SciMedCentral

Annals of Biometrics & Biostatistics

Research Article

A Monte Carlo Study of Multilevel Latent Class Regression

Luohua Jiang[1]* and Shuai Chen[2]

[1]*Department of Epidemiology and Biostatistics, School of Rural Public Health, Texas A&M Health Science Center, College Station, Texas, USA*

[2]*Department of Statistics, Texas A&M University, Collect Station, Texas, USA*

Corresponding author

Luohua Jiang, Department of Epidemiology and Biostatistics, Texas A&M Health Science Center, 219 SRPH Administration Building, TAMU 1266 College Station, TX 77843-1266, USA; Tel: 979-845-2675; Fax: 979-845-1877; Email: ljiang@srph.tamhsc.edu

Submitted: 28 June 2013

Accepted: 03 July 2013

Published: 07 July 2013

Keywords

- Latent class analysis
- Multilevel models
- Correlated random intercepts

Abstract

Multilevel latent class analysis (MLCA) has been built into a few statistical software to analyze nested data that do not satisfy the conditional independence assumption of simple Latent class analysis (LCA). Multilevel latent class regression (MLCR) is also available in those software packages to analyze the relationships between latent class membership and covariates. The impact of using simple latent class regression (SLCR) instead of MLCR for nested data has not been investigated empirically. In this study, we conduct Monte Carlo simulations to examine the influence of intra-class correlation (ICC) on the estimation bias and coverage of regression coefficients using MLCR vs. SLCR. We also evaluate the consequences of assuming perfect correlation when the random intercepts are actually not perfectly correlated. The results indicate that, with the increase of ICC, the biases in regression coefficients increased while the coverage probabilities decreased. Further, we find that the bias caused by the misspecification of perfect correlation assumption in MLCR estimation was slight, especially when the ICC was low. Thus, MLCR with perfect correlation might be a computationally efficient method without substantial loss of accuracy, and hence could be a reasonable substitute for MLCR procedure with ordinary correlation when computation burden is a concern.

INTRODUCTION

Latent class analysis (LCA) is a widely used statistical method in many fields. It assumes the subjects belong to some latent subgroups, referred as latent classes. Although the classes are not directly observed, they can be inferred from a set of observed categorical variables using LCA. Further, the relationships between latent class membership and covariates may be assessed using latent class regression (LCR). In simple LCA, it is assumed that subjects are independent conditional on the latent class membership. However, if the subjects are clustered in groups, the conditional independence assumption may not be met. Thus, multilevel techniques are in need to incorporate the intra-cluster dependence for those types of data.

In recent years, multilevel latent class analysis (MLCA) has been developed by a few groups [1-3] to apply LCA with nested data. They also extended MLCA to include Level 1 and Level 2 covariates in the model [1,3]. In multilevel LCR (MLCR), a two-level multinomial logistic regression is adopted, by introducing random intercepts across Level 2 (cluster) units. When the number of latent classes C is more than 2, the C-1 random intercepts are allowed to be correlated with one another. However, the computational burden of this model grows exponentially with C. Thus, Vermunt [1] suggested to model all the random intercepts using a common factor, which assumes the random intercepts are perfectly correlated.

To the best knowledge of the authors, the impact of using simple latent class regression (SLCR) instead of MLCR for nested data has not been investigated empirically. In this article, we present a Monte Carlo simulation study to examine the influence of intra-class correlation (ICC) on the estimation bias and coverage of regression coefficients using MLCR vs. SLCR. We also evaluate the consequences of assuming perfect correlation when the random intercepts are actually not perfectly correlated. More specifically, the performance of SLCR, MLCR with perfect correlation (MLCR-P), and MLCR with ordinary correlation (MLCR-O) are compared in a MLCR model with 3 latent classes.

