DESIGN-BASED METHODS FOR SURVEY DATA: ALTERNATIVE USES OF ESTIMATING FUNCTIONS

David Binder, Milorad Kovacevic, Georgia Roberts
Methodology Branch, Statistics Canada, Ottawa, ON, Canada K1A 0T6

Key Words: design-based variances, linearization methods, analysis of complex survey data, bootstrap methods

Abstract: The appropriateness of using design-based methods to analyze survey data is now well recognized. Research has shown that design-based methods offer some protection against model misspecification and informative sample designs. The use of design-based methods has grown now that software has been developed to make such methods more accessible to data analysts.

A variety of approaches are currently in use by analysts for estimating design-based variances of estimated model parameters, including linearization, balanced repeated replication, jackknifing and bootstrapping**.** The particular approach to bootstrapping popularly used with survey data - which we call the direct survey bootstrap - suffers from some of the same deficiencies as the standard bootstrap: in particular, the estimated variances can be unstable in certain circumstances. Recently, methods have been proposed for making inferences using an estimating function bootstrap in a model-based setting, which seem to provide more stable results. These methods have been adapted to produce different design-based estimating function survey bootstraps (EFSB). In this paper, through a simulation study motivated by a real-life analysis, we compare the properties of the direct survey bootstrap with these EFSB's and linearized versions of them. A linearized EFSB approach seems to recommend itself for the analysis of survey data.

A. INTRODUCTION

A.1 Approaches to Variance Estimation Used for Inference with Survey Data

With the increase in use of design-based methods for the analysis of survey data, considerable attention has been paid to determining adequate and accessible methods of obtaining design-based covariance estimates required for inference.

The traditional approach for producing variance estimates of estimated quantities is an analytical one - frequently referred to as the Taylor-linearization approach. It was the first to be implemented in commercial analysis software. SUDAAN, originated by Babu Shah, was one of the software leaders in making the Taylor-linearized method readily accessible. This approach builds on the standard formula for estimating the variance of a total, for a particular survey design, when the estimate of the total has the form of a weighted sum and where the weights are the inverse of inclusion probabilities. This standard formula is then adapted, through a linearization approach, to account for weight adjustments such as nonresponse and poststratification and to accommodate nonlinear statistics such as ratio estimates.

While the standard formula for the variance estimate of a total is straightforward for simple survey designs, it can be much more complicated for the more complex designs actually used in practice. However, it is often feasible to make the simplifying assumption that the true design can be approximated by that of a stratified multistage design with with-replacement sampling of primary sampling units at the first stage. Under this assumption, a variance estimation formula for a simple total has a form that is readily calculated. Yet, in more recent years, there has still been a move away from this analytical approach. A major reason for this is that the analytical approach requires development of a new formula for every estimator and for every weight adjustment - or requires further simplifying assumptions about negligible impact of ignoring some of the complexities of these adjustments.

In moving away from an analytical approach for variance estimation, there has been a move towards replication methods. This move has occurred for several reasons. First, there has been growing popularity of replication methods in other areas of statistics, which has lead to their consideration for design-based variance estimation. This has been due, in large part, to the greater computing power available to analysts. As well, for design-based variance estimation, the replication methods have the advantages of being able to accommodate the impact of weight adjustments and do not require the development of a new formula for every estimator. Versions of jackknifing, balanced repeated replication, and bootstrapping are all available for complex survey data. See Rust and Rao (1996) for the properties of some of these estimators.

A.2 Why Consider Other Bootstrapping Approaches Than the Direct Survey Bootstrapping Approach

The direct survey bootstrap approach, which is described in detail in Section B.1, is one of the replication methods that has gained popularity for use with survey data and is supported for use with several of Statistics Canada's social surveys. It has the advantage that, once bootstrap samples have been taken and bootstrap weights calculated, the user estimates the quantities of interest in exactly the same way with the full sample and with each of the bootstrap samples, and then combines these estimates to obtain variance estimates.

One disadvantage, however, with this direct bootstrap approach is that the estimating equations may not have a solution for one or more of the bootstrap samples, even though a solution is possible for the full sample. This seems to happen particularly frequently when working with a small sample, where some of the bootstrap samples lead to ill-conditioned matrices to be inverted. Thus, a bootstrapping approach that did not have this disadvantage would be preferable.

A search of the non-survey literature identified research on methods for bootstrapping the estimating function rather than

directly bootstrapping the quantities of interest. Hu and Kalbfleisch (2000), for example, obtained more stable variance estimates by this approach. An extension of this idea to the survey case seemed worthy of investigation.

Finally, while computers are more powerful, any bootstrapping approach can still be time-intensive if it involves a full iterative model-fitting with each separate bootstrap sample. Logistic regression modeling is an example of such a type of model. It seemed reasonable to examine variance approaches that could make use of the bootstrap samples but not require full iterative fits.

B. BOOTSTRAPPING VARIANCE ESTIMATION METHODS TO BE COMPARED

In this study, as is generally done when selecting a variance estimation approach with survey data, the assumption is made that the true design of the survey may be approximated by a stratified multistage design where there is with-replacement sampling of primary sampling units (psu's) at the first stage. The survey units forming the *b*-th bootstrap replicate are obtained by sampling $n_h - 1$ psu's independently with replacement from the n_h sampled psu's in each stratum**,** and the *b*-th bootstrap weight variable is created by adjusting the survey weight variable on each unit to account for the results of the replicate sampling and for any other adjustments done to the survey weight. In total, *B* bootstrap replicates and weights are created.

The objective is to get an estimate $\hat{\theta}$ of the finite population vector parameter θ, and an estimate of the covariance matrix of $\hat{\theta}$. To begin, we define the parameter θ as the solution of the population estimating equation $U(\theta) = \sum_U u_i(\theta) = 0$, where the $u_i(\theta)$ are suitably defined. As examples, when θ is the vector of coefficients of a linear regression model, $u_i(\theta) = x_i(y_i - x_i'\theta)$, while for the coefficients of a logistic regression model, $u_i(\theta) = x_i[y_i - p_i(\theta)]$, where $p_i(\theta) = \exp(x_i'\theta)/[1+\exp(x_i'\theta)]$. We can then produce the sample estimating function $\hat{U}(\theta) = \sum_s w_i u_i(\theta)$ which is a weighted sum over the sample of components $u_i(\theta)$, where w_i is the value of the survey weight variable on the *i*th unit in the sample. We then define $\hat{\theta}$ as the solution to the estimating equation $\hat{U}(\theta) = 0$, i.e. $\hat{U}(\hat{\theta}) = \sum_s w_i u_i(\hat{\theta}) = 0$.

To obtain an estimate of the covariance matrix of $\hat{\theta}$, several approaches are considered in this paper, and are described below. All make use of the *B* bootstrap replicates and weights described above.

B.1 Direct Approach to Bootstrapping for Survey Data

The direct survey bootstrap approach is one of the replication methods that has gained popularity for use with survey data. It is, for example, the main variance estimation approach that is supported for several of the Statistics Canada social surveys. The steps for obtaining the direct survey bootstrap variance estimate of the estimate $\hat{\theta}$ of vector θ are as follows:

i) Use the full sample to get estimate $\hat{\theta}$ of θ as described above.

ii) Calculate the estimate $\hat{\theta}^{(b_d)}$ using the *b*-th replicate weight variable - in the same way as $\hat{\theta}$ was calculated using the survey weight variable; that is, define $\hat{\theta}^{(b_d)}$ as the solution to the *b*-th bootstrap estimating equation $\hat{U}^{(b)}(\theta) = \sum_s w_i^{(b)} u_i(\theta) = 0$, where $w_i^{(b)}$ is the value of the *b*-th bootstrap weight variable for the *i*th unit in the sample. Do this for each of the *B* replicates.

iii) Calculate the direct bootstrap estimate of the covariance matrix of $\hat{\theta}$ as

$$\hat{V}_{direct} = \sum_{b=1}^{B} (\hat{\theta}^{(b_d)} - \hat{\theta})(\hat{\theta}^{(b_d)} - \hat{\theta})' / B$$

A problem that can arise with this approach is that, one or more of the bootstrap estimating equations may not have a solution; this is generally due to a matrix involving the bootstrap weight variable, which must be inverted, being ill conditioned for each of these resamples, even when the matrix for the full sample is not ill conditioned. This problem seems to be more prevalent when working with a small sample. The usual approach to dealing with this situation for variance estimation is to eliminate such bootstrap samples and use only those for which equation solutions can be obtained

