Published in final edited form as:

Nurs Res. 2011 ; 60(3): 208–212. doi:10.1097/NNR.0b013e31820a3dbe.

Application of a Hierarchical Model Incorporating Intra-family Correlation and Cluster Effects

An-Lin Cheng, PhD[Assistant Professor] and
School of Nursing, University of Missouri-Kansas City

Patricia J. Kelly, PhD, MPH, APRN[Professor]
School of Nursing, University of Missouri-Kansas City

Abstract

Background—Research interventions at the family level often include individual- and group-level data that can present an analytic challenge. The study that motivated this paper was an intervention study conducted with elementary school children and their parents. Randomization occurred at the school level, with families nested within schools. Repeated measurements collected from children and parents at different time points presented modeling challenges, including how to specify the covariance structure correctly among all measurements.

Objectives—To introduce a mixed model with random effects to model the correlations among family members, repeated measures, and the grouping effect.

Method—A hierarchical random-effect model was used that included both fixed and random effects; time and intervention-by-time variables were included as fixed effects, the school-specific variable was included as random effect, and the intrafamily correlation was modeled through a spatial autoregression covariance matrix. Comparisons were made between the performance of the proposed modeling method with other parsimony models using Akaike's Information Criterion (AIC).

Results—The proposed modeling method produced a 3% and 9% reduction of AIC values, respectively, compared to the two other models. The likelihood ratio test further confirmed that the full model is better than the other two models ($p < .0001$ for both models).

Discussion—The data suggest that using the proposed mixed model technique will produce a significantly better model fit for intrafamily correlation with a nested study design.

Keywords

mixed modeling; longitudinal data; cluster effect

The statistical technique of multilevel modeling has been used in education, nursing, health, and social science research for over 20 years (Draper, 1995; Estabrooks, Midodzi, Cummings, & Wallin, 2007; Greenland, 2000; Rasbash & Goldstein, 1994). Because of its utility in clinical trials, applications of this modeling method are often found in studies with randomized samples. Randomization occurs either within the groups of the hierarchical structure or among individual participants as a result of study designs and procedures (Murray, 1998). With multilevel modeling, it is possible to identify the two sources of variance induced by the design: within a cluster and between clusters. This powerful

analytic tool can be used to identify the hierarchy structure in the data set and to specify the cluster effects in the model, producing results with increased estimation accuracy (Demidenko, 2004).

One application of multilevel modeling is analyzing family research data. In many studies, family units are nested within communities or treatment groups (Herrera, Lee, Nanyonjo, Laufman, & Torres-Vigil, 2009). Challenges in analyzing such studies arise because the measurements are correlated between family members and repeated measurements from different time points are correlated for each individual. Knafl, Knafl, and McCorkle (2005) proposed and demonstrated a modeling technique that can take the intrafamilial correlations into account easily with a spatial correlation matrix. However, the issue of multilevel data structure is not addressed by Knafl et al. (2005). Proposed here is a multilevel model based on Knafl et al.'s (2005) method, and the method is extended to address a more complicated data structure.

Motivating Study

Familias en Acción was a community-based participatory action research study using a prospective randomized control design to examine intervention effects of violence-related attitudes. The 312 participants were third, fourth, and fifth grade students and their parents at 14 elementary schools in the Southwest United States (Kelly et al., 2010). Randomization occurred at the school level, with half of schools receiving the intervention in year one and the remaining schools receiving the intervention in year two in a delayed entry manner. The 10-session curriculum was implemented weekly during an existing district-wide afterschool program while one of the parents also participated in the program. Data were collected from students and their parents at three time points: baseline, postintervention, and 3 months postintervention. Some families had more than one child in the study. Institutional Review Board (IRB) approval was obtained from the two academic institutions of the investigators.

