GENERALIZED REGRESSION ESTIMATION FOR A TWO-PHASE SAMPLE OF TAX RECORDS

John Armstrong and Hélène St-Jean[1]

ABSTRACT

Data from tax returns are used to obtain estimates of Canadian economic production for small businesses. Most variables of interest are not available in machine-readable form and must be obtained from source documents. Sampling is needed to avoid prohibitive data capture costs. Estimates are required for domains defined using four digits of Standard Industrial Classification (SIC) code but only the first two digits of SIC can be reliably obtained from administrative sources. A two-phase sample design using Bernoulli selection is employed. The estimation methodology involves adjustment of Horvitz-Thompson weights to compensate for differences between actual and expected sample sizes within poststrata. This approach does not use all available information. Some economic variables, particularly gross revenue, are available in machine-readable form for all tax records. The possibility of improving estimates using this information is examined in the context of generalized regression estimation.

1. INTRODUCTION

The two-phase tax sample is part of a general strategy for production of annual estimates of Canadian economic activity at Statistics Canada. 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 use of tax data does not have a large impact on the quality of estimates for the business population, since the contribution of small businesses is relatively small in most industries.

The two-phase tax sample 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.

[1] John Armstrong and Hélène St-Jean, Statistics Canada, Business Survey Methods Division, 11 - RH Coats Bldg, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

Administrative files containing information on business taxfilers are provided to Statistics Canada by Revenue Canada, the Canadian government department responsible for tax collection. There are two reasons why sampling of tax records is used rather than simple tabulation from these administrative files. First, although taxfilers are classified by Revenue Canada using the SIC code system, only the first two digits of SIC can be determined with sufficient accuracy using business activity information reported on tax returns. The cost of improving the accuracy of SIC4 codes for all tax records would be substantial. Second, estimates are required for many variables that are not available in machine-readable form and must be obtained from source documents. The cost of transcription of this information for all records would be prohibitive.

A two-phase approach to sampling of tax records was adopted to facilitate accurate estimation of economic production at the SIC4 level. Bernoulli sampling is used at both phases because it provides important operational advantages. The estimation methodology currently used in production is based on the Horwitz-Thompson estimator. It incorporates ratio adjustments calculated within poststrata to account for differences between actual and expected sample sizes. This methodology involves use of population counts but does not employ all available auxiliary information. The work on the use of additional auxiliary information reported here was motivated by the potential to reduce sample sizes required to obtain specified levels of precision. The estimation problem for two-phase sampling can be formulated using generalized regression estimation. This framework facilitates extensions of the current estimator to employ additional auxiliary variables.

The two-phase sample design is briefly described in Section 2. Section 3 includes a description of the estimation methods currently used in production and a discussion of their original motivation. Much of the presentation in Sections 2 and 3 follows the discussion of sample design and estimation issues in Armstrong, Block and Srinath (1993). A derivation of the generalized regression estimator in terms of the calibration approach of Deville and Särndal (1992) is presented in Section 4. The current production estimator in situated in the general framework. Section 5 includes the results of an empirical study involving comparison of the Horvitz-Thompson estimator, the ratio-adjusted version used in production and two generalized regression estimators that use additional auxiliary information.

2. SAMPLING DESIGN

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. Revenue Canada provides Statistics Canada with taxfiler information that can be used to construct a sampling frame.

All taxfilers reporting business income 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.

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. Information concerning numbers of businesses owned by T1 taxfilers and ownership shares is not available from Revenue Canada. 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 are not provided as frame data. Data for these variables are captured from copies of tax return information for taxfilers in the second-phase sample.

Information about the population of taxfilers for a particular tax year is accumulated by Revenue Canada over a period of two calendar years as tax returns are received and processed. If sample selection for a particular tax year did not begin until a complete frame was available, data capture operations would lead to considerable additional delays before estimates could be produced. Bernoulli sampling (also called Poisson sampling) is used during selection of both first- and second-phase samples to reduce delays and provide a relatively uniform workload to operations staff.

