Analyzing multiple membership hierarchical data using PROC GLIMMIX

Elsa Vazquez Arreola, Arizona State University, Tempe, Arizona

Jeffrey R. Wilson, Arizona State University, Tempe, Arizona

ABSTRACT

Data collected based on a hierarchical structure are commonly seen in many fields such as education, social sciences, and medicine to name a few. For example, patients nested within doctors and doctors nested within hospitals or students nested within schools, schools nested within districts. Hierarchical data where lower level units are nested within a single upper level unit, for example, cases when patients only see one doctor are known as single membership. However, it is not always the case that lower level units are completely nested within only one upper level unit, as when a patient sees two or more doctors at a time; such nesting is known as multiple membership. Browne, Goldstein and Rasbash (2001) developed multiple membership multilevel models for continuous and binary data. We illustrate the use of PROC GLIMMIX in fitting multiple membership models to hierarchical binary data.

INTRODUCTION

In fields such as education, medicine, accounting, business among others, correlated data is a common issue. There may be observations collected at several time points from longitudinal studies, or data might be collected in a hierarchical structure. For example, in medicine, patients may be nested within doctors and doctors nested within hospitals; in education, students are usually nested within teachers, teachers within schools and schools within districts; in accounting, companies are nested within accounting firms, etc. When modeling outcomes of data with a hierarchical structure, it is imperative to account for the clustering and hence the correlation inherent at each level of the hierarchy. If we were to ignore such, one risks obtaining results on the parameter estimates and hypothesis tests that may not be valid, Wilson and Lorenz (2015).

When lower level units are completely nested within only one single level unit, data are said to be pure hierarchical data or to have a single membership structure. However, there may be situations in which some lower level units are nested within more than one higher level unit, such structure is known as multiple membership. Identifying whether data are pure hierarchical or multiple membership is the first step in conducting correct statistical analyses; once this is done, it is also important to use the correct model depending on the format of the data. We explore situations when there are two levels in the hierarchy, where level-2 units represent the higher level units and level-1 units represent the lower level units. We provide examples of situations where single or multiple membership data are present. We discuss models that can be fitted to each of these types of data and that account for the clustering at the level-2 units, hierarchical and multiple membership models. We give models for continuous and binary data and present code on PROC GLIMMIX to fit multiple membership models to binary outcomes.

REGRESSION MODELS FOR HIERARCHICAL DATA

HIERARCHICAL MODELS

Nested data structures are the norm rather than the exception in most research fields. For example, one may want to study performance of students from different elementary schools on a standardized test, in such case students would be nested within schools. Another example would be, if we were interested in understanding how accounting firms impact the probability of different companies of filing restatement in their annual reports, where companies would be nested within accounting firms. Also, in a study conducted to detect what influences the probability that an embryo would implant when using in-vitro fertilization for women with tubal disease (Hogan & Blazar, 1999), several embryos coming from the same woman were utilized, embryos were nested within women. In all above examples, units at the observational level are completely nested within only one higher level unit; students are completely nested within schools, companies are completely nested within accounting firms and embryos within women. Thus, there is a likely correlation between the outcomes at the observational level coming from the nesting at the higher level. Hierarchical models are useful in such situations.

Hierarchical models or mixed effects models as often referred to, consist of both fixed and random effects. Usually fixed effects are given by covariates in the models and random effects are the higher level units, the units at which the clustering exists. These models allow us to identify and account for the multiple sources of variation that may affect the response, (Zhu 2014). When one has continuous or binary data with a hierarchical structure, one can use generalized linear mixed effects to model the outcomes; linear mixed effects models for continuous responses and hierarchical logistic regression for binary data. Models with random intercepts for each type of outcome are shown below, in each of them, \( n \) represents the number of clusters or level-2 units, \( i = 1, \ldots, n \), and \( n_j \) represents the number of units per cluster \( i,j = 1, \ldots, n_j \); \( X_1, \ldots, X_p \) are fixed effects or covariates; \( \beta_1, \ldots, \beta_p \) are the coefficients for the
respective covariates.

