Appl. Statist. (2003)
52, *Part* 4, *pp.* 431–443

A novel bootstrap procedure for assessing the relationship between class size and achievement

James R. Carpenter

London School of Hygiene and Tropical Medicine, UK

and Harvey Goldstein and Jon Rasbash

Institute of Education, London, UK

[Received June 2000. Final revision April 2003]

Summary. There is on-going concern about the relationship between class size and achievement for children in their first years of schooling. The Institute of Education's class size project was set up to address this issue and began recruiting in the autumn of 1996. However, because of the non-normality of achievement measures, especially in mathematics, the results have hitherto been presented by using transformed achievement measures. This makes the interpretation difficult for non-statisticians. Ideally, the data would be modelled on the original scale and a bootstrap procedure used to ensure that inferences are robust to non-normality. However, the data are multilevel. In the paper we therefore propose a nonparametric residual bootstrap procedure that is suitable for multilevel models, show that it is consistent and present a simulation study which demonstrates its potential to yield substantial reductions in the difference between nominal and actual confidence interval coverage, compared with a parametric bootstrap, when the underlying distribution of the data is non-normal. We then apply our approach to estimate the relationship between class size and achievement for children in their reception year, after adjusting for other possible determinants.

Keywords: Class size project; Confidence interval; Multilevel model; Nonparametric bootstrap

1. Introduction

There is on-going concern about the relationship between class size and achievement for children in their first years of schooling. In the UK, debate has focused on the negative effects of large classes, whereas in the USA it has centred on the efficacy and cost-effectiveness of class size reductions (Blatchford *et al.*, 2002). The Institute of Education's class size project was set up to address this issue and began recruiting in the autumn of 1996. Students were assessed in mathematics and literacy before entering the reception class and at the end of each year. In addition, as described in more detail in Section 2, information was collected on a variety of plausible determinants of achievement.

An analysis of the 1996 cohort is given by Blatchford *et al.* (2002). However, because of the non-normality of the measures of achievement, especially those relating to mathematics, a normalizing transformation was applied before the analyses were carried out. The results are therefore not readily interpretable by educationalists; although key predictors are identified, their effect on measures of achievement that are familiar to educationalists is not transparent.

Address for correspondence: James R. Carpenter, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK.
E-mail: James.Carpenter@lshtm.ac.uk

 0035–9254/03/52431

If we wish to carry out an analysis using the original achievement scale, we should be wary of the assumption of normally distributed residuals. The bootstrap provides a natural way to address this in regression (Davison and Hinkley (1997), chapter 6). However, the class size project is a multilevel data set, with children nested within classes within schools within educational authorities. There is no well-established nonparametric bootstrap procedure for such data sets.

To address this problem, we propose a residual bootstrap for multilevel models. We show that, under certain regularity conditions, this is consistent, and we present a simulation study which confirms that the method gives a substantial reduction in confidence interval coverage error compared with a parametric bootstrap, even when the residuals follow a χ^2_1-distribution at all levels.

Having established the properties of our method, we then use it to estimate confidence intervals for the relationship between class size and achievement for children in the reception class, after adjusting for plausible determinants.

The plan for the remainder of this paper is as follows. Section 2 gives more detail about the class size project. Section 3 outlines the nonparametric bootstrap, relates it to other suggestions and shows its consistency. A substantial simulation study is presented in Section 4. Section 5 describes the analysis of the data that are described in Section 2, and some conclusions are drawn in Section 6.

2. The data

As discussed in Section 1, the Institute of Education's class size project arose out of concern over the effect of class size differences on pupils' educational achievements. The project recruited its first cohort in 1996, and a second cohort in 1997. The present analysis uses only the 1996 cohort. Data were collected before children started in the reception class, and then throughout the whole of key stage 1 (i.e. the reception year, year 1 and year 2, corresponding to ages 4–7 years). The data are multilevel: for our analysis we used complete data, which are available on 4621 pupils in 254 classes in 157 schools.

Here we look at the effect of class size on achievement in mathematics in reception classes. On entering their reception class, the Avon reception entry assessment (Avon Education Department and Institute of Education, 1996) was used to measure each child's literacy and mathematics ability. This was measured again at the end of the reception year. Literacy was assessed by using the literacy base-line component of the reading progress test (Reading Progress Test, 2000), whereas mathematics ability was assessed by a teacher-administered test that was devised and piloted by Blatchford *et al.* (2002). Both tests aimed to cover the curriculum that is experienced by the children in their reception year.

In addition, pupils' background details including age, sex, entitlement to free school meals, fluency in the English language, previous nursery education, term of starting school, attendance and special educational needs were collected.

Blatchford *et al.* (2002) presented separate models relating achievement in literacy and mathematics at the end of the reception year to class size, after adjusting for other determinants. Because of the non-normality of the end of reception year mathematics results (Fig. 1), the scores were normalized at the start of their analysis. They found a

> 'clear effect of class size difference on children's attainment in reception year, both before and after adjusting for possible confounding factors'.

