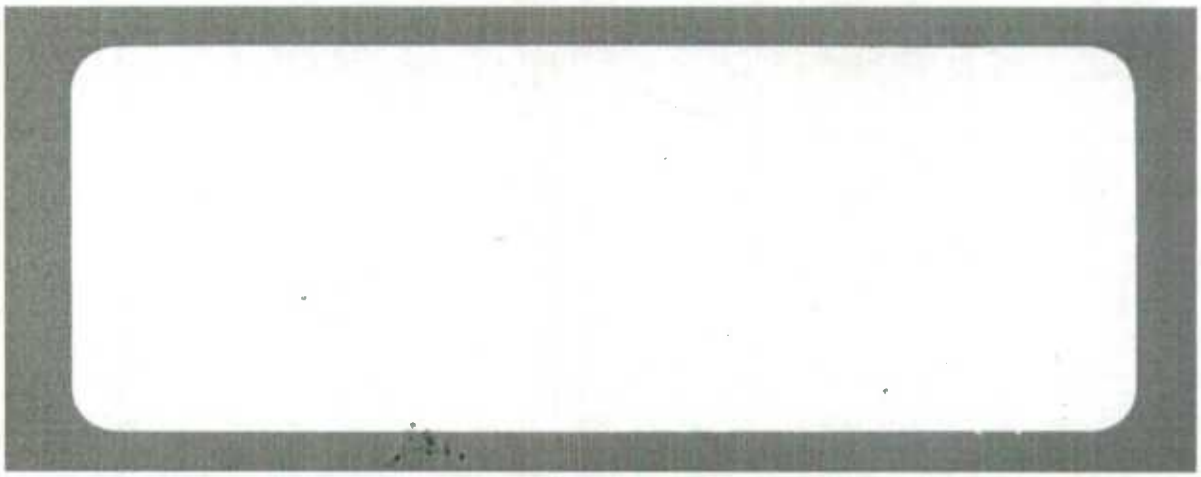
11-619E
no. 2013-001
c. 3

Methodology Branch

Direction de la méthodologie



Household Survey Methods Division

Division des méthodes d'enquêtes auprès des ménages

Statistics Canada Statistique Canada

Canada

WORKING PAPER
METHODOLOGY BRANCH

Variance Estimation for High Income Tables

HSMD - 2013 – 001E

Wei Qian

Household Survey Methods Division
Statistics Canada

January 2013

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada

Variance Estimation for High Income Tables

Wei Qian[1]

Abstract

Income Statistics Division has produced high income tables for 1982-2009, using the data in the Longitudinal Administrative Databank (LAD). These tables involve estimation of percentiles and quantities in percentile groups. Until now, there has not been any statement about the quality (cvs) of the estimates produced.

In this report, we propose a solution on how variance estimates for the variables in those high income tables could be obtained via Taylor linearization and the estimating equation approach. Data from P.E.I is used to illustrate the results obtained.

[1] I would like to thank Xuelin Zhang, Cynthia Bocci, Abdellatif Demnati, Ann Lim, Christian Nadeau, Wisner Jocelyn for providing many useful comments on the paper.

Estimation de la variance pour les tableaux des hauts revenus

Wei Qian

Résumé

La division de la statistique du revenu produit des tableaux d'information sur les personnes à hauts revenus couvrant la période 1982-2009, à partir de la banque de données administratives longitudinales (BDAL). La production de ces tableaux impliquent des estimations de quantiles ainsi que des estimations à l'intérieur de groupes définis par des quantiles. Jusqu'à maintenant, on ne s'est pas attardé à évaluer la qualité (cv) des estimations produites.

Dans le rapport ci-joint, on propose une façon d'obtenir des estimateurs de variance pour les statistiques produites, grâce à la méthode de linéarisation de Taylor et des équations d'estimation. Nous utilisons les données provenant de l'Ile du Prince Edouard(IPE) pour fins d'illustration.

Table of Contents

1. Introduction

In early 2012, Income Statistics Division (ISD) proposed publishing several "High Income Tables" through the Canadian Socio-Economic Information Management System (CANSIM). The estimates in the high income tables (also known as high income statistics) are obtained from the Longitudinal Administrative Databank (LAD). High income tables provide estimates on demography, income and taxation in groups defined by income percentiles for various levels of geography including Canada, provinces / territories, and regions (such as Census Metropolitan Area /Census Agglomeration (CMA/CA). In this paper, the quality, more specifically the sampling variance, of high income statistics is of interest and a method is proposed to provide appropriate associated variance estimates.

This paper is organized as follows. Section 2 gives an introduction of the LAD from which the high income statistics are obtained. Section 3 provides an overview of the parameters of interest and their estimators in the high income tables. In Section 4, linearization and re-sampling variance estimation methods are discussed, and linear variance estimators are derived by using a unified estimating equations approach. In Section 5, the linear variance estimators are evaluated, using tax data from the T1 family file (T1FF) and by comparing them to variance estimators obtained via the bootstrap method. In Section 6, the linear variance estimators are applied to the 1988 and 2009 high income tables and the coefficient of variation (CV) estimates are produced for selected estimates. The last section summarizes the findings and discusses future work.

2. Longitudinal Administrative Databank

High income tables are produced from data in the LAD which consists of a 20% random sample selected from the T1 family file (T1FF).

The T1FF is an annual cross-sectional file of all taxfilers and their families. Census families in the T1FF are created from personal income tax returns (T1) submitted to the Canada Revenue Agency (CRA). Both legal and common-law spouses are linked by the spousal Social Insurance Number (SIN) provided on their tax forms, or by matching by name, address, age, sex, and marital status. Children are identified through a similar algorithm and through supplementary files. Prior to 1993, non-filing children were identified from information on their parents' tax forms. Information from the Family Allowance Program was used to assist in the identification of children. Since 1993, information from the Child Tax Benefit Program has been used for this purpose.

The individuals on the LAD are selected using Bernoulli sampling with equal selection probability of 1/5, based on their SIN. Although there is no age restriction, people without a SIN can only be included in the family component. Once a person is selected, this individual will be on the LAD file for any subsequent year if he or she is on the T1FF file for that year. A unique LAD identification number allows individuals selected for the LAD to be linked across the years to create a longitudinal profile of each individual.

The LAD is augmented each year with a sample of new taxfilers so that it consists of approximately 20% of taxfilers every year. The sample has increased from 3,227,485 persons in 1982 to 5,158,895 in 2009 (an almost 60% increase). This increase reflects increases in the Canadian population and increases in the incidence of tax filing as a result of the introduction of the federal sales tax credit in 1986 and the Goods and Services Tax credit in 1989.

The LAD is organized into four levels of aggregation, namely the individual, spouse/parent, family, and child(ren) levels. The databank contains information on demographics, income, and other taxation data at the different levels of aggregation, with new data being added annually as the information becomes available. Changes in tax legislation and in the design of the T1 form itself have resulted in some variables not being available for all years as well as some minor definitional changes from one year to the next.

The LAD is also linked with the Longitudinal Immigration Database (IMDB) which contains immigration records from 1980 to 2007.

Why are the high income tables based on the LAD instead of the T1FF? The main reason is that high income tables contain longitudinal statistics on high income trends for Canadian taxfilers. The LAD is a longitudinal database, while the T1FF is cross-sectional. Once the definition of a variable is changed, the LAD is revised for all reference years to maintain its longitudinal consistency; this change is not applied to previous T1FFs. The longitudinal profile of the LAD has made it an important research tool for longitudinal studies on income.

More details about the LAD can be found in the *Longitudinal Administrative Data Dictionary* (Statistics Canada internal document, 2010).

3. Parameter Estimation

High income tables provide statistics on demography, income and taxation in groups defined by income percentiles for various levels of geography. Two sets of tables are generated for the percentiles: national level tables and local level tables. Statistics in the national level tables are always based on ranking taxfilers within the national (Canada-wide) income distribution, while statistics in the local level tables are based on ranking

taxfilers within the income distribution of a specific geographic area (Province or CMA/CA etc.). In the national level tables, since the percentile thresholds are based on Canada-wide, the estimation of parameters for provinces or CMA/CAs is domain estimation. ***This paper instead focuses on the local level tables***. From the methodology point of view, the statistics in the national level tables should be of better quality than those in the local ones since the national percentile estimates are less variable due to the larger sample size.

The local level tables include six tables whose percentile thresholds are defined by different income variables. Table 1 lists the income variables defining the percentile groups.

Table 1: Income variables defining the percentile group

Income Variable
Market income
Total income
After tax income
Market income with capital gains
Total income with capital gains
After tax income with capital gains

Table 2 summarizes the statistics generated for each percentile group in high income tables. The first item is the estimate providing information on the distribution of the income variables. The other items are the demographic, income and taxation characteristics. Items 15-20 are longitudinal characteristics, and their estimation depends on the sampling design of previous years. For longitudinal statistics, the associated variance estimation is much more complicated than that for cross-sectional statistics because of the dependence. In this paper, the computation is simplified by treating longitudinal indicators as cross-sectional ones. For large sample sizes, the variance associated with percentile estimators is very small and the estimates are close to the actual values. Their CVs may also provide an idea of the quality of the longitudinal statistics.

Table 2: A summary of statistics in high income tables

	Statistics
1	Income threshold value
2	Number of tax filers
3	Percentage, married males or females
4	Percentage married by sex
5	Median age
6	Median income
7	Average income
8	Share of income
9	Share of income, by sex
10	Median federal and provincial income taxes paid
11	Average federal and provincial income taxes paid
12	Share of federal and provincial income taxes paid
13	Percentage of income from wages and salaries
14	Percentage of income from wages and salaries, by sex
15	Percentage in the same quantile last year
16	Percentage in the same quantile five years ago
17	Percentage in top 5 percentiles last year
18	Percentage in top 5 percentiles five years ago
19	Percentage in top 5 percentiles at least once during the preceding five-year period
20	Percentage always in top 5 percentiles during the preceding five-year period

For methodological purposes, these parameters may be summarized into six categories: the percentile of a distribution, and, within a percentile group, the mean, median, share, ratio and a function of them such as the product of share and ratio. The estimators for the different types of parameters in the high income tables are given below.

