Multilevel Multilingual

Multilevel Models in Stata, R and Julia

Andrew Grogan-Kaylor

2024-03-20

Table of contents

1 Multilevel Multilingual **6**
1.1 Introduction 6
1.2 The Data 6
1.3 An Introduction To Equations and Syntax 6
1.3.1 Stata 7
1.3.2 R 7
1.3.3 Julia 7

2 Descriptive Statistics **8**
2.1 Descriptive Statistics 8
2.1.1 Stata 8
2.1.2 R 8
2.1.3 Julia 9

3 Unconditional Model **11**
3.1 The Equation 11
3.2 Run Models 11
3.2.1 Stata 11
3.2.2 R 12
3.2.3 Julia 14

4 Cross Sectional Model **15**
4.1 The Equation 15
4.2 Stata 15
4.2.1 Get The Data 15
4.2.2 Graph 15
4.2.3 Run The Model 15
4.3 R 18
4.3.1 Load The Needed Packages And Get The Data 18
4.3.2 Graph 18
4.3.3 Run The Model 19
4.4 Julia 20
4.4.1 Load The Needed Packages And Get The Data 20
4.4.2 Graph 20
4.4.3 Change Country To Categorical 21

4.4.4 Run The Model 21

List of Figures

4.1 Outcome by Parental Warmth (Stata) 16
4.2 Outcome by Parental Warmth (R) 19
4.3 Outcome by Parental Warmth (Julia) 21

List of Tables

1.1 Sample of Simulated Multilevel Data 6

1 Multilevel Multilingual

1.1 Introduction

Below, I describe the use of Stata, R, and Julia to estimate multilevel models.

> **Results Will Vary Slightly**
>
> Estimating multilevel models is a complex endeavor. The details of how this is accomplished are beyond the purview of this book. Suffice it to say that across software packages there will be differences in estimation routines, resulting in some numerical differences in the results provided by different software packages. Substantively speaking, however, results should agree across software.

1.2 The Data

The examples use the `simulated_multilevel_data.dta` file from *Multilevel Thinking*. Here is a direct link to download the data.

Table 1.1: Sample of Simulated Multilevel Data

country	HDI	family	id	group	physical_punishment	warmth	outcome
1	69	1	1.1	2	2	3	59.18
1	69	2	1.2	2	4	0	61.54
1	69	3	1.3	1	4	4	51.87
1	69	4	1.4	2	0	6	51.71
1	69	5	1.5	2	3	2	55.88
1	69	6	1.6	1	5	3	60.78

1.3 An Introduction To Equations and Syntax

To explain statistical syntax for each software, I consider the general case of a multilevel model with dependent variable `y`, independent variables `x` and `z`, clustering variable `group`, and a random slope for `x`. i is the index for the person, while j is the index for the `group`.

$$y = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + u_{0j} + u_{1j} \times x_{ij} + e_{ij} \tag{1.1}$$

1.3.1 Stata

In Stata `mixed`, the syntax for a multilevel model of the form described in Equation 1.1 is:

```
mixed y x || group: x
```

1.3.2 R

In R `lme4`, the general syntax for a multilevel model of the form described in Equation 1.1 is:

```
library(lme4)

lmer(y ~ x + z + (1 + x || group), data = ...)
```

1.3.3 Julia

In Julia `MixedModels`, the general syntax for a multilevel model of the form described in Equation 1.1 is:

```
using MixedModels

fit(MixedModel, @formula(y ~ x + z + (1 + x | group)), data)
```

2 Descriptive Statistics

2.1 Descriptive Statistics

2.1.1 Stata

```
use simulated_multilevel_data.dta // use data
```

We use `summarize` for *continuous* variables, and `tabulate` for *categorical* variables.

```
summarize outcome warmth physical_punishment HDI

tabulate group
```

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
     outcome |      3,000    53.46757     6.65179   33.39014   76.75101
      warmth |      3,000    3.524333    1.889956          0          7
physical_p~t |      3,000    2.494667    1.380075          0          5
         HDI |      3,000    64.76667    17.24562         33         87
```

```
  arbitrary |
      group |
   variable |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      1,507       50.23       50.23
          2 |      1,493       49.77      100.00
------------+-----------------------------------
      Total |      3,000      100.00
```

2.1.2 R

```r
library(haven) # read data in Stata format

df <- read_dta("simulated_multilevel_data.dta")
```

R's descriptive statistics functions rely heavily on whether a variable is a *numeric* variable, or a *factor* variable. Below, I convert two variables to factors (`factor`) before using `summary`[1] to generate descriptive statistics.

