Multilevel Modeling: Principles and Applications in R

Day 1: The Basics

Constantin Manuel Bosancianu

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

January 20, 2021

Welcome! It's great to have you in the workshop!

Workshop structure (1)

First two days showcase fundamental features of multilevel models (MLMs):

1. Ability to model (explain) variation in data at multiple levels
2. Ability to explain variation between groups in *estimates*
3. Ability to estimate effects for groups not encountered in our data

We go through: (1) estimation; (2) interpretation; (3) model assessment and checking; and (4) graphical display of quantities of interest and predictions based on the models.

Workshop structure (2)

The last day applies these lessons to a specific data configuration: longitudinal data.

Multiple observations over time for a large set of units ($N \gg T$).

Most of the lessons from the previous two days hold for MLMs applied to longitudinal data, so we focus on key differences.

Logistics

Lecture based on .pdf and pre-recorded video. Labs based on .Rdata & .R, and carried out "live" (via Zoom).

Over the 3 workshop days, time spent on lectures gradually decreases in favor of labs:

- ✓ D1: $\approx$ 100 min lecture & 100 min lab
- ✓ D2: $\approx$ 80 min lecture & 120 min lab
- ✓ D3: $\approx$ 60 min lecture & 140 min lab

Small part of the lab also devoted to questions about readings or video lectures.

Why MLM?

Value of MLM

MLMs are uniquely suited to capturing one type of social complexity: the way individuals/firms/NGOs act or think may be context-dependent.

An example which I focus on are the cross-country differences in the likelihood that lower-income people participate in politics.

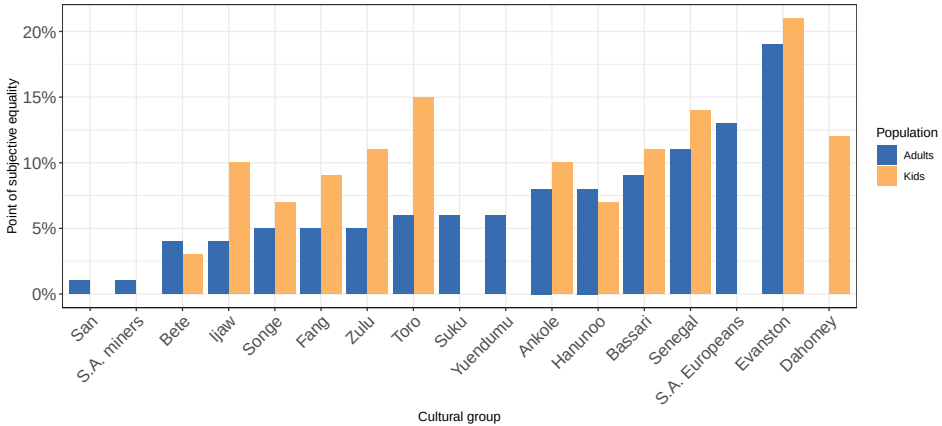
Many similar examples related to educational research, e.g. differences between schools in how much progress students make over a 4-year cycle.

Müller-Lyer illusion

Which one is longer?

We've known about this illusion since 1889. However, since 1966 we also know that not all cultures experience this in the same way (Segall, Campbell, & Herskovits, 1966).

Cross-cultural variance (1)



Adapted from McCauley and Henrich (2006)

Cross-national variance (2)

	Micro	Macro
Turnout	education	compulsory voting
	income	party polarization
Trust	education	post-communist country
Religiosity	age	income inequality
	gender	GDP

Trying to see the world like this trains your mind: how individual actions shape context, and how context, in turn, shapes individual action.

Reasons for using MLMs

Substantive: systematically account for how outcomes (or *effects*) vary across groups, beyond what can be explained by unit-level factors.

Statistical:

- ✓ obtain accurate SEs for estimates in instances of clustered data;
- ✓ *model* the heteroskedasticity (unobserved heterogeneity) in the data.

There is the *nuisance* element to deal with, but it's the second statistical reason that's most important.

Quick OLS recap

OLS mechanics

$$Y_i = \beta_0 + \beta_1 * X1_i + \cdots + \beta_k * Xk_i + \epsilon_i,\ \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

Here, Y is the dependent variable, $X1$ through Xk are the independent variables (IVs), and ϵ is the residual (error).

