Multilevel Modeling: Principles and Applications in R

Day 2: Random Slopes

Constantin Manuel Bosancianu

WZB Berlin Social Science Center
Institutions and Political Inequality
manuel.bosancianu@gmail.com

January 21, 2021

Today

Centering in MLMs: why it's done, and how to do it.

Today

Centering in MLMs: why it's done, and how to do it.

How to specify *random slopes* in such models.

Today

Centering in MLMs: why it's done, and how to do it.

How to specify *random slopes* in such models.

How to interpret and display effects from *cross-level* interactions.

Today

Centering in MLMs: why it's done, and how to do it.

How to specify *random slopes* in such models.

How to interpret and display effects from *cross-level* interactions.

Sample size considerations in MLM.

Centering predictors

Reasons for centering

So far, you encountered the concept in regression:

- ✓ to make intercept "more meaningful"

Reasons for centering

So far, you encountered the concept in regression:

- ✓ to make intercept "more meaningful"
- ✓ to be able to compare effect strength among predictors (*standardization*)

$$x - \bar{x} \tag{1}$$

Reasons for centering

So far, you encountered the concept in regression:

- ✓ to make intercept "more meaningful"
- ✓ to be able to compare effect strength among predictors (*standardization*)

$$x - \bar{x} \tag{1}$$

With clustered data, centering represents a technical solution to a unique problem.

Reasons for centering

So far, you encountered the concept in regression:

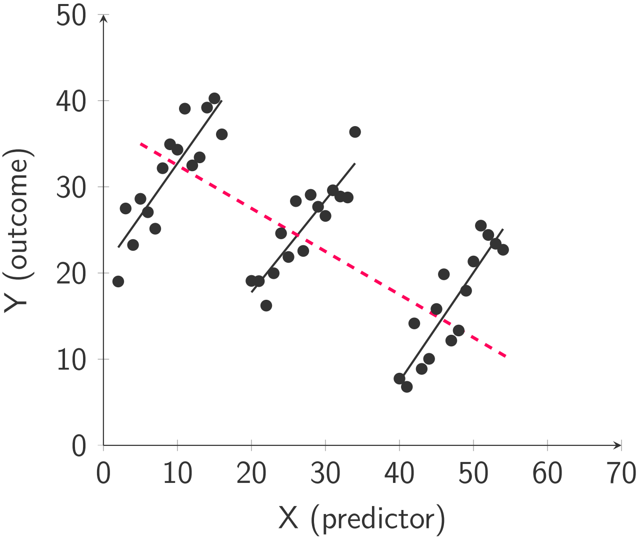
- ✓ to make intercept "more meaningful"
- ✓ to be able to compare effect strength among predictors (*standardization*)

$$x - \bar{x} \tag{1}$$

With clustered data, centering represents a technical solution to a unique problem.

Without a correction, the coefficient for an individual-level (unit-level, or L1) variable is a mix of "within-group" and "between-group" relationships.

Mix of patterns



Group-mean centering

For this problem, centering offers a solution: "artificially" erase the between-group variation in the individual-level variables.

Group-mean centering

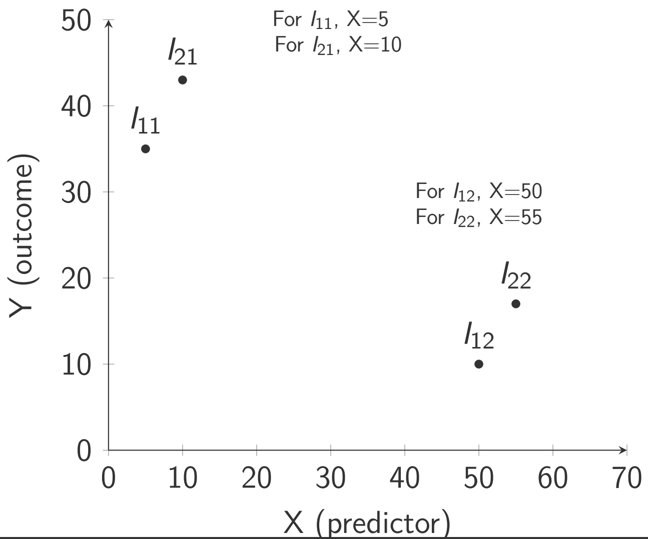
For this problem, centering offers a solution: "artificially" erase the between-group variation in the individual-level variables.

For each j group, one can center a variable X, with the formula

$$X_{centered} = X_{ij} - \overline{X_j} \tag{2}$$

$\bar{X}_j$ is the mean of variable X in the j group. This is called **group-mean centering** (or *centering within clusters*).

Centering variables (1)



Centering variables (2)

For group 1, $\overline{X}$=7.5. For group 2, $\overline{X}$=52.5.

	X raw	X centered
I_{11}	5	-2.5
I_{21}	10	2.5
I_{12}	50	-2.5
I_{22}	55	2.5

Centering variables (2)

For group 1, $\overline{X}$=7.5. For group 2, $\overline{X}$=52.5.

	X raw	X centered
I_{11}	5	-2.5
I_{21}	10	2.5
I_{12}	50	-2.5
I_{22}	55	2.5

With group-mean centering, the only thing that is left over is the relative position of individuals within a group, e.g. the distance between I_{11} and I_{21} remains 5 after centering.

Grand-mean centering

There is a second type of centering, meant for level 2 variables: **grand-mean centering**.

$$Z_{centered} = Z_j - \overline{Z} \quad (3)$$

In this procedure, we subtract from each group's value on Z the mean of the Zs of all groups.

Practical advice

Recommendations made by Enders and Tofighi (2007), depending on estimates of interest:

- ✓ $X_{L1} \rightsquigarrow Y$: group-mean centering for X

Practical advice

Recommendations made by Enders and Tofighi (2007), depending on estimates of interest:

- ✓ $X_{L1} \rightsquigarrow Y$: group-mean centering for X
- ✓ $X_{L2} \rightsquigarrow Y$: grand-mean centering for X

Practical advice

Recommendations made by Enders and Tofighi (2007), depending on estimates of interest:

- ✓ $X_{L1} \rightsquigarrow Y$: group-mean centering for X
- ✓ $X_{L2} \rightsquigarrow Y$: grand-mean centering for X
- ✓ $X_{L2} \times X_{L1}$: grand-mean for X_{L2}, group-mean for X_{L1}

Practical advice

Recommendations made by Enders and Tofighi (2007), depending on estimates of interest:

- ✓ $X_{L1} \rightsquigarrow Y$: group-mean centering for X
- ✓ $X_{L2} \rightsquigarrow Y$: grand-mean centering for X
- ✓ $X_{L2} \times X_{L1}$: grand-mean for X_{L2}, group-mean for X_{L1}
- ✓ $X_{L2} \times X_{L2}$: grand-mean for both

Practical advice

Recommendations made by Enders and Tofighi (2007), depending on estimates of interest:

- ✓ $X_{L1} \rightsquigarrow Y$: group-mean centering for X
- ✓ $X_{L2} \rightsquigarrow Y$: grand-mean centering for X
- ✓ $X_{L2} \times X_{L1}$: grand-mean for X_{L2}, group-mean for X_{L1}
- ✓ $X_{L2} \times X_{L2}$: grand-mean for both
- ✓ $X_{L1} \times X_{L1}$: group-mean for both

Practical advice

Recommendations made by Enders and Tofighi (2007), depending on estimates of interest:

- ✓ $X_{L1} \rightsquigarrow Y$: group-mean centering for X
- ✓ $X_{L2} \rightsquigarrow Y$: grand-mean centering for X
- ✓ $X_{L2} \times X_{L1}$: grand-mean for X_{L2}, group-mean for X_{L1}
- ✓ $X_{L2} \times X_{L2}$: grand-mean for both
- ✓ $X_{L1} \times X_{L1}$: group-mean for both

Unlike in standard regression, centering in MLMs can (and should!) change the magnitude of the estimate.