However, the results are only available on an unfamiliar normalized scale, limiting their accessibility by educationalists.

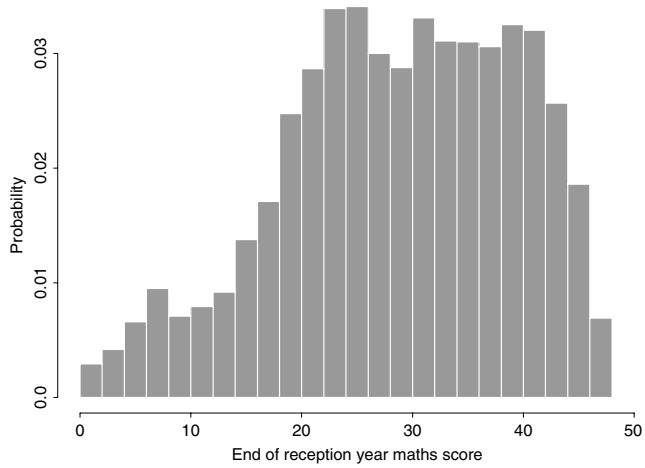

Fig. 1. Histogram of results of the mathematics evaluation at the end of the reception year: the maximum possible score is 48

Motivated by this issue, this paper describes a nonparametric residual bootstrap, suitable for this data set, and uses it to derive a confidence interval for the relationship between end of reception year mathematics score and class size, after adjusting for other determinants.

3. Bootstrap methods for multilevel models

In this section we review the parametric bootstrap for multilevel models (Section 3.1) and propose a nonparametric alternative in Section 3.2. Firstly though we briefly review bootstrap methods in regression.

Bootstrapping is now a well-established procedure for estimating uncertainty of parameter estimators in statistical models, and the nonparametric bootstrap is particularly useful (Davison and Hinkley, 1997; DiCiccio and Efron, 1996; Young, 1994; Efron and Tibshirani, 1993). In the context of ordinary least squares regression of an $n \times 1$ vector Y on an $n \times (1 + p)$ matrix X (where the first column contains 1s), three kinds of bootstrap can be employed (Davison and Hinkley (1997), pages 262–266).

The first is a *parametric residual* bootstrap. Y is regressed on X, giving estimates of the $(1 + p) \times 1$ column vector of regressors, $\hat{\beta}$, and the residual variance $\hat{\sigma}^2$. Then bootstrap residuals, say r_i^* $(i = 1, \ldots, n)$, are sampled from a normal distribution, with variance $\hat{\sigma}^2$. The ith bootstrap response is then $y_i^* = x_i\hat{\beta} + r_i^*$. The second, a *nonparametric residual* bootstrap, is similar, but the r_i^*s are now drawn with replacement from the suitably rescaled and centred set of empirical residuals $\hat{e}_i$.

Conversely, using the third approach, *case resampling*, bootstrap data sets can be generated before any modelling is done. This approach constructs the bootstrap data matrix by sampling with replacement rows from the $n \times (2 + p)$ matrix (Y, X).

Clearly, these approaches work quite differently and can be expected to yield different results. Whereas the residual resampling methods assume both that the model is correctly specified and variance homogeneity (so that the residuals are exchangeable), case resampling assumes neither of these; moreover it does not assume that the conditional mean of $Y|X = x$ is a linear function of x. Case resampling is thus more robust to heteroscedasticity and model misspecification but inefficient otherwise (Davison and Hinkley (1997), page 264). Further, with case resampling,

the simulated samples have different design matrices, as the xs are randomly sampled. This is undesirable, since in principle inference should be conditional on the xs. Lastly, whereas the means and variances of the set of $\hat{\beta}^*$s obtained by using residual resampling will be very close to those obtained by using ordinary least squares (Davison and Hinkley (1997), page 262), this is not generally true for the set of $\hat{\beta}^*$s obtained by case resampling, although the difference is often negligible in large data sets.

Now consider multilevel class size data, consisting of observations on pupils in classes in schools in education authorities. If we sample with replacement from pupils to create a bootstrap data set, we destroy the natural hierarchy of the data. To avoid doing this, we must resample at the highest level. However, often in educational data there will be few units at this level, so we cannot apply the case resampling bootstrap with confidence (see also Davison and Hinkley (1997), pages 100–102).

All this motivates us to develop a residual bootstrap approach. First we discuss a parametric version and then introduce a nonparametric version.

