ELSEVIER

Statistics & Probability Letters 56 (2002) 425–430

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

The BLUPs are not "best" when it comes to bootstrapping

Jeffrey S. Morris *

Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Box 447, Houston, TX 77030-4009, USA

Received June 2001; received in revised form November 2001

Abstract

In the setting of mixed models, some researchers may construct a semiparametric bootstrap by sampling from the best linear unbiased predictor residuals. This paper demonstrates both mathematically and by simulation that such a bootstrap will consistently underestimate the variation in the data in finite samples. © 2002 Published by Elsevier Science B.V.

Keywords: Bootstrap; Correlated data; Mixed models; Nested models

1. Introduction

Mixed models are used extensively in various settings to model continuous grouped data. Their flexibility and availability in major statistical software packages contribute to their popularity. Besides giving point estimates of fixed effects and variance components in the model, these packages also yield standard errors, confidence intervals, and test of hypotheses for these quantities. These rely on asymptotic results assuming a Gaussian model that may not be appropriate in some situations.

When the appropriateness of using asymptotic results is questioned, resampling procedures such as the bootstrap can be used to estimate these quantities. In the setting of mixed models, the parametric bootstrap (Efron and Tibshirani, 1993) is one alternative. In this method, a Monte Carlo simulation is used to generate a large number of bootstrap samples from the assumed parametric distribution of the data, with estimators from the data plugged in for the fixed effects and variance components. Standard errors and quantiles for statistics of interest are then estimated from these bootstrap samples.

The parametric bootstrap relies heavily upon the Gaussian assumption for the random effects in the mixed model. In fact, it can be shown that for inference on variance components, the parametric

* Tel.: +1-713-794-1720; fax: +1-713-745-4940.
E-mail address: jeffmo@odin.mdacc.tmc.edu (J.S. Morris).

0167-7152/02/$ - see front matter © 2002 Published by Elsevier Science B.V.
PII: S0167-7152(02)00041-X

bootstrap is inconsistent if these assumptions are violated. In cases where the normality of the random effects' distribution is not strongly believed, one may wish to use a nonparametric bootstrap, nonparametric in the sense that it does not directly depend upon the distributional assumptions on the random effects. This cannot be accomplished through simple resampling, since that would treat the data as independent, ignoring the correlation structure. Any valid nonparametric bootstrap procedure must account for the grouping structure in the data.

In the setting of a simple linear regression, one nonparametric bootstrap method proposed by Efron (1979,1982) is the *bootstrap based on residuals*. In this procedure, the bootstrap data set is constructed using samples taken with replacement from the estimated empirical distribution of the residuals, which are then added onto an estimate of the mean function obtained from the data. This idea can be extended to mixed models by sampling with replacement from predictors of the random effects and residuals in the model. The major issue is deciding which predictors to use. A natural choice is the best linear unbiased predictors (BLUPs), since they are in some sense optimal, are readily available in many statistical software packages, and have become popular in recent years for various applications. This may lead some to consider constructing a nonparametric bootstrap by simply resampling the BLUPs, as I and my colleagues have encountered while reviewing submitted papers.

While the BLUPs have some properties of optimality in predicting an individual's random effect, this optimality does not transfer over to bootstrapping. A BLUP-based bootstrap will consistently underestimate the variability in the data for small samples, which is the most likely scenario in which it will be used. Mathematical results are presented supporting this fact, and the poor performance of the BLUP bootstrap is demonstrated in the setting of a simple random effects model by simulation.

2. BLUPs for mixed models

The standard general linear mixed model, discussed in Searle et al. (1992), can be written as

$$\underline{y} = X\underline{\beta} + Z\underline{\gamma} + \underline{\varepsilon}, \tag{1}$$

where X and Z are the design matrices for the fixed effects, $\underline{\beta}$, and random effects, $\underline{\gamma}$, respectively. The residuals, $\underline{\varepsilon}$, are assumed to be mean zero random variables with covariance matrix R, and the random effects $\underline{\gamma}$ are assumed to be mean zero random variables with covariance matrix G. With these assumptions, the response vector $\underline{y}$ has mean $X\underline{\beta}$ and covariance matrix $V = ZGZ^{\mathrm{t}} + R$.

