Chapter 3 Linear Mixed Models with Covariates

Paul E. Johnson

February 1, 2016

1 Glossary

Since this book uses unfamiliar notation I'll never remember....
ξ_{ij}: "xi" individual row-level error term for group i row j,
ζ_j: "zeta" a group level random error, j indexes a grouping variable, $Var(\zeta_j) = \psi$
ϵ_{ij}: "epsilon" $Var(\epsilon_{ij}) = \theta$. Individual row-level error uncorrelated with ζ_j
$Cov(y_{ij}, y_{i'j}|\zeta_j) = 0$. Apart from ζ_j, the observed scores are "conditionally uncorrelated"
The combined error term has two variance components

$$\xi_{ij} = \zeta_j + \epsilon_{ij} \tag{1}$$

2 Highlights

1. Running example: Study babies grouped by mothers.
2. Standard deviation
   a) Overall: deviations of cases about overall (grand mean)
   b) Between standard deviation: Deviations of group means around overall (grand) mean
   c) Within standard deviation: Deviations of individual row values from group means
3. Linear Random Intercept model

$$y_{ij} = \beta_1 + \beta_2 x_{2ij} + \ldots + \beta_p x_{pij} + \xi_{ij},\ \xi_{ij} = \zeta_j + \epsilon_{ij} \tag{2}$$

   Two interesing rearrangements
   a) A two level residual

$$y_{ij} = \beta_1 + \beta_2 x_{2ij} + \ldots + \beta_p x_{pij} + (\zeta_j + \epsilon_{ij}) \tag{3}$$

b) A random intercept

$$y_{ij} = (\beta_1 + \epsilon_{ij}) + \beta_2 x_{2ij} + \ldots + \beta_p x_{pij} + \epsilon_{ij}) \tag{4}$$

4. Remember assumptions, all the errors are well behaved.
   a) Independently drawn observations within each level
   b) Uncorrelated with across levels
   c) Uncorrelated with predictors
   d) Homoskedastic

5. p. 130. The Population Model. When we let $\zeta_j = 0$ and $\epsilon_{ij} = 0$, we have the baseline population model. I think of this as the "most likely" or modal district. Some call it the "population averaged" or "marginal regression".

6. Intraclass correlation

$$\rho = \frac{\psi}{\psi + \theta} \tag{5}$$

   a) This is a theoretical quantity, the description of group effects versus individual random effects.
   b) We can get various estimates of these quantities, use as guides on guessing that the group random effects are important.
   c) p. 131. **Love that Graph!!**
   d) The "empty" or "null" model, AKA "Unconditional model". Include only the intercept and the random effects, no covariates. That gives a baseline view.

7.

8. p. 135 Variance explained: Things like R^2.
   a) Honestly: why don't we take the simple approach of calculating Pearson's R between observed and predicted values, and square that?
   b) p. 135: Snidjers and Bosker proposal compares the "null model" estimates and the fitted model estimates.

$$R^2 = \frac{\hat{\psi}_0 - \hat{\psi}_1 + \hat{\theta}_0 - \hat{\theta}_1}{\hat{\psi}_0 + \hat{\theta}_0} \tag{6}$$

   answers question: How much less unexplained variance is there in the fitted model 1 than in the empty model 0?
   c) p. 136: Raudenbush and Bryk instead propose focus on the reduction in variance components separately, look for proportional decline in $\hat{\psi}_1$ and $\hat{\theta}_0$ separately.

d) p. 137. *Here's a neat observation.* **This is worth the price of admission by itself.** "... adding level-1 covariates can reduce both variances, as we can see by comparing the estimates for the above model with the full model. The reason is that many level-1 covariates vary both within and between clusters and can be decomposed as $x_{ij} = (x_{ij} - \bar{x}_{.j}) + \bar{x}_{.j}$, where $x_{ij} - \bar{x}_{.j}$ only varies at level-1 and $\bar{x}_{.j}$only varies at level-2. Note that the estimated level-2 variance can increase when adding level-1 covariates, potentially producing a negative R_2^2."