MATERIALS AND METHODS

SLCR

In a SLCR model, let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iM})'$ be an observed response vector for the ith individual, where variable Y_{im} takes possible values 1, 2, ..., r_m, and c_i = 1,2, ..., C denote the latent class membership of the ith individual. Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ be a vector of explanatory variables for the ith individual. A LCR model can

be expressed as:

$$\Pr(\mathbf{Y}_1=\mathbf{y}_1,\ldots,\mathbf{Y}_n=\mathbf{y}_n)=\prod_{i=1}^{n}\sum_{c=1}^{C}\gamma_c(\boldsymbol{x}_i)\prod_{m=1}^{M}\prod_{k=1}^{r_m}\rho_{mk|c}^{I(y_{im}=k)},$$

where $\rho_{mk|c}$ = Pr ($Y_{im} = k \mid c_i = c$) is the conditional probability of response k to the mth item given class membership c; $\gamma_c(\boldsymbol{x}_i)$ is the class membership probabilities for the cth class given $\boldsymbol{x}_i$, which is related to γ_c through a multinomial logistic regression:

$$\gamma_c(\boldsymbol{x}_i)=\Pr(c_i=c|\boldsymbol{x}_i)=\frac{\exp(\boldsymbol{x}_i'\beta_c)}{1+\sum_{j=1}^{C-1}\exp(\boldsymbol{x}_i'\beta_j)},$$

where $\boldsymbol{\beta}_c$ is a vector of logistic regression coefficients, with the reference class as C.

MLCR

If we denote c_{ij} as the class membership of the ith individual coming from the jth cluster, $\gamma_{jc}(x_{ij},w_j)$ as the probability that c_{ij} =c with level 1 covariate x_{ij} and level 2 covariate w_j, assuming random intercept $u_{jc} \sim N(0, 1)$, a MLCR model can be written as:

Level 1 (individual):

$$\gamma_{jc}(x_{ij},w_j)=\Pr(c_{ij}=c|x_{ij},w_j)=\frac{\exp(\beta_{0jc}+\beta_{1c}x_{ij})}{1+\sum_{k=1}^{C-1}\exp(\beta_{0jk}+\beta_{1k}x_{ij})}$$

Level 2 (cluster): $\beta_{0jc}=\alpha_{0c}+\alpha_{1c}w_j+\sigma_c u_{jc}$.

The intra-class correlation (ICC) for class c in MLCR is defined as the proportion of the variance of the random effects out of the total variance, i. e. , i.e., $r_c=\frac{\sigma_c^2}{\sigma_c^2+\pi^2/3}$ [1].

When C is more than 2, MLCR-O allows the C-1 random intercepts u_{jc} to be correlated with one another. On the other hand, the MLCR-P uses a common factor to model all the random intercepts (i. e. , $u_{jc}= u_j$ for all c), which assumes the random intercepts are perfectly correlated.

Monte carlo simulation

In the Monte Carlo simulation study, we generated data from a 3-class MLCR model using the R 2.15. 2 package. The observed vector, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i3})'$, have 3 categorical variables, each of which has 5 categories. The model includes 2 covariate variables. One covariate is a Level 1 continuous variable with standard normal distribution, and the other is a Level 2 binary covariate. We assign Class 3 as the reference class, i. e. , $\alpha_{03} = \alpha_{13} = \beta_{13} = \sigma_3 = 0$. We set the regression parameters as $\alpha_{01} = \alpha_{02} = -1$, $\alpha_{11} = 0.5$, $\alpha_{12} = -0.5$, $\beta_{11} = 0.5$, $\beta_{12} = -0.5$. The random intercept u_{j1} is correlated with u_{j2} through a bivariate normal distribution:

$$\begin{pmatrix}u_{j1}\\u_{j2}\end{pmatrix}\sim BVN\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right).$$