B.2 Linearized Estimating Function (LEF) Bootstrap Approach

Our first step is to define $\hat{\theta}$ as the solution to the estimating equation $\hat{U}(\theta) = 0$, as described above. We then make a linear approximation to $\hat{U}(\hat{\theta})$ around $\hat{U}(\theta)$:

$$0 = \hat{U}(\hat{\theta}) \approx \hat{U}(\theta) + \frac{\partial \hat{U}(\theta)}{\partial \theta}(\hat{\theta} - \theta).$$

Rearranging this equation, we can obtain an expression for $\hat{\theta} - \theta$,

$$\hat{\theta} - \theta \approx -\left(\frac{\partial \hat{U}(\theta)}{\partial \theta}\right)^{-1} \hat{U}(\theta),$$

which then leads to a sandwich form expression to approximate the variance of $\hat{\theta}$:

$$V(\hat{\theta}) = \left(\frac{\partial \hat{U}(\theta)}{\partial \theta}\right)^{-1} V[\hat{U}(\theta)] \left[\left(\frac{\partial \hat{U}(\theta)}{\partial \theta}\right)'\right]^{-1}$$

It follows that the estimated variance of $\hat{\theta}$ will also have the sandwich form:

$$\hat{V}(\hat{\theta}) = \left(\frac{\partial \hat{U}(\theta)}{\partial \theta}\right)^{-1} \hat{V}[\hat{U}(\theta)] \left[\left(\frac{\partial \hat{U}(\theta)}{\partial \theta}\right)'\right]^{-1}$$

evaluated at $\theta = \hat{\theta}$.

We then make use of the bootstrap replicates and the estimating function to obtain the "meat" of the sandwich in the following way. We calculate the value of the b-th bootstrap estimating function at $\hat{\theta}$, i.e. $\hat{U}^{(b)}(\hat{\theta}) = \sum_s w_i^{(b)} u_i(\hat{\theta})$, for b=1,2,...,B. Then we define the LEF bootstrap variance estimate of $\hat{U}$ at $\theta = \hat{\theta}$ to be

$$\hat{V}_{BS}\left[\hat{U}(\theta)\right] = \sum_{b=1}^{B}\left[\hat{U}^{(b)}(\hat{\theta}) - \hat{U}(\hat{\theta})\right]\left[\hat{U}^{(b)}(\hat{\theta}) - \hat{U}(\hat{\theta})\right]' / B,$$

which finally leads to

$$\hat{V}_{LEF}(\hat{\theta}) = \left(\frac{\partial \hat{U}(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}}\right)^{-1} \hat{V}_{BS}\left[\hat{U}(\theta)\right]\left[\left(\frac{\partial \hat{U}(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}}\right)'\right]^{-1}.$$

It should be noted that we did not actually obtain an estimate of theta from each bootstrap sample. However, if such estimates are desired (such as for the production of diagnostics), the following may be done.

Recall that $$\hat{\theta} - \theta \approx -\left(\frac{\partial \hat{U}(\theta)}{\partial \theta}\right)^{-1} \hat{U}(\theta).$$

It then follows that a reasonable definition for the b-th bootstrap estimate of theta through the LEF approach would be

$$\hat{\theta}^{b_{LEF}} = \hat{\theta} - \left(\frac{\partial \hat{U}(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}}\right)^{-1} \hat{U}^{(b)}(\hat{\theta}).$$

If we consider the particular case of θ being the vector of coefficients of a logistic regression model, the values of the quantities given above are the following:

$$\hat{U}(\theta) = \sum_s w_i\left[y_i - p_i(\theta)\right], \text{ where}$$
$$p_i(\theta) = \exp(x_i'\theta)/\left[1+\exp(x_i'\theta)\right];$$
$$\hat{U}^{(b)}(\hat{\theta}) = \sum_s w_i^{(b)}\left[y_i - p_i(\hat{\theta})\right];$$
$$\frac{\partial \hat{U}(\theta)}{\partial \theta} = -\sum_s w_i x_i x_i' p_i(\theta)\left[1 - p_i(\theta)\right]; \text{ and}$$

$$\hat{\theta}^{b_{LEF}} = \hat{\theta} - \left\{\sum_s w_i x_i x_i' p_i(\hat{\theta})\left[1 - p_i(\hat{\theta})\right]\right\}^{-1}\left\{\sum_s w_i^{(b)}\left[y_i - p_i(\hat{\theta})\right]\right\}$$

When compared to the direct bootstrapping approach, there are advantages to the LEF bootstrap approach described above. For one, estimating equations are solved only once - using the full sample - rather than B+1 times, which can be a considerable time advantage, particularly when iterative procedures are used in the solution. There is also no problem with ill-conditioned matrices, provided that an ill-conditioned matrix is not encountered when fitting the model with the full sample; this is a particular advantage with small samples. It should also be noted that the bootstrap weights generated for use in the direct bootstrap approach may be used for implementing the LEF bootstrap approach.

Advantages are gained, but at the cost of disadvantages. One great advantage of the direct bootstrap approach is that you simply "turn the crank" as you apply it to different point estimates; it is a completely repetitive process. However, with the LEF bootstrap, it is necessary to do the "linearization math" for each different model. Yet, for software packages where Taylor linearization is already an option, the LEF bootstrap should be a relatively simple addition.

B.3 One Estimating Function Approach

The motivation for this approach is as follows:

Recall that, in the direct bootstrap approach, $\hat{\theta}^{(b_d)}$ satisfies $0 = \hat{U}^{(b)}(\hat{\theta}^{(b_d)}) = \sum_s w_i^{(b)} u_i(\hat{\theta}^{(b_d)})$.

It is readily seen that the right-hand side of this equation may be decomposed as:

$$\sum_s w_i^{(b)} u_i(\hat{\theta}^{(b_d)}) = \sum_s w_i u_i(\hat{\theta}^{(b_d)}) + \sum_s w_i^{(b)} u_i(\hat{\theta}) + R,$$

where $R = \sum_s (w_i - w_i^{(b)})\left[u_i(\hat{\theta}) - u_i(\hat{\theta}^{(b_d)})\right]$. In many situations, R is small. Thus, if R is dropped and the remainder of the right-hand side is set to 0 and the equation is solved for $\hat{\theta}^{(b_d)}$, you would be getting the solution to $\hat{U}(\theta) = -\hat{U}^{(b)}(\hat{\theta})$. Binder and Patak (1994) use a similar approach for developing this form of estimating functions.

Thus, the EF approach consists of defining the b-th bootstrap estimate of θ, say $\hat{\theta}^{(b_{EF})}$, to be the value of θ that is the solution of $\hat{U}(\theta) = -\hat{U}^{(b)}(\hat{\theta})$ and the EF bootstrap variance estimate of $\hat{\theta}$ to be

$$\hat{V}_{EF} = \sum_{b=1}^{B}(\hat{\theta}^{(b_{EF})} - \hat{\theta})(\hat{\theta}^{(b_{EF})} - \hat{\theta})' / B.$$

This approach, like the direct bootstrap, again requires the solution of a system of equations for every bootstrap sample, and again can encounter ill-conditioned matrices needing inversion - although not necessarily for the same bootstrap samples as the direct. In practice, such "bad" samples could be eliminated before estimation of the variance.

It should be noted that, if a Newton-Raphson approach is used to solve the system of equations for each bootstrap sample and if the full-sample estimate $\hat{\theta}$ is used as the starting value for θ, then the first-step estimate of θ by this approach is $\hat{\theta}^{b_{LEF}}$. It then follows that if this first-step estimate is used in the variance estimate, then $\hat{V}_{LEF}$ is obtained. It should also be noted that this first-step estimate of θ may be calculated for every bootstrap sample, so that the variance estimate would be based on the full number of bootstrap replicates.