Random slopes

MLM specification (1)

We used a very simple model yesterday:

$$
\begin{cases}
EFF_{ij} = \beta_{0j} + \beta_{1j} * EDU_{ij} + \beta_{2j} * URB_{ij} + \epsilon_{ij} \\
\beta_{0j} = \gamma_{00} + \gamma_{01} * CORR_j + \upsilon_{0j} \\
\beta_{1j} = \gamma_{10} \\
\beta_{2j} = \gamma_{20}
\end{cases}
\tag{4}
$$

MLM specification (1)

We used a very simple model yesterday:

$$\begin{cases} EFF_{ij} = \beta_{0j} + \beta_{1j} * EDU_{ij} + \beta_{2j} * URB_{ij} + \epsilon_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01} * CORR_j + \upsilon_{0j} \\ \beta_{1j} = \gamma_{10} \\ \beta_{2j} = \gamma_{20} \end{cases} \tag{4}$$

Our theory could extend to how the *effect* of a level-1 predictor varies based on a level-2 predictor.

MLM specification (1)

We used a very simple model yesterday:

$$
\begin{cases}
EFF_{ij} = \beta_{0j} + \beta_{1j} * EDU_{ij} + \beta_{2j} * URB_{ij} + \epsilon_{ij} \\
\beta_{0j} = \gamma_{00} + \gamma_{01} * CORR_j + \upsilon_{0j} \\
\beta_{1j} = \gamma_{10} \\
\beta_{2j} = \gamma_{20}
\end{cases}
\tag{4}
$$

Our theory could extend to how the *effect* of a level-1 predictor varies based on a level-2 predictor.

How does the effect of education vary?

MLM specification (2)

$$
\begin{cases}
EFF_{ij} = \beta_{0j} + \beta_{1j} * EDU_{ij} + \beta_{2j} * URB_{ij} + \epsilon_{ij} \\
\beta_{0j} = \gamma_{00} + \gamma_{01} * CORR_j + \upsilon_{0j} \\
\beta_{1j} = \gamma_{10} + \gamma_{11} * CORR_j + \upsilon_{1j} \\
\beta_{2j} = \gamma_{20}
\end{cases}
\tag{5}
$$

MLM specification (2)

$$
\begin{cases}
EFF_{ij} = \beta_{0j} + \beta_{1j} * EDU_{ij} + \beta_{2j} * URB_{ij} + \epsilon_{ij} \\
\beta_{0j} = \gamma_{00} + \gamma_{01} * CORR_j + \upsilon_{0j} \\
\beta_{1j} = \gamma_{10} + \gamma_{11} * CORR_j + \upsilon_{1j} \\
\beta_{2j} = \gamma_{20}
\end{cases}
\tag{5}
$$

If we choose to, we can also write up a model for the effect of urban residence on political efficacy.

MLM specification—extended form

The extended form:

$$
\begin{aligned}
EFF_{ij} = {} & \overbrace{\gamma_{00} + \gamma_{01} * CORR_j + \upsilon_{0j}}^{\beta_{0j}} + \overbrace{(\gamma_{10} + \gamma_{11} * CORR_j + \upsilon_{1j})}^{\beta_{1j}} * EDU_{ij} + \\
& + \gamma_{20} * URB_{ij} + \epsilon_{ij} = \\
= {} & \gamma_{00} + \gamma_{01} * CORR_j + \upsilon_{0j} + \gamma_{10} * EDU_{ij} + \gamma_{11} * EDU_{ij} * CORR_j + \\
& + \upsilon_{1j} * EDU_{ij} + \gamma_{20} * URB_{ij} + \epsilon_{ij} = \\
= {} & \gamma_{00} + \gamma_{10} * EDU_{ij} + \gamma_{20} * URB_{ij} + \gamma_{01} * CORR_j + \gamma_{11} * EDU_{ij} * CORR_j + \\
& + \upsilon_{1j} * EDU_{ij} + \upsilon_{0j} + \epsilon_{ij}
\end{aligned}
\tag{6}
$$

Extended form

The extended form is useful as that's the way `Stata`, `R`, and even `SPSS` ask you to specify the model. `Mplus` is an exception here.

Extended form

The extended form is useful as that's the way `Stata`, `R`, and even `SPSS` ask you to specify the model. `Mplus` is an exception here.

Fixed effects: γ_{00}, γ_{10}, γ_{20}, γ_{01}, and γ_{11}.

Extended form

The extended form is useful as that's the way `Stata`, `R`, and even `SPSS` ask you to specify the model. `Mplus` is an exception here.

Fixed effects: γ_{00}, γ_{10}, γ_{20}, γ_{01}, and γ_{11}.

Random effects: υ_{0j}, υ_{1j}, and ϵ_{ij}.

Extended form

The extended form is useful as that's the way `Stata`, `R`, and even `SPSS` ask you to specify the model. `Mplus` is an exception here.

Fixed effects: γ_{00}, γ_{10}, γ_{20}, γ_{01}, and γ_{11}.

Random effects: υ_{0j}, υ_{1j}, and ϵ_{ij}.

$$
\begin{cases}
EFF_{ij} = \beta_{0j} + \beta_{1j} * EDU_{ij} + \beta_{2j} * URB_{ij} + \epsilon_{ij} \\
\beta_{0j} = \gamma_{00} + \gamma_{01} * CORR_j + \upsilon_{0j} \\
\beta_{1j} = \gamma_{10} + \gamma_{11} * CORR_j + \upsilon_{1j} \\
\beta_{2j} = \gamma_{20}
\end{cases}
\tag{7}
$$

Only the highlighted quantities are being estimated in the model.

Our example—continued

	DV: political efficacy
(Intercept)	$2.961\ (0.041)^{***}$
Woman	$-0.131\ (0.008)^{***}$
Education (years)	$0.303\ (0.008)^{***}$
Income	$0.182\ (0.008)^{***}$
Urban residence	$0.062\ (0.009)^{***}$
Perceived absence of corruption	$0.439\ (0.083)^{***}$
Num. obs.	32020
Num. groups: country	31
Var: country (Intercept)	0.051
Var: Residual	0.474

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Our example—a random slope (1)

We first allow for a slope to vary, to see whether there is sufficient variation to be explained (a simpler specification than in Equation 6).

