Multilevel Regression Estimation Methods for Continuous Dependent Variables

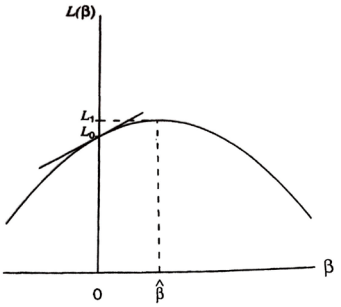
General Concepts of Maximum Likelihood Estimation

The most commonly used estimation methods for multilevel regression are maximum likelihood-based. Maximum likelihood estimation (ML) is a method developed by R.A.Fisher (1950) for finding the best estimate of a population parameter from sample data (see Eliason,1993, and Enders, Chapter 2, 2022, for accessible introductions). In statistical terms, the method maximizes the joint probability density function (pdf) with respect to some distribution. With independent observations, the joint probability of the distribution is a product function of the individual probabilities of events, so ML finds the likelihood of the collection of observations from the sample. In other words, it computes the estimate of the population parameter value that is the optimal fit to the observed data.

ML has a number of preferred statistical properties, including asymptotic consistency (approaches the parameter value with increasing sample size), efficiency (lower variance than other estimators), and parameterization invariance (estimates do not change when measurements or parameters are transformed in allowable ways). Distributional assumptions are necessary, however, and there are potential biases in significance tests when using ML. ML can be seen as a more general method that encompasses ordinary least squares (OLS), where sample estimates of the population mean and regression parameters are equivalent for the two methods under regular conditions. ML is applied more broadly across statistical applications, including categorical data analysis, logistic regression, and structural equation modeling.

Iterative Process

For more complex problems, ML is an iterative process [for multilevel regression, usually Expectation-Maximization (EM) or iterative generalized least squares (IGLS) is used] in which initial (or "starting") values are used first. The computer then computes the likelihood function, which represents a lack of fit, for that set of parameter "guesses." On the next step, another set of parameter estimates are used and so on until there is a "response surface" that represents the likelihood values for all of the guesses. Each step is called an *iteration*. The idea is similar to the idea of ordinary least squares (OLS) in regression in which the squared errors or residuals are minimized to obtain the best fit of the regression line to the data and the regression coefficients.



p.12. Agresti, A. Categorical data analysis, third edition. New York: Wiley.

Tangent lines can be drawn (first derivative) for any particular point on the curve (as in the point L_0 in the figure above), and when the slope of the tangent line equals 0 (second derivative), the maximum of the curve of possible estimates is found, and this point corresponds to the optimal sample estimate of the parameter. The computer stops when a certain selected criterion for closeness of fit has been reached (convergence) and values for the fit of the overall model and the parameter values are generated. More complex models, particularly those with more random slopes may take more iterations to converge. Software packages set a maximum number of iterations, which the user must often override. Generally, this is not a problem, but it is also not unusual for convergence to fail even when given a large number of iterations. In these instances, the researcher will need to make choices about the number of random effects estimated and possibly the number of predictors in the model. Missing data, dependent variable

distributions, and the number of random effects relative to group size may be additional factors in convergence difficulties.

Restricted or Residual Maximum Likelihood (REML)

The most common (and usually the default) estimation method for multilevel regression models is a variant of full maximum likelihood called either restricted or residual maximum likelihood (REML) by various authors.[1] Full maximum likelihood is typically also available as an estimation option, but it produces variance (random) effects estimates that are biased (usually underestimated; Longford, 1993), with more substantial biases occurring with smaller samples (fewer groups). REML differs from full ML in a couple of ways (Raudenbush & Bryk, 2002, Chapter 3). First, REML takes into account a degrees of freedom correction (much like the difference between sample and population variance formulas) for the variance effects based on the number of fixed effects in the model. Secondly, whereas full ML alternates estimation of variances (random effects) and fixed effects, REML estimates them separately by first estimating the variances iteratively and then estimating the fixed effects coefficients (see McNeish, 2017, for a very clear discussion of the differences between ML and REML estimation).

