Austin Biometrics and Biostatistics

Open Access

Austin Publishing Group

Research Article

Implementing Nonparametric Residual Bootstrap Multilevel Logit Model with Small Number of Level-2 Units

Wang Y[1]*, Song L[2] and Wang J[1]

[1]Departments of Pediatrics and Biostatistics, The George Washington University School of Medicine, USA

[2]Merck & Company, Inc., Merck Research Laboratory, Biostatistics and Research Decision Science, USA

***Corresponding author:** Wang Y, Children's National Health System, Department of Pediatrics and Biostatistics, The George Washington University School of Medicine, 111 Michigan Avenue, NW, Washington, DC 20010, USA

Received: February 01, 2017; **Accepted:** March 16, 2017; **Published:** March 27, 2017

Abstract

It is a challenge to analyze hierarchically structured data with either numerical or categorical response variables when number of groups (e.g., level 2 units) is small. Even sample size is large, a small number of groups would cause downward bias in standard errors of parameter estimates in multilevel modeling, thus the test statistics would be enlarged and the type I error would be inflated. Both parametric and nonparametric residual bootstrap approaches have been developed to deal with a small number of groups in multilevel modeling when the response variable is numeric. However, the corresponding approach is limited for multilevel modeling with categorical response variables, ex. binary outcome. To fill the gap, we have developed an approach by implementing nonparametric residual bootstrap multilevel logit model for binary data with small number of groups using SAS macro. With simulated data for modeling binary response variable with a small number of groups, our results showed explicit advantage of the nonparametric residual bootstrap approach over the approach using the default estimator -- Residual Pseudo-Likelihood (RSPL) –in SAS Proc Glimmix.

Keywords: SAS Proc Glimmix; SAS macro; Multilevel model; Residual Bootstrap Multilevel Logit Model; Bias

Introduction

In statistics, there are two important concepts: consistence and bias of a parameter estimate. Consistence means that a parameter estimate $\hat{\theta}$ converges to its unknown true parameter θ when sample size $n \to \infty$, while bias of a parameter estimate $\hat{\theta}$ refers to the difference between $\hat{\theta}$ and θ. Bias can be fixed bias due to the system itself or random bias/ variability due to sampling errors. Evidences show that Maximum Likelihood Estimates (MLEs) of variance components are generally downwardly biased [1,2] with multilevel logit model using generalized linear mixed procedure in SAS. When the number of groups (level-2 units) is small, the bias of variance components was downward for correlated binary data [3,4]. Of many factors which can result in the downward estimation of variance components, the estimators and the assumption are the most important. First, MLE estimates of variance are smaller than Ordinary Least Square (OLS) estimates because the denominator in the formula of MLE uses the sample size *n* instead of *n* minus the rank of an independent variable matrix in OLS; Second, both level-1 and level-2 residuals are assumed to have normal distributions. However, to ensure such an assumption holds, the number of groups (number of level-2 units) have to be large. Unfortunately, the assumption of a normal distribution for level-2 residuals may not hold in real research because the number of groups is often small. The downward bias of variance components implies the smaller standard error, resulting in bigger test statistics and therefore, inflating type I error in a hypothesis test. In addition, the downward bias of variance components will result in a shorter confidence interval, thus the claimed coverage probability of 95% Confidence Interval (CI) will be less than stated [5,6]. The ways to correct the downward bias of standard errors include, but not limited to: asymptotic bias correction, Jackknife, and bootstrap. Restricted or Residual Maximum Likelihood (REML) approach [7] used in the SAS Proc Mixed model for numerical data, and Residual Pseudo-likelihood (RSPL) analog to REML used in SAS Proc Glimmix for categorical data, belong to the category with asymptotic bias correction. REML takes into account of fixed effects to maximize the likelihood function while RSPL is casted in terms of Taylor expansion to maximize the pseudo likelihood [8]. Both REML and RSPL can reduce, to some extent, the bias in estimation of some complex variance. Firth [9] proposed another approach by adding a first order bias term of the maximum likelihood estimator to the score function to prevent the bias from occurring. In Jackknife procedure [10], the leave-one-out estimators are used for bias correction with order asymptotic bias smaller than $O(n^{-1})$ [11]. Bootstrap is a collection of methods following the bootstrap framework to improve the accuracy of inference. Three kinds of bootstrap methods are available: (1) case resampling through repeated sampling from original data with replacement, (2) parametric residual bootstrap with residual resample's randomly drawn from normal distributions with replacement, (3) nonparametric residual bootstrap with residuals randomly drawn from new transformed residuals with replacement. Although both of the parametric residual bootstrap and nonparametric residual bootstrap are often used, the non-parametric residual bootstrap is more preferred because it provides more accurate inferences through correction of standard error or variance term [12] than parametric residual bootstrap. Wang et al. [13] developed an approach to conduct the nonparametric residual bootstrap multilevel modeling to deal with small number of groups

