Multilevel Modelling: Introduction

Violetta Korsunova
Dep. of Sociology, HSE

Why use multilevel approach?

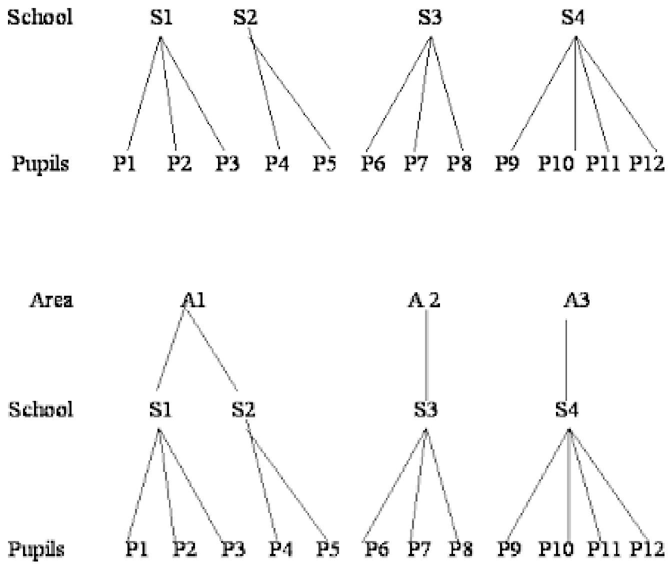
The basic assumption behind statistical inference is the independence of observations

Your units of analysis (respondents or whatever you study) must be unrelated.

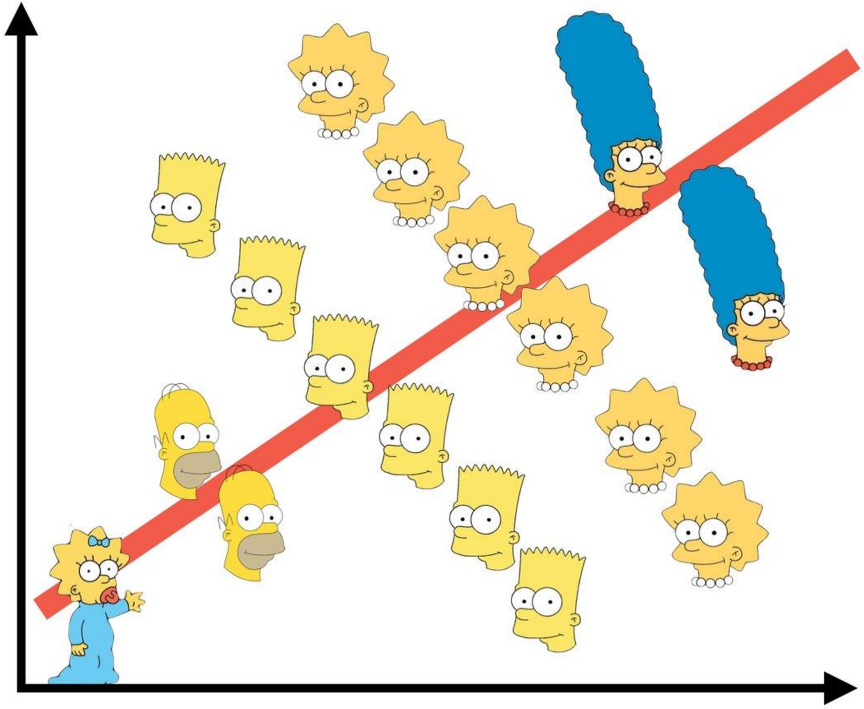
However, sometimes they are nested into particular groups, e.g. countries, schools, teams etc

In this case you can’t assume the observations are independent, and you should take into account the hierarchical structure of your data

Hierarchical structure



Simpson’s paradox



What does multilevel analysis do?