3.1. Parametric bootstrap for random-effects models

To fix our ideas, we discuss data from a two-level hierarchy, which would represent pupils (level 1) within classes (level 2). However, our approach is directly applicable to any level of hierarchy (so it extends to include schools (level 3)), although the notation becomes increasingly cumbersome. Let $j = 1, \ldots, J$ index the J classes and $i = 1, \ldots, I_j$ index the I_j pupils in the jth class, with data Y_{ij}. Denote by x_{ijk} and z_{ijk} $(k = 0, \ldots, K)$ the kth element of the $K + 1$ covariates for the fixed and random part of the model for observation Y_{ij}, where $x_{ij0} = z_{ij0} = 1$. It is not necessary in general for the dimension of the fixed and random parts of the model to be equal; we adopt this for notational simplicity. Then the normal errors model is

$$y_{ij} = \sum_{k=0}^{K} \beta_k x_{ijk} + \sum_{k=0}^{K} u_{jk} z_{ijk} + e_{ij}, \tag{1}$$

where the βs are fixed effects, the us are random class effects from a $(K+1)$-dimensional normal distribution with mean 0 and covariance matrix Σ and e_{ij} are independent identically distributed errors (at the pupil level) from an $N(0, \sigma_e^2)$ distribution and are independent of the us.

We assume that the parameters $(\beta, \Sigma, \sigma_e^2)$ have been estimated by maximum, or restricted maximum, likelihood (Goldstein, 1986, 1989). Then the parametric residual bootstrap proceeds as follows.

Step 1: simulate $e_{ij}^* \sim N(0, \hat{\sigma}_e^2)$ $(i = 1, 2, \ldots, I_j; j = 1, \ldots, J)$. Further, simulate the $1 \times (K+1)$ row vector $u_j^* = (u_{j0}^*, u_{j1}^*, \ldots, u_{jK}^*)$ from the $(K+1)$-dimensional normal distribution $N(0, \hat{\Sigma})$ for $j = 1, \ldots, J$.

Step 2: calculate the bootstrap data (y_{ij}^*, x_{ij}) by setting

$$y_{ij}^* = \sum_{k=0}^{K} \hat{\beta}_k x_{ijk} + \sum_{k=0}^{K} u_{jk}^* z_{ijk} + e_{ij}^*.$$

Step 3: refit the model to the bootstrap data to obtain the first set of bootstrap estimates $(\{\hat{\beta}_k^*\}_{k=0,\ldots,K}, \hat{\Sigma}^*, \hat{\sigma}_e^{2*})$.

Step 4: repeat steps 1–3 B times to obtain B sets of bootstrap parameter estimates for inference.

Although the parametric bootstrap can be useful for bias correction, particularly when the data are discrete (Rasbash *et al.*, 2000), it does not free inference from the assumption that the residuals have a normal distribution. Thus the bootstrap confidence intervals will not adequately

reflect any non-normality in the data, such as will be present if we are modelling educational test scores.

3.2. A nonparametric residual bootstrap for random-effects models

In the light of the foregoing discussion, what is needed is a generalization of the residual non-parametric bootstrap to the multilevel case. A crude residual bootstrap for model (1) would be as follows.

Step 1: obtain parameter estimates for model (1) from the data (either by maximum or restricted maximum likelihood (Goldstein, 1986, 1989)), and calculate the residuals at level 1, $\{\hat{e}_{ij}\}_{i=1,\ldots,I_j;j=1,\ldots,J}$, and level 2, $\{\hat{u}_j\}_{j=1,\ldots,J}$ (note that the $\hat{u}_j$ are $1 \times (K+1)$ row vectors).
Step 2: sample independently with replacement from these two sets, obtaining two new sets $\{e^*_{ij}\}_{i=1,\ldots,I_j;j=1,\ldots,J}$ and $\{u^*_j\}_{j=1,\ldots,J}$.
Step 3: the bootstrap data set is then $(y^*_{ij}, x_{ij}, z_{ij})$ $(i=1,2,\ldots,I_j;\ j=1,\ldots,J)$, where

$$y^*_{ij} = \sum_{k=0}^{K} \hat{\beta}_k x_{ijk} + \sum_{k=0}^{K} u^*_{jk} z_{ijk} + e^*_{ij} \quad (i=1,\ldots,I_j;\ j=1,\ldots,J).$$

Step 4: refit the model to the bootstrap data to obtain one set of bootstrap parameter estimates $(\{\hat{\beta}^*_k\}_{k=0,\ldots,K}, \hat{\Sigma}^*, \hat{\sigma}^{2*}_e)$.
Step 5: repeat steps 2–4 to obtain B bootstrap parameter estimates for each of the parameters in the model.

Note that, in step 2, we sample the $1 \times (K+1)$ row vectors $\{\hat{u}_j\}_{j=1,\ldots,J}$ with replacement, not the individual elements of these vectors.

Sampling from the $\{\hat{e}_{ij}\}$ and $\{\hat{u}_j\}$ independently breaks the correlation between the estimates of the us and es, so that the sets $\{e^*_{ij}\}$ and $\{u^*_j\}$ are uncorrelated, in line with the model assumptions. The drawback is that this will yield underdispersed bootstrap distributions of parameter estimates and downwardly biased variance parameter estimates. This is because the crude residuals' variance–covariance matrix is different from the maximum (restricted) likelihood estimate, as the crude residuals are 'best linear unbiased predictors' which are 'shrunk' towards 0 (Robinson, 1991). We therefore need to 'reflate' the residuals, so that their covariance matrix is equal to the maximum (restricted) likelihood estimate of the covariance matrix obtained from the model, before passing them back through the fitted model in step 3 above.

