Paper PO-14

Fitting multivariate random-effects models using SAS® PROC GLIMMIX

Lei Li, RTI International, Statistics and Epidemiology Unit, RTP, North Carolina

Namasivavayam Ambalavanan, University of Alabama at Birmingham, Department of Pediatrics, Birmingham, Alabama

Abhik Das, RTI International, Statistics and Epidemiology Unit, Rockville, Maryland

ABSTRACT

Multinomial ordinal data arise when measures of an outcome are scaled into ordered categories. The ordinal data can be analyzed using multinomial logit models (Agresti 2002). But the analysis is often complicated by the clustering nature of the data. One approach to handling intra-cluster correlation is logistic regression using a random intercept. Another approach is using random-intercept multivariate logit model, in which multiple logits of the outcomes are assumed to share a common random intercept.

The multivariate modeling approach can be further extended by incorporating a vector of Multivariate Normal random intercepts to allow more flexibility in the correlation structure among multivariate outcomes. In this paper we describe a continuation-ratio logit model with Multivariate Normal random intercepts and use SAS PROC GLIMMIX to fit the model to survival data among premature infants admitted into hospital neonatal intensive care units.

INTRODUCTION

Multinomial ordinal data arise when measures of outcomes are scaled into ordered categories and multinomial logit models can be used for the analysis of the data [1]. In neonatal clinical research, data collected are often multinomial ordinal, taking the form of presence or absence of an outcome of interest at a number of progressive time points. For example, the clinical investigation that motivated this work concerns with mortality and morbidity of extremely preterm infants in the first few hours after birh (i.e., 12 or 24 hours) and then afterwards until 120 days. The multinomial data to be analyzed are the numbers of deaths or incidences of morbidity in two consecutive time periods, less than 12 or 24 hours after birth, 12 or 24 hours – 120 days, and the number of survivors after 120 days. As frequently encountered in clinical research, the analysis of such data is complicated by the clustering nature of the data. The study recruited infants from multiple medical centers and large differences in mortality and morbidity among the centers have been observed (Vohr and Wright et al. 2004, Cotton and Oh et al. 2005, Walsh and Laptook 2007). Even after accounting for center differences in confounding factors such as birth weight, gestational age and antenatal steroid use, these differences persisted (Tyson and Younes et al. 1996).

One approach to modeling outcome data from multiple centers is logistic regression that uses a random intercept to represent extra variation among centers. But taking this approach of logistic regression with a random intercept, one will need to dichotomize the multinomial outcome and fit separate models for each recoded binary outcome. Another approach is random-intercept multivariate logit model, in which logits of the outcomes are assumed to share a common random intercept (Das, Poole and Bada 2004). Although this approach is statistically more efficient than fitting separate models for each outcome, it is not well suited for multinomial ordinal data such as the aforementioned infant mortality or morbidity data. It may be realistic to assume a common random intercept where the multiple outcomes may measure the same underlying phenomenon. But the magnitude of intra-center variation of mortality or morbidity can be very different over a number of progressive time periods. Furthermore, the assumption of a shared common random intercept prevents the estimation of the covariance among the multinomial outcomes. It can be of clinical importance to assess whether a center's performance on mortality and morbidity outcomes in an early period is predictive of its performance in a late period and identify factors associated with the center variations.

Coull and Agresti (2000) extended the multivariate approach by incorporating a vector of multivariate normal random intercepts in the continuation-ratio logit models for the analysis of categorical data. This model is a special case of the multivariate generalized linear mixed models and is not limited to the analysis of categorical data. In this paper, we describe the continuation-ratio logit model with multivariate normal random intercepts and apply it to model deaths among very premature infants admitted to hospital neonatal intensive care units (NICU) as examples. We will illustrate the modeling process using PROC GLIMMIX in the recently released SAS® 9.2 (SAS Institute 2009). The estimation results from logistic regressions with random intercepts are also obtained for comparison.

STATISTICAL MODEL AND DATA

MODEL

Suppose that the outcome of interest may fall into one of the J ordinal categories. Let the multinomial response probability for the *j*th category be p_j, j=1, 2, … , J. The corresponding probability conditional on the response falling into the *j*th category or afterwards is

$$\pi_j = p_j / (p_j + \ldots + p_J).$$

The continuation-ratio logit for the *j*th category, j=1, … , J-1, is defined as the logit of this conditional probability, that is,

$$\text{logit}(\pi_j) = \log(p_j / (p_{j+1} + \ldots + p_J)).$$

Although other forms of logit can be defined and modeled for an ordinal multinomial outcome, the use of the continuation-ratio logit is convenient in that a multinomial logit model can often be estimated by separately fitting a binary logistic regression model for each logit(π_j). The models can be estimated and interpreted in ways more familiar to epidemiologists.