**Linear Mixed Model with Random Intercept**

\[ Y_{ij} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \gamma_i + \epsilon_{ij} \]

Where \( Y_{ij} \) is the value of the response for subject \( j \) nested within unit \( i \), \( \beta_0 \) represents the marginal intercept; \( \gamma_i \) represents the random effects for each of the level-2 units with \( \gamma_i \sim N(0, \sigma^2_\gamma) \) and \( \epsilon_{ij} \sim N(0, \sigma^2_\epsilon) \)

**Hierarchical Logistic Regression with Random Intercept**

\[ \logit(P_{ij}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \gamma_i \]

Where \( \logit(P_{ij}) \) represents the logit of the probability of success for unit \( j \) nested within unit \( i \), \( \beta_0 \) depicts the marginal odds of success for a typical level-1 unit nested within a typical level-2 unit (marginal intercept) and \( \gamma_i \) denotes the random effects on the probability of success of cluster \( i \) with \( \gamma_i \sim N(0, \sigma^2_\gamma) \).

Although data where complete nesting is present is very common, there are situations where some level-1 units are nested within two or more level-2 units, and the nesting is not complete. In such cases, to avoid making invalid inferences, one should fit the so called multiple membership models. We explain when it is required to use such models and provide an example on how to fit them using PROC GLIMMIX.

**MULTIPLE MEMBERSHIP MODELS**

While nested data structures are very common in education, social sciences, medicine and health care among others, there are cases when the nesting is not complete as all units in one level are not completely nested in a higher level. There are those hierarchical structures where a unit is found to be contained in more than one upper level. One example, suppose we have high school students completing their 4 years of study at the high school level and taking the SAT, while some students only attend one school during a four year period, others attend more than one school, and each of those schools has a certain effect on the SAT scores. Another example pertains to a patient recovering from a surgery and one wants to account for the effect a nurse had on the recovery, in such a case what normally happens is that several nurses take care of the same patient, and all of them influence the recovery process. Yet another example in accounting, we could be interested in modeling the number of years a company has to restate their financial statements during a 10-year period. Usually firms are nested within accounting firms, but some companies may change from one firm to another during the 10 years, and when modeling this response, we would have to account for the effects of each of the accounting firms that the company belonged to. These examples give rise to multiple membership hierarchical models.

In multiple membership multilevel models, lower level observational units are nested within one or more higher level units at the same time, (Fielding & Goldstein, 2006). It is important to account for the effects on the responses for each of the higher levels of which the lower level unit is a member. To incorporate the multiple contributions of each of the higher level units, a widely used technique is to assign weights to these contributions. For example, in the patient recovering example, we would assign weights to each of the nurses; usually these weights are proportional to the number of higher level units. In the financial restatement example, a company might be a firm’s client for 3 years, a client for a second firm for 6 years, and a client of a third firm for one year, in cases like this it is more appropriate to assign weights depending on the time spent by the company with each of the accounting firms.

In 2001, Browne, Goldstein and Rasbash introduced multiple membership multilevel models for response variables that follow distributions from the exponential family such as continuous and binary data. Each of these models has a similar structure to the generalized linear mixed effects models, developed for each type of response; the only difference between the models for single membership data and multiple membership data, comes from the weights mentioned before. The multiple membership models for each type of response are shown below. Again, \( n \) represents the number of clusters or level-2 units, \( i = 1, \ldots, n \); \( n_i \) represents the number of units per cluster \( i, j = 1, \ldots, n_i \); \( X_1, \ldots, X_p \) are fixed effects or covariates; \( \beta_1, \ldots, \beta_p \) are the coefficients for the respective covariates and \( j \in cluster(i) \) is the set of level-2 units at which level-1 unit \( j \) is nested with.

**Multiple Membership Model for Continuous Data**

\[ Y_{ij} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \sum_{j \in cluster(i)} w_{ij}^{(2)} \gamma_{ij}^{(2)} + \epsilon_{ij} \]
Where $Y_{ij}$ is the outcome of subject $j$ nested within cluster $i$, $\beta_0$ represents the marginal intercept, $w_{ij}^{(2)}$ are the weights of each cluster $i$ on the outcome of lower level unit $j$ with $\sum w_{ij}^{(2)} = 1$, $\gamma_{i}^{(2)}$ represents the random effects of cluster $i$ on the outcome with $\gamma_{i}^{(2)} \sim N(0, \sigma^{2}_{\gamma})$ and $e_{ij} \sim N(0, \sigma^{2}_{e})$.

