Paper 151-31

Analysis of Large Hierarchical Data with Multilevel Logistic Modeling Using PROC GLIMMIX

Jia Li, Constella Group, LLC, Durham, NC
Toni Alterman, James A. Deddens, National Institute for Occupational Safety and Health, Cincinnati, OH

ABSTRACT

Studies that combine individual-level and aggregate data are common in epidemiologic research. Such studies are often subject to ecological fallacy which arises from confounding of the individual-level relationship due to heterogeneity of exposure variables and covariates within groups. One approach to address this concern is to use multilevel modeling. The advantage of using multilevel modeling is that it takes the hierarchical structure of the data into account by specifying random effects at each level of analysis, and thus results in a more conservative inference for the aggregate effect. In this study, we combined data from two databases for analysis. Data from the National Occupational Mortality Surveillance System (NOMS) containing individual-level information from death certificates was linked by occupation to the Occupational Information Network (O*NET) which contains job characteristics at the occupational level. We examined the adjusted association between job characteristics and select causes of death. A recently available generalized linear mixed models procedure, PROC GLIMMIX, was used to fit the multilevel logistic regression model to our data. Results are compared to those obtained from logistic regression modeling that ignores the hierarchical structure of the data. Results demonstrate the potential of drawing incorrect conclusions when multilevel modeling is not used. Problems encountered from use of PROC GLIMMIX with large data sets will be discussed.

INTRODUCTION

The National Occupational Mortality Surveillance System (NOMS) is based on death certificate data from 28 states with coded occupation information. Cause of death is coded to the 9th revision of the International Classification of Diseases (ICD-9). Occupation is coded according to the 1980 Bureau of the Census classification system. The database is maintained through a collaborative effort of the National Institute for Occupational Safety and Health (NIOSH), the National Center for Health Statistics (NCHS), and the National Cancer Institute (NCI). Demographic variables such as gender, race and age are also included in the database.

The Occupational Information Network (O*NET) system was developed by the Department of Labor (DOL) as a replacement for the Dictionary of Occupational Titles. A major goal of O*NET is to develop a common language for identifying and defining work attributes, work content, and outcomes that can form the basis of a classification system that would be applied across occupational groups. It is a unique, powerful source for continually updated information on skill requirements and occupational characteristics. The O*NET contains a large number of quantitative variables, and is a primary source of occupational information. O*NET 98 is the first version of the system. A crosswalk between 1,169 O*NET 98 occupation codes and 468 census occupation codes is provided with the data.

The purpose of this study was to examine the association between death from injury based on data from the NOMS and the job characteristic “exposure to hazardous equipment at work” from the O*NET, among blue-collar workers (U.S. Bureau of Census occupation codes 403-889). NOMS 1998 data was linked with the O*NET 98 data by census occupation code. For each census occupation that corresponded to more then one O*NET occupation, the mean of the exposure values was obtained and rounded to an integer. Decedents whose ICD-9 coded cause of death was 800-959 (WHO, 1977) were defined as having died from injury. Exposure to hazardous equipment was coded on a Likert scale from 0 to 4 in O*NET 98. Ratings of 3 and 4 were combined and coded as ‘high exposure’ and ratings of 0 to 2 as ‘low exposure’ for ease of interpretation. Those who died at less than 20 years of age were excluded from the study because they were assumed to have limited occupational experience. Decedents aged 60 or more were also excluded because they were more likely to have retired or not be working at their usual occupation at the time of their death. This resulted in a sample size of 48,662 decedents having had a usual occupation as a blue collar worker.

METHODS

Research involving a combination of both individual-level and aggregate data has a long history in epidemiology (Chambers & Skinner, 2003). One reason for this is that individual-level data may be limited, while aggregate data containing the relevant information may be readily available from administrative sources or population census. The aggregate data are usually in the form of means or percents for a set of groups into which the population has been partitioned. However, use of aggregate data alone to make inference about individual-level relationships can introduce bias, leading to the ecological fallacy (Steel & Holt, 1996). A study with an aggregate risk factor, but with outcome, group membership, and confounders assessed individually is partially ecologic. The ecological fallacy arises from confounding of the individual-level relationship due to heterogeneity in exposure variable of interest and other covariates within groups (Rothman & Greenland, 1998). One approach to address concerns regarding ecological fallacy has been the development of multilevel modeling (Goldstein, 2003). Multi-level modeling takes into account the hierarchical structure of the data (e.g. decedents clustered within occupations as in our data). Such data structure is subject to intra-class correlation, whereby individuals within the same group are more alike than individuals across groups. Analysis that ignores this intra-class correlation may underestimate the standard error of the regression coefficient of the aggregate risk factor, leading to overestimation of the significance of the risk factor. To more conservatively estimate the standard error, a separate random error term should be specified for each level of analysis (Blakely & Woodward, 2000).

