

TREATMENTS FOR LINK NONRESPONSE IN INDIRECT SAMPLING

Xiaojian Xu¹, Pierre Lavallée²

ABSTRACT

We consider overcoming the overestimation in using generalized weight share method (GWSM) caused by link nonresponse in Indirect Sampling. Adjustment methods incorporating link nonresponse in using GWSM have been constructed. A simulation study on a longitudinal survey is presented using these adjustment methods. The simulation results show that these adjustments to the GWSM perform well in both reducing estimation bias and variance. The advancement in bias reduction is significant.

KEY WORDS: Generalized Weight Share Method, nonresponse, Indirect Sampling, Longitudinal survey.

RÉSUMÉ

On considère la correction de la surestimation reliée à la Méthode généralisée du partage des poids (MGPP) causée par la non-réponse de liens dans le sondage indirect. Des méthodes d'ajustement pour la non-réponse de liens reliée à la MGPP ont été développées. On présente une étude par simulation utilisant ces méthodes d'ajustement sur une enquête longitudinale. Les résultats des simulations montrent que ces ajustements à la MGPP donnent de bons résultats en réduisant le biais d'estimation et la variance. L'avancement sur la réduction du biais est significatif.

MOTS CLÉS : Enquête longitudinale; méthodes généralisée du partage des poids; non-réponse; sondage indirect.

1. INTRODUCTION

Indirect Sampling refers to selecting samples from a population which is not, but it is related to, the target population of interest. Such sampling scheme is carried out often when we do not have a sampling frame for the target population, but have a sampling frame for another population that is related to it. We call the latter the sampling population.

There is a sizeable amount of literature concerning estimation problems that are associated with Indirect Sampling. We name a few here. Initially, estimation methods for production of cross-sectional estimates using longitudinal household survey are discussed in Ernst (1989). He presented a weight share method in the context of longitudinal survey, and also showed that this method provides an unbiased estimator for the total for any characteristic in the population of interest. Kalton and Brick (1995) concluded that such a method also provides minimal variance of estimated population total for some simple sampling schemes for a longitudinal household panel survey. Lavallée (1995) extended the weight share method in a completely general context of Indirect Sampling which includes longitudinal survey as a particular example. The new method is called the Generalized Weight Share Method (GWSM). This weighting scheme provides unbiased estimates, irrespective of sampling schemes for obtaining the sample from the sampling population. In implementing the GWSM, adjustments for a variety of nonresponse problems have to be done as any other weighting scheme. Lavallée (2001) provided an adjusted GWSM, incorporating possible total nonresponse problems in Indirect Sampling. With Indirect Sampling, one type of nonresponse is called link nonresponse that is associated with the situation where it is impossible or failed to determine whether a unit in the sampling population is related or not to a unit in the target population. Lavallée (2001) pointed out the problem of overestimation when using GWSM with link nonresponse, and left finding suitable adjustments for the GWSM with link nonresponse as a rather open question. This present study focuses on developing treatments of estimation bias caused by such link nonresponse.

¹ Department of Mathematics, Brock University, St. Catharines, Ontario, CANADA L2S 3A1, email: xxu@brocku.ca

² Social Survey Methods Division, Statistics Canada Ottawa, Ontario, Canada K1A 0T6, email: pierre.lavallee@statcan.ca.

In the present paper, the notation and the problem are described in Section 2. In Section 3, we propose a few methods to modify the GWSM, incorporating link nonresponse. A simulation study using a real life data set is presented in Section 4 while a closing remark is stated in Section 5.

2. NOTATION AND PROBLEM

Let us define the following notation:

U^B --- The target population without any known sampling frame.

U^A --- The population related to U^B , with a known sampling frame.

s^A --- A selected sample from U^A .

M^A --- The number of units in U^A .

m^A --- The number of units in s^A .

π_j^A --- The selection probability of the j th unit in U^A , with $\pi_j^A > 0$ and $\sum_{j=1}^{M^A} \pi_j^A = m^A$.