Consider a population of size N, such that $U = \{1, \dots, N\}$. Let X be the income variable defining the percentile group and Y be another variable (demographic, income or taxation) whose quantities are of interest. Let ξ_p denote the pth percentile of X and let γ_p denote the quantity of interest for Y in the top pth percentile group defined as $\{i \in U: x_i \geq \xi_p\}$. Then, (ξ_p, γ_p) are the parameters of interest. Let S be the sample drawn from U and $(\hat{\xi}_p, \hat{\gamma}_p)$ be estimators for (ξ_p, γ_p).

Since γ_p is a parameter of the population in a percentile group, the estimation of γ_p relies on the estimation of the percentile ξ_p. The weighted percentile estimator $\hat{\xi}_p$ is defined as

$$\hat{\xi}_p = \begin{cases} x_{(1)} & if\ \frac{1}{\hat{N}} w_1 > p \\ \frac{1}{2}\left(x_{(i)} + x_{(i+1)}\right) & if\ \frac{1}{\hat{N}} \sum_{j=1}^{i} w_j = p \\ x_{(i+1)} & if\ \frac{1}{\hat{N}} \sum_{j=1}^{i} w_j < p < \frac{1}{\hat{N}} \sum_{j=1}^{i+1} w_j \end{cases} \quad (1)$$

where $x_{(i)}$ is the ordered values of the income variable, w_i is the sampling weight (the inverse of the selection probability) associated with $x_{(i)}$ and $\hat{N} = \sum_{i \in S} w_i$. The weighted percentile estimator is consistent and its bias is negligible for large sample sizes. The estimate can be obtained from the SAS procedure PROC UNIVARIATE. Given the percentile estimate, $\hat{\gamma}_p$ is defined below for the different types of parameters.

Case 1. γ_p is an average in the top pth percentile group

For both continuous income variables and categorical demography variables, many parameters in these tables can be expressed as an average. For example, the percentage of male in the population is the average of an indicator variable indicating male or not. Item 3, 7 and 11 in Table 2 can be expressed as an average. The estimator for an average is

$$\hat{\gamma}_p = \frac{\sum_{i \in S} w_i I\{x_i \geq \hat{\xi}_p\} y_i}{\sum_{i \in S} w_i I\{x_i \geq \hat{\xi}_p\}}, \quad (2)$$

where $I\{x_i \geq \hat{\xi}_p\} = \begin{cases} 1 & if\ x_i \geq \hat{\xi}_p \\ 0 & \text{otherwise} \end{cases}$.

Case 2. γ_p is a median in the top pth percentile group

The top pth percentile group is treated as a sub-population. Let S_p be the corresponding sub-sample such that $S_p = \{i: i \in S, x_i \geq \hat{\xi}_p\}$. Let $y_{(i)}$ be the ordered values of the variable Y for sampled units in S_p, and $w_{(i)}$ be the sampling weight associated with the unit whose y-value is $y_{(i)}$. Then, the estimator of the median of Y based on S_p is

$$\hat{\gamma}_p = \begin{cases} y_{(1)} & if\ \frac{1}{\hat{N}_p} w_1 > 0.5 \\ \frac{1}{2}\left(y_{(i)} + y_{(i+1)}\right) & if\ \frac{1}{\hat{N}_p} \sum_{j=1}^{i} w_j = 0.5 \\ y_{(i+1)} & if\ \frac{1}{\hat{N}_p} \sum_{j=1}^{i} w_j < 0.5 < \frac{1}{\hat{N}_p} \sum_{j=1}^{i+1} w_j \end{cases} \quad (3)$$

where $\hat{N}_p = \sum_{i \in S_p} w_i$.

Case 3. γ_p is a ratio in the top pth percentile group

Some statistics may be expressed as a ratio of the totals (average) of two variables. In Table 2, the estimator of the percentages of income from wages and salaries in the pth percentile group is defined as

$$\hat{\gamma}_p = \frac{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\} y_i}{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\} x_i}, \quad (4)$$

where Y is the wage and salaries and X is the income variable.

The estimator of the percentage of married by sex, can also be expressed as a ratio as

$$\hat{\gamma}_p = \frac{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\} I\{\text{person } i \text{ is married and male(female)}\}}{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\} I\{\text{person } i \text{ is male (female)}\}},$$

where $I\{\text{person } i \text{ is married and male(female)}\} = \begin{cases} 1 & \text{if } i \text{ is married and male(female)} \\ 0 & \text{otherwise} \end{cases}$ and

$I\{\text{person } i \text{ is male(female)}\} = \begin{cases} 1 & \text{if } i \text{ is male(female)} \\ 0 & \text{otherwise} \end{cases}$.

Longitudinal statistics are also treated as ratios. For example, the estimator of the percentage in top 5 percentiles at least once during the preceding five-year period (Item 19) is

$$\hat{\gamma}_p = \frac{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\} I\{\text{in top 5\% at least once during the preceding 5 year period}\}}{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\} I\{\text{ person } i \text{ filed during the preceding 5 year period }\}}.$$

The longitudinal indicator in the numerator depends on the 5th percentile estimates of the last five years. In the case of large sample sizes, those 5th percentile estimates are very close to the actual 5th percentile. Therefore, the longitudinal indicators are treated as fixed and the above estimator becomes a ratio of two indicator variables.

Case 4. γ_p is a share in the top pth percentile group

The share of a percentile group reflects the degree of income inequality in a population. It is defined as the ratio of total income (or tax) for the persons in the percentile group over that for all persons in the population. The estimator is given as

$$\hat{\gamma}_p = \frac{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\} y_i}{\sum_{i\in S} w_i y_i}. \quad (5)$$

Case 5. γ_p is a product of share and ratio

The estimators of some parameters may not be as simple as the above cases, but they can be expressed as a function of them. For example, for men in a percentile group, the income share, γ_p, can be viewed as a product of share $\gamma_p^{(S)}$ and ratio $\gamma_p^{(R)}$:

$$\gamma_p = \frac{income\ of\ men\ in\ a\ percentile\ group}{income\ of\ all}$$

$$= \frac{income\ of\ all\ in\ a\ percentile\ group}{income\ of\ all} \times \frac{income\ of\ men\ in\ a\ percentile\ group}{income\ of\ all\ in\ percentile\ group}$$

$$= \gamma_p^{(S)} \times \gamma_p^{(R)}.$$

Accordingly, the estimator is defined as

$$\hat{Y}_p = \hat{Y}_p^{(S)} \times \hat{Y}_p^{(R)}. \tag{6}$$

where $Y_p^{(R)}$ and $\hat{Y}_p^{(R)}$ are given as **Case 3** and $Y_p^{(S)}$and $\hat{Y}_p^{(S)}$ are given in **Case 4**.

4. Variance Estimation

High income tables are intended to provide information about all Canadian taxfilers. The T1FF, serving as the sampling frame, is based on the T1 forms collected by the CRA. After 1992, the T1FF provides a very good coverage of the target population. In addition, most of the values on the T1FF are reported by the taxfilers or are derived from the reported values; only a very small portion is imputed. Therefore, non-response errors and measurement errors should be negligible. In this paper, only the sampling variance of high income statistics is considered and the associated variance or CV estimates are presented.

Two types of variance estimation methods are usually considered for household surveys: re-sampling and linearization. Bootstrap and jackknife are the two most popular re-sampling methods used for household surveys at Statistics Canada. The jackknife method is often used for surveys with multi-stage clustering design such as the Labour Force Survey (LFS). In this study, the jackknife is ruled out because it performs poorly for estimating the variance of non-smooth estimators such as sample percentiles. Bootstrap variance estimators are commonly used in household surveys, such as Survey of Labour and Income Dynamics (SLID). The advantage of the bootstrap method is that, 1) it works well for non-smooth estimators under simple sampling designs, and 2) it is easy to implement. It is not necessary to develop formulas for the different estimators. The bootstrap algorithm for Bernoulli sampling is very simple. The disadvantage of the bootstrap is the time and computational resources required. As stated previously, the LAD sample size now is more than 5 million records. Running the estimation process repeatedly on the LAD for all geography levels would take a tremendous amount of time. For example, for Ontario, more than 3 weeks was required to produce all local level tables. However, the bootstrap provides a tool to verify other variance estimators for some smaller domains; moreover, the bootstrap may be preferable for analytical purposes as the analysts can use the bootstrap samples to generate replicates of test statistics and then produce confidence interval estimates.

On the other hand, linearization methods have long been used in surveys and the theory is well developed. Standard variance estimation methods from textbooks can be used only for linear estimator, such as the Horvitz-Thompson (HT) estimator (see Särndal et al., 1991). For a smooth nonlinear estimator, Taylor linearization permits the nonlinear estimator to be approximated by a HT total estimator for a new variable - *linear variable*. Then, the variance of the nonlinear estimator may be approximated by the variance

of an HT total estimator which, in turn, can be estimated by the standard methods. For example, suppose $\hat{\theta}$ is a non-linear "smooth" estimator and Z is the associated linear variable. Then,

$$V(\hat{\theta}) \approx V(\sum_{i \in S} w_i z_i), \tag{7}$$

where S is the sample and z_i is the value for the linearized variable attached to unit i. The problem with the linearization method is that a linear variable must be found for each estimator and the linearization method is not easily generalized. For example, if a quantity in a low percentile group is of interest, the formula for variance estimation developed for the top percentile group cannot be reused. However, the linearization method does not require replication therefore the computation is fast. In addition, it provides consistent variance estimates. The linearization variance estimation method is discussed below.

