

Catalogue no. 11-522-XIE

**Statistics Canada International  
Symposium Series - Proceedings**

**Symposium 2006 :  
Methodological Issues in  
Measuring Population Health**

2006



**Statistics  
Canada**

**Statistique  
Canada**

**Canada**

## Bootstrap Methods for Analyzing Complex Sample Survey Data

J. N. K. Rao<sup>1</sup>

### Abstract

Application of standard methods to survey data without accounting for the design features and weight adjustments can lead to erroneous inferences. Bootstrap methods offer an attractive option to the analyst for taking account of the design features and weight adjustments. The data file consists of the full-sample final weights and associated bootstrap final weights for a large number of bootstrap replicates as well as the observed data on the sample elements. We show how such data files can be used to analyze survey data in a straightforward manner using weighted estimating equations. A one-step estimating function bootstrap method that avoids some difficulties with the bootstrap is also discussed.

KEY WORDS: Bootstrap weights; Design features; Logistic regression; Weighted estimating equations.

### 1. Introduction

The principal steps in a sample survey are survey design, data collection and processing, estimation and analysis of data. In this paper, we focus on the analysis of data from complex large-scale surveys, cross-sectional as well as longitudinal. Data obtained from such surveys typically involve stratified multi-stage cluster sampling, leading to dependencies among sample elements, and unequal probabilities of selection, leading to unequal design weights. Moreover, design weights are often calibrated to known population totals of auxiliary variables as well as subjected to unit non-response adjustments. As a result, application of standard statistical methods to survey data without accounting for the design features and weight adjustments could lead to incorrect inferences even for large samples: underestimation of standard errors of estimators, inflated type I error rates (probabilities of rejecting a true null hypothesis) and erroneous model diagnostics. This difficulty with standard methods motivated the development of new methods that account for the survey design and lead to asymptotically valid inferences. We refer the reader to the books by Skinner, Holt and Smith (1989), Chambers and Skinner (2003) and Lehtonen and Pahkinen (2004) for excellent accounts of such methods.

We focus on finite population parameters that can be formulated as solutions to “census” estimating equations (EE) and also on parameters of an underlying super-population model that leads to EE. The EE approach can handle a wide variety of finite population (or descriptive) parameters as well as model parameters. To make inference on the parameters of interest under stratified multi-stage cluster sampling, we can use either Taylor linearization methods or re-sampling methods that include the jackknife, balanced repeated replication (BRR) and the Rao-Wu bootstrap (Rao and Wu, 1988; Rao, Wu and Yue, 1992). An advantage of re-sampling methods is that a single standard error formula is used for all estimators, unlike the linearization which requires the derivation of a separate formula for each case. Moreover, linearization can become cumbersome in handling post-stratification and non-response adjustments, whereas it is relatively straightforward with re-sampling methods. Among the re-sampling methods, bootstrap offers an attractive option to the analyst for taking account of the survey design. It is easy to implement and more flexible than the jackknife and the BRR in terms of number of replicates (or re-samples), and provides valid inferences for both smooth and non-smooth statistics (such as the median) unlike the jackknife. The data file for bootstrap implementation will consist of the full-sample final weights and associated bootstrap final weights for a large number of bootstrap replicates as well as the observed data on the sample elements.

---

<sup>1</sup> J. N. K. Rao, Carleton University, School of Mathematics and Statistics, Ottawa, Canada, K1S 5B6

In this paper, we show how such data files can be routinely used to analyze survey data in a straightforward manner using sample weighted estimating equations (Sections 2 and 3). We also give a one-step estimating function (EF) bootstrap method that avoids a difficulty with the bootstrap (Section 4). Statistics Canada is currently using bootstrap methods for analyzing data from several large-scale surveys.