Intervention content comprised four program elements: *Conocimiento* (Acknowledgment), *Entendimiento* (Understanding), *Integración* (Integration), and *Movimiento* (Movement); designed to include physical, emotional, mental, and spiritual aspects related to individual, family, and community facets of life. Intervention activities were informed by traditional indigenous teachings, based on culturally rooted concepts stressing the values necessary to build and maintain harmonious and balanced relationships. The core relationship values were *respeto* (respect), *dignidad* (dignity), *confianza* (trust), and *cariño* (love). Facilitation was based on the use of an *espejo* (mirror) process of teaching using strategies such as storytelling, reflection, and guidance. The role of the facilitator was to be a teacher, guide, role model, and nurturer. Facilitators were trained to present material in light of personal experiences rather than dogmatic theory.

The analytic challenge was to correctly model the school (cluster) effect, the time effect, the intervention effect, the parent-to-child correlation, and the correlations due to repeated measures across time. A diagram of the data structure is presented in Figure 1. One of the outcome variables--program values--is a six-item survey developed from intervention content (A man who is really 'macho' is always ready to fight; A man who is really 'macho' never admits that he is wrong.) with 3-point Likert-scale responses and possible scores of 3 to 18 ($\alpha = .65$); higher scores indicated more positive attitudes about program issues. This scale will be used to illustrate the proposed modeling technique to determine the intervention effect.

Method

Model-Building

The modeling challenge of the data structure was twofold. First, there was a hierarchical structure, with families nested within schools. Second, within each family, measures from the child or children and from one of the parents were collected across different time points. To address these challenges, the model-building needed to incorporate two steps: first, to identify the fixed effect and, second, to identify the random effect. The role of the fixed effect is to answer the proposed research question; the role of the random effect is to account for the various correlations induced from the design to provide precise interpretations for the fixed effect. Built on the foundation of multilevel modeling techniques (Laird & Ware, 1982; Verbeke & Molenberghs, 2000), a general linear mixed model for the data can be conceptualized as follows:

$$Y_{ip:jfk:l} = T_j + TC_{jl} + U_{k:l} + U_{ip:jf:l} + \varepsilon_{ip:jfk:l} \quad (1)$$

with the observed value $Y_{ip:jfk:l}$ for the i th child or p th parents nested within f th family, k th school, and l th intervention group, and observation at the j th time as a function of two fixed effects; the time effect at time j (T_j) plus the intervention effect at time j (TC_{jl}) and two random effects: $U_{k:l}$ and $U_{ip:jf:l}$, plus the residual $\varepsilon_{ip:jfk:l}$. T_j represents the changes of the measurement at time j and TC_{jl} represents the changes of the measurement for intervention group at time j . $U_{k:l}$ is a random effect used here to account for the cluster effect (school), and $U_{ip:jf:l}$ is a random effect to account for the correlations within family.

Fixed effect – Step one—The first step of model building is to identify the fixed effect. The fixed effects are considered as part of the main effect of interest to answer the research hypotheses. In the model presented here (1), the main interest was in the intervention effect at different times, so the intervention effect was included at time j (TC_{jl}) in the model. TC_{jl} was included using dummy coding to present the intervention group at time j . In a longitudinal study without an intervention group, T_j s alone are enough to capture the change on outcome across time. It is also possible to include the location effects for inclusion in the model as fixed effects, if relevant to the study. Possible covariates relevant to the study's theoretical pathways can be included in the model as fixed effects in order to adjust for possible confounding effects.

Random effect – Step two—In a traditional linear regression model, the response variable (Y s) are assumed to be independent and distributed identically, with the error term following a normal distribution (Feldman & McKinlay, 1994; Zucker et al., 1995). However, in the motivating study, these assumptions were not met because of the correlations between the responses from children in the same school, correlations between responses from parents and children in the same family, and correlations of repeated measures from the same individual across different time points. Once a dependence is observed from the data set, the second step of model-building is to incorporate the correlation into the model. To apply Knafl et al.'s (2005) proposed spatial autoregression correlation matrix to model the correlation structure in the family data for a longitudinal study, three indicators must be created to present the complex data structure. First, a dummy variable is created to represent the parent response (parent as 1 and child as 0); second, an indicator is used to represent responses from different children from the same family (first child as 1, second child as 2, and so on); and third, an indicator is used for time domain (baseline as 1, postintervention as 2, and 3-month postintervention as 3). The artificial coding for child, parent, and time point enables the calculation of the distance between any

two measurements within the same family and later serves in the calculation of correlation for any two given measurements.