2.1 First-Phase Sample Selection

The first-phase sample is a sample of taxfilers selected from a frame created using Revenue Canada information. Strata are defined by SIC2, province and size (gross business income). The first-phase sample is a longitudinal sample. All taxfilers that are included in the first-phase sample in tax year T and are still in-scope for tax sampling in TY(T+1) (tax year T+1) are included in the first-phase sample for TY(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 Bernoulli 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 assigned a given taxfiler does not change from one tax year to the next. Denote the SIC2 codes used to define first-phase sampling strata within a province by $e=1,2,\ldots,E$ and denote size groups by $g=1,2,\ldots,G$. Let f_{egT} denote the first-phase sampling fraction for first-phase stratum eg, the stratum corresponding to SIC2 code e and size group g, for TY(T). Define the hash interval H_{egT} as $H_{egT} = (0, f_{egT})$.

Let R_i denote the hash number associated with taxfiler i and suppose that taxfiler i falls in first-phase stratum $eg(i,T)$ in TY(T). If taxfiler i is not in the first-phase sample of TY(T-1), taxfiler i is selected for the first-phase sample in TY(T) if $R_i \in H_{eg(i,T)T}$. Since taxfiler identifiers do not change over time, Bernoulli sampling facilitates selection of a longitudinal first-phase sample. Sample sizes obtained using this method are random variables.

First-phase selection probabilities must be updated from one year to the next. 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. Let $s1^T$ denote the first-phase sample for TY(T). The probability that taxfiler i is included in $s1^T$ is $p_{1i}^{T} = p_{1i}^{T-1} + P(i \in s1^T\ , i \notin s1^{T-1})$ where $P(i \in s1^T\ , i \notin s1^{T-1})$ is the probability that taxfiler i is included in the first-phase sample for TY(T) and is not included in the first-phase sample for TY(T-1).

If taxfiler i is a birth to the population for tax sampling in TY(T), we have $p_{1i}^{T} = f_{eg(i,T)T}$. Otherwise, noting that R_i does not vary between tax years and that the lower limit of the hash interval H_{egT} is zero for all e, g and T, it follows that $p_{1i}^{T} = max(\ p_{1i}^{T-1}, f_{eg(i,T)T})$.

2.2 Second-Phase Sample Selection

Let $J = \{j\}$ denote the population of businesses that is the target population for tax sampling. In order to select the second-phase sample, statistical entities are created using information about businesses corresponding to taxfilers in the first-phase sample. Each tax return includes income and expense data, as well as information about the percentage of the business owned by the taxfiler. A statistical entity, denoted by (i,j), is created for every taxfiler-business combination in the first-phase sample. Statistical entities 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.

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 Bernoulli 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 but does not change from one tax year to the next. Consequently, the hash number associated with each statistical entity does not change. Although the second-phase sample is not a longitudinal sample, the number of statistical entities in common between second-phase samples for consecutive tax years can be controlled by varying the overlap of hash intervals.

3. ESTIMATION

3.1 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. Note that adjusted data for business *j* is used during tabulation of estimates but is not employed for sample allocation or selection.

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

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

where J_i is a set containing the indices of the businesses wholly or partially owned by taxfiler *i*, p_{2i} is the probability that statistical entity *(i,j)* is selected for the second-phase sample and $y_j(d) = y_j$ if business *j* falls in domain *d* and is otherwise zero.

Noting that selection probabilities depend only on the taxfiler index *i*, $\hat{Y}_{H\text{-}T}(d)$ can be written as

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

where

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

Denoting expectations (and variances) with respect to the first and second phases of sampling by E_1 and E_2 (and V_1 and V_2), the variance of $\hat{Y}_{H\text{-}T}(d)$ can be derived as

$$\begin{aligned} V(\hat{Y}_{H-T}) &= V_1 E_2(\hat{Y}_{H-t}) + E_1 V_2(\hat{Y}_{H-T}) \\ &= V_1\Big(\sum_{i \in s1} y_i(d)/p_{1i}\Big) + E_1\Big(\sum_{i \in s1} p_{2i}(1-p_{2i})(y_i(d)/(p_{1i}p_{2i}))^2\Big) \\ &= \sum_i p_{1i}(1-p_{1i})(y_i(d)/p_{1i})^2 + \sum_i p_{1i}p_{2i}(1-p_{2i})(y_i(d)/(p_{1i}p_{2i}))^2 \\ &= \sum_i \left[(1-p_{1i}p_{2i})/(p_{1i}p_{2i})\right] y_i(d)^2 \quad . \end{aligned}$$

