# Applied Multilevel Regression Modeling

## Day 4: Model fit & Diagnostics

Constantin Manuel Bosancianu

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

August 1, 2019

# Yesterday

With notation, things become complicated pretty quickly when we allow L1 slopes to vary. Nonetheless, the subscript system has a logic!

Writing a model for slopes introduces cross-level interactions—they allow you to *explain* how a L1 effect varies across L2 groups.

Strive to present results from these interactions in a graphical format and, as much as possible, using quantities of interest rather than raw estimates.

Sample size matters. For cross-national research, it's typically L2 sample size, though for other designs it's the L1 one.

# Today

Estimation: Maximum Likelihood.

Model fit indices in MLM: *logLikelihood*, AIC, BIC, $R^2$. Their use in making model comparisons.

Model diagnostics at multiple levels of the hierarchy.

Presenting result in MLM.

# Estimation

# Estimation

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

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

In practice, maximum likelihood is the default estimation method, though it comes at a price of increased estimation time.

Lewis and Linzer (2005) outline a strategy, *only for clusters with large sample sizes*, to use least squares methods for estimation.

# Maximum likelihood

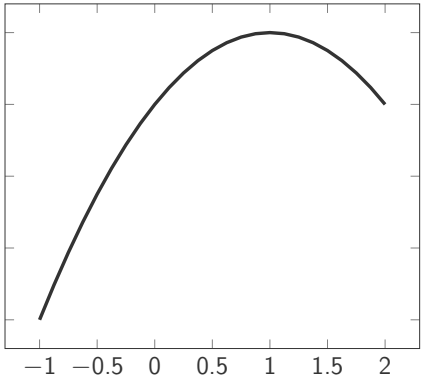
Unfortunately, we have to discuss this topic for a little bit as well.

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 ...

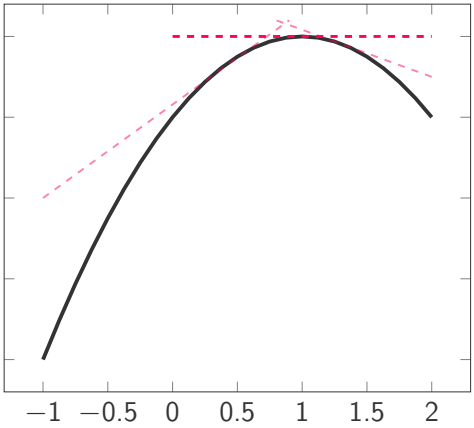
# Maximum likelihood

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



# Maximum likelihood

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



# Maximum likelihood

The second derivative to the likelihood function is used to determine if we found the minimum or maximum.[1]

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

The steeper the curve, the lower the standard errors.

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

# Maximum likelihood

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 its desirable properties in avoiding severe bias of variance components in situations of low sample sizes at L2.

# Maximum likelihood

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.

# 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 (2018), but comes with a steeper learning curve.

# 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.

All are by-products of ML-estimation, and are frequently reported in the case of GLMs.

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

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

[2]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.
- ✓ deviance: $-2 \times LL$.
- ✓ Akaike Information Criterion (AIC): $-2 \times LL + 2 \times k$.[3]
- ✓ Bayesian Information Criterion (BIC): $-2 \times LL + k \times log_e n$.[4]

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).

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

[4] $log_e n$ 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.

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.

Generaly speaking, the higher it is the better the model fits the data.[5]

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.[6]

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

[6] $log_e 0 = -\infty$ 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.

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

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.[7]

[7] 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.

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.[8]

[8] 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.

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

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

# $R^2$

In Equation 2, 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* $R^2$.

# 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 $\chi^2$ 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).

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

I test a few specifications from yesterday:

- ✓ 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

| Models | D.F. | AIC | BIC | Deviance | Chisq | Chisq. D.F. | p |
|---|---|---|---|---|---|---|---|
| M1 | 12 | 66987.36 | 67087.85 | 66963.36 | | | |
| M2 | 13 | 66985.51 | 67094.37 | 66959.51 | 3.8541 | 1 | 0.0496* |
| M3 | 14 | 66987.45 | 67104.69 | 66959.45 | 0.0573 | 1 | 0.8107 |
| M4 | 15 | 66984.08 | 67109.70 | 66954.08 | 5.3660 | 1 | 0.0205* |

Model fit comparison table from `anova()` function

$M4 > M3 \approx M2 > M1$

# Model assumptions and diagnostics

# Model assumptions (1)

Our brief intro on Monday briefly listed a set of assumption for OLS, grouped in a "stochastic" and "systematic" part.

