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

```
save.options = options()
options(na.action = "na.pass")
mm <- model.matrix(~math1st + ses + mathknow, data = dat)
in_sample <- apply(is.na(mm), 1, sum) == 0 # these rows aren't missing anything
options(save.options)
# remove those na
fit3 <- lmer( math1st ~ yearstea + mathknow + mathprep + housepov + (1 | schoolid/classid),
              dat, subset = in_sample)
summary(fit3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula:
## math1st ~ yearstea + mathknow + mathprep + housepov + (1 | schoolid/classid)
##    Data: dat
##  Subset: in_sample
##
## REML criterion at convergence: 10821
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5552 -0.6118 -0.0311  0.5863  3.8315
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  classid:schoolid (Intercept)   94.36   9.714
##  schoolid         (Intercept)  223.31  14.943
##  Residual                      1136.43  33.711
## Number of obs: 1081, groups:  classid:schoolid, 285; schoolid, 105
##
## Fixed effects:
```

```
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 532.29852    5.20495 228.85767 102.268  < 2e-16 ***
## yearstea      0.06193    0.14717 223.76570   0.421  0.67432
## mathknow      2.55143    1.44530 231.06560   1.765  0.07883 .
## mathprep     -0.75440    1.42809 203.20755  -0.528  0.59790
## housepov    -41.62117   14.08834 109.83230  -2.954  0.00383 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) yearst mthknw mthprp
## yearstea -0.264
## mathknow -0.052  0.030
## mathprep -0.666 -0.175  0.004
## housepov -0.568  0.077  0.082  0.032
```

```
wald.test(b = fixef(fit3), Sigma = summary(fit3)$vcov, Terms = 2:4)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 3.5, df = 3, P(> X2) = 0.32
```

a. Report if adding the predictors as a block is justified [must use WALD test, not LRT] **Answer:** The Wald test generates a p-value = 0.22, which shows that we have no reason to add classroom-level predictors as a block. But it might be reasonable to include `mathknow` since it is significant according to the t-test.
b. Report change in σ^2_η and change in σ^2_ϵ. The change in σ^2_η is 94.3625825-85.4593745 = 8.903208 and change in σ^2_ϵ is 1136.4309806-1146.8001472 = -10.3691666.
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).

```
fit4 <- lmer( math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
                housepov + (1 | schoolid/classid), dat, subset = in_sample)
summary(fit4)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
##     housepov + (1 | schoolid/classid)
##    Data: dat
##  Subset: in_sample
##
## REML criterion at convergence: 10729.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8581 -0.6134 -0.0321  0.5971  3.6598
##
```

```
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  classid:schoolid (Intercept)   93.89   9.689
##  schoolid         (Intercept)  169.45  13.017
##  Residual                     1064.96  32.634
## Number of obs: 1081, groups:  classid:schoolid, 285; schoolid, 105
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  539.63041    5.31209  275.39010 101.585  < 2e-16 ***
## ses           10.05076    1.54485 1066.56211   6.506 1.18e-10 ***
## minority     -16.18676    3.02605  704.47787  -5.349 1.20e-07 ***
## sex           -1.21419    2.09483 1022.42110  -0.580    0.562
## yearstea       0.01129    0.14141  226.80861   0.080    0.936
## mathknow       1.35004    1.39168  234.49768   0.970    0.333
## mathprep      -0.27705    1.37583  205.27111  -0.201    0.841
## housepov     -17.64850   13.21755  113.87814  -1.335    0.184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) ses    minrty sex    yearst mthknw mthprp
## ses      -0.121
## minority -0.320  0.162
## sex      -0.190  0.020 -0.011
## yearstea -0.259 -0.028  0.024  0.016
## mathknow -0.083 -0.007  0.115  0.007  0.029
## mathprep -0.631  0.053  0.001 -0.006 -0.172  0.004
## housepov -0.451  0.082 -0.178 -0.007  0.071  0.058  0.038
```

```
wald.test(b = fixef(fit4), Sigma = summary(fit4)$vcov, Terms = 2:4)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 85.1, df = 3, P(> X2) = 0.0
```