This variance is estimated by

$$\hat{V}(\hat{Y}_{H-T}(d)) = \sum_{i \in s2} \frac{(1-p_{1i}p_{2i})}{(p_{1i}p_{2i})^2} y_i(d)^2 \quad .$$

3.2 Poststratified Horvitz-Thompson Estimator

Sunter (1986) shows that the estimator analogous to $\hat{Y}_{H\text{-}T}(d)$ has a large variance in the case of a one-phase design using Bernoulli sampling. He considers a ratio form of the estimate, adjusted for differences between actual and expected sample sizes as suggested by Brewer, Early and Joyce (1972). He notes that the ratio form has a small bias and a variance that is considerably smaller than the unadjusted version. The methodology used to produce tax estimates incorporates ratio adjustments to account for differences between actual and expected sample sizes.

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 using a poststratified ratio adjustment. Following their notation, let $U=\{u\}$ denote a set of first-phase poststrata and suppose that 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

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

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

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

Similarly, let $V=\{v\}$ define a set of second-phase poststrata. An estimate of the number of taxfilers in second-phase poststratum v, based on the first-phase sample, is

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

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

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

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

$$W_i=W_{1i}W_{2i} \quad .$$

The poststratified estimate of the total of *y* over domain *d* is given by

$$\hat{Y}(d)=\sum_{i\in s2}\sum_{j\in J_i}\delta_{ij}W_i y_j(d) \quad .$$

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

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

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

$$\begin{aligned}\hat{V}(\hat{Y}(d)) = &\sum_{u}\sum_{v}(N_u/\check{N}_u)^2(\tilde{N}_v/\dot{N}_v)^2\sum_{i\in s2\cap u\cap v}\frac{(1-p_{1i})}{p_{1i}^2p_{2i}}(y_i(d)-\frac{\hat{Y}_u(d)}{\hat{N}_u})^2\\ &+\sum_{u}\sum_{v}(N_u/\check{N}_u)^2(\tilde{N}_v/\dot{N}_v)^2\sum_{i\in s2\cap u\cap v}\frac{(1-p_{2i})}{(p_{1i}p_{2i})^2}(y_i(d)-\frac{\hat{Y}_v(d)}{\hat{N}_v})^2 \quad ,\end{aligned}$$

where the estimates $\hat{N}_u$ and $\hat{N}_{..}$ 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.

4. GENERALIZED REGRESSION ESTIMATION

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.

Deville and Särndal (1992) derived the generalized regression estimator using calibration. Use of a calibration approach is convenient in the context of the two-phase tax sample. 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

$$\sum_{i\epsilon s1\cap u} W1_i \boldsymbol{x_i}=\boldsymbol{X_u}$$

for each first-phase poststratum u, where $\boldsymbol{x}_i$ is an $L_1 \times 1$ vector of auxiliary variables known for all units in the population and $\boldsymbol{X}_u$ is the vector of auxiliary variable totals for poststratum u. The adjusted weights minimize the distance measure $\sum_{i\epsilon s1}(g_{1i}-1)^2/p_{1i}$.

Weighting of the second-phase sample involves a calibration procedure 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

$$\sum_{i\epsilon s2\cap v} W_i \boldsymbol{z_i}=\tilde{\boldsymbol{Z}}_{\boldsymbol{v}}$$

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 $\tilde{\boldsymbol{Z}}_{\boldsymbol{v}}=\sum_{i\in s1\cap v} W_{1i}\boldsymbol{z}_{\boldsymbol{i}}$ is an estimate of the vector of auxiliary variable totals for post-stratum v, computed using the adjusted first-phase weights W_{1i}. The final weights minimize the distance measure $\sum_{i\in s2} W_{1i}(g_{2i}-1)^2/p_{2i}$.