Generally, you assume that a part of your dependent variable's variance is determined by the differences in contextual factors aka 2-level factors

It means that the intercepts and/or betas in your model vary across groups

Let's look at the illustration: http://mfviz.com/hierarchical-models/

Random intercept model

The common equation for a linear model goes as follows:

$$y_{ij} = b_0 + b_1 * x_{ij} + e_{ij}$$

Random intercept model implies a unique b0 in each group:

$$b_{0j} = \gamma_{00} + \eta_{0j}$$

where γ_{00} is the mean value of intercepts aka grand mean, and η_{0j} is the error term that discerns the intercept in group j from the grand mean

The final equation is:

$$y_{ij} = \gamma_{00} + \eta_{0j} + b_1 * x_{ij} + e_{ij}$$

Random slope model

Again, the common equation for a linear model goes as follows:

$$y_{ij} = b_0 + b_1 * x_{ij} + e_{ij}$$

Random slope model implies a unique b1 in each group:

$$b_{1j} = \gamma_{10} + \eta_{1j}$$

where γ_{10} is the mean value of slopes aka grand slope, and η_{1j} is the error term that discerns the slope in group j from the grand slope

The final equation is:

$$y_{ij} = b_0 + (\gamma_{10} + \eta_{1j}) * x_{ij} + e_{ij}$$

Random slope and intercept model

Finally, both intercepts and slopes can vary:

$$b_{0j} = \gamma_{00} + \eta_{0j}$$
$$b_{1j} = \gamma_{10} + \eta_{1j}$$

The full model is:

$$y_{ij} = \gamma_{00} + \eta_{0j} + (\gamma_{10} + \eta_{1j}) * x_{ij} + e_{ij}$$

or

$$y_{ij} = \gamma_{00} + \eta_{0j} + \gamma_{10} * x_{ij} + \eta_{1j} * x_{ij} + e_{ij}$$

Inserting 2-level variable: random intercept

These models imply that a part of the group-level error terms (η) is explained by some factors (Z)

$$b_{0j} = \gamma_{00} + \eta_{0j}$$
$$\eta_{0j} = \gamma_{01} * Z_j + e_{0j}$$

So the full equation:

$$y_{ij} = \gamma_{00} + \gamma_{01} * Z_j + e_{0j} + b_1 * x_{ij} + e_{ij}$$

Inserting 2-level variable: random slope

$$b_{1j} = \gamma_{10} + \eta_{1j}$$
$$\eta_{1\mathrm{j}} = \gamma_{11} * Z_j + e_{1j}$$

So the full equation:

$$y_{ij} = \gamma_{00} + \eta_{0j} + \gamma_{10} * x_{ij} + \gamma_{11} * Z * x_{ij} + e_{1j} * x_{ij} + e_{ij}$$

The $\gamma_{11} * Z_j * xi_j$ term is called ***cross-level interaction***

Inserting 2-level variable: random slope and intercept

As we combine both parts we can get the final model:

$$b_{0j} = \gamma_{00} + \eta_{0j}$$
$$\eta_{0j} = \gamma_{01} * Z_j + e_{0j}$$
$$b_{1j} = \gamma_{10} + \eta_{1j}$$
$$\eta_{1j} = \gamma_{11} * Z_j + e_{1j}$$

So the full equation:

$$y_{ij} = \gamma_{00} + \gamma_{01} * Z_j + e_{0j} + \gamma_{10} * x_{ij} + \gamma_{11} * Z_j * x_{ij} + e_{1j} * x_{ij} + e_{ij}$$

Intraclass correlation coefficient (ICC)

Shows how much variance is explained by 2-level factors:

$$ICC = \frac{\sigma^2(group)}{\sigma^2(group) + \sigma^2(individual)}$$

Should be at least 0.05 which means that 5% of the variance is due to group features

You may use multilevel approach with smaller ICC if you expect causal heterogeneity (but you have to prove it)

Number of groups

Do not use multilevel when you have < 20 groups

Be careful when you have < 50 groups:

- 10 groups for 1 second-level variable
- Cross-level interactions can be biased

How to do in R

Use data ‘imm23.Rdata’:

- MATH – math test score
- SEX – gender (‘male’, ‘female’)
- WHITE – race (‘non-white’, ‘white’)
- PUBLIC – type of school (‘private’, ‘public’)
- SES – family’s social-economic status
- HOMEWORK – number of hours per week spent on math
- PARENTED – parents education level
- MEANSES – mean SES in school
- SCHID – school’s ID

Which variables are 2-level?