Citation: Wang Y, Song L and Wang J. Implementing Nonparametric Residual Bootstrap Multilevel Logit Model with Small Number of Level-2 Units. Austin Biom and Biostat. 2017; 4(1): 1034.

for continuous response variables using a SAS macro. However, the corresponding modeling approaches are not available for categorical (e.g., binary) response variables. To the best of our knowledge, the present study is by far the first time to apply nonparametric residual bootstrap technique to modeling binary response variables in multilevel data with a small number of groups.

We gave a brief description of how to conduct on-parametric residual bootstrap multilevel modeling; and demonstrated how to implement nonparametric residual bootstrap multilevel logit model using our SAS macro, and then analyzed simulated data using residual bootstrap approach, as well as SAS Proc Glimmix with the default estimator RSPL. The results of the different modeling approaches were compared and findings were discussed.

Methods

Non-parametric residual bootstrap multilevel logit model

The following hypothesized parsimonious multilevel logit model was used for demonstration of modeling hierarchically structured data with a binary response variable:

$$Y_{ij} = g^{-1}(\alpha_j + \beta_{1j} x_{1ij} + \varepsilon_{ij}) \quad (1)$$

$$\alpha_j = \gamma_{00} + u_{0j} \quad (2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_{1j+} u_{1j} \quad (3)$$

where i=1, 2, …, i represents the i^{th} level-1 unit, j=1, 2, …, j represents the j^{th} group (level-2 units), g^{-1} is the inverse of a link function: a logit function for the logistic regression model. The level-2 residuals u_{0j} and u_{1j}have bivariate normal distributions with zero means and unknown variances and covariance; and u_{0j} and u_{1j} are independent of the level-1 residual ε_{1j} which is assumed to have a normal distribution with the mean of zero and unknown variance.

When the number of groups is small, non-parametric residual bootstrap approach is applied for multilevel modeling (RBMLM) [13]. The primary procedure of RBMLM is to transform both level-1 and level-2 residuals of a multilevel model, and then draw random samples of the transformed residuals with replacement to generate a large set of bootstrap samples. With a continuous response variable, the model has an identity link function. A new response variable is generated as the combination of the predicted value $\hat{y}$ and the transformed residual for each subject in each bootstrap sample. For multilevel logit model, the level-1 and level-2 residuals can be estimated from SAS Proc Glimmix procedure, and then transformed using the same approach described in Wang et al. [13]. A new response variable in the bootstrap sample is generated by summing up the predicted log odds (i.e., *logit*) and the transformedlevel-1 residual. However, the generated new response variable is continuous in scale; we, therefore, need to transform it to a binary measure (0 *vs* 1) for final modeling. The specific steps for conducting non-parametric residual bootstrap multilevel logit model are described below.

Step 1: Run a multilevel logit model with the simulated dataset, save the level-1 and level-2 residuals, and then rescale the residuals by centering to ensure they have zero means. Next, transform the rescaled residuals into new residuals (see Appendix 1 in Wang, Carpenter, & Kepler [13]).