To illustrate the above point, we conducted our analysis using two approaches.

In the first approach, we fit a multiple logistic regression model on the combined data with PROC LOGISTIC. The dependent variable is death from injury (yes/no); the risk factor of interest is exposure to hazardous equipment at work (high/low); confounders included are gender, race (white/black/other), age (continuous, centered) and a quadratic term for age. This model ignores the hierarchical structure of the data, and treats aggregate exposure as if it was measured at individual level. The model is expressed by the following equation

$$\log it(\pi_{ij}) \equiv \log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \mu + Exposure_i + Gender_{ij} + Race_{ij} + \beta_1 Age_{ij} + \beta_2 Age_{ij}^2$$

where π_{ij} is the expected probability of death from injury for the *j*th individual of the *i*th occupation conditional on the predictor variables. The SAS code is shown below.

```
proc logistic data=noms.combined descending;
  class exposure gender race / param=ref order=internal;
  model injury = exposure gender race age age*age;
run;
```

In the second approach, we fit a generalized linear mixed model (GLMM) on the same data with SAS GLIMMIX procedure. The GLIMMIX procedure for SAS 9.1 is available for downloading from the SAS website. In this approach the influence of job exposure on the outcome is included through both occupation-level fixed effect and random effect. The model is expressed by the following equation

$$\log it(\pi_{ij}) \equiv \log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \mu + Exposure_i + \gamma_j + Gender_{ij} + Race_{ij} + \beta_1 Age_{ij} + \beta_2 Age_{ij}^2$$

where γ_j ~ i.i.d. N(0, σ_g^2), and π_{ij} is the expected probability of death from injury for the *j*th individual of the *i*th occupation conditional on the predictor variables and the random effect. The SAS code is shown below.

```
proc glimmix data=noms.combined;
  class occupation exposure gender race;
  model injury (event='Yes') = exposure gender race age age*age
      / solution dist=binary link=logit ddfm=satterth oddsratio;
  random intercept / subject=occupation solution;
  nloptions tech=nrridg;
  ods exclude solutionr;
run;
```

The MODEL statement lists the same fixed effects as in the previous approach, and the RANDOM statement identifies the group structure in the mixed model. The model option DDFM=SATTERTH uses the Satterthwaite method to adjust for denominator degree of freedom for tests of the fixed effects. One draw back in using the Satterthwaite method is that it requires extra memory and extra computation time. When the number of observations per group is large as in our data, and when running the program on a PC, it is possible that it will take the program a long time to run, or the program may stop with an OUT OF MEMORY error message. Fortunately, SAS has an alternative algorithm to solve this problem, that is, to ask for the random effects solutions by specifying the SOLUTION option in the RANDOM statement. Since we do not want to see these random solutions, we use the ODS EXCLUDE SOLUTIONR statement to suppress them in the printout. The model option ODDSRATIO requests calculation of odds ratios and their confidence limits. Models fit with the GLIMMIX procedure usually require nonlinear optimization methods. One can control the optimization through options of the NLOPTIONS statement. In this study we used the TECH=NRRIDG option in the NLOPTIONS statement which specifies an optimization technique of Newton-Raphson with ridging to help with the convergence of the procedure.

RESULTS

Tables 1-3 are selected outputs from the first approach. The coefficient of exposure to hazardous equipment is 0.0769, and its standard error is 0.0299. The p-value for the Wald chi-square test is 0.0101, indicating a significant association between death from injury and exposure to hazardous equipment at work. The estimated mortality odds ratio (high exposure vs. low exposure) is 1.08 (95% CI: 1.02-1.15).