How to do in R

Use data ‘imm23.Rdata’:

- MATH – math test score
- SEX – gender (‘male’, ‘female’)
- WHITE – race (‘non-white’, ‘white’)
- PUBLIC – type of school (‘private’, ‘public’)
- SES – family’s social-economic status
- HOMEWORK – number of hours per week spent on math
- PARENTED – parents education level
- MEANSES – mean SES in school
- SCHID – school’s ID

Fitting the null model

```
> library(lme4)
> m0 = lmer(MATH ~ 1|SCHID, data = data)
> summary(m0)

Linear mixed model fit by REML ['lmerMod']
Formula: MATH ~ 1 | SCHID
   Data: data
REML criterion at convergence: 3798.7
Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.68160 -0.72864 -0.01926  0.73173  2.67329
Random effects:
 Groups   Name        Variance Std.Dev.
 SCHID    (Intercept) 26.12    5.111
 Residual             81.24    9.014
Number of obs: 519, groups:  SCHID, 23
Fixed effects:
            Estimate Std. Error t value
(Intercept)   50.759      1.151   44.09
```

Fitting the null model

```
> library(lme4)
> m0 = lmer(MATH ~ 1|SCHID, data = data)
> summary(m0)

Linear mixed model fit by REML ['lmerMod']
Formula: MATH ~ 1 | SCHID
   Data: data
REML criterion at convergence: 3798.7
Scaled residuals:
     Min       1Q    M                   Max
-2.68160 -0.72864 -0.              2.67329
Random effects:
 Groups   Name        Variance Std.Dev.
 SCHID    (Intercept) 26.12    5.111
 Residual             81.24    9.014
Number of obs: 519, groups:  SCHID, 23
Fixed effects:
            Estimate Std. Error t value
(Intercept)   50.759      1.151   44.09
```

Group variance

Fitting the null model

```
> library(lme4)
> m0 = lmer(MATH ~ 1|SCHID, data = data)
> summary(m0)

Linear mixed model fit by REML ['lmerMod']
Formula: MATH ~ 1 | SCHID
   Data: data
REML criterion at convergence: 3798.7
Scaled residuals:
     Min       1Q    M                       Max
-2.68160 -0.72864 -0.                   2.67329
Random effects:
 Groups   Name        Var   nce Std.Dev.
 SCHID    (Intercept) 26.12    5.111
 Residual             81.24    9.014
Number of obs: 519, groups:
Fixed effects:
            Estimate Std. E
(Intercept)   50.759      1.
```

Group variance

Individual variance

Fitting the null model

```
> library(lme4)
> m0 = lmer(MATH ~ 1|SCHID, data = data)
> summary(m0)

Linear mixed model fit by REML ['lmerMod']
Formula: MATH ~ 1 | SCHID
   Data: data
REML criterion at convergence: 3798.7
Scaled residuals:
     Min       1Q    Me                  Max
-2.68160 -0.72864 -0.                2.67329
Random effects:
 Groups   Name        Variance Std.Dev.
 SCHID    (Intercept) 26.12    5.111
 Residual             81.24    9.014
Number of obs: 519, groups:
Fixed effects:
            Estimate Std. E
(Intercept)   50.759      1.
```

Group variance

Individual variance

$$\text{ICC} = \frac{26.12}{26.12+81.24} = 0.24$$

Fitting the random intercept model

```
> m1 = lmer(MATH~SES+WHITE+HOMEWORK+PARENTED+SEX+(1|SCHID), data = data)
> screenreg(m1)
=====================================
                          Model 1
-------------------------------------
(Intercept)                 37.76 ***
                            (2.35)
SES                          0.20
                            (0.98)
WHITEwhite                   2.99 **
                            (1.07)
HOMEWORK                     2.17 ***
                            (0.27)
PARENTED                     2.24 ***
                            (0.57)
SEXfemale                   -0.44
                            (0.73)
-------------------------------------
AIC                       3677.71
BIC                       3711.73
Log Likelihood           -1830.86
Num. obs.                  519
Num. groups: SCHID          23
Var: SCHID (Intercept)       9.01
Var: Residual               65.25
=====================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

Fitting the random intercept and slope model

