Methodol Comput Appl Probab (2009) 11:21–27
DOI 10.1007/s11009-008-9072-8

Bootstrap Algorithms for Risk Models with Auxiliary Variable and Complex Samples

Giancarlo Manzi · Fulvia Mecatti

Received: 13 July 2007 / Revised: 26 November 2007 /
Accepted: 3 March 2008 / Published online: 19 April 2008

Abstract Resampling methods are often invoked in risk modelling when the *stability* of estimators of model parameters has to be assessed. The accuracy of variance estimates is crucial since the operational risk management affects strategies, decisions and policies. However, auxiliary variables and the complexity of the sampling design are seldom taken into proper account in variance estimation. In this paper bootstrap algorithms for finite population sampling are proposed in presence of an auxiliary variable and of complex samples. Results from a simulation study exploring the empirical performance of some bootstrap algorithms are presented.

Keywords Probability-proportional-to-size sampling · Ratio model · Ratio estimator · Regression estimator · Resampling methods

PACS 02.50.Tt · 02.70.Rr

Mathematics Subject Classification (2000) 62F40

JEL Classification C15 · C42 · C81

1 Introduction

Statistical modelling of the operational risk in finance is receiving increasing attention (De Fontnouvelle et al. 2006; Nyström and Skoglund 2002; Risk Management Group 2003).

G. Manzi
Department of Economics, Business and Statistics, University of Milan, Milan, Italy
e-mail: giancarlo.manzi@unimi.it

F. Mecatti (✉)
Department of Statistics, University of Milan–Bicocca, Milan, Italy
e-mail: fulvia.mecatti@unimib.it

The widely discussed case of the re-nationalisation of the British railways company *Rail-Track* in 2002 and some recent financial scandals are emblematic examples.

The operational risk management affects strategies, decisions and policies (Cruz 2002) and therefore the *stability* of estimators of model parameters states a judicious inferential issue. However the estimation framework is usually cumbersome to handle. Data are often collected under complex non-*iid* sampling plans or are produced under uncontrolled random mechanisms. As a consequence, the classical *iid* assumption as well as the sampling design with *equal* inclusion probability is routinely violated. In addition the estimators have generally a complex structure so that the analytical assessment of their accuracy is difficult. In this context, resampling methods for estimating the variance of estimators and for constructing confidence intervals appear as a natural solution.

Since Efron's original bootstrap applies to the classical *iid* framework, suitable modifications are required in order to address complex issues from survey sampling.

In the present paper bootstrap algorithms for finite population sampling are considered in presence of complex sample data and of an auxiliary variable. The latter could summarize past experience and/or known information related to the random mechanism providing the sample data.

In operational risk literature non-*iid* bootstrap techniques are seldom concerned. Efron's classical bootstrap is usually applied despite the complexity of the sampling design, especially when dealing with extreme value inference or when the sample size is small (Coleman 2003). As a starting point and with the main aim of providing some empirical evidence as a guidance for future research, in this paper we focus on a very simple ratio model, i.e. a linear regression for the N population values without intercept and with heteroscedastic errors

$$y_i = \beta x_i + \varepsilon_i \tag{1}$$

$i = 1 \cdots N$, where the ε_i are independent random variables with $E(\varepsilon_i) = 0$ and $E\left(\varepsilon_i^2\right) = \sigma_i^2$.

We assume that a sample s of size n is selected under a sampling design with probability proportional to the auxiliary variable x and without replacement (often referred as πPS sampling). Note that this allows a wider use of the auxiliary variable, which is involved at both the selection and the estimation stages. In addition *unequal* inclusion probabilities at the selection stage are acknowledged.

Three πPS bootstrap algorithms are presented in Section 2. In Section 3 a simulation study is illustrated and empirical results are presented. A discussion of simulation results is given in Section 4.

2 Bootstrap Algorithms

Two classic estimators for the regression coefficient β in model 1 are considered: the ratio estimator

$$\widehat{\beta}_{\mathrm{RA}} = \sum\nolimits_{i\in s} y_i \Big/ \sum\nolimits_{i\in s} x_i \tag{2}$$

and the regression estimator

$$\widehat{\beta}_{\mathrm{RE}} = \sum\nolimits_{i\in s} y_i x_i \Big/ \sum\nolimits_{i\in s} x_i^2. \tag{3}$$

In order to estimate the estimator's variance $V\left(\widehat{\beta}\right)$, the so-called *wild bootstrap* is generally proposed under simple random sampling (Helmers and Wegkamp 1998).