M^B --- The number of units in U^B .

N --- The number of clusters in U^B .

U_i^B --- The i th cluster of U^B with $\bigcup_{i=1}^N U_i^B = U^B$.

M_i^B --- The number of units in the i th cluster U_i^B .

$l_{j,ik}$ --- An indicator variable of link existence: $l_{j,ik} = 1$ indicates that there is a link between the j th unit in U^A and the k th unit in U_i^B , while $l_{j,ik} = 0$ otherwise.

$L_{j,i}^B$ --- The total number of links existing between unit j of U^A and units of U_i^B , i. e., $L_{j,i}^B = \sum_{k=1}^{M_i^B} l_{j,ik}$.

L_i^B --- The total number of links existing between the units of U^A and the units of U_i^B , i. e., $L_i^B = \sum_{j=1}^{M^A} L_{j,i}^B$.

y_{ik} --- The value of the characteristics for the k th unit of the i th cluster in population U^B .

Y^B --- The total of all y_{ik} 's, i. e. $Y^B = \sum_{i=1}^N \sum_{k=1}^{M_i^B} y_{ik}$.

Ω^B --- The clusters in U^B where there is at least one unit ik such that $l_{j,ik} = 1$ for some j th unit in s^A . The set Ω^B contains the clusters of U^B identified by the units from s^A .

n --- The number of clusters in Ω^B .

w_{ik} --- The estimation weight assigned to the k th unit of the i th cluster.

t_j --- The indicator variable of being selected in s^A : $t_j = 1$ indicates that the j th unit in U^A is in s^A , and 0 otherwise.

Ω^A --- The set of units in U^A that have links to some units in Ω^B .

t_j^L --- The indicator variable of being included in s^A for units in Ω^A : $t_j^L = 1$ indicates that the j th unit in Ω^A is in s^A , and 0 otherwise.

T^A --- The number of units in Ω^A .

Ω_i^A --- The set of units in U^A that have links to some units in U_i^B with $i \in \Omega^B$.

T_i^A --- The number of units in Ω_i^A .

s_i^A --- The set of units in s^A that have links to some units in U_i^B with $i \in \Omega^B$.

m_i^A --- The number of units in s_i^A .

$t_{j,i}^L$ --- The indicator variable of being included in s_i^A for units in Ω_i^A : $t_{j,i}^L = 1$ indicates that the j th unit in Ω_i^A is in s_i^A , and 0 otherwise.

Our goal is to estimate the total Y^B for the target population U^B that is divided into N clusters. In order to do so, we select a sample s^A from U^A with selection probability π_j^A . Then, we identify Ω^B using the non-zero links $l_{j,ik}$. All units of the clusters in Ω^B are surveyed, where y_{ik} and the set of links $l_{j,ik}$ are measured.

We are interested in determining an estimation weight w_{ik} to be assigned to each unit k of each surveyed cluster i . Such weights should be chosen in an appropriate manner so that the estimator of Y^B ,

$$\hat{Y}^B = \sum_{i=1}^n \sum_{k=1}^{M_i^B} w_{ik} y_{ik} \quad (1)$$

performs well in estimating Y^B .

We are interested in estimating the quantity Y^B using \hat{Y}^B . According to Horvitz and Thompson (1952), let w_{ik} be inverse of selection probability π_{ik} of the k th individual of U_i^B in the target population. Then \hat{Y}^B provides an unbiased estimator for Y^B . However, the computation for π_{ik} is often difficult or even impossible in the context of Indirect Sampling. Therefore, the GWSM is introduced to address this issue. The following steps describe the GWSM:

Step 1: Compute the initial weights w'_{ik} :

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}; \quad (2)$$

Step 2: Compute L_i^B :

$$L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik}; \quad (3)$$

Step 3: Obtain final weight w_i :

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{L_i^B}; \quad (4)$$

Step 4: Set $w_{ik} = w_i$ for all units k in the i th cluster.