These ϵ_1, ϵ_2, $\ldots \epsilon_n$ have, collectively, a normal distribution with mean 0 and constant variance (a key Gauss-Markov assumption).

A quick example

	DV: Political efficacy
(Intercept)	2.155 (0.106)***
Age (decades)	0.019 (0.011)
Gender: woman	−0.189 (0.034)***
Education (no. of years)	0.046 (0.006)***
Income: 2nd quartile	0.021 (0.049)
Income: 3rd quartile	0.071 (0.049)
Income: 4th quartile	0.274 (0.052)***
Residence: urban	0.154 (0.035)***
R^2	0.105
Adj. R^2	0.101
Num. obs.	1683
RMSE	0.703

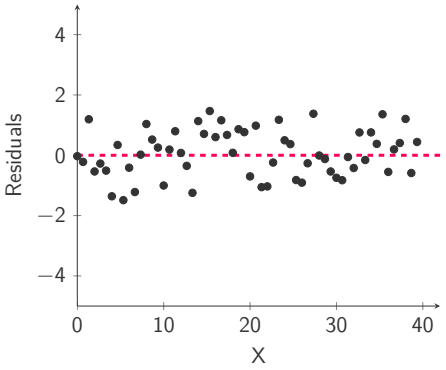
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

ISSP data, Citizenship module II, Belgium (2016).

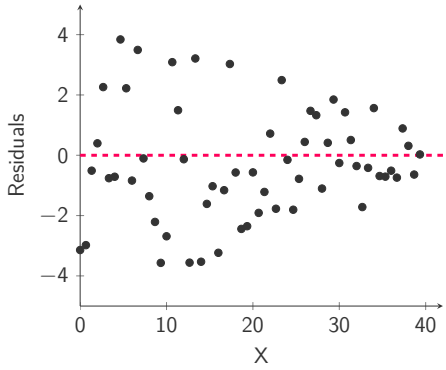
Political efficacy is an index constructed by averaging 4 attitudinal items, each measured on a 5-point scale.

One OLS assumption

Homoskedasticity: $\epsilon \sim \mathcal{N}(0, \sigma^2)$.[1]



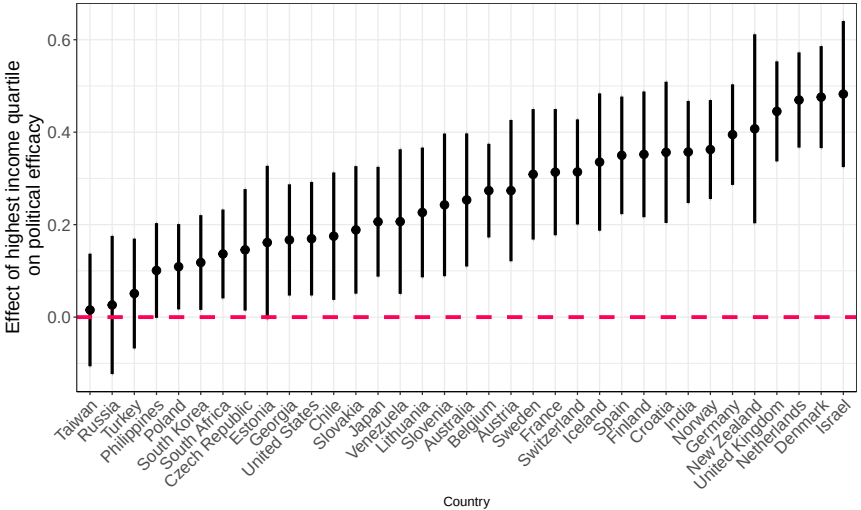
(a) Homoskedasticity



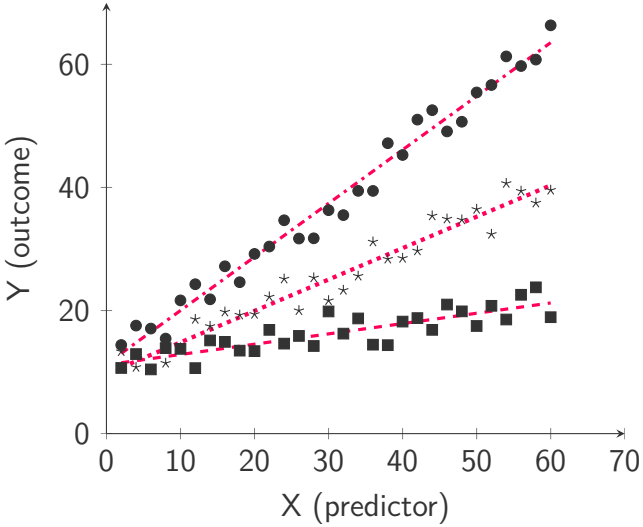
(b) Heteroskedasticity

[1] For an in-depth coverage, please consult the relevant sections in Berry (1993) or Fox (2016).