In addition we set different values of $\sigma_1 = \sigma_2 = \sigma$ to obtain ICCs at various levels. A value of σ = 0.416 generates data with ICC of 0.05, a value of σ = 0.6 generates data with ICC of 0.1, and a value of σ = 1 gives us data with ICC of 0.25. Finally, the conditional item response probabilities for m = 1,2,3 were chosen as:

$$(\rho_{m1|1},\rho_{m2|1},\rho_{m3|1},\rho_{m4|1},\rho_{m5|1}) = (0.05,\ 0.8,\ 0.05,\ 0.05,\ 0.05),$$

$$(\rho_{m1|2},\rho_{m2|2},\rho,_{m3|2},\rho_{m4|2},\rho_{m5|2}) = (0.05,\ 0.05,\ 0.1,\ 0.4,\ 0.4)$$

$$(\rho_{m1|3},\rho_{m2|3},\rho_{m3|3},\rho_{m4|3},\rho_{m5|3}) = (0.3,\ 0.3,\ 0.3,\ 0.05,\ 0.05).$$

We applied the SLCR, MLCR-P and MLCR-O approaches to estimate the regression coefficients of the two covariates using Mplus 7 software package [4]. Since MLCR-O adopts the true model, we expect MLCR-O would perform the best. We generated 500 replications of 3000 subjects with different ICCs. The 3000 subjects were grouped into 30 or 300 equally sized clusters, half of which were assigned with 0 for the Level 2 covariate, and the other half were assigned with 1.

RESULTS AND DISCUSSION

Table 1 presents the biases and 95% confidence interval coverage rates for the estimates of regression coefficients using different models. It clearly shows that SLCR had the largest bias and worst confidence interval coverage among the three methods, especially for the Level 2 covariate. With the increase of ICC, the biases in regression coefficients increased while the coverage probabilities decreased. SLCR also had worse performance

Table 1: Summary of the relative bias $\left(\frac{\hat{\theta}-\theta}{\theta}\right)$ and 95% confidence interval coverage for the regression coefficients.

# of groups	Group Size	ICC	Class	Coefficient	Relative Bias			95% CI Coverage Rate		
					SLCR	MLCR-P	MLCR-O	SLCR	MLCR-P	MLCR-O
30	100	0.05	**Class 1**	Level 1 Covariate	-.02	.00	.00	.93	.94	.95
				Level 2 Covariate	-.03	.02	-.03	.69	.94	.94
				Intercept	.03	-.00	.01	.84	.93	.94
			Class 2	Level 1 Covariate	.01	.00	-.01	.96	.95	.94
				Level 2 Covariate	-.01	-.01	-.01	.87	.94	.93
				Intercept	.02	.03	.01	.93	.94	.93
30	100	0.1	**Class 1**	Level 1 Covariate	-.05	-.00	.00	.92	.93	.93
				Level 2 Covariate	-.06	-01	.01	.59	.94	.95
				Intercept	.06	.00	-.00	.70	.94	.92
			Class 2	Level 1 Covariate	.01	.02	.01	.94	.93	.95

				Level 2 Covariate	-.02	-.02	-.01	.76	.94	.94
				Intercept	.05	.04	.00	.85	.91	.92
30	100	0.25	**Class 1**	Level 1 Covariate	-.14	-.01	.00	.68	.92	.93
				Level 2 Covariate	-.13	.05	-.00	.44	.93	.94
				Intercept	.13	-.00	.02	.51	.93	.91
			Class 2	Level 1 Covariate	.05	.09	.00	.93	.90	.95
				Level 2 Covariate	.09	.12	.03	.66	.92	.94
				Intercept	.11	.13	-.01	.75	.88	.93
300	10	0.05	**Class 1**	Level 1 Covariate	-.03	.00	-.00	.94	.95	.95
				Level 2 Covariate	-.05	-.01	-.01	.92	.96	.94
				Intercept	.04	-.01	.01	.94	.96	.96
			Class 2	Level 1 Covariate	.00	.00	-.01	.96	.95	.95
				Level 2 Covariate	.01	.01	-.01	.93	.94	.96
				Intercept	.03	.01	-.00	.95	.96	.97
300	10	0.1	**Class 1**	Level 1 Covariate	-.06	.01	.01	.92	.95	.94
				Level 2 Covariate	-.07	-.00	.01	.87	.95	.95
				Intercept	.06	-.01	-.00	.90	.95	.97
			Class 2	Level 1 Covariate	.02	.04	-.02	.93	.93	.94
				Level 2 Covariate	.02	.02	.01	.95	.95	.95
				Intercept	.04	.04	.01	.94	.95	.93
300	10	0.25	**Class 1**	Level 1 Covariate	-.16	-.01	.01	.63	.95	.95
				Level 2 Covariate	-.16	-.01	-.01	.75	.96	.96
				Intercept	.16	.02	-.00	.63	.96	.94
			Class 2	Level 1 Covariate	.04	.10	.01	.95	.92	.95
				Level 2 Covariate	.04	.08	-.00	.93	.95	.96
				Intercept	.15	.18	-.01	.87	.87	.94