Under πPS sampling we consider three bootstrap algorithms, all of them relying upon the notion of *bootstrap population*. We refer to a bootstrap population as a set formed by replicating w_i times every sampled unit $i \in s$. Hence the bootstrap population includes data from the original sample only according to a basic bootstrap principle. Furthermore:

- the bootstrap population has (random) size $N^* = \sum_{i \in s} w_i = \sum_{i=1}^{N} w_i \delta_i$ where δ_i denotes the (original) sample membership indicator of every population unit i, i.e. a random variable taking value 1 if $i \in s$ and 0 otherwise. Let π_i be the inclusion probability of unit i into the original sample s. Hence $E(\delta_i) = \pi_i$;
- the total X^* of the auxiliary variable into the bootstrap population is given by $\sum_{i \in s} w_i x_i$.

Weights w_i are evaluated by calibrating with respect to known features of the actual population generating the data, namely the size N and the total $X = \sum_{i=1}^{N} x_i$ of the auxiliary variable. Once the bootstrap population is arranged, the general form of the bootstrap algorithm consists of the following three steps:

1. *Resampling step*. A collection of B bootstrap samples s^*, each of the same size n as the original sample s, is selected from the bootstrap population under a given re-sampling design, with B chosen sufficiently large.
2. *Bootstrap distribution step*. Let $\widehat{\beta}^*$ be the *replication* of the estimate $\widehat{\beta}$ computed by applying either Eq. 2 or 3 to a bootstrap sample s^*. The collection of the B replications provides the bootstrap distribution of estimator $\widehat{\beta}$.
3. *Variance estimation step*. The variance $V^*\left(\widehat{\beta}^*\right)$ of the bootstrap distribution supplies the bootstrap estimate for the estimator variance $V\left(\widehat{\beta}\right)$.

The three bootstrap algorithms considered here differ either in the choice of w_i or in the re-sampling design which can or can not *mimic* the original πPS design. We start with the

1. *Holmberg's bootstrap* (Holmberg 1998), where the bootstrap population is given by setting $w_i = \pi_i^{-1}$, i.e. the inverse of the inclusion probability in the original sample. Assuming that for every population unit i the auxiliary variable takes on positive value and that the sampling design ensures $\pi_i = n\pi_i/X$, the resulting bootstrap population can be viewed as calibrated with respect to both X (uniformly) and N (on average). Hence we have: $X^* = X$ and $E(N^*) = N$. The re-sampling step is performed under the same πPS design yielding the original sample s.

In addition we propose:

2. a *simplified Holmberg's bootstrap* based upon the same bootstrap population as for the original Holmberg's algorithm and upon a *simple random* re-sampling *with equal probabilities* according to Mecatti (2000). Hence the bootstrap population is calibrated to both X and N whereas the bootstrap principle (according to which the re-sampling design should *mimic* the original sampling design) is unattended. On the other hand the simplification in the resampling step allows noticeable computational advantages;
3. a *model assisted πPS bootstrap* by assuming $\varepsilon_i = x_i^{1/2} u_i$ in model 1 and by using a suitable estimator $\widehat{F}(\cdot)$ of the original finite population distribution according to Rao, Kovar and Mantel (1990). The bootstrap population follows by setting:

$$w_i = N\left[\widehat{F}(y_i) - \widehat{F}(y_{i-1})\right] \text{ for all } i \in s. \tag{4}$$

The resampling step is still performed by mimicking the original πPS design.

Since the efficiency of a πPS sampling over a simple random sampling increases as the relationship between the study variable y and the auxiliary variable x approaches proportionality, the ratio model 1 is implicitly assumed when a πPS sampling design is chosen. Furthermore, estimator $\widehat{F}$ is proved to be both asymptotically *design*-unbiased and *model*-unbiased (Rao, Kovar and Mantel, 1990). As a consequence the proposed model assisted πPS bootstrap is expected to retain the design based properties from the Holmberg's bootstrap 1 and also to be calibrated with respect to the model assumption $y \overset{a}{\propto} x$ (where $\overset{a}{\propto}$ denotes approximated proportionality).

3 Empirical Results

A simulation study has been conducted with the twofold purpose of investigating the empirical performance of the three πPS bootstrap algorithms presented in Section 2 and of comparing them with the wild bootstrap, i.e. the methodology usually suggested in the literature.