The case of clustered data



Consequences of heterogeneity



In this instance, applying an overall slope ("naive pooling") to the data will generate heteroskedasticity.

This can be addressed with country dummies (*fixed effects*), and a lot of interactions.

Addressing heteroskedasticity

Fixed Effects

Solution: $J - 1$ dummy indicators for groups (LSDV approach). Computationally very fast, and conceptually accessible.

Problems with the strategy:

- ✓ Very cumbersome with large number of groups
- ✓ Cannot explain *why* slopes vary between groups

FEs are a very powerful solution, but sometimes at the cost of obscuring dynamics at play.

Cluster-corrected SEs

They implement a *post hoc* solution: adjust SEs (biased), while leaving $\hat{\beta}$ alone (*supposedly* unbiased).

If the heteroskedasticity is caused by effect heterogeneity, itself caused by a L2 dynamic at play $\Rightarrow$ model specification itself is incorrect (Freedman, 2006).

If this is the case $\hat{\beta}$s are incorrect (see also King & Roberts, 2015) and the Huber–White estimator is not very helpful.

MLMs: benefits

Multiple benefits to using an MLM, instead of the previous two strategies (Steenbergen & Jones, 2002):

- ✓ Combine multiple levels of analysis in a *single* specification (addresses misspecification concerns)
- ✓ Explore and *model* causal heterogeneity, using substantive variables, instead of treating it as a nuisance
- ✓ Gain the ability to make predictions for new *contexts*

MLMs: costs

These benefits come with costs (as always):

- ✓ Increased computational complexity (ML-based estimation)
- ✓ More stringent assumptions (operating at each level of the data)
- ✓ *Theoretically* more demanding, given potential linkages between predictors at different levels

Introducing MLMs

MLM as compromise solution

MLMs: compromise between *no pooling* and *complete pooling* approach.

No pooling: run regressions group-by-group and present estimates of interest.

Complete pooling: ignore group membership completely and run a single model on the entire sample.

Each approach comes with weaknesses, but MLMs manage to partly overcome these weaknesses.

Political efficacy example

ISSP *Citizenship II* module, data collected between 2014 and 2016: 35 countries and 48120 valid observations on political efficacy, education, and urban residence.

Outcome—political efficacy:

- ✓ "People like me. . . no say in what the government does"
- ✓ "Do not think government cares about much what people like me think"
- ✓ "Have a pretty good understanding of the important political issues"
- ✓ "Most people are better informed about politics [. . .] than I am"

Each item measured on 5-point Likert scale; final index is an average of the four (or fewer) items.

Complete pooling (1)

Predictors:

- ✓ education: number of years of full-time education (0–35)
- ✓ urban residence: dichotomous (1=big city, or suburbs of big city)

Complete pooling: fit model on the entire sample, without taking into account group membership.

Implication: information from the entire sample is used in estimating parameters.

Complete pooling (2)

	DV: Political efficacy
(Intercept)	2.315 (0.011)***
Education (years)	0.047 (0.001)***
Urban residence	0.083 (0.007)***
R^2	0.066
Adj. R^2	0.066
Num. obs.	48120
RMSE	0.749

