EFFECT OF NONRESPONSE ADJUSTMENTS ON VARIANCE ESTIMATES FOR THE NATIONAL POPULATION HEALTH SURVEY

Harold J. Mantel, Sylvain Nadon, and Douglas Yeo, Statistics Canada
Harold J. Mantel, 15F R.H. Coats Bldg., Statistics Canada, Ottawa, ON K1A 0T6, CANADA

Key Words: Bootstrap; Nonresponse; Longitudinal surveys; Variance estimation

1. Introduction

Statistics Canada's National Population Health Survey (NPHS) is a longitudinal household survey conducted by Statistics Canada, instituted in 1994 to survey the health of Canadians along with its determinants. Wave 1 of the survey provided a panel of approximately 17,000 respondents to be contacted every two years for up to twenty years. Data are collected during four collection periods in each survey year. Panel respondents were chosen by randomly selecting one person per surveyed household. The sample design of the NPHS is mostly based upon the Labour Force Survey (LFS) sampling methodology. The LFS design generally selects a stratified two-stage sample of dwellings selected within clusters (except in some rural, remote, and apartment strata) with six clusters selected in each stratum. NPHS strata were created by grouping LFS strata, keeping some or all of the LFS selected clusters, but selecting fresh sample dwellings within those clusters. In Quebec the NPHS sample is taken from households that participated in the *Enquête Sociale et de Santé* (ESS), a health survey conducted by Santé Québec in 1992-93. Its design was similar to that of the LFS. For more details on the NPHS design, see Tambay and Catlin (1995).

Estimation weights for NPHS are derived from the basic inverse probability of inclusion sampling weights by adjusting for survey nonresponse and calibrating to known population totals for age-sex categories by province. In a very small number of cases extreme weights are adjusted in an adhoc manner.

In order to provide estimates of sampling variances for statistics and analyses, the NPHS uses a version of the bootstrap method described in Rao, Wu and Yue (1992) and Yung (1997). Variance estimation for complex surveys such as the NPHS typically use a resampling method such as the bootstrap or jackknife, or methods based on Taylor expansions. Resampling methods have some advantages over the Taylor method in that the Taylor method requires deriviation of an appropriate Taylor approximation for each new type of statistic or analysis. On the other hand, resampling methods typically require much more computational power; however, recent advances in computer power have made resampling methods feasible even for complicated, iterative estimation procedures. Another advantage of resampling methods is that the variance estimation can conveniently be divided into two parts: (1) derivation of the replicate estimation weights, which only needs to be done once, and (2) calculation of replicate estimates and their variance. Thus analysts can be supplied with a file containing the sets of replicate weights, and need not know any of the details of the complex design in order to derive valid variance estimates for their analyses. In resampling methods it is also easy, in principle, to account for variability due to various adjustments to the weights, such as nonresopnse (NR) adjustments and calibration, simply by applying the same adjustment procedures to each individual set of replicate weights. For these reasons we decided to use a replication procedure. We chose the bootstrap over the jackknife for two reasons: (1) the bootstrap yields valid variance estimators from a reasonable number of replicates (*e.g.*, 500), whereas the jackknife requires a replicate for each sample cluster, and (2) as developed by Rao *et al.* (1992) the bootstrap has better properties for non-smooth statistics such as quantiles or the low income cutoff (LICO - see Kovačević and Yung, 1997)

In this paper we describe enhancements to the bootstrap method used for variance estimation in the NPHS. In particular, we describe how the effects of estimation of NR adjustment factors are incorporated into the bootstrap weights for the third wave of the survey (this was not done for the second wave), and compare bootstrap variance estimates derived from the enhanced method to those obtained with the old method. Section 2 describes the bootstrap variance estimation method used in the NPHS. Section 3 describes the NR adjustment procedures, and how they were included in the bootstrap. Section 4 presents an analysis of the effects of having excluded these adjustments from the bootstrap.

2. Bootstrap Method

In this section we describe in some detail the bootstrap method used in the NPHS. The bootstrap resampling method for the iid case 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. It is this Rao, Wu andYue version of the bootstrap that was implemented in the NPHS. It assumes L design strata, where stratum h contains N_h clusters of which $n_h \geq 2$

clusters are sampled with replacement. Typically the clusters are actually selected without replacement, but the without replacement assumption is usually a reasonable approximation. There is no restriction on subsampling within clusters. We let w_{hik} denote the inverse sampling probability weight for the kth sample individual from the ith sample cluster of stratum h.