As stated previously, the sampling design for the LAD is very simple: Bernoulli sampling with the selection probability of 0.2. As a result, the variance formula given by (7) can be simplified as

$$V(\hat{\theta}) \approx 4 \sum_{i \in U} z_i^2, \tag{8}$$

where U is the population. In the case where the number of individuals in the population is not available (the population counts in some small geographies may not be provided), the variance estimator is then given by

$$\hat{V}(\hat{\theta}) = 20 \sum_{i \in S} \hat{z}_i^2, \tag{9}$$

where $\hat{z}_i$ is a proper estimator of z_i since z_i may involve some unknown finite population quantities.

Binder (1983) introduced a unified estimating equations approach for estimating finite population parameters. The estimating equations approach assumes that the finite population is a sample from a superpopulation model and the sample is a subsample of the finite population. Any finite population parameter θ can be viewed as a solution of "census" estimating equations:

$$U(\theta) = \sum_{i \in U} u(\theta, y_i) = 0.$$

The estimator $\hat{\theta}$ can be found by solving the corresponding weighted estimating equations:

$$\hat{U}(\theta) = \sum_{i \in S} w_i u(\theta, y_i) = 0,$$

where $\hat{U}(\theta)$ is the HT total estimator of $U(\theta)$. Under regularity conditions, $\hat{\theta}$ is a consistent estimator of θ. For more details on the derivation of linear variables, see the appendix. For the case where θ is a parameter vector, u is a vector of the same dimension as θ.

Suppose θ_0 is the true value of θ. Taylor linearization around θ_0 leads to

$$\frac{1}{N}\left(\hat{U}(\hat{\theta}) - \hat{U}(\theta_0)\right) \approx \left(E[u(\gamma, Y)]|_{\gamma=\hat{\theta}} - E[u(\gamma, Y)]|_{\gamma=\theta_0}\right) \approx \left[\frac{\partial E[u(\theta; Y)]}{\partial \theta}\right]_{\theta=\theta_0} (\hat{\theta} - \theta_0).$$

where the expectation is under the superpopulation model. The conditions for the approximation are discussed in Randles (1982) and Shao and Rao (1994). Suppose that θ is a parameter vector. Then,

$$\hat{\theta} - \theta_0 \approx -\frac{1}{N}\left[\frac{\partial E[u(\theta;Y)]}{\partial \theta}\right]_{\theta=\theta_0}^{-1} \hat{U}(\theta_0).$$

Therefore, the variance of $\hat{\theta}$ is

$$V(\hat{\theta}) \approx V\left(\sum_{i\in S} w_i u_i^*\right)$$
$$= \sum_{k,l\in U}\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) u_k^* u_l^{*T} \qquad (10)$$

where $u_i^* = -\frac{1}{N}\left[\frac{\partial E[u(\theta;Y)]}{\partial \theta}\right]_{\theta=\theta_0}^{-1} u_i(\theta_0, y_i)$. Since u_i^* may involve unknown quantities, they can be replaced by the proper estimate $\hat{u}_i^*$. As a result, the variance estimator becomes

$$\hat{V}(\hat{\theta}) = \sum_{k,l\in S}\left(\frac{1}{\pi_k\pi_l} - \frac{1}{\pi_{kl}}\right)\hat{u}_k^*\hat{u}_l^{*T}. \qquad (11)$$

Given the formula in (9), the variance estimator in (11) becomes

$$\hat{V}(\hat{\theta}) = 20\sum_{i\in S}\hat{u}_i^*\,\hat{u}_i^{*T}.$$

Thus, it remains to find u, u_i^* and $\hat{u}_i^*$.

High income statistics involve the estimation of percentiles (non-smooth statistics) and quantities in the top pth percentile group. The application of estimating equations approach to non-smooth statistics is discussed in Binder and Kovacevic (1995) and Osier (2009).

Let $u_i = (u_{1i}, u_{2i})^T$ be estimating functions for (ξ_p, γ_p), where ξ_p is the pth percentile of the income variable X and γ_p is the quantity of interest for variable Y in the top pth percentile group defined by ξ_p. In this study, X and Y are different variables, while both reference papers only discussed the case where X and Y are the same. This difference leads to the estimation of their conditional distributions for which a nonparametric method in Borkowf et al. (1996) is used.

Assume that X is a nonnegative continuous variable[2]. For ξ_p, the pth percentile of the variable X, and its estimator $\hat{\xi}_p$, the estimating equation and linear variable are

$$u_{1i} = I\{x_i \le \xi_p\} - p,$$
$$u_{1i}^* = -\frac{1}{f(\xi_p)}\left[I\{x_i \le \xi_p\} - p\right], \quad \textit{and}$$

[2] Some individuals may have negative income values. Since we only consider estimating the parameters in the top percentile groups, setting these negative values to zero has little impact.

$$\hat{u}_{1i}^{*} = -\frac{1}{\hat{f}(\hat{\xi}_p)}\left[I\{x_i \le \hat{\xi}_p\} - p\right],$$

where $\hat{u}_{1i}^{*}$ needs the estimation of $f(\xi_p)$ – the probability density function of X at ξ_p .

Two possible methods can be used for the estimation of the density function for complex survey data. Francisco and Fuller (1991) use the density estimator

$$\hat{f}(x) = \frac{2z_{\alpha/2}\delta}{h_1 + h_2}$$

where

$$\delta^2 = mse\left\{\sum_{i\in S} w_i[I\{x_i \le x\} - p]\right\},$$

$z_{\alpha/2}$ is the $100\left(1-\frac{\alpha}{2}\right)$-th percentile from the standard normal distribution, and h_1 and h_2 are found by solving

$$\inf_{h_1}\left\{\frac{1}{\hat{N}}\sum_{i\in S} w_i[I\{x_i \le x - h_1\} - p] \le -z_{\alpha/2}\delta\right\}, \quad \text{and} \quad \inf_{h_2}\left\{\frac{1}{\hat{N}}\sum_{i\in S} w_i[I\{x_i \le x + h_2\} - p] \ge z_{\alpha/2}\delta\right\}.$$

Lohr and Buskirk (1999) propose a weighted kernel density estimator such that

$$\hat{f}(x) = \frac{1}{\hat{N}}\sum_{i\in S} w_i\phi_h(x - x_i),$$

where h is the bandwidth and

$$\phi_h(t) = \frac{1}{h\sqrt{2\pi}}\exp\left(-\frac{t^2}{2h^2}\right),$$

is the standard normal density rescaled by the bandwidth. The bandwidth is obtained by

$$h = 0.79\hat{Q}n^{-\frac{1}{5}},$$

where $\hat{Q}$ is the sample interquartile range (IQR). Note that the kernel density estimation is very sensitive to the choice of bandwidth, especially at the tail of the distribution.

In this paper, the method proposed by Francisco and Fuller (1991) is used. As suggested by Rao and Wu (1987), α is set to 0.05. Using the data from selected small domains, the variance estimates are shown to be very similar to the bootstrap variance estimates.

For a different variable Y and its corresponding quantity γ_p and estimator $\hat{\gamma}_p$, estimating functions and associated linearized variables are presented below for the different cases. More details on the derivation of the linear variables for **Case 1** is provided in the appendix, using an approach similar to that given in Binder and Kovacevic (1995).

Case 1. γ_p is an average in the top pth percentile group

The mean estimator $\hat{\gamma}_p$ is used not only for continuous income variables but also for categorical demography variables. For example, to estimate the percentage of male, we only need to create a variable indicating male or not, the percentage of male is the average of the indicator variable. The estimating function for γ_p is

$$u_{2i} = I\{x_i \geq \xi_p\}(y_i - \gamma_p),$$

and the associated linearized variable is

$$u_{2i}^* = \frac{1}{N(1-p)}\left\{(\gamma_p - E[Y|\xi_p])\left(I_{\{x \leq \xi_p\}} - p\right) + I_{\{x \geq \xi_p\}}(y_i - \gamma_p)\right\}.$$

By replacing all unknown quantities above replaced by proper estimators, the above formula becomes

$$\hat{u}_{2i}^* = \frac{1}{\hat{N}(1-p)}\left\{(\hat{\gamma}_p - \hat{E}[Y|\hat{\xi}_p])\left(I_{\{x \leq \hat{\xi}_p\}} - p\right) + I_{\{x \geq \hat{\xi}_p\}}(y_i - \hat{\gamma}_p)\right\}.$$

A nonparametric method is used to estimate $[Y|\xi_p]$, the conditional expected value of Y given X at ξ_p. The nonparametric estimator (Nadaraya-Watson kernel estimator with the normal kernel and the same bandwidth h for $\hat{\xi}_h$) is given by

$$\hat{E}[Y|x] = \frac{\sum_{i \in S} w_i y_i \phi_h(x - x_i)}{\sum_{i \in S} w_i \phi_h(x - x_i)}.$$

Note that if Y and X are the same variable, then $\hat{E}[Y|x] = x$.

Case 2. γ_p is a median in the top pth percentile group

Assume that Y is a continuous nonnegative variable. Denote $f_X(x)$ and $F_X(x)$ as the marginal density and cumulative distribution function (CDF) of X and $f_Y(y)$ and $F_Y(x)$ as the marginal density and CDF of Y. Denote $F_{X|Y}(x|y)$ as the conditional CDF of X given $Y = y$ and $F_{Y|X}(y|x)$ the conditional CDF of Y given $X = x$.

