CROSS-SECTIONAL INFERENCE BASED ON LONGITUDINAL SURVEYS: SOME EXPERIENCES WITH STATISTICS CANADA SURVEYS

Georgia Roberts, Milorad Kovacevic, Harold Mantel, Owen Phillips[1]
Statistics Canada

Abstract

This paper focuses on cross-sectional inference based on data from a longitudinal survey which carries some additional components to achieve cross-sectional representativity. When inferring about the differences in the cross-sectional populations at two different points in time, problems arise with variance estimation for the difference of the respective estimates, when the estimates are derived from such a survey. There are several factors contributing to these problems. Of these, the most important is the sample overlap at the two time points due to the underlying longitudinal survey design; this introduces a strong covariance component which must be included in the estimate of the variance of the difference. Also associated with the underlying longitudinal sample is the complexity introduced by longitudinally sampled individuals moving from one geographical part of the country to another, and thus being used to represent a different part of the cross-sectional population than that for which they were selected. The degree of complication that such factors introduce to the variance estimation problem is determined by the manner in which the longitudinal sample has been supplemented and adjusted in order to attain cross-sectional samples and by the available design information that may be used for cross-sectional inference.

The variance estimation problem is addressed for Canada's Survey of Labour and Income Dynamics (SLID) within a Taylor linearization approach as well as within the resampling framework with emphasis on the bootstrap method. For cross-sectional purposes, SLID combines two independent panels of longitudinal individuals sampled three years apart and also includes all members of the families and households with whom the originally selected longitudinal individuals live at a certain point in time. A numerical illustration based on SLID is included.

Key words: bootstrap, combining panels, Taylor linearization, variance estimation

1. Introduction

The objective of most cross-sectional surveys is to produce unbiased (or nearly unbiased) estimates of levels such as totals or means at a given time point, and, in the case of repeated surveys, to produce estimates of the net change that occurred in the population between two time points. These estimates are often accompanied by estimated measures of precision. The primary objective of longitudinal surveys is the production of longitudinal data series that are appropriate for studying the gross change in a population between collection dates, and for research on causal relationships among variables.

In order to improve the cost-effectiveness of surveys, statistical agencies very often derive cross-sectional estimates from longitudinal survey data assuming that the survey design takes this possibility into account, and that estimation procedures are developed to satisfy cross-sectional as well as longitudinal requirements. A good example of such 'double' utilization of a longitudinal survey is the Canadian Survey of Labour and Income Dynamics (SLID). It was originally designed

[1] Georgia Roberts, Milorad Kovacevic, Harold Mantel, Owen Phillips, Data Analysis Resource Center, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; e-mail: georgia.roberts@statcan.ca, milorad.kovacevic@statcan.ca, harold.mantel@statcan.ca, owen.phillips@statcan.ca.

to provide longitudinal estimates and analyses. However, recognizing the cross-sectional capabilities of SLID, Statistics Canada made it a principal survey for providing annual income data and used it to replace a classic cross-sectional Survey of Consumer Finances (SCF) as of 1998.

In order to achieve cross-sectional representativity, different approaches have been taken in different longitudinal surveys. SLID employs overlapping panels, each of six years duration and selected three years apart. The cross-sectional sample for a particular year also includes cohabitants of the longitudinal individuals from the two panels, i.e., all individuals that are living with the originally selected longitudinal individuals at a certain point in time. In this way, only households composed entirely of immigrants who have arrived since the last panel selection (at most three years out of date) are not represented in the sample. The elaborate cross-sectional weighting scheme that includes a non-response adjustment, an optimal combination of the two panels, adjustments for interprovincial migration and influential values, and post-stratification to a number of post-stratum totals completes the adjustments towards cross-sectional representativity of the population at a given time (Levesque and Franklin, 2000).

Point estimation of parameters of the cross-sectional population based on data from longitudinal surveys in general, and from SLID in particular, has been studied and documented (Lavallee 1995, Merkouris 1999, Levesque and Franklin 2000). However, variance estimation for these estimates hasn't received as much attention. In particular, the problem of formal comparison of the estimates from two years, which requires variance estimation for the difference of the estimates, is seldom addressed. This paper focuses on that problem. It is an extension of previous work by Roberts and Kovacevic (1999) on the comparison of cross-sectional prevalence rates estimated from the Canadian National Population Health Survey.

