MLM measures of fit

Constantin Manuel Bosancianu[1]

WZB Berlin Social Science Center
Institutions and Political Inequality

July 3, 2019

This will be a longer tutorial, focusing on measures of model fit commonly used in multilevel modeling, as well as on how to compare between models. Due to the estimation technique most commonly used (Maximum Likelihood), the measures of fit are reliant on the likelihood of the model, making interpretation slightly more difficult than in the case of a R^2 from OLS regression. In the following sections I will go through five of the most commonly used measures, after which I will focus on how we engage in model comparisons in the case of multilevel models.

1 logLikelihood

The first measure, produced directly by the estimation procedure, is the logarithm of the likelihood of the model. In very simple terms, glossing over some of the more technical details, the likelihood is the product of the densities evaluated at the observations.[2] Drawing on high school math, the logarithm of a product is the sum of logarithms of the individual terms that constitute the product.

$$log(abc) = log(a) + log(b) + log(c) \tag{1}$$

With a large sample size, the densities will usually all be much smaller than 1 (see Figure 1 on page 2), making the individual logarithms each smaller than 0 (the natural logarithm of any quantity smaller than 1 will be negative). The sum of these negative quantities will itself be negative, producing the usual numbers reported in multilevel regression tables for the logarithm of the likelihood. Most often times, then, the *logLikelihood* will be a very small negative number, e.g. -12075.45.[3] The *logLikelihood* is a measure of fit that indicates how well our model fits the data: smaller values (farther away from 0) indicate models which fit the data worse.

[1] You can reach me at bosancianu@icloud.com. If you spot any mistakes I'd be grateful if you sent me an email pointing it out; I'll update the document and credit the help offered. One example was a small mistake concerning a subscript in equation 5, which Daniele Mantegazzi spotted in 2015 (thanks!).

[2] This very simple definition belongs to Isabel Canette, senior statistician at StataCorp LP.

[3] At the end of this tutorial I'll indicate a situation where this no longer applies.

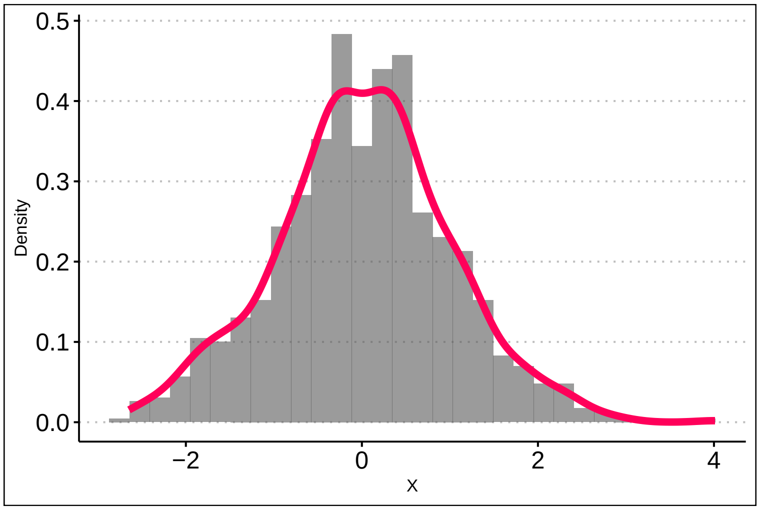

Figure 1: Normal density plot

2 Deviance

The *deviance* is computed as $-2 \times logLikelihood$, which makes it an indicator of model misfit: higher values (farther away from 0) indicate a worse fitting model. More about the deviance in a following section; for now, this simple formula will suffice.

3 AIC and BIC

The *Akaike Information Criterion* (AIC) is a simple extension of the logLikelihood.

$$AIC = -2 \times logLikelihood + 2 \times k \tag{2}$$

In equation 2, k represents the number of parameters estimated by our model. Unlike the deviance, the AIC includes a penalty for the number of parameters estimated in the model. When comparing two models this means that, *ceteris paribus*, the model with fewer parameters will receive a lower AIC score, suggesting a slightly better fit to the data.