when only a limited number of clusters were available in the data comparing to the scenario when a large number of clusters were collected. In general, these results are consistent with the simulation results of multilevel logistic regression models [5], suggesting the importance of using multilevel analysis techniques when you have clustered/correlated data that do not satisfy the conditional independence assumption, especially when the regression coefficients for level 2 covariates are of interest.

When the MLCR-P procedure was compared to the MLCR-O procedure, it appears that the loss of efficiency was not substantial especially when ICC was low. Even for the cases with an ICC of 0.25, MLCR-P was only slightly worse than MLCR-O in terms of biases and coverage rates. Meanwhile, MLCR-P procedure took much less computation time than MLCR-O (2.5 minutes vs. 30 minutes for each simulated dataset on a PC with CPU of Intel i5 2.40GHz).

CONCLUSIONS

Consistent with previous studies of multilevel logistic regression models [5] and multilevel data analysis techniques broadly [6], the results of this empirical study demonstrate the importance of using multilevel regressions, more specifically, MLCR when performing LCR for clustered/correlated data structure.

When the number of latent classes (C) is more than 2, the computational complexity of the estimation procedure for MLCR-O increases rapidly with the increase of C. To alleviate the computational intensity and reduce computation time, the perfect correlation assumption for the random intercepts may be adopted to use MLCR-P in those situations. However, attention should be paid to the possible bias brought by such misspecification. Based on our Monte Carlo simulation results, we conclude that the bias caused by the misspecification of perfect correlation among random intercepts in MLCR-P model estimation is slight, especially when the ICC is low. Therefore, MLCR-P might serve as a computationally efficient method without substantial loss of accuracy in parameter estimates, hence could be a reasonable substitute for MLCR-O procedure when computation burden is a concern.

ACKNOWLEDGEMENTS

Manuscript preparation was supported in part by American Diabetes Association (ADA #7-12-CT-36, L. Jiang).

REFERENCES

1. Vermunt JK. Multilevel latent class models. Sociol Methodol. 2003; 33: 213-39.

2. Vermunt JK Latent class and finite mixture models for multilevel data sets. Stat Methods Med Res 2008; 17: 33-51.

3. Asparouhov T, Muthén, BO. Multilevel mixture models. In Hancock GR & Samuelsen KM, editors. Advances in latent variable mixture models. Charlotte, NC: Information Age. 2008; 27-51.

4. Muthén LK, Muthén BO. Mplus User's Guide. 7nd ed. Los Angeles, CA: Muthen & Muthen; 2012.

5. Moineddin R, Matheson FI, Glazier RH A simulation study of sample size for multilevel logistic regression models. BMC Med Res Methodol 2007; 7: 34.

6. Hox J. Multilevel analysis: techniques and appliations. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Publishers; 2010.

Cite this article

Jiang L, Chen S (2013) A Monte Carlo Study of Multilevel Latent Class Regression. Ann Biom Biostat 1: 1004.