Multilevel analysis in R

Predicting school popularity

- Simulated data set
- 100 classes, 2000 pupils
- Continuous outcome: popularity rating
- Explanatory variables pupil level
  - sex (0 = boy, 1 = girl),
  - extraversion (1-10)
- Explanatory variables class level:
  - teacher experience (in years, 2-25)

Reading in the data to R

- Read in a .csv fiel
- Check the file
- View descriptive statistics

```
popular <- read.csv(file = "popular.csv")
head(popular)
summary(popular)
```

- We will do this together in a bit

lmer doesn't run this model so has to use lm()

0. Model ignoring multilevel structure

```r
lm(popular ~ 1, data = popular)
```

$$popularity_i = \gamma_{00} + e_i$$

```
Call:
lm(formula = popular ~ 1, data = popular)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0765 -0.9765  0.0235  0.9236  4.4235 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.07645    0.03091   164.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.383 on 1999 degrees of freedom
```

Model:	**M_0: intercept only – no ML**	**M_1: intercept only**
Fixed part	Coefficients (SE)	Coefficients (SE)
Intercept γ_{00}	5.08 (0.03)	
Gender γ_{10}		
Extr γ_{20}		
T exp γ_{01}		
Extr*texp γ_{21}		
Random part		
σ_e^2		
$\sigma_{u_0}^2$		

1. Intercept-only Model

```r
lmer(popular ~ 1 + (1|class), REML = FALSE,
       data = popular)
```

$$popularity_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

- The random part is given in the equation in between brackets: `(1|class)`.
- The 1 means that we are only making the intercept random.
- The intercept is conditional (i.e., given) on which class a student belongs to.

Reminder on full vs restricted maximum likelihood

- Full maximum likelihood estimation:

  Both the regression coefficients and variance components are included in the likelihood function

- Restricted maximum likelihood estimation: (REML)

  Only the variance components are included in the likelihood functions, regression coeffects estimated separately

Full vs restricted maximum likelihood

- Full maximum likelihood estimation:

  Tends to underestimate the variance components (in small samples)

- BUT:

  The overall chi square test based on the likelihood of two models to compare the models can only be used with FML

  (for the fixed part of the model)

And, bias tends to be small

1. Intercept-only Model with FML

```
lmer(popular ~ 1 + (1|class), REML = FALSE,
        data = popular)
```

$$popularity_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

- REML is default in lme4
- Simple to change
- `REML = FALSE`

1. Intercept-only Model: Output

```
lmer(popular ~ 1 + (1|class), REML = FALSE,
      data = popular)
```

$$popularity_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: popular ~ 1 + (1 | class)
   Data: popular

     AIC      BIC   logLik deviance df.resid
  6333.5   6350.3  -3163.7   6327.5     1997

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5662 -0.6983  0.0021  0.6758  3.3173

Random effects:
 Groups   Name        Variance Std.Dev.
 class    (Intercept) 0.6945   0.8333
 Residual             1.2218   1.1053
Number of obs: 2000, groups:  class, 100

Fixed effects:
            Estimate Std. Error t value
(Intercept)  5.07786    0.08696    58.4
```

Model:	**M_0: intercept only – no ML**		**M_1: intercept only**	
Fixed part	Coefficients (SE)		Coefficients (SE)	
Intercept γ_{00}	5.08	(0.03)	5.08	(0.09)
Gender γ_{10}				
Extr γ_{20}				
T exp γ_{01}				
Extr*texp γ_{21}				
Random part				
σ_e^2			1.22	
$\sigma_{u_0}^2$			0.69	

1. Intercept-only Model: ICC

```
lmer(popular ~ 1 + (1|class), REML = FALSE,
       data = popular)
```

$$popularity_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

Should we perform multilevel analysis?

$$\rho = \frac{\sigma_{u0}^2}{\sigma_e^2 + \sigma_{u0}^2} = \frac{0.69}{1.22 + 0.69} = 0.36$$

About one third of the variance is at the class level (this is unusually high)

$>$ 0.1 is kinda the cut-off.

Model:	M_0: intercept only – no ML	M_1: intercept only
Fixed part	Coefficients (SE)	Coefficients (SE)
Intercept γ_{00}	5.08 (0.03)	5.08 (0.09)
Gender γ_{10}		
Extr γ_{20}		
T exp γ_{01}		
Extr*texp γ_{21}		
Random part		
σ_e^2		1.22
$\sigma_{u_0}^2$		0.69

2. Fixed Model: 1st level predictors