The bootstrap variance estimator for an estimator $\hat{\theta}$ is calculated as follows:

(i) Weighting: Independently for each stratum, select a simple random sample with replacement of m_h clusters from the n_h sampled clusters. With m^*_{hi} denoting the number of times the *(hi)*-th cluster is selected ($\sum_i m^*_{hi} = m_h$), the bootstrap weights are defined as

$$w^*_{hik} = \left[1 - \left(\frac{m_h}{n_h - 1} \right)^{1/2} + \left(\frac{m_h}{n_h - 1} \right)^{1/2} \frac{n_h}{m_h} m^*_{hi} \right] w_{hik}.$$

For the NPHS, m_h was set to n_h-1, a commonly used value, ensuring that the bootstrap weights, w^*_{hik}, are nonnegative, and reducing the above equation to

$$w^*_{hik} = \frac{n_h}{n_h - 1} m^*_{hi} w_{hik}.$$

(ii) Estimation: Calculate $\hat{\theta}^*$, the bootstrap replicate of estimator $\hat{\theta}$, by replacing the survey weights w_{hik} by the bootstrap weights w^*_{hik} in the formula for $\hat{\theta}$.

(iii) Independently replicate steps (i) and (ii) a large number of times, say B, and calculate the corresponding replicate estimates, $\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(B)}$. The bootstrap variance estimator for $\hat{\theta}$ is then given by

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

where $\hat{\theta}^*_{(\cdot)} = (1/B) \sum_b \hat{\theta}^*_{(b)}$.

For the NPHS, as for most other surveys, the sampling weights are adjusted to account for Nonresponseand then calibrated to known population totals. Logically these adjustments to the weights ought to be considered part of the estimation process, *i.e.,* step (ii) above. However, it is practically convenient to consider them separately, and to include them as part of the weighting process, since they are identical for all estimates derived from the survey. The survey weights provided with the NPHS microdata files incorporate these adjustments. The files of bootstrap weights are similarly adjusted.

For the bootstrap weights derived for waves 1 and 2 of the NPHS, the bootstrap weight adjustment was applied after the NR adjustment, for simplicity. In other words, it was the NR-adjusted sampling weights that were adjusted for bootstrap subsampling as in step (i) above. These bootstrap-adjusted weights were then calibrated to population totals in the same way that the estimation weights were. Thus there is a component of the overall variance, that due to estimation of the NR adjustment factors, that is not reflected in variance estimates derived from these bootstrap weights. For cycle 3 of the survey, some of the NR adjustments were included in the bootstrap procedure, and it is the aim of this paper to analyse the effects of this enhancement.

Note 2.1. When the bootstrap subsampling is done before the NR adjustment, sample units are eligible to be selected in the bootstrap sample regardless of their response status. If the subsampling takes place after the NR adjustment, it is only the respondent units that are included in the bootstrap samples. There are a small number of clusters, generally smaller clusters, which are completely nonrespondent. Such clusters may still be included in the bootstrap sample of clusters in step (i) above when NR adjustment takes place after the subsampling, but not if it takes place before the subsampling. This means that the samples of clusters available for subsampling were slightly different for the two bootstrap procedures. For the longitudinal file there were 3,264 sample clusters available for the bootstrap when the NR adjustment was included, and 3,194 when it was not. For the health file the numbers were 4,266 and 4,114. The importance of this fact will be discussed further in Section 4.

3. Nonresponse Adjustment

In this section we describe the NR adjustments and how they were incorporated into the bootstrap.

Generally speaking, the approach taken to NR adjustment for the NPHS is to form NR adjustment classes which are thought to be homogeneous with respect to propensity to respond to the survey. Within adjustment classes the weights of the respondents are simply multiplied by a factor to make the sum of the adjusted weights for the respondents equal to the sum of the unadjusted weights of both respondents and nonrespondents. The NR classes are formed using variables that are available for both respondents and nonrespondents and that are good predictors of response status. For a cross-sectional survey the information available for formation of NR adjustment classes is usually quite limited. In a longitudinal survey, the survey variables observed for respondents to one wave of the survey can be used as potential predictors of

Nonresponseto subsequent waves of the survey. This is the approach taken for NPHS, using a CHAID (chi-square automatic interaction detection) algorithm to find good NR adjustment classes when such predictors are available from previous waves of the survey. To be more specific, the Knowledge-Seeker software was used to help determine the NR adjustment classes. When response predictors from previous waves are not available, as in the case of top-up samples, simple geographic classes are used. A more detailed description of NR adjustment for NPHS can be found in Tambay, Şchiopu-Kratina, Mayda, Stukel and Nadon (1998).