```r
df$country <- factor(df$country)

df$group <- factor(df$group)

summary(df)
```

```
    country          HDI            family            id               group   
 1      : 100   Min.   :33.00   Min.   :  1.00   Length:3000        1:1507  
 2      : 100   1st Qu.:53.00   1st Qu.: 25.75   Class :character   2:1493  
 3      : 100   Median :70.00   Median : 50.50   Mode  :character           
 4      : 100   Mean   :64.77   Mean   : 50.50                              
 5      : 100   3rd Qu.:81.00   3rd Qu.: 75.25                              
 6      : 100   Max.   :87.00   Max.   :100.00                              
 (Other):2400                                                               
 physical_punishment     warmth         outcome     
 Min.   :0.000       Min.   :0.000   Min.   :33.39  
 1st Qu.:2.000       1st Qu.:2.000   1st Qu.:48.78  
 Median :3.000       Median :4.000   Median :53.64  
 Mean   :2.495       Mean   :3.524   Mean   :53.47  
 3rd Qu.:3.250       3rd Qu.:5.000   3rd Qu.:58.06  
 Max.   :5.000       Max.   :7.000   Max.   :76.75  
```

2.1.3 Julia

```julia
using Tables, MixedModels, MixedModelsExtras, StatFiles, DataFrames, CategoricalArrays, Datal

df = DataFrame(load("simulated_multilevel_data.dta"))
```

[1]`skimr` is an excellent new alternative library for generating descriptive statistics in R.

Similarly to R, Julia relies on the idea of *variable type*. I use `transform` to convert the appropriate variables to *categorical* variables.

```
@transform!(df, :country = categorical(:country))

@transform!(df, :group = categorical(:group))
```

```
describe(df) # descriptive statistics
```

```
8×7 DataFrame
 Row   variable             mean      min      median   max      nmissing  eltyp
       Symbol               Union…    Any      Union…   Any      Int64     Union

   1   country                        1.0               30.0            0  Union
   2   HDI                  64.7667   33.0     70.0     87.0            0  Union
   3   family               50.5      1.0      50.5     100.0           0  Union
   4   id                             1.1               9.99            0  Union
   5   group                          1.0               2.0             0  Union
   6   physical_punishment  2.49467   0.0      3.0      5.0             0  Union
   7   warmth               3.52433   0.0      4.0      7.0             0  Union
   8   outcome              53.4676   33.3901  53.6426  76.751          0  Union
                                                                1 column omitted
```

3 Unconditional Model

An *unconditional* multilevel model is a model with no independent variables. One should always run an unconditional model as the first step of a multilevel model in order to get a sense of the way that variation is apportioned in the model across the different levels.

3.1 The Equation

$$\text{outcome}_{ij} = \beta_0 + u_{0j} + e_{ij} \tag{3.1}$$

The Intraclass Correlation Coefficient (ICC) is given by:

$$\text{ICC} = \frac{var(u_{0j})}{var(u_{0j}) + var(e_{ij})} \tag{3.2}$$

In a two level multilevel model, the ICC provides a measure of the amount of variation attributable to Level 2.

3.2 Run Models

3.2.1 Stata

```
use simulated_multilevel_data.dta // use data
```

```
mixed outcome || country: // unconditional model
```

```
Performing EM optimization ...

Performing gradient-based optimization:
Iteration 0:  Log likelihood = -9856.1548
Iteration 1:  Log likelihood = -9856.1548
```

```
Computing standard errors ...

Mixed-effects ML regression                     Number of obs    = 3,000
Group variable: country                         Number of groups =    30
                                                Obs per group:
                                                             min =   100
                                                             avg = 100.0
                                                             max =   100
                                                Wald chi2(0)     =     .
Log likelihood = -9856.1548                     Prob > chi2      =     .