BLUP is a method of predicting the random effects $\underline{\gamma}$ in a linear mixed model. The BLUPs are *linear* functions of the data, and are *unbiased* in that the expectation of the predictors, $\hat{\underline{\gamma}}$, is the same as the expectation of the random effects. Note that this does not mean that $E(\hat{\underline{\gamma}}|\underline{\gamma}) = \underline{\gamma}$ for all $\underline{\gamma}$ (see Robinson, 1991). They are called *predictors* rather than estimators because the quantities they represent are random variables, not fixed parameters, and they are *best* in the sense that they minimize the generalized mean square error of prediction, $E(\hat{\underline{\gamma}} - \underline{\gamma})^{\mathrm{t}}B(\hat{\underline{\gamma}} - \underline{\gamma})$, with B being any positive-definite symmetric matrix.

There are many derivations for the BLUPs for random effects in a general mixed model setting (see Robinson, 1991). The BLUPs of the random effects $\hat{\underline{\gamma}}$ can be obtained simultaneously along with the best linear unbiased estimators (BLUE) of the fixed effects $\hat{\underline{\beta}}$ by solving Henderson's mixed model equations (1950), which are

$$X^{\mathrm{t}}R^{-1}X\underline{\hat{\beta}} + X^{\mathrm{t}}R^{-1}Z\underline{\hat{\gamma}} = X^{\mathrm{t}}R^{-1}\underline{y},$$

$$Z^{\mathrm{t}}R^{-1}X\underline{\hat{\beta}} + (Z^{\mathrm{t}}R^{-1}Z + G^{-1})\underline{\hat{\gamma}} = Z^{\mathrm{t}}R^{-1}\underline{y}. \tag{2}$$

These equations can be solved to obtain plug-in formulas for $\underline{\hat{\beta}}$ and $\underline{\hat{\gamma}}$,

$$\underline{\hat{\beta}} = (X'V^{-1}X)^{-}X^{\mathrm{t}}V^{-1}\underline{y}, \tag{3}$$

$$\underline{\hat{\gamma}} = GZ^{\mathrm{t}}V^{-1}(\underline{y} - X\underline{\hat{\beta}}). \tag{4}$$

In practice, the covariance matrices G and R are usually unknown, and must be estimated from the data. In this case, $\underline{\hat{\gamma}}$ are known as "estimated BLUPs".

Although the principles discussed in this paper transfer to more complex mixed models, the problems with the BLUP bootstrap will be illustrated using the simple random effects model with a balanced design. This model can be written as

$$y_{ij} = \mu + \gamma_i + \varepsilon_{ij}, \tag{5}$$

where $i=1,\ldots,r$ and $j=1,\ldots,n$. The overall fixed mean is μ, and the random effects γ_i and residuals ε_{ij} are independent and identically distributed $\mathrm{N}(0,\sigma_\gamma^2)$ and $\mathrm{N}(0,\sigma_\varepsilon^2)$, respectively, with the γ_i and ε_{ij} mutually independent.

In this case, simple calculations show that the BLUE of μ and the BLUPs of γ_i are

$$\hat{\mu} = \hat{y}_{\cdot\cdot} = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n} y_{ij}, \tag{6}$$

$$\hat{\gamma}_i = \left(\frac{n\sigma_\gamma^2}{n\sigma_\gamma^2 + \sigma_\varepsilon^2}\right)(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}). \tag{7}$$

BLUP predictors for the residuals would simply be $\hat{\varepsilon}_{ij} = y_{ij} - \bar{y}_{i\cdot}$.

3. The BLUP bootstrap

Following is one algorithm for a BLUP-based bootstrapping procedure for the simple balanced random effects model. Algorithms could be constructed for more complex mixed models using a similar approach.

(1) Fit mixed model to the data to obtain maximum likelihood (ML) estimates for μ, $\hat{\mu}$, restricted maximum likelihood (REML) estimates for σ_γ^2 and σ_ε^2, and estimated BLUPs for the random effects γ_i and residuals ε_{ij}.

(2) Take a sample of size r with replacement from the estimated BLUPs $\{\hat{\gamma}_1,\ldots,\hat{\gamma}_r\}$. Call these γ_i^*.

(3) Take a sample of size $r \times n$ with replacement from the predictors for the residuals $\{\hat{\varepsilon}_{11},\ldots,\hat{\varepsilon}_{ij}\}$. Alternatively, the residuals could be kept in groups by individuals, so that for each individual,

a random i^* is selected from $\{1,\ldots,r\}$, then a sample of size n is taken with replacement from $\{\hat{\varepsilon}_{i^*1},\ldots,\hat{\varepsilon}_{i^*n}\}$. Call these samples ε^*_{ij}.

