Multilevel model analysis using R

Nicolae-Marius JULA (mariusjula@yahoo.com)
Nicolae Titulescu University of Bucharest

ABSTRACT

The complex datasets cannot be analyzed using only simple regressions. Multilevel models (also known as hierarchical linear models, nested models, mixed models, random coefficient, random-effects models, random parameter models or split-plot designs) are statistical models of parameters that vary at more than one level.[1]

Multilevel models can be used on data with many levels, although 2-level models are the most common.

Multilevel models, or mixed effects models, can be estimated in R. There are several packages available in CRAN. In this paper we are presenting some common methods to analyze these models.

Keywords: Multilevel analysis, R, CRAN, package
Jel Classification: B23, C23, C33, C87

INTRODUCTION

Multilevel models are usually used in statistical analysis of data that have a hierarchical or clustered structure. One can find such data in various fields, like in educational research (schools – classes), social studies (families – members), medical research (patients nested within hospitals) and so on. Clustered data may also appear as a result of the particular research design. For instance, in large scale survey studies the data collection is usually organized in multistage sampling design that results in a clustered or a stratified design. Of course, this approach is not used exclusively in statistic studies, but the usual practice of these models are in this field of statistics or in a related one. There was a period of time when statisticians ignored this multilevel structure and they performed the analyses by simply disaggregating all the data to the lowest level and then using the common standard analyzing models.

This approach is not problems free. One of them is related to sampling variance, the so-called **design effect** (*deff*), detailed by Kish[2] in 1965. The

1. Bryk, Stephen W. Raudenbush, Anthony S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2. ed., [3. Dr.] ed.). Thousand Oaks, CA [u.a.]: Sage Publications. ISBN 0-7619-1904-X

2. Kish, L. (1965): *Survey Sampling.* Wiley, New York.

design effect can be seen as the loss of effectiveness by the use of cluster sampling, instead of simple random sampling. The design effect is basically the ratio of the actual variance, under the sampling method actually used, to the variance calculated under the assumption of simple random sampling.

As Turner stated: "The interpretation of a value of (the design effect) of, say, 3.0, is that the sample variance is 3 times bigger than it would be if the survey were based on the same sample size but selected randomly. An alternative interpretation is that only one-third as many sample cases would be needed to measure the given statistic if a simple random sample were used instead of the cluster sample with its (design effect) of 3.0"[1]

The design effect can be calculated:

$$DEFF = 1 + \delta (n - 1) \quad (1)$$

Where:

DEFF - design effect,

δ - intraclass correlation for the statistic in question,

n - average size of the cluster

Looking at this equation, *DEFF* equals 1 only when either the intraclass correlation is zero ($\delta = 0$), or the cluster size is one ($n = 1$). In all other situations *DEFF* is larger than one, which denotes that standard statistical formulas will underestimate the sampling variance, meaning that we may obtain significance tests with an inflated alpha level (type I error rate). There were tests conducted by Tate and Wongbundhit[2] and they stated that estimates of the regression in multilevel models are unbiased, but have a larger sampling variance compared to OLS methods resulted estimators. Using significance tests in multilevel structure models without considering this aspect could lead to misinterpretations.

MULTILEVEL REGRESSION MODEL

Multilevel models (also known as hierarchical linear models, nested models, mixed models, random coefficient, random-effects models, random parameter models, or split-plot designs) are statistical models of parameters that vary at more than one level. The models assume hierarchical data, in which the dependent variable is measured at the lowest level and the independent (explanatory) variables are measured at all available levels.

1. Turner, AG. (1996): *Sampling Topics for Disability Surveys*. United Nations Statistics Division, Technical Notes, December, http://www.undp.org/popin/demotss/tcndec96/tony.htm

2. Tate, R. & Wongbundhit, Y. (1983): *Random versus Nonrandom Coefficient Models for Multilevel Analysis*. Journal of Educational Statistics, 8, 103-120.