## 2. Weighted Estimating Equations

We focus on stratified multi-stage designs with large number of strata  $L$ , and relatively few primary sampling units (clusters)  $n_h (\geq 2)$ , sampled within each stratum  $h (= 1, \dots, L)$ . We assume that sub-sampling within sampled clusters  $i (= 1, \dots, n_h)$  is performed to ensure unbiased estimation of cluster totals. The basic design weight attached to the element  $hik$  in the sample  $s$  is denoted by  $d_{hik}$ . The weights  $d_{hik}$  are calibrated to known population totals of auxiliary variables and adjusted for unit non-response. The resulting final weights are denoted by  $w_{hik}$ ,  $hik \in s$  and reported in the data file along with the observed responses  $y_{hik}$  and predictor variables  $x_{hik}$  (possibly vector-valued).

Many parameters of interest can be formulated as the solutions to census estimating equations

$$U(\theta) = \sum u_{hik}(\theta) = 0, \quad (1)$$

where the summation is over the population elements  $hik$  (Binder, 1983). For example, the choices  $y_{hik} - \theta$ ,  $I(y_{hik} \leq \theta) - \frac{1}{2}$  and  $x_{hik}(y_{hik} - x'_{hik}\theta)$  for  $u_{hik}(\theta)$  in (1) give the solution  $\theta_N$  as population mean  $\bar{Y}$ , population median and population linear regression coefficient, respectively. No super-population model is assumed and the parameters are descriptive including the regression coefficient which measures how much of the variation in the response  $y$  is explained by the predictor variables  $x$ . However, the regression coefficient may be motivated by a linear regression model  $E_m(y_{hik}) = \mu_{hik} = x'_{hik}\theta$  with constant error variance, where  $\theta$  is the model parameter and  $E_m$  denotes model expectation. Similarly, a census logistic regression coefficient is obtained by letting  $u_{hik}(\theta) = x_{hik}(y_{hik} - \mu_{hik})$ , where  $\log\{\mu_{hik}/(1 - \mu_{hik})\} = x'_{hik}\theta$ .

We are interested in the descriptive (census) parameters  $\theta_N$  as well as the model parameters  $\theta$  when a model is specified. Note that  $\theta_N$  is a consistent estimator of the model parameter  $\theta$  under a specified model. So an estimator  $\hat{\theta}$  of both  $\theta_N$  and  $\theta$  (under a model) is obtained by estimating the population total  $U(\theta)$  using the final weights  $w_{hik}$  and solving the weighted estimating equations

$$\hat{U}(\theta) = \sum_{hik \in s} w_{hik} u_{hik}(\theta) = 0. \quad (2)$$

Typically, the Newton-Raphson (NR) algorithm is used to find the solution  $\hat{\theta}$  iteratively from (2). The  $(r+1)$ -th step of the NR iteration is given by

$$\hat{\theta}_{r+1} = \hat{\theta}_r + [\hat{J}(\hat{\theta}_r)]^{-1} \hat{U}(\hat{\theta}_r), \quad (3)$$

where  $\hat{\theta}_r$  is the solution at the  $r$ -th iteration, and  $\hat{J}(\hat{\theta}_r)$  and  $\hat{U}(\hat{\theta}_r)$  are the values of  $\hat{J}(\theta) = -\partial \hat{U}(\theta) / \partial \theta'$  and  $\hat{U}(\theta)$  evaluated at  $\theta = \hat{\theta}_r$ . Iterating the algorithm to convergence produces the solution  $\hat{\theta}$ .

The estimator  $\hat{\theta}$  is design-consistent for  $\theta_N$  and design and model consistent for  $\theta$ . In the case of non-response, we need to assume that the non-response adjustment leads to consistent estimators. We refer the reader to Thompson (1997, Chapter 5) for an excellent account of estimating equations for survey data.