Our example—a random slope (1)

We first allow for a slope to vary, to see whether there is sufficient variation to be explained (a simpler specification than in Equation 6).

$$
\begin{cases}
EFF_{ij} = \beta_{0j} + \beta_{1j}EDU_{ij} + \beta_{2j}URB_{ij} + \beta_{3j}FEM_{ij} + \beta_{4j}INC_{ij} + \epsilon_{ij} \\
\beta_{0j} = \gamma_{00} + \gamma_{01}CORR_j + \upsilon_{0j} \\
\beta_{1j} = \gamma_{10} + \upsilon_{1j} \\
\beta_{2j} = \gamma_{20} \\
\beta_{3j} = \gamma_{30} \\
\beta_{4j} = \gamma_{40}
\end{cases}
\tag{8}
$$

Our example—a random slope (1)

We first allow for a slope to vary, to see whether there is sufficient variation to be explained (a simpler specification than in Equation 6).

$$
\begin{cases}
EFF_{ij} = \beta_{0j} + \beta_{1j}EDU_{ij} + \beta_{2j}URB_{ij} + \beta_{3j}FEM_{ij} + \beta_{4j}INC_{ij} + \epsilon_{ij} \\
\beta_{0j} = \gamma_{00} + \gamma_{01}CORR_j + \upsilon_{0j} \\
\beta_{1j} = \gamma_{10} + \upsilon_{1j} \\
\beta_{2j} = \gamma_{20} \\
\beta_{3j} = \gamma_{30} \\
\beta_{4j} = \gamma_{40}
\end{cases}
\tag{8}
$$

Not yet explaining *why* the slope of education varies between countries.

Our example—a random slope (2)

Extended form of the model is easy to produce:

$$\begin{aligned} EFF_{ij} = \gamma_{00} + \gamma_{10} EDU_{ij} + \gamma_{20} URB_{ij} + \gamma_{30} FEM_{ij} + \gamma_{40} INC_{ij} + \gamma_{01} * CORR_j + \\ + \upsilon_{1j} * EDU_{ij} + \upsilon_{0j} + \epsilon_{ij} \end{aligned} \tag{9}$$

Our example—a random slope (2)

Extended form of the model is easy to produce:

$$
\begin{aligned}
EFF_{ij} = \gamma_{00} + \gamma_{10} EDU_{ij} + \gamma_{20} URB_{ij} + \gamma_{30} FEM_{ij} + \gamma_{40} INC_{ij} + \gamma_{01} * CORR_j + \\
+ \upsilon_{1j} * EDU_{ij} + \upsilon_{0j} + \epsilon_{ij}
\end{aligned}
\tag{9}
$$

```
model2 <- lmer(poleff ~ 1 + educCWC + urbanCWC + femaleCWC + incCWC +
                  cpiCGM + (1 + educCWC| cnt),
                data=ISSPsub)
```

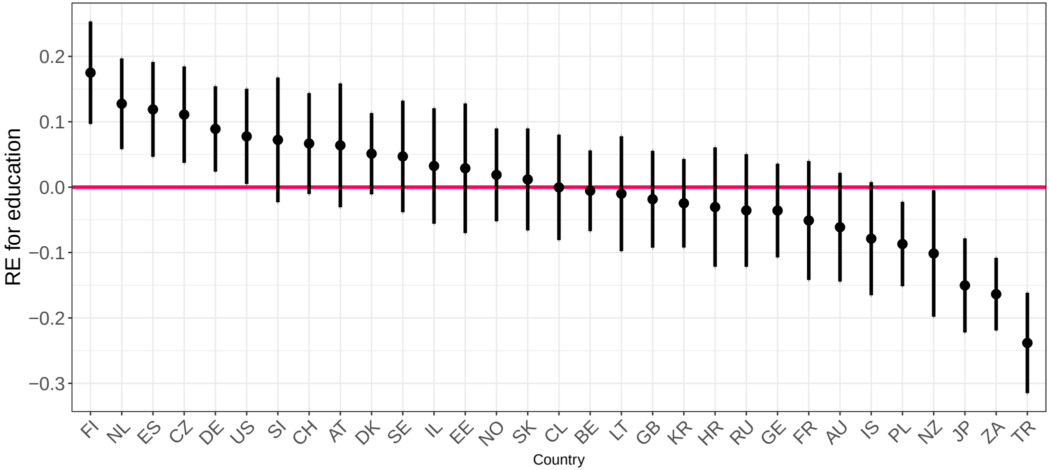
`lmer()` formula implements extended form of model almost verbatim.

Our example—a random slope (3)

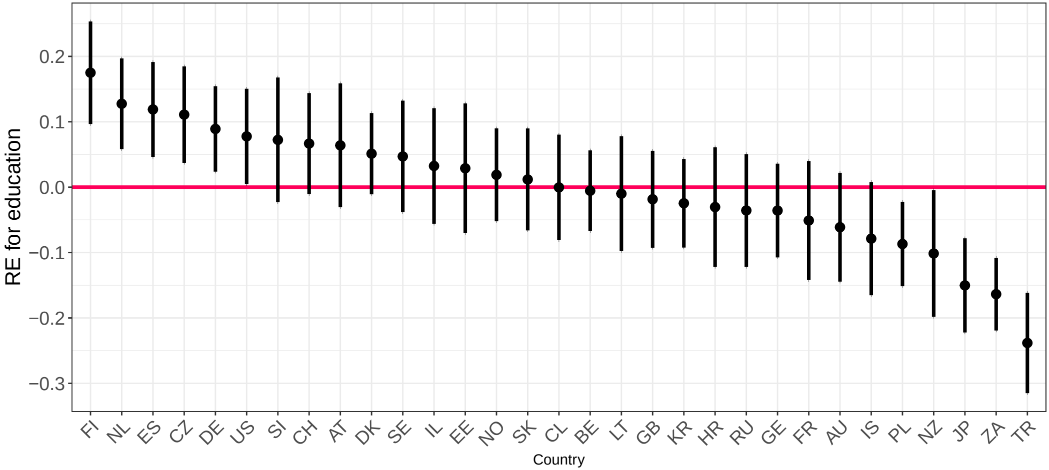
	DV: political efficacy
(Intercept)	2.961 (0.041)***
Education (years)	0.306 (0.020)***
Urban residence	0.065 (0.009)***
Woman	−0.132 (0.008)***
Income	0.182 (0.008)***
Perceived absence of corruption	0.494 (0.078)***
Num. obs.	32020
Num. groups: cnt	31
Var: country (Intercept)	0.051
Var: country Education	0.010
Cov: country (Intercept) Education	−0.009
Var: Residual	0.471