```r
lmer(popular ~ gender + extrav + (1|class),
        REML = FALSE, data = popular)
```

$$popularity_{ij} = \gamma_{00} + \gamma_{10} sex_{ij} + \gamma_{20} extr_{ij} + u_{0j} + e_{ij}$$

Model:	**M_1: intercept only**		**M_2: level 1 predictors**	
Fixed part	Coefficients (SE)		Coefficients (SE)	
Intercept γ_{00}	5.08	(0.09)	2.14	(0.12)
Gender γ_{10}			1.25	(0.04)
Extr γ_{20}			0.44	(0.02)
T exp γ_{01}				
Extr*texp γ_{21}				
Random part				
σ_e^2	1.22		0.59	
$\sigma_{u_0}^2$	0.69		0.62	
Explained var				
Level 1			0.52	
Level 2			0.11	

2. Fixed Model: 1st level predictors: Output

```r
lmer(popular ~ gender + extrav + (1|class),
        REML = FALSE, data = popular)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: popular ~ 1 + gender + extrav + (1 | class)
   Data: popular

     AIC      BIC   logLik deviance df.resid
    4944     4972    -2467     4934     1995

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2113 -0.6578 -0.0048  0.6739  2.9771

Random effects:
 Groups   Name        Variance Std.Dev.
 class    (Intercept) 0.6204   0.7876
 Residual             0.5915   0.7691
Number of obs: 2000, groups:  class, 100

Fixed effects:
            Estimate Std. Error t value
(Intercept)  2.14138    0.11696   18.31
gender       1.25314    0.03741   33.50
extrav       0.44151    0.01615   27.34

Correlation of Fixed Effects:
       (Intr) gender
gender -0.100
extrav -0.706 -0.085
```

Model:	**M_1: intercept only**		**M_2: level 1 predictors**	
Fixed part	Coefficients (SE)		Coefficients (SE)	
Intercept γ_{00}	5.08	(0.09)	2.14	(0.12)
Gender γ_{10}			1.25	(0.04)
Extr γ_{20}			0.44	(0.02)
T exp γ_{01}				
Extr*texp γ_{21}				
Random part				
σ_e^2	1.22		0.59	
$\sigma_{u_0}^2$	0.69		0.62	
Explained var				
Level 1			0.52	
Level 2			0.11	

3. Fixed Model: 1st and 2nd level predictors

```
lmer(popular ~ gender + extrav + texp +
         (1|class),
         REML = FALSE, data = popular)
```

$$popularity_{ij} = \gamma_{00} + \gamma_{10} sex_{ij} + \gamma_{20} extr_{ij} + \gamma_{01} texp_j + u_{0j} + e_{ij}$$

Model:	**M_2: level 1 predictors**		**M_3: level 1 and 2 predictors**	
Fixed part	Coefficients (SE)		Coefficients (SE)	
Intercept γ_{00}	2.14	(0.12)	0.81	(0.17)
Gender γ_{10}	1.25	(0.04)	1.25	(0.04)
Extr γ_{20}	0.44	(0.02)	0.45	(0.02)
T exp γ_{01}			0.09	(0.01)
Extr*texp γ_{21}				
Random part				
σ_e^2	0.59		0.59	
$\sigma_{u_0}^2$	0.62		0.29	
Explained var				
Level 1	0.52		0.52	
Level 2	0.11		0.58	

3. Fixed Model: 1st and 2nd level predictors

```
lmer(popular ~ gender + extrav + texp +
         (1|class),
         REML = FALSE, data = popular)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: popular ~ 1 + gender + extrav + texp + (1 | class)
   Data: popular

     AIC      BIC   logLik deviance df.resid
  4874.3   4907.9  -2431.1   4862.3     1994

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1794 -0.6492 -0.0067  0.6708  3.0103

Random effects:
 Groups   Name        Variance Std.Dev.
 class    (Intercept) 0.2888   0.5374
 Residual             0.5914   0.7690
Number of obs: 2000, groups:  class, 100

Fixed effects:
            Estimate Std. Error t value
(Intercept) 0.809326   0.168828   4.794
gender      1.254095   0.037265  33.653
extrav      0.454484   0.016154  28.134
texp        0.088409   0.008676  10.190

Correlation of Fixed Effects:
       (Intr) gender extrav
gender -0.040
extrav -0.592 -0.090
texp   -0.801 -0.037  0.141
```

Model:	**M_2: level 1 predictors**		**M_3: level 1 and 2 predictors**	
Fixed part	Coefficients (SE)		Coefficients (SE)	
Intercept γ_{00}	2.14	(0.12)	0.81	(0.17)
Gender γ_{10}	1.25	(0.04)	1.25	(0.04)
Extr γ_{20}	0.44	(0.02)	0.45	(0.02)
T exp γ_{01}			0.09	(0.01)
Extr*texp γ_{21}				
Random part				
σ_e^2	0.59		0.59	
$\sigma_{u_0}^2$	0.62		0.29	
Explained var				
Level 1	0.52		0.52	
Level 2	0.11		0.58	

4. Random Coefficient Model

