A complication in doing all this is that the package nlme (lme) is supplanted by the new and improved lme4 (lmer); both are widely used so I try to do both tracks in separate Rogosa R-sessions

Stat 209 Lab: Linear Mixed Models in R
This lab covers the Linear Mixed Models tutorial by John Fox.
Lab prepared by Karen Kapur.

1 Introduction

1. The normal linear model is given by

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i,$$

$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$

   Equivalently, we have $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where $\epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I_n})$.

2. Mixed effects models include random effects terms which arise from grouped data. For example, for data on individuals over time, each individual represents a group and a model may include a random effect for each individual.
   $y_{ij} = \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_p x_{pij} + b_{i1} z_{1ij} + b_{i2} z_{2ij} + \ldots + b_{iq} z_{qij} + \epsilon_{ij}$, with $b_{ik} \sim \text{Normal}(0, \psi^2_{\text{k}}), \text{cov}(\text{b}_{\text{ik}}, \text{b}_{\text{ik}'}) = \psi^2_{\text{kk}'}, \epsilon_{\text{ij}} \sim \text{Normal}(0, \lambda^2_{\text{ijj}}), \text{cov}(\epsilon_{\text{ij}}, \epsilon_{\text{ij}'}) = \sigma^2 \lambda_{\text{ijj}'}$.

   (a) The value i indexes the group, j the observation, $j = 1, \ldots, n_i$.

   (b) The coefficients $b_{i1}, \ldots, b_{iq}$ are assumed distributed as multivariate normal, $\text{Normal}_{\text{q}}(\mathbf{0}, \mathbf{\Psi})$.

   (c) The covariance of errors among observations in group i is given by $\sigma^2 \lambda_{ijj'}$. If observations are sampled independently within groups, we have $\lambda_{ijj'} = 0$ for $j \neq j'$. On the other hand, if observations are longitudinal observations on an individual over time, certain assumptions may be made about $\lambda_{ijj'}$.

2 Getting Started in R

For the first part of the lab we will be using the MathAchieve and MathAchSchool datasets from the nlme library. I assume the user is using the RGui.

1. Install the nlme package (if it is not installed already) by selecting Packages− >Install Package(s). Select "nlme" and click ok. The packages should automatically install.

current installations of R will have nlme already present, do >library() to confirm

OR... as is done in the lme4 (lmer) version of Lab2 get the data from the MEMSS package and install and load lme4. Don't load both lme4 and the older nlme in the same session

2. To check that the library was installed correctly, simply type

```
> library(nlme)
```

3 Application

1. Read in the data. MathAchieve consists of School Id, Minority Status, Sex, SES (Socio-Economic Status), MathAch (Score on math achievement test), and MeanSES.

```
> library(nlme)
> data(MathAchieve)
> MathAchieve[1:10,]
Grouped Data: MathAch ~ SES | School
   School Minority    Sex    SES MathAch MEANSES
1    1224       No Female -1.528   5.876  -0.428
2    1224       No Female -0.588  19.708  -0.428
3    1224       No   Male -0.528  20.349  -0.428
4    1224       No   Male -0.668   8.781  -0.428
5    1224       No   Male -0.158  17.898  -0.428
6    1224       No   Male  0.022   4.583  -0.428
7    1224       No Female -0.618  -2.832  -0.428
8    1224       No   Male -0.998   0.523  -0.428
9    1224       No Female -0.888   1.527  -0.428
10   1224       No   Male -0.458  21.521  -0.428
```

2. Read in the school data. The school data consists of School Id, Size, Sector (Catholic or Public), PRACAD, DILCLIM, HIMINTY, MEANSES

```
> data(MathAchSchool)
> MathAchSchool[1:10,]
     School Size   Sector PRACAD DISCLIM HIMINTY MEANSES
1224   1224  842   Public   0.35   1.597       0  -0.428
1288   1288 1855   Public   0.27   0.174       0   0.128
1296   1296 1719   Public   0.32  -0.137       1  -0.420
1308   1308  716 Catholic   0.96  -0.622       0   0.534
1317   1317  455 Catholic   0.95  -1.694       1   0.351
1358   1358 1430   Public   0.25   1.535       0  -0.014
1374   1374 2400   Public   0.50   2.016       0  -0.007
```

note: the backarrow <- is equivalent to =
it is a holdover from S and still used by many
who learned R that way. We use =
as in the Rogosa session

```
1433    1433   899 Catholic    0.96   -0.321       0    0.718
1436    1436   185 Catholic    1.00   -1.141       0    0.569
1461    1461  1672   Public    0.78    2.096       0    0.683
```