a. Report if justified statistically as a block of predictors [must use WALD test, not LRT] The wald test gives a p-value less than 0.05, which justifies the significance of adding a block of individual predictors.
b. Report change in variance components for all levels The change in σ^2_η is 93.8853485-85.4593745 = 8.425974, the change in σ^2_ζ is 169.4480999-280.6812733 = -111.2331734 and change in σ^2_ϵ is 1064.9564422-1146.8001472 = -81.8437049.
c. Give a potential reason as to why the school level variance component drops from prior model Individual predictors are correlated with school-level effect.
d. WRITE OUT THIS MODEL using your chosen notation (include assumptions). $MATH1ST_{ijk} = b_0 + b_1 SES_{ijk} + b_2 MINORITY_{ijk} + b_3 SEX_{ijk} + b_4 YEARSTEA_{jk} + b_5 MATHKNOW_{jk} + b_6 MATHPREP_{jk} + b_7 HOUSEPOV_k + \zeta_{0k} + \eta_{0jk} + \varepsilon_{ijk}$, 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*. 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)
e. Report the model fit or lack of fit

```
fit5.1 <- lmer( math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
                housepov + (1 | schoolid/classid) + (0 + yearstea || schoolid),
                dat, subset = in_sample)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00805459 (tol = 0.002, component 1)
```

```
summary(fit5.1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
##     housepov + (1 | schoolid/classid) + (0 + yearstea || schoolid)
##    Data: dat
##  Subset: in_sample
##
## REML criterion at convergence: 10729.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8482 -0.6147 -0.0322  0.5979  3.6603
##
## Random effects:
##  Groups           Name        Variance  Std.Dev.
##  classid.schoolid (Intercept) 9.247e+01  9.6159
##  schoolid         (Intercept) 1.684e+02 12.9758
##  schoolid.1       yearstea    1.008e-02  0.1004
##  Residual                     1.065e+03 32.6361
## Number of obs: 1081, groups:  classid:schoolid, 285; schoolid, 105
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  539.59885    5.30780  266.47954 101.662  < 2e-16 ***
## ses           10.04528    1.54492 1066.09816   6.502 1.21e-10 ***
## minority     -16.16715    3.02635  702.61831  -5.342 1.24e-07 ***
## sex           -1.21060    2.09480 1022.21558  -0.578    0.563
## yearstea       0.01128    0.14192  122.87743   0.079    0.937
## mathknow       1.33106    1.39155  234.33195   0.957    0.340
## mathprep      -0.26584    1.37588  204.90504  -0.193    0.847
## housepov     -17.72082   13.21686  113.58577  -1.341    0.183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) ses    minrty sex    yearst mthknw mthprp
## ses      -0.121
## minority -0.320  0.162
## sex      -0.191  0.020 -0.010
## yearstea -0.258 -0.027  0.023  0.015
## mathknow -0.082 -0.007  0.115  0.006  0.028
## mathprep -0.632  0.053  0.001 -0.006 -0.172  0.003
## housepov -0.450  0.082 -0.179 -0.007  0.070  0.057  0.037
```

```
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00805459 (tol = 0.002, component 1)
```

```
fit5.2 <- lmer( math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
                housepov + (1 | schoolid/classid) + (0 + yearstea + mathknow || schoolid),
                dat, subset = in_sample)
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(fit5.2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
##     housepov + (1 | schoolid/classid) + (0 + yearstea + mathknow ||
##     schoolid)
##    Data: dat
##  Subset: in_sample
##
## REML criterion at convergence: 10729.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8484 -0.6150 -0.0323  0.5980  3.6601
##
## Random effects:
##  Groups           Name        Variance  Std.Dev.
##  classid.schoolid (Intercept) 9.261e+01  9.6234
##  schoolid         (Intercept) 1.686e+02 12.9828
##  schoolid.1       yearstea    9.821e-03  0.0991
##  schoolid.2       mathknow    0.000e+00  0.0000
##  Residual                     1.065e+03 32.6342
## Number of obs: 1081, groups:  classid:schoolid, 285; schoolid, 105
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  539.59988    5.30889  266.37427 101.641  < 2e-16 ***