We now describe a procedure for doing this. For convenience we outline the procedure by using the level 2 (class) residuals, but analogous operations can be carried out at all levels. First we fit the model and calculate the 'class' residuals. This gives an estimate of the class effects $\hat{u}_j = (\hat{u}_{j0}, \ldots, \hat{u}_{jK})$ for $j=1,\ldots,J$. Let these form the rows of a $J \times (K+1)$ matrix $\hat{U}$. Then write the empirical covariance matrix of the estimated residuals at level 2 in model (2) as

$$S = \hat{U}^{\mathrm{T}}\hat{U}/J$$

and the corresponding maximum (restricted) likelihood estimate of the covariance matrix of the random class residuals at level 2 as R. The empirical covariance matrix is estimated using the number of classes (i.e. level 2 units), J, as divisor rather than $J-1$. We assume that the estimated residuals have been centred. We seek a transformation of the residuals of the form

$$\hat{U}^* = \hat{U}A \tag{2}$$

where A is a matrix of order equal to the number of random coefficients at level 2 such that

$$\hat{U}^{*\mathrm{T}}\hat{U}^*/J = A^{\mathrm{T}}\hat{U}^{\mathrm{T}}\hat{U}A/J = A^{\mathrm{T}}SA = R. \tag{3}$$

To form A we note the following. Write the Cholesky decomposition of S, in terms of a lower triangular matrix, as

$$S = L_S L_S^{\mathrm{T}} \tag{4}$$

and the Cholesky decomposition of R as

$$R = L_R L_R^{\mathrm{T}}. \tag{5}$$

We have

$$L_R L_S^{-1} \hat{U}^{\mathrm{T}} \hat{U} (L_R L_S^{-1})^{\mathrm{T}} / J = L_R L_S^{-1} S (L_S^{-1})^{\mathrm{T}} L_R^{\mathrm{T}} = L_R L_R^{\mathrm{T}} = R. \tag{6}$$

Thus, we can choose the lower triangular matrix

$$A = (L_R L_S^{-1})^{\mathrm{T}}.$$

In Appendix A, we apply the work of Shao *et al.* (2000) (who proposed a case resampling bootstrap) to show that this method will give asymptotically correct coverage of bootstrap percentile confidence intervals (for a description of percentile intervals, see Carpenter and Bithell (2000)).

We note, however, that A is not unique since

(a) we can choose either upper or lower Cholesky decompositions of S and R and
(b) expressions (3)–(6) are unaltered when we substitute $L_S W_S$ for L_S and $L_R W_R$ for L_R where W_S and W_R are orthogonal matrices. In this case we have the more general form $A = (L_R W_R W_S^{-1} L_S^{-1})^{\mathrm{T}}$.

Although these choices will not affect asymptotic consistency, which depends on second-order properties, some choices may provide better confidence interval coverage in finite samples. Specifically, as in the case of residual bootstrapping in ordinary least squares, we are aiming to estimate the true cumulative distribution function with the empirical distribution function of the residuals (Davison and Hinkley (1997), page 261). Removing the skewness of the residuals will not aid this; rather, leaving the residuals' skewness as little changed as possible while correcting their variance should minimize the difference between the cumulative distribution function and the empirical distribution function, and thus minimize the difference between actual and nominal coverage error. Thus the transformation that achieves this in the multilevel setting is preferable.

The new set of transformed class level residuals $\hat{U}^*$, given by model (2), now have covariance matrix equal to that estimated from the model, and we sample vectors of residuals (corresponding to rows of $\hat{U}^*$) with replacement, as described in the nonparametric residual bootstrap algorithm above. To complete the residual bootstrap, we repeat this process at every level of the model, with sampling being independent across levels.

3.3. Some remarks

In a single-level (i.e. ordinary least squares) model our rescaling approach is extremely close to the nonparametric residual bootstrap for linear regression that was proposed by Davison and Hinkley (1997), pages 257–262. To see this, note that, if n is the number of observations, in this case the sample variance of the residuals, $\Sigma \hat{e}_j^2/n$, will be a downwardly biased estimate of the true residual variance; an unbiased estimate is $\Sigma \hat{e}_j^2/(n-p)$ (where p is the number of parameters in the model). Our procedure forms new residuals $\hat{r}_j = \hat{e}_j \sqrt{\{n/(n-p)\}}$, whose sample variance is the unbiased estimate of residual variability; the variance of the resampling residuals

that was proposed by Davison and Hinkley (1997), page 262, algorithm 6.1, only approximates this.

Davison and Hinkley (1997), pages 100–101, also discussed a simple residual bootstrap and gave a shrinkage correction, with the same aim as ours, for a simple variance components model.

To conclude this section, note that the procedure can be applied when subsets of the random effects have different variances, e.g. if the pupil level variances (level 1) differ between the sexes. All that is required is exchangeability within the groups.