**Multiple Membership Model for Binary Data**

$$logit(P_{ij}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \sum_{j \in \text{cluster}_i} w_{ij}^{(2)} \gamma_{i}^{(2)}$$

Where $logit(P_{ij})$ is the log of the probability of success for unit $j$ nested within cluster $i$, $\beta_0$ represents the intercept of a typical level-1 unit nested within a typical level-2 unit. $w_{ij}^{(2)}$ is the weight corresponding to level-2 unit $i$ for level-1 unit $j$, and the weights for each level-1 unit $j$ sum up to 1; $\gamma_{i}^{(2)}$ is the random effect provided by level-2 unit $i$ with $\gamma_{i}^{(2)} \sim N(0, \sigma^{2}_{\gamma})$.

**DATA EXAMPLE**

To illustrate the use of PROC GLIMMIX in fitting multiple membership models, we analyzed data extracted from the National Survey of Children with Special Health Care Needs 2009/2010. Data was collected in each state around the United States where about 750 households were randomly sampled per state, and only one child per household was included in the study. The outcome of interest was whether the children’s health conditions caused financial problems for their families, thus our response variable was binary and we modeled the probability that children’s conditions created financial hardship at their household. We also included children’s gender (SEX), whether the primary language spoken at the children’s home was English (PLANGUAGE), whether the child was Hispanic (HISPANIC) and the number of adults at the household (TOTALADULT) as covariates in our model. After deleting observations where either the outcome or the covariates were missing, our dataset contained 33,805 observations. There was information on 20 different diagnosed conditions for each child in the dataset, which we used as our second level units. Since some children in the survey had been diagnosed with two or more conditions, we utilized diagnosed conditions as our multiple membership units. Table 1 presents the conditions included in the study and the number assigned to them.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Number Assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention Deficit Disorder or Attention Deficit Hyperactive Disorder (ADD or ADHD)</td>
<td>1</td>
</tr>
<tr>
<td>Depression</td>
<td>2</td>
</tr>
<tr>
<td>Anxiety Problems</td>
<td>3</td>
</tr>
<tr>
<td>Behavioral or conduct problems</td>
<td>4</td>
</tr>
<tr>
<td>Autism, Asperger’s Disorder, pervasive developmental disorder, or other autism spectrum disorder</td>
<td>5</td>
</tr>
<tr>
<td>Developmental delay</td>
<td>6</td>
</tr>
<tr>
<td>Intellectual disability or mental retardation</td>
<td>7</td>
</tr>
<tr>
<td>Asthma</td>
<td>8</td>
</tr>
<tr>
<td>Diabetes</td>
<td>9</td>
</tr>
<tr>
<td>Epilepsy or seizure disorder</td>
<td>10</td>
</tr>
<tr>
<td>Migraine or frequent headaches</td>
<td>11</td>
</tr>
<tr>
<td>Head Injury, concussion, or traumatic brain injury</td>
<td>12</td>
</tr>
<tr>
<td>A heart problem, including congenital heart disease</td>
<td>13</td>
</tr>
<tr>
<td>Blood problems such as anemia or sickle disease</td>
<td>14</td>
</tr>
<tr>
<td>Cystic Fibrosis</td>
<td>15</td>
</tr>
<tr>
<td>Cerebral Palsy</td>
<td>16</td>
</tr>
<tr>
<td>Muscular Dystrophy</td>
<td>17</td>
</tr>
<tr>
<td>Down Syndrome</td>
<td>18</td>
</tr>
</tbody>
</table>
Table 1: Diagnosed conditions and number assigned when manipulating dataset

<table>
<thead>
<tr>
<th>Condition</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthritis or other joint problems</td>
<td>19</td>
</tr>
<tr>
<td>Allergies</td>
<td>20</td>
</tr>
</tbody>
</table>

An important aspect in fitting multiple membership models using PROC GLIMMIX, is the structure of the dataset analyzed. There must be only one observation per children, in which the conditions diagnosed appear in different columns. The highest number of diagnoses per child was 16, so we had 16 columns for conditions and 16 columns for weights in each observation. In our case, we assigned weights to the random effects of diagnosed conditions, based on the number of diagnoses per child; if a child was diagnosed with only one condition, 1 was the weight for the random effects of that condition; if a child was diagnosed with 2 conditions, the random effects for each condition had weights of ½ each, and so on. Figure 1 and Figure 2 show some observations of the dataset to which we fit the multiple membership model. On Figure 1, we can observe that child with id 3, had conditions 2 and 3, while child with id 19 had conditions 6, 14 and 20 and the child with id 64 had a total of 5 diagnosed conditions. In Figure 2, we can observe the weights for the random effects on diagnosed conditions.