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Random effects

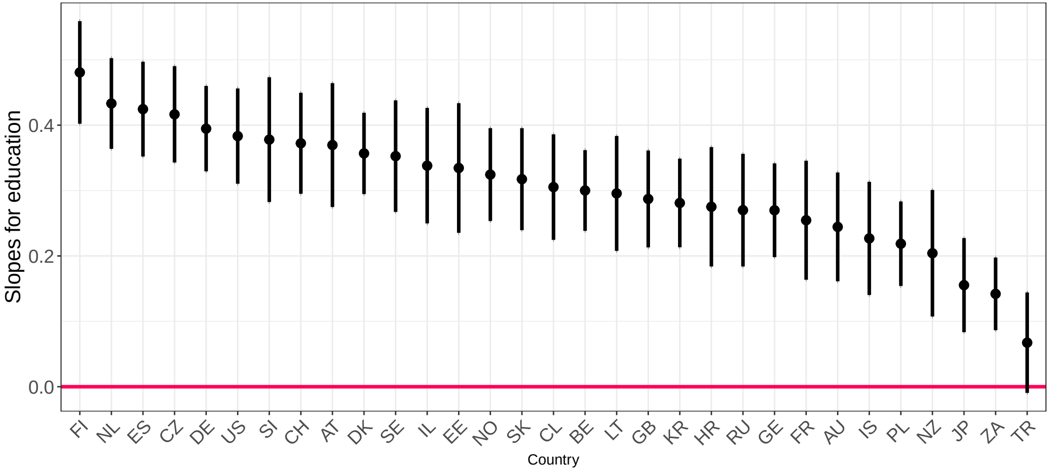


Random effects

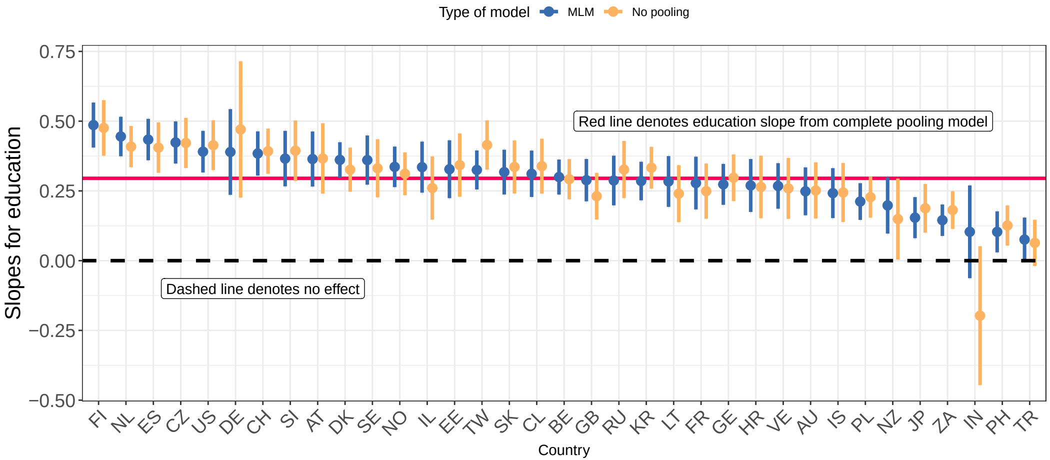


REs are deviations from the fixed-effect for education, not effects themselves!

Actual slopes



Comparison: no pooling & MLM with low N



10% of sample for IN and DE was used for demonstration

Adding predictors for random slopes (1)

So far, we have only added a L2 predictor for the random intercept: $\beta_{0j} = \gamma_{00} + \gamma_{01} CORR_j + \upsilon_{0j}$.

The slope for education was not explained by anything in the previous specification: $\beta_{1j} = \gamma_{10} + \upsilon_{1j}$.

Adding predictors for random slopes (1)

So far, we have only added a L2 predictor for the random intercept: $\beta_{0j} = \gamma_{00} + \gamma_{01} CORR_j + \upsilon_{0j}$.

The slope for education was not explained by anything in the previous specification: $\beta_{1j} = \gamma_{10} + \upsilon_{1j}$.

We can also explain *systematically* why effects vary across contexts, by adding predictors for the slopes as well.

Adding predictors for random slopes (1)

So far, we have only added a L2 predictor for the random intercept: $\beta_{0j} = \gamma_{00} + \gamma_{01} CORR_j + \upsilon_{0j}$.

The slope for education was not explained by anything in the previous specification: $\beta_{1j} = \gamma_{10} + \upsilon_{1j}$.

We can also explain *systematically* why effects vary across contexts, by adding predictors for the slopes as well.

$$\beta_{1j} = \gamma_{10} + \gamma_{11} CORR_j + \upsilon_{1j} \tag{10}$$

Adding predictors for random slopes (1)

A comparison of 3 models:

- ✓ RI model with L2 predictor for intercept (Equation 4)

Adding predictors for random slopes (1)

A comparison of 3 models:

- ✓ RI model with L2 predictor for intercept (Equation 4)
- ✓ RI+RS model with L2 predictor for intercept, but not for slope (Equation 8)

Adding predictors for random slopes (1)

A comparison of 3 models:

- ✓ RI model with L2 predictor for intercept (Equation 4)
- ✓ RI+RS model with L2 predictor for intercept, but not for slope (Equation 8)
- ✓ RI+RS model with L2 predictor for both intercept *and* slope (Equation 6)

Model comparison

	RI	RI + RS with no pred.	RI + RS with pred.
(Intercept)	2.962^{***}	2.962^{***}	2.962^{***}
	(0.045)	(0.046)	(0.045)
Education	0.288^{***}	0.292^{***}	0.292^{***}
	(0.008)	(0.021)	(0.019)
Perceived absence of corruption	0.300^{**}	0.354^{***}	0.300^{**}
	(0.092)	(0.090)	(0.092)
Education * Perceptions			0.111^{**}
			(0.039)
Num. obs.	36486	36486	36486
Num. groups: country	35	35	35
Var: country (Intercept) ($\sigma^2_{\upsilon_{0j}}$)	0.072	0.072	0.072
Var: Residual ($\sigma^2_{\epsilon_{ij}}$)	0.476	0.473	0.473
Var: country Education ($\sigma^2_{\upsilon_{1j}}$)		0.014	0.011
Cov: country (Intercept) Education ($\rho\sigma^2_{\upsilon_{0j}}\sigma^2_{\upsilon_{1j}}$)		-0.008	-0.006

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Other predictors excluded, along with model fit statistics.

Small extensions—practice

As a starting point, use the structure of the RI+RS model with predictors for the slope (Equation 6). Write down 2 models:

- ✓ The specification in RI+RS model with predictors for slope, *plus* `GINI` as predictor for the intercept (M1)

Small extensions—practice

As a starting point, use the structure of the RI+RS model with predictors for the slope (Equation 6). Write down 2 models:

- ✓ The specification in RI+RS model with predictors for slope, *plus* `GINI` as predictor for the intercept (M1)
- ✓ The specification above, *plus* `GINI` as another predictor for the slope of education (in addition to corruption) (M2)

Small extensions—practice

As a starting point, use the structure of the RI+RS model with predictors for the slope (Equation 6). Write down 2 models:

- The specification in RI+RS model with predictors for slope, *plus* `GINI` as predictor for the intercept (M1)
- The specification above, *plus* `GINI` as another predictor for the slope of education (in addition to corruption) (M2)

How many parameters are estimated in each model?[1]

[1] If `GINI` would have been a predictor for the slope of `URB`, how many parameters would have been estimated?

Small extensions—results

	M1		M2	
	β	SE	β	SE
(Intercept)	2.962	$(0.046)^{***}$	2.962	$(0.046)^{***}$
Education	0.292	$(0.019)^{***}$	0.292	$(0.017)^{***}$
Urban residence	0.051	$(0.008)^{***}$	0.051	$(0.008)^{***}$
Perceived absence of corruption	0.273	$(0.107)^{*}$	0.309	$(0.107)^{**}$
Gini (10-point)	−0.057	(0.104)	0.019	(0.107)
Education * Perceptions	0.111	$(0.039)^{**}$	0.050	(0.040)
Education * Gini			−0.123	$(0.039)^{**}$
Var: country (Intercept)		0.075		0.074
Var: country Education		0.011		0.008
Cov: country (Intercept) Education		−0.008		−0.006
Var: Residual		0.473		0.473

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Sample sizes same as for previous specifications. Model fit measures and additional predictors excluded.

Cross-level interactions

Cross-level interactions

Centering helps here, as it reduces the collinearity between main terms and interaction.[2]

[2] Kam and Franzese Jr. (2007) point out that this happens because we are effectively changing what we are estimating, i.e. it's not so much a solution as a re-specification.

Cross-level interactions

Centering helps here, as it reduces the collinearity between main terms and interaction.[2]

Education has been group-mean centered and standardized (2SD). Perceptions of corruption and Gini have been grand-mean centered and standardized (2SD).

[2] Kam and Franzese Jr. (2007) point out that this happens because we are effectively changing what we are estimating, i.e. it's not so much a solution as a re-specification.