The formulation of a multivariate continuation-ratio logit model for clustered data in the settings of multi-center studies, experiments on animal litters, or longitudinal repeated measurement studies features a cluster-specific vector of random intercepts [7]. Specifically, denote the vector of the continuation-ratio logits for a cluster or individual s by $\boldsymbol{\pi}_s = (\pi_{s1}, \pi_{s2}, \ldots, \pi_{s(J-1)})$, s=1,…,N, the model is of the form

$$\text{logit}(\boldsymbol{\pi}_s) = \boldsymbol{\alpha}_s + \mathbf{X}_s\boldsymbol{\beta}$$

where $\boldsymbol{\alpha}_s$ is a vector of J-1 random intercepts and assumed to follow a multivariate Normal distribution, MN($\mathbf{0}$,$\boldsymbol{\Sigma}$), and $\boldsymbol{\beta}$ is a vector of coefficients for the covariate matrix $\mathbf{X}_s$ representing fixed effects to be estimated. The variance-covariance matrix $\boldsymbol{\Sigma}$ specifies the correlation structure among the continuation-ratio logits.

The estimation of this model usually involves complex multidimensional integral approximation, and computational techniques such as multiple quadrature or Monte Carlo EM algorithms may be used. Because this model is a special case of the multivariate generalized linear mixed models, we used PROC GLIMMIX to estimate the model parameters. The model estimation methods are based on restricted pseudo-likelihood in PROC GLIMMIX.

DATA AND MODEL FITTING

The Neonatal Research Network (NRN) of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) of the U.S. National Institutes of Health supports a registry of infants between 22 0/7 to 28 6/7 weeks (<29 weeks) inclusive gestational age or between 401 grams to 1000 grams inclusive birth weight admitted to the NICUs of participating centers. The data used for our example analysis were taken from the NRN registry. All inborn infants during 1998 and 2007 with birth weights of 401-1000 grams in the registry were included in our analysis. One clinical center was excluded due to small number of infants. The outcome of interest was the early death in the first 24 hours after birth and the late death between the first 24 hours and 120 days. Identifying predictors of neonatal survival has been of great interests to clinicians who desire to make evidence-based decisions on clinical management (Tyson and Parik et al. 2008, Higgins, Delivorial-Papadopoulos and Raju 2005, Ambalavanan and Carlo 2001, 2005). Two major research questions are to be addressed. First, the effects of potential predictors of infant survival will be estimated and the differences in the estimated effects on the early death and the late death will be tested. Second, the variance-covariance structure between a center's early death rate and

its late death rate will be assessed and the correlation will be tested. The potential predictors of death included neonatal characteristics such as gestational age, birth weight, sex and race, pregnancy complications such as multiple birth and antepartum hemorrhage, and labor and delivery information such as use of antenatal steroid and antibiotics and C-section. In addition, the effect of respiratory support at age 24 hours, which is unlike other predictors and available only among infants who survived beyond the first 24 hours, was examined.

The input data file must be properly prepared to make use of the multivariate modeling feature of PROC GLIMMIX, one record corresponding to each dependent variable for an individual. In the case of our analysis of the early death and the late death, the dependent variables are the continuation-ratio logits, logit(π_1), the log-odds of the early death, and logit(π_2), the log-odds of the late death conditional on survival during the first 24 hours. We created a binary dependent variable of death and an indicator of the time period corresponding to the dependent variable in the input data to represent the two logits. All infants had records for the dependent variable with its value being logit(π_1) and the value of the indicator being the first 24 hours, and infants who lived longer than 24 hours also had records for the dependent variable with its value being logit(π_2) and the value of the indicator being the 24 hours - 120 days.

The indicator of the time period was treated as a class random-effect variable in our models. Various forms of the variance-covariance matrix available in the PROC GLIMMIX were specified for them to model the correlation structure between the continuation-ratio logits. Because the coefficient vector for the predictors represents fixed effects common for the two logits, the interaction terms between the indicator of the time period and the predictors were also included in the model to obtain the fixed-effects estimates unique to each of the two logits. We also included the predictor, respiratory support at age 24 hours, in our models. Because data on this predictor were collected at the 24 hours of life, the value of this predictor was set to be zero in all records for the values of the dependent variable corresponding to logit(π_1) in the input data and only its effects on logit(π_2) were estimated.