```
> m2 = lmer(MATH~SES+WHITE+HOMEWORK+PARENTED+SEX+(1+HOMEWORK|SCHID), data = data)
> screenreg(m2)
==============================================
                                  Model 1
----------------------------------------------
(Intercept)                         40.07 ***
                                    (2.49)
SES                                  0.34
                                    (0.88)
WHITEwhite                           2.52 *
                                    (0.98)
HOMEWORK                             1.83 *
                                    (0.82)
PARENTED                             1.57 **
                                    (0.52)
SEXfemale                           -0.21
                                    (0.66)
----------------------------------------------
AIC                               3608.30
BIC                               3650.82
Log Likelihood                   -1794.15
Num. obs.                          519
Num. groups: SCHID                  23
Var: SCHID (Intercept)              45.64
Var: SCHID HOMEWORK                 13.64
Cov: SCHID (Intercept) HOMEWORK    -21.81
Var: Residual                       50.71
==============================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

Comparing models

```
> anova(m1, m2)

refitting model(s) with ML (instead of REML)
Data: data
Models:
m1: MATH ~ SES + WHITE + HOMEWORK + PARENTED + SEX + (1 | SCHID)
m2: MATH ~ SES + WHITE + HOMEWORK + PARENTED + SEX + (1 + HOMEWORK| SCHID)
Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
m1  8 3682.5 3716.5 -1833.2   3666.5
m2 10 3614.7 3657.2 -1797.4   3594.7 71.747      2  2.633e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The second model is significantly better

```
> ranef(m2) #Random effects
$SCHID
      (Intercept)   HOMEWORK
6053    1.3929274  0.4982606
6327   10.1677477 -6.1556950
6467   -3.9073555  3.3376482
7194    4.9603475 -3.6008679
7472    4.1458875 -4.5327887
7474    9.5276497 -3.7584765
7801    7.4586558 -4.2529180
7829    5.5010243 -4.3901776
7930   -6.9919231  4.9685279
24371  -5.4619278  2.3408758
24725  -9.2301555  3.1575251
```

The beta-coefficient for HOMEWORK in school 60~~6~~53 is 1.83 (β from model)+0.498 = 2.328

What is the beta in school 7474?

Inserting 2-level predictor

```
> m3 = lmer(MATH~SES+WHITE+HOMEWORK+PARENTED+SEX+MEANSES+(1+HOMEWORK|SCHID), data = data)
> screenreg(m3)
=============================================
                                Model 1
---------------------------------------------
(Intercept)                       40.61 ***
                                  (2.51)
SES                                0.16
                                  (0.88)
WHITEwhite                         2.49 *
                                  (0.98)
HOMEWORK                           1.86 *
                                  (0.82)
PARENTED                           1.49 **
                                  (0.52)
SEXfemale                         -0.26
                                  (0.66)
MEANSES                            2.59
                                  (1.44)
---------------------------------------------
AIC                             3604.68
BIC                             3651.45
Log Likelihood                 -1791.34
Num. obs.                        519
Num. groups: SCHID                23
Var: SCHID (Intercept)            46.45
Var: SCHID HOMEWORK               13.56
Cov: SCHID (Intercept) HOMEWORK  -22.41
Var: Residual                     50.74
=============================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