Using first- and second-phase "g-weights", the generalized regression estimator can be written as

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

Let $\breve{\boldsymbol{X}}_{\boldsymbol{u}}=\sum_{i\in s1\cap u} \boldsymbol{x}_{\boldsymbol{i}}/p_{1i}$ 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

$$g_{1i}=1+\boldsymbol{\lambda}_{\boldsymbol{u}}'\boldsymbol{x}_{\boldsymbol{i}} \ ,$$

where $\boldsymbol{\lambda}_{\boldsymbol{u}}'=(\boldsymbol{X}_{\boldsymbol{u}}-\breve{\boldsymbol{X}}_{\boldsymbol{u}})'\boldsymbol{M}_{\boldsymbol{u}}^{-1}$ and $\boldsymbol{M}_{\boldsymbol{u}}^{-1}=(\sum_{i\in s_1\cap u}\boldsymbol{x}_{\boldsymbol{i}}\boldsymbol{x}_{\boldsymbol{i}}'/p_{1i})^{-1}$. For second-phase poststratum v, denote the estimate of $\tilde{\boldsymbol{Z}}_{\boldsymbol{v}}$ based on initial second-phase weights by $\dot{\boldsymbol{Z}}_{\boldsymbol{v}}=\sum_{i\in s2\cap v} W_{1i}\boldsymbol{z}_{\boldsymbol{i}}/p_{2i}$. The second-phase g-weight is

$$g_{2i}=1+\boldsymbol{\lambda}_{\boldsymbol{v}}'\boldsymbol{z}_{\boldsymbol{i}}$$

where $\boldsymbol{\lambda}_{\boldsymbol{v}}'=(\tilde{\boldsymbol{Z}}_{\boldsymbol{v}}-\dot{\boldsymbol{Z}}_{\boldsymbol{v}})'\boldsymbol{M}_{\boldsymbol{v}}^{-1}$ and $\boldsymbol{M}_{\boldsymbol{v}}^{-1}=(\sum_{i\in s_2\cap v} W_{1i}\boldsymbol{z}_{\boldsymbol{i}}\boldsymbol{z}_{\boldsymbol{i}}'/p_{2i})^{-1}$.

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

$$V(\hat{Y}_{GREG}(d)) \approx \sum_{i}\frac{1-p_{1i}}{p_{1i}}(E_{1i}(d))^2 + E_1[\sum_{i\in s_2}\frac{1-p_{2i}}{p_{2i}}(W_{1i}E_{2i}(d))^2] \ ,$$

where $E_{1i}(d) = y_i(d) - \boldsymbol{x}_i{}'\boldsymbol{B}_u(\boldsymbol{d})$ for each taxfiler in first-phase poststratum *u* and ***B***$_u$***(d)***, the vector of estimated coefficients from the regression of *y(d)* on **x** that would be obtained if *y(d)* was available for all taxfilers in first-phase poststratum *u*, is given by

$$\boldsymbol{B}_u(\boldsymbol{d}) = (\sum_{i \in u} \boldsymbol{x}_i \boldsymbol{x}_i{}')^{-1} (\sum_{i \in u} \boldsymbol{x}_i y_i(d)) \quad .$$

Similarly, $E_{2i}(d) = y_i(d) - \boldsymbol{z}_i{}'\boldsymbol{B}_v(\boldsymbol{d})$ for each taxfiler in second-phase poststratum *v* and ***B***$_v$***(d)***, 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 taxfilers in the component of the first-phase sample falling in second-phase poststratum *v*, is given by

$$\boldsymbol{B}_v(\boldsymbol{d}) = (\sum_{i \in s1 \cap v} W_{1i} \boldsymbol{z}_i \boldsymbol{z}_i{}')^{-1} (\sum_{i \in s1 \cap v} W_{1i} \boldsymbol{z}_i y_i(d)) \quad .$$

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

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

Refer to Appendix A for more details concerning the derivation of the approximate variance of $\hat{Y}_{\mathrm{GREG}}$*(d)* and this variance estimator.