```r
lmer(popular ~ 1 + extrav + gender + texp + (extrav|class),
        REML = FALSE, data = popular)
```

$popularity_{ij} =$
$\gamma_{00} + \gamma_{10} sex_{ij} + \gamma_{20} extr_{ij} + \gamma_{01} texp_j + u_{0j} +$
$u_{1j} sex_{ij} + u_{2j} extr_{ij} + e_{ij}$

We will run this model on Monday.

Take notice on how we add a random slope for extrav

Significant slope variation for *extraversion*, not for *sex*

Model:	**M$_4$: with random slope**	
Fixed part	Coefficients (SE)	
Intercept γ_{00}	0.76	(0.20)
Gender γ_{10}	1.25	(0.04)
Extr γ_{20}	0.45	(0.03)
T exp γ_{01}	0.09	(0.01)
Extr*texp γ_{21}		
Random part		
σ_e^2	0.55	
$\sigma_{u_0}^2$	1.32	
$\sigma_{u_2}^2$	0.03	
$\sigma_{u_{02}}^2$	-0.19	

5. Full Multilevel Regression Model

```
lmer(popular ~ 1 + extrav + gender + texp +
       extrav*texp + (extrav|class),
       REML = FALSE, data = popular)
```

$popularity_{ij} =$
$\gamma_{00} + \gamma_{10} sex_{ij} + \gamma_{20} extr_{ij} + \gamma_{01} texp_j +$
$\gamma_{21} extr_{ij} * texp_j + u_{0j} + u_{2j} extr_{ij} + e_{ij}$

Notice how the cross-level interaction was added

Model:	**M_4: with random slope**		**M_5: cross-level interaction**	
Fixed part	Coefficients (SE)		Coefficients (SE)	
Intercept γ_{00}	0.74	(0.20)	-1.21	(0.27)
Gender γ_{10}	1.25	(0.03)	1.24	(0.04)
Extr γ_{20}	0.45	(0.02)	0.80	(0.04)
T exp γ_{01}	0.09	(0.01)	0.23	(0.02)
Extr*texp γ_{21}			-0.03	(0.00)
Random part				
σ_e^2	0.55		0.55	
$\sigma_{u_0}^2$	1.30		0.45	
$\sigma_{u_2}^2$	0.03		0.00	
$\sigma_{u_{02}}^2$	-0.19		-0.03	

Let’s try it!

- Download from BB
  - Starter R script (or start your own)
  - Popular.csv data set
- Read the data into R
- Request descriptive statistics
- Run Model 0 and 1 (random intercept)
  - Do your results match the slides? If yes, continue.
  - Calculate the ICC
- Try running the next models (as time permits)
  - We will continue on Monday

A quick note on reporting of results

- Follow APA guidelines. Example:

  $b_{gender} = 1.25$, $t(1996) = 34.45$; $p = .012$

Never ever say p = 0, there is no p=0.
p < 0.01
says Beth

- If asked for equations, write your own and don't copy-paste a screen shot from the slides

Monday

- We will continue with this example in groups of 3
- Including:
  - Assumption practice
  - More time on model comparisons
  - Calculating explained variance
  - Completing excel file

Answer / excel sheet

Name data	Model 0: no random intercept		Model 1: random intercept		Model 2: Fixed level 1		Model 3: Fixed level 1,2		Model 4: Random slope and covariance		Model 5: Rand. Sl. + cov. + Predictor of slope	
	Par. Est.	*SE*	*Par. Est.*	*SE*	*Par. Est.*	*SE*	*Par. Est.*	*SE*	*Par. Est.*	*SE*	*Par. Est.*	*SE*
FIXED												
mean/intercept												
var1												
var2												
...												
cross level int.												
RANDOM												
VAR(e(ij))												
VAR(u(0j)) ***												
VAR(u(1j)) ***												
Covar(u(0j),u(1j))												
VAR(u(2j)) ***												
Covar(u(0j),u(2j))												
	Fit	*par*	*Fit*	*par*	*Fit*	*par*	*Fit*	*par*	*Fit*	*par*	*Fit*	*par*
Likelihood												
Deviance												
Diff Dev *,***												
AIC **												
Explained variance												
R2 level 1												
R2 level 2												
R2 cross level interaction												

* cut off value for test with 1 df is 3,84;

** lowest AIC is best model ;

*** in test of random parameters (both Wald as well as difference of deviances), p has to be divided by 2

Interpret model parameter estimates of model with lowest AIC

Adjust number of rows and columns as needed