4. Simulation study

Here we report the results of a simulation study to compare the parametric and new nonparametric residual bootstrap. Let $j = 1, \ldots, J$ index level 2 units (e.g. classes) and $i = 1, \ldots, I_j$ index level 1 units (e.g. pupils). Here there are the same number of level 1 units for each level 2 unit (i.e. pupils within each class), so all $I_j = I$. We use the model

$$y_{ij} = \alpha + u_{j\alpha} + (\beta + u_{j\beta})x_{ij} + e_{ij}, \tag{7}$$

where $\alpha = 3$, $\beta = 5$, the x_{ij} are simulated from a normal distribution with mean 0 and variance 500 and we simulate non-normal random effects as follows.

For each $j = 1, \ldots, J$ we draw a sample (z_{j1}, z_{j2}) from the bivariate normal distribution

$$N\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right\},$$

so that the covariance of (z_{j1}, z_{j2}), denoted σ_{z_1,z_2}, is 0.5. We then set $u_{j\alpha} = z_{j1}^2 - 1$ and $u_{j\beta} = z_{j2}^2 - 1$. This means that both $u_{j\alpha}$ and $u_{j\beta}$ have marginal $(\chi_1^2 - 1)$-distributions whose mean is 0 and variance is 2. It is routine to show that $\mathrm{cov}(u_{j\alpha}, u_{j\beta}) = 2\sigma_{z_1,z_2}^2$, which is $2 \times 0.5^2 = 0.5$ in this case.

Lastly, we draw $e_{ij} \sim \chi_1^2 - 1$ independently of the us, so that the random terms at level 2 are independent of those at level 1.

Initially, we had $J = 20$ level 2 units each with $I = 10$ level 1 units. At each replication, a 'data set' was simulated from the model, and then, using both a parametric and a nonparametric residual approach, bootstrap percentile intervals (Carpenter and Bithell (2000) and Davison and Hinkley (1997), page 202) were constructed.

We implemented the parametric bootstrap as described in Section 3.1. Specifically, at each replication, we obtained restricted maximum likelihood estimates of the variance of e_{ij}, denoted $\hat{\sigma}_e^2$, and the 2×2 covariance matrix of the us, denoted $\hat{\Sigma}$, by fitting model (7) to the simulated data set. The parametric bootstrap then drew $e_{ij}^* \sim N(0, \hat{\sigma}_e^2)$, and

$$u_j^* = \begin{pmatrix}u_{j,\alpha}^*\\u_{j,\beta}^*\end{pmatrix} \sim N\left\{\begin{pmatrix}0\\0\end{pmatrix}, \hat{\Sigma}\right\},$$

as in step 1 of the algorithm in Section 3.1.

Likewise, we implemented the nonparametric bootstrap as described in Section 3.2, after the residuals had been 'reflated' by using our proposal. Note, in particular, after having rescaled the set of 1×2 level 2 residual vectors $\{\hat{u}_j\}_{j=1,\ldots,J}$, we sample individual vectors with replacement from this set, as described in step 2 of the algorithm in Section 3.2.

For each 90% bootstrap confidence interval we used 999 simulations, in line with Davison and Hinkley (1997), page 202. A total of 500 replicates were used as this was sufficient for a clear picture to emerge.

Table 1. Coverage of nominal 90% bootstrap confidence intervals for parameters in model (7), for various sample sizes

Parameter	*Coverages (%) for the following sample sizes and methods:*					
	20 level 2, 10 level 1		*40 level 2, 20 level 1*		*80 level 2, 40 level 1*	
	Parametric	*Nonparametric*	*Parametric*	*Nonparametric*	*Parametric*	*Nonparametric*
α	84	85	86	86	93	92
β	83	84	86	85	89	90
$\sigma^2_{u_\alpha}$	53	66	48	69	49	80
$\sigma^2_{u_\beta}$	46	64	52	74	48	79
$\mathrm{cov}(u_\alpha, u_\beta)$	61	67	65	73	69	79
σ^2_e	53	78	50	87	50	87

Subsequently the experiment was repeated with two larger sample sizes:

(a) $J = 40$ level 2 units, each with $I = 20$ level 1 units, and
(b) $J = 80$ level 2 units, each with $I = 40$ level 1 units.

These larger sample sizes correspond more closely to our data set.

Table 1 shows the results. The coverage is always better for the nonparametric bootstrap, with the most dramatic improvements being for the random parameters. Moreover, the coverage error decreases as the sample size increases, in line with theory. Perhaps because of the skewness of the χ^2_1-distribution, the coverage error does not disappear for the largest sample sizes that are considered here.

We also examined, for each of the three sample sizes, the average of the parameter estimates from the 500 simulated data sets, and the average of the parametric and nonparametric bootstrap parameter estimates.

After allowing for multiple testing (to ensure an overall type I error of 5% or less), the average of the parameter estimates from the simulated data sets was within sampling variation of the true parameter values.