------------------------------------------------------------------------------
     outcome | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _cons |   53.46757   .3539097   151.08   0.000     52.77392    54.16122
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects parameters  |   Estimate   Std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
country: Identity            |
                  var(_cons) |   3.348734   .9702594      1.897816    5.908906
-----------------------------+------------------------------------------------
               var(Residual) |   40.88284   1.060908       38.8555    43.01597
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 169.64        Prob >= chibar2 = 0.0000
```

```
estat icc // ICC
```

```
Intraclass correlation

------------------------------------------------------------------------------
                       Level |        ICC   Std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
                     country |   .0757091   .0203761      .0442419    .1265931
------------------------------------------------------------------------------
```

3.2.2 R

```r
library(haven)

df <- read_dta("simulated_multilevel_data.dta")
```

```r
library(lme4) # estimate multilevel models

fit0 <- lmer(outcome ~ (1 | country),
             data = df) # unconditional model

summary(fit0)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: outcome ~ (1 | country)
   Data: df

REML criterion at convergence: 19712.5

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.97650 -0.68006  0.00936  0.67580  3.03510

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept)  3.478   1.865
 Residual             40.883   6.394
Number of obs: 3000, groups:  country, 30

Fixed effects:
            Estimate Std. Error t value
(Intercept)    53.47       0.36   148.5
```

```r
library(performance)

performance::icc(fit0) # ICC
```

```
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.078
  Unadjusted ICC: 0.078
```

3.2.3 Julia

```julia
using Tables, MixedModels, MixedModelsExtras,
StatFiles, DataFrames, CategoricalArrays, DataFramesMeta

df = DataFrame(load("simulated_multilevel_data.dta"))
```

```julia
@transform!(df, :country = categorical(:country))
```

```julia
m0 = fit(MixedModel,
         @formula(outcome ~ (1 | country)), df) # unconditional model
```

```
Linear mixed model fit by maximum likelihood
 outcome ~ 1 + (1 | country)
   logLik   -2 logLik     AIC       AICc        BIC
 -9856.1548 19712.3097 19718.3097 19718.3177 19736.3288

Variance components:
            Column   Variance Std.Dev.
country  (Intercept)   3.34871 1.82995
Residual              40.88285 6.39397
 Number of obs: 3000; levels of grouping factors: 30

  Fixed-effects parameters:

               Coef.  Std. Error       z  Pr(>|z|)

(Intercept)  53.4676    0.353908  151.08    <1e-99
```

```julia
icc(m0) # ICC
```

```
0.07570852291396266
```

4 Cross Sectional Model

4.1 The Equation

Recall the general model of Equation 1.1, and the syntax outlined in Section 1.3. Below in Equation 4.1, we consider a more substantive example.

$$\text{outcome}_{ij} = \beta_0 + \beta_1 \text{warmth}_{ij} + \beta_2 \text{physical punishment}_{ij} + \beta_3 \text{group}_{ij} + \beta_4 \text{HDI}_{ij} + u_{0j} + u_{1j} \times \text{warmth}_{ij} + e_{ij} \tag{4.1}$$

4.2 Stata

4.2.1 Get The Data

```
use simulated_multilevel_data.dta
```

4.2.2 Graph

```
twoway scatter outcome warmth, ///
  xtitle("warmth") ytitle("outcome") ///
  title("Outcome by Parental Warmth")