Whats so neat? This explains why we can take a predictor and divide it into the within and between effects.

e) p. 137: Small R^2 does not mean "this is the wrong model." The true model might have high error variance

f) p. 137: **Warning**: Calculate the ICC with the null model, 0. Calculate ICC with model that has predictors. The ICC may change in hard-to-understand ways. Both variances may shrink, but if the relative shrinkage differs, then ICC may rise or fall after adding predictors.

9. p. 138: Hypothesis Tests

a) z tests on individual fixed coefficients

b) LR test can also be used to test individual or groups of coefficients

c) Tests like $H_0 : \beta_2 = \beta_3 = 0$

i. This is the same test we would do in ordinary regression. Could fit model with and without those and do LR test on difference.

ii. p. 139 creates a Wald type test that is usually similar/identical. See the testparm function in stata.

iii. The LR test requires calculation of both models.

iv. The Wald test can be calculated if we just fit the bigger model.

d) p. 142. LR test is used to decide if a variance of a random effect is actually 0.

10. Between and within effects.

a) Between regression: regression of means on means. xtreg , re

i. p. 145. SEE big table, comparing RE, Between, Within estimates.

b) Within-mother effecs.

i. re-code each variable as deviation from the group mean

$(y_{ij} - \bar{y}_{.j}) = \beta_2(x_{2ij} - \bar{x}_{2.j}) + \beta_3(x_{3ij} - \bar{x}_{3.j}) + \ldots + \epsilon_{ij} - \bar{\epsilon}_{.j}$

A. WOW! Look, the intercept β_1 and the group random effect ζ_j "canceled out", disappeared!

B. This is IDENTICAL to a Fixed Effects regression, that includes all of the group dummy-variable intercepts as predictors.

C. The fixex effects model has lots of dummy variable coefficient estimates.

D. Output has "sigma_u" and "sigma_e". Sigma u looks like a random effect standard deviation, but it is not, it is just the deviation observed among intercepts (p. 147). I am pretty sure.

c) p. 147 Random intercept model "can be expressed as a weighed average of the within estimator and the between estimator."

i. Goes into discussion of random effect fixed effects estimator from FGLS, which I expect is not well understood here because many readers don't know about FGLS.

d) p. 149. Confounding: Omitted covariates, level-2 "endogeneity". This is interesting, not entirely understandable to me.

i. Mothers who smoke engage in other risky behaviors. What is the damage of these unmeasured variables?

ii. Smoking is measured, it gets blamed for damage caused by other behaviors.

iii. Within model not affected because all level-2 variables that don't vary are "canceled out" or "absorbed into the intercept."

iv. The between estimate is probably corrupted, however.

v. p. 150. Ecological fallacy plot just like mine

vi. p. 150, "cluster level confounding": "can be described as correlations between the level-1 variable of interest–such as smoking x_{2ij}–and the random intercept ζ_j, which represents the effects of omitted level-2 covariates."

vii. p. 151: compositional effect: clustering of high SES within a school.

viii. **CLAIM**: p. 151 The "so-called contextual effect $\beta_2^B - \beta_2^2$, an additional increase of the second shcool's mean $\bar{y}_{.j}$ after allowing for the within (or compositional) effect β_2^W. The contextual effect could be due to nonrandom assignment of the high SES students to better schools (confounding), as well as direct peer effects."

ix. P. 151. **WARNING**, same as in DeLeuw and Kreft book. Don't make mistake (many have) of including a predictor ($x-$*within group mean*) without including the *within group mean* in the model as well.

x. p. 151. TODO. Find "compositional effect" term/definition in broader literature.

11. p. 152, Allowing different within and between (continues point p. 151)

a) Fit a model including both group mean and x deviation from group mean

$$y_{ij} = \beta_1 + \beta_2^W(x_{2ij} - \bar{x}_{2.j}) + \beta_2^B \bar{x}_{2.j} + \dots \tag{7}$$

b) PJ: explain to class about mean-centering. Help to explain point bottom p.152 that it is NOT necessary to subtract term, it only affects cluster mean estimate. Like intercept interpretation with mean centered regression.