The means of the 500×999 bootstrap parameter estimates, from the parametric and nonparametric bootstrap, were also within sampling variation of each other. However, for several parameters in the largest sample size, and the covariance in the middle sample size, there was a detectable difference between the average of the parameter estimates from the 500 simulated data sets and the average of the 500×999 bootstrap parameter estimates, with the bootstrap parameter estimates less than 1.5% higher. Moreover, the percentage of data sets for which the estimation, which does not allow negative variance components, fails increases from around 0.5% to 1.5% as the sample size increases. This indicates that this difference arises because we do not allow negative variance estimates, an observation that is reinforced by other simulations that we have done.

Finally, note that there is little difference between the parametric and nonparametric bootstrap in terms of computational load.

The simulations suggest that the nonparametric approach is always preferable to the parametric and confirm that for large data sets, such as the class size project data that are analysed here, the coverage error of bootstrap confidence intervals is likely to be small, even when the residuals at all levels are highly non-normal.

5. Analysis of class size data

As mentioned earlier, Blatchford *et al.* (2002) derived a model for the effect of class size on mathematics achievement, after adjusting for a range of other possible determinants. Their response variable is a transformed measure of achievement, but here we use the actual mathematics test score at the end of the reception year.

As there is no reason to believe that the relationship between class size and achievement should be linear, yet more complex polynomials tend to impose too rigid constraints, especially at the extreme class sizes, Blatchford *et al.* (2002) explored various regression spline approaches, which modify the basic polynomial with the addition of smoothly joining local polynomials at selected 'knots'.

In this reanalysis, we followed this approach. We began with a variance components model and a quadratic relationship between achievement and class size. To avoid an upturn at high class sizes, we explored up to two knots at a variety of class sizes. The simplest relationship which captured the main features had a single knot at 25 years,

$$\beta_0 + \beta_1 \text{ class size} + \beta_2 \text{ class size}^2 + \beta_3(\text{class size} - 25)_+^2,$$

where $(z)_+ = z$ if $z \geqslant 0$ and $(z)_+ = 0$ otherwise.

We then adjusted for the determinants of class size that were identified by Blatchford *et al.* (2002), retaining parameters with estimates that were significant at the 5% level (with the exception of sex, whose effect is of particular interest). As we are working on a different scale, it turns out that a slightly simpler model results.

The estimated coefficients, with class size centred at 30, are given in Tables 2 and 3. We see that, after adjusting for preschool attainment in mathematics and literacy, term of entry, eligibility for free school meals, sex and age at entry, there is still a strong effect of class size, particularly for those with below average attainment in mathematics at the start of the year. Fig. 2 shows the estimated effect of class size on mathematics score, for a girl with average pre-reception mathematics ability and literacy, at a school with average pre-reception mathematics ability and literacy, who started school in the autumn at the average age, and who was ineligible for free school meals.

We calculated confidence intervals as follows. First, we calculated equitailed 95% normal theory intervals by using the asymptotic covariance matrix of the parameter estimates. This is shown by the broken curve in Fig. 2. We then used the nonparametric bootstrap to obtain 999 sets of bootstrap parameter estimates. Using these bootstrap estimates, we calculated the bootstrap sample estimate of the covariance matrix of the parameter estimates and used this instead of the asymptotic covariance matrix. As an alternative, we used the bootstrap parameter estimates to calculate a 95% bootstrap percentile interval at a number of class sizes. Both methods for constructing the bootstrap confidence intervals give similar results, although the percentile bootstrap intervals are not always symmetric about the mean. In Fig. 1, we therefore plot the curve given by joining together the 95% bootstrap percentile intervals calculated at a number of class sizes.

The nonparametric bootstrap confidence interval is close to the normal theory interval, but slightly narrower, particularly at smaller class sizes. We conclude that there is no evidence of an improved achievement in mathematics until the class size drops below 25, but then the improvement is quite marked. This is more remarkable when it is considered that, in relation to home influences and other within-child factors, the influence of school experience and, within that, class size, might be expected to be relatively small.

We can also use the nonparametric bootstrap to calculate a confidence interval for the

Table 2. Fixed effects estimates for the model for the end of reception year attainment in mathematics†

Variable	*Coefficient (standard error)*
Intercept	32.8 (1.601)
Class size − 30	0.590 (0.4615)
(Class size − 30)^2	6.48×10^{-2} (2.894×10^{-2})
Pre-reception mathematics score‡	0.268 (0.1670)
(Pre-reception mathematics score‡)^2	1.34×10^{-2} (1.133×10^{-2})
Pre-reception literacy score‡	0.222 (9.994×10^{-3})
(Pre-reception literacy score‡)^2	-2.15×10^{-3} (4.204×10^{-4})
Entry in spring or summer term§	−5.53 (0.4338)
Eligible for free school meals§	−0.856 (0.2941)
Boy§	2.24×10^{-3} (0.1932)
Age at start of reception‡	1.53 (0.4203)
Lowest quartile of pre-reception mathematics score§	−0.243 (0.5287)
Highest quartile of pre-reception mathematics score§	-9.90×10^{-2} (0.5208)
Pre-reception mathematics score − school average	0.617 (0.1597)
(Pre-reception mathematics score − school average)^2	-2.84×10^{-2} (8.667×10^{-3})
Interactions	
Class size − 30 and lowest quartile of pre-reception mathematics score§	−0.166 (7.216×10^{-2})
Class size − 30 and highest quartile of pre-reception mathematics score§	6.66×10^{-2} (7.464×10^{-2})
Spline term	
If class size > 25, (class size − 25)^2	-6.35×10^{-2} (5.650×10^{-2})

†The response is the mathematics score, which lies between 0 and 48.
‡Variable is centred on its sample average.
§Dummy variable, taking on the value 1 for the given category.