There are several choices of different spatial autoregression correlation matrixes (Littell, Milliken, Stroup, & Wolfinger, 1996) to fit such data. Knafl et al. (2005) suggested that an anisotropic power correlation matrix can best model the dynamics of such data structure, which can be expressed as:

$$\text{Corr}(U_{ip:jf:l}, U_{ip':j'f:l}) = \sigma^2 \prod_{i=1}^c \rho_i^{d(i,p,j)} = \sum_{ip:jf:l}$$

where σ^2 and ρ_i are unknown correlation parameters for each coordinate, c is the number of coordinates, and $d(i,p,j)$ is the Euclidean distance between the i_{th} coordinate, $i = 1, \dots, c$, of the p_{th} and j_{th} observations in the input data set. The correlations decrease as the distances $d(i,p,j)$ increase and the correlation between any two measurements is the function of the three correlation parameters for each coordinate. The random effect $U_{k:l}$, allows for correlation among individuals within a school and for random variation among the participants ($\varepsilon_{ip:jfk:l}$). Under the general linear mixed model, random effects are assumed to be independent and distributed as $U_{k:l} \sim N(0, \sigma_{k:l}^2)$, $U_{ip:jf:l} \sim N(0, \Sigma_{ip:jf:l})$ and

$\varepsilon_{ip:jfk:l} \sim N(0, \sigma_e^2)$. The estimations of fixed effect and random effect can be achieved through the specification of explanatory variables, random effects (Singer, 1998), and repeated measures (Littell, Henry, & Ammerman, 1998) in SAS 9.2. More description and examples of the method can be found in Knafl et al. (2005).

Analysis

Using the data from the *Familias en Acción* intervention study, the proposed method was evaluated by comparing three different models: the model with proposed spatial covariance structure within multilevel modeling, the multilevel modeling ignoring the spatial feature, and the naïve model without accounting for any multilevel factor. Since the preliminary analyses showed that the outcome variable *program values* was not distributed normally, a Box-Cox transformation of the score was performed on the outcome variable (Box & Cox, 1964). The transformed scores approximated a normal distribution and the finding interpretation was based on the transformed score. The aforementioned models for the outcome variable were fitted to identify the time effect and intervention effect. Gender was included in the models as a significant covariate. Akaike's Information Criterion (AIC) was used for model comparisons calculated as: $AIC = -2L + 2d$, where L was the restricted maximum likelihood value, n was the number of observations, and d was the number of estimated covariance parameters. The best model was assumed to be the one with the lowest values of AIC. Variance components were estimated by using the stabilized Newton-Raphson algorithm implemented in PROC MIXED for restricted maximum likelihood (REML) estimation (Littell et al., 1996). Under some regularity assumptions, the REML estimation of the fixed effect is consistent and robust as long as the mean is specified correctly in the model (Liang & Zeger, 1986). All these models were estimated in the context of linear mixed model using PROC MIXED, SAS v9.2. The program syntax corresponding to each of the analysis models is shown in Table 1.

Results

Gender was included in the models as a significant covariate (Table 2). The time effects were $T_j = 1.3957, 1.4197$, and 1.4042 , at baseline, postintervention, and 3 month postintervention, respectively. The intervention effects were $TC_{jt} = 0.0113, 0.0106$, and

0.0251, at the three time points, respectively. All models provided the same quantitative result; that is, there was a significant intervention effect on program values at 3 months postintervention. Students in the intervention group increased the program values score by 0.0251 with 95% confidence interval of (0.0015, 0.0488) in the desired direction. However, examination of AIC values (Table 2) showed that the proposed full multilevel model with spatial covariance provided the best model fit when compared to models produced without accounting for intrafamily correlation and the naïve model. The full model produced a 3% and 9% reduction of AIC values compared to the two other models, respectively. The likelihood ratio test (Casella & Berger, 2001; Mood, Graybill, & Boes, 1974) further confirmed that the full model was significantly better than the other two models ($p < .0001$ for both models). The results from the full model showed that the correlation of outcome within schools $\sigma^2_{k,l} = .0002$. The intersibling correlation was 0.3594, the parent-child correlation was 0.5295, and the temporal correlation was 0.4504; intrafamily correlations were all statistically significant at .05 level.