The paper is organized into five sections. Section 2 contains a description of the problem and details some of its causes. Two approaches to variance estimation as a practical solution to the problem are given in Sections 3 and 4. Section 5 contains a numerical illustration and some concluding remarks.

2. Problem Description

Statistics Canada conducted the Survey of Consumer Finances (SCF) annually beginning in 1971 to provide income data for families and individuals. Its output consisted of estimates of a variety of income distribution parameters at the national and provincial levels for a number of different subpopulations. Due to the near independence of the samples in consecutive years, inference about net change from year to year was straightforward and computable from the reported annual estimates of levels and their standard errors. Since the survey contents of the SCF and SLID are almost identical, Statistics Canada decided to replace the SCF by SLID starting in 1998. The main reason was a gain in efficiency. Also the extensive demographic, socio-economic and labour content of SLID would allow different perspectives on income distributions through a better fitting of a variety of models.

The longitudinal underpinnings of SLID introduce complexities that cause difficulties when it comes to estimation of the variance of the difference of estimates in any two years (that are not more than 6 years apart). Some of these complexities are the following:

i) The cross-sectional SLID sample in any year contains all longitudinal individuals and their cohabitants who are in-scope for cross-sectional purposes. Thus, the cross-sectional samples are not independent at the two time points and have a large degree of overlap. Longitudinal

individuals are in-scope cross-sectionally for a given year if they are still members of Canada's ten provinces as of December 31 of the reference year, excluding those who live in institutions, military barracks or on Indian reserves .

ii) The cohabitants of the originally selected longitudinal individuals generally stay with these individuals for more than a year and thus increase the overlap of the samples.

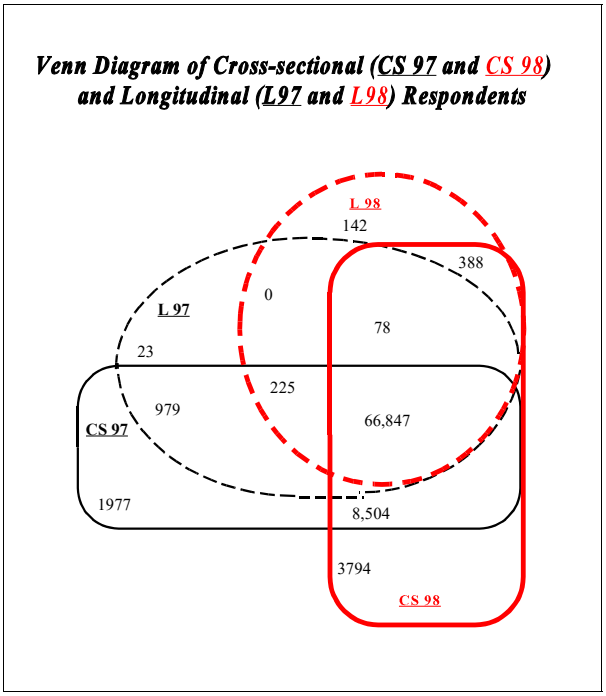
iii) At each of the time points after 1995, the cross-sectional samples contain two longitudinal panels that were selected three years apart. Each panel represents the entire survey population at the time of its selection. These overlapping panels are optimally combined to represent the cross-sectional population in a reference year. The optimality criterion was based on minimizing the variance of an estimated total and resulted in 'panel allocation factors' calculated at the level of province for each reference year. These factors were then applied to individual weights. The overlapping panels can be thought of as a special case of a dual frame survey (Merkouris, 1999).

iv) For cross-sectional purposes the longitudinal individuals who changed province of residence after being selected into the longitudinal sample are considered to be part of the sample for the province in which they reside at the time of the cross-section. However, for variance estimation these individuals must be considered to be part of their original province, stratum and cluster.

v) The cross-sectional samples are affected by longitudinal non-response because of the way in which longitudinal individuals are included in the cross-sectional samples.

Many of these complexities are accounted for through SLID's elaborate weighting scheme so that point estimation of cross-sectional parameters and their net change over time is consistent. Bootstrap weights specifically created for each cross-sectional sample also account for most of these complexities. However, estimation of the variance for the difference of two estimates obtained in different years is not straightforward due to the sample overlap.