Cross-level interactions

Centering helps here, as it reduces the collinearity between main terms and interaction.[2]

Education has been group-mean centered and standardized (2SD). Perceptions of corruption and Gini have been grand-mean centered and standardized (2SD).

Interpreted in the same way as interactions in regular multiple regression (Brambor et al., 2005).

[2] Kam and Franzese Jr. (2007) point out that this happens because we are effectively changing what we are estimating, i.e. it's not so much a solution as a re-specification.

Cross-level interactions

Centering helps here, as it reduces the collinearity between main terms and interaction.[2]

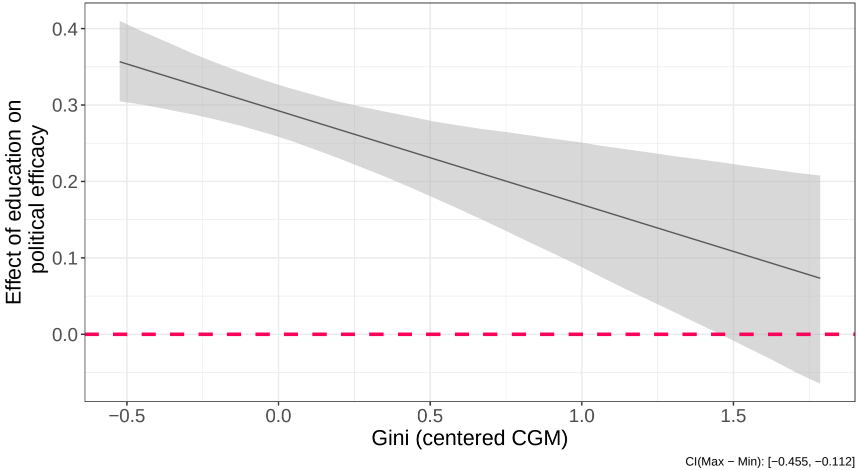
Education has been group-mean centered and standardized (2SD). Perceptions of corruption and Gini have been grand-mean centered and standardized (2SD).

Interpreted in the same way as interactions in regular multiple regression (Brambor et al., 2005).

Always turn to graphs to present cross-level interactions.

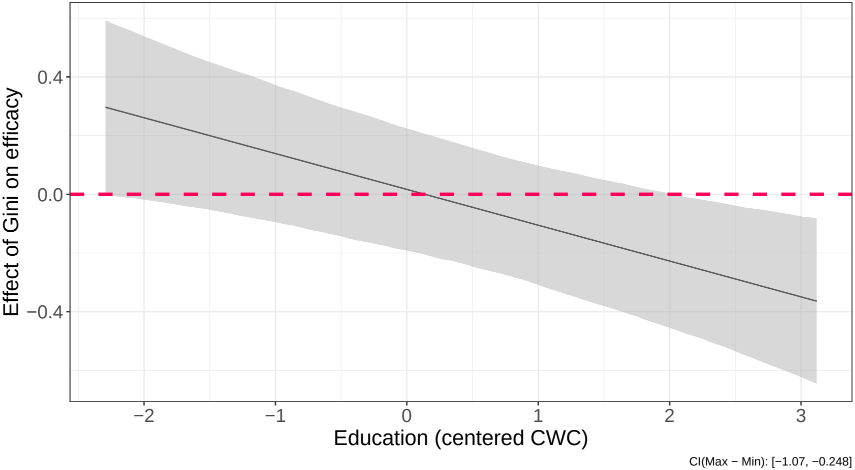
[2] Kam and Franzese Jr. (2007) point out that this happens because we are effectively changing what we are estimating, i.e. it's not so much a solution as a re-specification.

Education × Gini



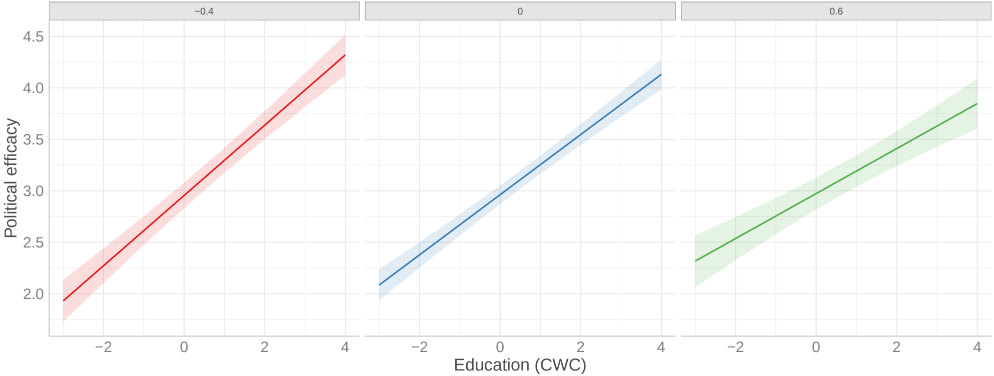
$\beta_{education}$ depicted at varying levels of GINI

Symmetry of interpretation



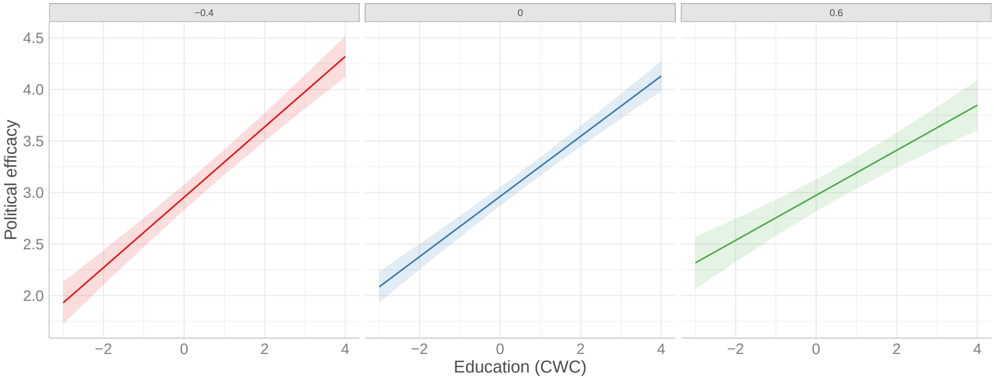
β_{Gini} depicted at varying levels of `EDU`