Step 2: Draw a random sample with replacement from the transformed level-1 and level-2 residuals, separately.

Step 3: Use the transformed Level-2 residuals to estimate adjusted fixed coefficients.

Step 4: Use the adjusted fixed coefficients to estimate the predicted log odds (i.e., *logit*), and then generate a new response variable by summing up the estimated log odds and the transformed level-1 residual.

Let's use the model shown in Eqs 1-3 to further describe Steps 3 and 4. After level-2 residuals are transformed, they are used to generate the adjusted fixed coefficients using Eqs. (4) and (5); and then generate a new continuous response variable using Eq. (6).

$$\hat{\alpha}_j^* = \hat{\gamma}_{00} + \hat{u}_{0j}^* \quad (4)$$

$$\hat{\beta}_{1j}^* == \hat{\gamma}_{10} + \gamma_{11} z_{1j} + \hat{u}_{1j}^* \quad (5)$$

$$y_{ij}^* = \hat{\alpha}_j^* + \hat{\beta}_{1j}^* x_{1ij} + \hat{\varepsilon}_{ij}^* \quad (6)$$

Then Step 5 follows, in which a new binary response variable y_{ij}^{**} is generated by turning the generated numeric response variable y_{ij}^* into a probability $\hat{pr}_{ij}$ using the inverse of logit function, and then comparing $\hat{pr}_{ij}$ with a random number r_{ij} that was drawn from the uniform distribution of (0,1) for each individual. If $pr_{ij} > r_{ij}$, then let $y_{ij}^{**} = 1$, else $y_{ij}^{**} = 0$. Then, refit the model shown in Eqs 1-3, using the new binary variable y_{ij}^{**} as the response variable and save the model parameter estimates.

Step 6: Repeat Steps 2-6 for a total of (B-1) times (B=500 bootstrap samples in the present study) and append the B sets of model parameter estimates. The mean and standard deviation of the empirical distribution of the bootstrap estimates for particular parameter would be the bootstrap parameter estimate and its standard error, respectively.

On the basis of the SAS macro RBMLM developed by Wang et al. [13], we developed a SAS macro for nonparametric residual bootstrap multilevel logit model. Since Proc Glimmix used to analyze data with a generalized linear mixed model can provide residuals similar to Proc Mixed, it can be used to obtain the residual terms for logit models which cannot be obtained using traditional logistic regression. A multilevel dataset with 2500 observations consisting 50 groups (level-2 units) with 50 cases in each group was simulated for demonstrating the parsimonious multilevel logit model shown in Eqs. 1-3.

Simulation

First, the model was estimated using SAS Proc Glimmix procedure with the default estimator of Residual Pseudo-Likelihood (RSPL). Then, three sub-datasets were randomly selected from the simulated dataset with different number of groups (20, 15, and 10), in which the group size remained unchanged (i.e., 50 cases per group). Then both Proc Glimmix with the default estimator RSPL and our SAS macro were applied to model each of the three sub-datasets, respectively; and the parameter estimates were compared to "true" parameters. Assuming the dataset was simulated for a "target populations," the regression coefficients (e.g., fixed-effect of x_1, cross-level interaction effect of $x_1 * z_1$) estimated from the dataset using RSPL are considered "true" parameters. If a regression coefficient estimated using a sub-dataset sampled from the original dataset deviated from the

Table 1: Selected results using nonparametric residual bootstrap multilevel logit model *vs*. Residual Pseudo-likelihood (RSPL) through sampling with 10 Groups.

