WP. 19
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (ii): Implementing editing strategies and links to other parts of processing

PERFORMANCE OF RESAMPLING VARIANCE ESTIMATION TECHNIQUES WITH IMPUTED SURVEY DATA

Supporting Paper

Submitted by the National Institute of Statistics (INE), Spain[1]

I. INTRODUCTION

1. In this paper we study the performance of jackknife and bootstrap techniques when estimating the variance and the true coverage of a confidence interval of a survey estimator under imputation. This performance is measured through the Monte Carlo bias and mean square error of the variance estimator and the coverage rate of the confidence interval.

2 The imputation methods used are ratio and mean imputation. The variable under study is turnover.

3 We work with artificial populations taken from large samples used in one structural business survey and one short-term business survey. In the first case, we draw a random sample from the business population. In the second case, we draw a stratified random sample from the business population.

4 We consider the jackknife variance estimation based on adjusted imputed values proposed by Rao and Shao (1992. This takes into account the fact that some data are imputed values. If a respondent is deleted, a re-imputation is done using the response set reduced by one unit.

5 In the case of imputation by the mean, we also use the bootstrap procedure proposed by Shao and Sitter (1996).

6 In the following section we describe the Monte Carlo study and present the results. Finally, we give some conclusion.

II. MONTE CARLO STUDY

7. Firstly, we consider as population the industrial businesses with 20 or more employees of the Structural Industrial Business Survey. We simulate 200,000 simple random samples from this population drawn without replacement of sizes n=100, 500, 1000 and 5000. The population size is the 16,438 businesses.

[1] Prepared by Felix Aparicio-Pérez and Dolores Lorca (fapape@ine.es; mdlorca@ine.es).

8. Secondly, a sample of 9,414 businesses from the retail trade Index Survey is taken as a sampling frame. Here, we also simulate 200,000 stratified random samples drawn from this population mimicking the true survey design of size n=800, 1500, 2200 and 3000.

9. The turnover is imputed using ratio and mean imputation. In the first case, the auxiliary variable is the total expenses in which has a 99% correlation with the turnover and the second case is the same month year ago turnover.

10. In the case of data from Structural Industrial Business Survey, a loss of about 30 per cent is simulated assuming that the response mechanism is uniform. In the case of data from Retail Trade Index Survey, missing data are generated following a distribution similar to the true missing value pattern observed in the survey. Missing data are also simulated for auxiliary variables used for ratio imputation, in order to study the effect of these covariates in the results. Accordingly, missing values in the auxiliary variables are simulated on a simple random sampling basis with 0, 10, 30 and 100 percent rates.

11. Within each replication, we compute the percentage relative bias, the relative mean square error and the coverage rate of the 95 percent confident interval based on normal approximation.

III. RESULTS

III.1 Structural Industrial Business Survey

TABLE 1. Imputed variable: turnover. Auxiliary variable: total expense

size	Ratio:	100%		90%			70%			0%		
	Mean:	0%		10%			30%			100%		
	RB(%)	MSE	COVR(%)	RB(%)	MSE	COVR(%)	RB(%)	MSE	COVR(%)	RB(%)	MSE	COVR(%)
100	0.72	4.28	59.2	0.68	4.23	58.8	0.14	4.22	58.6	0.29	4.52	56.9
500	1.22	1.84	69.9	0.32	1.82	69.4	0.38	1.82	68.9	-0.90	1.96	65.8
1000	0.36	1.26	76.7	-0.07	1.25	76.5	-0.50	1.25	75.9	-2.12	1.35	72.6
5000	-2.22	0.47	90.3	-2.41	0.47	90.3	-4.25	0.47	89.8	-12.0	0.51	87.3

TABLE 2. Imputed variable: turnover. Imputation method:mean imputation

Sample size	BOOTSTRAP		
	Percentile		
	RB(%)	MSE	COVR(%)
100	-0.42	4.47	60.6
500	1.00	2.00	68.3
1000	4.24	1.45	75.6
5000	27.15	0.76	91.9

III.2 Retail Trade Index Survey

TABLE 3. Imputed variable: turnover. Auxiliary variable: same month year ago turnover

Sample size	Ratio	100%		90%			70%			0%		
	Mean	0%		10%			30%			100%		
	RB(%)	MSE	COVR(%)	RB(%)	MSE	COVR(%)	RB(%)	MSE	COVR(%)	RB(%)	MSE	COVR(%)
800	-19.8	1.15	61.7	-19.7	1.14	61.9	-19.9	1.13	61.9	-19.6	1.13	62.0
1500	-0.63	0.89	77.1	0.17	0.88	77.5	0.03	0.88	77.7	0.06	0.88	77.7
2200	-1.21	0.79	81.5	-0.58	0.79	81.8	-0.03	0.79	82.0	-0.15	0.79	81.9
3000	-1.92	0.60	85.0	-0.75	0.61	86-0	0.15	0.61	85.5	0.31	0.61	85.6

IV. CONCLUSIONS

12. As expected, the performance of the jackknife variance estimator is better for larger sample sizes and for ratio imputation. For the case III.2, this latter is not appreciable but in additional simulations not shown in this paper, it can be seen that an increase in the rate of missing values leads to a better performance of ratio imputation.

13. The jackknife variance estimation performs poorly. This shows that strong skewness and kurtosys of imputed variable can influence considerably the results. Even when we consider imputation within survey design strata (case III.2) the performance does not improve. Only the increase of the sample size seems to be the way to improve the jackknife performance.

14. Regarding the table 2 we see that the percentile bootstrap performs better than the jackknife for coverage rate of the confidence intervals and the reverse is true for mean square errors and bias of the variance.

REFERENCES

Aparicio-Pérez, F. and Lorca, D., (2002). Performance of Jackknife Variance Estimation using several imputation methods. UNECE work session on Statistical Data Editing. Helsinki.

Aparicio-Pérez, F. and Lorca, D., (2003). Performance of Bootstrap Techniques with Imputed Survey Data. UNECE work session on Statistical Data Editing. Madrid.

Lorca, D and Aparicio-Pérez, F (2003). Variance Estimation under imputation in short-term business surveys. 54 th Session of the International Statistical Institute (ISI). Berlin.

Rancourt, E., Lee, H. and Särndal, C. (1994). Variance estimation under more than one imputation method. International Conference on establishment Surveys,pp 374-379.

Rao, J.N.K and Shao, J (1992). Jackknife variance estimation with survey data under hot deck imputation. Biometrika, 79, pp 811-822.

Saigo, Shao and Sitter (2001). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed Data. Survey Methodology, vol 27,No. 2,pp.189-196.

Shao and Sitter (1996) Bootstrap for imputed survey data. Journal of American Statistical Association, 91,1278-1288.