The estimating function for γ_p, the median of Y in the top pth percentile group is

$$\mu_{2i} = I\{x_i \geq \xi_p\}[I\{y_i \leq \gamma_p\} - 0.5],$$

and the associated linearized variable is

$$u_{2i}^* = \frac{1}{[1 - F_{X|Y}(\xi_p|\gamma_p)]f_Y(\gamma_p)}\{[0.5 - F_{Y|X}(\gamma_p|\xi_p)](I\{x_i \leq \xi_p\} - p) + I\{x_i \geq \xi_p\}[I\{y_i \leq \gamma_p\} - 0.5]\},$$

where $f_Y(y)$, $F_{X|Y}(x|y)$ and $F_{Y|X}(y|x)$ have all been defined previously. After replacing all the population quantities by their estimates, the formula becomes

$$\hat{u}_{2i}^* = \frac{1}{[1 - \widehat{F_{X|Y}}(\hat{\xi}_p|\hat{\gamma}_p)]\hat{f}_Y(\hat{\gamma}_p)}\{[\widehat{F_{Y|X}}(\hat{\gamma}_p|\hat{\xi}_p) - 0.5](I\{x_i \leq \hat{\xi}_p\} - p) + I\{x_i \geq \hat{\xi}_p\}[I\{y_i \leq \hat{\gamma}_p\} - 0.5]\},$$

where $\widehat{F_{X|Y}}$, $\hat{f}_Y$, and $\widehat{F_{Y|X}}$ are the estimators of $F_{X|Y}$, f_Y and $F_{Y|X}$ respectively.

For the estimation of the conditional distribution $F_{X|Y}$, one can follow Borkowf et al. (1997),

$$F_{X|Y}(\xi_p|\gamma_p) = P(X \le \xi_p|Y = \gamma_p)$$
$$= P(F_X(X) \le p|F_Y(Y) = F_Y(\gamma_p)),$$

which leads to

$$\widehat{F_{X|Y}}(\xi_p|\gamma_p) = \frac{\sum_{i\in S} w_i I\{\widehat{F_X}(x_i) \le p, |\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}}{\sum_{i\in S} w_i I\{|\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}},$$

where $\hat{\delta}^2 = \frac{0.8}{n}\widehat{F_Y}(\hat{\gamma}_p)(1 - \widehat{F_Y}(\hat{\gamma}_p))$ and $z_{\alpha/2}$ is the $100\left(1 - \frac{\alpha}{2}\right)$-th percentile from the standard normal distribution. Similarly, the conditional CDF $F_{Y|X}(\gamma_p|\xi_p)$ is given by

$$\widehat{F_{Y|X}}(\hat{\gamma}_p|\hat{\xi}_p) = \frac{\sum w_i I\{\widehat{F_Y}(y_i) \le \widehat{F_Y}(\hat{\gamma}_p), |\widehat{F_X}(x_i) - p| \le z_{\alpha/2}\hat{\delta}^*\}}{\sum w_i I\{|\widehat{F_X}(x_i) - p| \le z_{\alpha/2}\hat{\delta}^*\}},$$

where $\hat{\delta}^{*2} = 0.8p(1-p)/n$.

Using the approach used by Francisco and Fuller (1991), the marginal density estimator for $h(\gamma_p)$ is given by

$$\widehat{F_Y}(\hat{\gamma}_p) \approx \frac{\sum_{i\in S} w_i I\{|\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}}{\hat{N}(Y_{max} - Y_{min})},$$

where $Y_{max} = \max\{y_i : i \in S, |\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}$ and $Y_{min} = \min\{y_i : i \in S, |\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}$. Hence,

$$[1 - \widehat{F_{X|Y}}(\hat{\xi}_p|\hat{\gamma}_p)]\widehat{f_Y}(\hat{\gamma}_p) = \frac{\sum_{i\in S} w_i I\{\widehat{F_X}(x_i) > p, |\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}}{\hat{N}(Y_{max} - Y_{min})}.$$

Case 3. γ_p is a ratio in the top pth percentile group

Within a percentile group, the proportion of income total from wages and salary is a ratio. Let Y be the variable in the numerator and Z be the variable in the denominator in this ratio, the estimating function and linear variable for the ratio are

$$\mu_{2i} = I\{x_i \ge \xi_p\}(y_i - \gamma_p z_i),$$

$$\mu_{2i}^* = \frac{1}{E[ZI(X \ge \xi_p)]}\left\{(E[Y|\xi_p] - \gamma_p\xi_p)\left(I_{\{x_i\le\xi_p\}} - p\right) + I_{\{x_i\ge\xi_p\}}(y_i - \gamma_p z_i)\right\} \quad and$$

$$\hat{\mu}_{2i}^* = \frac{1}{\hat{E}[ZI(X \ge \hat{\xi}_p)]}\left\{(\hat{E}[Y|\hat{\xi}_p] - \hat{\gamma}_p\hat{\xi}_p)\left(I_{\{x_i\le\hat{\xi}_p\}} - p\right) + I_{\{x_i\ge\hat{\xi}_p\}}(y_i - \hat{\gamma}_p z_i)\right\}.$$

where $\hat{E}[ZI(X \ge \hat{\xi}_p)] = \frac{\sum_i w_i z_i I(x_i \ge \hat{\xi}_p)}{\sum_i w_i}$.

Case 4. γ_p is a share for the top pth percentile group

The estimating function and linear variable for the share are

$$\mu_{2i} = [I\{x_i \ge \xi_p\} - \gamma_p]y_i$$

$$\mu_{2i}^* = \frac{1}{\mu_y}\{E[Y|\xi_p](I\{x_i \le \xi_p\} - p) + (I\{x_i \ge \xi_p\} - \gamma_p)y_i\}, \qquad and$$

$$\hat{\mu}_{2i}^{*} = \frac{1}{\hat{\mu}_y}\{\hat{E}[Y|\hat{\xi}_p](I\{x_i \le \hat{\xi}_p\} - p) + (I\{x_i \ge \hat{\xi}_p\} - \hat{\gamma}_p)y_i\},$$

where $\hat{\mu}_y = 1/\widehat{N}\sum_{i\in S} w_i y_i$.

Case 5. γ_p is a product of share and ratio

When γ_p is a product of a ratio (as defined in **Case 3**) and a share (as defined in **Case 4**) such that $\gamma_p = \gamma_p^{(S)}\gamma_p^{(R)}$, the estimating function and linear variable are

$$\mu_{2i} = \gamma_p^{(S)}\mu_{2i}^{(S)} + \gamma_p^{(R)}\mu_{2i}^{(R)},$$

$$\mu_{2i}^{*} = \gamma_p^{(S)}\mu_{2i}^{(S)*} + \gamma_p^{(R)}\mu_{2i}^{(R)*} \quad and$$

$$\hat{\mu}_{2i}^{*} = \hat{\gamma}_p^{(S)}\hat{\mu}_{2i}^{(S)*} + \hat{\gamma}_p^{(R)}\hat{\mu}_{2i}^{(R)*},$$

where $\mu_{2i}^{(R)}$, $\mu_{2i}^{(R)*}$, and $\hat{\mu}_{2i}^{(R)*}$ have been previously given for a ratio in **Case 3** and $\mu_{2i}^{(S)}$, $\mu_{2i}^{(S)*}$, and $\hat{\mu}_{2i}^{(S)*}$ for a share in **Case 4.**

A special case: γ_p is the count in the top pth percentile group

The count in the top pth percentile group where $\gamma_p = N(1-p)$ and $\hat{\gamma}_p = \widehat{N}(1-p)$ is a special case of parameter of interest. The variance of $\hat{\gamma}_p$ is

$$V(\hat{\gamma}_p) = (1-p)^2 V(\widehat{N}) = (1-p)^2 V\left(\sum_{i\in S} w_i\right) = (1-p)^2 \sum_{i\in U} \frac{(1-\pi_i)}{\pi_i} = 4N(1-p)^2.$$

Hence, the corresponding CV estimate is given by

$$\widehat{CV}(\hat{\gamma}_p) = \frac{\sqrt{\hat{V}(\hat{\gamma}_p)}}{\hat{\gamma}_p} = \frac{\sqrt{4\widehat{N}(1-p)^2}}{\widehat{N}(1-p)} = \frac{2}{\sqrt{\widehat{N}}}$$

5. Evaluation of the performance of variance estimators

In this section, the variance estimators are evaluated. The linear variance estimates are compared to both the approximate true variance calculated from the T1FF and the bootstrap variance estimates. This evaluation is only done for Prince Edward Island (P.E.I.) as it yields the largest variances at the provincial level.

Suppose θ is the parameter of interest and $\hat{\theta}$ is a consistent estimator of θ. The relative bias (RB) in Table 3 is defined as

$$RB(\hat{\theta}) = \frac{\hat{\theta} - \theta}{\theta} \times 100\%.$$

The approximate CV (ACV) of $\hat{\theta}$ in Table 3 is defined as

$$ACV(\hat{\theta}) = \frac{\sqrt{AV(\hat{\theta})}}{\theta} \times 100\%,$$

where $AV(\hat{\theta})$ is computed by the formula given in (10). The true parameter values are computed from the T1FF. The CV estimator on the sample is defined as

$$\widehat{CV}(\hat{\theta}) = \frac{\sqrt{\hat{V}(\hat{\theta})}}{\hat{\theta}} \times 100\%.$$

The bootstrap CV estimate is based on 1,000 bootstrap replicates (bootstrap weights) and defined as

$$\widehat{CV}^b(\hat{\theta}) = \frac{\sqrt{\hat{V}^b(\hat{\theta})}}{\hat{\theta}} \times 100\%.$$

The replicates were generated, using the pseudo-population approach (see Beaumont and Patak, 2012).

Variance estimates for selected high income statistics are produced. The income variable is the *total income*. Since the parameter estimators have large sample properties, it is expected that the CVs for other provinces should be smaller than that for P.E.I.

Table 3 lists the approximate CV and CV estimates generated from the linearization method and bootstrap method for the top 1% and 5% income group in P.E.I. Small differences are observed between the approximate CV and the other two CV estimates. The differences observed in the top 5% income group are smaller than those in the top 1% income group as the sample size in the top 5% income group is larger. CV estimates from two methods are very similar with largest difference being 0.8% for the product of share and ratio and these differences become smaller as the income group becomes larger.

The quality of the estimates in the top 5% income group is better than that in the top 1% group. The CVs and RBs generally decrease from the top 1% group to the top 5%. Exceptions include the RBs for the percentages of male and married; however, it should be noted that both decrease when the group of records under consideration is expanded.