12. p. 154. NEW focus: "However, ζ_j may be correlated with another within-mother covariate x_{3ij} and the inconsitency in estimating the corresponding regression coefficient β_3 can be transmitted to the estimator for β_2^W."

    a) Suggested fix (Mundlak, 1978). Include the within-group means as predictors for ALL of the individual level predictors. Nice use of foreach loop p. 154.

    b) p 156. Here's a happy message and a warning. "A great advantage of clustered or multilevel data is that we can investigate and address level-2-endogeneity of level-1 covariates (correlation between ζ_j and x_{ij}). However, the approaches considered in this chapter do not produce consistent estimates of the coefficients of level-2 covariates and the random-intercept variance in this case. Furthermore, the approaches cannot handle level-2 endogeneity of level-2 covariates (correlation between ζ_j and x_j). Both problems are addressed in an approach suggested by Hausman and Taylor (1981), which we describe in section 5.2."

13. Hausman endogeneity test.

    a) Demin Wu was professor at KU, he is the Wu in "Durbin-Wu-Hausman" test.

    b) Commonly described as a check for endogenous predictors. Are some predictors correlated with the error term in a way that causes spurious correlation in the estimates (radios cause mental illness).

    c) Specification test: Can we say there is a difference between the fixed effects model (estimates $\hat{\beta}^W$) and the FGLS estimates $\hat{\beta}^{FGLS}$. I

    d) "if the random-intercept model is correctly specified"... 2 estimators should be same.

14. Fixed Effects vs Random Effects

    a) Fixed Effects model: dummy variables BLOCK any group-level fixed predictors.

    b) Random Effects model: allows group-level fixed predictors. "An advantage of the random-effects model is that it can be used to estimate the effects of cluster-level covariates, in contrast to the fixed-effects model, although consistent estimation requires both level-1 and level-2 exogeneity." (p. 159).

    c) Please see table 159.

Questions	Answers	
Inference for population clusters	No	Yes
Minimum n of clusters required	Any	> 10 or 20
Assumptions for ζ_j		Level-2 exogeneity, constant variance
Cluster-level covariates allowed	No	Yes
Inference for clusters in particular sample?	Yes	No, but can get EB statements about ζ_j. (PJ: PLS view may help here)
Minimum cluster size required	Need large for estimate of fixed intercepts	
Parsimonious	No, one param per intercept	Yes, one variance param per random effect (PJ: PLS view may alter this)
Within-cluster effects of covariates	Yes	Yes, by including cluster means.

15. p. 160: Residual diagnostics: Spot outliers.

    a) standardized residuals

    b) xtmixed shows how to get standardized residuals at level 2 and 2. See below, "predict xxx, reses". RSES:

16. More on Statistical Inference.

    a) This book takes "old school" approach of characterizing mixed model estimation as an exercise in Generalized Least Squares (GLS)

    b) OLS

$$\hat{\beta}^{OLS} = (X'X)^{-1}X'y \tag{8}$$

    c) GLS: Suppose V is a "variance covariance" matrix of the error terms. Aitken's GLS estimator has been recognized since the 1930s.

$$\hat{\beta}^{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y \tag{9}$$

    d) FGLS (Feasible): IGLS: (Iterative GLS) If V is unknown, we estimate it from data, say the residuals of $\hat{\beta}^{OLS}$. Then fit the GLS model, get the residuals, re-estimate V, re fit. As Bates mentioned in many of his projects, this math can be done in more graceful way by solving $\hat{\beta}^{GLS}$ for V and then iterating on V.

e) p. 165. Theorize to make V as simple as possible. Note, as long as the clusters are not correlated to each other, then V is **block diagonal**:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & 0 & 0 & 0 \\ 0 & \mathbf{V}_2 & 0 & 0 \\ & & \ddots & \\ & & & \mathbf{V}_J \end{bmatrix} \tag{10}$$

f) Then within each block, in a cross-sectional case, the covariances between rows are all of the "exchangeable" variety. We don't have any reason to say row 1 is different than row 3

$$\mathbf{V}_j = \begin{bmatrix} \psi + \theta & \psi & \psi \\ \psi & \psi + \theta & \psi \\ \psi & \psi & \psi + \theta \end{bmatrix} \tag{11}$$

g) Maximum Likelihood Analysis would derive estimates of $\hat{\beta}$ and $\hat{\psi}$ and $\hat{\theta}$.

h) REML: an adjusted method for estimating variance parameters. I can't describe any better than the authors, p. 166. TODO: Find a more clear explanation that I actually understand.