For purposes of illustration, suppose we are interested in estimating the variance of the difference of estimates for reference years, 1997 and 1998. Here we illustrate some difficulties in determination of SLID individuals who are in the cross-sectional samples in these two years through graphical presentation of the composition of the cross-sectional and longitudinal SLID samples for 1997 and 1998. There are 78,532 individuals with positive 1997 cross-sectional weights, and 79,611 with positive 1998 cross-sectional weights. An individual with a positive cross-sectional weight in a particular year is cross-sectionally in scope and belongs to a responding household in that year. The total number of cross-sectional individuals common to both years is 75,351, of which 66,847 are longitudinal individuals. The remaining 8,504 common individuals are the cohabitants who were with longitudinal individuals in both years. The common individuals represent 96% of the cross-sectional sample in 1997 and 95% in 1998. There were 101 (23+78) longitudinal individuals for 1997 that were not in scope cross-sectionally in 1997; 78 of them remained longitudinal in 1998 and also became cross-sectionally valid in 1998, while the remaining 23 individuals are probably longitudinal non-respondents in 1998. Another 225 individuals, who were in the longitudinal samples in both years and in the 1997 cross-sectional sample, were lost for cross-sectional estimation in 1998, most likely by moving out of scope (due to moving into institutions or out of the ten provinces, or dying). It is also interesting to observe that 530 (142+388) individuals had positive longitudinal weights in 1998 but had zero longitudinal and cross-sectional weights in 1997, most likely due to wave nonresponse in 1997. Only 388 of these were cross-sectionally in scope in 1998. Most of cross-sectional individuals in SLID stayed in the province where they were originally selected: in 1997 only 3.2% lived in a different province and in 1998 only 3.9%.



In order to address the problem of variance estimation for the difference of cross-sectional estimates obtained in 1997 and 1998, we now introduce some notation. Let s_t be the individuals on the cross-sectional SLID sample at time t, where t=1 for 1997 and t=2 for 1998. Suppose that we are interested in the mean after-tax income within a domain of the population at each time point. The mean income within the domain at time t may be estimated by $\hat{\theta}_t = \hat{Y}_t / \hat{X}_t$, with $\hat{Y}_t = \Sigma_{s_t} w_{ti} y_{ti}$ and $\hat{X}_t = \Sigma_{s_t} w_{ti} x_{ti}$, where w_{ti} is the cross-sectional weight of the ith individual in s_t (who will be called called the itth individual); $y_{ti} = income$ if the itth individual is in the domain, and $y_{ti} = 0$ otherwise; and $x_{ti} = 1$ if the itth individual is in the domain, and $x_{ti} = 0$ otherwise. Then $\hat{\Delta} = \hat{\theta}_1 - \hat{\theta}_2$ estimates the net change in the mean income between the two time periods. The main problem addressed in this paper is the estimation of the variance of $\hat{\Delta}$. In the next two sections we present two possible methods, Taylor linearization and a pseudo-coordinated bootstrap method.

3. Variance Estimation: Taylor Linearization Approach

3.1 Linearization of $\hat{\Delta}$

One possible approach to obtaining a design-based variance estimate of $\hat{\Delta}$ is Taylor linearization. In developing this approach, for ease of presentation, adjustments to the final weights will be ignored. Since $\hat{\Delta}$ is a non-linear function of the data from both samples s_t, t=1,2, the first step is to linearize $\hat{\Delta}$ by expansion into a Taylor series around the true net change in means. Assuming that the remainder term is negligible for a sufficiently large sample, the following approximation holds:

$$\hat{\Delta} \approx \Delta + \frac{1}{X_1}\left[(\hat{Y}_1 - Y_1) - \theta_1 (\hat{X}_1 - X_1)\right] - \frac{1}{X_2}\left[(\hat{Y}_2 - Y_2) - \theta_2 (\hat{X}_2 - X_2)\right], \quad (1)$$

where $\theta_t = Y_t / X_t$, t=1,2. This implies that

$$Var(\hat{\Delta}) \approx Var\left[\frac{1}{X_1}\sum_{s_1} w_{1i}(y_{1i} - \theta_1 x_{1i}) - \frac{1}{X_2}\sum_{s_2} w_{2i}(y_{2i} - \theta_2 x_{2i})\right]. \quad (2)$$