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

No pooling

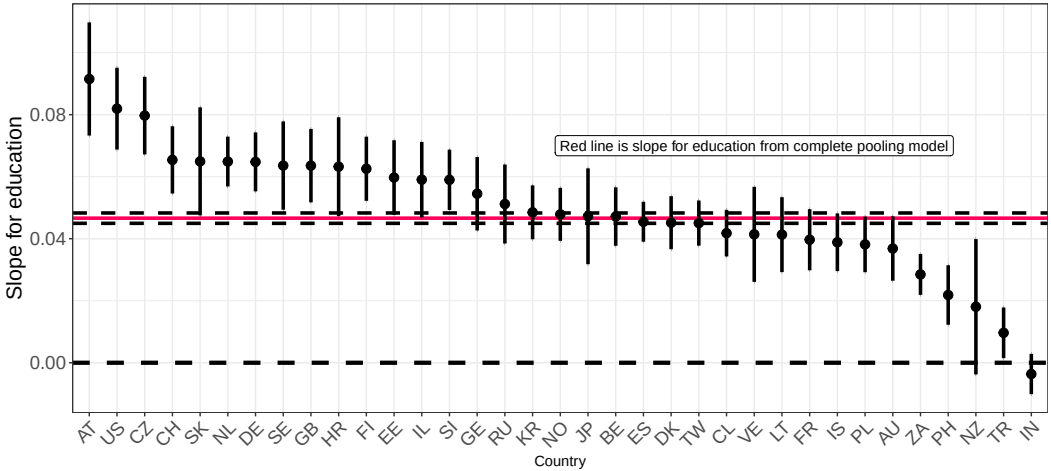
No pooling: fit same model separately for each group.[2]

Implication: the estimate for group j is obtained using information from just that group.

Great for understanding how an effect varies across contexts, and perhaps get clues as to *why*.

[2] A variant of this uses country dummies in the model.

Slopes for education



Why the difference in CIs between, say, New Zealand and South Africa?

No vs. *Complete*

Complete pooling eliminates variation in estimates between groups, but minimizes uncertainty.

No pooling maximizes variation in estimates between groups, but results in maximum uncertainty.

Depending on group size, *no pooling* might also make groups seem more different from each other than they really are (Gelman & Hill, 2007).

Multilevel estimator (1)

$$\hat{\alpha_j}^{MLM} \approx \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\bar{y}_{all}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \tag{2}$$

- ✓ $\bar{y}_j$: unpooled estimate
- ✓ $\bar{y}_{all}$: completely pooled estimate
- ✓ n_j: sample size in group j
- ✓ σ_y^2: within-group variance in outcome
- ✓ σ_α^2: variance in average level of outcome between groups

Multilevel estimator (2)

$$\hat{\alpha_j}^{MLM} \approx \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\bar{y}_{all}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \tag{3}$$

Smaller $n_j \Rightarrow$ MLM estimate for group j is pulled closer to $\bar{y}_{all}$.

Greater homogeneity of groups $\Rightarrow$ greater differences in means between groups $\Rightarrow \sigma_y^2$ is lower and σ_α^2 is higher $\Rightarrow$ MLM estimate is pulled closer to $\bar{y}_j$.

MLM notation

From OLS to MLM (1)

$$EFF_i = \beta_0 + \beta_1 * EDU_i + \beta_2 * URB_i + \epsilon_i, \text{ with } \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{4}$$

MLM extension is not that different.

$$EFF_{ij} = \beta_{0j} + \beta_{1j} * EDU_{ij} + \beta_{2j} * URB_{ij} + \epsilon_{ij} \tag{5}$$

i indexes level-1 units, while j indexes level-2 groups.

The equation denotes that the intercepts and slopes for each of the j groups are getting their own statistical specification.

From OLS to MLM (2)

For now, we only want a meaningful specification for the intercept.

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * CORR_j + v_{0j} \tag{6}$$

Interpretation: the level of political efficacy in a country (controlling for level-1 factors) is associated with the level of (perceived) corruption in the country.

From OLS to MLM (3)

Though we could write a similar model for either, or both, of the level-1 slopes, let's keep them fixed for now.

$$\begin{cases} \beta_{1j} = \gamma_{10} \\ \beta_{2j} = \gamma_{20} \end{cases} \tag{7}$$

The implication: the effect of education on the level of political efficacy, and of urban residence on efficacy, *is identical* for each of the J groups.

From OLS to MLM (4)

Taken together, we have 4 equations, at 2 levels.

$$\begin{cases} EFF_{ij} = \beta_{0j} + \beta_{1j} * EDU_{ij} + \beta_{2j} * URB_{ij} + \epsilon_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01} * CORR_j + v_{0j} \\ \beta_{1j} = \gamma_{10} \\ \beta_{2j} = \gamma_{20} \end{cases} \tag{8}$$

Subscripts vary depending on whether the variable is measured at level-1 (`URB` or `EDU`) or at level-2 (`CORR`).