The linearization method and bootstrap method produce very similar confidence interval (CI) estimates. This implies that the asymptotic normality assumption of the estimators is satisfied. It should be noted that the linearization intervals may be slightly shorter than the bootstrap intervals, which is very common for those two methods.

Table 3. Comparison of CV estimates and 95% Confidence Interval (C.I.) for linearization and Bootstrap for estimates for P.E.I (2009)

						CV			95% Confidence Interval					
									Population		Linear		Bootstrap	
IncGrp	Measure	Variables	Parameter	Estimate	RB	$ACV(\hat{\theta})$	$\widehat{CV}(\hat{\theta})$	$\widehat{CV}^b(\hat{\theta})$	lower	upper	lower	upper	lower	upper
Top 1%	Threshold	Income	131115	130824	-0.20%	1.90%	1.80%	2.00%	125914	135734	126209	135439	125486	135660
	Total	Counts	1098	1089	-0.80%	0.60%	0.60%	0.60%	1067	1111	1076	1102	1068	1098
	Mean	% of Male	78.20%	80.30%	2.70%	3.20%	3.00%	3.00%	73.9%	86.7%	75.6%	85.0%	75.57%	85.03%
		% of Married	82.40%	83.50%	1.30%	2.80%	2.70%	2.90%	78.5%	88.5%	79.1%	87.9%	79.01%	88.28%
		Income	214589	212371	-1.00%	5.70%	4.60%	4.60%	188031	236711	193224	231518	193297	231597
		Tax	65700	63132	-3.90%	7.40%	6.10%	6.00%	52360	73904	55564	70680	55301	70335
	Median	Income	166331	166192	-0.10%	3.00%	2.40%	2.50%	156406	175978	158700	173684	158136	172947
		Tax	48666	48923	0.50%	3.60%	3.70%	3.90%	45456	52390	45567	52279	46883	53789
		Age	53	53.5	0.90%	1.10%	1.90%	2.20%	52	55	51.5	55.5	52	56
	Ratio	Wage in Income	59.70%	63.00%	5.50%	4.80%	4.80%	4.90%	54.5%	71.5%	57.1%	68.9%	56.43%	68.56%
	Share	Income	6.60%	6.50%	-1.50%	5.00%	4.20%	4.30%	4.5%	8.2%	5.2%	7.7%	5.90%	8.1%
		Tax	11.40%	11.00%	-3.50%	6.50%	5.30%	5.30%	9.4%	12.6%	9.9%	12.1%	9.83%	12.10%
	Share By Ratio	Income share by male		5.30%			5.0%	5.8%						
Top 5%	Threshold	Income	78128	78062	-0.10%	0.90%	0.80%	0.90%	76675	79449	76838	79286	76829	79550
	Total	Counts	5489	5445	-0.80%	0.60%	0.60%	0.60%	5337	5553	5381	5509	5375	5500
	Mean	% of Male	69.30%	72.20%	4.20%	1.80%	1.80%	1.70%	66.0%	78.4%	69.7%	74.7%	69.83%	74.82%
		% of Married	80.00%	82.30%	2.90%	1.40%	1.30%	1.50%	77.3%	87.3%	80.2	84.4%	80.40%	84.45%
		Income	119222	118778	-0.40%	2.40%	2.00%	2.10%	113092	124464	114122	123434	113993	123622
		Tax	32280	31608	-2.10%	3.40%	2.90%	2.90%	29080	34136	29	33405	29709	33338
	Median	Income	97240	97491	0.30%	1.20%	1.10%	1.20%	95134	99848	95389	99593	95417	100040
		Tax	24817	24776	-0.20%	1.50%	1.30%	1.30%	232	25479	24	25385	24	25351
		Age	50	50	0.00%	0.60%	0.60%	1.00%	49	51	49.4	50.6	49	50
	Ratio	Wage in Income	70.80%	77.20%	9.00%	1.60%	1.30%	1.60%	68.6%	75.9%	69.7%	74.7%	69.36%	74.72%
	Share	Income	18.50%	18.30%	-1.10%	1.90%	1.60%	1.70%	17.5%	19.1%	17.7%	18.9%	17.69%	18.86%
		Tax	28.10 %	27.60%	-1.70%	3.40%	2.10%	2.10%	25.7%	29 %	25	29.6%	25.66%	29.6%
	Share By Ratio	Income share by male		13.60%			2.30%	2.70%						

Population CI for : $\hat{\theta} \pm 1.96\sqrt{mse(\hat{\theta})}$ where $mse(\hat{\theta}) = (\hat{\theta} - \theta)^2 + AV(\hat{\theta})$, Linear CI for θ : $\hat{\theta} \pm 1.96\sqrt{\hat{V}(\hat{\theta})}$, Bootstrap CI is based on 1000 bootstrap replicates $(2\hat{\theta} - \hat{\theta}^*_{(0.025)}, 2\hat{\theta} - \hat{\theta}^*_{(0.975)})$.

6. Estimates and CVs for selected domains and income groups

Tables 4 and 5 give the high income estimates (blue) and associated CVs (red) for reference years 2009 and 1988 respectively. These CV estimates were obtained using the linearization method.

From 1988 to 2009, the number of taxfilers has increased by 47%. In 2009, the T1FF covered approximately 75% of the Canadian population. At the Canada level, the income threshold for the top 1% has doubled between 1988 and 2009. Men or married persons always comprise the majority of the highest income group. The share of tax paid by the top 1% group increased from 13.3% to 18.4%.

For the top percentile groups, the CV for a median estimate was usually smaller than the corresponding mean estimate. The CVs for the percentages of remaining in the top 5% in last five years were all less than 5%.

In general, estimates in both tables are of good quality. The quality of estimates usually depends on the number of sampled units in the group of interest. For example, large CVs (>15%) are observed in some female top 1% groups that are often very small. For most cases, the 2009 estimates are of better quality than the 1988 estimates.

7. Summary and future work

In this paper, two methods for variance estimation have been considered for statistics in the high income tables: linearization and bootstrap. The estimates of CVs and CIs for these two methods are very close. However, in practice, the linearization method is employed as the computing time for the bootstrap method is extreme. Note that the linearization method requires the first derivative for each estimator.

Researchers are often interested in the LAD data at the provincial level or lower. For small or medium samples, the bootstrap method may be preferred because of its simplicity. For complex statistics, users of the data can produce their variance estimates or estimate their distributions easily by using the bootstrap replicates. Therefore, the feasibility of providing users with bootstrap weights when the micro data are released might be investigated.

The LAD contains information on individuals and families, while the high income tables only provide estimates on individuals. The potential of producing estimates at the family level can be investigated in the future. Since Poisson sampling is not efficient, areas such as the weighting strategy and the calibration on demographic total can be studied to improve the efficiency especially for small domain estimates.

References:

Beaumont, J.F. and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review,* 80, 127-148.

Binder, D., and Kovacevic, M.(1995). Estimating some measures of income inequality from survey data. *Survey Methodology*, 21,137-145.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Borkowf, C.B., Gail, M.H., Carroll, R., and Gill, R.D.(1997). Analyzing bivaraite continuous data grouped into categories defined by empirical quantiles of marginal distributions. Biometrics, 53, 1054 – 1069.

Francisco, C.A. and Fuller, W.A. (1991). *Quantile estimation with a complex survey design*. The Annals of Statistics, 19, 454 – 469.

Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under regression superpopulation model. *Journal of the American Statistical Association*, 77, 89 - 96.

Lohr, S. and Buskirk, T. (1999). Density estimation with complex survey data. SSC Annual Meeting, June, 1999, *Proceedings of the Survey Methods Section.*

Osier,G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, 167-195.

Randles. R.H. (1982). On the Asymptotic Normality of Statistics with Estimated Parameters. *Annals of Statistics*, 10, 462-474.

Rao, J.N.K. and Wu, C.F.J. (1987). Methods for Standard Errors and Confidence Intervals from Survey Data: Some Recent Work. *Proceedings of the 46th session, International Statistical Institute*, 3, 5-19.

Särndal, C.E., Swensson, B. and Wretman, J. (1991). Model Assisted Survey Sampling. New York: Wiley.

Shao, J. and Rao, J.N.K. (1993). Standard Errors for Low Income Proportions Estimated from Stratified Multi-Stage Samples. *Sankhya*, 55B, 393-414.

Statistics Canada. (2010). Longitudinal Administrative Data Dictionary. Catalogue no. 12-585-X.

Cc: Babyak, Colin - HSMD/DMEM
Subject: RE: CHMS sampling paper

Bonjour Christine,

Est-ce que tu as le numéro de référence pour le document de travail? J'aimerais envoyer une copie de ce document à mon client sous peu.

Merci!

Suzelle

From: Gambino, Jack - HSMD/DMEM
Sent: March-18-13 3:29 PM
To: Cousineau, Christine - HSMD/DMEM
Cc: Giroux, Suzelle - HSMD/DMEM; Babyak, Colin - HSMD/DMEM
Subject: FW: CHMS sampling paper

Merci, Suzelle.

Christine: Un autre Working Paper. J'ai déjà signé le formulaire.

From: Giroux, Suzelle - HSMD/DMEM
Sent: March 18, 2013 8:49 AM
To: Gambino, Jack - HSMD/DMEM
Cc: Babyak, Colin - HSMD/DMEM
Subject: RE: CHMS sampling paper

Bonjour Jack,

Here are the final versions we have updated with your comments.

<< File: Cycle2_Sampling_Documentation_Final_English_April17_Release.docx >>
<< File:
Cycle2_Echantillon_Documentation_Definitive_Francais_Diffusion_Avril17.docx >>

Let me know if everything is to your satisfaction.

The next step will be to ask Christine Cousineau to format these two documents in a bilingual Branch Working Paper. Can I go ahead and ask her to do so?

Thank you!

Suzelle

From: Gambino, Jack - HSMD/DMEM
Sent: March-14-13 12:15 PM
To: Giroux, Suzelle - HSMD/DMEM
Cc: Babyak, Colin - HSMD/DMEM
Subject: RE: CHMS sampling paper

I prefer electronic!