Sample s_t can be expressed as $s_t = \bigcup_{k=1}^{10} s_{tk}$ where s_{tk} represents those observations in s_t forming the cross-sectional sample for province k at time t. It then follows that

$$Var(\hat{\Delta}) \approx Var\left[\sum_{k=1}^{10} \hat{Z}_1(s_{1k}) - \sum_{k=1}^{10} \hat{Z}_2(s_{2k})\right], \quad (3)$$

where $\hat{Z}_t(s_{tk}) = \sum_{s_{tk}} w_{ti} Z_{ti}$, and $Z_{ti} = X_t^{-1}(y_{ti} - \theta_t x_{ti})$, t=1,2.

If we ignore, for the moment, longitudinal individuals who are residing in a different province than the one for which they were selected, the provincial samples s_{tk} are independent due to design of SLID, where there was independent sample selection in different provinces . This then implies that

$$Var(\hat{\Delta}) \approx \sum_{k=1}^{10} Var\left[\hat{Z}_1(s_{1k}) - \hat{Z}_2(s_{2k})\right]. \quad (4)$$

The kth provincial component of this variance, $Var\left[\hat{Z}_1(s_{1k}) - \hat{Z}_2(s_{2k})\right]$, can be expanded further as

$$Var\left[\hat{Z}_1(s_{1k}) - \hat{Z}_2(s_{2k})\right] = Var\left[\hat{Z}_1(s_{1k})\right] + Var\left[\hat{Z}_2(s_{2k})\right] - 2\,Cov\left[\hat{Z}_1(s_{1k}), \hat{Z}_2(s_{2k}\right]. \quad (5)$$

The problem of estimating the variance of $\hat{\Delta}$ then reduces to estimating the terms on the right hand side of (5).

3.2 Notation and Assumptions Required for Variance Estimation

The following detailed notation is required for explanation of the variance estimation:
H_{tk} = # of strata in the cross-sectional sample in province k at time t,
n_{tkh} = # of sampled clusters in the hth stratum in province k at time t,
n_{tkhc} = # of sampled individuals in cth cluster of hth stratum in province k at time t,
w_{tkhci} = weight on the ith individual in cth cluster of hth stratum in province k at time t, and
$z_{tkhci} = \hat{X}_t^{-1}(y_{tkhci} - \hat{\theta}_t x_{tkhci})$.
It should be noted that the strata and weights are those in use after the combining of the two panels. See Levesque and Franklin (2000) and Merkouris (1999) for details.

The following standard assumptions for variance estimation for data from a survey with a stratified multistage design are considered to hold for each of the cross-sectional SLID samples:

i) The design of each cross-sectional sample is approximately stratified with selection of psu's with replacement.
ii) Each psu is selected at most once (because of small sampling fractions).
iii) $n_{tkh}\sum_{i=1}^{n_{tkhc}} w_{tkhci} z_{tkhci} = n_{tkh} z_{tkhc}$ (*i.e.*, $n_{tkh}\times$ weighted cluster total) is approximately unbiased as an estimator for the stratum total Z_{tkh} for any z variable and for any value of t, k, h, and c.

Under these assumptions, there is a straightforward approach to estimate a stratum total and the variance of stratum total at each time point. As well, if the same psu's are represented in a stratum at both time points, there is a straightforward approach to estimating a covariance between stratum totals at the two time points. In particular, under these assumptions:

i) An (approximately) unbiased estimate for Z_{tkh} is $\hat{Z}_{tkh} = \sum_{c=1}^{n_{tkh}} z_{tkhc}$.
ii) An (approximately) unbiased estimate of the variance of $\hat{Z}_{tkh}$ is

$$\hat{var}[\hat{Z}_{tkh}] = n_{tkh}/(n_{tkh} - 1)\sum_{c=1}^{n_{tkh}} (z_{tkhc} - \bar{Z}_{tkh})^2,$$

where $\bar{Z}_{tkh} = \hat{Z}_{tkh}/n_{tkh}$.
iii) If, at times t=1 and t=2, the same psu's are observed in a stratum sample, (which implies that

