on the variance estimator for calibration estimators advocated by Deville and Särndal (1992, p. 380) is

$$\hat{S}_1 = \sum_{i\in s2} \frac{(1-p_{2i})}{(p_{1i}p_{2i})^2} (g_{1i}g_{2i}q_{2i})^2.$$

Ignoring variability due to the estimation of regression coefficients during first-phase weighting, the second term in the variance expression can be written as

$$V_1 E_2(\hat{Y}_{\text{GREG}}(d)) = V_1\left(\sum_{i\in s1} w_{1i} y_i(d)\right)$$

$$= \sum_i \frac{(1 - p_{1i})}{p_{1i}} Q_{1i}^2.$$

An estimator of this term is

$$\hat{S}_2 = \sum_{i\in s2} \frac{(1 - p_{1i})}{p_{1i}^2 p_{2i}} (g_{1i} q_{1i})^2.$$

REFERENCES

ARMSTRONG, J., BLOCK, C., and SRINATH, K.P. (1993). Two-phase sampling of tax records for business surveys. *Journal of Business and Economic Statistics*, 11, 407-416.

BANKIER, M., RATHWELL, S., and MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census of Population. *Statistics Sweden, Workshop on the Uses of Auxiliary Information in Surveys.*

BREWER, K.R.W., EARLY, L.J., and JOYCE, S.F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 231-239.

CHOUDHRY, G.H., LAVALLÉE, P., and HIDIROGLOU, M. (1989). Two-phase sample design for tax data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 646-651.

DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 34, 911-934.

DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

HIDIROGLOU, M.A., SÄRNDAL, C.-E., and BINDER, D.A. (1993). Weighting and estimation in establishment surveys. Paper presented at the International Conference on Establishment Surveys, Buffalo, New York.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

RAO, J.N.K. (1968a). Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.

RAO, J.N.K. (1968b). Some small sample results in ratio and regression estimation. *Journal of the Indian Statistical Association*, 6, 160-168.

ROYALL, R.M., and EBERHARDT, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, Series C, 37, 43-52.

SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SAUTORY, O. (1991). La macro SAS: CALMAR. Unpublished manuscript, Institut national de la statistique et des études économiques, Paris.

STATISTICS CANADA (1980). *Standard Industrial Classification*. Catalogue No. 12-501E, Statistics Canada.

SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: A useful technique. *Journal of Official Statistics*, 2, 161-168.

WU, C.F.J. (1982). Estimation of variance of the ratio estimator. *Biometrika*, 69, 183-189.

WU, C.F.J., and DENG, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study. In Box, G.E.P. *et al.* (Eds.), *Scientific Inference, Data Analysis and Robustness*, New York: Academic Press, 245-277.