Computational Statistics and Data Analysis 59 (2013) 70–81

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bootstrapping for highly unbalanced clustered data

Mayukh Samanta [*,1], A.H. Welsh [2]

The Australian National University, Canberra, Australia

ARTICLE INFO

Article history:
Received 8 August 2011
Received in revised form 4 July 2012
Accepted 5 September 2012
Available online 12 September 2012

Keywords:
Bootstrap
Clustered data
Fast and robust bootstrap
Quasi-likelihood estimation
Robust estimation
Unbalanced data
Variance components

ABSTRACT

We apply the generalized cluster bootstrap to both Gaussian quasi-likelihood and robust estimates in the context of highly unbalanced clustered data. We compare it with the transformation bootstrap where the data are generated by the random effect and transformation models and all the random variables have different distributions. We also develop a fast approach (proposed by Salibian-Barrera et al. (2008)) and show that it produces some encouraging results. We show that the generalized bootstrap performs better than the transformation bootstrap for highly unbalanced clustered data. We apply the generalized cluster bootstrap to a sample of income data for Australian workers.

1. Introduction

Consider independent observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_g$, from g clusters, where each $\boldsymbol{y}_i$ is an m_i-vector which is linearly related to a known $m_i \times p$ design matrix of explanatory variables $\boldsymbol{X}_i$, $i = 1, \ldots, g$, and is subject to $c > 1$ sources of variation. There are three different situations in this context: (1) balanced case: $m_i = m$, all clusters have the same number of observations; (2) mildly unbalanced case: $m_i \neq m_{i'}$ for $i \neq i'$ but cluster size does not vary much; (3) highly unbalanced case: the m_i have a very large variance e.g., one cluster has 3 observations, one cluster has 5 observations and another cluster has 100 or even more observations. This paper will focus on bootstrapping in the third, highly unbalanced case.

Motivation for considering highly unbalanced clustered data is provided by the Australian Workplace Industrial Relations Survey (AWIRS) described in Section 5. Observations are made on employees clustered within workplaces which vary in size from $m_i = 1$ to $m_i = 75$ employees. To model and make inferences based on these data, we need methods that can handle highly unbalanced clusters. Recent literature has considered various aspects of bootstrapping in linear mixed effect models with an emphasis on the variance components (Field and Welsh, 2007; Field et al., 2008, 2010). Field et al. (2008) estimated the sampling distribution of the mean and variance parameters of the linear mixed effect model using semiparametric bootstraps which make use of the structure of the model. They considered unbalanced clusters and data with multiple levels of random effects, and established that these estimators have different distributions under transformation and random effect models (see (17) and (18) below) unless all the random variables have Gaussian distributions. Pang and Welsh (in press) and Field et al. (2010) applied the generalized cluster bootstrap approach of Chatterjee and Bose (2005) in which estimators defined by estimating equations are bootstrapped using a randomly weighted estimating equation. The generalized cluster

* Corresponding author.
E-mail addresses: m.samanta@uq.edu.au (M. Samanta), Alan.Welsh@anu.edu.au (A.H. Welsh).
1 Mayukh Samanta is a Postdoctoral Research Fellow.
2 A.H. Welsh is E.J. Hannan Professor of Statistics in Centre for Mathematics and its Applications, The Australian National University, Canberra ACT 0200, Australia.

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.09.004

bootstrap and the transformation bootstrap (Field and Welsh, 2007) seem to perform best of the bootstraps considered. Theoretically, the generalized cluster bootstrap is more widely applicable than the transformation bootstrap which applies only under the transformation model. On the other hand, the generalized cluster bootstrap performs well for mildly unbalanced data but, unlike the transformation bootstrap, is not specifically designed to handle highly unbalanced data. Motivated by these considerations, we have explored the generalized cluster and transformation bootstraps in the context of extremely unbalanced clustered data. We applied both the generalized cluster and transformation bootstraps to the Australian workplace survey data with highly unbalanced cluster size. To our surprise, the generalized cluster bootstrap continues to perform well, even under extreme imbalance, and therefore seems to be the bootstrap method of choice.

We initially applied the bootstraps with fully iterated estimators (discussed in detail in Section 3) to solve the estimating equations, but this is computationally intensive, especially with robust estimates. Hence we also considered the fast and robust bootstrap (FRB) developed by Salibian-Barrera et al. (2008). They described and applied this method in linear regression and multivariate location scatter models and we have applied it to the random effect model. In a simulation study, for 100 samples of 18 clusters, the time taken by the bootstrap with full iteration for the estimation of the parameters is about 8 h whereas FRB takes about one hour. Hence, a substantial amount of time is saved using FRB. Although this is an approximate approach for estimating the parameters, the results are very encouraging compared to the fully iterated approach. We show in simulation experiments under both random effect and transformation models that the fast generalized cluster bootstrap is more attractive than the transformation bootstrap for highly unbalanced clustered data.