Sample SAS Code

```
Data modelingdata;

Set babies; /* one row for each baby */
response = death24;
respsuppt1=0; /* a covariate available only after 24th hour */
dthtime="Early(<24hours)        "; /* indicator of time period */
output;
response = death24_120;
 respsuppt1=respsuppt;
dthtime="Late(24 hrs - 120 days)";
if death24_120>. then output;

keep center patientid response dthtime
gestage gestage_rc birthwt bwt100 wtle700g gwksle24 antibio aphemor
mhyper edu_hs male antester apgar5_4 multbrth csect intub white
respsuppt1;

run;

%macro joint_m(cs, catcov, predictors);
proc glimmix data=modelingdata pconv=.0000002 ic=Q;
class center dthtime &catcov;
model response (event="1") = dthtime &predictors
                           / noint dist=binary solution;
random dthtime / subject=center type=&cs g;
*nloptions absfconv=.0001;
run;
%mend;
```

We produced estimation results from a number of models. First, we fitted two separate logistic regression models with random intercepts by the medical center for the early death and for the late death excluding infants who died in the first 24 hours. Second, we jointly modeled the early death and the late death by fitting the multivariate random-effects continuation-ratio logit models. The forms of variance-covariance matrix for the random intercepts specified

included variance components, compound symmetry, main diagonal and unstructured. And finally, we added the predictor, respiratory support at age 24 hours, in the multivariate random-effects continuation-ratio logit models.

RESULTS

The frequency distributions of the infant deaths and the predictors are summarized in Table 1. There are about equal numbers of the early deaths and the late deaths in the data.

The estimated coefficients for the predictors and variance-covariance estimates are summarized in Table 2. The multivariate logit model and the two univariate logistic regression models with random intercepts show very similar results. The estimated coefficients are close and their standard error estimates are essentially identical. It can be seen that the predictors have different effects on the early death than on the late death in terms of the magnitude of strength and level of significance. For example, the effect of gestational age is much stronger on the early death than on the late death. The effect of antepartum hemorrhage is significant on the early death but not on the late death. Multiple birth and intubation have significant and negative effects on the early death, reducing the log-odds of the early death, but they show positive effects on the late death.

The estimated variance-covariance matrix for the random intercepts in the multivariate random-effects logit model shown was obtained under the unstructured form. The estimated variances are .23 for the early period and .19 for the late period, and the ratios to their standard error estimates are much greater than two. Thus there are strong clustering effects within medical centers of deaths and this effect appears stronger during the first 24 hours of life than during the 24 hours and 120 days. The estimated covariance between the early period and the late period, however, is small and the ratio to its standard error estimate is less than one, suggesting that the chance of the late death in a medical center (given survival beyond the early period) is not correlated with the chance of the early death. This also implies that the multivariate random-effects logit model should be essentially equivalent to the two univariate logistic regression models with random intercepts in this case. A comparison of -2log-likelihood values calculated for these models, also shown in the table, suggests that the multivariate model fits the data more adequately though.

Table 3 shows the estimation results from the multivariate random-effects logit model that included an additional predictor, respiratory support at 24 hours. As indicated in the estimated variance-covariance matrix in the previous multivariate random-effects logit model, a diagonal variance-covariance matrix for the random intercepts was specified. The respiratory support at 24 hours has a strong effect on the late death.

CONCLUSION

The multivariate models with random intercepts used in this paper assumed multivariate Normal distributed random effects. The extension from the univariate random effects to multivariate random effects afforded us the flexibility in the specification of the correlation structure among multiple responses. This extension is useful when the correlations among responses are of substantive research interests or are believed to be negative. With improved SAS PROC GLIMMIX the models can be readily estimated and well-suited for the analysis of clustered multinomial ordinal data.

One issue we encountered in using PROC GLIMMIX concerns with the stability of the computational algorithm. Because the numerical estimation procedure involves two loops of iteration process, the default convergence criteria sometimes can not be met for the variance-covariance estimates between the two loops and thus the algorithm is terminated without yielding final parameter estimates. The default value of the convergence criteria has to be decreased to overcome this.

REFERENCES

Agresti, A. Categorical data analysis. 2nd edition. 2002. John Wiley and Sons, Inc., New Jersey.

SAS Institute. SAS/STAT 9.2 User's Guide. SAS Institute: Cary, NC, 2009.

Vohr, B.R., L.L. Wright, et al., Center differences and outcomes of extremely low birth weight infants. Pediatrics 2004. 113: 781-9.