It follows (see Theorem in Section 3 of Lavallée, 2001) that

$$\hat{Y}^B = \sum_{i=1}^n \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{L_i^B} \sum_{k=1}^{M_i^B} y_{ik} \quad (5)$$

is an unbiased estimator for Y^B , provided that all links $l_{j,ik}$ can be correctly identified and that $L_i^B > 0$ for all clusters in U^B . The estimation weights assigned in (5) are

$$w_{ik} = \begin{cases} \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{L_i^B}, & \text{for all units } k \text{ in cluster } i \text{ when } i \text{ in } \Omega^B; \\ 0, & \text{when } i \text{ is not in } \Omega^B. \end{cases} \quad (6)$$

When link nonresponse occurs, as indicated in Lavallée (2001), L_i^B can not be determined. Traditionally, using only the observed links in (5) results in overestimation of Y^B , since some link components are actually missing in summation of the total L_i^B . Our study focuses on such a problem and tries to adjust the estimation weights w_{ik} by estimating L_i^B so as to obtain a better performance in the estimation of Y^B .

3. ADJUSTMENTS FOR BIASED ESTIMATION

As indicated in Section 1, biases in the estimation using GWSM occur due to link nonresponse problems. In this situation, not all of the composition in L_i^B can be identified or observed. Although the links between the units in s^A and units in U^B can be normally determined in practice, the part of links outside s^A is often difficult or even impossible to identify. We say that such units are having missing links with U^B . Let $\Delta^A = \Omega^A \setminus s^A$ be the set of units with possible missing links. Then,

$$L_i^B = \sum_{j \in s^A} \sum_{k=1}^{M_i^B} l_{j,ik} + \sum_{j \in \Delta^A} \sum_{k=1}^{M_i^B} l_{j,ik}. \quad (7)$$

If we carry out the GWSM without taking these missing links into account, this corresponds to using the total of observed links L_i^{B*} instead of L_i^B , where

$$L_i^{B*} = \sum_{j \in s^A} \sum_{k=1}^{M_i^B} l_{j,ik} + \sum_{j \in \Delta_0^A} \sum_{k=1}^{M_i^B} l_{j,ik}, \quad (8)$$

with Δ_0^A being a subset of Δ^A , and containing only the units whose links are observed. The cost is overestimation of Y_B in using (5) since

$$L_i^B \geq L_i^{B*}$$

We suggest some methods for adjusting the GWSM for link nonresponse by estimating the total number of links, L_i^B .

3.1 Estimating L_i^B by proportional adjustment for each individual cluster (Method 1)

To address the link nonresponse problem, we focus on estimating L_i^B using the known information about the links within s^A . To compute the weights in (6) using GWSM, we only need to estimate L_i^B for those $i \in \Omega^B$. For any $i \in \Omega^B$,

$$L_i^B = \sum_{j=1}^{T_i^A} L_{j,i}^B. \quad (9)$$

A general estimator for this total can be expressed as $\hat{L}_i^B = \sum_{j=1}^{T_i^A} w_{j,i}^L L_{j,i}^B$, where $w_{j,i}^L$ is a random weight that takes the value $w_{j,i}^L = 0$ if j is not in the sample s_i^A . For each $i \in \Omega^B$, we use the known link information between s_i^A and U_i^B to estimate the link information between Ω_i^A and U_i^B . The expectation of \hat{L}_i^B is

$$E(\hat{L}_i^B) = \sum_{j=1}^{T_i^A} E(w_{j,i}^L) L_{j,i}^B. \quad (10)$$

By comparing (9) and (10), it can be observed that \hat{L}_i^B is unbiased for L_i^B for any weighting scheme with $E(w_{j,i}^L) = 1$ for all j .