$n_{1kh} = n_{2kh}$), an (approximately) unbiased estimate of the covariance of $\hat{Z}_{1kh}$ and $\hat{Z}_{2kh}$ is given by

$$\hat{cov}[\hat{Z}_{1kh}, \hat{Z}_{2kh}] = n_{1kh}/(n_{1kh} - 1) \sum_{c=1}^{n_{1kh}} (z_{1khc} - \bar{Z}_{1kh})(z_{2khc} - \bar{Z}_{2kh}).$$

3.3 Application

These results can then be readily applied to the cross-sectional SLID samples for 1997 and 1998. By the design of SLID, cross-sectional samples for those two years should consist of the same strata and psu's within each province at both time points, even though there are several reasons to expect that the individuals within a particular psu would not be exactly the same at the two time points (such as nonresponse of a longitudinal individual to the income questions at one of the time points or a longitudinal person entering an institution between the two time points). The following variance and covariance estimates would follow in a straightforward manner from the results above:

$$\hat{var}\left[\hat{Z}_t(s_{tk})\right] = \sum_{h=1}^{H_{tk}} n_{tkh}/(n_{tkh} - 1) \sum_{c=1}^{n_{tkh}} (z_{tkhc} - \bar{Z}_{tkh})^2, \text{ and}$$

$$\hat{cov}\left[\hat{Z}_1(s_{tk}), \hat{Z}_2(s_{2k})\right] = \sum_{h=1}^{H_{1k}} n_{1kh}/(n_{1kh} - 1) \sum_{c=1}^{n_{1kh}} (z_{1khc} - \bar{Z}_{1kh})(z_{2khc} - \bar{Z}_{2kh}),$$

while z_{1khc} and z_{2khc} would consist of weighted sums over different individuals if the *khc*-th psu contained different individuals at the two time points.

3.4 Accounting for Movers Between Provinces

In the development above, it was assumed that individuals continue to reside in the province for which they were selected into the sample. Modifications need to be made to the Taylor linearization variance approach when there are movers, that is, people who, for either time point, are cross-sectionally representing a different province than the one for which they were drawn into the sample. This can be done by first decomposing s_{tk} into $s_{tk} = s_{t1k} \cup s_{t2k} \cup ... \cup s_{t10k}$ where s_{tjk} are those people in s_{tk} who were selected into the sample in province *j*. Then, $Var(\hat{\Delta})$ can be expanded in the s_{tjk}, and terms be grouped according to the province of selection. Making use of the fact that independent sampling was done by province, formulae similar to those in 3.3 above may be developed readily for calculating the required variances and covariances among the s_{tjk} domains. While theoretically straightforward, implementation could be tedious if many of the s_{tjk}, $j \neq k$ are non-empty.

4. Variance Estimation: Bootstrap Methods

Replication methods for variance estimation are becoming increasingly popular for analysis of data from complex surveys. Methods suitable for data from stratified multistage survey designs are now available, and their properties have been investigated both theoretically and empirically. One attractive feature of these methods is that the relatively difficult task of deriving replicate survey weights only needs to be done once by the methodologists most familiar with the survey design and weighting. In particular, complexities due to multistage sampling, multiple frame estimation, interprovincial migration of longitudinal panel members, adjustments to the weights to account for non-response, *etc.*, can be incorporated into these replicate weights. Use of the replicate weights by any analyst to derive valid design based variance estimates is then relatively simple, and does not require any direct knowledge of the complex survey design or weighting procedures.

In this section we first briefly describe a bootstrap method, called the coordinated bootstrap, which is suitable for overlapping samples on two occasions. We then describe an approximation to the coordinated bootstrap, called the pseudo-coordinated bootstrap, which may be used when coordinated bootstrap weights are unavailable.

4.1 Coordinated Bootstrap Method

In this subsection we describe a coordinated bootstrap method for estimation of the variance of the difference of two cross-sectional estimates. The bootstrap resampling method for iid samples has been extensively studied (see Efron, 1982). It was extended by Rao and Wu (1988) to stratified multistage designs and again by Rao, Wu and Yue (1992) to include nonsmooth statistics. Yung (1997) contains a concise description of the procedure. To summarize, for each bootstrap replicate a sample of PSUs is drawn with replacement from the set of sampled PSUs in each stratum. Sampling weights of each sample unit are then adjusted to reflect this resampling; this is called the bootstrap adjustment to the sampling weights. Any further adjustments to the sampling weights, such as nonresponse adjustments or calibration of the weights, should also be applied to each bootstrap replicate to produce what we will call a set of bootstrap weights. The bootstrap variance estimator for a weighted estimator $\hat{\theta}$ is then calculated as