Inserting cross-level interaction

```
> m4 = lmer(MATH~SES+WHITE+HOMEWORK+PARENTED+SEX+MEANSES+MEANSES*HOMEWORK+(1+HOMEWORK|SCHID), data = data)
> screenreg(m4)
=============================================
                               Model 1
---------------------------------------------
(Intercept)                      40.50 ***
                                 (2.54)
SES                               0.17
                                 (0.88)
WHITEwhite                        2.50 *
                                 (0.98)
HOMEWORK                          1.92 *
                                 (0.85)
PARENTED                          1.49 **
                                 (0.52)
SEXfemale                        -0.26
                                 (0.66)
MEANSES                           1.42
                                 (2.88)
MEANSES:HOMEWORK                  0.70
                                 (1.50)
---------------------------------------------
AIC                            3603.82
BIC                            3654.84
Log Likelihood                -1789.91
Num. obs.                       519
Num. groups: SCHID               23
Var: SCHID (Intercept)           48.24
Var: SCHID HOMEWORK              14.18
Cov: SCHID (Intercept) HOMEWORK -23.46
Var: Residual                    50.73
=============================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

Analysing model fit

```
> library(sjstats)
> r2(m4)

# R2 for Mixed Models

  Conditional R2: 0.604
     Marginal R2: 0.277
```

Analysing model fit

```
> library(sjstats)
> r2(m4)

# R2 for Mixed Mode

  Conditional R2: 0.604
     Marginal R2: 0.277
```

R2 for fixed+random effects

Analysing model fit

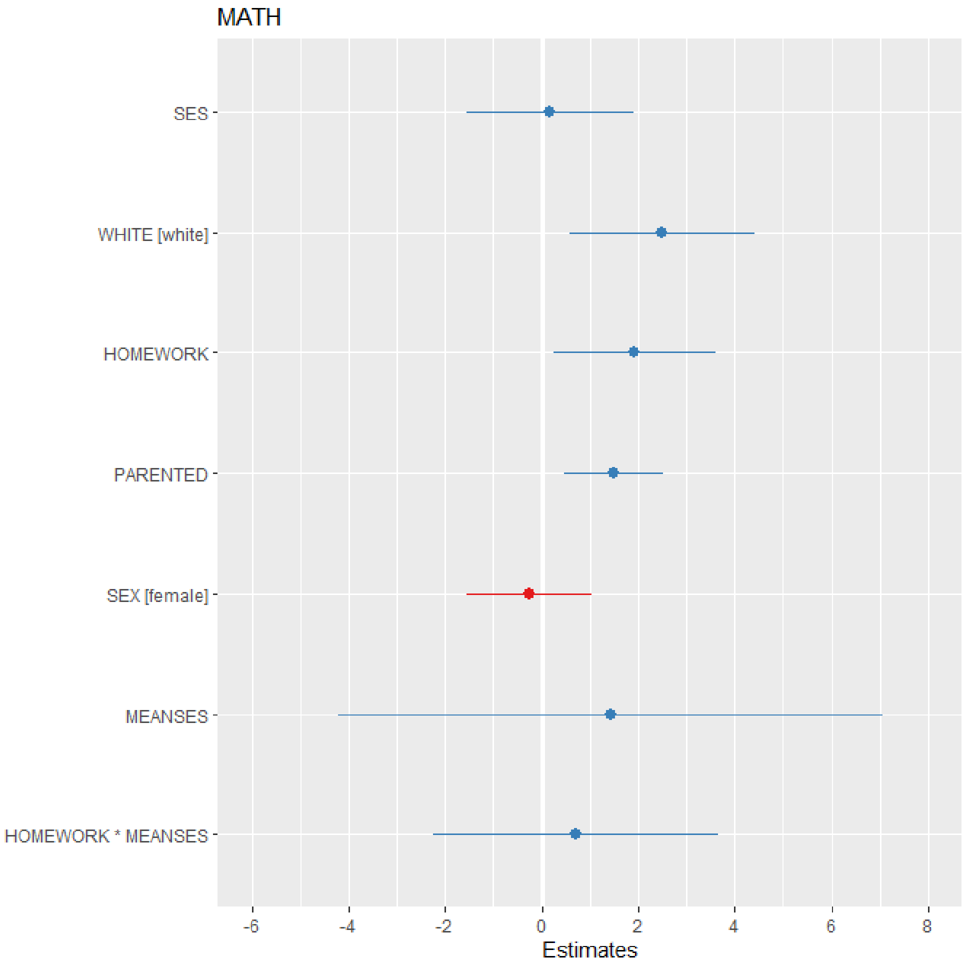
```
> library(sjstats)
> r2(m4)