3. Turns out that the mean SES was calculated incorrectly for the student data. Therefore, we will re-calculate it.

```
> attach( MathAchieve )
> mses <- tapply( SES, School, mean)
> detach( MathAchieve )
```

4. Create a data frame with the variables of interest.

```
> Bryk <- as.data.frame( MathAchieve[, c("School", "SES", "MathAch" ) ] )
> names(Bryk) <- c("school", "ses", "mathach")
```

5. Add additional variables. Use the school name to make sure variables are assigned the appropriate position.

```
> Bryk$meanses <- mses[as.character(Bryk$school)]
> Bryk$cses <- Bryk$ses - Bryk$meanses
>
> sector <- MathAchSchool$Sector
> names(sector) = row.names( MathAchSchool )
> Bryk$sector <- sector[ as.character(Bryk$school) ]
```

4 Examining the Data

We ask whether math achievement depends on socio-economic status, whether it varies by sector, and whether it varies randomly across schools in the same sector. Note that in the tutorial, it is explored whether a linear fit is appropriate. That part of the analysis has been skipped here.

1. We regress math achievement scores against socio-economic status for each school. We create separate lmList objects for Catholic and public schools. Here we are not fitting any random effects, showing that getting the ordinary least squares coefficients can be obtained for different groups by using the lmList method.

```
> remove( sector )
> attach( Bryk)
> cat.list <- lmList( mathach ~ ses|school,
                                subset = sector == 'Catholic', data = Bryk )
> pub.list <- lmList( mathach ~ ses|school,
                                subset = sector == 'Public', data = Bryk )
```

2. The function intervals() in R plots 95% confidence intervals by default.

```
> plot(intervals( cat.list ), main = 'Catholic')
> plot(intervals( pub.list ), main = 'Public')
```

   Since SES is centered at zero, the intercept parameter value represents the math achievement for an average SES. We can see from these plots that there is substantial school-to-school variation.

3. We make boxplots of the coefficients to compare Catholic and public schools.

```
> cat.coef <- coef( cat.list )
> pub.coef <- coef( pub.list )
> par( mfrow = c(1,2) )
> boxplot( cat.coef[,1], pub.coef[,1],
      main = 'Intercepts', names = c('Catholic', 'Public') )
> boxplot( cat.coef[,2], pub.coef[,2],
      main = 'Slopes', names = c('Catholic', 'Public') )
```

5 Fitting a Hierarchical Linear Model

We group within schools. Within a school, math achievement is regressed on cses.

$$\mathrm{mathach_{ij}} = \alpha_{\mathrm{0i}} + \alpha_{\mathrm{1i}}\mathrm{cses_{ij}} + \epsilon_{\mathrm{ij}}$$

for individual j in school i. We consider that school intercepts and slopes depend on sector and the average level of SES.

$$\alpha_{0i} = \gamma_{00} + \gamma_{01}\mathrm{meanses_i} + \gamma_{02}\mathrm{sector_i} + \mathrm{u_{oi}}$$

$$\alpha 1i = \gamma_{10} + \gamma_{11}\mathrm{meanses_i} + \gamma_{12}\mathrm{sector_i} + \mathrm{u_{1i}}$$

Hence the model is given by

$$\mathrm{mathach_{ij}} = \gamma_{00}+\gamma_{01}\mathrm{meanses_i}+\gamma_{02}\mathrm{sector_i}+\gamma_{10}\mathrm{cses_{ij}}+\gamma_{11}\mathrm{meanses_i cses_{ij}}+\gamma_{12}\mathrm{sector_i cses_{ij}}+\mathrm{u_{0i}}+\mathrm{u_{1i} cses}$$

In terms of the notation given in the introduction, we have

$$\mathrm{mathach_{ij}} = \beta_1+\beta_2\mathrm{meanses_i}+\beta_3\mathrm{sector_i}+\beta_4\mathrm{cses_{ij}}+\beta_5\mathrm{meanses_i cses_{ij}}+\beta_6\mathrm{sector_i cses_{ij}}+\mathrm{b_{i1}}+\mathrm{b_{i2} cses_{ij}}+\epsilon_{ij}$$

We place no restriction on the covariances of the random coeficients, but assume that individual errors are independent within schools, with constant variance.