Table1. Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Exposure	1	6.6177	0.0101
Gender	1	216.9798	<.0001
Race	2	898.8020	<.0001
Age	1	3433.7309	<.0001
Age*Age	1	27.1749	<.0001

Table 2. Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.3860	0.0200	4826.0120	<.0001
Exposure	High	1	0.0769	0.0299	6.6177	0.0101
Gender	Female	1	-0.5435	0.0369	216.9798	<.0001
Race	Black	1	-1.0437	0.0348	898.2554	<.0001
Race	Other	1	-0.1277	0.0800	2.5477	0.1105
Age		1	-0.1123	0.00192	3433.7309	<.0001
Age*Age		1	-0.00061	0.000118	27.1749	<.0001

Table3. Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Exposure High vs Low	1.080	1.018	1.145
Gender Female vs Male	0.581	0.540	0.624
Race Black vs White	0.352	0.329	0.377
Race Other vs White	0.880	0.752	1.030

Tables 4-7 are selected outputs from the second approach. The estimated between occupation variance is 0.0409. The coefficient of exposure to hazardous equipment is 0.0334, and its standard error is 0.0571. The p-value for the Wald chi-square test is 0.5594, indicating no statistically significant association between death from injury and exposure to hazardous equipment at work. The estimated mortality odds ratio (high exposure vs. low exposure) is 1.03 (95% CI: 0.92-1.16).

Table 4. Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	Occupation	0.04090	0.01025

Table 5. Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Exposure	1	135.3	0.34	0.5594
Gender	1	4744	92.06	<.0001
Race	2	48655	406.74	<.0001
Age	1	48655	3390.14	<.0001
Age*Age	1	48655	23.62	<.0001

Table 6. Solutions for Fixed Effects

Effect	Exposure	Gender	Race	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept				-1.4228	0.03152	178.7	-45.13	<.0001
Exposure	High			0.03339	0.05705	135.3	0.59	0.5594
Exposure	Low			0	.	.	.	.
Gender		Female		-0.4056	0.04228	4744	-9.59	<.0001
Gender		Male		0	.	.	.	.
Race			Black	-1.0090	0.03539	48655	-28.51	<.0001
Race			Other	-0.1093	0.08036	48655	-1.36	0.1737
Race			White	0	.	.	.	.
Age				-0.1128	0.001937	48655	-58.22	<.0001
Age*Age				-0.00058	0.000118	48655	-4.86	<.0001

Table 7. Odds Ratio Estimates

Effect	Exposure	Gender	Race	_Exposure	_Gender	_Race	Estimate	DF	95% Confidence Limits	
Exposure	High			Low			1.034	135.3	0.924	1.157
Gender		Female			Male		0.667	4744	0.614	0.724
Race			Black			White	0.365	48655	0.340	0.391
Race			Other			White	0.896	48655	0.766	1.049
Age							0.893	48655	0.890	0.897
Age*Age							0.999	48655	0.999	1.000

CONCLUSION

Aggregate data are often easier to obtain than data on individuals, and may offer valuable clues about individual behavior. Combining individual-level and aggregate data is an efficient approach in epidemiologic research. Multilevel modeling has the advantage of taking the hierarchical structure of such combined data into account. By specifying random effects at each level of analysis, inference made on the aggregate fixed effect is more conservative.

Running PROC GLIMMIX on our large dataset which contains more then 48,000 observations with Satterthwaite calculations is computationally intensive because of the extra memory needed. The consequence is that either it takes the program a very long time to run, or the program may stop with OUT OF MEMORY error message. An algorithm for the Satterthwaite calculations that requires less memory is available in SAS which asks for the random effects solutions in the RANDOM statement by specifying the SOLUTION option.

REFERENCES

Blakely, T., Woodward, A. 2000. “Ecological effects in multi-level studies”. *Journal of Epidemiology and Community Health* 54:367-374.

Chambers, R.L., Skinner, C.J. 2003. *Analysis of survey data*. New York: John Wiley.

Goldstein, H. 2003. *Multilevel statistical models*. New York: John Wiley.

Rothman, K.J., Greenland, S. 1998. *Modern epidemiology*. Philadelphia: Lippincott-Raven.

Steel, D., Holt, D. 1996. “Analyzing and adjusting aggregation effects: the ecological fallacy revisited”. *International Statistical Review* 64:39-60.

WHO. 1977. *International classification of disease, 9th revision*. Geneva, Switzerland: World Health Organization.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Jia Li
Constella Group, LLC
5555 Ridge Ave
Cincinnati, OH 45213
Work Phone: (513) 841-4455
Email: jli@constellagroup.com