# R2 for Mixed Mode

  Conditional R2: 0.604
     Marginal R2: 0.277
```

R2 for fixed+random effects

R2 for fixed effects

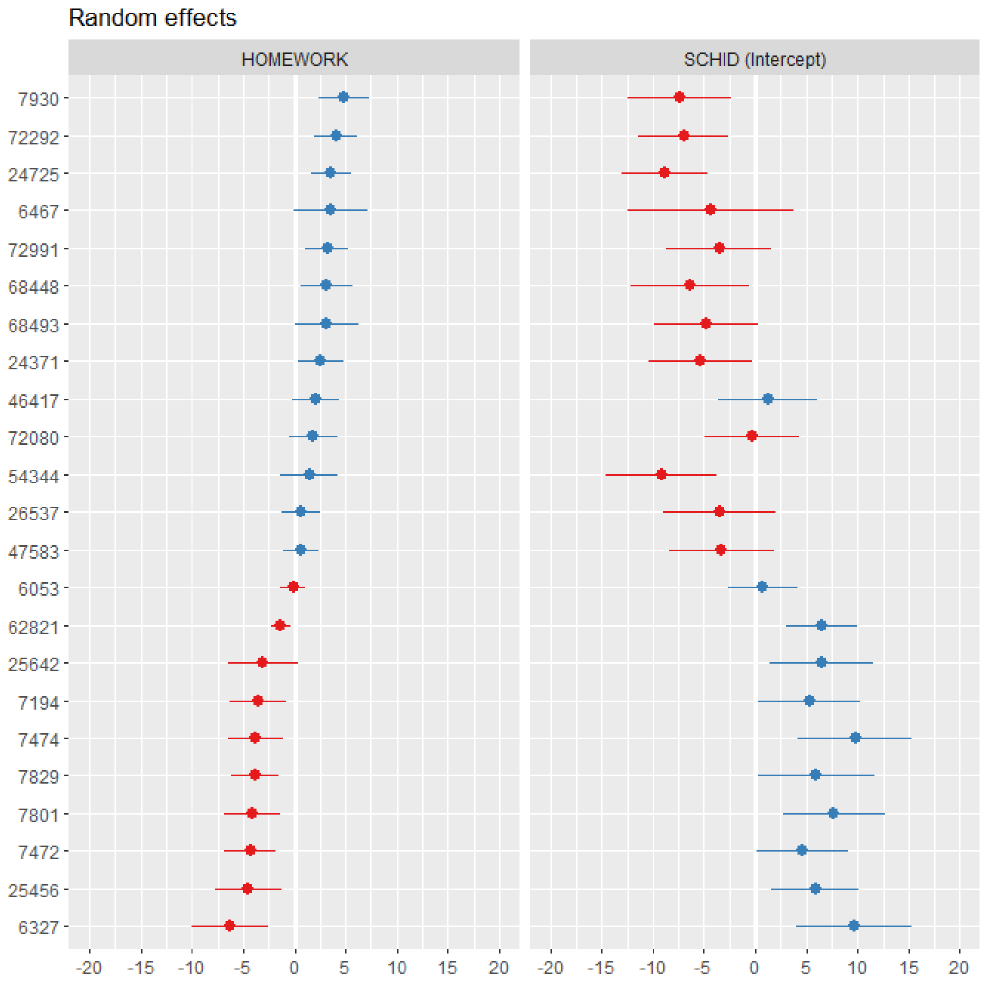
Visualization: FE