Table 3. Estimated covariance matrix for the model for the end of reception year attainment in mathematics

Component of variance	*Estimated covariance term (standard error)*		
School level			
Intercept	10.5 (2.821)		
Pre-reception mathematics score	−0.128 (0.1468)	4.54×10^{-2} (1.625×10^{-2})	
(Pre-reception mathematics score)^2	−0.100 (3.249×10^{-2})	2.40×10^{-3} (2.309×10^{-3})	1.19×10^{-3} (6.641×10^{-4})
Class level			
Between-class variance	12.4 (2.107)		
Pupil level			
Between-pupil variance	39.7 (0.8696)		

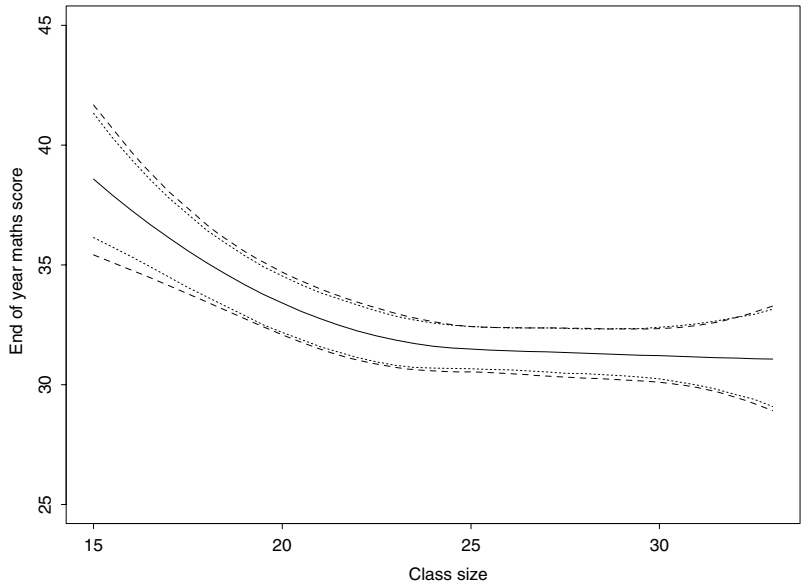

Fig. 2. Effect of class size on mathematics score, with 95% nonparametric bootstrap percentile confidence interval (·····) and 95% normal theory confidence interval (-------): the maximum possible score is 48; for details, see the text

intraclass correlation, $\sigma^2_{\text{class}}/(\sigma^2_{\text{class}} + \sigma^2_{\text{pupil}})$. This reflects the breakdown between pupil and class variability, and is important for planning the size of future studies. The approximate normal theory 95% interval, derived by using the delta method (Pawitan (2001), page 89), is (0.17,0.3), which is similar to the nonparametric bootstrap percentile interval, (0.18,0.29) in this case.

6. Conclusions

The relationship between class size and achievement has important policy implications. Although the possibility remains that further confounding factors exist, which could modify the conclusions, and that the results may not generalize to other parts of the UK where education policy and practice may vary, our view is that the class size project provides the most extensive *prima facie* evidence for a causal effect of class size on achievement.

Further, using our proposal for a nonparametric residual bootstrap for multilevel models, we have been able to construct an accurate confidence interval for the relationship between class size and achievement, without having to work on a transformed scale. This is important, as it makes the results far more accessible to educationalists and thus enables a more informed discussion of the issues.

The simulation study that is reported in Section 4 gives us confidence in the results: it shows that even when the distribution of the residuals is χ^2_1 at all levels, provided that the sample size is large, as it is in the class size study (4621 pupils from 254 classes in 157 schools) the bootstrap will give accurate confidence intervals.

Thus the residual bootstrap provides a robust alternative to a parametric bootstrap which should be used in preference to the parametric bootstrap, even when departures from normality appear slight. It is implemented in MLwiN, version 1.10.0007 (`http://www.multilevel.ioe.ac.uk`).

Acknowledgements

The class size data were modelled by using MLwiN (`http://www.multilevel.ioe.ac.uk`) and the simulation study was carried out with R (`http://cran.r-project.org`). We thank Peter Blatchford for permission to use the class size data. We are grateful to the referees, whose comments have led to a greatly improved manuscript, and Anthony Davison for his comments on a preliminary draft.