In the MLM framework they are roughly the same, with 3 added complications (assuming a 2-level structure):

- ✓ We now have *two* sets of predictors and two sets of residuals (a L1 and L2 specification and error structure);
- ✓ The L2 error structure can be comprised by multiple sets of residuals;
- ✓ Violations of assumptions at a specific level can *spill over* into the other level's estimation process.

# Model assumptions (2)

I assume that standard exploratory investigations (histograms, scatterplots, outlier detection) on the predictors have been performed; they will not be covered here.

I assume a very simple structure for the model:

$$\begin{cases} EFF_{ij} = \beta_{0j} + \beta_{1j} * EDU_{ij} + \epsilon_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01} * CORR_j + v_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} * CORR_j + v_{1j} \end{cases} \tag{3}$$

For this particular phenomenon, political efficacy, "very simple" likely also means "misspecified".

# Assumptions: stochastic

Referring to the errors at the two levels of the model:

1. $\epsilon_i$ are normally distributed with mean 0 and constant variance $\sigma^2_\epsilon$ within each $j$ unit;
2. $v_{0j}$ and $v_{1j}$ are multivariate normally distributed (each with mean 0 and variances $\sigma^2_{v_0}$ and $\sigma^2_{v_1}$);
3. Errors are independent of each other: $Cov(\epsilon_{ij}, v_{0j}) = 0$ and $Cov(\epsilon_{ij}, v_{1j}) = 0$.

In the first assumption, by implication, it is also the case that overall $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\epsilon)$.

# Assumptions: stochastic & systematic

1. $Cov(EDU_{ij}, \epsilon_{ij}) = 0$ (in the case of multiple predictors, each covariance has to be 0);
2. $Cov(CORR_j, v_{0j}) = 0$ and $Cov(CORR_j, v_{1j}) = 0$ (in the case of multiple predictors, the covariance between each and each random effect has to be 0);
3. $Cov(EDU_{ij}, v_{0j}) = 0$ and $Cov(EDU_{ij}, v_{1j}) = 0$ (in the case of multiple predictors, each set of covariances should be 0);
4. $Cov(CORR_j, \epsilon_{ij}) = 0$ (in the case of multiple L2 predictors, each covariance should be 0);

# Consequences of L1 assumption violations

Excluding a meaningful covariate at L1 ($Cov(EDU_{ij}, \epsilon_{ij}) \neq 0$) $\Rightarrow$ bias in L2 estimates.

Excluding a meaningful covariate at L2 that is associated with an L1 covariate ($Cov(EDU_{ij}, v_{0j}) \neq 0$) $\Rightarrow$ bias in L1 estimates.[9]

Measurement error in L1 covariate *may* bias L2 estimates if:

- ✓ measurement error in $EDU_{ij}$ varies systematically between $j$ countries;
- ✓ this variance is associated with a L2 covariate in the specification.

[9] This is one case where centering L1 covariates is helpful, as it wipes out the association between the random effect and the group-level average of $EDU_{ij}$.

# Diagnosing L1 assumption violations

Start by examining the dispersion in the residuals for the $j$ groups. The Bartlett–Kendall test can tell you whether they are roughly comparable ($H_0$: variances are equal $\Rightarrow$ you'd like to see $p > 0.05$).[10]

Examine normality of errors with density plots for the residuals in each of the $j$ groups, or Q-Q plots. The same graphical tools can be used on the L2 residuals.

Examine plots of residuals against predictors at L1, to assess the desired lack of association.

[10] If things aren't good, it may suggest issues with the L1 specification.

# Consequences of L2 assumption violations

Omitting a predictor at L2 biases the estimate of other predictors for that random effect. Without CWC centering, also biases estimates of predictors for *other* random effects.

Measurement error in L2 covariate biases estimates for other L2 covariates as well.

# Diagnosing L2 assumption violations

Standard plots could be used to diagnose heteroskedasticity in the distribution of these residuals.

Similarly standard plots can be used to diagnose normality.[11]

[11] This is complicated by the inability to observe the *true* slopes in the groups, but rather estimated slopes and their uncertainty.

# Presenting results in MLM

# What to showcase

It's important to know this, both for your own work, and when assessing the results encountered in the literature.

When in doubt, take any MLM-based analysis from the last decade from a top journal and see what they report there.

Let's take the models we worked with yesterday, and test 3 of them, gradually increasing the complexity.

