Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Multi Level Models

When there are [say] subjects within hospitals and [say] hospitals within regions, one can argue for several 'new' variances [and maybe covariances] .

Now, let us consider a multi level model with more than two levels. The right hand side of the model equation will be of the form:

$$X\beta + Z^{(4)}\boldsymbol{u}^{(4)} + Z^{(3)}\boldsymbol{u}^{(3)} + Z^{(2)}\boldsymbol{u}^{(2)}$$

where $\boldsymbol{u}^{(4)}$ provides for the region components where $\boldsymbol{u}^{(3)}$ provides for the hospital components and $\boldsymbol{u}^{(2)}$ provides for the subject components. It is anticipated that a correct formulation of such models will enable the estimates [and standard errors] of the primary regression coefficients [the vector β] to have the desired properties. The estimates can then be interpreted as being 'adjusted' for regions, hospitals and subjects. Intermediate models with some components missing will have interpretations reflecting 'partial' adjustment or 'crude' description.

For this example, we would want the regression coefficients that are 'between region' comparisons to reflect region differences in their estimates and standard errors, 'between hospital' comparisons that are 'within regions' comparisons to reflect the such differences, 'between subjects' comparisons that are 'within hospitals' to reflect such differences and lastly 'within subjects' comparisons to be handled correctly.

The predictions of components will satisfy additional constraints [to ensure identifiability of all concerned] and the interpretations will change [as always]. We will have two more variances to estimate and interpret. $\sigma^2_{u^{(3)}}$ and $\sigma^2_{u^{(4)}}$. Normality assumptions for [$\boldsymbol{u}^{(3)}$ and $\boldsymbol{u}^{(4)}$] are in play here too. Components for slopes [and more elaborate] are also available. All of these matters will require assessments [in principle] with inference methods and graphical displays.

Lets try out a study of conditional means [assuming normality of the errors]

$$\boldsymbol{y} = X\beta + Z^{(3)}\boldsymbol{u}^{(3)} + Z^{(2)}\boldsymbol{u}^{(2)} + \epsilon$$

We will illustrate with a study of math achievement scores (Anderson et al 2009 from West et al): 1,190 first-grade students from 312 classrooms in 107 schools are used for this example. The outcome of interest is mathgain which measures change in student math achievement scores from the spring of kindergarten to the spring of first grade. Students (Level 1) are nested within classrooms (Level 2), and classrooms are nested within schools (Level 3). We can see that mathprep is the same for all students in the same classroom and hence mathprep is a part of between classrooms. The value of housepov is the same for all classrooms within a given school and hence is a part of between schools. mathgain and sex are specific to each student and so they are within classrooms. The data is in classroom.dta

```
. mixed mathgain mathkind sex minority ses housepov || schoolid: || classid:

Mixed-effects ML regression                     Number of obs     =      1,190

------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
       schoolid |        107          2       11.1         31
        classid |        312          1        3.8         10
------------------------------------------------------------

                                                Wald chi2(5)      =     451.39
Log likelihood = -5694.8221                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
    mathgain |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mathkind |  -.4706098   .0222209   -21.18   0.000     -.514162   -.4270576
         sex |  -1.230705   1.654031    -0.74   0.457    -4.472547    2.011137
    minority |  -7.745651   2.372263    -3.27   0.001     -12.3952   -3.096101
         ses |   5.232545   1.241811     4.21   0.000     2.798641    7.666449
    housepov |  -11.30137   9.822189    -1.15   0.250    -30.55251    7.949768
       _cons |   284.9108   10.98678    25.93   0.000     263.3771    306.4445
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
schoolid: Identity           |
                  var(_cons) |    73.7545   25.13606      37.81751    143.8415
-----------------------------+------------------------------------------------
classid: Identity            |
                  var(_cons) |    81.3245   28.93938      40.48748    163.3511
-----------------------------+------------------------------------------------
               var(Residual) |   732.0155   34.50772       667.412    802.8723
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 62.78                 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

The same R analysis