$$v_B(\hat{\theta}) = \frac{1}{B}\sum_b \left(\hat{\theta}^*_{(b)} - \hat{\theta}^*_{(\cdot)}\right)^2 \tag{6}$$

where $\hat{\theta}^*_{(b)}$ is the estimate of θ based on the bth set of bootstrap weights, and $\hat{\theta}^*_{(.)}$ is the mean of $\hat{\theta}^*_{(b)}$ over the B bootstrap replicates. Alternatively, $\hat{\theta}$ is often substituted for $\hat{\theta}^*_{(.)}$ in (6).

The same method may be used for multistage sampling on two occasions with overlapping samples. The following procedure is used for each bootstrap replicate. For sample PSUs that are common to the two occasions by design, the bootstrap samples for the two occasions must be "coordinated"; *i.e.*, the same bootstrap samples of PSUs should be used for each occasion. For the sample PSUs that are chosen independently on either occasion, bootstrap samples of PSUs should also be chosen independently. Bootstrap adjustments to the sampling weights would be applied as usual, and any further adjustments to the weights would be applied independently in each sample. Now, if $\hat{\theta} = \hat{\Delta}$ is the difference between two cross-sectional estimates, one from each of the samples, then its variance can be estimated consistently from (6) using these coordinated sets of bootstrap weights.

4.2 Pseudo-Coordinated Bootstrap Method

Although the coordinated bootstrap offers a neat solution to the problem of variance estimation for the difference of two cross-sectional estimates, it cannot be applied when the bootstrap samples were drawn independently for each of the two samples, as is often the case for cross-sectional files produced from longitudinal surveys. Recalling that $Var(\hat{\Delta}) = Var(\hat{\theta}_1) + Var(\hat{\theta}_2) - 2Cov(\hat{\theta}_1, \hat{\theta}_2)$, we propose here a method to produce approximate coordinated bootstrap weights which may be used for estimation of the covariance of the two cross-sectional estimates. Because of the approximations and assumptions involved it is recommended that the original bootstrap weights, $w_{ti(b)}$, be used for estimation of the variances of the cross-sectional estimates.

In the coordinated bootstrap approach, for individuals in PSUs that are common to the two samples the bootstrap adjustment of the basic sampling weights would be the same for both samples. Thus

for an individual in the overlap of the two samples, the ratio of the *b*th coordinated bootstrap weight to the final estimation weight should be approximately the same for both samples, with any differences in these ratios due only to differences in the other adjustments to the weights. If we also assume that individuals not in the sample overlap were sampled independently of the overlap, and independently on each occasion, then their contribution to the covariance should be zero. Under these conditions the procedure described below should yield reasonable results. For SLID, cross-sectional individuals who are not in the overlap are not independent of the overlap; however, the number of such individuals is relatively small.

From the *b*th set of bootstrap weights associated with s_1 we define a set of pseudo-coordinated bootstrap (PCB) replicate weights as follows:

$$w_{1,1i(b)} = \begin{cases} w_{1i(b)} & i \in s_1,\ i \in s_2 \\ w_{1i} & i \in s_1,\ i \neg\in s_2 \\ 0 & i \neg\in s_1 \end{cases} \qquad w_{1,2i(b)} = \begin{cases} w_{2i} w_{1i(b)} / w_{1i} & i \in s_1,\ i \in s_2 \\ w_{2i} & i \neg\in s_1,\ i \in s_2 \\ 0 & i \neg\in s_2 \end{cases} \tag{7}$$