Level 1 regression can be seen as:

$$Y_t = a_0 + a_1X_t + e_t \tag{1}$$

Where:

Y_t – response variable,

a_0 – intercept,

a_1 – slope,

X_t – explanatory variable

e_t – residual

For example, let *J* be the number of groups and a different number of individuals Nj in each group. On the individual (lowest) level we have the dependent variable Yij and the explanatory variable Xij, and on the group level we have the explanatory variable Zj. Thus, a separate regression equation in each group can be written as following:

$$Yij = b0j + b1j\ Xij + eij. \tag{2}$$

The bj are modeled by explanatory variables at the group level:

$$b0j = g00 + g01\ Zj + u0j, \tag{3}$$

$$b1j = g10 + g11\ Zj + u1j. \tag{4}$$

Substitution of (3) and (4) in (2) gives:

$$Yij = g00 + g10\ Xij + g01\ Zj + g11\ ZjXij + u1j\ Xij + u0j + eij \tag{5}$$

In general, there will be more than one explanatory variable at the lowest level and also more than one explanatory variable at the highest level. Assume that we have *P* explanatory variables X at the lowest level, specified by the subscript p (p=1,P), and *Q* explanatory variables Z at the highest level, specified by the subscript q (q=1,Q).

Then, equation (5) becomes the more general equation:

$$Yij = g00 + gp0\ Xpij + g0q\ Zqj + gpq\ ZqjXpij + upj\ Xpij + u0j + eij \tag{6}$$

Another way to define a multilevel regression is:

For level 2:

$$Y_{ij} = a_0 + a_iX_{ij} + alpha_i + beta_j + e_{ij} \tag{7}$$

Where:

$alpha_i$ -> specific effect at level 1

$beta_j$ -> specific effect at level 2

$alpha_i$ and/or $beta_j$ can be analyzed as fixed or random. In the equation (7), if j=t we have a panel structure.

For level higher than 2, the second index may differ from “t”.
ex. Y_{ijk}, where:
k - product; j - firm; i - branch;
a_i -> constant a_i=a_1 (for any i)
-> not constant (variable)
-> constant for some explicative variables and not for others

The estimators used in multilevel analysis are Maximum Likelihood (ML), having standard errors estimated from the inverse of the information matrix. These standard errors are used in the Wald test (the test *Z = parameter / (st.err. param.)* is referred to the standard normal distribution to create a p-value for the null-hypothesis that in the population that specific parameter is null)

ACCURACY OF FIXED/RANDOM PARAMETERS AND THEIR STANDARD ERRORS

For the fixed parameters, the estimates for the regression coefficients appear generally unbiased, for OLS and GLS, as well as for ML estimation. OLS estimates seem to have a larger sampling error; Kreft[1] estimates that they are about 90% efficient.

There were simulation done by Van der Leeden & Busing[2] and Mok[3], and analytic work by Snijders & Bosker[4], which suggest that a large number of groups seems more important than a large number of individuals in the group for the precision of the results.

For the random parameters, the estimates of the residual error at the lowest level are mostly accurate. The group level variance components are generally underestimated (FML somewhat more that with RML). The findings stated that GLS variance estimates are less accurate than ML ones, and for accurate estimates many groups (>100) may be needed[5]

1. Kreft, Ita G.G. (1996): *Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies*. California State University, Los Angeles.

2. Van Der Leeden, R. & Busing, F. (1994): *First Iteration versus IGLS/RIGLS Estimates in Two-level Models: a Monte Carlo Study with ML3*. Department of Psychometrica and research Methodology, Leiden University, Leiden.

3. Mok, M. (1995): *Sample Size requirements for 2-level Designs in Educational Research. Multilevel Models Project*, University of London, London.

4. Snijders, T.A.B. & Bosker, R. (1993): *Modeled Variance in Two-level Models*. Journal of Educational Statistics, 18, 273-259.