Cotten, C.M., W. Oh, et al., Prolonged hospital stay for extremely premature infants: risk factors, center differences, and the impact of mortality on selecting a best-performing center. J. Peerinatol. 2005. 25: 650-5.

Walsh, M., A. Laptook, et al., A cluster-randomized trial of benchmarking and multimodal quality improvement to improve rates of survival free of bronchopulmonary dysplasia for infants with birth weights of less than 1250 grams. Pediatrics 2007. 119: 876-90.

Tyson, J.E., N. Younes, et al., Viability, morbidity, and resource use among newborns of 501- to 800-g birth weight. National Institute of Child Health and Human Development Neonatal Research Network. Jama 1996. 276: 1645-51.

Das, A., Poole, W. K. and H. S. Bada, A repeated measures approach for simultaneous modeling of multiple neurobehavioral outcomes in newborns exposed to cocaine in utero. American Journal of Epidemiology 2004. 159: 891-899.

Coull, B. A. and A Agresti. Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. Biometrics 2000. 56: 73-80.

Tyson, J.E., N.A. Parikh, et al., Intensive care for extreme prematurity--moving beyond gestational age. N Engl J Med 2008. 358: 1672-81.

Higgins, R.D., M. Delivoria-Papadopoulos, and T.N. Raju, Executive summary of the workshop on the border of viability. Pediatrics 2005. 115: 1392-6.

Ambalavanan, N. and W.A. Carlo, Comparison of the prediction of extremely low birth weight neonatal mortality by regression analysis and by neural networks. Early Hum Dev 2001. 65: 123-37.

Ambalavanan, N., W.A. Carlo, et al., Prediction of death for extremely low birth weight neonates. Pediatrics 2005. 116: 1367-73.

ACKNOWLEDGMENTS

This work was done on behalf of the National Institute of Child Health and Development Neonatal Research Network (NICHD NRN).

CONTACT INFORMATION

Lei Li, Ph.D.
Statistics and Epidemiology Unit
RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709
Phone: (919)541-8821
Fax: (919)541-5966
Email: lei@rti.org

Table 1. In-born infants with birth weights of 401 to 1000 grams, NICHD Neonatal Network, 1998-2007 (n=15422).

Variable[1]		n	%
Deaths	Died <=24 hours	2533	16.4
	Died 24 hours-120 days	2566	16.6
Gestational age in weeks	<=23	2869	18.6
	24	2584	16.8
	25	2729	17.7
	26	2550	16.5
	27	1914	12.4
	>=28	2774	18.0
Birth weight in grams	400-499	1348	8.7
	500-599	2385	15.5
	600-699	2918	18.9
	700-799	2833	18.4
	800-899	2877	18.7
	900-1000	3061	19.9
Gender	Female	7713	50.0
	Male	7709	50.0
Race	White	8051	52.2
	Others	7310	47.4
Mother's education[2]	Less than high school	2037	13.2
	High school or more	4647	30.1
	Not available	8738	56.7
Multiple birth		3764	24.4
C-section		8888	57.6
Intubation at birth		10554	68.4
Use of antenatal steroids		11643	75.5
Maternal antibiotics		10037	65.1
Antepartum hemmorrage		2676	17.4
Maternal hypertension		3907	25.3
Respiratory support at 24 hours[2]		11821	76.7

Notes:
[1] Missing data: 2 on gestational age, 35 on antibiotics, 13 on hemorrage, 13 on hypertension, 29 on antenatal steroids, 1 on multiple birth, 14 on C-section, 10 on intubation, 61 on race;
[2] Respiratory support form required only for infants survived > 24 hours.

Table 2. Estimation results from two separate logistic regression models with random intercepts of early death (<=24 hours) and late death (24 hours - 120 days) and the results from the multivariate random-effects continuation-ratio logit model