# What to showcase

| | M1 | M2 | M3 |
|---|---|---|---|
| (Intercept) | 2.962 (0.041)$^{***}$ | 2.959 (0.042)$^{***}$ | 2.963 (0.041)$^{***}$ |
| Female | −0.144 (0.008)$^{***}$ | −0.145 (0.008)$^{***}$ | −0.145 (0.008)$^{***}$ |
| Education | 0.364 (0.008)$^{***}$ | 0.368 (0.020)$^{***}$ | 0.363 (0.018)$^{***}$ |
| Urban residence | 0.061 (0.008)$^{***}$ | 0.063 (0.008)$^{***}$ | 0.063 (0.008)$^{***}$ |
| Perceptions of corruption | 0.455 (0.091)$^{***}$ | 0.484 (0.085)$^{***}$ | 0.484 (0.085)$^{***}$ |
| Gini (10-point) | 0.059 (0.110) | −0.020 (0.103) | 0.073 (0.109) |
| Education * Gini | | | −0.111 (0.043)$^{*}$ |
| AIC | 67643.617 | 67524.823 | 67525.168 |
| BIC | 67710.610 | 67608.564 | 67617.284 |
| Log Likelihood | −33813.808 | −33752.411 | −33751.584 |
| Num. obs. | 32020 | 32020 | 32020 |
| Num. groups: country | 31 | 31 | 31 |
| Var: country (Intercept) | 0.052 | 0.053 | 0.052 |
| Var: Residual | 0.481 | 0.479 | 0.479 |
| Var: country Education | | 0.010 | 0.008 |
| Cov: country (Intercept) Education | | −0.010 | −0.008 |

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

# To include

You'll want to show these elements in a table of results:

- ✓ Fixed effects: if running out of space, focus on predictors of interest and exclude controls
- ✓ Random effects: as many as the model has
- ✓ Model fit statistics: LL, AIC, BIC[12]
- ✓ Sample size: level 1 and level 2

Some software reports these in a single table, while others make you collect the information from a few different tables.

[12] R doesn't report the deviance, but it can be easily computed and added.

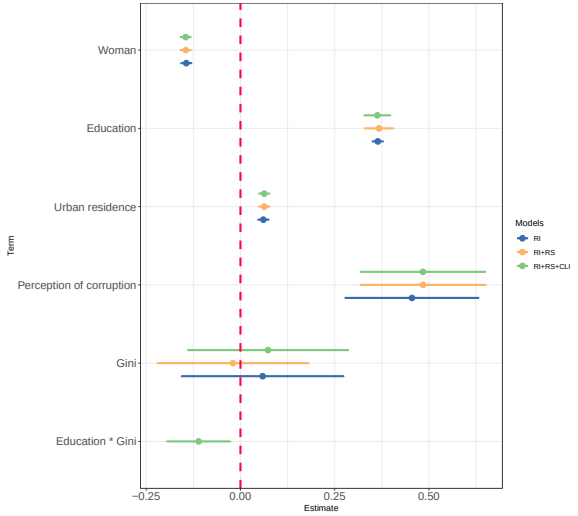
# Dot-and-whisker plots

Graphical display, which has the advantage of being easy to explain.

Can be handled automatically through packages like `dotwhisker`, though for maximum control a `tidyr`+`dplyr`+`ggplot2` pipeline can be established.

Limitation: pretty cumbersome for more than 3–4 models, and larger model specifications, but could be tailored to display only predictors of interest.

# Dot-and-whisker plots



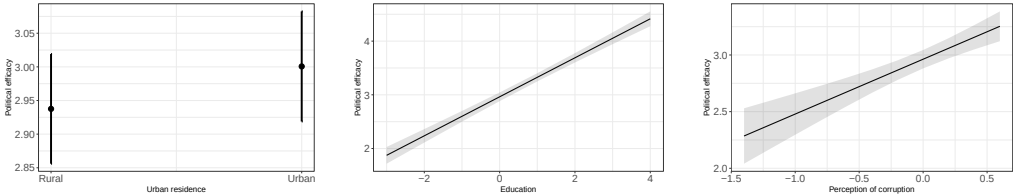

Intercept has been suppressed from the plot, as it "stretches" the horizontal axis.

Model fit statistics could be added in a text box, though perhaps only 1–2 indices for each model.

Sample sizes at the two levels of analysis could fit in a figure note.

# Marginal effects

Though they are not efficient in terms of space, they can be used in combination with other methods. Particularly good at highlighting one or two predictors, as well as at depicting interactions.




These include only fixed-effect uncertainty, which is the standard for reporting. Including random-effects uncertainty as well will make the CIs larger (sometimes, by a lot).

Thank you for the kind attention!

# References


Elff, M., Heisig, J. P., Schaeffer, M., & Shikano, S. (2018). *No Need to Turn Bayesian in Multilevel Analysis with Few Clusters: How Frequentist Methods Provide Unbiased Estimates and Accurate Inference.* Retrieved from `https://doi.org/10.31235/osf.io/z65s4`

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.

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