We can similarly define PCB weights, $w_{2,1i(b)}$ and $w_{2,2i(b)}$, corresponding to the *b*th set of bootstrap weights associated with s_2. If PSU identifiers were available, then we could replace the PCB adjustment factor $w_{1i(b)}/w_{1i}$ in (7) by $\Sigma_{j \in PSU(i)} w_{1j(b)} / \Sigma_{j \in PSU(i)} w_{1j}$, which would be more stable. If we have *B* replicates in each set of bootstrap weights then it may be reasonable to construct *B*/2 sets of PCB weights based on s_1 bootstraps and *B*/2 based on s_2; however, we may have as many as *B* sets based on each sample. If the original bootstrap weights are benchmarked to some population totals, then we may wish to similarly benchmark the PCB weights, assuming that the benchmarking procedure is known. The covariance of two cross-sectional estimates, $\hat{\theta}_1$ and $\hat{\theta}_2$, would then be estimated by

$$cov_B(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{B_{PC}} \sum_b \left(\hat{\theta}^*_{1(b)} - \hat{\theta}^*_{1(\cdot)} \right) \left(\hat{\theta}^*_{2(b)} - \hat{\theta}^*_{2(\cdot)} \right) \tag{8}$$

where the summation is over the B_{PC} sets of PCB weights, and $\hat{\theta}^*_{1(b)}$ and $\hat{\theta}^*_{1(b)}$ are calculated using, respectively, either $w_{1,1(b)}$ and $w_{1,2(b)}$ from (7), or $w_{2,1(b)}$ and $w_{2,2(b)}$.

4.2.1 Pseudo-Coordinated Bootstrap for Non-independent Non-overlap

If the assumption of independence of the sampling of individuals not in the overlap is not reasonable, then the above procedure would tend to underestimate the magnitude of the covariance. However, the procedure could be modified in various ways.

The first approach to accounting for dependence of the non-overlapping part of the sample is based on identifying PSUs within the samples. For s_2 individuals whose PSU intersects the common sample, PCB weights based on s_1 bootstrap weights could be constructed by multiplying w_{2i} by $\Sigma_{j \in PSU(i)} w_{1j(b)} / \Sigma_{j \in PSU(i)} w_{1j}$. For PSUs that do not intersect the common sample at all, it might be reasonable to assume that such PSUs from s_1 are sampled independently of those in s_2. Alternatively, if such PSUs from s_1 can be linked to corresponding PSUs from s_2, then a similar type of adjustment can be used to construct PCB weights.

For a second, somewhat simpler approach, if θ is a smooth function of population totals, then some

of the extra covariance due to the non-overlapping parts of the samples could be captured using a linear approximation. Suppose for example that $\theta = \theta(X)$ where X is the population total of a variable x. If we write $\hat{X}_k = \hat{X}_k^{\mathrm{o}} + \hat{X}_k^{\mathrm{no}}$, where the superscript "o" denotes the overlap part of the sample, and the superscript "no" denotes the non-overlap part, then we may write an approximation:

$$\mathrm{Cov}(\hat{\theta}_1,\hat{\theta}_2) = \left(\frac{\partial\theta_1}{\partial X_1}\right)\left(\frac{\partial\theta_2}{\partial X_2}\right)\Big(\mathrm{Cov}(\hat{X}_1^{\mathrm{o}},\hat{X}_2^{\mathrm{o}}) + \mathrm{Cov}(\hat{X}_1^{\mathrm{no}},\hat{X}_2^{\mathrm{o}}) + \mathrm{Cov}(\hat{X}_1^{\mathrm{o}},\hat{X}_2^{\mathrm{no}}) + \mathrm{Cov}(\hat{X}_1^{\mathrm{no}},\hat{X}_2^{\mathrm{no}})\Big).$$

If we now define PCB weights based on s_1 bootstraps as

$$w_{1,1i(b)} = \begin{cases} w_{1i(b)} & i\in s_1 \\ 0 & i\neg\in s_1 \end{cases} \qquad w_{1,2i(b)} = \begin{cases} w_{2i}w_{1i(b)}/w_{1i} & i\in s_1,\ i\in s_2 \\ w_{2i} & i\neg\in s_1,\ i\in s_2 \\ 0 & i\neg\in s_2 \end{cases}$$

then these weights could be used to estimate $\mathrm{Cov}(\hat{X}_1^{\mathrm{o}},\hat{X}_2^{\mathrm{o}})$ and $\mathrm{Cov}(\hat{X}_1^{\mathrm{no}},\hat{X}_2^{\mathrm{o}})$. Similarly defined PCB weights based on s_2 bootstraps could be used to estimate $\mathrm{Cov}(\hat{X}_1^{\mathrm{o}},\hat{X}_2^{\mathrm{o}})$ and $\mathrm{Cov}(\hat{X}_1^{\mathrm{o}},\hat{X}_2^{\mathrm{no}})$. However, estimation of the component $\mathrm{Cov}(\hat{X}_1^{\mathrm{no}},\hat{X}_2^{\mathrm{no}})$ requires PCB weights that simultaneously adjust the weights for both of the non-overlapping parts of the samples.