```
classroom <- read.csv("classroom.csv")
summary(lmer(mathgain ~ mathkind + sex + minority + ses + housepov + (1|schoolid) + (1|
classid), classroom, na.action = "na.omit", REML = FALSE))
```

Now let us consider a logistic regression with three levels:

$\text{logit}(\boldsymbol{p}) = X\beta + Z^{(3)}\boldsymbol{u}^{(3)} + Z^{(2)}\boldsymbol{u}^{(2)}$ where $\boldsymbol{p} = Pr(\boldsymbol{y} = \mathbf{1})$

Consider a study of patient, physician and hospital and how they are related to whether a patient's lung cancer goes into remission after treatment. Part of a larger study of treatment outcomes and quality of life in patients with lung cancer. [remission.dta] [from ats.ucla.edu]

Here is the analysis of a three level logistic regression with components for doctors [level 2] and for hospitals [level 3]. Patients are level 1. In this example, doctors are nested within hospitals, meaning that each doctor belongs to one and only one hospital and patients are nested within doctors so each patient has only one doctor.

For this illustration, we will consider the following explanatory variables :

1) between patient: age - age in years, los - length of stay in hospital [days], famhx - family history, canst - cancer stage, il6 - Interleukin 6

2) between doctor: experience - years of doctor's experience

3) between hospital: medicaid
did - doctor ID hid - hospital ID

```
. melogit remission age los i.famhx il6 i.canst experience medicaid || hid: ||
did:,intpoints(25)

Mixed-effects logistic regression               Number of obs     =      8,525

------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
            hid |         35        134      243.6        377
            did |        407          2       20.9         40
------------------------------------------------------------

Integration method: mvaghermite                 Integration pts.  =         25

                                                Wald chi2(9)      =     532.18
Log likelihood = -3582.1085                     Prob > chi2       =     0.0000
------------------------------------------------------------------------------
   remission |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0159199   .0060628    -2.63   0.009    -.0278027    -.004037
         los |  -.0440304   .0364266    -1.21   0.227    -.1154253    .0273644
             |
       famhx |
        yes  |   -1.30622   .0954458   -13.69   0.000    -1.493291    -1.11915
         il6 |  -.0585866   .0117307    -4.99   0.000    -.0815784   -.0355947
             |
       canst |
          II |  -.3224089   .0785303    -4.11   0.000    -.4763255   -.1684923
         III |  -.8614403   .1026131    -8.40   0.000    -1.062558   -.6603224
          IV |  -2.160296   .1655747   -13.05   0.000    -2.484817   -1.835776
             |
  experience |    .125612   .0277104     4.53   0.000     .0713005    .1799234
    medicaid |   1.009479   .6618775     1.53   0.127    -.2877774    2.306735
       _cons |  -2.239379   .6752311    -3.32   0.001    -3.562808   -.9159503
-------------+----------------------------------------------------------------
hid          |
   var(_cons)|   .2324021   .1578259                      .061402    .8796242
-------------+----------------------------------------------------------------
hid>did      |
   var(_cons)|   3.995935   .4213358                     3.249876    4.913263
------------------------------------------------------------------------------
LR test vs. logistic model: chi2(2) = 2470.00             Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Attempted R analysis. glmer does not enable nAGQ>1 with more than 2 levels.

```
remission <- read.csv("remission.csv")
summary(glmer(remission ~ age + los + famhx + il6 + factor(canst) + experience +
medicaid+(1|hid) +(1|did), data=remission, family=binomial,glmerControl(optimizer="bobyqa",
optCtrl = list(maxfun = 100000))))
```

The next example will enable us consider logistic regression again and also proportion odds.

The Television School and Family Smoking Prevention and Cessation Project (TVSFP) study [Flay et al., 1988].
The study involved seventh-grade students from 135 classrooms from 28 schools, where the schools