As a first approach, we adopt the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), also called π estimator (Särndal, Swensson, and Wretman, 1992). Note that, by the definition of $\Omega_i^A, \Omega_i^A \supset s_i^A$ for all i . We imitate a procedure for estimating the number of links in Ω_i^A using that in s_i^A . We then see s_i^A as a ‘‘sample’’ selected from the ‘‘population’’ Ω_i^A . Let $\pi_{j,i}^L$ be the ‘‘probability’’ of j (which is in Ω_i^A) being included in s_i^A . Then, let

$$w_{j,i}^L = \begin{cases} 1/\pi_{j,i}^L, & j \text{ is in } s_i^A, \\ 0, & j \text{ is in } \Omega_i^A/s_i^A. \end{cases} \quad (11)$$

According to Corollary 3.1 in Cassel, Särndal, and Wretman (1977), this weighting scheme provides an unbiased estimator for L_i^B . We have

$$\hat{L}_i^B = \sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_j^L}{\pi_{j,i}^L}. \quad (12)$$

It leads us with an asymptotically unbiased (proof follows) estimator of Y^B :

$$\tilde{Y}^B = \sum_{i=1}^n \frac{\sum_{j=1}^{M_i^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{\sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_{j,i}^L}{\pi_{j,i}^L}} \sum_{k=1}^{M_i^B} y_{ik}. \quad (13)$$

In order to show its unbiasedness, we use the technique of Taylor linearization. We first obtain

$$\frac{1}{\hat{L}_i^B} \approx \frac{2L_i^B - \hat{L}_i^B}{(L_i^B)^2}.$$

Then,

$$\begin{aligned} E(\tilde{Y}^B) &\approx \sum_{i=1}^n E_{t_j} \left[E_{t_{j,i}^L} \left(\frac{1}{(L_i^B)^2} \left(2L_i^B - \sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_{j,i}^L}{\pi_{j,i}^L} \right) \sum_{j=1}^{M_i^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \middle| \Omega^B \right] \sum_{k=1}^{M_i^B} y_{ik} \\ &= \sum_{i=1}^n E_{t_j} \left(\frac{1}{L_i^B} \sum_{j=1}^{M_i^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \sum_{k=1}^{M_i^B} y_{ik} \\ &= E_{t_j}(\hat{Y}^B). \end{aligned} \quad (14)$$

According to Lavallée (1995), $E_{t_j}(\hat{Y}^B) = Y^B$, and so \hat{Y}^B is an approximately unbiased estimator of Y^B .

Now, we need to compute $\pi_{j,i}^L$. It is a function of π_j^A yet it depends on how s_i^A affects on U_i^B , and therefore on Ω_i^A . Such an effect is difficult to track and varies from case to case. However, we may give a general estimate of it. We propose that the selection probability $\pi_{j,i}^L$ be roughly estimated by the proportion which s_i^A takes within Ω_i^A , namely

$$\hat{\pi}_{j,i}^{L(1)} = \frac{m_i^A}{T_i^A}. \quad (15)$$

Therefore,

$$\hat{L}_i^{B(1)} = \sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_j^L}{\hat{\pi}_{j,i}^{L(1)}} = \frac{T_i^A}{m_i^A} \sum_{j=1}^{m_i^A} L_{j,i}^B. \quad (16)$$

and

$$\begin{aligned}
\hat{Y}^{B(1)} &= \sum_{i=1}^n \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{\frac{T_i^A}{m_i^A} \sum_{j=1}^{M_i^A} L_{j,i}^B} \sum_{k=1}^{M_i^B} y_{ik} \\
&= \sum_{i=1}^n \frac{m_i^A}{T_i^A} \frac{\sum_{j=1}^{M^A} \frac{L_{j,i}^B}{\pi_j^A}}{\sum_{j=1}^{M_i^A} L_{j,i}^B} \sum_{k=1}^{M_i^B} y_{ik}.
\end{aligned} \tag{17}$$