Although the point estimator is the same for both the census and the model parameters, inferences generally differ. In the case of census parameter, we estimate the design variance of  $\hat{\theta}$  while in the case of model parameter we estimate the design-model variance (or the total variance)  $V(\hat{\theta}) = E_m V_d(\hat{\theta}) + V_m E_d(\hat{\theta}) \approx E_m V_d(\hat{\theta}) + V_m(\theta_N)$ , where  $E_d, V_d$  denote design expectation and design variance and  $V_m$  denotes model variance. A robust estimator of the first term  $E_m V_d(\hat{\theta})$  is obtained by estimating the design variance  $V_d(\hat{\theta})$ , but the estimation of the second term  $V_m(\theta_N)$  requires the specification of the model covariance structure. However, for the multi-stage cluster sampling designs, the second term is small relative to the first term if  $n/N$  is negligible, where  $n$  and  $N$  denote the total number of sampled clusters and the total number of clusters in the population. In this case, the estimator of design variance can also be used to estimate the total variance.

### 3. Rao-Wu Bootstrap

In the Rao-Wu bootstrap method, a bootstrap sample is obtained by drawing a simple random sample of  $m_h = n_h - 1$  clusters with replacement from the  $n_h$  sample clusters, independently for each stratum  $h$ .

We select a large number,  $B$ , of independent bootstrap samples which can be represented in terms of bootstrap frequencies  $m_{hi}(b) = \text{number of times } (hi) \text{-th sample cluster is selected in the } b \text{-th bootstrap sample } (b = 1, \dots, B)$ . Typically,  $B = 500$  bootstrap samples are used in Statistics Canada. The bootstrap design weights are simply given by  $d_{hik}(b) = d_{hik} \{n_h / (n_h - 1) m_{hi}(b)\}$ . Now replace  $d_{hik}$  by  $d_{hik}(b)$  in the method for getting the final weights from  $d_{hik}$  to get the bootstrap final weights  $w_{hik}(b)$ . Data file of weights will consist of  $\{w_{hik}, w_{hik}(1), \dots, w_{hik}(B)\}$ . We refer the reader to Girard (2006) for practical issues in generating the bootstrap weights in the context of the Canadian National Longitudinal Survey of Children and Youth.

Substituting the bootstrap weights  $w_{hik}(b)$  for  $w_{hik}$  in the weighted estimating equations (2), we get the bootstrap estimating equations

$$\hat{U}_b(\theta) = \sum_{hik \in s} w_{hik}(b) u_{hik}(\theta) = 0 \quad (4)$$

It is sufficient to perform a one-step NR iteration of (4) with the full sample estimate  $\hat{\theta}$  as the starting value to get the bootstrap estimates  $\hat{\theta}(b)$  given by

$$\hat{\theta}(b) = \hat{\theta} + [\hat{J}_b(\hat{\theta})]^{-1} \hat{U}_b(\hat{\theta}), \quad (5)$$

where  $\hat{J}_b(\theta) = -\partial \hat{U}_b(\theta) / \partial \theta'$ , provided  $\hat{J}_b(\hat{\theta})$  is invertible. However, Binder et al. (2004) found that it is possible to have several bootstrap samples  $b$  for which  $\hat{J}_b(\hat{\theta})$  is ill-conditioned and hence not invertible.

Assuming that  $\hat{\theta}(b)$  is available for all the bootstrap samples, the Rao-Wu bootstrap estimator of the covariance matrix of  $\hat{\theta}$  (denoted variance estimator of  $\hat{\theta}$ ) is given by

$$v_B(\hat{\theta}) = B^{-1} \sum_{b=1}^B [\hat{\theta}(b) - \hat{\theta}][\hat{\theta}(b) - \hat{\theta}]' \quad (6).$$

Square root of the diagonal elements of (6) will give bootstrap standard errors of the components of the vector  $\hat{\theta}$ . Another bootstrap variance estimator is obtained by substituting  $\hat{\theta}(\cdot) = B^{-1} \sum_b \hat{\theta}(b)$  for  $\hat{\theta}$  in (6). The two bootstrap variance estimators are asymptotically equivalent, but standard errors from (6) will be slightly larger than those from the alternative bootstrap variance estimator. The bootstrap variance estimators lead to over-estimation of the design variance  $V_d(\hat{\theta})$ , but the over-estimation will be small if the overall first stage sampling fraction  $n/N$  is small even if the first stage sampling fractions  $n_h/N_h$  for some strata  $h$  are not negligible (Shao, 2002). Moreover, over-estimation of the design variance might in fact be useful in estimating the total variance of  $\hat{\theta}$  and hence making inference on the model parameters  $\theta$ .