Appendix A

We consider maximum likelihood estimates to which restricted maximum likelihood estimates are asymptotically equivalent. Shao *et al.* (2000) proposed a case resampling bootstrap for hierarchical data, a stratified version of strategy 1 of Davison and Hinkley (1997), page 100. Under certain regularity conditions, Shao *et al.* (2000) showed that, using this resampling plan, bootstrap percentile confidence intervals are asymptotically consistent. Their arguments also hold in our situation provided that the expectation, over the bootstrap distribution, of the score equations is 0.

To see that this is true, suppose that we construct the bootstrap data Y_{ij}^* for the ith response on the jth person by adding to the mean $x_{ij}^{\mathrm{T}}\hat{\beta}$ the sum of

(a) a sample from the overall, bottom level residuals and
(b) a sample from the person-specific residuals,

where both have been mean centred and rescaled so that $E_* Y_{ij}^* = x_{ij}^{\mathrm{T}}\hat{\beta}$ and, collecting together the bootstrap data on an individual, $\mathrm{cov}(Y_j^*) = \hat{V}$. Now consider the score equations for fitting the model by maximum likelihood. The log-likelihood is

$$\log(L) = -(Y - X\beta)^{\mathrm{T}} V^{-1} (Y - X\beta) - \log|V|,$$

or equivalently, if $S = (Y - X\beta)(Y - X\beta)^{\mathrm{T}}$,

$$\log(L) = -\mathrm{tr}(V^{-1}S) - \log|V|.$$

Considering the first formulation, the score equations for β are

$$-2X^{\mathrm{T}} V^{-1} (Y - X\beta).$$

So, if we replace Y by residual bootstrap data Y^* and set $\beta = \hat{\beta}$, then $Y^* - X\hat{\beta}$ is a vector and the expectation over the bootstrap distribution of the score is 0 at $\beta = \hat{\beta}$. This gives consistency of $\hat{\beta}^*$ for $\hat{\beta}$.

For the variance terms, first note that the residuals were rescaled before resampling so that $E_* S^* = \hat{V}$. Then, using the second formulation of the log-likelihood above we have that the score for a typical variance parameter, say γ_k, is

$$-\mathrm{tr}\left\{ \frac{\partial}{\partial \gamma_k}(V^{-1}S) \right\} + \mathrm{tr}\left(V \frac{\partial V^{-1}}{\partial \gamma_k} \right).$$

If $S = S^*$ then this is equal to 0 if

$$E_* \mathrm{tr}\left\{ \frac{\partial}{\partial \gamma_k}(V^{-1}S^*) \right\} = \mathrm{tr}\left(V \frac{\partial V^{-1}}{\partial \gamma_k} \right),$$

when $V = \hat{V}$, so $\gamma_k = \hat{\gamma}_k$. However, this is true because V is symmetric, so

$$\begin{aligned} E_* \mathrm{tr}\left\{ \frac{\partial}{\partial \gamma_k}(V^{-1}S^*) \right\} &= E_* \mathrm{tr}\left(S^{*\mathrm{T}} \frac{\partial V^{-1}}{\partial \gamma_k} \right) \\ &= \mathrm{tr}\left(E_* S^{*\mathrm{T}} \frac{\partial V^{-1}}{\partial \gamma_k} \right) \\ &= \mathrm{tr}\left(\hat{V} \frac{\partial V^{-1}}{\partial \gamma_k} \right). \end{aligned}$$

The result follows.

References

Avon Education Department and Institute of Education (1996) *Avon Reception Entry Assessment*. London: Institute of Education.

Blatchford, P., Goldstein, H., Martin, C., and Browne, W. (2002) A study of class size effects in English school reception year classes. *Br. Educ. Res. J.*, **28**, 169–185.

Carpenter, J. and Bithell, J. (2000) Bootstrap confidence intervals: when, which, what?: a practical guide for medical statisticians. *Statist. Med.*, **19**, 1141–1164.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

DiCiccio, T. J. and Efron, B. (1996) Bootstrap confidence intervals. *Statist. Sci.*, **11**, 189–212.

Efron, B. and Tibshirani, R. (1993) *An Introduction to theBootstrap*. New York: Chapman and Hall.

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43–56.

Goldstein, H. (1989) Restricted unbiased iterative generalised least squares estimation. *Biometrika*, **76**, 622–623.

Pawitan, Y. (2001) *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Oxford: Oxford University Press.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. and Lewis, T. (2000) *A User's Guide to MLwiN*. London: Institute of Education.

Reading Progress Tests (2000) *Literacy Baseline*. London: Hodder and Stoughton.

Robinson, G. K. (1991) That BLUP is a good thing: the estimation of random effects. *Statist. Sci.*, **6**, 15–51.

Shao, J., Kübler, J. and Pigeot, I. (2000) Consistency of the bootstrap procedure in individual bioequivalence. *Biometrika*, **87**, 573–585.

Young, G. A. (1994) Bootstrap: more than a stab in the dark? *Statist. Sci.*, **9**, 382–415.