B.4 An Alternative Estimating Function Approach

Recall that $\hat{U}(\theta)$ and $\hat{U}^{(b)}(\hat{\theta})$ are defined respectively as $\hat{U}(\theta) = \sum_s w_i u_i(\theta)$ and $\hat{U}^{(b)}(\theta) = \sum_s w_i^{(b)} u_i(\theta)$. Define the b-th bootstrap estimate of θ, say $\hat{\theta}^{(b_{EF2})}$, to be the value of θ that is the solution of $\hat{U}(\theta) = U^{(b)}(\hat{\theta})$. Then estimate the

variance of $\hat{\theta}$ by

$$\hat{V}_{EF2} = \sum_{b=1}^{B} (\hat{\theta}^{(b_{EF2})} - \hat{\theta})(\hat{\theta}^{(b_{EF2})} - \hat{\theta})' / B .$$

This approach, like the direct bootstrap and the EF method described in B.3, requires the solution of a system of equations for every bootstrap sample, and again can encounter ill-conditioned matrices needing inversion - although not necessarily for the same bootstrap samples as for these other methods. In practice, such "bad" samples could be eliminated before estimation of the variance, which is what we did.

Instead of solving the system of equations exactly for each bootstrap sample, if a Newton-Raphson approach is followed, the full-sample estimate $\hat{\theta}$ may be used as the starting value, and the first-step estimate could be taken as the estimate of θ. The first-step estimate of θ by this approach is

$$\hat{\theta}^{b_{LEF2}} = \hat{\theta} + \left(\frac{\partial \hat{U}(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right)^{-1} \hat{U}^{(b)}(\hat{\theta}). \qquad \text{(B.1)}$$

It should be noted that if this estimate is used in the variance formula, then $\hat{V}_{LEF}$ is obtained, even though $\hat{\theta}^{(b_{LEF2})} \neq \hat{\theta}^{(b_{LEF})}$. Rao and Tausi (2004) proposed a similar first-step estimate with a jackknife approach to variance estimation.

C. DESCRIPTION OF SIMULATION STUDY

C.1 Model

In order to compare the properties of the variance estimators presented in Section B, we carried out an extensive simulation study. We simulated a model of the relationship between ***the loss of independence among seniors*** (***LOSS***) and several factors associated with their health status, living arrangements and habits, according to a model fitted to the data from the first two cycles (1994/95 and 1996/97) of the Canadian National Population Health Survey (NPHS), as presented in Martel, Bélanger and Berthelot (2002).

The model that we simulated relates the probability of a senior losing his independence to his age, sex, body mass index, number of chronic diseases and smoking habits, in the following form:

$$\begin{aligned} \text{logit}(LOSS) = & \\ & -3.284 + 0.382 * SEX + 1.388 * AGEGR \\ & + 0.624 * UNDERWGT - 0.109 * OVERWGT \\ & + 0.641 * CHRDS + 0.484 * SMOK \end{aligned} \qquad \text{(C.1)}$$

All variables in the model are binary: *LOSS* (1, if person has lost his independence within the last two years, 0 if still independent), *SEX* (0 for women, 1 for men), *AGEGR* (0 for age in [65,75), 1 for age in [75 +)), *UNDERWGT* (1 for *BMI* $\leq$ 18.5, 0 otherwise), *OVERWGT* (1 for *BMI* $\geq$25, 0 otherwise), *CHRDS* (1 if at least one of 10 chronic conditions is present, 0 otherwise), *SMOK* (1 if presently smokes daily or if quit recently, 0 otherwise). All variables other than *LOSS* are measured at the start of the two-year period.

Note that the reference values for all variables included in the logistic model (C.1) are 0. The variables related to *BMI* (body mass index) originate from a variable *BMIGR* with three categories (0 for *BMI* $\leq$ 18.5, 1 for 18.5<*BMI* < 25, and 2 for *BMI* $\geq$ 25). The ten chronic conditions considered by Martel et al. (2002) were asthma, arthritis, back problems, bronchitis/emphysema, diabetes, heart disease, cancer, effects of stroke, urinary incontinence, and glaucoma/cataracts.

C.2 The Finite Population

We simulated a finite population of 2.5 million individuals which has some of the characteristics of the Canadian NPHS subpopulation of elderly people, aged 65 and more. The variables were generated as Bernoulli random variables using the joint probabilities estimated from the NPHS sample. (To conserve space, the values of these probabilities are not given here, but may be obtained from the authors.)

After having simulated values for *SEX*, *AGEGR*, etc., the dependent variable *LOSS* was also created as a Bernoulli variable with probability equal to

$$p_x = p(LOSS = 1 \mid x) = \left[1 + \exp(-x'\theta)\right]^{-1},$$

where x and $\theta = (\beta_1, \beta_2, ..., \beta_7))$ are defined by model (C.1). The proportion of individuals in the simulated finite population who lost their independence is 0.1009.

C.2.1. "Natural" cluster structure

Several additional variables, not used in model (C.1), were also simulated. The idea was to create variables which could be used to define a "natural" cluster structure of the simulated finite population. By a natural structure we mean clusters of similar individuals. These additional binary variables are: *URBRUR* (0 if area of residence is urban, and 1 otherwise), *EDU* (0 if education is less then postsecondary and 1 otherwise), *INC* (0 for low income and 1 for medium/high income). These variables were simulated using the conditional probabilities estimated from NPHS within age-sex groups. For the purpose of this simulation study, a fictional propensity score of 'accessibility to medical care facilities' was defined for each individual as

$$\begin{aligned} p(ACCESS) = \{1 + \exp[& -(-2 - 2.5 * URBRUR + 0.8 * EDU \\ & + 0.75 * INC - 1.0 * LOSS)]\}^{-1} \end{aligned} \qquad \text{(C.2)}$$

As described below, *ACCESS* was used as a variable in the sample design. The variable *LOSS* is included in (C.2) to make the sample design of the simulation study informative, essentially meaning that the distribution of the sampled units is different to some extent from the distribution that would be obtained by sampling directly from the model (see Binder and Roberts, 2001, 2003). A possible control for the amount of informativeness is the coefficient for LOSS in expression (C.2).

The individual records were placed into four classes according to the value of their accessibility score, as shown in Table 1.

Table 1. *Four accessibility score classes*

Class	Accessibility	*p*(access)
1	Low	$p < 0.015$
2	Medium Low	$0.015 \le p < 0.048$
3	Medium High	$0.048 \le p < 0.15$
4	High	$p \ge 0.15$

We then arranged the finite population into clusters in two different ways - one with smaller cluster sizes between 10 and 50 records and the other with larger clusters containing between 50 and 150 individual records. These arrangements will be called Clustering S and Clustering L respectively. These clusterings were accomplished in the following manner:

The records were ordered according to accessibility class, but with random order within class. For Clustering S, they were then assigned to clusters whose sizes were generated as random integers between 10 and 50, using the uniform distribution U(10, 50). Clustering S resulted in 83,378 clusters. For Clustering L, assignment was done similarly, but the cluster sizes were generated as random integers between 50 and 150, resulting in 25,009 clusters For example, for Clustering L, the first random number was 79 implying that the first 79 individual records belong to the first cluster; the second random number was 113, so that the next 113 individuals belong to the second cluster, etc. In this way, individuals with similar values of "accessibility" were placed in the same cluster.

C.3 Sample design

A sample of clusters was selected without replacement with the probability of selection proportional to the cluster size by the Sampford method as implemented in SAS procedure SURVEYSELECT.

The original sampling weights produced by this procedure were then calibrated to the known totals of *AGEGR*, *SEX* and *URBRUR*. The calibration was done by an iterative raking procedure on the marginal totals. After the calibration, the weights of the records from the same clusters were not all equal as was the case with the original sampling weights.

We considered two sample sizes - 25 clusters and 50 clusters. We then selected 500 samples of 25 clusters from each of the two population arrangements, and also 500 samples of 50 clusters.

It should be noted that there was no stratification and that no subsampling was done within clusters.

C.4 Bootstrapping

For each cluster sample of size *n* (where *n* = 25 or 50), we produced 500 bootstrap replicates by taking a simple random sample with replacement of size *n*-1 clusters. The bootstrap weights in the b*th* bootstrap replicate were then obtained by first adjusting the original sampling weights to reflect the repetition of some and the exclusion of other clusters, following the formula

$$w_{ij}^{(b)} = w_{ij}\, k_i^{(b)}\, \frac{n}{n-1},$$

where w_{ij} is the original sampling weight of the *jth* individual from the i*th* cluster, $k_i^{(b)}$ is the number of repetitions of the i*th* cluster in the b*th* bootstrap replicate. Note that $\sum_i k_i^{(b)} = n-1$. Then these weights were calibrated to the known totals of *AGEGR*, *SEX* and *URBRUR*, in the same way as the original full sample weights were calibrated.