Since *y(d)* is available only for units in *s2*, the best available estimate of ***B***$_u$***(d)*** is

$$\hat{\boldsymbol{B}}_u(\boldsymbol{d}) = (\sum_{i \in s2 \cap u} W_i \boldsymbol{x}_i \boldsymbol{x}_i{}')^{-1} (\sum_{i \in s2 \cap u} W_i \boldsymbol{x}_i y_i(d)) \quad .$$

Similarly, the best available estimate of ***B***$_v$***(d)*** is

$$\hat{\boldsymbol{B}}_v(\boldsymbol{d}) = (\sum_{i \in s2 \cap v} W_i \boldsymbol{z}_i \boldsymbol{z}_i{}')^{-1} (\sum_{i \in s2 \cap v} W_i \boldsymbol{z}_i y_i(d)) \quad .$$

The sample residuals needed to compute the variance estimator are $e_{1i}(d) = y_i(d) - \boldsymbol{x}_i{}'\hat{\boldsymbol{B}}_u(\boldsymbol{d})$

and $e_{2i}(d) = y_i(d) - \boldsymbol{z}_i{}'\hat{\boldsymbol{B}}_v(\boldsymbol{d})$.

If a single auxiliary variable with value one for all taxfilers is employed during both first- and second-phase weighting, $g_{1i}=N_u/\check{N}_u$ for all taxfilers in first-phase poststratum u, $g_{2i}=\tilde{N}_v/\dot{N}_v$ for all taxfilers in second-phase poststratum v and $\hat{Y}_{GREG}(d)$ is equivalent to $\hat{Y}(d)$. In addition, there is only one minor difference between the respective variance estimators in this special case. The second-phase g-weight appears in the leading term of $\hat{V}(\hat{Y}(d))$ but does not appear in $\hat{V}(\hat{Y}_{GREG}(d))$.

If y is strongly correlated with $\mathbf{x}$ 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 $\mathbf{x}$ 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 $\mathbf{x}$ and z within poststrata that include at least one sampled unit falling in domain d.

The correlation between $y(d)$ and $\mathbf{x}$ and z within a poststratum that includes at least one sampled unit falling in domain d will be low if domain d includes only a small proportion of all the sampled units in the poststratum. This situation may arise for two reasons in the context of the two-phase tax sample. First, poststrata may be defined to include many domains. 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 SIC4 domain. Second, if the SIC codes used for stratification contain errors and poststrata are formed by combining sampling strata, domains may be divided between a number of poststrata.

The g-weights associated with the generalized regression estimator and, consequently, generalized regression estimates, can be negative. In the special case in which $\hat{Y}_{GREG}(d)$ is equivalent to $\hat{Y}(d)$, all g-weights will be non-negative. Use of additional auxiliary information can lead to negative weights.

5. EMPIRICAL STUDY

In order to compare the performance of $\hat{Y}_{H\text{-}T}(d)$, $\hat{Y}(d)$ and $\hat{Y}_{GREG}(d)$, 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 $\hat{Y}_{GREG}(d)$, it will be called $\hat{Y}_{GREG\text{-}TPH}(d)$ in subsequent discussion. (TPH is an abbreviation for two-phase Hájek.) Two other generalized regression estimators were considered. In both cases, $\mathbf{x}$ and z contained a variable with value one for all taxfilers. One generalized regression estimator involved calibration on taxfiler revenue during second-phase weighting. The second estimator involved calibration on taxfiler revenue at both phases of weighting. Estimates of domain totals computed using these two estimators are denoted by $\hat{Y}_{GREG\text{-}R2}(d)$ and $\hat{Y}_{GREG\text{-}R1R2}(d)$, 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 universe 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 9884 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 the approximate variance estimator for $\hat{Y}_{GREG}(d)$ when the sample size is very small (Rao 1968). 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 $\hat{Y}_{GREG\text{-}TPH}(d)$, $\hat{Y}_{GREG\text{-}R2}(d)$ and $\hat{Y}_{GREG\text{-}R1R2}(d)$ 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 $\hat{Y}_{H\text{-}T}(d)$. Since GREG-R1R2 did not produce any negative estimates, no measures were taken to modify the negative weights associated with the estimator.