REML is more commonly used in multilevel regression because its sampling variance estimates are less biased with fewer groups (Browne, 1998). Although the two methods usually produce similar results, with closer correspondence as the number of groups increases (i.e., they are asymptotically equivalent), it is wisest to use REML as a default method because of its general better performance (Hox & McNeish, 2020). With sufficient sample sizes, REML produces good estimates of fixed and random effects and their standard errors, but smaller samples and distributions can impact random effects estimates and significance tests (Maas & Hox, 2004; 2005; McNeish & Stapleton, 2016a). Although fixed effects estimates using ML and REML themselves are not impacted by nonnormal distributions, standard error estimation and significance tests for both the fixed and random effects may be sensitive to violations of normality assumptions.[2]

Bayesian Estimation

The Bayes statistical approach is usually considered to be a philosophical departure from maximum likelihood (or any other classic or “frequentist”) estimation approach. While maximum likelihood estimation can be thought of as an assessment of the probability of the observed data given a certain parameter value, Bayesian analysis can be thought of as assessing the probability of the parameter given the data. In the Bayesian approach, the parameters are thought of as random, uncertain values. Estimates are obtained by using a prior distribution, which is an initial assumed distribution that can be anything from a minimally useful distribution (e.g., normal distribution with very wide variance) to something very specific (e.g., normal distribution with specific mean and narrow variance), weighted by a likelihood function, to arrive at a posterior distribution. Depending on the type of priors used, the posterior distribution is more influenced or less influenced by the prior distribution. The process is implemented with an iterative algorithmic search called a Markov chain Monte Carlo process (or Gibbs sampler) that searchers for one parameter at a time, on each step sampling at random possible values from the prior distribution given the chosen constraints. At the end of the process, the probable values are chosen and uncertainty intervals are derived around them (called “credible” or “credibility” or “posterior probability” intervals). The interpretation of the intervals with Bayesian modeling is that there is a 95% probability that the true parameter falls within that range. The maximum likelihood (and other frequentist) intervals represent the probability that 95% of intervals constructed from all the other samples in the sampling distribution will contain the true population value. In many simple cases and in many practical applications, the Bayesian and the frequentist approaches will arrive at very similar values and the same conclusions. Bayesian estimation is a very popular idea in multilevel (and other) modeling but has not become that widely implemented by researchers in practice yet. Bayesian estimation can be useful in

[1] You may see maximum likelihood abbreviated as ML, MLE, FIML, or FML and you may see restricted (or residual) maximum likelihood abbreviated as REML or RML.
[2] See the subsequent "Diagnostics" and "Robust Standard Errors" handouts for more information.

circumstances when multilevel models have difficulty converging or are likely to have biased estimates. Small samples with multilevel models (few groups) is one area in which Bayesian estimation may be useful, although its performance compared with the restricted maximum likelihood approach (particularly with corrections, such as Kenward-Roger adjustments) is heavily dependent the use of correct and informative priors (McNeish, 2016). The choice of prior can sometimes lead to very different results or conclusions (Depaoli, 2014; Gelman, 2006; Lambert et al., 2005), so a potentially difficult challenge is finding good priors, requiring strong theoretical or solid outside empirical information, particularly for complex models with many interdependent parameters.

References and Recommended Readings

Browne, W.J. (1998). *Applying MCMC methods to multilevel models.* Bath, UK: University of Bath.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1*, 473–514.

Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in medicine, 17*, 1261-1291.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association, 76*, 341-353.

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(2), 239–252.

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice* (Vol. 96). Newbury Park, NJ: Sage Publications.

Enders, C. K. (2022). *Applied missing data analysis.* Guilford Publications.

Fisher, R. A. (1950). *Contributions to mathematical statistics.* New York: Wiley.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). Bayesian Analysis, 1(3), 515-534.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43–56.

Goldstein,H. (1987). Multilevel covariance component models. *Biometrika, 74*, 430-31

Goldstein, H. (1989). Restricted unbiased iterative generalized least-squares estimation. *Biometrika, 76*, 622-623.

Hox, J., & McNeish, D. (2020). Small samples in multilevel modeling. n R. Van de Schoot & M. Miočević (Eds.), Small sample size solutions, (215-225). Routledge.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? a simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine, 24*(15), 2401–2428.

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22.

Lindley, D.V., & Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34*, 1–41.

Mason, W. M., Wong, G. M., Entwistle, B.: Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.) *Sociological methodolog*y, San francisco, Jossey-Bass, 1983, 72–103

Maas, C. J., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*(2), 127-137.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86-92.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 750-773.

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. Multivariate behavioral research, 52(5), 661-670.

McNeish, D. M., & Stapleton, L. M. (2016a). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*(2), 295-314.

McNeish, D., & Stapleton, L. M. (2016b). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, *51*, 495–518.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). The analysis of repeated measurements: A comparison of mixed-model Satterthwaite F tests and a nonpooled adjusted degrees of freedom multivariate test. *Communications in Statistics-Theory and Methods*, *28*, 2967–2999.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of education*, 59, 1-17.

Zeger, S.L., and Karim, M.R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. J*ournal of the American Statistical Association, 86*, 79–86.