	True Parameter	Bootstrap		RSPL	
		Estimate (S.E)	Coverage Probability	Estimate (S.E)	Coverage Probability
Intercept	0.306	0.198 (0.950)	95,2%	0.238 (0.966)	95.9%
Fixed effect of x_1	-1.125	-1.142 (1.553)	97.6%	-1.154 (1.288)	92.5%
Cross-level Interaction Effect of x_1*z_1	1.146	1.332 (2.184)	96.9%	1.245 (1.739)	90.7%

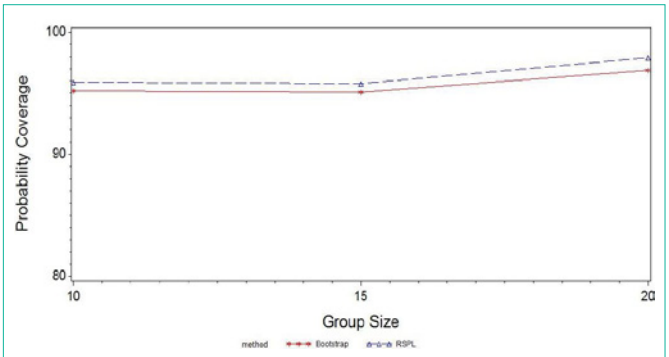

Figure 1: Probability coverage of 95% CI of intercept.

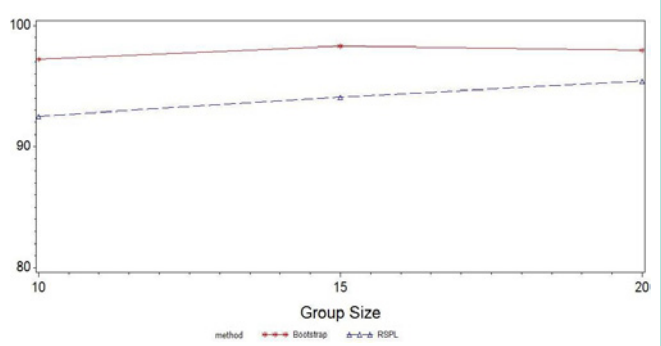

Figure 2: Probability coverage of 95% CI of X1.

corresponding "true" coefficient, the deviation would be considered bias. The extent of bias was evaluated by coverage probability of the 95% C.I.

To estimate the coverage probabilities for parameter estimates, we repeated the sampling and modeling process 500 times, and the percentage of the 95% CIs for each fixed coefficient estimate covering the corresponding "true" parameter was calculated for each model. If a parameter's standard error is unbiased, we would expect its coverage probability to be at least 0.95, otherwise the standard error is downward biased.

Results

Table 1 shows the model results using the sub-dataset with only 10 groups (level-2 units). As it can be seen, the fixed effect of x_1 and the cross-level interaction effect of x_1*$z_1$estimated from bootstrap or RSPL are similar to each other and close to the corresponding "true" coefficients. However, the coefficients estimated from bootstrap approach had better coverage probabilities than those estimated from RSPL: the coverage probabilities of both fixed effect of x_1 and cross-level interaction effect of x_1*z_1 were all greater than 0.95 for bootstrap approach, but less than 0.95 for RSPL.

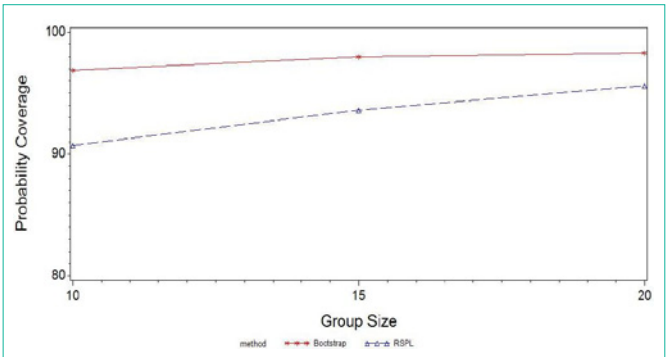

Figure 3: Probability coverage of 95% CI of X1*Z1.

