Project A - Model Selection and Notation

Chongjun Liao

5/7/2022

0. We will use the classroom.csv data for this project.

a. math1st will be the outcome of interest for this first part
   i. Recall that `math1st = mathkind + mathgain`
b. Read in the data (R: store as `dat`)
c. Fit all models using REML
d. It's best if you use `lmerTest::lmer` rather than `lme4::lmer` to call the MLM function. The former provides p-values for fixed effects in the summary.
e. There are 2 common error messages one can get from lmer calls: failed to converge (problem with hessian: negative eigenvalue; max|grad| = ...); and singularity. They may both be problematic in a real problem, but the latter suggests that a variance component is on the boundary of the parameter space.
   1. In your discussion/writeup, consider the latter to be a "convergence problem" and ignore the former.

```
dat <- read.csv("~/Documents/GitHub/mlm_final_project/data/classroom.csv")
dat <- dat %>%
  mutate(math1st = mathkind + mathgain)
```

1. Estimate an Unconditional Means Model (UMM) with random intercepts for both schools and classrooms (nested in schools).

```
fit1 <- lmer( math1st ~ (1 | schoolid/classid), dat)
summary(fit1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math1st ~ (1 | schoolid/classid)
##    Data: dat
##
## REML criterion at convergence: 11944.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.1872 -0.6174 -0.0204  0.5821  3.8339
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  classid:schoolid (Intercept)   85.46   9.244
##  schoolid         (Intercept)  280.68  16.754
##  Residual                     1146.80  33.864
```

```
## Number of obs: 1190, groups:  classid:schoolid, 312; schoolid, 107
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  522.540      2.037 104.407   256.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a. Report the ICC for schools and the ICC for classrooms **Answer:** The ICC for schools is 0.2447517 and the ICC for classrooms is 0.0745198.

b. **WRITE OUT THIS MODEL** using your preferred notation, but use the same choice of notation for the remainder of your project

   i. Be mindful and explicit about any assumptions made. $MATH1ST_{ijk} = b_0 + \zeta_{0k} + \eta_{0jk} + \varepsilon_{ijk}$, with $\zeta_{0k} \sim N(0, \sigma^2_{\zeta_0})$, $\eta_{0jk} \sim N(0, \sigma^2_{\eta_0})$ and $\varepsilon_{ijk} \sim N(0, \sigma^2_{\varepsilon})$, independently of one another, j represents classrooms and k represents *schools*.

2. ADD ALL School level predictors

```
fit1 <- lmer( math1st ~ (1 | schoolid/classid), dat)
fit2 <- lmer( math1st ~ housepov + (1 | schoolid/classid), dat)
summary(fit2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math1st ~ housepov + (1 | schoolid/classid)
##    Data: dat
##
## REML criterion at convergence: 11927.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.1142 -0.6011 -0.0350  0.5600  3.8154
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  classid:schoolid (Intercept)   82.36   9.075
##  schoolid         (Intercept)  250.93  15.841
##  Residual                     1146.95  33.867
## Number of obs: 1190, groups:  classid:schoolid, 312; schoolid, 107
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  531.294      3.341 102.809 159.024   <2e-16 ***
## housepov     -45.783     14.236 111.063  -3.216   0.0017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr)
## housepov -0.810
```

```
anova(fit1,fit2)
```

```
## refitting model(s) with ML (instead of REML)

## Data: dat
## Models:
## fit1: math1st ~ (1 | schoolid/classid)
## fit2: math1st ~ housepov + (1 | schoolid/classid)
##      npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fit1    4 11956 11976 -5973.9    11948
## fit2    5 11948 11973 -5968.8    11938 10.125  1   0.001463 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a. Report if adding the predictors as a block is justified There is only one school-level predictor which is `housepov`, its p-value is $0.0017029 < 0.05$, and I do a LRT on model with and without the school-level predictor, the p-value is $0.0014627 < 0.05$. So it is reasonable to add school-level predictor.
b. Report change in σ^2_ζ. The change in σ^2_ζ is 280.6812733-250.9258585 = 29.7554148.

3. ADD ALL Classroom level predictors

a. Report if adding the predictors as a block is justified [must use WALD test, not LRT]
b. Report change in σ^2_η and change in σ^2_ϵ.
c. Give a potential reason as to why σ^2_ϵ is reduced, but not σ^2_η?

4. ADD (nearly) ALL student level predictors (but not `mathgain` or `mathkind`, as these are outcomes in this context).

a. Report if justified statistically as a block of predictors [must use WALD test, not LRT]
b. Report change in variance components for all levels
c. Give a potential reason as to why the school level variance component drops from prior model
d. WRITE OUT THIS MODEL using your chosen notation (include assumptions).

5.a. Try to add a random slope for each teacher level predictor (varying at the school level; one by one separately- not all together) b. Report the model fit or lack of fit c. Retry the above, allowing the slopes to be correlated with the random intercepts (still one by one) d. Report anything unusual about the variance components (changes that are in a direction you didn't expect) and any potential explanation for why those changes occurred (hint: what did you add to the model?).

6. Question:

a. Why is it a bad idea to include a classroom-level variable with random slopes at the classroom level?