5. Idem 7

ACCURACY AND SAMPLE SIZE

It is generally accepted that increasing sample sizes at all levels, estimates and their standard errors improve. Kreft suggests a "rule of thumb", which she calls the '30/30 rule.' To be statistically safe, researchers should use a sample of at least 30 groups with at least 30 individuals per group. From the various simulations presented above, this rule is of better use for fixed parameters. Some specialists suggest that the numbers should be modified as follows: if there is strong interest in cross-level interactions, the number of groups should be larger, (a 50/20 rule – 50 groups with 20 individuals/group); if there is stronger interest in the random part, or in the variance and/or covariance components, the amount of used groups should be considerably larger, which leads to a 100/10 rule (100 groups with 10 individuals/group). One should take into account the costs attached to data collection, so if the number of groups is increased, than the number of individuals per group might decreases.

MULTILEVEL ANALYSIS IN R

The widely used package in R for multilevel analysis is *lme4*. It is not installed by default, so one should call:

```
install.packages("lme4")
```

We use for this paper a modified example from Harvey Goldstein - Datasets used in Multilevel Statistical Models; 3rd edition 2003 (http://www.bristol.ac.uk/cmm/team/hg/msm-3rd-ed/datasets.html), converted into *csv* file format for ease of use.

The used dataset can be found here: http://www.bristol.ac.uk/cmm/team/hg/msm-3rd-ed/jsp-728.xls. There was a new column inserted "School_class", generated as a random number from 1 to 4.

For a complete listing, see Annex.

The used data set has the following colums:

- math_yr_3 – obtained result by a student in 3rd year
- math_yr_1 – obtainded result by a student in 1st year
- Gender - gender of the student (1 for masculine)
- Social_class
- School_class – randomized class associated to schools (1 to 4)
- School_ID

- Normal_score_yr_3 – average score for students in 3^{rd} year
- Normal_score_yr_1 – average score for students in 1^{st} year

First, we used a simple regression (OLS):

```
lm(formula = math_yr_3 ~ math_yr_1 + Gender + Social_class +
    School_ID + School_class, data = my.lmm.data)
             coef.est coef.se
(Intercept)  14.54     0.98
math_yr_1     0.64     0.03
Gender       -0.33     0.36
Social_class -0.71     0.40
School_ID     0.02     0.01
School_class -0.08     0.16
---
n = 728, k = 6
residual sd = 4.82, R-Squared = 0.47
```

There are several approaches when using multilevel analysis. We are presenting here the use of *lmer* function for two cases: varying intercept and varying slope.

1. Fit a varying intercept model with *lmer*

Group level variables can be specified using the syntax: (1|School_ID), which tells *lmer* to fit a linear model with a varying-intercept group effect using the variable School_ID

> MLL.Example.6 <- lmer(math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID), data = my.lmm.data)
> display(MLL.Example.6)

```
lmer(formula = math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID),
    data = my.lmm.data)
            coef.est coef.se
(Intercept) 14.10     0.73
math_yr_1    0.65     0.02
Gender      -0.35     0.34

Error terms:
 Groups    Name        Std.Dev.
 School_ID (Intercept) 1.81
 Residual              4.45
---
number of obs: 728, groups: School_ID, 48
AIC = 4309.5, DIC = 4286.9
deviance = 4293.2
```

Multiple group effects can be fitted with multiple group effect terms.
> MLL.Example.7 <- lmer(math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID) + (1 | School_class), data = my.lmm.data)
> display(MLL.Example.7)

```
lmer(formula = math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID) +
    (1 | School_class), data = my.lmm.data)
            coef.est coef.se
(Intercept) 14.10     0.73
math_yr_1    0.65     0.02
Gender      -0.35     0.34

Error terms:
 Groups       Name        Std.Dev.
 School_ID    (Intercept) 1.81
 School_class (Intercept) 0.00
 Residual                 4.45
---
number of obs: 728, groups: School_ID, 48; School_class, 4
AIC = 4311.5, DIC = 4286.9
deviance = 4293.2
```