3.2 Estimating L_i^B by overall proportional adjustment (Method 2)

In the previous approach, the information regarding m_i^A and T_i^A are needed for every i . Suppose that we ignore the variation of Ω_i^A among all i . Then, we simply propose to use

$$\tilde{L}_i^B = \sum_{j=1}^{T^A} \frac{L_{j,i}^B t_j^L}{\pi_j^L}, \tag{18}$$

that uses link information in s^A to estimate the link information in T^A , where t_j^L is the indicator variable for being in s^A from Ω^A . Now, we need to compute π_j^L . Again, it is a function of π_j^A and yet it depends on the complexity of the impact of s^A on Ω^B , hence on Ω^A . So the computation is difficult and varies from case to case without a general form. But generally, we may give a rough estimate of it. One way to do so is to estimate π_j^L using the proportion of units in s^A

which take in Ω^A , i.e., $\tilde{\pi}_j^L = \frac{m^A}{T^A}$. This leads to

$$\hat{L}_i^{B(2)} = \frac{T^A}{m^A} \sum_{j=1}^{m^A} L_{j,i}^B \tag{19}$$

For simple random designs with or without stratification, $\hat{L}_i^{B(2)}$ provides an unbiased estimator for L_i^B . For more complex designs, it provides a model-based unbiased estimator under the assumption that, for any cluster i , the average of total existing links associated with all units in the sample s^A is the same as that of existing links associated with all units in U^A , i.e.,

$$\frac{\sum_{j=1}^{m^A} L_{j,i}^B}{m^A} = \frac{\sum_{j=1}^{M^A} L_{j,i}^B}{T^A}. \tag{20}$$

The estimation weights are provided by

$$w_{ik}^{(2)} = \frac{m^A}{T^A} \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{\sum_{j=1}^{m^A} L_{j,i}^B t_j}, \text{ for all units } k \text{ in cluster } i. \tag{21}$$

It follows that Y^B can be estimated by

$$\hat{Y}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^n \frac{\sum_{j=1}^{M^A} \frac{L_{j,i}^B}{\pi_j^A}}{\sum_{j=1}^{m^A} L_{j,i}^B} \sum_{k=1}^{M^B} y_{ik} . \quad (22)$$

4. SIMULATION STUDY

For longitudinal surveys, the production of cross-sectional estimates at a particular survey wave after the initial wave is a practical example of an Indirect Sampling problem. Since the population changes over time, the target population is not exactly the same as the initial population from which the longitudinal sample was selected.

In this section, we will use the Survey of Labour and Income Dynamics (SLID) as an example to demonstrate the performance of the estimators that we introduced in Section 3. SLID is an annual longitudinal survey of individuals that belong to households (clusters). The individuals are selected within panels that are selected every three years, and kept for six years, so that there are always two concurrent panels. The sample design for SLID is detailed in Lévesque and Franklin (2000). The terminology that we use in this paper follows Lavallée (1995): a longitudinal person is an individual that was selected at the initial wave; a cohabitant is a person living with a longitudinal person at the current wave; an initially-present cohabitant is an individual that was present in the population at the initial wave and is now selected; and an initially-absent cohabitant is an individual that was not present in the population (e.g. newborn or immigrant) at the initial wave but is now selected.

In the present context, U^A is the population at the initial wave of the longitudinal survey, while U^B is the population at any of the following wave after the initial wave. The sample s^A contains the longitudinal individuals. Since the two populations U^A and U^B are mainly the same, but at different time points, the links between U^A and U^B are mainly one-to-one. That is, we have $l_{j,ik} = 1$ if the individual j of U^A (initial wave) corresponds to the individual k of household i of population U^B (current wave), 0 otherwise. As a consequence, $L_{j,i}^B$ is a binary variable: its value is 1 if individual j lives in household i at the current wave, and 0 otherwise. The quantity L_i^B is the total number of longitudinal persons and initially-present cohabitants at the initial wave who live in household i at the current wave.