(4) Construct the bootstrap data set according to the structure of the model, i.e. $y^*_{ij} = \hat{\mu} + \gamma^*_i + \varepsilon^*_{ij}$.

(5) Fit the mixed model to the bootstrap data, obtaining estimates of μ and the variance components σ^2_γ and σ^2_ε.

(6) Return to step (2), iterate B times.

This approach seems to be a natural generalization of Efron's "bootstrapping the residuals", and is a procedure that some may use in an ad hoc fashion as a nonparametric bootstrap for mixed models. I and my colleagues have encountered some researchers employing this method without justification while reviewing papers submitted to statistical journals. In reality, this method is actually a "semiparametric" bootstrap, since it depends on the structure of the model, but not the distributional assumptions. Use of this method implies an underlying confidence in the assumption that the empirical distributions of the estimated BLUPs effectively mimic the distribution of the random effects in the model, an assumption that is not true. It is untrue even if the variance components in the model are known and we can obtain the actual BLUPs. The BLUPs are, of course, mean zero, but for the simple balanced random effects model, straightforward calculations (see the appendix) show that their variance is given by

$$\mathrm{Var}(\hat{\gamma}_i) = \left(\frac{n\sigma^2_\gamma}{n\sigma^2_\gamma + \sigma^2_\varepsilon}\right)(1 - r^{-1})\sigma^2_\gamma, \tag{8}$$

$$\mathrm{Var}(\hat{\varepsilon}_{ij}) = (1 - n^{-1})\sigma^2_\varepsilon. \tag{9}$$

We see clearly that the variances of the BLUPs are biased downwards as estimators of the variance components. This effect is more pronounced for small sample sizes in r or n, and for cases where the residual variance σ^2_ε is much larger than σ^2_γ, such that σ^2_ε is not negligible compared with $n\sigma^2_\gamma$. This bias will cause the BLUP bootstrap to underrepresent the variation in the data, resulting in confidence bounds for the fixed effects that are too narrow, and confidence bounds for the variance components that are miscentered. While it is true that this problem disappears asymptotically, it is in the small sample cases that people may be most likely to use such a bootstrap procedure. A similar phenomenon can be shown for more general mixed models using the results of Speed (1991) and McGilchrist and Yau (1995).

4. Simulation

A simulation study was performed to examine the performance of the BLUP bootstrap for the simple balanced random effects model for some different sample sizes. In all cases, the fixed mean μ was assumed to be zero without loss of generality, and the variance components σ^2_γ and σ^2_ε were assumed to be one.

Under each scenario, 500 data sets were generated, and 90% confidence intervals were constructed for the fixed mean and both variance components using standard asymptotic results and the BLUP bootstrap. In performing the BLUP bootstrap, $B = 500$ bootstrap samples were taken. Calculations were done using the lme function in S-PLUS. The asymptotic confidence intervals were the default

Table 1
Coverage probabilities from simulation (500 runs) of nominal 90% confidence intervals using (1) asymptotic results and (2) the BLUP bootstrap ($B = 500$)

		$n = 5$			$n = 10$			$n = 30$		
r	Method	μ	σ_γ^2	σ_ε^2	μ	σ_γ^2	σ_ε^2	μ	σ_γ^2	σ_ε^2
5	1	0.840	0.966	0.946	0.850	0.992	0.960	0.808	0.962	0.978
	2	0.868	0.728	0.748	0.866	0.648	0.820	0.810	0.584	0.930
10	1	0.866	0.994	0.968	0.846	0.982	0.990	0.866	0.972	0.980
	2	0.846	0.704	0.738	0.806	0.672	0.872	0.842	0.670	0.926
30	1	0.894	0.990	0.966	0.884	0.974	0.976	0.892	0.980	0.980
	2	0.858	0.724	0.608	0.862	0.750	0.774	0.886	0.794	0.920

intervals give by lme, which are approximate confidence intervals that rely on the empirical information matrices and conditional t-tests (see Pinheiro and Bates, 2000). In the bootstrap procedures, the residuals were sampled in groups, as described in step (3) above, as this improved their performance. On each bootstrap data set, ML estimators were used for the fixed effects, and REML estimators for the variance components.