Two more sub-datasets with 15 and 20 groups (Level-2 units), respectively, were randomly selected from the simulated dataset, and the coverage probabilities of the multilevel model regression coefficients were estimated using both bootstrapping and RSPL for each of the datasets using the aforementioned approach. Figure 1 shows the coverage probability curve for intercept coefficient, while Figure 2 for fixed effect of x_1, and Figure 3 for the cross-level interaction effect of x_1*z_1 by number of groups. The coverage probability of 95% CI for the intercept coefficient was high (>0.95) and was only slightly better with an increase of number of groups for both the bootstrap and RSPL approaches. With respect to the fixed effect and cross-level interaction, the coverage probabilities were all greater than 0.95 for bootstrap approach, but less than 0.95 for RSPL when the number of groups was small. The difference between the two modeling approaches tended to diminish when number of groups increases. Our findings provide evidence that non-parametric residual bootstrap approach works better than RSPL for modeling multilevel data with binary response variable when the number of groups is small.

Discussion

In application of multilevel modeling, researchers often encounter a challenge to analyze data when the number of groups (higher level units or clusters) is small, since the standard errors of parameter estimates are biased downward if without correction. When a response variable is continuous, residual bootstrapping technique has been applied for bias correction caused by the small number of groups [12,14]. A SAS macro has been developed by Wang and collegues for non-parametric residual multilevel modling [13]. However, such an innovative analytical approach and corresponding computer programs are only applicable to multilevel modeling with continuous response variables. When the response variable is binary, application of the residual bootstrapping is not straightforward because the traditional logistic regression does not have a residual term. The level-1 residuals can be obtained through deducting the residuals

from the level-2 residuals. In the present study, we demonstrated how to conduct nonparametric residual bootstrap multilevel logit modeling using simulated data. Our results provide evidence that the nonparametric residual bootstrap approach produces more accurate parameter estimates than the RSPL (the default estimator in SAS Proc GLIMIX) for modeling hierarchically distributed data when the number of groups is small. Furthermore, the difference between the two estimation approaches diminished when the number of groups increased. Our SAS macro has been developed by the authors for nonparametric residual bootstrap multilevel logit model and the macro is available.

References

1. Claassen EA. A Reduced Bias Method of Estimating Variance Components in Generalized Linear Mixed Models. 2014.
2. Pinheiro JC, Chao EC. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. Journal of Computational and Graphical Statistics. 2006; 15: 58-81.
3. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American statistical Association. 1993; 88: 9-25.
4. Breslow NE, Lin X. Bias correction in generalized linear mixed models with a single component of dispersion. Biometrika. 1995; 82: 81-91.
5. Couton J, Stroup W. On the small sample behavior of generalized linear mixed models with complex experiments. CRC Press. 2013.
6. Stroup WW. Generalized linear mixed models: modern concepts, methods and applications. Proceedings of the 25th Conference Applied Statistics in Agriculture. 2013.
7. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. Biometrika. 1971; 58: 545-554.
8. Wolfinger R, O'connell M. Generalized linear mixed models a pseudo-likelihood approach. Journal of statistical Computation and Simulation. 1993; 48: 233-243.
9. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993; 80: 27-38.
10. Chernick MR. The jackknife: a resampling method with connections to the bootstrap. Wiley Interdisciplinary Reviews: Computational Statistics 2012; 4: 224-226.
11. Quenouille MH. Notes on bias in estimation. Biometrika. 1956; 43: 353-360.
12. Carpenter JR, Goldstein H, Rasbash J. A novel bootstrap procedure for assessing the relationship between class size and achievement. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2003; 52: 431-443.
13. Wang J, Carpenter JR, Kepler MA. Using SAS to conduct nonparametric residual bootstrap multilevel modeling with a small number of groups. Computer methods and programs in biomedicine. 2006; 82: 130-143.
14. Maas CJM, Hox JJ. Robustness issues in multilevel regression analysis. Statistics Neerlandica. 2004; 58: 127-137.

Austin Biom and Biostat - Volume 4 Issue 1 - 2017
Submit your Manuscript | www.austinpublishinggroup.com

Citation: Wang Y, Song L and Wang J. Implementing Nonparametric Residual Bootstrap Multilevel Logit Model with Small Number of Level-2 Units. Austin Biom and Biostat. 2017; 4(1): 1034.