Journal of Biometrics & Biostatistics

Wang, J Biom Biostat 2016, 7:2
http://dx.doi.org/10.4172/2155-6180.1000290

Opinion Open Access

# Linear and Non-Linear Mixed Models in Longitudinal Studies and Complex Survey Data

**Ke-Sheng Wang***

*Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, PO Box 70259, Lamb Hall, Johnson City, TN 37614-1700, USA*

***Corresponding author:** Kesheng Wang, Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, PO Box 70259, Lamb Hall, Johnson City, TN 37614-1700, USA, Tel: +1 423 439 4481; Fax: +1 423 439 4606; E-mail: wangk@etsu.edu

**Rec date:** March 15, 2016; **Acc date:** March 17, 2016; **Pub date:** March 24, 2016



## Introduction

Correlated data are fairly common in health and social sciences. For example, clustered data arise when subjects are nested in clusters such as classrooms, hospitals, and neighborhoods; while longitudinal data result from multiple measures for the same subject over long period of time; whereas repeated measures data are involved in multiple measurements for the same subject over time or other dimension. Observations for the same cluster/subject are likely to be correlated (non-independent). Mixed models (also known as multilevel models or hierarchical models) including both fixed effects and random effects have been developed to deal with correlated data [1-3]. The random effect of one variable has a prior distribution with variance and varies randomly within the population; whereas the fixed effect of the variable is the average effect in the entire population, expressed by the regression coefficient.

Linear mixed models (LMMs) are extensions of linear regression models, which describe the relationship between a continuous response variable and independent variables, with fixed effects and random effects. Generalized linear mixed models (GLMMs) or non-linear mixed models (NLMMs) are extensions of LMMs to allow response variables from different distributions, such as binary, ordinal or count responses. Alternatively, GLMMs can be considered as extensions of generalized linear models (e.g., logistic regression or Poisson regression) to include both fixed and random effects. Mixed models can be fitted using statistical software packages such as SAS, SPSS, R, Stata, MlwiN, HLM and WinBUGS [1-5].

## Complex Survey Data

Secondary data analysis of large survey data sets may serve as an economical means to test specific hypotheses that have not been adequately examined and to confirm new findings or to aid in the development of new research questions. It is important to consider cluster variables, for example, neighborhoods and census blocks, as random factors in multilevel analysis of complex survey data.

For example, 7595 adults participated in one of five cross-sectional surveys conducted by the Stanford Heart Disease Prevention Program were used to determine whether socioeconomic and food-related physical characteristics of the neighbourhood are associated with body mass index (BMI; $kg/m^2$). The SAS PROC MIXED was used to estimate fixed effect coefficients for individual-level and neighbourhood-level variables while adjusting for random intercepts between neighbourhoods [6]. Once the data was stratified using the 137 metropolitan/micropolitan statistical areas (MMSA) in the U.S., a two-level hierarchical model (individual level and MMSA-level) was developed using the SAS PROC GLIMMIX to estimate the relationship between coronary heart disease prevalence and occupational structure in U.S. metropolitan areas using the data from the 2006-2008 Behavioral Risk Factor Surveillance System (BRFSS) [7]. Using a multilevel linear mixed-effects modeling approach, Richmond-Bryant et al. [8] evaluated the relationship between lead in blood (PbB) levels and lead in ambient air (PbA) levels among children. The data for PbB were from the National Health and Nutrition Examination Survey (NHANES) III (1988-1994) and NHANES (1999-2008) and PbA data were from the U.S. Environmental Protection Agency. The geographical location (census block group) was treated as a random factor with random effect.

Recently, based on the GLMMs, a new method using a weighted composite likelihood was proposed to estimate the effect of individual-level covariates on an outcome when adjusted for neighborhood-level confounding in complex survey data [9]. This method was evaluated via simulation studies and applied to investigate racial/ethnic disparities in oral health care using the data from the 2008 Florida BRFSS survey, where the neighborhood as zip code was merged into the BRFSS data [9,10].

## Longitudinal Studies

In longitudinal data, the dependent variable is measured several times and the individual subject is used as a random factor. The LMMs or GLIMMS can be used to account for repeated measures in longitudinal studies and also for random effects.

For example, to evaluate the performance of NLMMs, both the logit and probit NLMMs were evaluated for the modeling of mediated, binary longitudinal data using SAS PROC NLMIXED [11]. Furthermore, Q-Gen software was developed to apply for flexible LMMs that can account for confounding variables and random effects in longitudinal studies of gene set analysis of gene expression data [12]. In another study, the associations of 13 single nucleotide polymorphisms (SNPs) with fasting plasma glucose level and blood hemoglobin A1c content were examined in a 5-year longitudinal genetic epidemiological study of atherosclerotic, cardiovascular and metabolic diseases using a GLMM [4] with adjustment for age, gender and BMI [13]. Statistical analysis was performed with R software version 3-0-2 (The R Project for Statistical Computing) and JMP Genomics version 6.0 (SAS Institute, Inc., Cary, NC, USA).