Results of comparison of the GREG-TPH and H-T estimators are presented in Table 1 and Table 2. The relative performance of these estimators are very similar for both variables of interest. The GREG-TPH estimator performs better than the H-T estimator for the 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 only 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. Comparison of the relative sizes of GREG-TPH and H-T estimators provides more compelling evidence to prefer GREG-TPH in practice. The GREG-TPH estimate of total expenses was larger then the H-T estimate for over 93% of the SIC4 domains. Actual two-phase tax sample sizes are typically lower than expected sample sizes for various operational reasons. Use of the GREG-TPH estimator provides an automatic non-response adjustment.

The GREG-TPH estimator is compared to GREG-R2 and GREG-R1R2 using total expenses as the variable of interest 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.

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 used to determine domain membership. A comparison of Tables 5 and 6 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.

6. CONCLUSIONS

Generalized regression estimation provides a convenient framework for the use of auxiliary information. It can be applied to the two-phase tax sample selected by Statistics Canada to obtain annual estimates of the economic activity of small businesses. The two-phase tax sample involves Bernoulli sampling at both phases of selection because Bernoulli sampling has considerable operational advantages. The estimation method currently used in production incorporates poststratified ratio adjustments during both first- and second-phase weighting to compensate for differences between actual and expected sample sizes. It can be derived as a generalized regression estimator.

In an empirical study, the generalized regression estimator currently used in production (GREG-TPH) performed much better then the Horvitz-Thompson estimator. Two other generalized regression estimators were compared to GREG-TPH. The alternative estimators produced improvement 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. The possibility that generalized regression estimators using more auxiliary information than GREG-TPH can be employed to produce estimates for certain industries where domains of interest are large and SIC codes used for stratification during first- and second-phase sample selection are relatively accurate is under investigation.

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 Jean Leduc for helpful comments on a earlier draft of this paper.

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*. (in press).

Bankier, M., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the

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. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 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.

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

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.

Rao, J.N.K. (1968). 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. *Sankhya*, Ser. C, 37, 43-52.

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 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.

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

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

$$V(\hat{Y}_{GREG}(d)) = E_1 V_2(\hat{Y}_{GREG}(d)) + V_1 E_2(\hat{Y}_{GREG}(d)) \quad .$$

First, consider the variance of the estimator with respect to the second phase of sampling, conditional on the results of first-phase calibration. The generalized regression estimator can be written as

$$\begin{aligned}\hat{Y}_{GREG}(d) &= \sum_{i\in s2} W_{1i} W_{2i} y_i(d) \\ &= \sum_{v} \sum_{i\in s2\cap v} W_{1i}(y_i(d) - \boldsymbol{z}_i{}'\hat{\boldsymbol{B}}_v) + \sum_{v} \tilde{Z}_v \hat{\boldsymbol{B}}_v\end{aligned}$$

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

$$\begin{aligned}E_1 V_2(\hat{Y}_{GREG}) &\approx E_1 V_2 (\sum_{i\in s2} W_{1i} E_{2i}(d)) \\ &= E_1 (\sum_{i\in s2} \frac{(1-p_{2i})}{p_{2i}} (W_{1i} E_{2i}(d))^2)\end{aligned}$$

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

$$\hat{Q1} = \sum_{i\in s2} \frac{(1-p_{2i})}{p_{1i}^2 p_{2i}^2} (g_{1i} g_{2i} e_{2i}(d))^2 \quad .$$

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

$$\begin{aligned}V_1 E_2(\hat{Y}_{GREG}(d)) &= V_1 (\sum_{i\in s1} W_{1i} y_i(d)) \\ &= \sum_{i} \frac{(1-p_{1i})}{p_{1i}} E_{1i}(d)^2 \quad .\end{aligned}$$

An estimator of this term is

$$\hat{Q2} = \sum_{i\in s2} \frac{(1-p_{1i})}{p_{1i}^2 p_{2i}} (g_{1i} e_{1i}(d))^2 \quad .$$

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

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

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