Summary of Multilevel Estimation Methods for Continuous Dependent Variables

Estimation Method	Description	Algorithms	Comments
Full Information Maximum Likelihood (ML or FML: Goldstein, 1986; Longford, 1987)	Estimates variances and covariances (e.g., τ_0^2 and τ_{01}) assuming known values for the regression coefficients (rather than estimates of the population values). Estimation uses fixed coefficients in the likelihood function. Between group variance estimates, (e.g., τ_0^2 and τ_{01}), underestimated.	• Iterated Generalized Least Squares (IGLS; Goldstein, 1986; 1987) • Gauss-Newton scoring method	As long as the residuals are normally distributed IGLS is a good algorithm for ML. ML is usually preferable for comparing deviances of models differing in the fixed effects, because it includes regression coefficients in the likelihood function (i.e., use ML to compare deviances to test subsets of predictor variables)
Restricted or Residual Maximum Likelihood (REML or RML: Goldstein, 1989; Mason, Wong, Entwistle, 1983; Raudenbush & Bryk, 1986)	Estimates of variances and covariances assuming regression coefficients are unknown. Estimation done separately for random and fixed coefficients. Unbiased estimate of between group variances. ML and REML have similar results if the number of groups is large, but preferable to ML for small samples (Browne, 1998; Browne & Draper, 2006; Longford, 1993; McNeish & Stapleton, 2016b)	• Restricted Iterated Generalized Least Squares (RIGLS: Goldstein, 1986, 1987, 1989) • Expectation Maximization (EM: Bryk & Raudenbush, 1992) • Fisher Scoring (Longford, 1993)-equivalent to IGLS • Gauss-Newton scoring method	REML is the default, but ML can be requested under "basic specifications" in HLM or by using /METHOD = ML in SPSS or REML = FALSE in R. Deviance comparisons used only for testing random coefficients (i.e., do not use REML and compare deviances to test subsets of fixed coefficients).
Empirical Bayes Estimates (EB: Dempster, Rubin, Tsutakawa, 1981; Lindley & Smith, 1972)	EB estimates are used in the estimation of particular group intercept or slopes (e.g., plotting, or assumption checking). The REML estimation in multilevel models can be interpreted from an EB perspective (see Raudenbush & Bryk, 2002, Chapter 13). Each group intercept or slope value is based on information from the grand mean (γ_{00}) and the average slope value (e.g., γ_{10}) across all groups as well as the slope estimate from each group (as in the OLS estimation) weighted by the reliability. Groups with larger sizes are more heavily weighted toward the OLS estimate and groups with smaller sizes are weighted more heavily toward the full sample estimates (γ_{00} and γ_{10}).	• Empirical Bayes or "shrinkage" estimates of intercepts and slopes are derived from the REML or ML process (see Raudenbush & Bryk, 2002, p. 47, Chapter 13) • When requesting plots for random slopes or intercept and slope estimates for each group	Bayes estimates are output in a "residual" file in HLM if requested under Basic Specifications. Can be used to assess normality assumption and are the values used in plots in HLM and the R method I demonstrated (the SPSS method I demonstrated uses OLS estimates, which are not shrunken).
Robust standard error estimates (Liang & Zeger, 1986)	Estimation of the standard errors for the fixed effects that corrects for biases (usually underestimation) due to non-normality of the dependent variable. Gives the best estimates when there is a large number of groups. The distribution of the dependent variable (or rather the residual distribution) is taken into account in computation of the standard errors.	• HLM uses a General Estimating Equation (GEE) approach (Liang & Zeger, 1986; Burton, Gurrin, & Sly, 1998) • Other packages such as Mlwin use a Huber-White, Eicker-Huber-White, or "sandwich estimator" (Eicker, 1963; Huber, 1967; White, 1980)	As long as the number of groups is moderately large (relative to the number of coefficients estimated) the robust estimates are usually preferable. Use with caution if there are less than 100 groups.
Bayesian Estimation (MCMC; Zeger & Karim, 1991)		• Markov chain Monte Carlo (MCMC) e.g., Gibbs sampler (Zeger & Karim, 1991) • Statistical packages such as BUGS/OpenBugs or RStan (using Gibbs/MCMC) use EB approaches more explicitly in model estimation. The Bayes factors approach is available in Mplus and MLWin.	Multilevel Bayesian estimation can be found in the R package `brms` and Mplus. Mplus offers a Bayes factors (Kass & Raftery, 1995) approach used to compare the relative likelihood of two hypotheses rather than compare an obtained sample coefficient to a null hypothesis value.