The estimation of coefficients $\boldsymbol{\beta}$ of logistic model (C.1) involves the inversion of a matrix formed from the sample observations. If, for a selected sample, this matrix was ill-conditioned, the sample was eliminated and another one chosen, until the desired number of samples was attained. However, given a selected sample, there was still the possibility that a bootstrap replicate drawn from the sample could produce an ill-conditioned matrix when solving the equation required for the direct bootstrap or for one of the two variants of the estimating function bootstrap. In such a case, this replicate is simply excluded for that particular variance estimation procedure and that estimate of the variance is then based on a reduced number of replicates. It should be noted that, for the linearized estimating function approach, this problem does not arise since the matrix that is inverted is the same matrix that is inverted when estimating $\boldsymbol{\beta}$ from the full sample.

C.5 Measures for Comparing Variance Estimation Methods

The coefficients of the logistic regression model were estimated from each of the simulated samples. The variances of these estimates were estimated using the replication methods described in the previous sections: direct bootstrap (Direct), linearized estimating function bootstrap (LEF), and two versions of the estimating function bootstrap (EF and EF2). The methods are compared with respect to accuracy, stability and the coverage properties of the resulting interval estimates, and also with respect to the distributional properties of the corresponding bootstrap estimates of the model coefficients. The measures used for the comparisons are described below. In these sections, β is used to represent any one of the 7 coefficients of the logistic model, since each coefficient is examined in the same way. As well, $\hat{\beta}$ is used to denote an estimate of β, in general, while $\hat{\beta}_k$ denotes the estimate of β based on the *k*th sample. The letter M is used to denote any one of the variance estimation methods being compared. Finally, n_s denotes the number of samples being used (which is 500).

C.5.1 Accuracy

The accuracy of method M for coefficient β is measured by the relative bias calculated as

$$relBIAS(\hat{V}_M) = \frac{\sum_k \hat{V}_{M,k} / n_s - EMSE(\hat{\beta})}{EMSE(\hat{\beta})} \qquad (C.3)$$

where $EMSE(\hat{\beta})$ is the empirical mean square error of $\hat{\beta}$ over the n_s independent samples, i.e.,

$$EMSE(\hat{\beta}) = \sum_k (\hat{\beta}_k - \beta)^2 / n_s .$$

It is considered as being a 'true' MSE of the estimated coefficient.

Our special concerns would be the methods that have appreciable negative relative bias because such methods would lead to overstated precision and significance.

C.5.2 Stability

The stability of method M is assessed by the Relative Root Mean Square Error of the estimated variance:

$$relRMSE(\hat{V}_M) = \frac{\sqrt{\sum_k \left[\hat{V}_{M,k} - EMSE(\hat{\beta})\right]^2 / n_s}}{EMSE(\hat{\beta})}. \qquad (C.4)$$

A bootstrap method should have a value of Relative Root Mean Square Error close to zero.

C.5.3 Coverage Properties

To evaluate the effectiveness of normal-theory confidence intervals, empirical coverage rates for method *M* were computed for nominal confidence coefficients of 100(1-α)%=90 and 95 percent, by

$$coverage_{1-\boldsymbol{\alpha}}(\hat{V}_M) = \frac{\sum_k I\left\{\left|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}\right| / \sqrt{\hat{V}_{M,k}} \le z_{\boldsymbol{\alpha}/2}\right\}}{n_s}$$

where $\boldsymbol{I}\{\mathbf{a}\}$=1 if **a** is true, and 0 otherwise, and $z_{\alpha/2}$ is the upper α/2th standard normal percentile. Upper and lower tail error rates were also computed but are not reported here.

C.5.4 Distributional Properties of Bootstrap Replicate Coefficient Estimates

In order to assess the distributional properties of the coefficient estimates obtained from the bootstrap replicates based on the different methods of variance estimation, we compared the moments of the distributions of the coefficient estimates obtained from the bootstrap replicates, averaged over the n_s samples, to the distributions of the coefficient estimates obtained from the n_s samples. The justification for doing this comparison is that, when using bootstrapping for carrying out inference on a parameter θ, it is generally assumed that the distribution of $\hat{\theta} - \theta$ can be approximated by that of $\hat{\theta}^{(b)} - \hat{\theta}$.

First, based on n_s samples, we estimated the first four moments of the distribution of $\hat{\beta} - \beta$: $E(\hat{\beta} - \beta)$, $E(\hat{\beta} - \beta)^2$, $skewness(\hat{\beta} - \beta) = skewness(\hat{\beta})$, $kurtosis(\hat{\beta} - \beta) = kurtosis(\hat{\beta})$. These estimates were obtained by using the SAS UNIVARIATE procedure with the coefficient estimates from the n_s samples from which β was subtracted. These estimates are denoted by $\hat{E}(\hat{\beta} - \beta)$, $\hat{E}(\hat{\beta} - \beta)^2$, $\hat{skewness}(\hat{\beta})$, $\hat{kurtosis}(\hat{\beta})$.

Next, for method M and for each of the n_s samples, we had up to 500 bootstrap replicates (depending on the method) which resulted in that many bootstrap estimates $\hat{\beta}_M^{(b)}$ of the model parameter β. These bootstrap estimates from the same sample (from which $\hat{\beta}$ estimated from that sample was subtracted) were then used to estimate the first four moments - namely the mean, MSE, skewness, and kurtosis - again using PROC UNIVARIATE. Then, the means of these estimated moments were calculated over the n_s samples. We denote these resultant quantities by $\hat{E}_{M,rep}(\hat{\beta}_M^{(b)} - \hat{\beta})$, $\hat{E}_{M,rep}(\hat{\beta}_M^{(b)} - \hat{\beta})^2$, $\hat{skewness}_{M,rep}(\hat{\beta}_M^{(b)} - \hat{\beta})$, and $\hat{kurtosis}_{M,rep}(\hat{\beta}_M^{(b)} - \hat{\beta})$.

Finally, for method M, the estimates of each of the moments were compared to the full sample estimates through the calculation of the following quantities:

$$Q1(M) = \left[\hat{E}_{M,rep}(\hat{\boldsymbol{\beta}}_M^{(b)} - \hat{\boldsymbol{\beta}}) - \hat{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] / |\boldsymbol{\beta}|$$

(Relative difference in biases)

$$Q2(M) = \left[\hat{E}_{M,rep}(\hat{\boldsymbol{\beta}}_M^{(b)} - \hat{\boldsymbol{\beta}})^2 - \hat{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2\right] / \hat{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2$$

(Scaled difference in MSE's)

$$Q3(M) = \left[\hat{skewness}_{M,rep}(\hat{\boldsymbol{\beta}}_M^{(b)} - \hat{\boldsymbol{\beta}}) - \hat{skewness}(\hat{\boldsymbol{\beta}})\right]$$

(Difference in skewness)

$$Q4(M) = \left[\hat{kurtosis}_{M,rep}(\hat{\boldsymbol{\beta}}_M^{(b)} - \hat{\boldsymbol{\beta}}) - \hat{kurtosis}(\hat{\boldsymbol{\beta}})\right]$$

(Difference in kurtosis)

A bootstrap method should have a good match to the full sample estimates for all four moments, not just for the first two moments. Good matches would be indicated by values of $Q1$ to $Q4$ close to 0. It should be noted that Q2(M) and $RelBIAS(\hat{V}_M)$ are the same.

D. RESULTS OF SIMULATION STUDY

This section describes the results of the comparisons of the different bootstrap-based variance estimation methods with respect to their accuracy, stability, coverage and distributional properties.

As pointed out in Section C.4, the variance estimates based on the different bootstrap-based methods could potentially use different numbers of bootstrap replicates. Chart 1 presents the degree of rejection of bootstrap replicates by the different methods for the four different arrangements of cluster size (S -10 to 50 units, or L - 50 to 150 units)and number of clusters

(small -25 or large -50 clusters) examined. When both the number of clusters and the size of clusters is small (yielding an average sample size of 750 units), the EF2 method has the greatest degree of replicate rejections. When either the number of clusters or the cluster size is increased, the amount of rejection declines considerably, for the EF2 method and also for the direct and EF methods. There is very little rejection with any of the methods when both the number of clusters and the cluster size are large (and the average sample size was 5000 units). The LEF method has no rejection under any condition.