For the third wave of the NPHS, we can distinguish two different types of data file: longitudinal and cross-sectional. The longitudinal file contains data for all waves of the survey, but only for individuals who were fully respondent to all waves (alternative definitions are possible, and indeed used, allowing varying degrees of partial response or nonresponse to different waves of the survey). The cross-sectional file includes all respondents to the third wave of the survey, regardless of their response status for previous waves of the survey. This includes the top-up samples that were added specifically to improve cross-sectional representativity.

For the bootstrap, the NR adjustment classes were taken as fixed and it is only the adjustment factors that were recalculated and reapplied for each bootstrap sample. However, this was not possible when a bootstrap sample included some nonrespondents but no respondents from a particular adjustment class since there were then no respondents to absorb the weight of the nonrespondents. In that case the problem class would be collapsed with one or more other NR adjustment classes. Normally the NR adjustment classes would be carefully chosen in light of the survey data; however, in the context of a resampling method like the bootstrap a more automated solution is required. We handled this problem by prespecifying, for each NR adjustment class, how it would be collapsed with others in case this problem arose. In actual fact the problem occured only a few times in 500 bootstrap samples.

3.1. Adjustment classes for the longitudinal file. For the longitudinal file, adjustments for nonresponseare cumulative. At the first wave of the survey, simple classes based on stratum and season of data collection were used. Nonrespondents to wave 1 were then excluded from the sample for wave 2 and the NR adjustment classes for wave 2 were based on wave 1 variables available for both respondents and nonrepondents to wave 2. Similarly, the NR classes for wave 3 were based on variables from waves 1 and 2. The overall wave 3 longitudinal NR adjustment factors were then just the products of the factors calculated for the three waves of nonresponse.

Only the last two of these three NR adjustments was included in the bootstrap procedure. Recovering the information needed to include the first adjustment into the bootstrap would have been too difficult and error-prone. Effective inclusion of such weight adjustment procedures into the bootstrap requires some advance planning, so that the information needed to replicate the procedure is available. After the NR adjustment, final calibration of the weights was done for each bootstrap sample as it had been for the complete sample.

The average longitudinal NR adjustment factor for wave 2 was 1.115 at the Canada level. Provincial averages ranged from 1.070 to 1.145, while individual NR adjustment classes had adjustment factors as high as 1.830. For wave 3 the response rates were generally better, with the average adjustment being 1.081, provincial averages ranging from 1.046 to 1.092, and the largest individual adjustment factor being 1.538.

3.2. Adjustment classes for the cross-sectional file. For the cross-sectional file the NR adjustments were done separately for the longitudinal panel and for the top-up samples. The panel part of the sample included all respondents to the first wave of the survey, regardless of response status for the second wave. Variables collected at wave 2 were used as potential predictors of nonresponse, but when these were not available wave 1 variables were used as proxies if possible. There were three different top-up samples: a general top-up sample for attrition, and samples of newborns and of immigrants to account for new entrants to the population. The first stage of the top-up sample for attrition consisted of the nonrespondent dwellings from the first wave of the survey; thus there was no need to apply a wave 1 NR adjustment to the panel part of the sample. NR adjustments were done separately within each of the top-ups. Because of small sample sizes, the NR adjustments for the top-ups used simple classes based on province.

For the cross-sectional files there was the additional problem of extreme weights which could arise when a longitudinal panel member moved from a province where the sampling rate was relatively low/high to one where it was high/low. In these cases, which were very few in number, adhoc adjustments were made to the weights for these units. In the bootstrap the same adhoc adjustment factors were applied to these problem cases. Finally the weights were calibrated as for the complete sample.

The average NR adjustment for the cross-sectional file was 1.143 at the Canada level, while provincial averages ranged from 1.081 to 1.188. The largest individual adjustment factor was 2.441 which occured in

one of the top-up samples. The largest adjustment factor for the panel part of the sample was 1.881.

4. Empirical investigation.

In this section we investigate the effects of including and not including the estimation of the NR adjustment factors in the bootstrap. We compare variance estimates derived from the two bootstrap methods, and also look at the effects on the NR adjustment factors and on the weights. We tried to compare the variance estimates analytically using approximate variance expressions based on Taylor linearization; however, the expressions are very complex and no real insight was gained by this. Comparison of the expressions may be possible by considering the asymptotic order of different terms, but it is not clear what the most appropriate asymptotic framework is, and the result can depend on the framework chosen. However, we are considering the question further. In the meantime, a small empirical comparison is presented here.