Mixed models have been extended to analyze multiple outcomes in longitudinal study. For example, a novel application of multivariate GLMMs (mGLMM) was proposed to analyze multiple longitudinal kidney function outcomes collected over 3 years on a cohort of 110 renal transplantation patients [14]. This longitudinal data included 3 correlated continuous outcomes (blood urea nitrogen, serum creatinine, and estimated glomerular filtration rate) in assessment of patients' kidney function over time to monitor disease progression. As

stated, the mGLMM can incorporate random effects for single outcome and also for shared or separate random effects for multiple outcomes.

## Concluding Remarks

There is undoubtedly a growing interest in the development and application of mixed models in analysis of correlated data. The mixed models can vary in terms of levels, experimental designs (e.g., case-control, cross-sectional, longitudinal with repeated measures, and cross-classified), type of the outcome variable (e.g., binary, categorical, count, and continuous), and number of outcomes (e.g., univariate and multivariate) [3].

Some practical issues should be considered in analysis of clustered data such as outlier detection, model selection, estimation methods, and evaluation of model fit. Furthermore, missing values are common in clustered or longitudinal data sets, especially in longitudinal studies due to dropout. Generally, it is assumed that missing data are missing at random (MAR); however such assumption is untestable in most health and social sciences research. Moreover, Bayesian statistical methods have recently made great inroads into many areas of science including analysis of correlated data. For example, a new algorithm for fast Bayesian mixed model association (BOLT-LMM) was developed in genome-wide association study of quantitative traits, which can improve speed and power of existing mixed model association methods [15]. In addition, further development of mixed models will be prospective. For example, one recent study extended the NLMMs to the multivariate non-linear mixed models (MNMMs) by jointly analyzing the maternal antibody decay for 4 infectious diseases (measles, mumps, rubella, and varicella) [16].

In short, mixed models are commonly used to deal with correlated data or hierarchical data in health and social sciences. Also acknowledging random effects can increase the power in estimation of the fixed effects, avoid the serious inflation of the type I error rates, and prevent false positive results.

## References

1. Li B, Lingsma HF, Steyerberg EW, Lesaffre E (2011) Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. BMC Med Res Methodol 11: 77.
2. West BT, Welch KB, Galecki AT (2014) Linear Mixed Models: A Practical Guide Using Statistical Software Second Edition.
3. Ene M, Leighton EA, Blue GL, Bell BA (2015) Multilevel Models for Categorical Data Using SAS®PROC GLIMMIX: The Basics. SAS Global Forum, Dallas, USA.
4. Dean CB, Nielsen JD (2007) Generalized linear mixed models: a review and some extensions. Lifetime Data Anal 13: 497-512.
5. Austin PC (2010) Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. Int J Biostat 6: p: 16.
6. Wang MC, Kim S, Gonzalez AA, MacLeod KE, Winkleby MA (2007) Socioeconomic and food-related physical characteristics of the neighbourhood environment are associated with body mass index. J Epidemiol Community Health 61: 491-498.
7. Michimi A, Ellis-Griffith G, Nagy C, Peterson T (2013) Coronary heart disease prevalence and occupational structure in U.S. metropolitan areas: a multilevel analysis. Health Place 21: 192-204.
8. Richmond-Bryant J, Meng Q, Davis A, Cohen J, Lu SE, et al. (2014) The influence of declining air lead levels on blood lead-air lead slope factors in children. Environ Health Perspect 122: 754-760.
9. Brumback BA, Zheng HW, Dailey AB (2013) Adjusting for confounding by neighborhood using generalized linear mixed models and complex survey data. Stat Med 32: 1313-1324.
10. Brumback BA, Cai Z, Dailey AB (2014) Methods of estimating or accounting for neighborhood associations with health using complex survey data. Am J Epidemiol 179: 1255-1263.
11. Blood EA, Cheng DM (2012) Non-linear mixed models in the analysis of mediated longitudinal data with binary outcomes. BMC Med Res Methodol 12: 5.
12. Turner JA, Bolen CR, Blankenship DM (2015) Quantitative gene set analysis generalized for repeated measures, confounder adjustment, and continuous covariates. BMC Bioinformatics 16: 272.
13. Yamada Y, Matsui K, Takeuchi I, Oguri M, Fujimaki T (2015) Association of genetic variants of the a-kinase 1 gene with type 2 diabetes mellitus in a longitudinal population-based genetic epidemiological study. Biomed Rep 3: 347-354.
14. Jaffa MA, Gebregziabher M, Jaffa AA (2015) Analysis of multivariate longitudinal kidney function outcomes using generalized linear mixed models. J Transl Med 13: 192.
15. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, et al. (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 47: 284-290.
16. Goeyvaerts N, Leuridan E, Faes C, Van Damme P, Hens N (2015) Multi-disease analysis of maternal antibody decay using non-linear mixed models accounting for censoring. Stat Med 34: 2858-2871.