Use actual predictions of outcome



Panels vary based on different levels of `GINI`

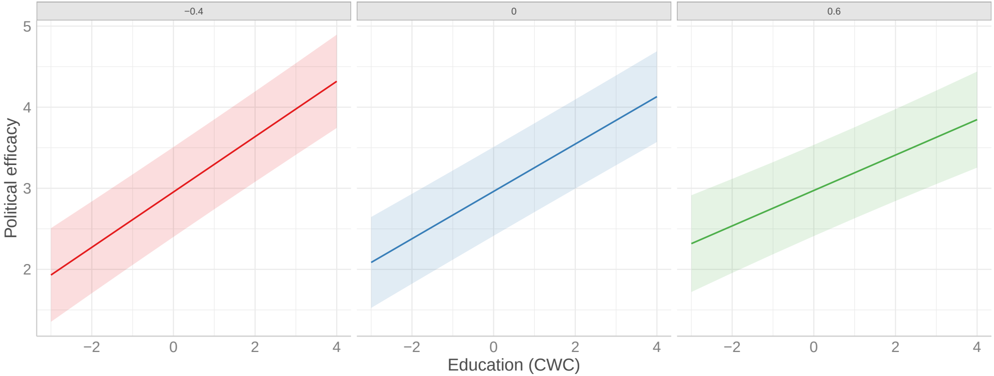
Use actual predictions of outcome



Panels vary based on different levels of `GINI`

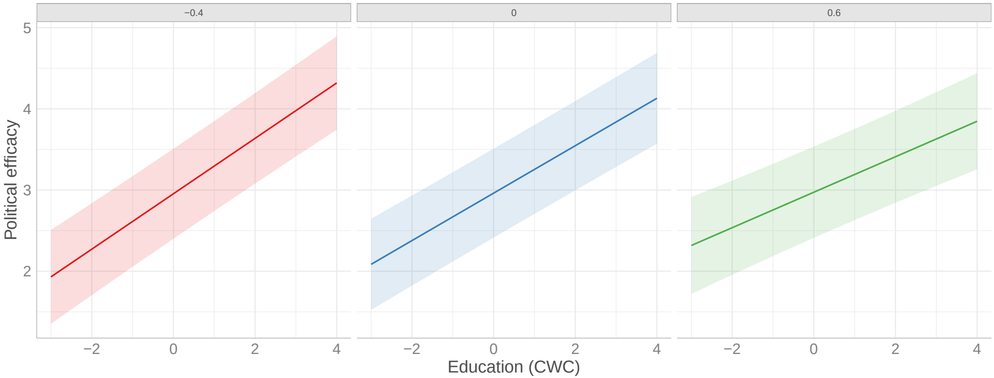
However, this incorporates only fixed-effect uncertainty.

. . . with random-effect uncertainty



Panels vary based on different levels of `GINI`

. . . with random-effect uncertainty



Panels vary based on different levels of `GINI`

Much wider CIs when incorporating both FE & RE uncertainty.

Sample size

Sample size (1)

As we covered briefly yesterday, this is maximum likelihood estimation. Its desirable properties kick in only in large samples.

Sample size (1)

As we covered briefly yesterday, this is maximum likelihood estimation. Its desirable properties kick in only in large samples.

Usually, the concern is with the level 2 sample size. The minimum is something like 30 (Stegmüller, 2013), although a more desirable threshold is 50.

Sample size (1)

As we covered briefly yesterday, this is maximum likelihood estimation. Its desirable properties kick in only in large samples.

Usually, the concern is with the level 2 sample size. The minimum is something like 30 (Stegmüller, 2013), although a more desirable threshold is 50.

Even for low sample sizes ML estimates will be unbiased, but will likely be inefficient.[3]

[3] There is a bit of a difference here between different types of ML-based estimators. REML is generally better than FIML.

L2 sample size

It's fair, though, to take a more nuanced view of things.

Estimate	Continuous DV	Binary DV
L1 fixed-effects point estimates	5	10
L2 fixed-effects point estimates	15	30
Fixed-effects std. errors	30	50
L1 RE estimate	10	30
L2 RE estimate	10/30 (REML/FIML)	10/50 (REML/FIML)
L2 RE std. error	50	100

Level 2 minimum sample size requirements (McNeish & Stapleton, 2016)

L1 group sample size

In *vanilla* cross-national research it is typically not a concern, ranging from a few hundred to a few thousands.

L1 group sample size

In *vanilla* cross-national research it is typically not a concern, ranging from a few hundred to a few thousands.

Strength of MLM: can give an estimate for small groups, borrowing power from larger groups.

L1 group sample size

In *vanilla* cross-national research it is typically not a concern, ranging from a few hundred to a few thousands.

Strength of MLM: can give an estimate for small groups, borrowing power from larger groups.

Interesting case: very many groups, but limited sample size at L1 in each group, e.g. a household survey.

L1 group sample size

In *vanilla* cross-national research it is typically not a concern, ranging from a few hundred to a few thousands.

Strength of MLM: can give an estimate for small groups, borrowing power from larger groups.

Interesting case: very many groups, but limited sample size at L1 in each group, e.g. a household survey.

You're facing a limit on the number of L1 parameters that you can allow to vary. For HH surveys, depending on the context, usually only 1 random slope.

Model fit

Model fit

Unlike OLS-based regression, where we usually look at a single measure of fit (R^2), in multilevel models we have about 4 measures of fit.

Model fit

Unlike OLS-based regression, where we usually look at a single measure of fit (R^2), in multilevel models we have about 4 measures of fit.

All are by-products of ML-estimation, and are also frequently reported in the case of GLMs (you might already be familiar with them).

Model fit

Unlike OLS-based regression, where we usually look at a single measure of fit (R^2), in multilevel models we have about 4 measures of fit.

All are by-products of ML-estimation, and are also frequently reported in the case of GLMs (you might already be familiar with them).

R^2 had a straightforward set of bounds: between 0 and 1, with higher values suggesting that the model fits the data better.[4]

[4] The interpretation of the R^2 is a bit less straightforward: it is the percentage of variance in the DV explained by the IVs.

Model fit

Unlike OLS-based regression, where we usually look at a single measure of fit (R^2), in multilevel models we have about 4 measures of fit.

All are by-products of ML-estimation, and are also frequently reported in the case of GLMs (you might already be familiar with them).

R^2 had a straightforward set of bounds: between 0 and 1, with higher values suggesting that the model fits the data better.[4]

Unfortunately, we don't have a similarly "well-behaved" measure for MLMs.

[4] The interpretation of the R^2 is a bit less straightforward: it is the percentage of variance in the DV explained by the IVs.

Model fit—indices

4 measures of fit:

- ✓ logLikelihood (LL): the logarithm of the likelihood of the model

Model fit—indices

4 measures of fit:

- ✓ logLikelihood (LL): the logarithm of the likelihood of the model
- ✓ deviance: $-2 \times LL$

Model fit—indices

4 measures of fit:

- ✓ logLikelihood (LL): the logarithm of the likelihood of the model
- ✓ deviance: $-2 \times LL$
- ✓ Akaike Information Criterion (AIC): $-2 \times LL + 2 \times k$.[5]

[5] k is the number of parameters estimated by the model.

Model fit—indices

4 measures of fit:

- ✓ logLikelihood (LL): the logarithm of the likelihood of the model
- ✓ deviance: $-2 \times LL$
- ✓ Akaike Information Criterion (AIC): $-2 \times LL + 2 \times k$.[5]
- ✓ Bayesian Information Criterion (BIC): $-2 \times LL + k \times log_e n$.[6]

All are usually provided as part of the estimation output by most software.

[5] k is the number of parameters estimated by the model.
[6] $log_e n$ is the sample size on which the model is estimated.

Model fit—indices

4 measures of fit:

- logLikelihood (LL): the logarithm of the likelihood of the model
- deviance: $-2 \times LL$
- Akaike Information Criterion (AIC): $-2 \times LL + 2 \times k$.[5]
- Bayesian Information Criterion (BIC): $-2 \times LL + k \times log_e n$.[6]

All are usually provided as part of the estimation output by most software.

We will also discuss a version of the R^2 devised for MLM (Snijders & Bosker, 1999).

[5] k is the number of parameters estimated by the model.
[6] $log_e n$ is the sample size on which the model is estimated.

Relative fit

The values for the 4 measures are meaningless in the absolute—a deviance of 110,034.45 doesn't tell you very much.

Relative fit

The values for the 4 measures are meaningless in the absolute—a deviance of 110,034.45 doesn't tell you very much.