The simulation considers an artificial set up according to McCarthy and Snowden (1985) where estimators $\widehat{\beta}_{\mathrm{RA}}$ and $\widehat{\beta}_{\mathrm{RE}}$ are expected to have optimal inferential properties. Two artificial populations of size N = 100 have been produced by giving x a Chi Square distribution with 9 and with 18 degrees of freedom. This leads to an asymmetrical and an almost symmetrical population with increasing variability. Model errors ε_i have been generated by a Normal distribution with zero mean and variance equals x_i. The Monte Carlo distribution of estimators $\widehat{\beta}_{\mathrm{RA}}$ and $\widehat{\beta}_{\mathrm{RE}}$ are provided by 10,000 simulation runs under the Rao–Sampford rejective πPS design (Rao 1965; Sampford 1967) for increasing sample fraction $f = n/N$ equals 0.1, 0.2 and 0.25. Note that the Rao–Sampford design although ensuring $\pi_i = nx_i/X$ does not allow simulations with large sample sizes, namely f>0.25, for the number of sample rejections increases at an exponential rate as the sample fraction increases.

For every Monte Carlo run, bootstrap algorithms have been performed by setting B = 200. Empirical properties of the bootstrap algorithms and of the bootstrap variance estimators have been evaluated by computing the following relative measures:

the Average Relative Error (Lahiri 2003):

$$\mathrm{ARE} = 100 \times \frac{E_{\mathrm{mc}}\left[V_{*}\left(\widehat{\beta}^{*}\right)\right] - V_{\mathrm{mc}}\left(\widehat{\beta}\right)}{V_{\mathrm{mc}}\left(\widehat{\beta}\right)} \tag{5}$$

the Relative Mean Square Error:

$$\mathrm{RMSE} = 100 \times \frac{E_{\mathrm{mc}}\left[V_{*}\left(\widehat{\beta}^{*}\right) - V_{\mathrm{mc}}\left(\widehat{\beta}\right)\right]^{2}}{V_{\mathrm{mc}}\left(\widehat{\beta}\right)} \tag{6}$$

where $\widehat{\beta}$ denotes estimators from the original sample, * denotes bootstrap quantities, expectations and variances are provided via Monte Carlo.

Simulation results concerning bootstrap estimates of the variance of the ratio estimator $\widehat{\beta}_{\mathrm{RA}}$ are shown in graphs arranged in Fig. 1. Graphs in Fig. 2 display simulation results regarding bootstrap estimates of the variance of the regression estimator $\widehat{\beta}_{\mathrm{RE}}$. Relative measures 5 and 6 concerning πPS bootstrap algorithms 1, 2 and 3 (Section 2) and the wild bootstrap, are plotted against the three simulated levels of sample fraction and paired for the two simulated populations (a and c, b and d).