Once we gave our dataset the correct structure, since our response variable was binary, we analyzed it fitting a multiple membership logistic regression model using PROC GLIMMIX. The following code was used:

```plaintext
proc glimmix data=final_data2;
  class disease1-disease16 PLANGUAGE(ref='0') SEX(ref='0') HISPANIC(ref='0');
  effect diseases=multimember(disease1-disease16/weight=(weight1-weight16) details);
  model response(event='1')=PLANGUAGE SEX HISPANIC TOTADULTR /dist=binary link=logit solution;
  random diseases/solution;
run;
```

The CLASS statement lets us specify categorical variables. In the EFFECT statement we define the contributions of each diagnosed condition to the probability that a child’s health conditions cause financial problems for his or her family, this by using the MULTIMEMBER keyword followed by the CLASS variables disease1 to disease16 and their respective weights. Our effect is called diseases. We then used the MODEL statement to specify all the fixed effects, the response variable and to define it through the DIST and LINK keywords as a binary outcome, where DIST=binary and LINK=logit. In the RANDOM statement, we declared diseases, our multimember effect as a level of clustering.

After running the code, we proceeded with our analysis. We were able to conclude that there were significant random
effects of diagnosed conditions on the probability of households facing financial problems because of child’s health conditions. The z-score for the variance of the random effects of diagnosed conditions was 2.924. We also found that primary language spoken at home, child’s gender, whether the child was Hispanic and the number of adults per child’s household had a significant impact on the probability that children’s health conditions caused financial hardship for their families. We observed that for a child whose primary language spoken home was English, the odds that those child’s health conditions caused financial problems for his or her family were smaller than the odds of a child whose primary language spoken at home was not English. Also, boys’ families were more likely than girls’ families to have financial problems because of the child’s health condition. Hispanic families were more likely than non-Hispanic families to face economic hardship because of their child’s health conditions. As the number of adults living in the household increased, the probability that a child’s health conditions caused financial problems for his or her family decreased. Partial output for this analysis is shown in Figure 3. Figure 4 shows the estimated random effects of each diagnosed condition.

![Partial Output of PROC GLIMMIX for multiple membership logistic regression model](image)

**Figure 3:** Partial Output of PROC GLIMMIX for multiple membership logistic regression model
Figure 4: Estimated random effects of diagnoses on the probabilities of having financial problems

It is worth noting that, PROC GLIMMIX can also be used to fit multiple membership models to continuous data. One just has to change the DIST and LINK option in the MODEL statement to DIST=normal and LINK=identity.

CONCLUSIONS

Hierarchical datasets are found in many research fields. Although, pure hierarchical or single membership data where lower level units are completely nested within only one higher level unit are ubiquitous. Multiple membership data, where lower level units are not completely nested within a single higher level unit are also very common. Accounting for the multiple membership structure of the data is very important to draw valid conclusions while conducting a statistical analysis. We provide sample code to fit models using PROC GLIMMIX that account for this structure when modeling binary outcomes and suggest changes to this code, so similar models can be fit to continuous response variables.

REFERENCES


ACKNOWLEDGMENTS

The authors would like to thank the Data Resource Center for Child & Adolescent Health, for collecting and facilitating the use of the National Survey of Children with Special Health Care Needs 2009/2010 dataset.

RECOMMENDED READING

- PROC GLIMMIX: The GLIMMIX Procedure :: SAS/STAT® User’s Guide

CONTACT INFORMATION

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

Name: Elsa Vazquez Arreola
E-mail: Maria.Vazquez@asu.edu
Web: https://www.linkedin.com/in/elsavazquez89/

Name: Jeffrey R Wilson
E-mail: Jeffrey.Wilson@asu.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.