For a longitudinal individual, the link is one-to-one. For cohabitants, there is a quite large possibility that their links cannot be identified a few years after the initial wave, i.e., it might be difficult to determine whether a cohabitant is initially-present, or initially absent. The larger the proportion of cohabitants takes in the sample, the larger this possibility goes. For instance, in panel 3 of SLID, cohabitants took 7.8% out of 47,377 individuals in the year of 2000, which is one year after the initial year. It went up to 13.87% in the year of 2002, which is three years later, and 15.22% in the year of 2003, four years later. We can see that the problem of links identification might not be negligible for such a big proportion of cohabitants.

Taking advantage of the availability of observed information, we implemented the approach of estimating L_i^B by the two kinds of proportional adjustments that we proposed in Section 3. In order to test the performance of the estimates obtained by these approaches, we carried out a simulation study using SLID data. Cross-sectional estimations for four income variables of interest have been produced for the year of 2003. These four variables are “total income before taxes”, “total income after taxes”, “earnings” (includes wages and salaries before deductions and self-employment income), and “wages and salaries before deductions” (also called employment income). We are interested in the total of the population incomes for each of these variables. These four quantities of interest have been estimated at both the national level and the provincial level.

It is advisable to make maximum use of every constraints that would help to calculate the values of T_i^A or T^A . For a longitudinal survey, the total number of links in a cluster i are less than or equal to the total number of individuals in this cluster, and greater than or equal to the number of longitudinal individuals in this cluster. Since T_i^A can be difficult to measure, assuming that the household composition is relatively stable through time, we can then assume that $T_i^A \approx M_i^B$.

For the simulations, we first assumed that the links between all units selected in the initial wave (1999) and the units in the target population in 2003 were correctly specified. Then, we computed the different totals using the GWSM. We used them as our target totals, i.e., the "true" values.

Second, we randomly throw away some links associated with the initially-present cohabitants. This had the effect of increasing the number of initially-absent individuals, which is the same as saying that we created some nonresponse in the identification of the links. For the purpose of the simulations, we removed 50% of the links. Without any adjustment, we recalculated the estimates using the GWSM. We used them as our estimation benchmarks, i.e., the "placebo" values.

Third, we estimated the same quantities using GWSM with the suggested proportional adjustments described in Section 3, to see whether the estimates are close enough to the "truth" and how much improvement these adjustments make.

This simulation study using SLID data demonstrates that the proposed methods perform very well in overcoming the overestimation problems that arises from link nonresponse.

Denote

$$w_i^{mean} = \frac{\sum_{j=1}^{m^A} L_{j,i}^B \frac{1}{\pi_j^A}}{\sum_{j=1}^{m^A} L_{j,i}^B}. \quad (23)$$

Then, using Method 1 and Method 2 of Section 3, we estimate Y^B by

$$\hat{Y}_{mean}^{B(1)} = \sum_{i=1}^n \frac{m_i^A}{T_i^A} w_i^{mean} \sum_{k=1}^{M_i^B} y_{ik}, \quad (24)$$

and

$$\hat{Y}_{mean}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^n w_i^{mean} \sum_{k=1}^{M_i^B} y_{ik}, \quad (25)$$

respectively.

Because T_i^A is intractable in practice, we simply approximated T_i^A by M_i^B , as suggested earlier. Similarly, because of the same problem, we approximated T^A by M^A . Even if this is not in favour of our simulation results, these alternatives turned out to produce satisfactory results in the context of our longitudinal population. Note that w_i^{mean} is the average weight of longitudinal persons who live in i th household at the current year. Because the mean is a central measure, it is also reasonable to use the median weight:

$$w_i^{median} = \text{the median of } \frac{1}{\pi_j^A}, j = 1, 2, \dots, m^A, \quad (26)$$

to enhance the robustness of the estimates. We then estimate Y^B using

$$\hat{Y}_{median}^{B(1)} = \sum_{i=1}^n \frac{m_i^A}{T_i^A} w_i^{median} \sum_{k=1}^{M_i^B} y_{ik}, \quad (27)$$

and

$$\hat{Y}_{median}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^n w_i^{median} \sum_{k=1}^{M_i^B} y_{ik}. \quad (28)$$

The comparison for the proposed estimates (24), (25), (27) and (28) are presented in Figures 1a and 1b below, for "total income after taxes" and for "earnings". These figures show that our estimates using both Method 1 and Method 2 perform very well in terms of reducing bias. Method 1 does work better than Method 2 overall, yet such improvement from Method 1 to Method 2 are much less compared to that made by moving from without adjustment to Method 2. Since

Method 2 provides us a high quality and involves much less information than Method 1, Method 2 is recommended. Similar results are obtained for “total income before taxes” and “wages and salaries before deductions”.