4.1. Comparison of variances. To compare variances we calculated coefficients of variation (CVs) for 51 totals using the two different versions of the bootstrap for both the longitudinal and health files. We also conducted logistic regression analyses to regress the probability of onset of back problems onto several potential predictors (34 predictors for the longitudinal file, 8 for the cross-sectional).

We had expected that inclusion of the NR adjustments in the bootstrap would increase the bootstrap estimated variances, since an extra source of variability was being accounted for. To our surprise this seemed not to be the case. The adjusted bootstrap variance estimates were sometimes smaller than the unadjusted estimates. In addition, the differences turned out to be fairly small.

We speculated that these unexpected results may have been due to the fact that the sets of bootstrap samples for the two procedures had been selected independently. They had to be selected independently since the samples of clusters available for subsampling were slightly different for the two procedures, as explained above in Note 2.1. We then decided to compare results of the two procedures when they were based on the same bootstrap sample of clusters. We therefore selected a bootstrap subsample of clusters restricting to those sample clusters that had at least one respondent, and applied the two different weighting procedures. The differences between the two sets of variance estimates were now smaller, as expected, but the adjusted estimates were still sometimes smaller than the unadjusted estimates.

In order to explore this phenomenon further we replicated the entire procedure 50 times, and for comparison we also considered the bootstrap weights before the final calibration to age/sex/province population totals. Thus we obtained 50 replicate sets of four bootstrap weight files, each set being based on common bootstrap subsamples of clusters. The four files in each set were as follows: ADJCAL - NR adjustment factors recalculated for each bootstrap sample and final calibration applied; NOACAL - NR adjustment factors fixed and final calibration applied; ADJNOC - NR adjustment factors recalculated for each bootstrap sample but no calibration of the weights; NOANOC - NR adjustment factors fixed and no calibration. Each bootstrap weight file contained 500 bootstrap weights.

We calculated the mean differences between pairs of sets of CV estimates, as well as the mean relative differences, *i.e.* the mean of the difference divided by the unadjusted estimate, and histograms of both of these. For the regressions we also looked at whether a decision on the significance of a regressor, at the 5% or 1% significance level, would be different for the adjusted and unadjusted procedures, though not even one such difference was found in 50 replicates.

Summary results of the simulations are presented in Tables 4.1 and 4.2. We still found that the adjusted variance estimates for some of the parameters were smaller than the unadjusted estimates, and the Monte Carlo errors now indicated that most of these differences were statistically significant.

The first row of these tables, labelled ADJCAL-NOACAL, summarize the differences between the new bootstrap procedure and the old one, *i.e.*, with final calibration applied after the NR adjustments. It can be seen that, although the differences between the two procedures are statistically significant, these differences are small both absolutely and relatively, and the direction of the differences does not consistently favour either the adjusted or the unadjusted bootstrap. The practical insignificance of the differences is underlined by the fact that in 50 replicates, not a single difference in significance of any of the regressors was found, either at the 5% or the 1% level.

A possible heuristic explanation for the unadjusted variance estimates to be sometimes larger is that the NR adjustments and final calibration adjustments may interact in such a way that when the NR adjustment factors were not included in the bootstrap, the variability due to the calibration adjustment factors was exagerated.

In an attempt to more clearly see the effect of recalculating the NR adjustment factors on the estimates of variance, the third row of the tables compares the adjusted and unadjusted procedures leaving out the final

calibration of the weights. However, results are very similar to those observed in the first row, with very small differences and no consistent direction.

The effect of final weight calibration on the variances of estimators is summarized in the second and fourth rows of the tables, ADJCAL-ADJNOC and NOACAL-NOANOC, which compare compare the variances of estimates using calibrated weights to those of estimates using uncalibrated weights, both with and without recalculation of the adjustment factors included. The effects of calibration are much larger and more significant than those of recalculation of the NR adjustment factors, generally, but not always, leading to a reduction in the variance of totals. In some cases the variance is reduced by a factor of two or more. For the logistic regressions calibration is less likely to lead to a reduction of the variance, but the effect of calibration is still much larger.

4.2. Other comparisons. In order to get another perspective on the effects of adding calculation of the NR adjustment factors into the bootstrap procedures, we looked at the means and variances of the recalculated NR adjustment factors over 500 bootstrap replicates.