Using the estimator  $\hat{\theta}$  and the associated bootstrap variance estimator, one can also obtain large-sample confidence intervals on the components of  $\theta$  and perform Wald and quasi-score tests (Rao, Scott and Skinner, 1998, Rao, 1999) of hypotheses on  $\theta$ . It is assumed that the number of first stage sample clusters,  $n$ , is large.

Beaumont and Bocci (2006) applied the bootstrap method to find  $P$ -values for testing a linear hypothesis of the form  $H_0: A\theta = c$  under a linear regression model  $y_{hik} = x'_{hik}\theta + \varepsilon_{hik}$ , where  $A$  is a specified matrix with full rank,  $c$  is a specified vector and  $\varepsilon_{hik}$  are the model errors. A survey weighted  $F$ -statistic,  $F(c)$ , for testing  $H_0$  can be obtained by using the WEIGHT statement in the procedure REG in SAS, provided all the final sample weights are positive. Treating this statistic as a  $F$ -variable leads to erroneous  $P$ -values even for large samples. Bootstrap weights can be used to find valid  $P$ -values. We simply use the SAS WEIGHT statement with  $w_{hik}$  replaced by the bootstrap weights  $w_{hik}(b)$  and  $c$  in  $H_0$  replaced by  $A\hat{\theta}$  to get bootstrap weighted  $F$ -statistic  $F_b(A\hat{\theta})$  for each bootstrap sample  $b = 1, \dots, B$ . The bootstrap  $P$ -value is then given by the proportion of bootstrap  $F$ -values  $F_b(A\hat{\theta})$  exceeding the full sample  $F$ -value  $F(c)$ . Null hypothesis is rejected if the bootstrap  $P$ -value is less than a prescribed level, say 0.05. A simulation study suggested that the proposed method has good power properties and maintains size of the test. It would be useful to extend this method to other models and hypotheses. Note that the method is designed for testing hypotheses.

#### 4. Estimating Function Bootstrap

In Section 3, we noted that the direct Rao-Wu bootstrap method can run into difficulty with ill-conditioned matrices when performing NR iterations to get the bootstrap estimates of  $\theta$ . To overcome this problem, Rao and Tausi (2004) proposed an estimating function (EF) bootstrap approach, motivated by the work of Hu and Kalbfleisch (2000) for the non-survey case. We approximate the estimating function  $\hat{U}(\theta)$  by the bootstrap estimating function  $\hat{U}_b(\hat{\theta})$  and solve

$$\hat{U}(\theta) = \hat{U}_b(\hat{\theta}) \quad (7)$$

for  $\theta$  using one-step NR with  $\hat{\theta}$  as the starting value to get

$$\tilde{\theta}(b) = \hat{\theta} - [\hat{J}(\hat{\theta})]^{-1} \hat{U}_b(\hat{\theta}) \quad (8)$$

Note that for all  $b = 1, \dots, B$  the inverse matrix in (8) is based on the full sample and hence remains the same, so that repeated inversion for each bootstrap sample is not needed. Thus the problem with ill-conditioned matrices for the direct bootstrap method is avoided and also computations are simplified.

To get the EF bootstrap variance estimator of  $\hat{\theta}$ , we simply replace  $\hat{\theta}(b)$  in (6) by :

$$v_{EFB}(\hat{\theta}) = B^{-1} \sum_{b=1}^B [\tilde{\theta}(b) - \hat{\theta}][\tilde{\theta}(b) - \hat{\theta}]' \quad (9)$$

Binder, Kovacevic and Roberts (2004) approximated  $\hat{U}(\theta)$  by  $-\hat{U}_b(\hat{\theta})$  instead of approximating by  $\hat{U}_b(\hat{\theta})$ , but the resulting EF bootstrap variance estimator (6) based on the one-step NR is algebraically identical to the Rao-Tausi bootstrap variance estimator. Hu and Kalbfleisch (2000) have shown that approximating  $\hat{U}(\theta)$  by  $\hat{U}_b(\hat{\theta})$  gives confidence intervals with good properties.