Table 1 contains the coverage probabilities of the 90% intervals using the two methods for the simulated conditions. First, note the performance of the asymptotic results. The confidence intervals on the mean have undercoverage problems for the smaller sample sizes ($r = 5$ and 10). For both variance components, the asymptotic confidence intervals are extremely wide and overconservative, resulting in coverage much greater than the nominal 90%. It is clear why some would search for alternatives to these asymptotic results for mixed models, especially for smaller sample sizes.

Use of the BLUP bootstrap does not correct these problems, however. We see that, for the fixed mean, the BLUP bootstrap has undercoverage problems similar to the asymptotic results, while the intervals for the variance components have severe undercoverage problems. For the variance component σ_γ^2, the undercoverage was extreme for all sample sizes considered, while for the residual variance σ_ε^2, the undercoverage was seen for $n = 5$ and 10, but not for $n = 30$. The BLUP bootstrap intervals for the variance components were miscentered, as expected from our theoretical results. In at least 96% of the intervals for σ_γ^2 and 91% of the intervals for σ_ε^2 not containing the true value, the true variance component was to the right of the upper confidence bound.

A question remains: When, if ever, could the BLUP bootstrap be used as an alternative to the asymptotic results and obtain reasonable coverage? The answer depends on the parameter of interest. In inference for μ, r is the important quantity to consider. For the smallest sample size considered, $r = 5$, the BLUP bootstrap had comparable or slightly better properties than the asymptotic results, and could reasonably be used. However, for moderate to large sample sizes, $r > 10$, the asymptotic results dominate the BLUP bootstrap. For σ_ε^2, our simulations (some not shown) indicate that the undercoverage is not bad for $n \geqslant 15$. For σ_γ^2, there are undercoverage problems that persist even in quite large sample sizes. A simulation done at $r = 50$, $n = 50$ (not shown) yielded a coverage of 0.845.

5. Conclusion

Bootstrapping the BLUPs seems like a natural extension of Efron's "bootstrapping the residuals" to mixed models, and is likely to be tried by some researchers as a nonparametric bootstrapping method. In this paper, it has been demonstrated that in the setting of a simple balanced random effects model, this procedure results in underestimation of the variation in the data, causing standard error estimates biased downwards and intervals with undercoverage problems. These problems are most evident in smaller samples, which is the setting in which the procedure is most likely to be used. This effect transfers to more general mixed models as well. The BLUPs, while readily available and optimal for prediction, are not optimal in a bootstrap, and should not be blindly used.

Acknowledgements

I wish to thank Raymond J. Carroll and Naisyin Wang for their helpful editorial comments concerning this paper.

Appendix A

This appendix gives the details behind the calculations of the variances of the BLUPs for the simple random effects models. Given the representation of the BLUPs for the random effects γ_i in (7), their variance is

$$\mathrm{Var}(\hat{\gamma}_i) = \left(\frac{n\sigma_\gamma^2}{n\sigma_\gamma^2 + \sigma_\varepsilon^2} \right)^2 \mathrm{Var}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}). \tag{A.1}$$

By substituting the parameters from the model given by (5) and simplifying, we see that $\mathrm{Var}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) = \mathrm{Var}(\gamma_i - \bar{\gamma}_\cdot) + \mathrm{Var}(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})$, which simplifies to $n^{-1}(1 - r^{-1})(n\sigma_\gamma^2 + \sigma_\varepsilon^2)$. Substituting back into (A.1) yields (8).

The variance of the BLUPs for the residuals is $\mathrm{Var}(y_{ij} - \bar{y}_{i\cdot}) = \mathrm{Var}(\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot})$, which is $(1 - n^{-1})\sigma_\varepsilon$, just as in (9).

References

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Statist. 7, 1–26.

Efron, B., 1982. The Jackknife, the Bootstrap, and Other Resampling Plans. SIAM, Philadelphia.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Henderson, C.R., 1950. Estimation of genetic parameters. Ann. Math. Statist. 21, 309–310 (abstract).

McGilchrist, C.A., Yau, K.K.W., 1995. The derivation of BLUP, ML, REML estimation methods for generalized linear mixed models. Comm. Statist. A—Theory Methods 24, 2963–2980.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and Splus. Springer, New York.

Robinson, G.K., 1991. That BLUP is a good thing: the estimation of random effects. Statist. Sci. 6, 15–32.

Searle, S.R., Casella, G., McCullogh, C.E., 1992. Variance Components. Wiley, New York.

Speed, T., 1991. Comment on that BLUP is a good thing: the estimation of random effects. Statist. Sci. 6, 42–44.