```
> library(sjPlot)
> plot_model(m4, type = 'est')
```

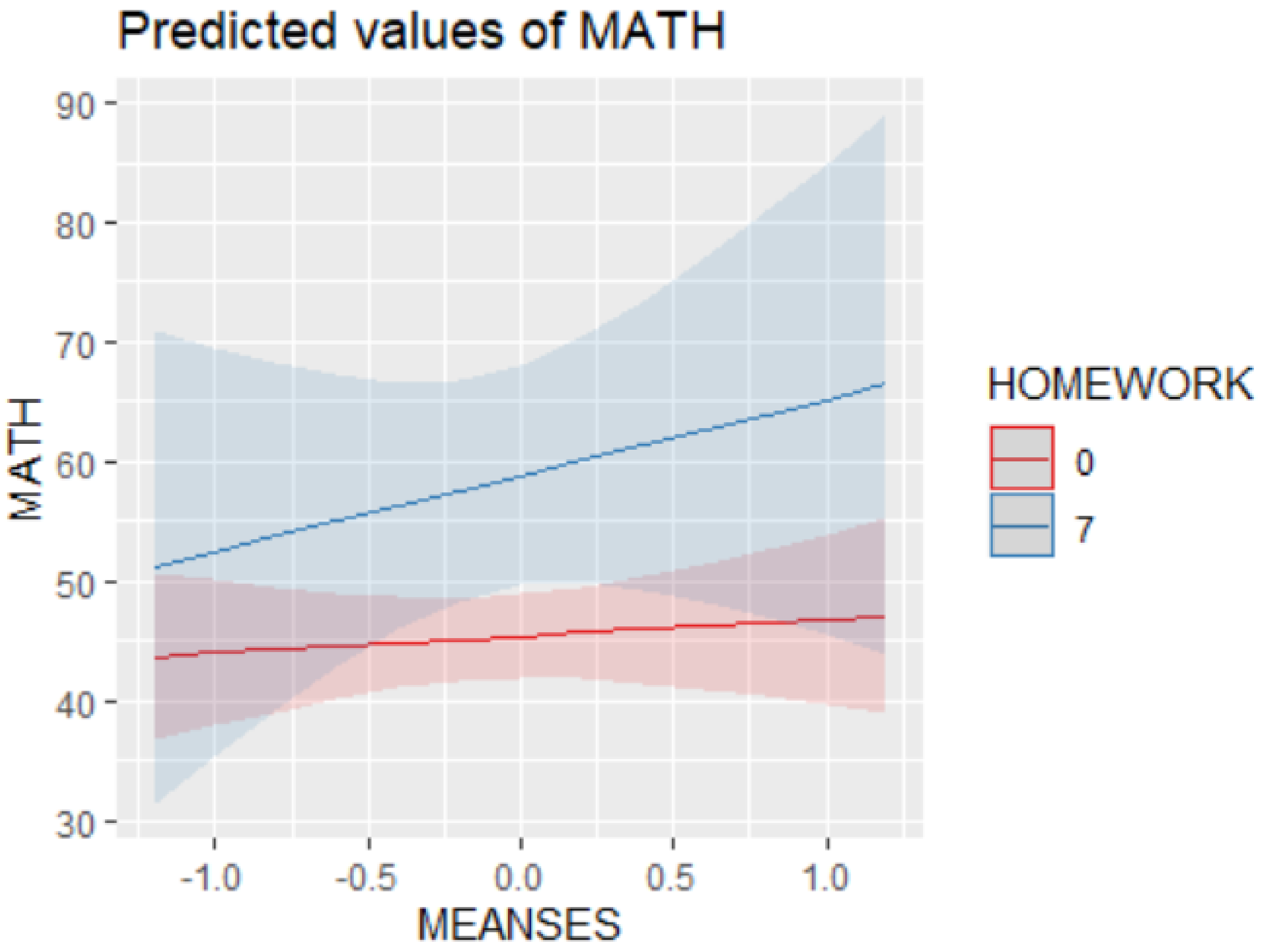


Visualization: RE

```
> plot_model(m4, type = 're',
             sort.est = 'HOMEWORK')
```



```
plot_model(m4, type = "pred",terms = c("MEANSES", "HOMEWORK[0,7]"))
```



Lab

- Vary the effect of parents' education among the schools. Check if the random slope model is better.
- Check if the type of school affects the results in math.
- Check if the effect of parental education differs depending on the school type.
- Indicate which school has the lowest impact of parents’ education.
- Plot the effects of your final model.
- Plot the interaction term and comment on the results
- Calculate R2.