The nested group effect terms can be fitted using the following syntax:
> MLL.Example.8 <- lmer(math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID/School_class), data = my.lmm.data)
> display(MLL.Example.8)

```
lmer(formula = math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID/
School_class),
    data = my.lmm.data)
            coef.est coef.se
(Intercept) 14.10     0.73
math_yr_1    0.65     0.02
Gender      -0.35     0.34

Error terms:
 Groups                Name        Std.Dev.
 School_class:School_ID (Intercept) 0.00
 School_ID             (Intercept) 1.81
 Residual                          4.45
---
number of obs: 728, groups: School_class:School_ID, 179; School_
ID, 48
AIC = 4311.5, DIC = 4286.9
deviance = 4293.2
```

Here the (1|School_ID/School_class) means that we want to fit a mixed effect term for varying intercepts 1| by schools and for classes that are nested within schools.

2. Fit a varying slope model with lmer

To analyze the effect of different student level indicators as they vary across School_class-es, as an alternative to fitting unique models by school (or School_ID/Social_class), a varying slope model can be fitted. The random effect term can be modified to include variables before the grouping terms:

(1 + Gender|School_ID/School_class)

Is interpreted by R to fit a varying slope and varying intercept model for schools and classes nested within schools, as well as to allow the slope of the open variable to vary by School_ID.

```
> MLL.Example.9 <- lmer(math_yr_3 ~ math_yr_1 + (1 + Gender| School_
ID/School_class),
         data = my.lmm.data)
> display(MLL.Example.9)
```

```
lmer(formula = math_yr_3 ~ math_yr_1 + (1 + Gender | School_ID/
School_class),
    data = my.lmm.data)
            coef.est coef.se
(Intercept) 13.96     0.72
math_yr_1    0.65     0.02

Error terms:
 Groups                Name        Std.Dev. Corr
 School_class:School_ID (Intercept) 0.24
                       Gender      0.48     -1.00
 School_ID             (Intercept) 1.76
                       Gender      0.11     1.00
 Residual                          4.44
---
number of obs: 728, groups: School_class:School_ID, 179; School_
ID, 48
AIC = 4318.1, DIC = 4288.1
deviance = 4294.1
```

CONCLUSIONS

This paper is just an introduction to the multilevel modeling in R. The used data is half generated data and the obtained results are just for exemplification of the functions.

Using R for multilevel modeling is an easy and powerful way to obtain the needed results. The high flexibility of accepted data formats recommend the use of R environment as an alternative to other commercial solutions. There are also other packages dealing with regression analysis and the on-going support of the community should help the analysts find the best available solution for their analyses.

References

1. Douglas Bates, Martin Maechler, Ben Bolker and Steven Walker (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-6. http://CRAN.R-project.org/package=lme4
2. Goldstein, M., (1999): Multilevel Statistical Methods, London, Institute of Education, Multilevel Models Project, April
3. Kish, L. (1965): Survey Sampling. Wiley, New York.
4. Knowles, J. E.(2013): Getting Started with Multilevel Modeling in R, http://jaredknowles.com/journal/2013/11/25/getting-started-with-mixed-effect-models-in-r
5. Kreft, Ita G.G. (1996): Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies. California State University, Los Angeles.
6. Mok, M. (1995): Sample Size requirements for 2-level Designs in Educational Research. Multilevel Models Project, University of London, London.
7. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
8. Snijders, T.A.B. & Bosker, R. (1993): Modeled Variance in Two-level Models. Journal of Educational Statistics, 18, 273-259.
9. Stephen B., Raudenbush W., Anthony S. (2002): Hierarchical Linear Models: Applications and Data Analysis Methods (2. ed., [3. Dr.] ed.). Thousand Oaks, CA [u.a.]: Sage Publications. ISBN 0-7619-1904-X
10. Tate, R. & Wongbundhit, Y. (1983): Random versus Nonrandom Coefficient Models for Multilevel Analysis. Journal of Educational Statistics, 8, 103-120.
11. Turner, AG. (1996): Sampling Topics for Disability Surveys. United Nations Statistics Division, Technical Notes, December, http://www.undp.org/popin/demotss/tcndec96/tony.htm
12. Van Der Leeden, R. & Busing, F. (1994): First Iteration versus IGLS/RIGLS Estimates in Two-level Models: a Monte Carlo Study with ML3. Department of Psychometrica and research Methodology, Leiden University, Leiden.