MLM specification—extended form

Extended form obtained by plugging in last 3 equations into the first one.

$$EFF_{ij} = \gamma_{00} + \gamma_{01} * CORR_j + v_{0j} + \gamma_{10} * EDU_{ij} + \gamma_{20} * URB_{ij} + \epsilon_{ij} = \\ = \gamma_{00} + \gamma_{10} * EDU_{ij} + \gamma_{20} * URB_{ij} + \gamma_{01} * CORR_j + v_{0j} + \epsilon_{ij} \tag{9}$$

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

Random effects: v_{0j}, and ϵ_{ij}.

The models which incorporate both began to be known as *mixed-effects models*.[3]

[3]The distinction between *fixed* and *random* comes from the experimental design literature. Gelman and Hill (2007) reject the name, which they consider confusing.

Fixed vs. random

Fixed-effects are interpreted in exactly the same way as coefficients in regression.

Random-effects, though, despite their name, are not interpreted as effects. They are, rather, variances of residuals.

The latter are useful to report—to the extent that they gradually become smaller, they indicate improvements in model fit.[4]

[4] A popular measure of model fit for multilevel models, called R^2, is based on these random-effects (Snijders & Bosker, 1999).

Practical example

First model (1)

```
model1 <- lmer(formula = poleff ~ 1 + educCWC + urbanCWC + cpiCGM + (1 | cnt),
               data = df_sub,
               REML = TRUE)
```

The formula is similar to OLS specification—L1 & L2 predictors mixed together, as in the extended form.

`(1 | cnt)`: random effects part. `1` means only a random intercept, varying between `cnt`.

REML = restricted maximum likelihood (type of estimation)

First model (2)

	DV: Political efficacy
(Intercept) (γ_{00})	2.952^{***}
	(0.044)
Education (no. of years) (γ_{10})	0.345^{***}
	(0.006)
Urban residence (γ_{20})	0.048^{***}
	(0.007)
Perceived absence of corruption (γ_{01})	0.272^{*}
	(0.089)
Num. obs.	48120
Num. groups: country	35
Var: country (Intercept) (υ_{0j})	0.067
Var: Residual (ϵ_{ij})	0.492

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

ICC

Are MLMs needed *always*?

Truthful answer is ***no***.

Go back to the formula for the estimator (Equation 2): The more heterogeneous the groups, the estimate is pulled toward $\bar{y}_{all}$, making a complete pooling model more attractive.

In extreme situations (groups are completely homogenenous inside, or are completely identical to each other), there's no need to invest in a MLM.

How can we tell when an MLM will be useful?

Clustering

We need a measure of how much clustering there is in a two-stage sample. This ties into the issue of *design effect* (Snijders & Bosker, 1999).

$$SE = \frac{SD}{\sqrt{n}} \tag{10}$$

Design effect: by how much do we have to adjust n to correct for the lack of independence among observations?

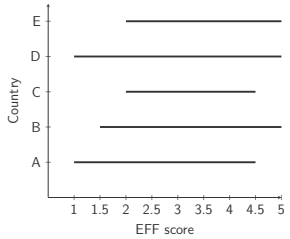
$$DE = 1 + (n - 1)\rho \tag{11}$$

Formula above is valid for equal group sizes. The more homogeneous the groups (ρ is higher), the design effect is larger $\Rightarrow$ the effective sample size is smaller!

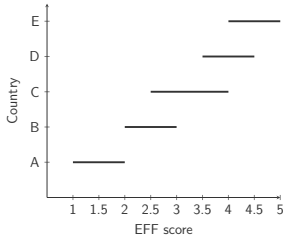
Clustering

The ICC (intraclass correlation coefficient): $\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_y^2}$. In this formula, $\sigma_{total}^2 = \sigma_\alpha^2 + \sigma_y^2$ (between-group variance and within-group variance).

Denotes the share of total variance that is between groups (can also be expressed as a correlation coefficient of individuals *within* the same group).