For MLM, we can use them to compare two or more models with each other, and determine which is the best fitting model.

Relative fit

The values for the 4 measures are meaningless in the absolute—a deviance of 110,034.45 doesn't tell you very much.

For MLM, we can use them to compare two or more models with each other, and determine which is the best fitting model.

The glitch is that, for the first three measures (LL, deviance, and AIC) we can only compare models which have been estimated on identical samples.

logLikelihood

The *likelihood* is a by-product of the estimation procedure. It's mathematical definition is a bit abstract: it is the product of the density evaluated at the observations.

logLikelihood

The *likelihood* is a by-product of the estimation procedure. It's mathematical definition is a bit abstract: it is the product of the density evaluated at the observations.

Generaly speaking, the higher it is the better the model fits the data.[7]

[7] Will distribute a small document about measures of model fit, which will have more detail.

logLikelihood

The *likelihood* is a by-product of the estimation procedure. It's mathematical definition is a bit abstract: it is the product of the density evaluated at the observations.

Generaly speaking, the higher it is the better the model fits the data.[7]

Typically, <u>but not always</u>, you'll see likelihoods in between 0 and 1.

[7] Will distribute a small document about measures of model fit, which will have more detail.

logLikelihood

The *likelihood* is a by-product of the estimation procedure. It's mathematical definition is a bit abstract: it is the product of the density evaluated at the observations.

Generaly speaking, the higher it is the better the model fits the data.[7]

Typically, but not always, you'll see likelihoods in between 0 and 1.

The logarithm of the likelihood will, in this case, be between $-\infty$ and 0.[8]

[7] Will distribute a small document about measures of model fit, which will have more detail.
[8] $log_e 0 = -\infty$ and $log_e 1 = 0$. This is because $e^0 = 1$ and $e^{-\infty} = 0$.

Deviance

Simple formula: $-2 \times LL$.

It's an indicator of misfit: the higher the number, the worse the model fit.

Deviance

Simple formula: $-2 \times LL$.

It's an indicator of misfit: the higher the number, the worse the model fit.

A loglikelihood of -100 is worse than a loglikelihood of -50, which means that a deviance of 200 (-100×-2) is worse than a deviance of 100.

AIC

The trouble with the *logLikelihood* and the deviance is that they don't take into account how many predictors we have in the model.

AIC

The trouble with the *logLikelihood* and the deviance is that they don't take into account how many predictors we have in the model.

Like with OLS-based R^2 the more predictors we add, the lower the deviance, even if those predictors are not statistically significant.

AIC

The trouble with the *logLikelihood* and the deviance is that they don't take into account how many predictors we have in the model.

Like with OLS-based R^2 the more predictors we add, the lower the deviance, even if those predictors are not statistically significant.

The AIC introduces a penalty for this: $-2 \times LL + 2 \times k$. In this case, the more parameters estimated, the worse the fit (if the value of the deviance is constant).

AIC

The trouble with the *logLikelihood* and the deviance is that they don't take into account how many predictors we have in the model.

Like with OLS-based R^2 the more predictors we add, the lower the deviance, even if those predictors are not statistically significant.

The AIC introduces a penalty for this: $-2 \times LL + 2 \times k$. In this case, the more parameters estimated, the worse the fit (if the value of the deviance is constant).

Can be used for comparisons of non-nested models, but with great care.[9]

[9] See model fit document distributed at the end of the session.

BIC

Compared to the deviance, the BIC implements corrections for both number of parameters estimated, and the sample size on which the model is tested.

BIC

Compared to the deviance, the BIC implements corrections for both number of parameters estimated, and the sample size on which the model is tested.

This allows for comparisons between models tested on different samples, e.g. when adding a variable with missing observations reduces the sample size for model estimation.

BIC

Compared to the deviance, the BIC implements corrections for both number of parameters estimated, and the sample size on which the model is tested.

This allows for comparisons between models tested on different samples, e.g. when adding a variable with missing observations reduces the sample size for model estimation.

In practice, I would suggest engaging in such comparisons with care.[10]

[10] A simple mathematical correction can't cover all empirical configurations of data.

R^2

A measure introduced by Snijders and Bosker (1999), but I'll present formulas similar to those from Luke (2004), because they are slightly simpler.

$$R_1^2 = 1 - \frac{\left(s^2_{\upsilon_{0j}} + s^2_{e_{ij}}\right)_{Model\ 2}}{\left(s^2_{\upsilon_{0j}} + s^2_{e_{ij}}\right)_{Model\ 1}} \tag{11}$$

R^2

A measure introduced by Snijders and Bosker (1999), but I'll present formulas similar to those from Luke (2004), because they are slightly simpler.

$$R_1^2 = 1 - \frac{\left(s_{\upsilon_{0j}}^2 + s_{e_{ij}}^2\right)_{Model\ 2}}{\left(s_{\upsilon_{0j}}^2 + s_{e_{ij}}^2\right)_{Model\ 1}} \tag{11}$$

$$R_2^2 = 1 - \frac{\left(s_{\upsilon_{0j}}^2 + \frac{s_{e_{ij}}^2}{n}\right)_{Model\ 2}}{\left(s_{\upsilon_{0j}}^2 + \frac{s_{e_{ij}}^2}{n}\right)_{Model\ 1}} \tag{12}$$

R^2

In Equation 12, the n is the Level 1 sample size.

R^2

In Equation 12, the n is the Level 1 sample size.

They are great for capturing changes at each level of the hierarchy, which makes them more detailed than other measures presented so far.

R^2

In Equation 12, the n is the Level 1 sample size.

They are great for capturing changes at each level of the hierarchy, which makes them more detailed than other measures presented so far.

In some instances, though, adding a predictor might result in an *larger* variance of residuals, which translates into a *negative* R^2.

Model comparisons

Likelihood ratio test (LRT)

Checking which deviance, AIC, or BIC is lower, to identify the better fitting model, is fairly subjective.

How do we know whether “low” is “low enough”?

Likelihood ratio test (LRT)

Checking which deviance, AIC, or BIC is lower, to identify the better fitting model, is fairly subjective.

How do we know whether "low" is "low enough"?

We can use a likelihood ratio test: $Deviance_{smaller\ model}$ - $Deviance_{larger\ model}$ has a χ^2 distribution, with $k_{larger\ model} - k_{smaller\ model}$ degrees of freedom.

Likelihood ratio test (LRT)

Checking which deviance, AIC, or BIC is lower, to identify the better fitting model, is fairly subjective.

How do we know whether "low" is "low enough"?

We can use a likelihood ratio test: $Deviance_{smaller\ model}$ - $Deviance_{larger\ model}$ has a χ^2 distribution, with $k_{larger\ model} - k_{smaller\ model}$ degrees of freedom.

Important: Use only with FIML estimation (not REML).

Likelihood ratio test (LRT)

Checking which deviance, AIC, or BIC is lower, to identify the better fitting model, is fairly subjective.

How do we know whether "low" is "low enough"?

We can use a likelihood ratio test: $Deviance_{smaller\ model}$ - $Deviance_{larger\ model}$ has a χ^2 distribution, with $k_{larger\ model} - k_{smaller\ model}$ degrees of freedom.

Important: Use only with FIML estimation (not REML).

Important: Use only for nested models.

Nested models

The second model has to have all the variables of the first model, and a few extra ones (at least one more).

Nested models

The second model has to have all the variables of the first model, and a few extra ones (at least one more).