ANNEX

```
>
> my.lmm.data<-read.csv(file.choose())
>
> library(lme4)  # load library
> library(arm)  # convenience functions for regression in R
>
> head(my.lmm.data)
  math_yr_3 math_yr_1 Gender Social_class School_class School_ID Normal_score_yr_3 Normal_score_yr_1
1        39        36      1            0            4         1          1.802743          1.551093
2        11        19      0            1            1         1         -2.290740         -0.980330
3        32        31      0            1            3         1         -0.041320          0.638187
4        27        23      0            0            3         1         -0.749930         -0.459870
5        36        39      0            0            1         1          0.743105          2.149517
6        33        25      1            1            1         1          0.162541         -0.181760
>
> OLSexamp <- lm(math_yr_3 ~ math_yr_1 + Gender + Social_class + School_ID + School_class, data =
my.lmm.data)
> display(OLSexamp)
lm(formula = math_yr_3 ~ math_yr_1 + Gender + Social_class +
    School_ID + School_class, data = my.lmm.data)
             coef.est coef.se
(Intercept)  14.54     0.98
math_yr_1     0.64     0.03
Gender       -0.33     0.36
Social_class -0.71     0.40
School_ID     0.02     0.01
School_class -0.08     0.16
---
n = 728, k = 6
residual sd = 4.82, R-Squared = 0.47
> #summary(OLSexamp)
>
>
> MLL.Example <- glm(math_yr_3 ~ math_yr_1, data = my.lmm.data)
> display(MLL.Example)
glm(formula = math_yr_3 ~ math_yr_1, data = my.lmm.data)
            coef.est coef.se
(Intercept) 13.84     0.69
math_yr_1    0.65     0.03
---
  n = 728, k = 2
  residual deviance = 16945.4, null deviance = 31834.8 (difference = 14889.4)
  overdispersion parameter = 23.3
  residual sd is sqrt(overdispersion) = 4.83
```

```
>
>
> #Fit a varying intercept model
>
> MLL.Example.2 <- glm(math_yr_3 ~ math_yr_1 + Gender, data = my.lmm.data)
> display(MLL.Example.2)
glm(formula = math_yr_3 ~ math_yr_1 + Gender, data = my.lmm.data)
            coef.est coef.se
(Intercept) 14.01     0.71
math_yr_1    0.65     0.03
Gender      -0.36     0.36
---
  n = 728, k = 3
  residual deviance = 16922.1, null deviance = 31834.8 (difference = 14912.7)
  overdispersion parameter = 23.3
  residual sd is sqrt(overdispersion) = 4.83
>
> AIC(MLL.Example.2)
[1] 4364.317
>
>
> anova(MLL.Example, MLL.Example.2, test = "F")
Analysis of Deviance Table

Model 1: math_yr_3 ~ math_yr_1
Model 2: math_yr_3 ~ math_yr_1 + Gender
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1       726      16945
2       725      16922  1   23.321 0.9992 0.3178
>
> MLL.Example.3 <- glm(math_yr_3 ~ math_yr_1 + Gender + School_class, data = my.lmm.data)
> display(MLL.Example.3)
glm(formula = math_yr_3 ~ math_yr_1 + Gender + School_class,
    data = my.lmm.data)
             coef.est coef.se
(Intercept)  14.15     0.82
math_yr_1     0.65     0.03
Gender       -0.36     0.36
School_class -0.05     0.16
---
  n = 728, k = 4
  residual deviance = 16919.5, null deviance = 31834.8 (difference = 14915.4)
  overdispersion parameter = 23.4
  residual sd is sqrt(overdispersion) = 4.83
```

```
>
> AIC(MLL.Example.3)
[1] 4366.204
>
>
> anova(MLL.Example, MLL.Example.3, test = "F")
Analysis of Deviance Table

Model 1: math_yr_3 ~ math_yr_1
Model 2: math_yr_3 ~ math_yr_1 + Gender + School_class
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1       726      16945