Discussion

The hierarchical model was proposed and applied here to analyze complicated familial correlated data from a randomized controlled school-level intervention study. A significant intervention effect was identified at 3 months postintervention; that is, the student and parents who were in the intervention group showed more positive attitudes about program issues than those in the comparison group. The proposed model was compared to a less complicated model that has been used widely as an analysis tool. Even though all three models produced the same statistical conclusion, the AIC confirmed that the proposed model produced the best model fit. The example presented here is for a sophisticated data structure; however, the dimensions of the spatial covariance can be added or deleted easily to model different data structures. This method has potential applications in community-based research, intervention studies involving families, organization level intervention studies, and longitudinal and group randomized trials.

There are several limitations of this study. First, the intention of this paper was to introduce a powerful statistical modeling tool to account for correlations that are the result of a complex study design and to demonstrate the modeling tool with a motivating study. The purpose was not to investigate the properties of the statistical model such as robustness or type I and type II errors. To understand the properties of the proposed model, further statistical simulation studies must be conducted. Second, the method presented here is applicable to normally distributed outcome variable only, if the outcome variable follows other distribution such as a binary, Poisson, or gamma distribution, the statistical technique is of limited use to account for the multilevel correlation. Further statistical method development is necessary for these distributions. Third, the method presented here is robust when the data structure construct contains more upper-level clusters than lower-level clusters. Therefore, when designing such studies, researchers should ensure that an adequate number of upper-level clusters exist for meaningful statistical analysis results. Fourth, the fact that the Box-Cox transformation was applied to the outcome variable might limit interpretations of the result. However, this is one of the commonly used methods to correct non-normality.

Acknowledgments

Funding for this work was received from the National Institute of Health, National Institute of Nursing Research, Grant # R01NR8565-4.

References

- Box EP, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*. 1964; 26:211–252.
- Casella, G.; Berger, RL. *Statistical inference*. 2. Belmont, CA: Duxbury Press; 2001.
- Demidenko, E. *Mixed models: Theory and applications*. Hoboken, NJ: Wiley; 2004.
- Draper D. Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*. 1995; 20:115–147.
- Estabrooks CA, Midodzi WK, Cummings GG, Wallin L. Predicting research use in nursing organizations: A multilevel analysis. *Nursing Research*. 2007; 56:S7–S23. [PubMed: 17625477]
- Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: Precision, sample size, and a unifying model. *Statistics in Medicine*. 1994; 13:61–78. [PubMed: 9061841]
- Greenland S. Principles of multilevel modeling. *International Journal of Epidemiology*. 2000; 29:158–167. [PubMed: 10750618]
- Herrera AP, Lee JW, Nanyonjo RD, Laufman LE, Torres-Vigil I. Religious coping and caregiver well-being in Mexican-American families. *Aging & Mental Health*. 2009; 13:84–91. [PubMed: 19197693]
- Kelly PJ, Lesser J, Cheng AL, Ocos-Sanchez M, Martinez E, Pineda D, et al. A prospective randomized controlled trial of an interpersonal violence prevention program with a Mexican-American community. *Family & Community Health*. 2010; 33:207–215. [PubMed: 20531101]
- Knafl GJ, Knafl KA, McCorkle R. Mixed models incorporating intra-familial correlation through spatial autoregression. *Research in Nursing & Health*. 2005; 28:348–356. [PubMed: 16028268]
- Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Littell RC, Henry PR, Ammerman CB. Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science*. 1998; 76(4):1216–1231. [PubMed: 9581947]
- Littell, RC.; Milliken, GA.; Stroup, WW.; Wolfinger, RD. *SAS system for mixed models*. Cary, NC: SAS Publishing; 1996.
- Mood, AM.; Graybill, FA.; Boes, DC. *Introduction to the theory of statistics*. 3. New York, NY: McGraw-Hill; 1974.
- Murray, DM. *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press; 1998.
- Rasbash J, Goldstein H. Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*. 1994; 19:337–350.
- Singer JD. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*. 1998; 23:323–355.
- Verbeke, G.; Molenberghs, G. *Linear mixed models for longitudinal data*. New York, NY: Springer; 2000.
- Zucker DM, Lakatos E, Webber LS, Murray DM, McKinlay SM, Feldman HA, ... & the CATCH Study Group. Statistical design of the child and adolescent trial for cardiovascular health (catch): Implications of cluster randomization. *Controlled Clinical Trials*. 1995; 16:96–118. [PubMed: 7789139]