$$V(\epsilon_i) = \sigma^2 \mathbf{I}_{n_i}$$

1. Re-order the levels of the factor sector to have value 0 for public and 1 for Catholic.

```
> Bryk$sector <- factor( Bryk$sector, levels = c('Public', 'Catholic') )
```

OR using lme4 fit using lmer

2. Now fit the linear mixed model.

```
> bryk.lme.1 <- lme( mathach ~ meanses*cses + sector*cses,
     random = ~ cses | school, data = Bryk )
> summary(bryk.lme.1)
```

3. Discussion of interpretation of coefficients.

   - The fixed-effect coefficient estimate of 12.128 represents the average level of math achievement in public schools since public schools are the baseline for comparison with the factor sector.
   - The coefficient for the sectorCatholic variable represents the additional average level of math achievement in catholic schools. Hence average levels of math achivement is higher in catholic schools than in public schools.
   - The coefficient for cses represents the estimated average slope for SES in public schools. The coefficient labeled cses:Catholic represents the additional average slope for SES in Catholic schools. We see that the average slope for SES in public schools is larger than the average slope for SES in Catholic schools.
   - The coefficient for meanSES represents a school's average math achievement to their average level of SES.

This is where we stopped in the class presentation; these additional exercises are useful R-practice

- The coefficient for meanses:cses gives the average change in the within-school SES slope a one-increment in the meanSES.

4. It is interesting to note how the regression for a single school relates to the hierarchical mixed effects model. We pull out data from a single school and so some illustrative examples.

   (a) First, pull out the school data.

   ```
   Bryk.6469 <- subset( x = Bryk, subset = school == 6469 )
   ```

   (b) Now, fit a regression of mathach on cses. Fit a different random effects model from above, ignoring all predictors except for cses (and intercept)

   ```
   lm.6469 <- lm( mathach ~ cses , data = Bryk.6469)
   bryk.lme.cses <- lme( mathach ~ cses, random = ~ cses | school,
     data = Bryk)
   ```

   (c) Now compare the coefficients from lm.6369 and bryk.lme.cses. The lm.6469 coefficient of cses should not deviate too much from the bryk.lme.cses coefficient of cses (taking into account the estimated sd).

   ```
   summary(lm.6469)
   summary(bryk.lme.cses)
   ```

   (d) Next, we consider the presence of predictors besides cses. We take the residuals from the fixed effects.

   ```
   bryk.lm <- lm( mathach ~ meanses*cses + cses*sector, data = Bryk)
   mathach.nofixed <- bryk.lm$resid
   mathach.nofixed.6469 <- subset( mathach.nofixed, subset = school == 6469
   ```

   (e) If we do a linear regression on cses, then we would expect a school's parameter estimates to be centered about zero with sd given by the standard deviation of the random effects.

   ```
   lm.6469.nofixed <- lm( mathach.nofixed.6469 ~ cses, data = Bryk.6469 )
   summary(lm.6469.nofixed)
   summary(bryk.lme.cses)
   ```

5. We compare the coefficient estimates from the OLS fit to the fixed effects from the lme model. The coefficient estimates of the lme model should approximate the coefficient estimates from the OLS fit.

6. It is often of interest to determine whether there is evidence that the variances of the random effects in the model are different from 0. We can test these hypotheses by deleting random-effects terms from the model and examining the change in log likelihood. We must be careful to compare models identical in their fixed effects.

```
> bryk.lme.2 <- update( bryk.lme.1, random = ~ 1 | school )
   # omitting random effect of cses
> anova( bryk.lme.1, bryk.lme.2)
           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
bryk.lme.1     1 10 46523.66 46592.45 -23251.83
bryk.lme.2     2  8 46520.79 46575.82 -23252.39 1 vs 2 1.124098    0.57

> bryk.lme.3 <- update( bryk.lme.1, random = ~ cses - 1 | school )
   # omitting random intercept
> anova(bryk.lme.1, bryk.lme.3)
           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
bryk.lme.1     1 10 46523.66 46592.45 -23251.83
bryk.lme.3     2  8 46740.23 46795.26 -23362.11 1 vs 2 220.5634  <.0001
```

We see that there is strong evidence that the average level of math achievement (represented by the intercept) varies from school to school, but that the coefficient of SES does not vary significantly, once differences between Catholic and public schools are taken into account and the average level of SES in schools is held constant.