(a) Variation mainly within



(b) Variation mainly between

Calculating ICC

Derived based on a *null* model (with no predictors):

$$\begin{cases} EFF_{ij} = \beta_{0j} + e_{ij} \\ \beta_{0j} = \gamma_{00} + v_{0j} \end{cases} \tag{12}$$

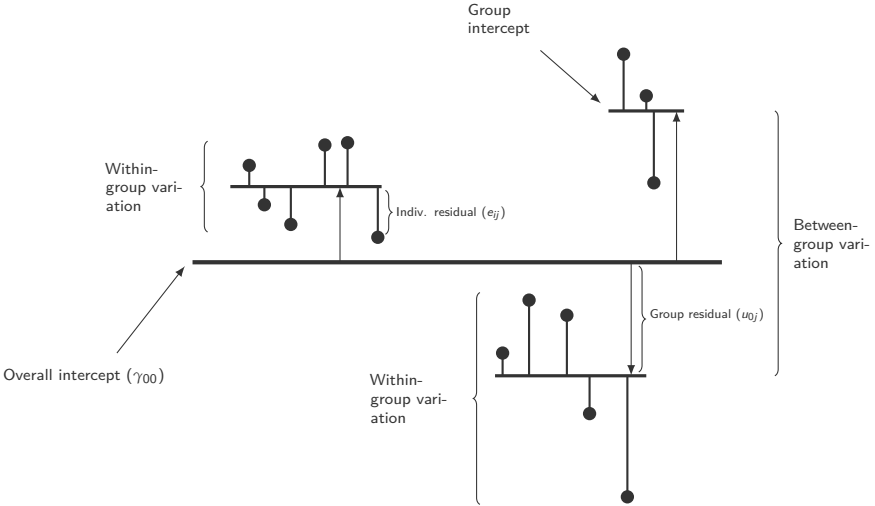
$$EFF_{ij} = \gamma_{00} + v_{0j} + e_{ij} \tag{13}$$

Directly from the model output in R you can compute:

$$ICC = \frac{\sigma^2_{v_{0j}}}{\sigma^2_{v_{0j}} + \sigma {\epsilon_{ij}}^2} \tag{14}$$

This is the way Luke (2004) introduces the ICC.

Visualizing the 3 elements



Adapted from Merlo et al. (2005).

Rules of thumb?

Not clear whether there are any, apart from those dictated by common sense.

With a model with very low ICC (e.g., below 0.05), it will be hard to find any group-level predictor that is statistically significant.

In the literature you can encounter mention of minimum values of 0.1–0.15 for the ICC value.

ICC for our example (1)

```
model2 <- lmer(formula = poleff ~ 1 + (1 | cnt),
               data = df_null,
               REML = TRUE)
```

Null model is a specification that contains no substantive predictors (see Equation 13).

It specifies only a random intercept: `(1 | cnt)`.

ICC for our example (2)

	DV: Political efficacy
(Intercept) (γ_{00})	2.94374***
	(0.04775)
Var: country (Intercept) ($\sigma^2_{v_{0j}}$)	0.08167
Var: Residual ($\sigma^2_{\epsilon_{ij}}$)	0.53158
Num. obs.	50954
Num. groups: countries	36

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

ICC for our example: $\frac{0.08167}{0.08167+0.53158} \approx 13.3\%$ of variance in outcome is at level-2 (between countries).

Estimation

Estimation

For fixed effects: (1) least squares, (2) maximum likelihood, or (3) Bayesian.

For variance components: (1) maximum likelihood, or (2) Bayesian.

In practice, maximum likelihood (ML) is the default estimation method (in the `lme4` package), but it comes at a price of increased estimation time.

Lewis and Linzer (2005) outline a strategy, *only for data where clusters have large sample sizes*, to use least squares methods for estimation.[5]

[5] A combination of OLS and FGLS (*feasible generalized least squares*).

Maximum likelihood (1)

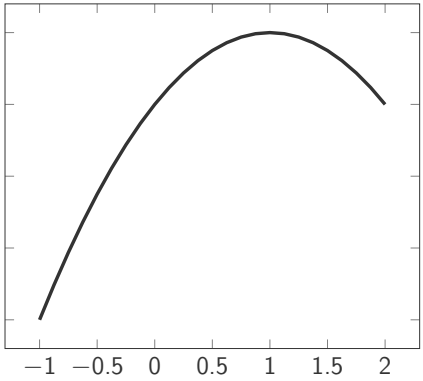
Unfortunately, we have to discuss this topic for a little bit as well.

Maximum likelihood does the estimation in a different way than OLS—it tries to find the coefficients which maximize the probability (likelihood) that we would get the data we observe.

This is not a simple calculation like with OLS, but a iterative procedure: update coefficients $\Rightarrow$ check if they're better $\Rightarrow$ update $\Rightarrow$ check again ...