quietly graph export scatter.png, replace
```

4.2.3 Run The Model

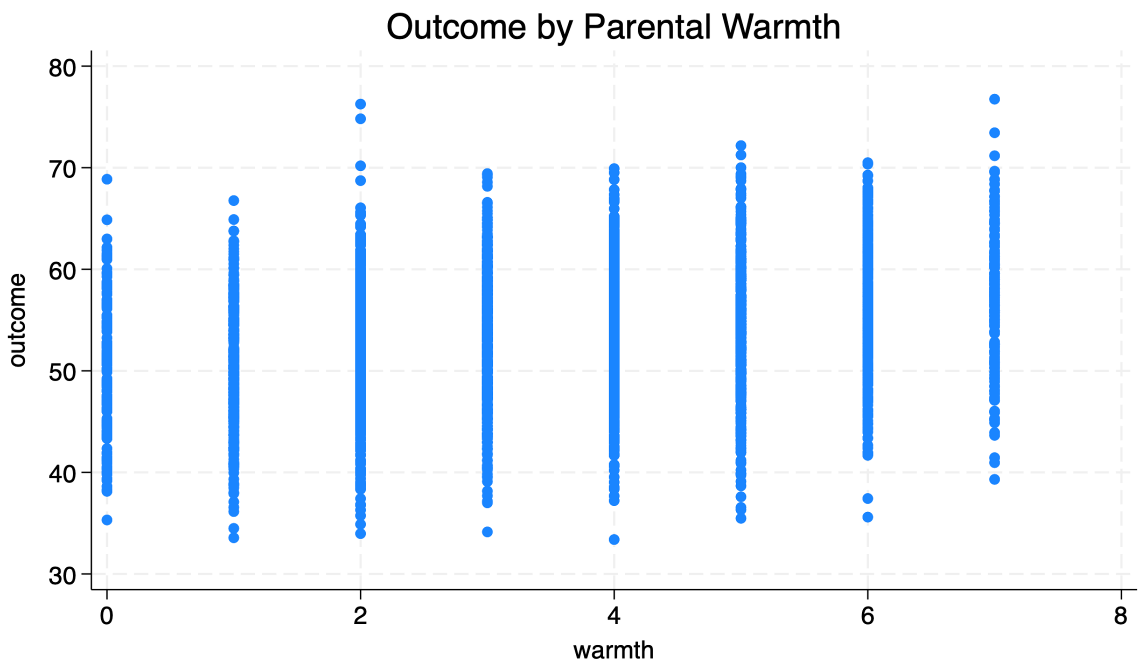

Figure 4.1: Outcome by Parental Warmth (Stata)

```
mixed outcome warmth physical_punishment group HDI || country: warmth
```

```
Performing EM optimization ...

Performing gradient-based optimization:
Iteration 0:  Log likelihood =  -9668.198
Iteration 1:  Log likelihood = -9667.9551
Iteration 2:  Log likelihood = -9667.9534
Iteration 3:  Log likelihood = -9667.9533
Iteration 4:  Log likelihood = -9667.9532

Computing standard errors ...

Mixed-effects ML regression                          Number of obs    =  3,000
Group variable: country                              Number of groups =     30
                                                     Obs per group:
                                                                  min =    100
                                                                  avg =  100.0
                                                                  max =    100
                                                     Wald chi2(4)     = 401.26
Log likelihood = -9667.9532                          Prob > chi2      = 0.0000

------------------------------------------------------------------------------------
            outcome | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
--------------------+---------------------------------------------------------------
             warmth |   .9616447   .0581825    16.53   0.000     .8476091     1.07568
physical_punishment |  -.8453802   .0798155   -10.59   0.000    -1.001816   -.6889448
              group |   1.084344   .2200539     4.93   0.000     .6530461    1.515642
                HDI |    .010557   .0204522     0.52   0.606    -.0295286    .0506426
              _cons |   49.87963   1.436612    34.72   0.000     47.06392    52.69534
------------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects parameters  |   Estimate   Std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
country: Independent         |
               var(warmth) |   1.83e-06   .0000173      1.76e-14    190.9774
                var(_cons) |   3.370262   .9633726      1.924651    5.901676
-----------------------------+------------------------------------------------
            var(Residual) |   36.01906   .9346936      34.23291    37.89842
------------------------------------------------------------------------------
```

```
LR test vs. linear model: chi2(2) = 198.01                Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

4.3 R

4.3.1 Load The Needed Packages And Get The Data

```r
library(haven)

df <- read_dta("simulated_multilevel_data.dta")
```

4.3.2 Graph

```r
library(ggplot2)

ggplot(df,
       aes(x = warmth,
           y = outcome)) +
  geom_point() +
  labs(title = "Outcome by Parental Warmth")
```