We begin this paper with formal definitions of the bootstraps and estimating equations in Section 2. We discuss the fully iterated and fast bootstrap implementation strategies in Section 3. We then present simulation experiments and results in Section 4 and an application to a set of survey data in Section 5. We end the paper with discussion and concluding remarks in Section 6.

2. Bootstraps and estimators

Consider $\boldsymbol{y}_i|\boldsymbol{X}_i \sim$ independent $(\boldsymbol{X}_i\boldsymbol{\mu}, \boldsymbol{V}_i)$, where $\boldsymbol{\mu}$ is an unknown regression parameter and $\boldsymbol{V}_i^{1/2}$ is a $m_i \times m_i$ symmetric, positive definite dispersion matrix which is a function of unknown, non-negative scale parameters, $\sigma_1, \ldots, \sigma_c$. Let $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_p, \sigma_1^2, \ldots, \sigma_c^2)^T$ denote the vector of all $s = p + c$ unknown parameters. For estimation of $\boldsymbol{\theta}$ we consider the general class of estimators defined by the estimating equations

$$g^{-1/2} \sum_{i=1}^{g} \boldsymbol{\psi}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\theta}) = \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{\psi}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\theta}) = (\boldsymbol{\psi}_1(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\theta})^T, \boldsymbol{\psi}_2(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\theta})^T)^T$, $\boldsymbol{\psi}_1$ is a p-vector of estimating functions corresponding to the regression parameter $\boldsymbol{\mu}$, and $\boldsymbol{\psi}_2$ is a c-vector of estimating functions corresponding to $\sigma_1^2, \ldots, \sigma_c^2$.

We consider two bootstraps which performed best in Field et al. (2008, 2010).

Generalized cluster bootstrap. For a row-wise exchangeable triangular array of random weights $\{w_{gi}\}$, the generalized cluster bootstrap estimator $\hat{\boldsymbol{\theta}}^*$ (Chatterjee and Bose, 2005; Pang and Welsh, in press) is a solution to the estimating equation

$$g^{-1/2} \sum_{i=1}^{g} \boldsymbol{w}_{gi} \boldsymbol{\psi}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\theta}) = 0. \tag{2}$$

As emphasized by Pang and Welsh (in press), it is important that the mean and variance of the weights are asymptotically equal to one but otherwise a wide variety of weights can be considered; the choice $\boldsymbol{w}_g = (w_{g1}, \ldots, w_{gg})^T \sim$ multinomial$(g, \boldsymbol{1}_g/g)$ produces the cluster bootstrap discussed in Field and Welsh (2007) while w_{gi} distributed as independent standard exponential random variables produces another bootstrap.

Transformation bootstrap. Given estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{V}}$ where $\hat{\boldsymbol{V}}_i$ is $\boldsymbol{V}_i$ with the unknown parameters replaced by estimates, let $\tilde{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{V}}_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}})$. Then sample independently with replacement from the elements of $\tilde{\boldsymbol{\epsilon}}^T = (\tilde{\boldsymbol{\epsilon}}_1^T, \ldots, \tilde{\boldsymbol{\epsilon}}_g^T)$ to produce $\tilde{\boldsymbol{\epsilon}}_1^*, \ldots, \tilde{\boldsymbol{\epsilon}}_g^*$, and construct bootstrap observations

$$\boldsymbol{y}_i^* = \boldsymbol{X}_i\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{V}}_i^{1/2}\tilde{\boldsymbol{\epsilon}}_i^*. \tag{3}$$

(If there is no intercept in the model, then we also need to center $\tilde{\boldsymbol{\epsilon}}$.)

The transformation bootstrap in Field and Welsh (2007) standardized the $\tilde{\boldsymbol{\epsilon}}$ to $\hat{\boldsymbol{\epsilon}} = \tilde{\boldsymbol{\epsilon}}/(\tilde{\boldsymbol{\epsilon}}^T\tilde{\boldsymbol{\epsilon}}/n)^{1/2}$ before resampling. Field et al. (2010) showed that applying that transformation bootstrap (with standardization) to robust estimators under contamination produces bootstrap estimates that are too small and that it is better to omit the standardization step.

Field et al. (2008, 2010) also considered the random effect bootstrap and showed that it does not perform well under either the random effect or transformation models, even in a mildly unbalanced situation, so we have not included it in our present study of highly unbalanced clustered data.