Figure 1a

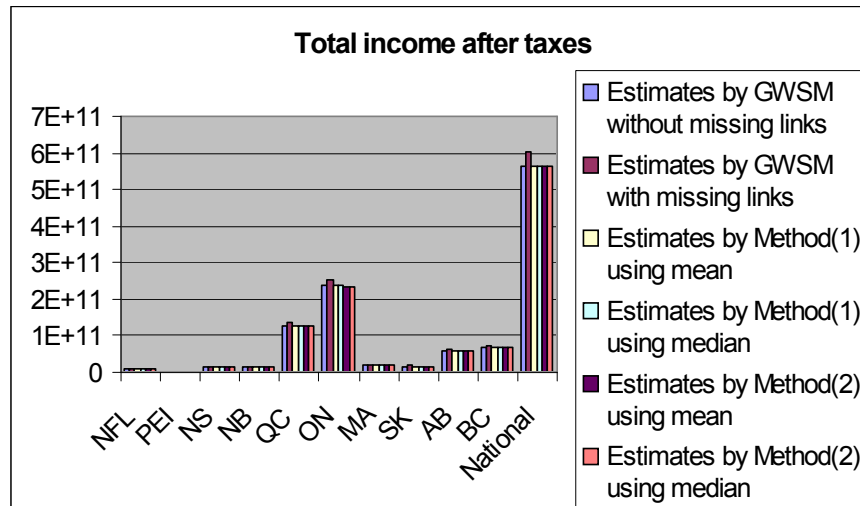
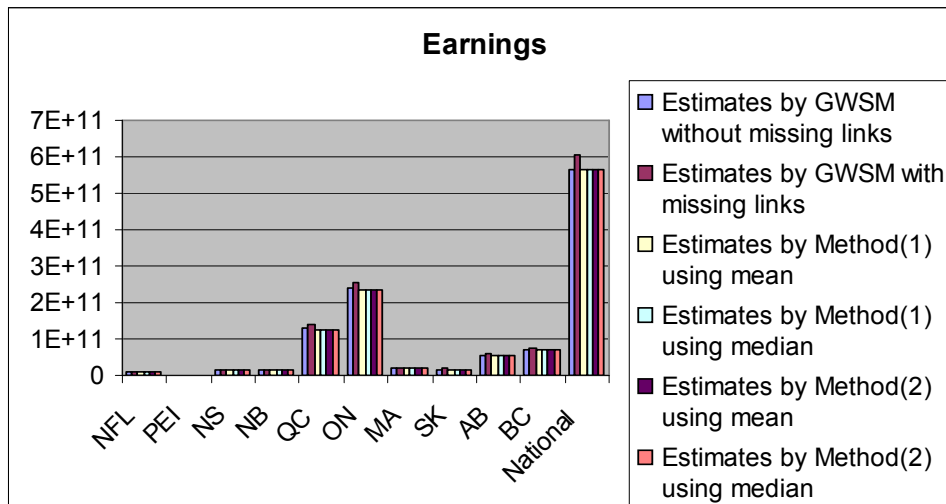


Figure 1b



We now focus on Method 2 using the mean, which gives the estimate $\hat{Y}_{mean}^{B(2)}$, to analyse how its variance performs in terms of estimating Y^B . We used the bootstrap technique to estimate the variance of $\hat{Y}_{mean}^{B(2)}$ at both the national level and the provincial level. The improvement of reducing the variance is not as large as reducing bias. However, it is found in this simulation study that the proposed method provides a smaller variance as well compared to applying GWSM without an adjustment for missing links. See Figure 2a and 2b below.