5. Illustration

The proposed methods are applied to SLID data where the average after-tax incomes for individuals aged 16 and over with income for 1997 and 1998 are compared. There were 60,901 and 62,272 such individuals in the 1997 and 1998 cross-sectional samples, respectively. The averages, their difference and the corresponding standard errors obtained by the proposed methods are given in the Table below.

Estimated averages of income-after-tax, their difference and standard errors

	Estimates	Standard Errors		
		Taylor	Bootstrap	
			Coordinated	Pseudo-Coordinated
$\hat{\theta}_{97}$	20285	137	132	
$\hat{\theta}_{98}$	21125	142	137	
$\hat{\theta}_{97}-\hat{\theta}_{98}$	-840	79	82	81

For application of the Taylor method all longitudinal individuals, and so their cohabitants, were associated with their province of residence at their time of selection. Also, it was assumed that the weights of all individuals from a stratum were multiplied by the same panel allocation factor (PAF). In such a case the stratum total can be estimated unbiasedly and the basic assumptions for variance estimation by Taylor linearization method as stated in Section 3 are satisfied.

This, however, may not be exactly true since the weights of individuals that joined the population after the selection of the first panel are not modified by the PAF, meaning that within a Panel 2 stratum some individual weights may be multiplied and some may not. However, the number of such individuals represents less then 0.6% of the Panel 2 size.

The bootstrap calculations are based on 500 replicates. The bootstrap weights that were produced for SLID for the 1997 and 1998 cross-sectional samples are already coordinated. The PCB weights

for this empircal comparison are defined as in (7), using an individual level PCB adjustment factor, with no subsequent benchmarking, and based on the assumption of independence of individuals not in the sample overlap. The first 250 sets PCB weights were based on the first 250 sets of bootstrap weights for s_1, while the second 250 were based on the second 250 sets of bootstrap weights for s_2.

The estimate of $\mathrm{Cov}(\hat{\theta}_{97}, \hat{\theta}_{98})$ based on the Taylor linearization method was 16346, while that based on coordinated bootstrap was 14806, and that based on the pseudo-coordinated bootstrap was 14838.

The preferred method for variance estimation in this set-up is the coordinated bootstrap, as it can take explicit account of all of the complexities of the survey design and estimation. The pseudo-coordinated bootstrap performed well in our example. Some additional empirical investigation is needed to assess its properties. Standard errors estimated by Taylor method are very close to those obtained by the bootstrap methods despite the approximations involved, including the ingnoring of weight adjustments.

Acknowledgment The authors would like to acknowledge the computational input by Michael Lo.

REFERENCES:

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

Lavallee, P. (1995). Cross-Sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method. *Survey Methodology,* Vol.21, 1, 25-32.

Levesque, I. and Franklin, S. (2000). Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics 1997 Reference Year. *Income Statistics Division Paper No.75F0002MIE-00004*, Statistics Canada

Merkouris, P. (1999). Cross-Sectional Estimation in Multiple-Panel Household Surveys. *Methodology Branch Working Paper HSMD 99-004E*, Statistics Canada

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling Inference with complex Survey Data. *Journal of American Statistical Association*, 83, 231-241.

Rao, J.N.K. , Wu, C.F.J.. and Yue (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217.

Roberts, G. and Kovacevic, M. (1999). Comparison of Cross-Sectional Estimates from Two Waves of a Longitudinal Survey. *Proceedings of the Survey Methods Section, Statistical Society of Canada,* 153-158

Stukel, D.M., Mohl, C.A., and Tambay, J.-L. (1997). Weighting for Cycle Two of Statistics Canada's National Population Health Survey. *Proceedings of the Survey Methods Section, Statistical Society of Canada,* 111-116.

Yung, W. (1997). Variance Estimation for Public Use Microdata Files. *Proceedings of Symposium 97 "New Directions in Surveys and Censuses", Statistics Canada*, 91-95.