M1: EFF $\Leftarrow$ EDU $+$ URB

M2: EFF $\Leftarrow$ EDU $+$ URB $+$ INC

M3: EFF $\Leftarrow$ EDU $+$ INC

Nested models

The second model has to have all the variables of the first model, and a few extra ones (at least one more).

M1: EFF $\Leftarrow$ EDU + URB

M2: EFF $\Leftarrow$ EDU + URB + INC

M3: EFF $\Leftarrow$ EDU + INC

M1 is nested in M2. M3 is nested in M2. No nesting relationship between M1 and M3.

LRT in practice (1)

I compare a few specifications we saw today:

- ✓ L1 predictors, CPI at L2, and random slope for education;
- ✓ L1 predictors, CPI at L2, and cross-level interaction between education and CPI;

LRT in practice (1)

I compare a few specifications we saw today:

- ✓ L1 predictors, CPI at L2, and random slope for education;
- ✓ L1 predictors, CPI at L2, and cross-level interaction between education and CPI;
- ✓ L1 predictors, CPI and GINI at L2, and cross-level interaction between education and CPI;
- ✓ L1 predictors, CPI and GINI at L2, and cross-level interactions between education and CPI & education and GINI;

LRT in practice (1)

I compare a few specifications we saw today:

- ✓ L1 predictors, CPI at L2, and random slope for education;
- ✓ L1 predictors, CPI at L2, and cross-level interaction between education and CPI;
- ✓ L1 predictors, CPI and GINI at L2, and cross-level interaction between education and CPI;
- ✓ L1 predictors, CPI and GINI at L2, and cross-level interactions between education and CPI & education and GINI;

Because by default these models were estimated with REML, they will have to be re-estimated with FIML first.

LRT in practice (2)

Models	k	AIC	BIC	Deviance	Chisq	Chisq. D.F.	p
M1	12	66982.55	67083.04	66958.55	NA	NA	NA
M2	13	66980.72	67089.59	66954.72	3.825	1	0.050
M3	14	66982.67	67099.90	66954.67	0.059	1	0.809
M4	15	66979.33	67104.94	66949.33	5.332	1	0.021

Model fit comparison table from `anova()` function

LRT in practice (2)

Models	k	AIC	BIC	Deviance	Chisq	Chisq. D.F.	p
M1	12	66982.55	67083.04	66958.55	NA	NA	NA
M2	13	66980.72	67089.59	66954.72	3.825	1	0.050
M3	14	66982.67	67099.90	66954.67	0.059	1	0.809
M4	15	66979.33	67104.94	66949.33	5.332	1	0.021

Model fit comparison table from `anova()` function

$Verdict$: $M4 > M3 \approx M2 > M1$

Data management for MLM

Data management

To wrap up, a “lighter” topic: how to manage data for an MLM analysis.

Data management

To wrap up, a “lighter” topic: how to manage data for an MLM analysis.

If you already have a data set which was built from the ground up with MLM in mind (e.g. CSES, WVS, ESS), half the work is already done.

Data management

To wrap up, a "lighter" topic: how to manage data for an MLM analysis.

If you already have a data set which was built from the ground up with MLM in mind (e.g. CSES, WVS, ESS), half the work is already done.

You'll want to add some more group-level data to it (e.g. income inequality, unemployment rate), which means you'll have to build an additional data set yourself.

Data management (1)

Merging will be done with the `join` family of function from `dplyr`.

```
merged_data <- left_join(l1data, l2data, by = "ID_var")
```

Data management (1)

Merging will be done with the `join` family of function from `dplyr`.

```
merged_data <- left_join(l1data, l2data, by = "ID_var")
```

The two data sets have to have a variable called the same, with the same categories (if we continue with the country example, these categories can be the country names).

Data management (1)

Merging will be done with the `join` family of function from `dplyr`.

```
merged_data <- left_join(l1data, l2data, by = "ID_var")
```

The two data sets have to have a variable called the same, with the same categories (if we continue with the country example, these categories can be the country names).

Biggest danger: `merged_data` ends up having duplicated rows, because there is more than one observation in `l2data` that can be matched with an observation from `l1data`.

Data management (2)

country	X_1	X_2
Albania	1	10000
Algeria	1	8000
Belgium	0	22000

Data 1

country	X_3	X_4
Albania	25	1
Algeria	40	0
Argentina	35	5
Belgium	120	9
Bulgaria	25	6

Data 2

Data management (2)

country	X_1	X_2
Albania	1	10000
Algeria	1	8000
Belgium	0	22000

Data 1

country	X_3	X_4
Albania	25	1
Algeria	40	0
Argentina	35	5
Belgium	120	9
Bulgaria	25	6

Data 2

Here, the `country` variable is the same, and we can merge the data without a problem, since all the categories in the first data set are also present in the second (Albania, Algeria, Belgium).[11]

[11] Had this not been the case, the merging procedure wouldn't have worked.

Data management (3)

If, however, you have to construct the data by yourselves, a good practice is to pay very close attention to the ID variable (in the past example, this was `country`).

Data management (3)

If, however, you have to construct the data by yourselves, a good practice is to pay very close attention to the ID variable (in the past example, this was `country`).

As the data sets become more and more complex (pupils in classrooms, in schools, in countries), you have to make sure this variable is as clear and clean as possible.

Data management (3)

If, however, you have to construct the data by yourselves, a good practice is to pay very close attention to the ID variable (in the past example, this was `country`).

As the data sets become more and more complex (pupils in classrooms, in schools, in countries), you have to make sure this variable is as clear and clean as possible.

Also, try to keep the level 1 and level 2 data sets separate, up until the actual moment of running the analyses—it makes it easier to graphically examine the variables.

Wide format

Religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k
Agnostic	27	34	60	81	76	137	122
Atheist	12	27	37	52	35	70	73
Buddhist	27	21	30	34	33	58	62
Catholic	418	617	732	670	638	1116	949
Evangelical Protestant	575	869	1064	982	881	1486	949
Hindu	1	9	7	9	11	34	47
Hist. Black Protestant	228	244	236	238	197	223	131
Jehovah's Witness	20	27	24	24	21	30	15
Jewish	19	19	25	25	30	95	69
DK/ref	15	14	15	11	10	35	21

Easy to look at breakdowns and examine associations, but not easy to feed into R's functions.

Long format

Religion	Income	Frequency
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Tidy data: (1) every observation is a row; (2) every variable is a column; (3) every data set contains a single type of observation.

dplyr & tidyr

Worth investing a lot of time in mastering these 2 packages:

- ✓ `mutate`: creating new variables;
- ✓ `rename`: renaming variables;
- ✓ `case_when`: recoding;
- ✓ `left_join`: data merging (only one of the functions in the family);
- ✓ `filter`: data subsetting;
- ✓ `select`: selecting columns (or excluding them).

`pivot_wider()` and `pivot_longer()` from `tidyr` do data reshaping in a very flexible way.

Thank you for the kind attention!

References I

Brambor, T., Clark, W. R., & Golder, M. (2005). Understanding Interaction Models: Improving Empirical Analyses. *Political Analysis*, *14*(1), 63–82.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138.

Kam, C. D., & Franzese Jr., R. J. (2007). *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor, MI: University of Michigan Press.

Luke, D. A. (2004). *Multilevel Modeling*. Thousand Oaks, CA: Sage Publications.

McNeish, D. M., & Stapleton, L. M. (2016). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychology Review*, *28*(2), 295–314.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Stegmüller, D. (2013). How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches. *American Journal of Political Science*, *57*(3), 748–761.