The *Bayesian Information Criterion* (BIC), on the other hand, introduces a penalization both for number of parameters estimated and for the sample size.

$$BIC = -2 \times logLikelihood + k \times ln(N) \tag{3}$$

In equation 3, k continues to represent the number of parameters estimated by the model, while N is the sample size.[4] We can see from the formula that larger sample sizes result in larger penalties and,

[4] Singer and Willett (2003) recommend using the level 1 sample size.

therefore, in poorer fitting models. The ability of the BIC to incorporate both penalties has sometimes led to its use in comparing models estimated on different samples. Although this can sometimes be acceptable, in truth we know little about the behavior of the BIC under different model specifications and sample configurations. Because of this, my advice here is to avoid using this measure in these instances.

4 R^2

A final measure, proposed by Snijders and Bosker (1999) under the name R^2, assesses the extent to which a more complex model reduces the prediction error (interpreted here as the magnitude of the residuals), compared to a less complex model (Luke, 2004, p. 35). Considering two models, M_1 and M_2, where M_2 includes all of the parameters in M_1 plus a few more, and that $var(u_{0j}) = \tau_0^2$ and $var(r_{ij}) = \sigma^2$, then we can write the formula[5] for R^2 at the level 1 as

$$R_1^2 = 1 - \frac{(\sigma^2 + \tau_0^2)_{model2}}{(\sigma^2 + \tau_0^2)_{model1}} \tag{4}$$

From equation 4 we can see that the better M_2 manages to explain the variance in the residuals, the lower is the value of the fraction from the equation. At the extreme, when M_2 perfectly explains the residuals, $R_1^2 = 1 - 0 = 1$. A similar formula gives us the R^2 at the level 2:

$$R_2^2 = 1 - \frac{(\frac{\sigma^2}{n} + \tau_0^2)_{model2}}{(\frac{\sigma^2}{n} + \tau_0^2)_{model1}} \tag{5}$$

The n in equation 5 designates the average number of level 1 units that are nested in level 2 units. The major benefit of the R^2 is that it is able to offer much more precise information than other indicators, as it captures changes at both levels of the hierarchy. On the other hand, in some instances adding a predictor might result, paradoxically, in a *larger* variance of residuals, and therefore in a *negative* R^2.

5 Usage of the indicators

As you get started with using these you might find useful a few rough guidelines regarding the use of these indicators in practical applications. To begin with, except for the BIC, the indicators should only be used to compare two models which have been estimated on the same sample. Without such a restriction the researcher has no way of knowing whether the reduction in logLikelihood or AIC is due to the addition of a predictor in the model, or due to a different sample on which the model is tested. Theoretically speaking, the BIC could be used for comparing models estimated on different samples,

[5] In this tutorial I presented formulas as they are reported by Luke, pp. 35–36, as these are more accessible to a new audience than the ones offered by Snijders and Bosker (1999).

as its formula includes a correction for sample size. In practice, though, this is rarely done, for the reasons outlined in the section which deals with this indicator.

A second guideline refers to the type of models which can be compared using these indicators. A conservative practice is to use the deviance and the AIC only for *nested* models. By *nested* models, I mean a configuration of two models whereby one includes all of the parameters of the other one, as well as a few extra ones. For example, if $M_1 : Y = \beta_0 + \beta_1 X_1 + e$ and $M_2 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$, then M_1 is nested in M_2.

In truth, this is one of the more contentious issues in multilevel modeling, about which disagreement still persists. Hox (2010, p. 50), Luke (2004, p. 35) and Burnham and Anderson (2002, p. 88) suggest that AIC can be used for non-nested models, while Hamaker, van Hattum, Kuiper, and Hoijtink (2011, p. 252) endorse this view, heavily citing Burnham and Anderson for support. Pinheiro and Bates (2000, p. 259) also take the view that AIC and BIC can be used for non-nested models. On the other hand, Ripley (2004) flatly states that AIC comparisons are only useful for nested models, a view which he expresses outside of his writings as well.[6] Singer and Willett (2003, p. 122) suggest that one can use AIC and BIC to compare non-nested models, but urge researchers to only use them when more traditional methods of model comparison don't work. Furthermore, they advise caution when using them. At the same time, in a handout/presentation from 2003, the same authors only say that one can "supposedly" use these information criteria to engage in such comparisons.[7] It does seem that the weight of the evidence tends to favor the "yes, we can!" camp. I would advise some restraint, though, when engaging in such AIC-based comparisons of non-nested models: (1) don't compare models run with different R functions, as they might produce likelihoods that aren't comparable; (2) don't compare models in which one uses a transformation of the outcome, while the other one doesn't.[8]