Figure 2a

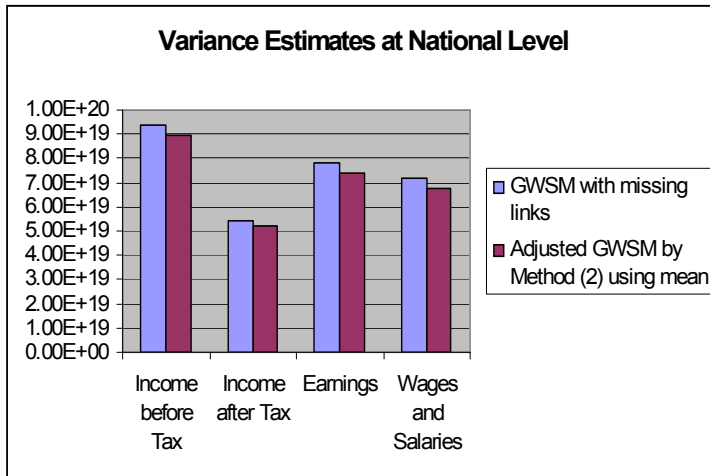
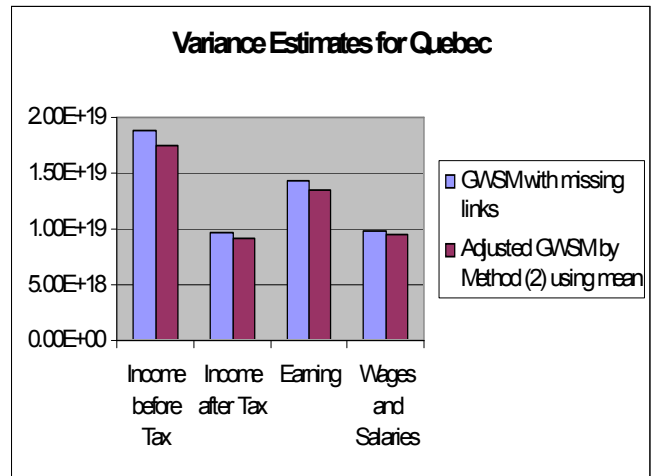


Figure 2b



5. CLOSING REMARK

We have constructed adjustment methods that are simple to use to address the problem of link nonresponse in Indirect Sampling. The simulation results show that the adjustment methods incorporating the link nonresponse performs well in terms of both reducing the estimation bias and providing an overall improvement to the variance. The advancement in bias reduction seems significant.

ACKNOWLEDGEMENTS

The research of Xiaojian Xu is funded by Mathematics of Information Technology and Complex Systems (MITACS) and the Province of Alberta Graduate Fellowship. She also thanks Dr. Douglas Wiens at the University of Alberta and Dr. Milorad Kovacevic at Statistic Canada for their support and encouragement throughout this study.

REFERENCES

- Cassel, C. M., Särndal, C. E., and Wretman, J. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley.
- Deville, J.-C., and Särndal, C. E. (1992). "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association*, **87**, 376-382.
- Draper, N. R., and Smith, H. (1998). *Applied Regression Analysis*, 3rd ed. New York: Wiley.
- Horvitz, D. G., and Thompson, D. J. (1952). "A Generalization of Sampling without Replacement from a Finite Universe". *Journal of the American Statistical Association*, **47**, 663-685.
- Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households using Weight Share Method". *Survey Methodology*, **21**, 25-32.
- Lavallée, P. (2001). "Correcting for Non-response in Indirect Sampling". *Proceedings of Statistics Canada's Symposium 2001*.
- Lévesque, I., Franklin, S. (2000). *Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics, 1997 Reference Year*. Publication of Statistics Canada, Catalogue 75F0002MIE-00004, Ottawa.
- Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.