Covariate	logistic regressions with random intercepts				Multivariate random-effects logit model			
	Died <=24 hours		24 hours - 120 days		Died <=24 hours		24 hours - 120 days	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept	-2.2253	0.2221	-3.0989	0.2096	-2.2300	0.2221	-3.0955	0.2095
Gestational age: <=23 weeks	2.2535	0.1667	0.8833	0.1246	2.2553	0.1667	0.8860	0.1246
24 weeks	1.2026	0.1592	0.6019	0.1048	1.2030	0.1592	0.6030	0.1048
25 weeks	0.6190	0.1602	0.2499	0.0998	0.6186	0.1602	0.2508	0.0998
26 weeks	0.0953	0.1709	0.0470	0.1012	0.0950	0.1709	0.0461	0.1012
27 weeks	-0.2280	0.1999	0.0212	0.1076	-0.2286	0.1999	0.0206	0.1076
Birth weights: 401-499 grams	2.4843	0.1781	2.5125	0.1333	2.4839	0.1780	2.5139	0.1333
500-599 grams	1.7083	0.1651	1.7240	0.1073	1.7079	0.1651	1.7247	0.1073
600-699 grams	1.0134	0.1620	1.1614	0.0983	1.0123	0.1620	1.1620	0.0983
700-799 grams	0.5939	0.1639	0.6785	0.0966	0.5935	0.1639	0.6789	0.0966
800-899 grams	0.3779	0.1645	0.1999	0.0991	0.3773	0.1645	0.2000	0.0991
Gender: male	0.4108	0.0674	0.4458	0.0502	0.4107	0.0674	0.4459	0.0502
Race: white	0.3201	0.0728	0.1870	0.0536	0.3217	0.0728	0.1855	0.0536
Multiple birth	-0.1883	0.0802	0.2421	0.0581	-0.1880	0.0802	0.2419	0.0581
C-section	-0.3327	0.0786	-0.1491	0.0569	-0.3317	0.0786	-0.1494	0.0569
Intubation at birth	-1.9102	0.0750	0.2897	0.0766	-1.9086	0.0750	0.2877	0.0766
Use of antenatal steroids	-1.1505	0.0795	-0.3887	0.0694	-1.1493	0.0795	-0.3883	0.0694
Maternal antibiotics	-0.1772	0.0793	-0.1456	0.0615	-0.1786	0.0793	-0.1454	0.0615
Antepartum hemmorrage	0.1466	0.0831	0.0676	0.0660	0.1461	0.0831	0.0669	0.0660
Maternal hypertension	-1.0867	0.1140	-0.3755	0.0703	-1.0875	0.1140	-0.3749	0.0703
Education: High school or more	0.2235	0.1146	-0.1838	0.0804	0.2250	0.1146	-0.1834	0.0804
Respiratory support at 24 hours			0.5950	0.1476			0.5933	0.1476
-2 Res Log Pseudo-Likelihood	103433.6		64966.4		168347.3			
Variance Parameter	0.2307	0.0870	0.1811	0.0660	0.2305	0.0869	0.1851	0.0674
Covariance Parameter					0.0482	0.0545		

Table 3. Estimation results from the multivariate random-effects continuation-ratio logit model: including an additional covariate available for the late death, respiratory support at 24 hours.

Covariate	Died <=24 hours		24 hours - 120 days	
	Estimate	S.E.	Estimate	S.E.
Intercept	-2.2253	0.2221	-3.0989	0.2096
Gestational age: <=23 weeks	2.2535	0.1667	0.8833	0.1246
24 weeks	1.2026	0.1592	0.6019	0.1048
25 weeks	0.6190	0.1602	0.2499	0.0998
26 weeks	0.0953	0.1709	-0.0470	0.1012
27 weeks	-0.2280	0.1999	-0.0212	0.1076
Birth weights: 401-499 grams	2.4843	0.1781	2.5125	0.1333
500-599 grams	1.7083	0.1651	1.7240	0.1073
600-699 grams	1.0134	0.1620	1.1614	0.0983
700-799 grams	0.5939	0.1639	0.6785	0.0966
800-899 grams	0.3779	0.1645	0.1999	0.0991
Gender: male	0.4108	0.0674	0.4458	0.0502
Race: white	0.3201	0.0728	0.1870	0.0536
Multiple birth	-0.1883	0.0802	0.2421	0.0581
C-section	-0.3327	0.0786	-0.1491	0.0569
Intubation at birth	-1.9102	0.0750	0.2897	0.0766
Use of antenatal steroids	-1.1505	0.0795	-0.3887	0.0694
Maternal antibiotics	-0.1772	0.0793	-0.1456	0.0615
Antepartum hemmorrage	0.1466	0.0831	0.0676	0.0660
Maternal hypertension	-1.0867	0.1140	-0.3755	0.0703
Education: High school or more	0.2235	0.1146	-0.1838	0.0804
Not available	0.4507	0.1037	-0.1060	0.0741
Respiratory support at 24 hours	0.0000	.	0.5950	0.1476
-2 Res Log Pseudo-Likelihood	168372.6			
Variance Parameter	0.2307	0.0870	0.1854	0.0675
Covariance Parameter	0	n.a.		