Finally, a measure of caution should be used with respect to the AIC/BIC and FIML or REML estimation. When using FIML estimation, models that differ either in fixed effects or random effects can be compared with the AIC or BIC. When relying on REML estimation, though, only models that differ in random effects can be compared (Hox, 2010, p. 51).

6 Trivia: can the logLikelihood be positive as well?

In a few instances, yes, it can. If you remember from the introductory part of this document, the likelihood is computed as the product of the densities evaluated at the observations. When taking the

[6] See the discussion at `http://r.789695.n4.nabble.com/Nested-AIC-td794191.html`, where Thomas Lumley seems to also endorse the view that AIC-based comparisons for non-nested models are somewhat risky. Somewhere else, Brian Ripley takes a more blunt approach toward the statements made by Burnham and Anderson in their book: `https://stat.ethz.ch/pipermail/r-help/2003-June/035526.html`.

[7] See `http://gseacademic.harvard.edu/alda/Handouts/ALDA%20Chapter%204.pdf`.

[8] In his response to the discussion at `http://r.789695.n4.nabble.com/Nested-AIC-td794191.html` Thomas Lumley outlines a few more situations where AIC should not be used for non-nested models.

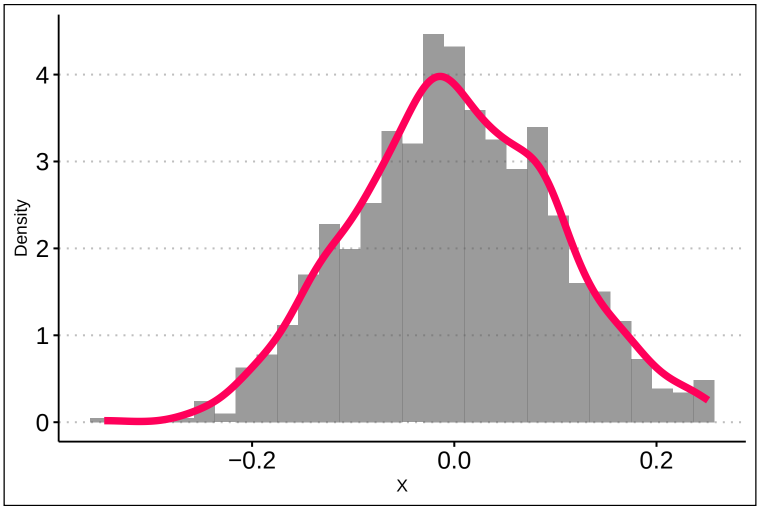

Figure 2: Normal density plot with very low variance

logarithm of this quantity, we obtain a sum of logarithms of densities. As long as the distribution looks similar to the one presented in Figure 1, with a normal degree of dispersion, the *logLikelihood* should be negative on account of each element in that sum being negative. Consider what would happen, though, if the distribution would have a very low variance, as the one presented in Figure 2.

In this instance, a considerable number of elements have a density above 1, which means that their logarithm will have a positive value. If our distribution has a lot of these elements, then it's possible that the logLikelihood will be positive.

References

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). New York: Springer.

Hamaker, E. L., van Hattum, P., Kuiper, R. M., & Hoijtink, H. (2011). Model Selection Based on Information Criteria in Multilevel Modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 231–255). New York: Routledge.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications*. Routledge.

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

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Ripley, B. D. (2004). Selecting Among Large Classes of Models. In N. Adams, M. Crowder, D. J. Hand, & D. Stephens (Eds.), *Methods and Models in Statistics: In Honour of Professor John Nelder, FRS* (pp. 155–170). London: Imperial College Press.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (1st ed.). London: Sage.