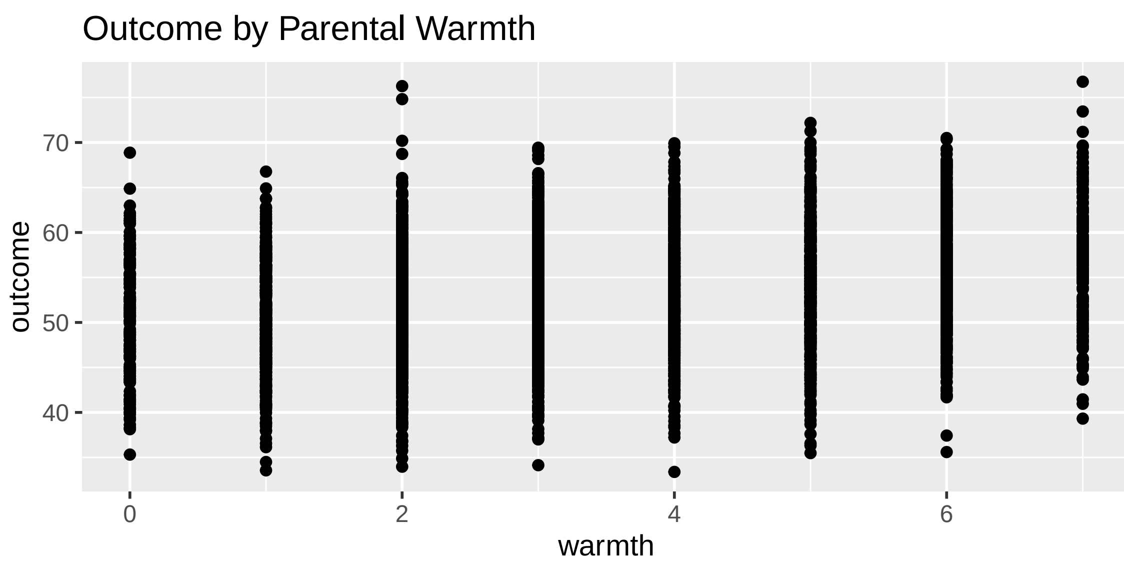

Figure 4.2: Outcome by Parental Warmth (R)

4.3.3 Run The Model

```
fit1 <- lmer(outcome ~ warmth + physical_punishment +
               group + HDI +
               (1 + warmth || country),
             data = df)

summary(fit1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: outcome ~ warmth + physical_punishment + group + HDI + ((1 |
    country) + (0 + warmth | country))
   Data: df

REML criterion at convergence: 19350.3

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.4496 -0.6807  0.0016  0.6864  3.1792 

Random effects:
```

```
 Groups    Name        Variance  Std.Dev.
 country   (Intercept)  3.611568 1.90041
 country.1 warmth       0.001876 0.04331
 Residual              36.049124 6.00409
Number of obs: 3000, groups:  country, 30

Fixed effects:
                    Estimate Std. Error t value
(Intercept)         49.88754    1.48203  33.662
warmth               0.96155    0.05875  16.367
physical_punishment -0.84556    0.07986 -10.588
group                1.08471    0.22017   4.927
HDI                  0.01044    0.02116   0.493

Correlation of Fixed Effects:
            (Intr) warmth physc_ group
warmth      -0.126
physcl_pnsh -0.135 -0.025
group       -0.218 -0.010 -0.019
HDI         -0.925 -0.006  0.008 -0.001
```

4.4 Julia

4.4.1 Load The Needed Packages And Get The Data

```
using Tables, MixedModels, StatFiles, DataFrames, CategoricalArrays, DataFramesMeta

df = DataFrame(load("simulated_multilevel_data.dta"))
```

4.4.2 Graph

```
using StatsPlots

@df df scatter(:outcome, :warmth,
               title = "Outcome by Parental Warmth",
               ylabel = "outcome",
               xlabel = "parental warmth")
```