From: Giroux, Suzelle - HSMD/DMEM
Sent: March 14, 2013 10:59 AM
To: Gambino, Jack - HSMD/DMEM
Cc: Babyak, Colin - HSMD/DMEM
Subject: RE: CHMS sampling paper

Thank you Jack! We will make the updates to the paper (both in English and French).

Is it OK if I send you the electronic versions once the changes are made instead of printing another paper copy?

Let me know!

Thanks!

Suzelle

From: Gambino, Jack - HSMD/DMEM
Sent: March-14-13 10:55 AM
To: Giroux, Suzelle - HSMD/DMEM
Cc: Babyak, Colin - HSMD/DMEM
Subject: CHMS sampling paper

I've gone over the new working paper—very well done! I've signed the form. I do have some minor comments:

1. 2.1.4: I believe a CMA has to have a core of 50K and an overall population of 100K.
2. 2.3.1: "households where all persons were under age 3" sounds strange/funny; how about "households where all *in-scope* persons were under age 3"?
3. 3.1.3, para. 6 (and elsewhere): the increase in efficiency was 6.6 *percentage points*, not 6.6%! (you could also say 9.6% since 6.6/68.6 = .0962) [I realize this is another one of my losing battles on language.]

Jack Gambino
Director / Directeur
Household Survey Methods / Méthodes d'enquêtes auprès des ménages
Statistics Canada / Statistique Canada
R.H. Coats 16 O
613-951-0334 | fax 613-951-3100
gambino@statcan.gc.ca

Table 4: CVs for selected estimates in High income tables for RY2009, (red indicates the CV)

GEO	IncGrp	Threshold	NFilers	Age	%					XTIRC					TAX			WAGE			%			
				Med	Male	Female	Married	Mar_ Male	Mar_ Female	Avg	Med	Shr	Shr_ Male	Shr_ Female	Avg	Med	Shr	Shr	Shr_ Male	Shr_ Female	Top5_ Ym1	Top5_ Ym5	Top5_ Ever	Top5_ Always
Canada	Top 0.1%	673800	25250	53	86.3%	13.7%	85.4%	88.0%	68.7%	1512737	1026128	3.8%	3.3%	0.5%	543308	384053	6.8%	62.0%	65.0%	41.7%	96.6%	92.2%	98.4%	88.6%
		0.7%	0.04%	0.3%	0.5%	3.2%	0.5%	0.5%	2.3%	1.5%	1.0%	1.5%	1.5%	4.9%	1.6%	1.1%	1.4%	1.3%	1.3%	4.7%	0.2%	0.3%	0.1%	0.4%
	Top 1%	198000	252320	51	79.2%	20.8%	82.8%	86.8%	67.9%	425140	278804	10.7%	8.7%	1.9%	146898	94375	18.4%	62.4%	65.6%	47.8%	94.1%	84.3%	96.8%	79.6%
		0.2%	0.04%	0.1%	0.2%	0.8%	0.2%	0.2%	0.6%	0.6%	0.3%	0.6%	0.6%	1.5%	0.7%	0.4%	0.6%	0.5%	0.5%	1.3%	0.1%	0.2%	0.1%	0.2%
Prince Edward Island	All Filers	0	108910	48	47.6%	52.4%	57.5%	60.5%	54.8%	32471	26193	1	54.5%	45.5%	5756	3035	1	64.1%	64.7%	63.3%	5.1%	3.2%	9.9%	2.9%
			0.6%	0.3%	0.6%	0.6%	0.5%	0.7%	0.8%	0.6%	0.5%	0	0.8%	0.9%	1.1%	1.5%	0	0.5%	0.8%	0.7%	2.5%	3.1%	1.7%	3.2%
	Top 1%	130800	1090	53.5	80.3%	19.7%	83.5%	86.3%	72.1%	212370	166192	6.5%	5.3%	1.3%	63132	48923	11.0%	63.0%	64.5%	56.4%	91.7%	82.2%	95.1%	75.9%
		1.8%	0.6%	1.9%	3.0%	12.1%	2.7%	2.7%	8.2%	4.6%	2.3%	4.2%	5.0%	15.0%	6.1%	3.0%	5.3%	4.8%	5.3%	10.9%	1.7%	2.5%	1.2%	3.0%
	Top 5%	78100	5450	50	72.2%	27.8%	82.3%	86.5%	71.3%	118778	97491	18.3%	13.6%	4.7%	31608	24735	27.5%	72.2%	72.4%	72.2%	76.4%	57.3%	85.7%	46.9%
		0.8%	0.6%	0.8%	1.8%	4.6%	1.3%	1.3%	3.3%	2.0%	1.3%	1.6%	2.3%	5.7%	2.9%	1.0%	2.1%	1.8%	3.3%	2.1%	2.0%	2.8%	1.2%	3.0%
	Top 10%	63400	10895	50	62.9%	37.1%	79.6%	85.0%	70.4%	94302	78062	29.1%	19.4%	9.7%	23746	18871	41.3%	75.0%	72.9%	79.3%	43.5%	28.5%	57.1%	24.9%
		0.6%	0.6%	0.6%	1.5%	2.5%	1.0%	1.0%	2.0%	1.4%	0.9%	1.0%	1.6%	3.4%	2.0%	0.9%	1.2%	1.2%	1.6%	1.7%	2.6%	3.1%	1.8%	3.2%
Manitoba	All Filers	0	888670	46	48.1%	51.9%	56.9%	59.2%	54.8%	36355	28097	1	56.9%	43.1%	7094	3277	1	68.1%	71.0%	64.3%	5.1%	3.3%	9.1%	3.1%
			0.2%	0.1%	0.2%	0.2%	0.2%	0.3%	0.3%	0.3%	0.2%	0	0.3%	0.4%	0.7%	0.5%	0	0.2%	0.3%	0.3%	0.9%	1.0%	0.6%	1.1%
	Top 1%	156900	8890	52	80.3%	19.7%	84.0%	87.5%	69.7%	309297	214939	8.5%	6.9%	1.6%	108749	70209	15.3%	65.5%	68.3%	53.8%	93.9%	82.6%	96.9%	77.3%
		1.0%	0.2%	0.5%	1.1%	4.3%	0.9%	0.9%	3.2%	3.0%	1.3%	2.8%	3.6%	8.1%	4.1%	1.6%	3.5%	2.1%	2.1%	7.9%	0.5%	0.8%	0.3%	1.0%
	Top 5%	90100	44440	50	73.3%	26.7%	80.4%	84.5%	69.1%	150479	112024	20.7%	15.7%	5.0%	45198	30691	31.9%	74.1%	76.0%	68.4%	76.6%	57.9%	85.5%	47.2%
		0.3%	0.2%	0.3%	0.6%	1.6%	0.5%	0.5%	1.2%	1.3%	0.4%	1.0%	1.4%	3.0%	2.1%	0.6%	1.4%	0.8%	0.9%	1.9%	0.7%	0.9%	0.4%	1.0%
	Top 10%	72200	88880	49	67.5%	32.5%	78.0%	82.5%	68.5%	115123	90145	31.7%	22.4%	9.2%	32102	22874	45.3%	77.5%	78.4%	75.5%	44.2%	28.9%	56.7%	25.1%
		0.2%	0.2%	0.2%	0.5%	1.0%	0.4%	0.4%	0.8%	0.9%	0.3%	0.6%	0.9%	1.8%	1.5%	0.5%	0.8%	0.5%	0.6%	1.0%	0.9%	1.0%	0.6%	1.1%
St. John's (NL)	All Filers	0	146900	45	47.5%	52.5%	55.1%	58.2%	52.3%	40935	30277	1	56.9%	43.1%	8354	3906	1	72.4%	74.6%	69.5%	5.1%	3.3%	9.2%	3.0%
			0.5%	0.3%	0.5%	0.5%	0.5%	0.6%	0.7%	0.7%	0.7%	0	0.7%	0.9%	1.2%	1.2%	0	0.4%	0.6%	0.6%	2.2%	2.6%	1.5%	2.7%
	Top 1%	199900	1470	51	80.3%	19.7%	86.4%	88.6%	77.6%	347401	281267	8.5%	6.7%	1.8%	121906	91491	14.6%	68.3%	69.6%	63.6%	92.0%	82.6%	95.8%	78.0%
		2.6%	0.5%	1.3%	2.6%	10.6%	2.1%	2.1%	6.4%	4.9%	3.7%	4.4%	6.3%	14.1%	6.3%	3.3%	5.2%	3.5%	3.8%	9.0%	1.5%	2.3%	1.0%	2.7%
	Top 5%	106000	7345	49	78.5%	21.5%	83.1%	86.2%	71.5%	176794	135857	21.6%	17.0%	4.6%	53819	38493	32.2%	76.4%	78.6%	68.1%	77.4%	60.5%	88.1%	48.8%
		1%	0.5%	1%	1.2%	4.5%	1.1%	1.1%	3.2%	2.3%	1.4%	1.7%	2.6%	6.3%	3.2%	1.6%	2.1%	1.4%	1.5%	3.7%	1.7%	2.2%	0.9%	2.5%
	Top 10%	81400	14690	47	72.4%	27.6%	80.7%	84.3%	71.1%	134009	105970	32.7%	24.4%	8.3%	38185	27432	45.7%	79.9%	81.2%	76.0%	44.8%	30.0%	60.1%	25.6%
		0.4%	0.5%	0.5	1.1%	2.8%	0.8%	0.8%	2.0%	1.6%	0.9%	1.0%	1.7%	3.9%	2.4%	1.4%	1.3%	0.9%	1.1%	1.9%	2.2%	2.6%	1.5%	2.7%
Sherbrooke (QC)	All Filers	0	153890	49	47.7%	52.3%	54.9%	57.7%	52.3%	33714	25797	1	56.0%	44.0%	7140	2633	1	61.4%	63.5%	58.7%	5.1%	3.5%	8.5%	3.3%
			0.5%	0.3%	0.5%	0.5%	0.5%	0.6%	0.7%	0.6%	0.8%	0	0.7%	0.8%	1.1%	1.4%	0	0.5%	0.8%	0.7%	2.2%	2.5%	1.5%	2.6%
	Top 1%	150000	1540	51	77.6%	22.4%	83.4%	87.9%	68.1%	268908	216134	8.0%	6.5%	1.5%	112243	89816	15.7%	41.0%	44.8%	24.8%	94.8%	83.6%	96.0%	80.1%
		2.2%	0.5%	1.7%	2.7%	9.5%	2.3%	2.2%	7.4%	3.8%	3.6%	3.4%	3.8%	12.0%	4.6%	4.5%	3.8%	5.8%	6.2%	13.2%	1.2%	2.0%	0.9%	2.2%
	Top 5%	81800	7695	50	71.5%	28.5%	76.7%	80.8%	66.5%	136018	104789	20.2%	15.0%	5.1%	48273	33040	33.8%	61.6%	62.4%	59.2%	79.5%	62.8%	87.4%	53.4%
		0.8%	0.5%	0.6%	1.4%	3.6%	1.3%	1.3%	3.0%	1.9%	1.1%	1.5%	1.8%	4.9%	2.6%	1.4%	1.7%	2.0%	2.4%	3.6%	1.6%	2.1%	1.0%	2.4%
	Top 10%	65700	15390	49	66.2%	33.8%	74.0%	79.0%	64.4%	104344	81769	31.0%	21.6%	9.3%	34230	24363	47.9%	68.4%	67.2%	71.1%	45.9%	31.3%	58.1%	28.1%
		0.6%	0.5%	0.8%	1.2%	2.3%	1.0%	1.0%	2.1%	1.3%	0.6%	0.9%	1.3%	3.2%	2.0%	0.8%	1.0%	1.3%	1.7%	2.0%	2.2%	2.5%	1.5%	2.6%