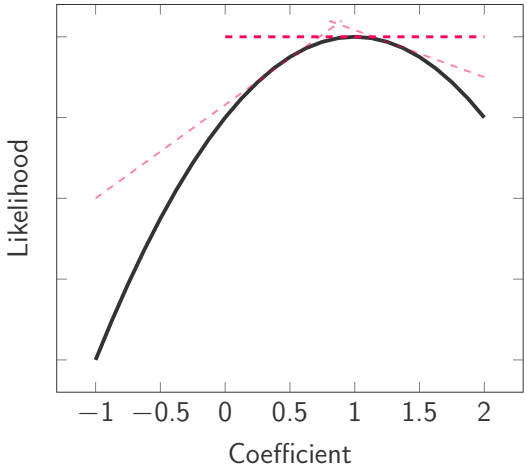
Maximum likelihood (2)

Convergence of the algorithm is reached when the change in coefficients is extremely small.



Maximum likelihood (3)

Setting the first derivative to the likelihood function to 0 gives you the coefficients.



Maximum likelihood (4)

The second derivative to the likelihood function is used to determine if we found the minimum or maximum.[6]

This second derivative is also used to determine the curvature of the likelihood function.

The steeper the curve, the lower the standard errors.

[6] If it's negative, we found the maximum; if it's positive, it's a minimum.

Maximum likelihood (5)

The mathematical details are not very important right now. What you should care about is that the estimation takes much, much longer than OLS.

There are 2 "flavors" of ML:

- ✓ Full Information ML (FIML)
- ✓ Restricted ML (REML)

Second one is default in `R`, partly due to a desirable property: it avoids severe bias of variance components when we have low sample sizes at L2.

Maximum likelihood (6)

FIML includes both fixed- and random-effects in the likelihood function to be maximized $\Rightarrow$ it produces unbiased estimates of fixed-effects.

REML includes only the random-effects in the first stage. The fixed-effects are estimated in a second stage $\Rightarrow$ it produces unbiased estimates of random-effects.

In practice, you would mostly use REML (the bias to the fixed effects is very small).

Alternatives to ML

In the past, GLS (generalized least squares) was also used—its main benefit is that it's much faster than ML.

However, we've since discovered that variance components from GLS are imprecise, and that the coefficients tend to be biased.

A second avenue is Bayesian estimation. Many valid reasons to use it, even when considering the results of Elff, Heisig, Schaeffer, and Shikano (2020), but comes with a steeper learning curve.

Thank you for the kind attention!

References I

Berry, W. D. (1993). *Understanding Regression Assumptions*. Thousand Oaks, CA: Sage Publications.

Elff, M., Heisig, J. P., Schaeffer, M., & Shikano, S. (2020). Multilevel Analysis with Few Clusters: Improving Likelihood-Based Methods to Provide Unbiased Estimates and Accurate Inference. *British Journal of Political Science*(forthcoming), 1–15. doi: 10.1017/S0007123419000097

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Freedman, D. A. (2006). On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, *60*(4), 299–302.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

King, G., & Roberts, M. E. (2015). How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It. *Political Analysis*, *23*(2), 159–179.

References II

Lewis, J. B., & Linzer, D. A. (2005). Estimating Regression Models in Which the Dependent Variable Is Based on Estimates. *Political Analysis*, *13*(4), 345–364.

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

McCauley, R. N., & Henrich, J. (2006). Susceptibility to the Müller-Lyer Illusion, Theory-Neutral Observation, and the Diachronic Penetrability of the Visual Input System. *Philosophical Psychology*, *19*(1), 1–23.

Merlo, J., Chaix, B., Yang, M., Lynch, J., & Råstam, L. (2005). A Brief Conceptual Tutorial of Multilevel Analysis in Social Epidemiology: Linking the Statistical Concept of Clustering to the Idea of Contextual Phenomenon. *Journal of Epidemiology and Community Health*, *59*(6), 443-449.

Segall, M. H., Campbell, D. T., & Herskovits, M. J. (1966). *The Influence of Culture on Visual Perception*. Indianapolis, IN: Bobbs-Merrill Company.

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

Steenbergen, M. R., & Jones, B. S. (2002). Modeling Multilevel Data Structures. *American Journal of Political Science*, *46*(1), 218–237.