Results for the other measures are presented in Charts 2-6. Each of these charts contains four graphs, one for each combination of cluster size (S or L) and number of clusters (25 or 50). In each graph, the 6 model coefficients other than the intercept are along the horizontal axis and the value of the measure is along the vertical axis. A separate line shows the results for each method. Note that the LEF2 method is given a separate line only in Chart 6 because, in the other charts, its values are the same as the LEF method.

D.1 Accuracy (and Q2)

Chart 2. presents the results on accuracy of variance estimation for the four compared methods as measured by the relative bias defined in equation (C.3). Apart from the case of large cluster size with large number of clusters, the LEF method is the least biased. The method with systematically large positive bias in these three situations is EF2, thus showing a strong tendency towards overestimation of variance. Direct BS and the EF bootstrap methods performed at almost identical levels of accuracy. For large cluster size and large number of clusters, the methods had similar performances. Overall the LEF bootstrap method showed the smallest relative bias.

D.2 Stability

The stability of the variance estimation procedures as measured by the relative MSE of the estimated variances is presented in Chart 3. The ranking of the methods remains the same as that seen for the relative bias. The LEF bootstrap method is the most stable. The Direct bootstrap is next to LEF. Again, the Direct method and EF show almost identical patterns of stability.

D.3 Coverage Properties

Charts 4 and 5 present the two-sided interval coverage rates estimated for 95% and 90% nominal confidence respectively, along with the confidence bounds. The coverage rates vary over the variables. However, the ranking of the methods is quite consistent. The LEF bootstrap method usually overstates the true coverage while the EF2 generally understates. The coverage rates obtained by the Direct method are overall the closest to the nominal rates. There is a great similarity between the rates obtained by the EF method and the Direct.

D.4 Distributional Properties of Bootstrap Replicate Coefficient Estimates

The comparison of the distributional properties of the parameter estimates, $\hat{\beta}_M^{(b)}$, based on different methods, with the distributional properties of the full sample estimates is summarized in Charts 6 A) to D). The quantities used for making the comparisons are the values $Q1$ to $Q4$ defined in Section C.5.4.

The first moments of the distribution of the LEF bootstrap estimates of the parameters are the closest to the first moments of the distribution of the full sample. This can be explained by the way these estimates are derived from the full sample estimates. The quantity Q2 is the relative bias of the variance estimates and is discussed in section D.1. Similarity of the skewness and the kurtosis of the LEF bootstrap estimates and those based on the full sample distribution dominates over the other methods. The overall finding regarding the distributional properties of different bootstrap estimates is that the LEF estimates outperform the estimates obtained by the other methods.

E. DISCUSSION AND RECOMMENDATIONS

E.1 Discussion

As can be seen from Chart 1, when samples consist of a small number of small clusters, methods other than LEF reject a considerable number of bootstrap samples. Increasing the size of cluster decreases the rate of rejection somewhat, but a greater improvement is achieved for these methods when the number of clusters is increased, so that when you have both larger clusters and a greater number of clusters, rejection is negligible. At least a portion of the improvement over the situation of a small number of small clusters is likely attributable to a larger number of ultimate units in the samples.

LEF and the first-step version of EF are equivalent. LEF and the first-step version of EF2 (B.1), called LEF2, yield equivalent variance estimates although the coefficient estimates from the same bootstrap sample differ in value.

Overall, the Direct and EF methods have very similar properties.

For a smaller number of clusters sampled (i.e. 25) LEF is usually the best (or close to it).

For more clusters sampled (i.e. 50) the other methods evaluated perform as well as LEF, with the exception of EF2, which generally performed more poorly.

The survey design used here, being only single-stage, is simpler than that frequently used for analytic surveys,. As well, cluster size averaging 30 units is in a realistic range, while clusters of average size 100 would be generally larger than what would be seen for many analytic surveys. Yet, the results of this simulation study have led to interesting new

findings regarding the behavior of the LEF bootstrap in comparison to the other methods studied in this paper.

E.2 Recommendations

Based on the results of this simulation study, we recommend the LEF method. For the properties evaluated, it performed the best overall. As well, LEF is computationally much faster than the Direct, EF or EF2 methods since there are no new matrices to invert after having inverted the matrices for the full-sample estimates and there is no iteration involved. It could also be readily implemented with commercial analysis software if a Newton-Raphson approach is supported, provided that the software allows the specification of the estimating function, the starting values and the matrix of derivatives may be specified, and the solution after one Newton-Raphson iteration may be obtained.

References

Binder, D.A. and Patak, Z. (1994), "Use of Estimating Functions for Estimation From Complex Surveys," *Journal of American Statistical Association*, 89, 1035-1043.

Binder, D.A. and Roberts, G. R. (2001) "Can informative designs be ignorable?" Newsletter of the Survey Research Methods Section, American Statistical Association, Issue 12. (January, 2001)

Binder, D.A. and Roberts, G. (2003). "Statistical Inference for Survey Data Analysis," 2003 Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, 568-572.

Hu, F. and Kalbfleisch, J.D. (2000), "The estimating function bootstrap," *Canadian Journal of Statistics*, 28, 449-481.

Martel, L., Bélanger, A., and Berthelot, J.-M. (2002), "Loss and Recovery of Independence among Seniors," *Health Reports*, 13, 35-48.

Rao, J.N.K. and Tausi, M. (2004). "Estimating Function Jackknife Variance Estimation under Stratified Multistage Sampling," *Communications in Statistics*, 33, 2087 – 2095.

Rust, K.F. and Rao, J.N.K. (1996) Variance Estimation for Complex Surveys Using Replication Techniques. *Statistical Methods in Medical Research*, 5, 215-238.

Chart 1. Number of samples (out of 500) by range of rejection of bootstrap replicates