Table 5: CVs for selected estimates in High income tables for RY1988, (red indicates the CV)

GEO	IncGrp	Threshold	NFilers	age	%					XTIRC					TAX			WAGE			%			
				Med	Male	Female	Married	Mar_ Male	Mar_ Female	Avg	Med	Shr	Shr_ Male	Shr_ Female	Avg	Med	Shr	Shr	Shr_ Male	Shr_ Female	Top5_ Ym1	Top5_ Ym5	Top5_ Ever	Top5_ Always
Canada	Top 0.1%	271911	17165	51	90.4%	9.6%	84.5%	87.2%	59.1%	580838	390955	2.6%	2.4%	0.2%	235354	155323	4.8%	66.9%	68.7%	47.9%	95.5%	85.3%	98.5%	75.9%
		0.8%	0.05%	0.5%	0.5%	4.7%	0.7%	0.6%	4.1%	2.0%	1.2%	2.0%	2.0%	7.1%	2.2%	1.4%	2.1%	1.3%	1.4%	6.4%	0.3%	0.5%	0.2%	0.7%
	Top 1%	94198	171615	48	87.6%	12.4%	82.8%	86.4%	57.5%	180623	127675	8.1%	7.2%	0.9%	65269	43546	13.3%	58.8%	60.7%	42.9%	91.8%	76.6%	96.0%	67.2%
		0.2%	0.05%	0.2%	0.2%	1.3%	0.2%	0.2%	1.2%	0.7%	0.3%	0.7%	0.7%	2.2%	0.9%	0.4%	0.8%	0.6%	0.6%	2.1%	0.1%	0.2%	0.1%	0.3%
Prince Edward Island	All Filers	0	79840	37	51.0%	49.0%	59.2%	61.2%	57.1%	17503	13910	1	62.7%	37.3%	3088	1703	1	63.3%	63.7%	62.6%	5.0%	3.3%	9.2%	2.8%
			0.7%	0.4%	0.7%	0.7%	0.6%	0.8%	0.9%	0.7%	0.6%	0.0%	0.7%	1.2%	1.4%	1.6%	0.0%	0.7%	0.9%	0.8%	3.0%	3.4%	1.9%	3.6%
	Top 1%	69748	800	47	90.0%	10.0%	90.6%	91.7%	81.3%	115248	85508	6.6%	6.0%	0.6%	36453	25368	11.8%	51.1%	51.3%	48.2%	88.6%	70.9%	95.7%	57.4%
		2.4%	0.7%	1.9%	2.5%	22.1%	2.3%	2.3%	10.7%	6.2%	2.3%	5.7%	6.4%	28.8%	8.3%	3.1%	7.2%	7.7%	8.1%	25.3%	2.4%	3.7%	1.2%	4.8%
	Top 5%	42840	3995	45	83.2%	16.8%	83.0%	89.3%	51.5%	64130	52265	18.3%	15.6%	2.7%	17664	13537	28.6%	61.2%	61.7%	58.2%	74.0%	53.3%	84.1%	41.0%
		0.8%	0.7%	1.1%	1.4%	7.2%	1.4%	1.2%	7.5%	2.6%	1.3%	2.1%	2.4%	9.1%	3.8%	1.5%	2.7%	2.8%	3.1%	6.2%	2.4%	2.9%	1.3%	3.4%
	Top 10%	34287	7985	43	78.6%	21.4%	81.0%	86.8%	59.8%	51127	42840	29.2%	23.7%	5.5%	13176	10284	42.7%	67.9%	67.2%	70.9%	42.0%	26.6%	54.6%	21.1%
		0.8%	0.7%	0.7%	1.2%	4.3%	1.1%	1.0%	4.0%	1.7%	0.7%	1.2%	1.5%	5.5%	2.7%	1.2%	1.5%	1.7%	2.0%	2.9%	3.0%	3.4%	1.9%	3.6%
Manitoba	All Filers	0	741800	39	50.5%	49.5%	56.1%	59.6%	52.5%	19364	14841	1	63.0%	37.0%	3785	1689.5	1	67.5%	69.9%	63.4%	5.1%	3.3%	9.5%	2.9%
			0.2%	0.2%	0.2%	0.2%	0.2%	0.3%	0.3%	0.3%	0.3%	0	0.2%	0.4%	0.5%	0.8%	0	0.2%	0.3%	0.3%	1.0%	1.1%	0.6%	1.2%
	Top 1%	76608	7420	51	88.6%	11.4%	84.4%	87.6%	59.8%	129475	98492	6.7%	6.0%	0.7%	43645	32041	11.5%	52.9%	54.6%	37.2%	92.1%	78.5%	95.4%	71.2%
		0.8%	0.2%	1%	0.8%	6.5%	1.0%	0.9%	5.7%	2.8%	1.3%	2.6%	2.6%	9.0%	3.3%	1.6%	2.9%	3.0%	3.2%	7.8%	0.6%	1.0%	0.4%	1.2%
	Top 5%	47922.5	37090	46	86.0%	14.0%	81.0%	85.8%	51.8%	71393	57724	18.4%	16.0%	2.4%	21615	16644	28.6%	68.6%	71.0%	52.5%	74.4%	54.7%	84.3%	42.9%
		0.2%	0.2%	0.3%	0.4%	2.6%	0.5%	0.5%	2.7%	1.1%	0.4%	0.9%	0.9%	3.4%	1.4%	0.5%	1.0%	1.0%	1.1%	2.4%	0.8%	1.0%	0.4%	1.1%
	Top 10%	39432	74185	44	81.5%	18.5%	78.8%	84.1%	55.2%	57334	47922	29.6%	24.7%	4.9%	16451	13180	43.5%	74.6%	76.2%	66.5%	43.0%	27.4%	56.5%	22.6%
		0.2%	0.2%	0.2%	0.4%	1.6%	0.4%	0.4%	1.5%	0.7%	0.2%	0.5%	0.5%	2.1%	1.0%	0.4%	0.6%	0.6%	0.7%	1.2%	1.0%	1.1%	0.6%	1.2%
St. John's (NL)	All Filers	0	104060	35	49.8%	50.2%	56.5%	60.0%	53.1%	20341	15812	1	63.5%	36.5%	4225	2303	1	75.2%	76.9%	72.3%	5.1%	3.5%	9.1%	3.1%
			0.6%	0.4%	0.6%	0.6%	0.5%	0.7%	0.8%	0.8%	0.7%	0	0.7%	1.2%	1.4%	1.8%	0	0.4%	0.6%	0.6%	2.6%	2.8%	1.6%	2.9%
	Top 1%	86010	1045	47	90.9%	9.1%	88.5%	91.6%	57.9%	169135	117040	8.4%	7.7%	0.6%	58398	39976	13.9%	65.6%	68.0%	34.8%	93.8%	84.0%	98.3%	75.4%
		2.3%	0.6%	1.9%	2.0%	19.7%	2.2%	2.0%	17.5%	6.9%	4.5%	6.3%	6.2%	26.0%	8.0%	4.4%	6.8%	4.2%	4.0%	34.6%	1.5%	2.2%	0.7%	2.8%
	Top 5%	50268	5205	45	86.3%	13.7%	84.6%	88.6%	59.4%	83029	62631	20.4%	18.0%	2.4%	25862	18115	30.6%	73.6%	76.2%	54.2%	81.1%	56.8%	88.2%	45.8%
		0.8%	0.6%	1.0%	1.1%	6.9%	1.2%	1.1%	6.1%	3.1%	1.4%	2.4%	2.4%	9.6%	3.9%	1.5%	2.7%	1.7%	1.7%	7.1%	1.7%	2.5%	1.0%	2.7%
	Top 10%	40080	10415	43	81.9%	18.1%	82.9%	88.3%	58.2%	63655	50261	31.3%	26.5%	4.8%	18667	13858	44.2%	78.3%	80.1%	68.4%	45.9%	28.3%	56.6%	23.3%
		0.6%	0.6%	0.6%	0.9%	4.3%	0.9%	0.8%	3.9%	2.1%	0.8%	1.4%	1.5%	5.9%	2.8%	1.1%	1.6%	1.1%	1.1%	3.3%	2.6%	2.8%	1.7%	2.9%
Sherbrooke (QC)	All Filers	0	87070	38	49.2%	50.8%	48.0%	53.7%	42.4%	19406	15530	1	62.4%	37.6%	4666	2497	1	71.9%	74.4%	67.7%	5.0%	3.4%	9.8%	2.8%
			0.7%	0.4%	0.7%	0.7%	0.7%	0.9%	1.1%	0.6%	0.8%	0	0.7%	1.1%	1.0%	2.0%	0	0.5%	0.7%	0.7%	2.8%	3.1%	1.7%	3.3%
	Top 1%	77537	875	48	90.3%	9.7%	73.1%	77.2%	35.3%	112174	98450	5.8%	5.3%	0.5%	43655	38611	9.4%	46.0%	47.3%	32.8%	95.2%	78.7%	98.7%	73.1%
		2.3%	0.7%	1.7%	2.2%	20.8%	4.2%	3.9%	29.5%	3.3%	3.0%	3.0%	3.4%	22.1%	4.2%	3.4%	3.7%	6.9%	7.1%	24.8%	1.4%	2.7%	0.6%	3.0%
	Top 5%	47444	4355	45	86.6%	13.4%	73.8%	78.6%	42.7%	67910	57356	17.5%	15.4%	2.1%	24483	20407	26.2%	67.8%	68.9%	60.1%	73.4%	52.9%	86.3%	39.0%
		0.7%	0.7%	0.8%	1.2%	8.0%	1.8%	1.7%	9.6%	1.6%	1.1%	1.3%	1.7%	8.6%	2.1%	1.5%	1.5%	2.1%	2.2%	6.3%	2.4%	2.8%	1.2%	3.2%
	Top 10%	38213	8715	44	81.6%	18.4%	70.7%	77.4%	41.3%	55139	47439	28.4%	23.7%	4.7%	18895	15927	40.5%	74.4%	75.0%	71.3%	42.0%	26.4%	56.3%	20.4%
		0.7%	0.7%	0.6%	1.0%	4.6%	1.4%	1.3%	6.0%	1.2%	0.7%	0.8%	1.2%	5.1%	1.6%	1.1%	1.0%	1.3%	1.4%	3.0%	2.9%	3.1%	1.7%	3.4%