2       724      16920  2   25.951 0.5552 0.5742
>
> table(my.lmm.data$Gender, my.lmm.data$School_class)

      1   2   3   4
  0 102 100  98  87
  1  91  82  79  89
>
>
> MLL.Example.4 <- glm(math_yr_3 ~ math_yr_1 + Gender + School_ID : School_class, data = my.lmm.data)
> display(MLL.Example.4)
glm(formula = math_yr_3 ~ math_yr_1 + Gender + School_ID:School_class,
    data = my.lmm.data)
                       coef.est coef.se
(Intercept)            13.88     0.75
math_yr_1               0.65     0.03
Gender                 -0.36     0.36
School_ID:School_class  0.00     0.00
---
  n = 728, k = 4
  residual deviance = 16915.2, null deviance = 31834.8 (difference = 14919.7)
  overdispersion parameter = 23.4
  residual sd is sqrt(overdispersion) = 4.83
>
>
> #Fit a varying intercept model with lmer
>
> MLL.Example.6 <- lmer(math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID ), data = my.lmm.data)
> display(MLL.Example.6)
lmer(formula = math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID),
    data = my.lmm.data)
            coef.est coef.se
(Intercept) 14.10     0.73
math_yr_1    0.65     0.02


Gender      -0.35     0.34

Error terms:
 Groups    Name        Std.Dev.
 School_ID (Intercept) 1.81
 Residual              4.45
---
number of obs: 728, groups: School_ID, 48
AIC = 4309.5, DIC = 4286.9
deviance = 4293.2
>
>
> #We can fit multiple group effects with multiple group effect terms.
>
> MLL.Example.7 <- lmer(math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID) + (1 | School_class), data =
my.lmm.data)
> display(MLL.Example.7)
lmer(formula = math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID) +
    (1 | School_class), data = my.lmm.data)
            coef.est coef.se
(Intercept) 14.10     0.73
math_yr_1    0.65     0.02
Gender      -0.35     0.34

Error terms:
 Groups       Name        Std.Dev.
 School_ID    (Intercept) 1.81
 School_class (Intercept) 0.00
 Residual                 4.45
---
number of obs: 728, groups: School_ID, 48; School_class, 4
AIC = 4311.5, DIC = 4286.9
deviance = 4293.2
>
> #nested group effect terms
> display(MLL.Example.8)
lmer(formula = math_yr_3 ~ math_yr_1 + Gender + (1 | School_ID/School_class),
    data = my.lmm.data)
            coef.est coef.se
(Intercept) 14.10     0.73
math_yr_1    0.65     0.02
Gender      -0.35     0.34

Error terms:
 Groups                 Name        Std.Dev.
 School_class:School_ID (Intercept) 0.00
```

```
 School_ID              (Intercept) 1.81
 Residual                           4.45
---
number of obs: 728, groups: School_class:School_ID, 179; School_ID, 48
AIC = 4311.5, DIC = 4286.9
deviance = 4293.2
>
> #Fit a varying slope model with lmer
> MLL.Example.9 <- lmer(math_yr_3 ~ math_yr_1  + (1 + Gender| School_ID/School_class),
+                    data = my.lmm.data)
> display(MLL.Example.9)
lmer(formula = math_yr_3 ~ math_yr_1 + (1 + Gender | School_ID/School_class),
    data = my.lmm.data)
            coef.est coef.se
(Intercept) 13.96     0.72
math_yr_1    0.65     0.02

Error terms:
 Groups                 Name        Std.Dev. Corr
 School_class:School_ID (Intercept) 0.24
                        Gender      0.48     -1.00
 School_ID              (Intercept) 1.76
                        Gender      0.11     1.00
 Residual                           4.44
---
number of obs: 728, groups: School_class:School_ID, 179; School_ID, 48
AIC = 4318.1, DIC = 4288.1
deviance = 4294.1
```