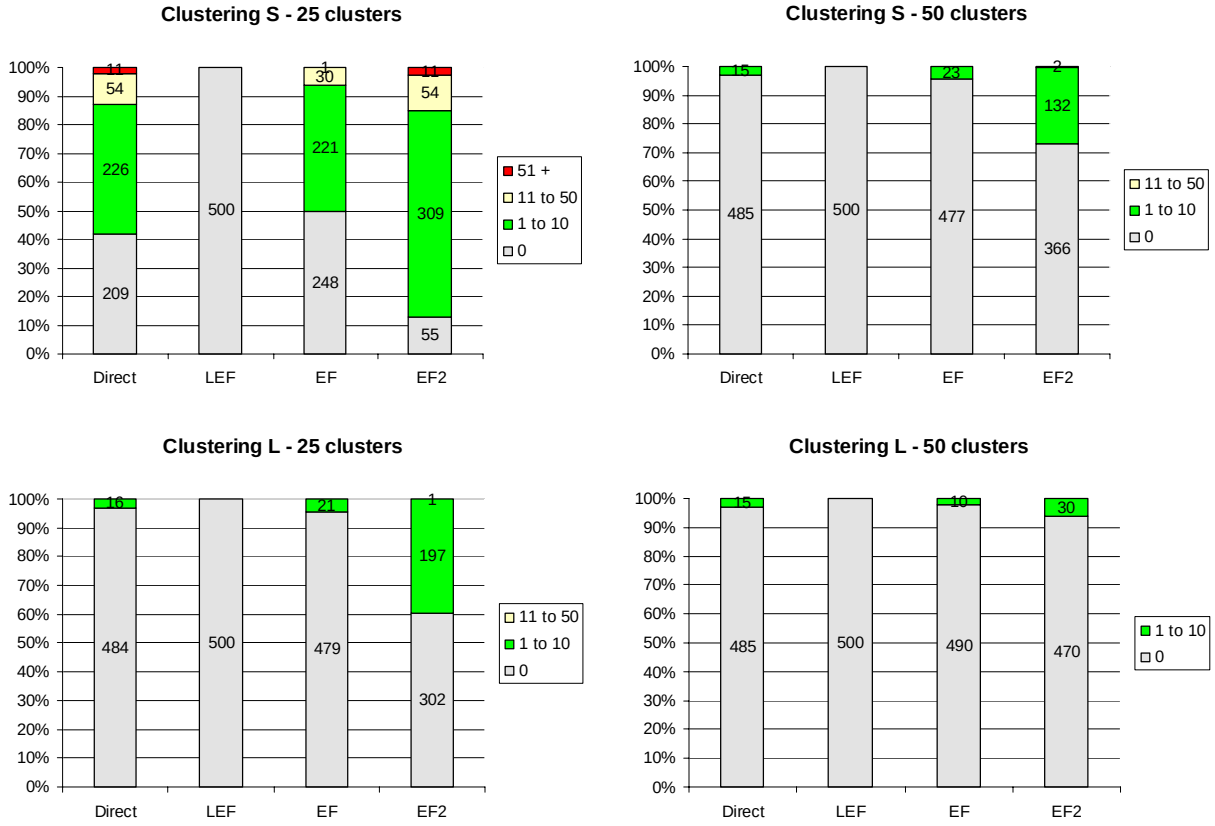

Chart 2. Relative bias of variance estimates

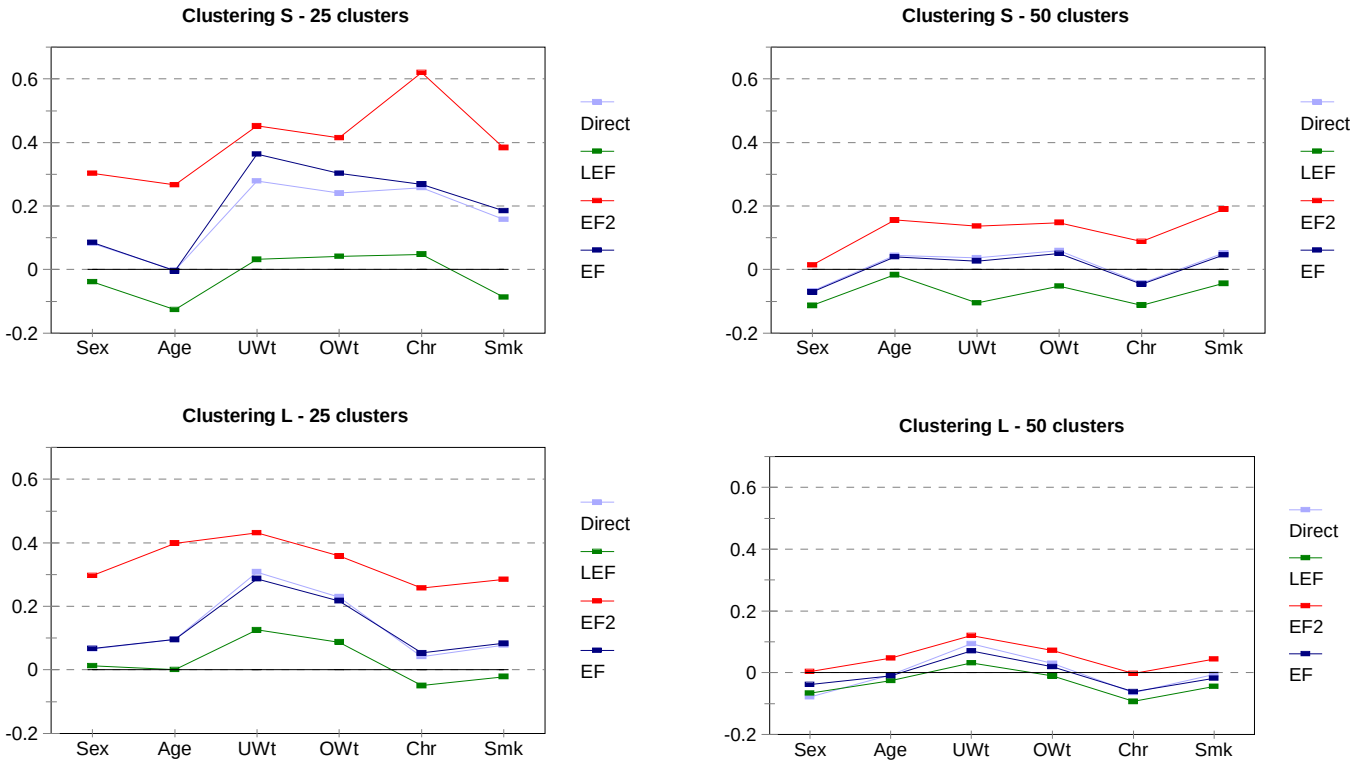

Chart 3. Relative mean square error of variance estimates

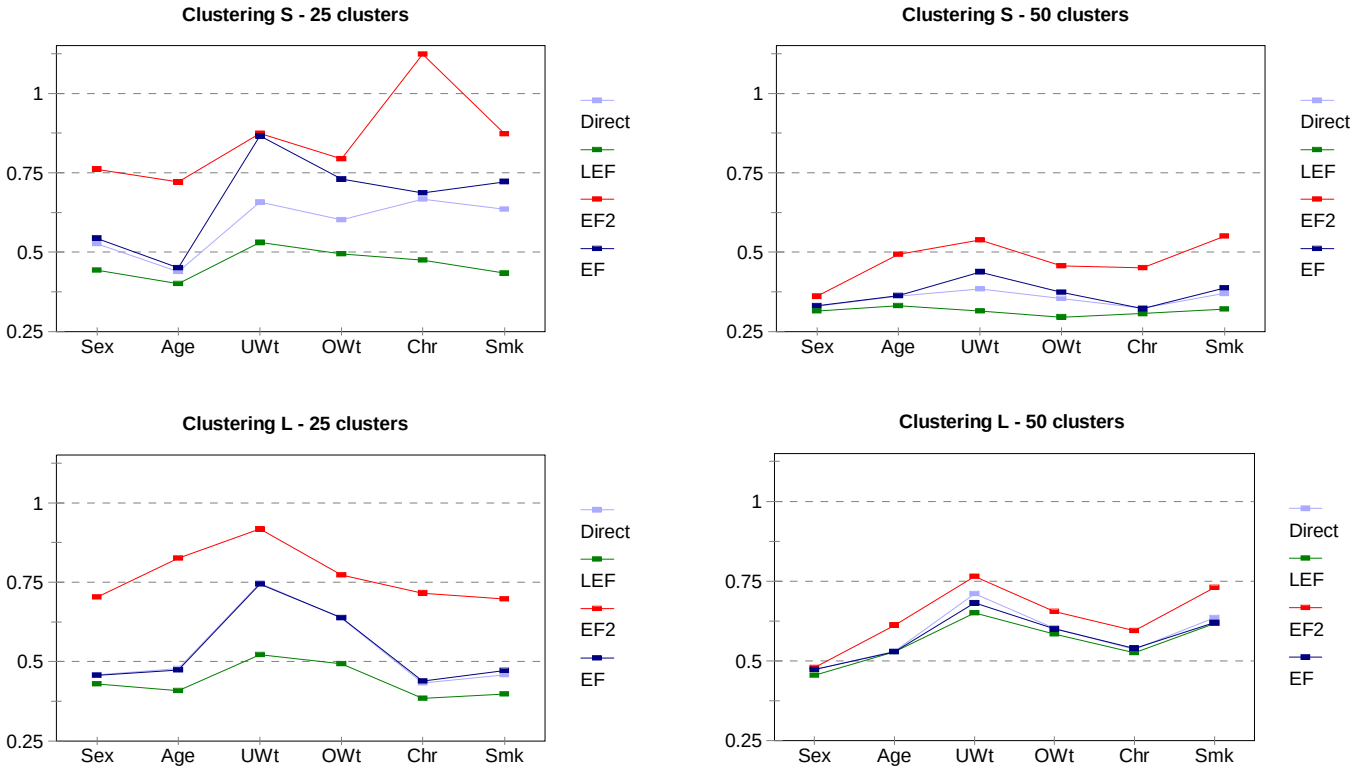