Figure 1.
Familias project data structure at baseline, postintervention, and 3 months postintervention

Table 1**SAS Syntax Example for the Models**

```

/* FULL MODEL WITH SPATIAL COVARIANCE */
PROC MIXED DATA=FAMILIA COVTEST ;
  CLASS SCHOOL FAMID TIMEPOINT PARENT KID ;
  MODEL OUTCOME=TIME1 TIME2 TIME3 INTEGTIME1 INTEGTIME2 INTEGTIME3
  GENDER/ NOINT SOLUTION CL ;
  RANDOM INT /SUB=SCHOOL;
  REPEATED /SUBJECT=FAMID TYPE=SP(POWA)(KID PARENT TIMEPOINT) ;
RUN;

/* MODEL WITHOUT SPATIAL COVARIANCE */
PROC MIXED DATA=FAMILIA COVTEST ;
  CLASS SCHOOL FAMID TIMEPOINT PARENT KID ;
  MODEL OUTCOME =TIME1 TIME2 TIME3 INTEGTIME1 INTEGTIME2 INTEGTIME3
  GENDER /NOINT SOLUTION CL;
  RANDOM INT/SUBJECT=SCHOOL ;
  REPEATED INT/SUBJECT=FAMID TYPE=CS;
RUN;

/*NAIVE MODEL*/
PROC MIXED DATA=FAMILIA COVTEST ;
  CLASS SCHOOL FAMID TIMEPOINT PARENT KID ;
  MODEL OUTCOME =TIME1 TIME2 TIME3 INTEGTIME1 INTEGTIME2 INTEGTIME3
  GENDER/NOINT SOLUTION CL ;
RUN;

```


Table 2

Results of Different Model Specifications

	Full model with spatial covariance	Model without spatial covariance	Naïve model
Main Effect			
Baseline	1.3957 (0.0148)*	1.3961 (0.0133)*	1.3945 (0.0117)*
Postintervention	1.4197 (0.0149)*	1.4208 (0.0135)*	1.4177 (0.0119)*
3 months Postintervention	1.4042 (0.0149)*	1.4044 (0.0133)*	1.4033 (0.0118)*
Intervention Effect			
Baseline	0.0113 (0.0118)	0.0123 (0.0117)	0.0102 (0.0086)
Postintervention	0.0106 (0.0121)	0.0090 (0.0121)	0.0084 (0.0092)
3 months Postintervention	0.0251 (0.0121)*	0.0245 (0.0120)*	0.0208 (0.0090)*
Covariate			
Gender	−0.0168 (0.0068)*	−0.0169 (0.0057)*	−0.0158 (0.0056)*
Random Effect			
School	0.0002 (0.0001)	0.0001 (0.0001)	NA
Child	0.3594 (0.1205)*	NA	NA
Parent	0.5295 (0.2043)*	NA	NA
Time point	0.4504 (0.0276)*	NA	NA
Family	NA	0.0022 (0.0003)*	NA
Residual	0.0089 (0.0004)*	0.0067 (0.0003)*	0.0091 (0.0003)*
Model comparison			
AIC	−2828.8	−2747.6	−2587.0

Notes.

*
 $p < .05$