For the cross-sectional file the average bootstrap CV of the adjustment factors, weighted by the total sample weight within each class, was about 3.3%, while the average absolute relative bias was about 0.3%. CVs and biases for some of the individual NR adjustment factors were much larger, particularly for some of the adjustment classes in the top-up which sometimes contained just a few sample elements. For the panel part of the survey the largest CV was 17.7% and the largest absolute relative bias was 3.1%. Overall then, the variability in the adjustment factors was not very large, though there were small pockets of high variability.

The situation for the longitudinal file was similar. For the wave 2 NR adjustments the average CV was 2.3% and the average absolute relave bias was 0.1%. The largest CV was 21.8% and the largest relative bias was 3.1%. For wave 3 the corresponding averages were 2.5% for the CVs and 0.2% for the relative biases, while the maximum values were 38.2% and 7.2%, respectively. Again it can be said that the overall variability of the adjustment factors is not very large.

Finally, we compared directly the new bootstrap weights (ADJCAL) to the old weights (NOACAL) for the longitudinal file. We looked at relative differences of the weights, *i.e.*, $(w_{ij}^{*} - w_{ij})/w_{i}$, where the subscript i indicates the sample unit and j denotes the jth bootstrap replicate. We found the overall mean of the absolute values of these quantities to be 0.027, and the mean of the within sample unit standard deviations to be 0.048. Thus incorporating the adjustment had an appreciable, though not very large, effect on individual weights.

4.3. Conclusion. In summary, the effect of including calculation of NR adjustment factors in the bootstrap seems to be practically insignificant, at least for the empirical examples considered here for Statistics Canada's National Population Health Survey. In contrast, the effects of final calibration are larger by an order of magnitude.

REFERENCES

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Kovačević, M. and Yung, W. (1997). Variance estimation for measures of income inequality and polarization—An Empirical Study. *Survey Methodology*, 23, 41-52.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.

Tambay, J.-L. and Catlin, G. (1995). Sample design of the National Population Health Survey. *Health Reports* (Statistics Canada Catalogue No. 82-003), Volume 7 (1), 33-42.

Tambay, J.-L., Şhiopu-Kratina, I, Mayda, J., Stukel, D. and Nadon, S. (1998). Treatment of Nonresponse in Cycle Two of the National Population Health Survey. *Survey Methodology*, 24, 147-156.

Yung, W. (1997). Variance estimation for public use microdata files. *Proceedings of the Statistics Canada Symposium*, 91-95.

Table 4.1. Differences of average percentage CVs for 51 totals

	Longitudinal File				Cross-sectional File			
	mean ave diff[1]	mean ave rel diff[2]	max. ave rel diff pos/neg	no. sig diff[3] pos/neg	mean ave diff	mean ave rel diff	max. ave rel diff pos/neg	no. sig diff pos/neg
ADJCAL-NOACAL	-0.0051	-0.0028	0.0078 -0.0198	9 26	-0.0020	-0.0002	0.0125 -0.0109	22 21
ADJCAL-ADJNOC	-0.1184	-0.0822	0.0573 -0.5519	11 37	-0.1321	-0.1452	0.0164 -0.6607	7 38
ADJNOC-NOANOC	-0.0113	-0.0064	0.0116 -0.0248	8 40	-0.0029	-0.0051	0.0107 -0.0430	15 33
NOACAL-NOANOC	-0.1245	-0.0860	0.0587 -0.5648	11 38	-0.1330	-0.1499	0.0152 -0.7053	8 39

1. mean ave difference: mean over 51 totals of average differences over 50 replications
2. rel diff: (CV1- CV2)/CV2
3. no. sig diff - number of totals (out of 51) for which the average difference divided by the Monte Carlo standard deviation of the differences was greater than 2.

Table 4.2. Differences of average percentage CVs for logistic regression parameters

	Longitudinal File (34 parameters)				Cross-sectional File (8 parameters)			
	mean ave diff[1]	mean ave rel diff	max. ave rel diff pos/neg	no. sig diff pos/neg	mean ave diff	mean ave rel diff	max. ave rel diff pos/neg	no. sig neg diff
ADJCAL-NOACAL	0.2025	0.0011	0.0051 -0.0030	19 7	-0.0000	-0.0004	0.0013 -0.0038	2 3
ADJCAL-ADJNOC	-22.08	-0.0176	0.8094 -0.6579	17 17	1.2849	0.0374	0.0940 -0.0621	6 2
ADJNOC-NOANOC	0.1311	0.0003	0.0036 -0.0028	13 8	-0.0155	-0.0009	0.0023 -0.0066	2 4
NOACAL-NOANOC	-22.15	-0.0183	0.8057 -0.6564	17 17	1.2694	0.0371	0.0939 -0.0628	6 2