Chart 4. Coverage rate - 95% nominal confidence

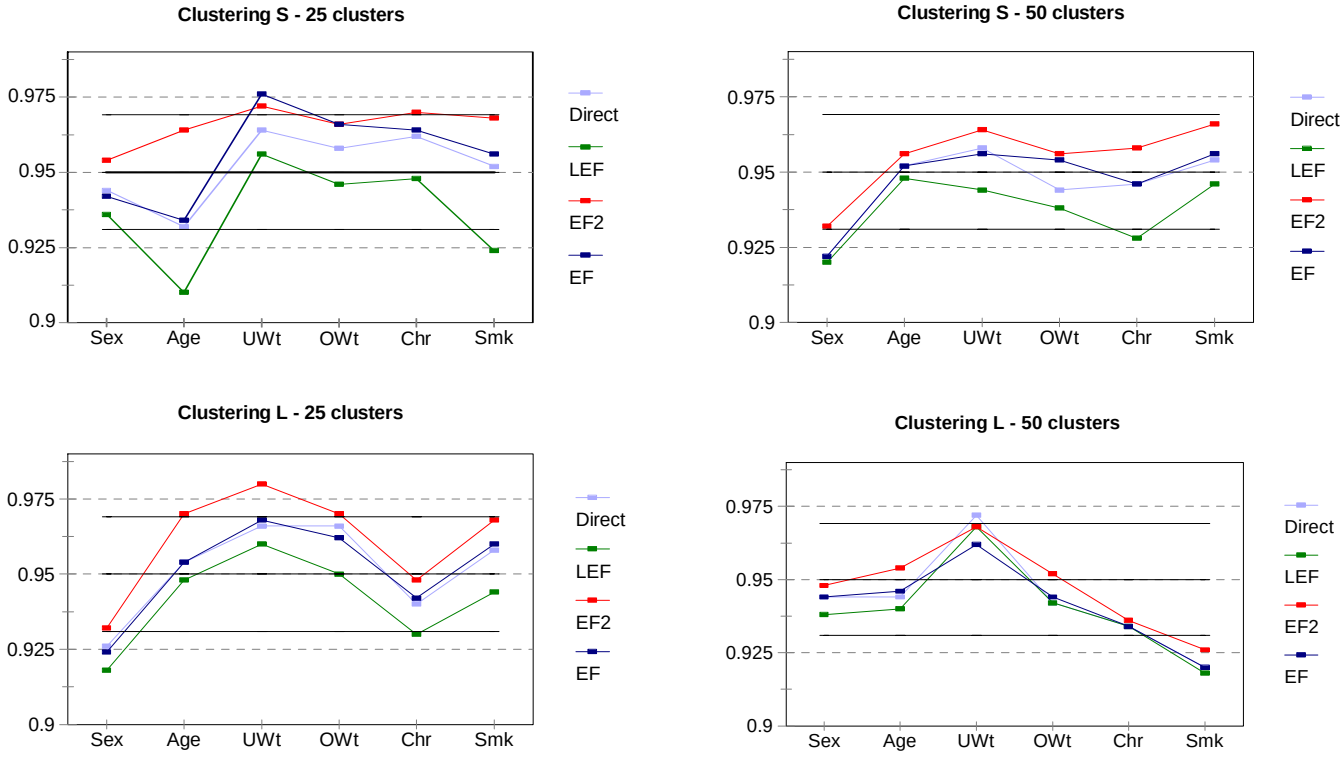

Chart 5. Coverage rate - 90% nominal confidence

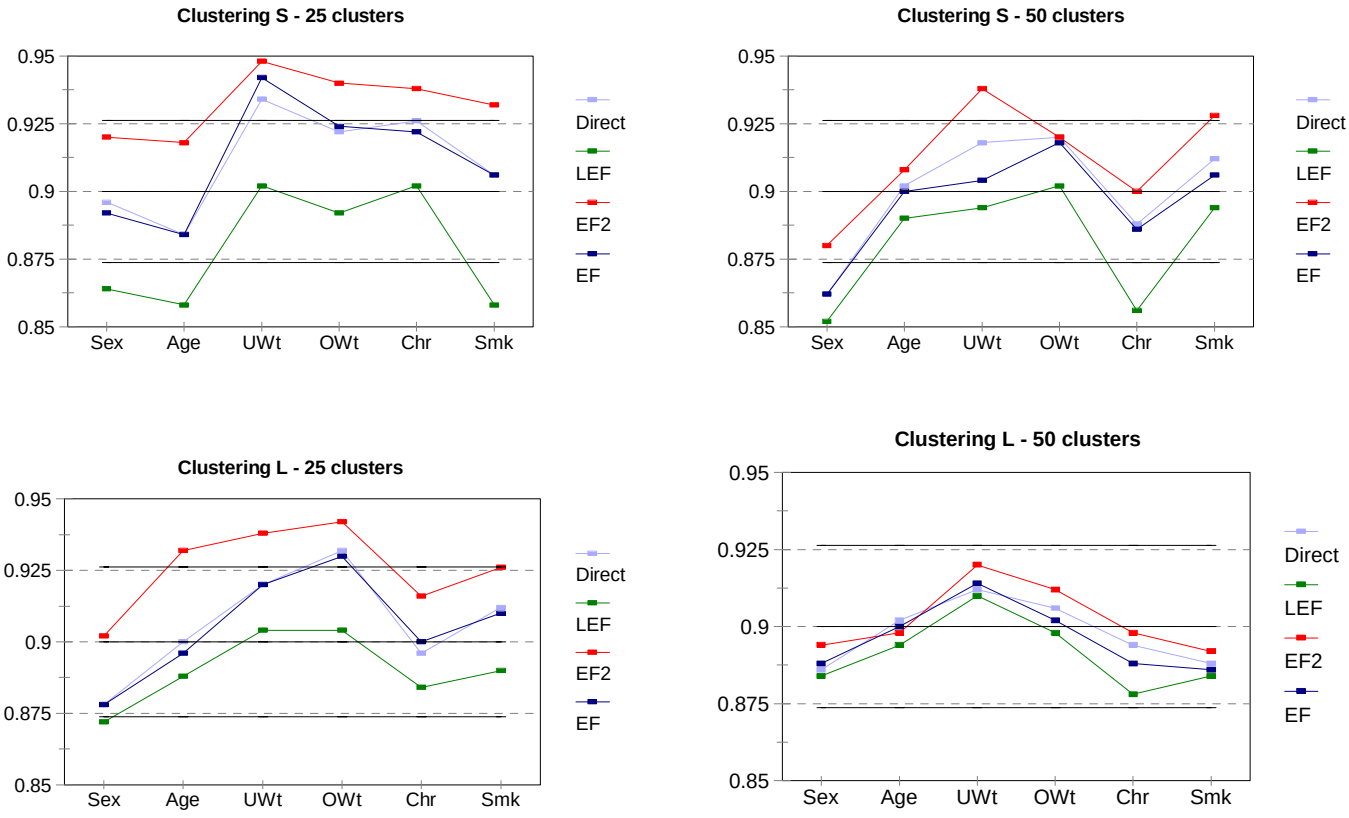

Chart 6: A) First moments: Value of Q1

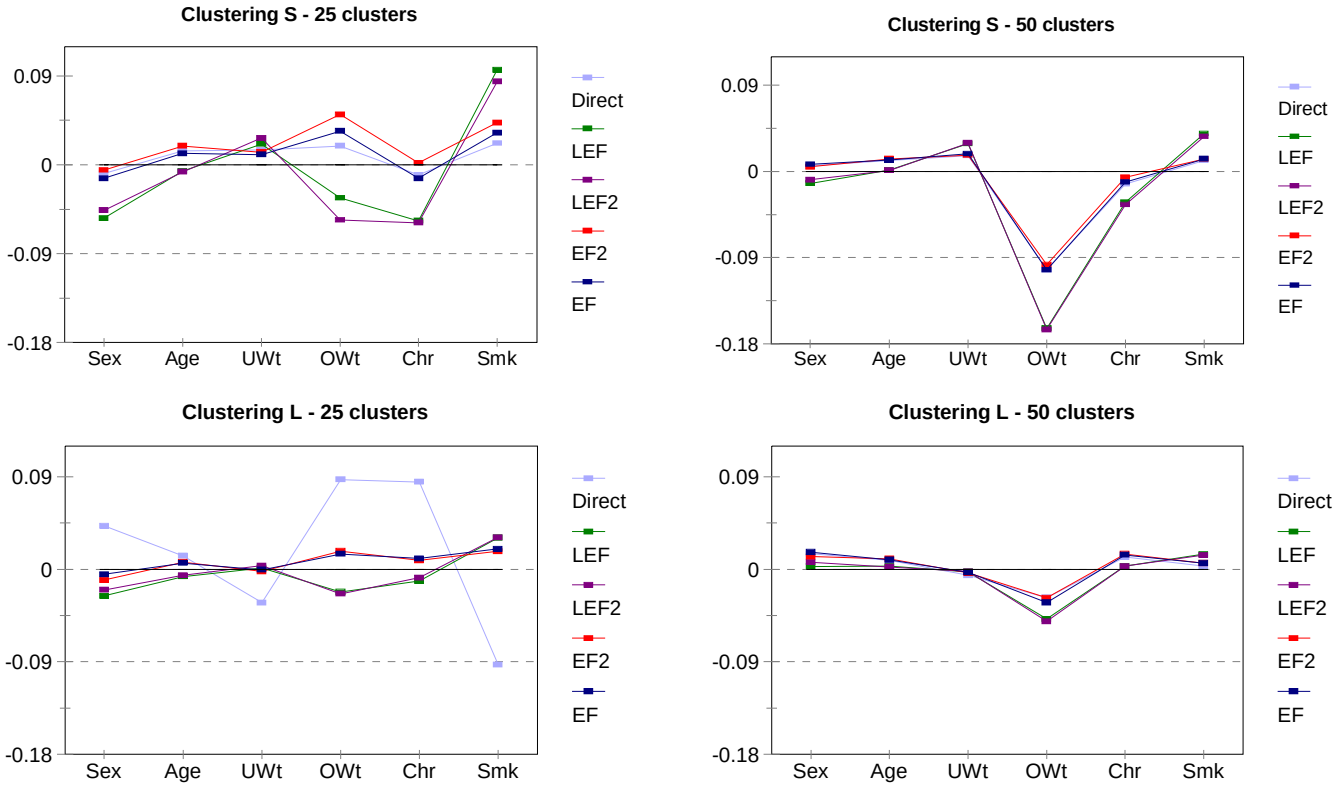