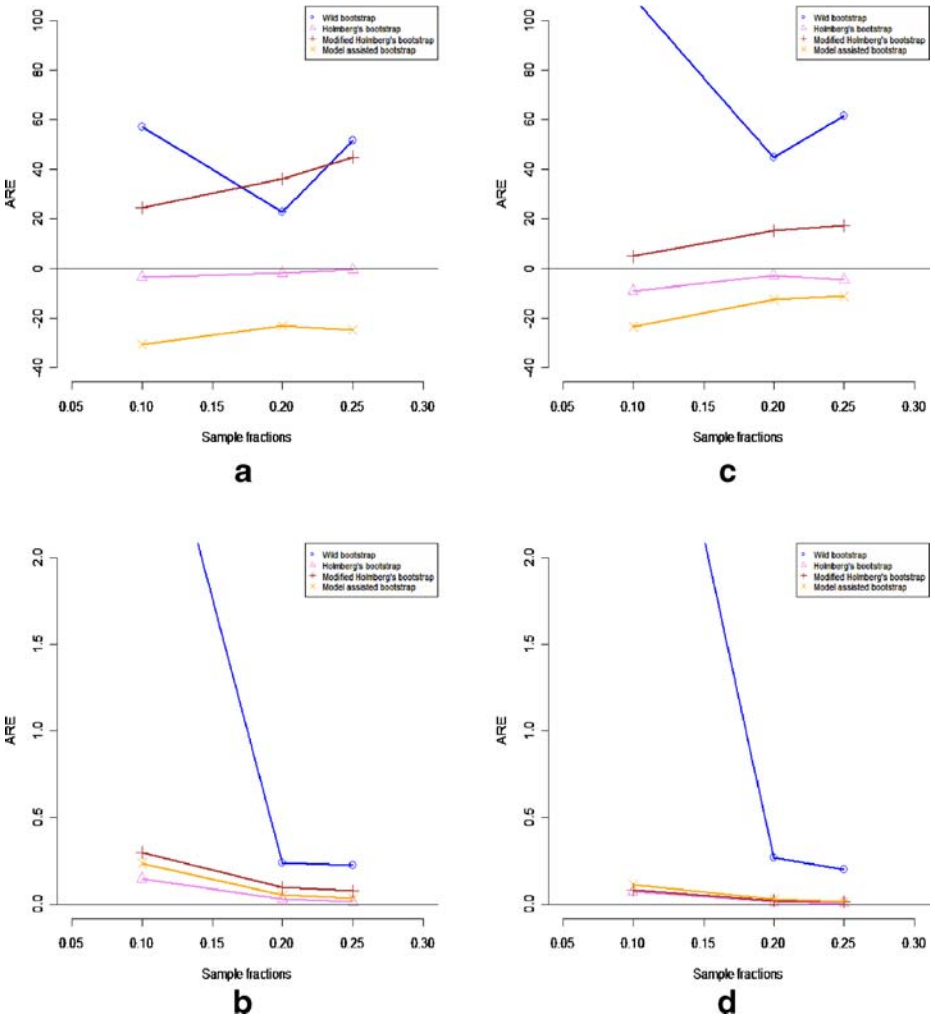

Fig. 1 ARE and RMSE of bootstrap estimates of $V\left(\widehat{\beta}_{\mathrm{RA}}\right)$ for increasing sample fraction and for two simulated populations. **a** ARE for population: $x \approx \chi^2_9$, $y_i = x_i + N(0, x_i)$. **b** RMSE for population: $x \approx \chi^2_9$, $y_i = x_i + N(0, x_i)$. **c** ARE for population: $x \approx \chi^2_{18}$, $y_i = x_i + N(0, x_i)$. **d** RMSE for population: $x \approx \chi^2_{18}$, $y_i = x_i + N(0, x_i)$

4 Discussion

For the simple ratio model 1, simulation results highlight that the three proposed πPS bootstrap algorithms represent valid alternatives to the wild bootstrap when the auxiliary variable is used at both the estimation and at the selection stage. They also encourage further research by considering more complex models and estimators suitable for operational risk as well as real data set.

The line referring to the wild bootstrap is located distinctly above the others in all the graphs except for two ambiguous cases limited to the simplified Holberg's bootstrap. As a consequence the wild bootstrap, as compared to the proposed πPS bootstrap algorithms, shows the worst performance in terms of biasedness (ARE) as well as of efficiency and