## ses           10.04520    1.54490 1066.10175   6.502 1.21e-10 ***
## minority     -16.16787    3.02653  702.69530  -5.342 1.24e-07 ***
## sex           -1.21085    2.09475 1022.23211  -0.578    0.563
## yearstea       0.01124    0.14193  122.94561   0.079    0.937
## mathknow       1.33223    1.39179  234.31811   0.957    0.339
## mathprep      -0.26601    1.37610  204.91987  -0.193    0.847
## housepov     -17.71968   13.22054  113.50872  -1.340    0.183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) ses    minrty sex    yearst mthknw mthprp
## ses      -0.121
## minority -0.320  0.162
## sex      -0.190  0.020 -0.010
## yearstea -0.258 -0.027  0.023  0.015
```

```
## mathknow -0.082 -0.007  0.115  0.006  0.028
## mathprep -0.631  0.053  0.001 -0.006 -0.172  0.003
## housepov -0.450  0.082 -0.179 -0.007  0.070  0.057  0.037
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
fit5.3 <- lmer( math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
                housepov + (1 | schoolid/classid) + (0 + yearstea + mathknow + mathprep || schoolid),
                dat, subset = in_sample)
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(fit5.3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
##     housepov + (1 | schoolid/classid) + (0 + yearstea + mathknow +
##     mathprep || schoolid)
##    Data: dat
##  Subset: in_sample
##
## REML criterion at convergence: 10729.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8485 -0.6149 -0.0323  0.5980  3.6600
##
## Random effects:
##  Groups           Name        Variance  Std.Dev.
##  classid.schoolid (Intercept) 9.270e+01 9.628e+00
##  schoolid         (Intercept) 1.684e+02 1.298e+01
##  schoolid.1       yearstea    9.678e-03 9.838e-02
##  schoolid.2       mathknow    0.000e+00 0.000e+00
##  schoolid.3       mathprep    5.133e-07 7.164e-04
##  Residual                     1.065e+03 3.263e+01
## Number of obs: 1081, groups:  classid:schoolid, 285; schoolid, 105
##
## Fixed effects:
##              Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  539.60082    5.30864  266.37268 101.646  < 2e-16 ***
## ses           10.04524    1.54490 1066.09969   6.502 1.21e-10 ***
## minority     -16.16848    3.02636  702.64771  -5.343 1.24e-07 ***
## sex           -1.21071    2.09476 1022.22241  -0.578    0.563
## yearstea       0.01124    0.14193  122.42627   0.079    0.937
## mathknow       1.33172    1.39180  234.34326   0.957    0.340
## mathprep      -0.26642    1.37615  204.92027  -0.194    0.847
## housepov     -17.71647   13.21784  113.58401  -1.340    0.183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) ses    minrty sex    yearst mthknw mthprp
```

```
## ses      -0.121
## minority -0.320  0.162
## sex      -0.191  0.020 -0.010
## yearstea -0.258 -0.027  0.023  0.015
## mathknow -0.082 -0.007  0.115  0.006  0.028
## mathprep -0.632  0.053  0.001 -0.006 -0.172  0.003
## housepov -0.450  0.082 -0.179 -0.007  0.070  0.057  0.037
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

c. Retry the above, allowing the slopes to be correlated with the random intercepts (still one by one)

```
fit5.c.1 <- lmer( math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
                housepov + (1 | schoolid/classid) + (yearstea || schoolid),
                dat, subset = in_sample)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00967674 (tol = 0.002, component 1)
```

```
fit5.c.2 <- lmer( math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
                housepov + (1 | schoolid/classid) + (yearstea + mathknow|| schoolid),
                dat, subset = in_sample)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues
```

```
## Warning: Model failed to converge with 1 negative eigenvalue: -4.6e-02
```

```
fit5.c.3 <- lmer( math1st ~ ses + minority + sex + yearstea + mathknow + mathprep +
                housepov + (1 | schoolid/classid) + (yearstea + mathknow + mathprep || schoolid),
                dat, subset = in_sample)
```

```
## boundary (singular) fit: see ?isSingular
```

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?