Roberts, Rao and Ren (2006) considered marginal logistic regression models with binary response data obtained from a longitudinal survey having a complex design. They adapted the generalized estimating equations (GEE) method of Liang and Zeger (1986) to the case of longitudinal survey data and estimated the model parameters as the solutions to survey-weighted GEE. They used the odds ratio approach to model the working covariance structure for the repeated measurements on the sample individuals. Inferences on the model parameters were based on the EF bootstrap standard errors which account for within subject dependence of repeated measurements as well as clustering of subjects and other survey design features. They applied the theory to longitudinal data from Statistics Canada's National Population Health Survey (NPHS). The NPHS began in 1994/95 and it collects information every two years from the same sample individuals. A stratified multi-stage cluster sampling design was used to select households within clusters, and then one household member 12 years or older was chosen to be the longitudinal respondent.

## References

- Beaumont, J.-F., and Bocci, C. (2006). "A Practical Bootstrap Method for Testing Hypotheses from Survey Data", unpublished report, Ottawa, Canada: Statistics Canada.
- Binder, D. (1983), "On the Variance of Asymptotically Normal Estimators from Complex Surveys", *International Statistical Review*, 51, pp. 279-293.
- Binder, D., Kovacevic, M, and Roberts, G. (2004), "Design-based Methods for Survey Data: Alternative Uses of Estimating Functions", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp.
- Chambers, R., and Skinner, C. J. (2003), *Analysis of Survey Data*, Chichester: Wiley.
- Girard, C. (2006), "How to Avoid Getting All Tied Up Bootstrapping a Survey: A Walk-Through Featuring the National Longitudinal Survey of Children and Youth", unpublished report (draft version), Ottawa, Canada: Statistics Canada.
- Hu, F., and Kalbfleisch, J.D. (2000), "The Estimating Function Bootstrap", *Canadian Journal of Statistics*, 28, pp. 449-499.
- Lehtonen, R., and Pahkinen, E. (2004), *Practical Methods for Design and Analysis of Complex Surveys, 2<sup>nd</sup> Edition*, Chichester: New York.

- Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models". *Biometrika*, 73, pp. 13-22.
- Rao, J. N. K., and Wu, C. F. J. (1988), "Resampling Inference with Complex Survey Data", *Journal of the American Statistical Association*, 83, pp. 231-241.
- Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992), "Some Recent Work on Resampling Methods", *Survey Methodology*, 18, pp. 209-217.
- Rao, J. N. K., Scott, A. J., and Skinner, C. J. (1998), "Quasi-Score Tests with Survey Data", *Statistica Sinica*, 8, 1059-1070.
- Rao, J. N.K. (1999), "Some Current Trends in Sample Survey Theory and Methods", *Sankhya B*, 61, pp. 1-57.
- Rao, J. N. K., and Tausi, M. (2004), "Estimating Function Jackknife Variance Estimators Under Stratified Multistage Sampling", *Communications in Statistics*, 33, pp. 2087-2095.
- Roberts, G., Ren, Q., and Rao, J. N. K. (2006), "Using Marginal Mean Models with Data from a Longitudinal Survey Having a Complex Design: Some Advances in Methods", to appear in *Methodology of Longitudinal Surveys*, Chichester: Wiley.
- Shao, J. (2002), "Resampling Methods for Variance Estimation in Complex Surveys with Imputed Data", in R. M. Groves et al. (eds) *Survey Nonresponse*, New York: Wiley, pp. 303-314.
- Skinner, C. J., Holt, D., and Smith, T. M. F. (1989), *Analysis of Complex Surveys*, New York: Wiley.
- Thompson, M. E. (1997), *Theory of Sample Surveys*, London: Chapman and Hall.