were randomized to one of four study conditions: (a) a social-resistance classroom curriculum, (b) a media (television) intervention, (c) a social- resistance classroom curriculum combined with a mass-media intervention, and (d) a no- treatment control group. These conditions enable the assessment of the interaction of social-resistance classroom curriculum (resist = yes(1) or no(0)) by mass-media intervention (tele = yes(1) or no(0)). A tobacco and health knowledge scale (THKS) was used in classifying subjects as knowledgeable or not. Data from 1600 students with pre and post-intervention data were available. [in flay.dta] Here, 'level 3' would be the school and 'level 2' will be the classroom within the school.

```
use flay.dta
gen tr=tele*resist
table post resist tele,row
-----------------------------------
          |     tele and resist
          | ---- 0 ---    ---- 1 ---
     post |    0     1       0     1
----------+------------------------
        0 |   20    11      25    11
        1 |   97    51      85    55
        2 |  129    78     105    86
        3 |   89   106      91   114
        4 |   62    83      71    80
        5 |   21    35      32    31
        6 |    2    16       5     5
        7 |    1             2     1
          |
    Total |  421   380     416   383
-----------------------------------

gen fail=(post<=2)
table fail resist tele,row

-----------------------------------
          |     tele and resist
          | ---- 0 ---    ---- 1 ---
     fail |    0     1       0     1
----------+------------------------
        0 |  175   240     201   231
        1 |  246   140     215   152
          |
    Total |  421   380     416   383
-----------------------------------

. melogit fail pre resist tele tr || school: || class:

Mixed-effects logistic regression               Number of obs      =      1600

------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
         school |       28         18       57.1        137
          class |      135          1       11.9         28
------------------------------------------------------------

Integration method: mvaghermite                 Integration points =         7

                                                Wald chi2(4)       =     91.42
Log likelihood =  -1027.851                     Prob > chi2        =    0.0000
------------------------------------------------------------------------------
        fail |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pre |   -.395364   .0463324    -8.53   0.000    -.4861738   -.3045541
```

```
      resist |  -1.038282   .2447573    -4.24   0.000    -1.517998   -.5585667
        tele |  -.3325112    .235739    -1.41   0.158    -.7945512    .1295289
          tr |   .4643693   .3426676     1.36   0.175    -.2072469    1.135986
       _cons |   1.246478   .1956927     6.37   0.000     .8629273    1.630029
-------------+------------------------------------------------------------------
school       |
   var(_cons)|   .0628837   .0616755                      .009198    .4299174
-------------+------------------------------------------------------------------
school>class |
   var(_cons)|   .1649057   .0813311                      .0627227    .4335572
------------------------------------------------------------------------------
LR test vs. logistic regression:     chi2(2) =    17.61   Prob > chi2 = 0.0001

Note: LR test is conservative and provided only for reference.
```

Compare the 3 level analysis with a 'naive' analysis:

```
logit fail pre resist tele tr
Logistic regression                                     Number of obs   =       1600
                                                        LR chi2(4)      =     139.23
                                                        Prob > chi2     =     0.0000
Log likelihood = -1036.6576                             Pseudo R2       =     0.0629

------------------------------------------------------------------------------
        fail |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pre |  -.3996965   .0440785    -9.07   0.000    -.4860887   -.3133043
      resist |  -.9725355   .1499846    -6.48   0.000      -1.2665    -.678571
        tele |  -.3155662   .1434481    -2.20   0.028    -.5967192   -.0344132
          tr |   .4127295   .2098317     1.97   0.049     .0014668    .8239921
       _cons |   1.217092   .1411995     8.62   0.000     .9403458    1.493838
------------------------------------------------------------------------------
```

Here is an example of proportional odds with three levels.

```
gen post2=post
replace post2=5 if post2>5
meologit post2 resist tele tr || school: || class:

Mixed-effects ologit regression                 Number of obs      =      1600

-----------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+------------------------------------------
         school |       28         18       57.1        137
          class |      135          1       11.9         28
-----------------------------------------------------------

Integration method: mvaghermite                 Integration points =         7

                                                Wald chi2(3)       =     13.26
Log likelihood = -2626.5587                     Prob > chi2        =    0.0041
------------------------------------------------------------------------------
       post2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      resist |   .8127511    .250087     3.25   0.001     .3225896    1.302913
        tele |   .2231896   .2453706     0.91   0.363     -.257728    .7041073
          tr |  -.4179726   .3509259    -1.19   0.234    -1.105775    .2698296
-------------+----------------------------------------------------------------
       /cut1 |  -2.873268   .2088559   -13.76   0.000    -3.282618   -2.463918
       /cut2 |  -.9323697   .1777779    -5.24   0.000    -1.280808   -.5839314
       /cut3 |   .2949082   .1761042     1.67   0.094    -.0502497    .6400661
       /cut4 |   1.445256   .1801406     8.02   0.000     1.092186    1.798325
       /cut5 |   2.817702   .1933811    14.57   0.000     2.438682    3.196722
-------------+----------------------------------------------------------------
```

```
school       |
   var(_cons)|   .1018239   .0575867                      .0336087    .3084951
-------------+------------------------------------------------------------------
school>class |
   var(_cons)|   .1632315   .0667987                       .073193    .3640312
------------------------------------------------------------------------------
LR test vs. ologit regression:       chi2(2) =    39.52   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Again, we can compare the 3 level analysis with a 'naive' analysis:

```
ologit post2 resist tele tr
Ordered logistic regression                       Number of obs   =       1600
                                                  LR chi2(3)      =      48.60
                                                  Prob > chi2     =     0.0000
Log likelihood = -2646.3178                       Pseudo R2       =     0.0091

------------------------------------------------------------------------------
       post2 |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      resist |   .7820853   .1263871     6.19   0.000      .534371      1.0298
        tele |   .2201797   .1235115     1.78   0.075    -.0218984     .4622577
          tr |  -.3844389   .1776945    -2.16   0.031    -.7327137    -.0361641
-------------+----------------------------------------------------------------
       /cut1 |  -2.783793   .1419802                     -3.062069   -2.505517
       /cut2 |  -.8922774   .0922024                     -1.072991    -.711564
       /cut3 |   .2720694   .0892523                      .0971381    .4470007
       /cut4 |    1.36275   .0956417                      1.175296    1.550205
       /cut5 |   2.691513   .1171768                       2.46185    2.921175
------------------------------------------------------------------------------
```

Now, let us consider a(nother dental) longitudinal study with 3 levels. The study [veneer.dta] was investigating the impact of veneer placement on subsequent gingival (gum) health among adult patients (Ocampo, 2005). Ceramic veneers were applied to selected teeth to hide discoloration. The treatment process involved removing some of the surface of each treated tooth, and then attaching the veneer to the tooth with an adhesive. The veneer was placed to match the original contour of the tooth as closely as possible. The investigators were interested in studying whether varying amounts of contour difference (CDA) due to placement of the veneer might affect gingival health in the treated teeth over time. One measure of gingival health was the amount of GCF in pockets of the gum adjacent to the treated teeth. GCF was measured for each tooth at visits 3 months and 6 months post- treatment. Each patient could have different numbers of treated teeth, and the particular teeth that were treated could differ by patient.
Patient (Level 3) : patient = Patient ID variable (Level 3 ID)
age = Age of patient when veneer was placed; constant for all observations on the same patient Tooth (Level 2) : tooth = Tooth number (Level 2 ID) base_cgf= Baseline measure of cgf for the tooth; constant for all observations on the same tooth
cda = Average contour difference in the tooth after veneer placement; constant for all observations on the same tooth Time-Varying (Level 1) : time = Time points of longitudinal measures (3 = 3 Months, 6 = 6 Months)
outcome: cgf = Gingival crevicular fluid adjacent to the tooth, collected at each time point

```
. mixed gcf time base_gcf cda age tbg tc ta || patient: time, cov(unstruct) || tooth:
Mixed-effects ML regression                     Number of obs     =        110

------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
        patient |         12          2        9.2         12
```

```
          tooth |         55          2        2.0          2
------------------------------------------------------------

                                                Wald chi2(7)      =      11.24
Log likelihood = -421.82522                     Prob > chi2       =     0.1283