Appendix: Deriving linear variables

Let $U = \{1,2, \dots, N\}$ be the index of a finite population of size N and (x_i, y_i) be the value of the variable vector (X, Y) attached to unit i, for $i \in U$, where (X, Y) are two nonnegative variables and X is continuous. Let ξ_p be the p[th] percentile of X for the finite population and γ_p be a quantity of interest of Y for units in the sub-population $U_p = \{i: i \in U, x_i \geq \xi_p\}$. Let S be a sample selected from U under a certain design. Denote $(\hat{\xi}_p, \hat{\gamma}_p)$ as the sample estimators of (ξ_p, γ_p).

Further, assume that $\{(x_i, y_i) | i \in U\}$ is a random sample from a super-population model with joint probability distribution function $F(x, y)$ and joint density function $f(x, y)$. Define the joint empirical distribution function (EDF) F_N based on the finite population as

$$F_N(x, y) = \frac{1}{N} \sum_{i \in U} I_{\{x_i \leq x, y_i \leq y\}}.$$

If N is unknown, the pseudo joint EDF $\hat{F}_N$ based on the sample is given as

$$\hat{F}_N(x, y) = \frac{1}{\hat{N}} \sum_{i \in S} w_i I_{\{x_i \leq x, y_i \leq y\}},$$

where $I_A = 1$ if $i \epsilon A$; otherwise $I_A = 0$.

Under regularity conditions and certain conditions for complex sampling design (see, Isaki and Fuller, 1982 and Krewski and Rao, 1981),

$$F_N(x, y) \xrightarrow{p} F(x, y) \ for \ any \ (x, y);$$

$$\hat{F}_N(x, y) - F_N(x, y) \xrightarrow{p} 0 \ for \ any \ (x, y).$$

For the estimating equations approach, finite population parameters $\theta = (\xi_p, \gamma_p)$ are viewed as a solution of "census" estimating equations for super-population model parameters,

$$U(\theta) = \sum_{i \in U} u(\theta; x_i, y_i) = 0 \ \ or \ \int u(\theta; x, y) dF_N$$

and the estimators $\hat{\theta} = (\hat{\xi}_p, \hat{\gamma}_p)$ can be found by solving the corresponding sample weighted estimating equations

$$\hat{U}(\theta) = \sum_{i \in S} w_i u(\theta; x_i, y_i) = 0 \ \ or \ \int u(\theta; x, y) d\hat{F}_N = 0.$$

Suppose θ_0 is the true parameter value, it is already known that

$$\hat{\theta} - \theta_0 \xrightarrow{p} 0,$$

where "$\xrightarrow{p}$" means convergence in probability under the design.

Thus, the weighted estimating equations can be decomposed as

$$0 = \int u(\hat{\theta}; x, y) d\hat{F}_N$$

$$= \int [u(\hat{\theta}; x, y) - u(\theta_0; x, y)] dF + \int u(\theta_0; x, y) d\hat{F}_N + R,$$

where the remainder $R = \frac{1}{N}\{(\hat{U}(\hat{\theta}) - E[u(\gamma, Y)]|_{\gamma=\hat{\theta}}) - (\hat{U}(\theta_0) - E[u(\theta_0, Y)])\}$. Randles (1982) and Shao and Rao (1993) have shown $R = o_p(|\hat{\theta} - \theta_0|)$ under some regularity conditions. Thus, R is negligible for large samples.

The above approximation leads to

$$\hat{\theta} - \theta_0 \approx -\left[\frac{\partial E[u(\theta; X, Y)]}{\partial \theta}\right]^{-1}_{\theta=\theta_0} \int u(\theta_0; x, y) d\hat{F}_N$$

$$\approx \sum_{i \in S} w_i u^*(\theta_0; x_i, y_i)$$

where

$$u^*(\theta_0; x, y) = -\frac{1}{N}\left[\frac{\partial E[u(\theta; X, Y)]}{\partial \theta}\right]^{-1}_{\theta=\theta_0} u(\theta_0; x, y).$$

For the first case where ξ_p is the p[th]-percentile of X and γ_p is the average of Y for units in U_p, the top p[th] percentile group, the linear variables are derived below.

The estimating function for (ξ_p, γ_p) are $u = (u_1, u_2)^T$ where

$$u_1 = I_{\{x \le \xi_p\}} - p,$$

$$u_2 = I_{\{x \ge \xi_p\}}(y - \gamma_p).$$

Then,

$$\frac{\partial E[u(\theta; X, Y)]}{\partial \theta} = \begin{pmatrix} \frac{\partial E[u_1]}{\partial \xi_p} & \frac{\partial E[u_1]}{\partial \gamma_p} \\ \frac{\partial E[u_2]}{\partial \xi_p} & \frac{\partial E[u_2]}{\partial \gamma_p} \end{pmatrix}$$

where $\frac{\partial E[u_1]}{\partial \xi_p} = f(\xi_p)$, $\frac{\partial E[u_1]}{\partial \gamma_p} = 0$,

$$\frac{\partial E[u_2]}{\partial \xi_p} = \frac{\partial E\left[I_{\{x \ge \xi_p\}} Y\right]}{\partial \xi_p} + \gamma_p f(\xi_p) = \frac{\partial \int_{\xi_p}^{\infty} \int_0^{\infty} y f(x, y) dy dx}{\partial \xi_p} + \gamma_p f(\xi_p)$$

$$= -\int_0^{\infty} y f(\xi_p, y) dy + \gamma_p f(\xi_p) = (\gamma_p - E[Y|\xi_p]) f(\xi_p).$$

and $\frac{\partial E[u_2]}{\partial \gamma_p} = P(x \ge \xi_p) = 1 - p$.

In a partition matrix,

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}$$

where A_{11} and A_{22} are non-singular, the inverse of A is given by

$$A^{-1} = \begin{pmatrix} A_{11}^{-1} & 0 \\ -A_{22}^{-1}A_{21}A_{11}^{-1} & A_{22}^{-1} \end{pmatrix}.$$

Therefore,

$$u_1^* = -\frac{1}{Nf(\xi_p)}\left[I_{\{x \le \xi_p\}} - p\right]$$

$$u_2^* = \frac{1}{N(1-p)}\left\{\left(\gamma_p - E[Y|\xi_p]\right)\left[I_{\{x \le \xi_p\}} - p\right] + I_{\{x \ge \xi_p\}}(y - \gamma_p)\right\}.$$

By replacing the unknown quantities in u_1^* and u_2^* by their proper estimates, the following formula are derived

$$\hat{u}_{1i}^* = -\frac{1}{\hat{N}\hat{f}(\hat{\xi}_p)}\left[I_{\{x_i \le \hat{\xi}_p\}} - p\right]$$

$$\hat{u}_{2i}^* = \frac{1}{\hat{N}(1-p)}\left\{\left(\hat{\gamma}_p - \hat{E}[Y|\hat{\xi}_p]\right)\left[I_{\{x_i \le \hat{\xi}_p\}} - p\right] + I_{\{x_i \le \hat{\xi}_p\}}(y_i - \hat{\gamma}_p)\right\}.$$

For the estimation of $E[Y|\xi_p]$, the non-parametric estimator (weighted Nadaraya-Watson kernel estimator) is

$$\hat{E}[Y|x] = \sum_{i \in S} l_i(x) y_i$$

where $l_i(x) = \frac{w_i K\left(\frac{x_i - x}{h}\right)}{\sum_{k \in S} w_k K\left(\frac{x_k - x}{h}\right)}$ with $h = 0.79 n^{-\frac{1}{5}} \hat{Q}$ ($\hat{Q}$ is the sample IQR) and $K(t) = \exp\left(-\frac{t^2}{2}\right)$.

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010532207