Chart 6: B) Second moments: Value of Q2

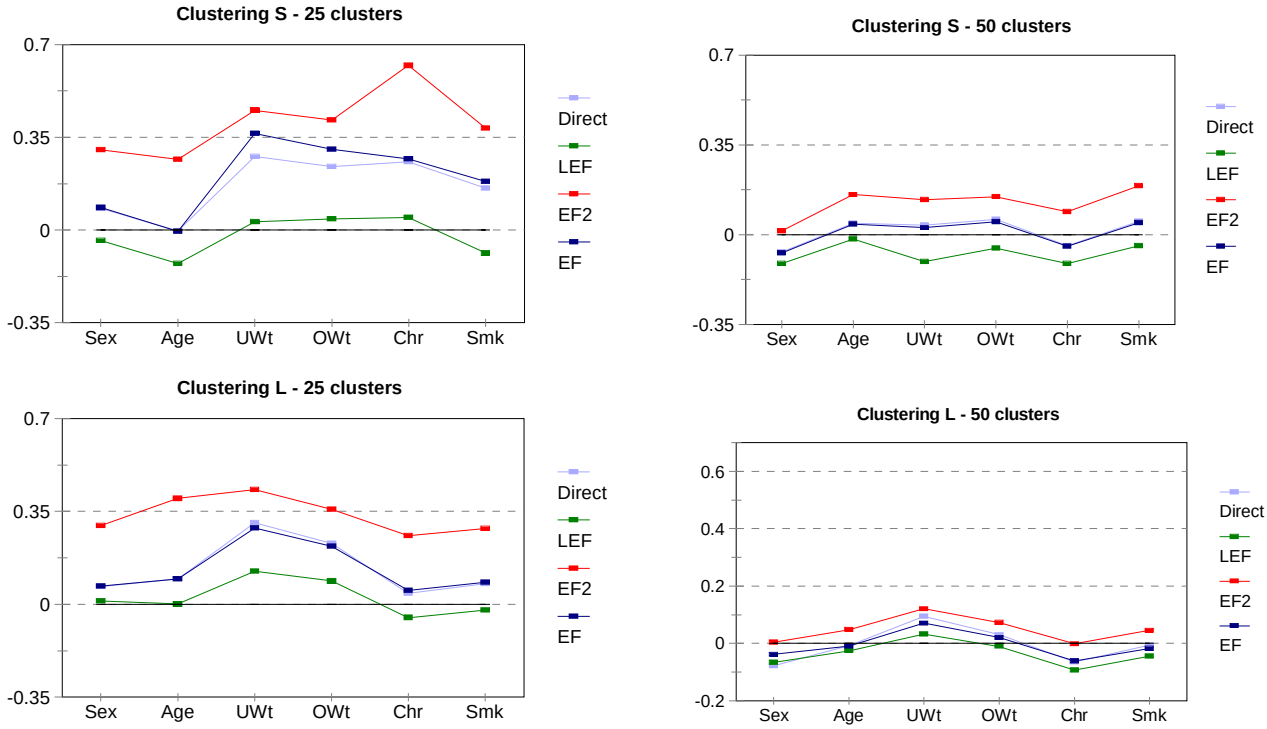

Chart 6: C) Third moments (skewness): Value of Q3

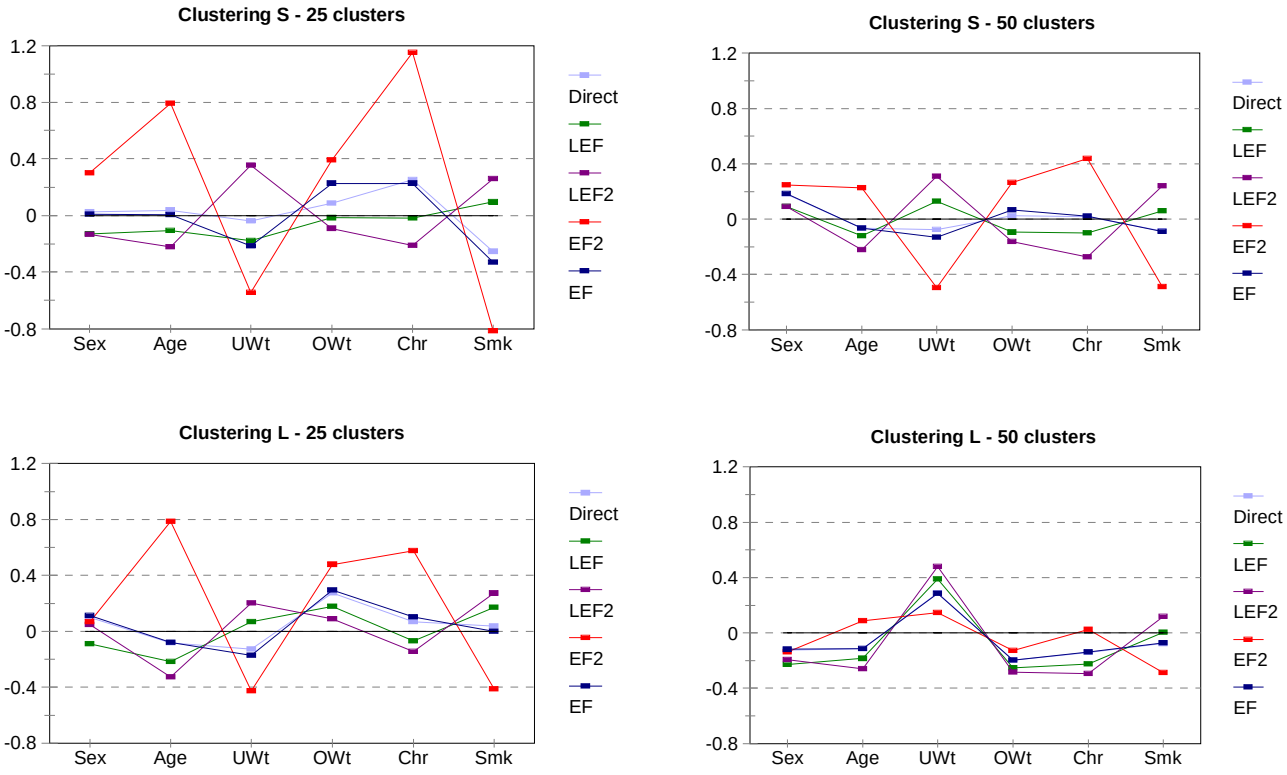

Chart 6: D) Fourth moments (kurtosis): Value of Q4