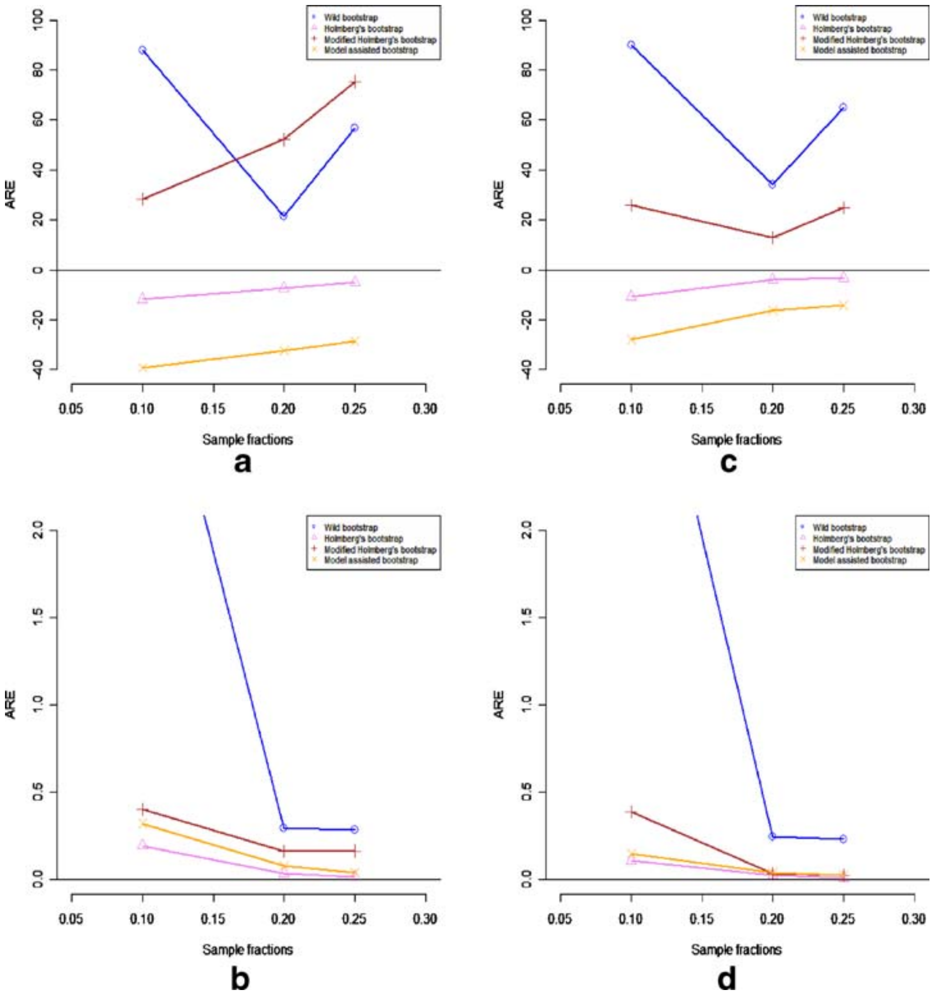

Fig. 2 ARE and RMSE of bootstrap estimates of $V\left(\widehat{\beta}_{\mathrm{RE}}\right)$ for increasing sample fraction and for two simulated populations. **a** ARE for population: $x \approx \chi_9^2$, $y_i = x_i + N(0, x_i)$. **b** RMSE for population: $x \approx \chi_9^2$, $y_i = x_i + N(0, x_i)$. **c** ARE for population: $x \approx \chi_{18}^2$, $y_i = x_i + N(0, x_i)$. **d** RMSE for population: $x \approx \chi_{18}^2$, $y_i = x_i + N(0, x_i)$

stability (RMSE) of the variance estimators. By comparing the three lines referring to the proposed πPS bootstrap algorithms, the Holmberg's bootstrap 1 emerges as the best performer. It uniformly provides variance estimates nearly unbiased, more efficient and more stable than all the competitors, yet for low sample size.

The simplification proposed with the modified Holmberg's bootstrap 2 leads to substantial computational advantages but results in significant biases. On the other hand, the simplification at the resampling step does not convey noticeable stability losses. Moreover the simplified Holmberg bootstrap supplies conservative variance estimates while both the πPS competitors tend to underestimate.

Finally, the proposed model assisted πPS bootstrap algorithm 3 shows an intermediate performance. As a result that does not seem to balance the major complexity and the heavier computational needs than both the πPS competitors. To this respect empirical evidence suggests further research, especially with more complex models and with real data.

As an overall evaluation, differences in shape and variability of the simulated populations appear not to affect the pattern of the relative behaviour of the compared bootstrap solutions. Finally, results from the three πPS algorithms tend to approach as the sample size increases.

References

Coleman R (2003) Op risk modelling for extremes. Part 2: statistical methods. Oper Risk 4(1):6–9

Cruz MG (2002) Modelling, measuring and hedging operational risk. Wiley, New York

De Fontnouvelle P, DeJesus-Rueff V, Jordan J, Rosengren E (2006) Capital and risk: new evidence on implications of large operational losses. J Money Credit Bank 38(7):1819–1846

Helmers R, Wegkamp M (1998) Wild bootstrapping in finite populations with auxiliary information. Scand J Statist 25(2):383–399

Holmberg A (1998) A bootstrap approach to probability proportional to size sampling. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp 378–383

Lahiri D (2003) On the impact of bootstrap in survey sampling and small-area estimation. Stat Sci 18:199–210

McCarthy PJ, Snowden CB (1985) The bootstrap and finite population sampling. Vital and health statistics, series 2, volume 95. Public Health Service Publication, U.S. Government Printing Office, Washington, DC, pp 1–23

Mecatti F (2000) Bootstrapping unequal probability samples. Stat Appl 12(1):67–77

Nyström K, Skoglund J (2002) Quantitative operational risk management. Swedbank Working Paper

Rao JNK (1965) On two simple schemes of unequal probability sampling without replacement. J Indian Stat Assoc 3:173–180

Rao JNK, Kovar JG, Mantel HJ (1990) On estimating distribution function and quantiles from survey data using auxiliary information. Biometrika 77:365–375

Risk Management Group (2003) The 2002 loss data collection exercise for operational risk: summary of the data collected. Report to the Basel Committee on Banking Supervision, Bank for International Settlement. Available at http://www.bis.org/bcbs/qis/ldce2002.pdf

Sampford MR (1967) On sampling without replacement with unequal probabilities of selection. Biometrika 54:499–513