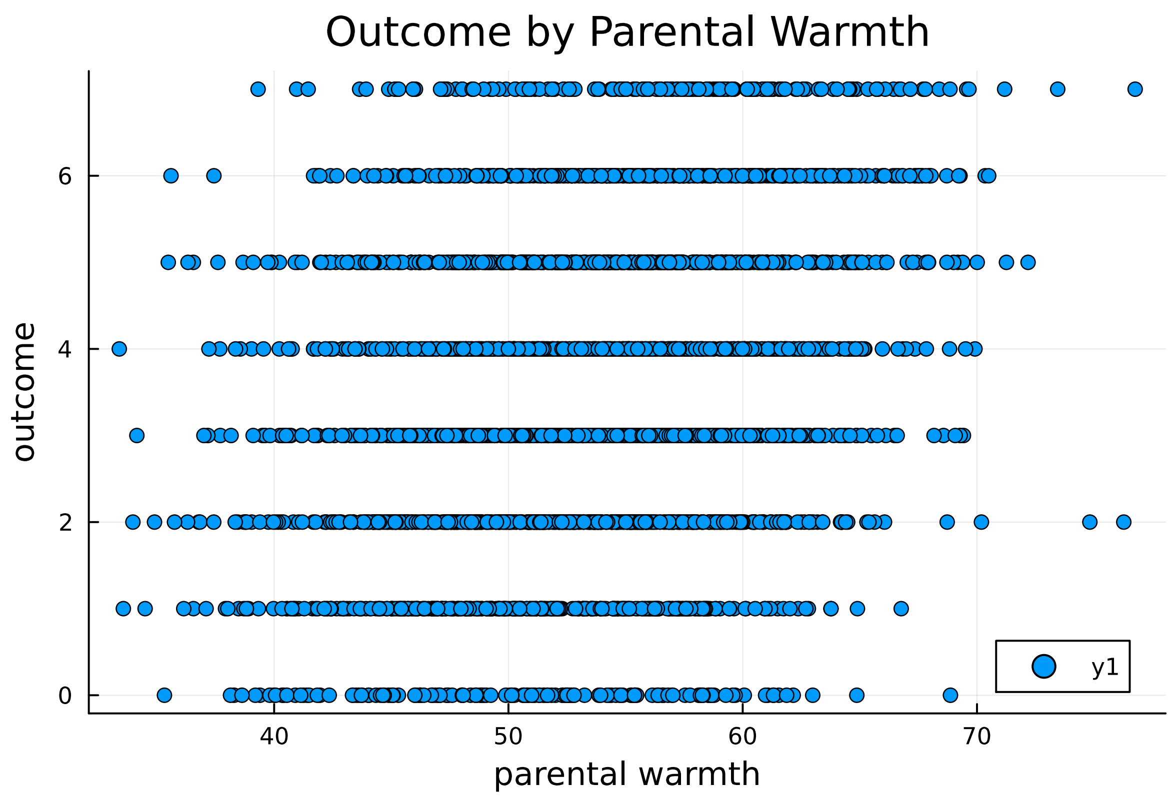

Figure 4.3: Outcome by Parental Warmth (Julia)

4.4.3 Change Country To Categorical

```
@transform!(df, :country = categorical(:country))
```

4.4.4 Run The Model

```
m1 = fit(MixedModel, @formula(outcome ~ warmth + physical_punishment +
                group + HDI +
                (1 + warmth | country)), df)
```

```
Linear mixed model fit by maximum likelihood
 outcome ~ 1 + warmth + physical_punishment + group + HDI + (1 + warmth | country)
   logLik   -2 logLik     AIC       AICc        BIC
```

```
 -9667.9392 19335.8783 19353.8783 19353.9385 19407.9357

Variance components:
            Column    Variance   Std.Dev.   Corr.
country  (Intercept)   3.2369484 1.7991521
         warmth        0.0001080 0.0103903 +1.00
Residual              36.0187144 6.0015593
 Number of obs: 3000; levels of grouping factors: 30

  Fixed-effects parameters:

                          Coef.  Std. Error       z  Pr(>|z|)

(Intercept)          49.9018      1.43435     34.79    <1e-99
warmth                0.961545    0.0582135   16.52    <1e-60
physical_punishment  -0.845389    0.0798149  -10.59    <1e-25
group                 1.08524     0.220055     4.93    <1e-06
HDI                   0.0101984   0.0204401    0.50    0.6178
```