17. p. 167. Bias of the mis-specified pooled OLS estimators.

18. p. 169. Power and sample size determination. This is a pretty big problem, I don't have energy to write the whole lecture here. In practice, the real question is "if you can collect N observations, would you rather divide them among J or P groups, knowing that the sample within each group must shrink if you opt for more groups.

    a) In past, I Did work on MLPowSim and PINT as estimators, literature on this is pretty clear. We'd rather have a handful of observations within lots of groups, rather than lots and lots of observations within a few groups.

    Admittedly, I need to justify this with some actual math.

3 Stata code

1. p. 125. xtsum gives SUPER CONVENIENT group level summaries of the outcome.

```
xtset momid

xtsum birwt smoke black
```

2. p. 126. What is the purpose here:

```
egen pickone = tag(momid)
summarize black if pickone == 1
```

3. p. 127. children per mother

```
egen num = count(birwt), by(momid)
tabulate num if pickone == 1
```

4. p. 127. Between-within-overall summary of a categorical variable

   ```
   xtset momid
   xttab smoke
   ```

   I don't understand what is "Within". Maybe you can tell me what this is good for.

5. p. 131. Estimation. Lets concentrate on xtmixed.

   ```
   xtmixed y x1 x2 x3 x4 x5 || groupvar: , mle
   ```

   gives maximum likelihood estimates.

   a) replace mle with "reml" for REML estimates.

   b) `vce(robust)`. There is is again. vague.

6. p. 139. Stata testparm

   xtreg -> testparm.

   a) Can get robust version with robust standard errors.

7. p. 140. LR test for variance component

8. ```
   xtset
   estimates store full
   xtset momid
   xtreg ...[all same stuff], mle
   lrtest full .
   /* "." is last fitted model. could instead do
   estimates store small
   lrtest full small
   ```

9. Predicted values: Stata margins command.

   a) `margins` only works if model used factor notation

   b) This is like predict with "newdata", the rockchalk package has a lot of stuff like this.

   c) Fit a model with lots of predictors, then set some variables in margins to get predicted values for them.

   d) margins i.x1#i.x2

   e) Note the output uses "Delta method" standard errors.

10. marginsplot, xdimension(education)

11. p. 154. Uses `lincom` to test that 2 fixed params are equal.
   ```

```
lincom mn_smok - dev_smok
```

12. p. 154. Generate a LOT of group-mean-centered variables

    a) Run lots of egen statements like

    ```
    egen mn_male = mean(male), by(momid)
    ```

    b) Use `foreach` loop as follows

    ```
    foreach var of varlist male mage kessner* novisit pretri*
       {
        egen mn_'var' = mean('var'), by(momid)
    }
    ```

    I'm pretty sure the single quotes are not correct, I think I've seen others do a grave on the left, `‘var’`.

13. p. 148 Hausman test. Use xj for predcitors and zzz for grouping var

    ```
    xtset zzz
    xtreg y x1 x2 x3 x4 x5, fe
    estimates store fixed
    xtreg y x1 x2 x3 x4 x5, re
    estimates store random
    hausman fixed random
    ```

    in example, RHS say “strong evidenc e of mis-specification”

14. Standardized residuals.

    a) Start like this:

    ```
    xtmixed y x1 x2 x3 || momid: , mle
    ```

    b) Get the level 2 standardized residuals with some hackery

    ```
    predict lev2, reffects
    predict comp_se, rses
    generate diag_se = sqrt(exp(2*[lns1_1_1]_cons - comp_se^2)
    replace lev2 = lev2/diag_se
    ```

    c) Level 1 standardized residuals come out for free if you ask correctly

    ```
    predict lev1, rstandard
    ```

    d) Then use histograms or such to see them. Inspect the data structure, see some frustrating aspects of the Stata “only one data set at a time” mentality. Second command here is workaround for that.

    ```
    histogram lev1, normal xtitle(Std level 1)
    histogram lev2 if idx==1, normal xtitle(Std level 2)
    ```

Key point here is that the "idx" variable exists and is used to grab one obs from each group.

15. p. 163. Again with "`vce(robust)`". Need to get to the bottom of what that's doing.