------------------------------------------------------------------------------
         gcf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |  -6.105889   6.831365    -0.89   0.371    -19.49512    7.283341
    base_gcf |  -.3176078   .2834754    -1.12   0.263    -.8732095    .2379938
         cda |  -.8844184    1.04596    -0.85   0.398    -2.934463    1.165626
         age |  -.9792811   .5524562    -1.77   0.076    -2.062075    .1035132
         tbg |   .0673982    .056224     1.20   0.231    -.0427988    .1775953
          tc |   .1298332   .2122767     0.61   0.541    -.2862215    .5458879
          ta |   .1105614   .1511103     0.73   0.464    -.1856093    .4067321
       _cons |   70.47211   26.10902     2.70   0.007     19.29937    121.6449
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
patient: Unstructured        |
                   var(time) |   36.70701   15.97778      15.64008    86.15072
                  var(_cons) |   447.1178   209.3122      178.6255    1119.181
              cov(time,_cons) |  -122.2297   56.09959     -232.1829   -12.27657
-----------------------------+------------------------------------------------
tooth: Identity              |
                  var(_cons) |   45.14037   15.68253      22.84774    89.18402
-----------------------------+------------------------------------------------
               var(Residual) |   47.48507   10.21156      31.15359    72.37792
------------------------------------------------------------------------------
LR test vs. linear model: chi2(4) = 86.33                 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

R handles nested and crossed components differently from Stata. For the correct R nested analysis, one needs a unique code for each tooth.

```
veneer <- read.csv("veneer.csv")
veneer$tooth2 <- as.numeric(paste(factor(veneer$patient), factor(veneer$tooth),sep=""))
summary(lmer(gcf ~ time + base_gcf + cda + age + tbg + tc + ta +(time | patient) + (1 |
tooth2), data = veneer, REML = F))
```

One more example from West et al [2015] to illustrate crossed [student and teacher] components. See sdf.dta

For crossed components, you have to fool Stata [which is set up for nested components]

```
use sdf.dta
mixed math year || _all: R.studid || _all: R.tchrid
```

The R code looks the 'same' as for nested components. lmer uses the codes to determine which is which.

```
sdf <- read.csv("sdf.csv")
summary(lmer(math ~ year + (1|studid) + (1|tchrid), school_data_final,REML = F))
```

Further on nested and crossed [adapted from lme4 manual : Bates(2012)]

"Consider an investigation [penicillin.csv] to assess the variability between samples of penicillin by the B. subtilis method. In this test method a bulk-innoculated nutrient agar medium is poured into a Petri dish of approximately 90 mm. diameter, known as a plate. When the medium has set, six small hollow cylinders or pots (about 4 mm. in diameter) are cemented onto the surface at equally spaced intervals. A few drops of the penicillin solutions to be compared are placed in the respective cylinders, and the

whole plate is placed in an incubator for a given time. Penicillin diffuses from the pots into the agar, and this produces a clear circular zone of inhibition of growth of the organisms, which can be readily measured. The diameter of the zone is related in a known way to the concentration of penicillin in the solution.
The variation in the diameter is associated with the plates and with the samples. Because each plate is used only for the six samples we are not interested in the contributions of specific plates as much as we are interested in the variation due to plates and in assessing the potency of the samples after accounting for this variation. Thus, we will condition on the plates. Also, we are more interested in the sample-to-sample variability in the penicillin samples than in the potency of a particular sample. Thus, we wish to condition on samples. In this experiment, each sample is used on each plate. We say that the sample and plate are crossed, as opposed to nested. By itself, the designation “crossed” just means that the factors are not nested. If we wish to be more specific, we could describe these factors as being completely crossed, which means that we have at least one observation for each combination of a level of sample and a level of plate.

```
lmer(diameter ~ 1 + (1|plate) + (1|sample), Penicillin)
```

The conditional distribution for a particular sample, say sample F, has less variability than the conditional distribution for a particular plate, say plate m.
Some presentations/texts leave the impression that one can only define components with respect to factors that are nested. This is the origin of the terms “multilevel”, referring to multiple, nested levels of variability, and “hierarchical”, also invoking the concept of a hierarchy of levels. Some references do describe the use of models with non-nested conditioning, but such models tend to be treated as a special case. The blurring of "mixed-effects" models with the concept of multiple, hierarchical levels of variation results in an unwarranted emphasis on “levels” when defining a model and leads to considerable confusion. It is perfectly legitimate to define models having random effects associated with non-nested factors. In the lme4 package, there is nothing special done for models with components that are nested. The same computational methods are used whether the factors form a nested sequence or are partially crossed or are completely crossed. A case of a nested sequence of “grouping factors” for the random effects (including the trivial case of only one such factor) is detected but this information does not change the course of the computation. It is available to be used as a diagnostic check. When the user knows that the grouping factors should be nested, the user can check if they are indeed nested."