**Multilevel Models with Binary and other Noncontinuous Dependent Variables**

Multilevel models with binary or count dependent variables can be understood in terms of the generalized linear modeling approach described by McCullagh and Nelder (1989) in which the predicted score is transformed.  In the case of logistic regression, this is a log transformation of predicted probabilities of a binary variable on the left of the equation with a standard linear regression equation on the right.

$$\ln\left(\frac{\hat{P}}{1-\hat{P}}\right) = \beta_0 + \beta_1 x_{ij}$$

Estimates from a single-level logistic equation produce regression coefficients that can be easily transformed into odds ratios, where OR = $e^{\beta}$ representing the odds of $Y = 1$ for each unit increase in $X$. Different link functions are used such as an ordered logit or probit link function for ordinal data, a Poisson log link function for counts, or a variety of other link functions for different data characteristics. A general way to state these models is to use the simple Greek letter eta, $\eta$, to represent the transformed predicted value. Then various link functions can be used to compute $\eta$ in the level-1 equation:

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij}$$

Multilevel regression models for non-continuous outcomes can be conceptualized similarly, sometimes referred to as hierarchical generalized linear models, although their accurate estimation is not as simple or straightforward as it is for the single-level generalized models. There are several good introductions to the various estimation issues, so I will not go into much detail here about estimation (Hedeker, 2005; Chapters 10 & 14, Raudenbush & Bryk, 2002; Chapter 6, Hox, 2002; O'Connell, 2022). Multilevel models can be estimated for binary, ordinal, multicategorical, and count variables.

**Software**

Some comments on software are necessary because not all software implementations of generalized multilevel models are the same. There are now a variety of software packages that can estimate models with noncontinuous outcomes. In SPSS, multilevel models for binary outcomes can be analyzed with `GENLINMIXED` (starting with Version 19), but I do not recommend their estimation approach. HLM and R both have high quality estimation methods. Several other software packages, including `SAS PROC GLIMMIX` (Version 9.2 or higher), Stata `xtmixed` and `GLLAMM`, and MixOR/Super Mixed also have features for estimation methods described below.

HLM has three approaches available for binary and count data, penalized quasi-likelihood (PQL), Laplace approximation, and adaptive quadrature (added in HLM 7). PQL is the only available estimator in `SPSS GENLINMIXED` procedure (Version 19 or higher). PQL estimates are used for the tests of random effects, but for fixed effects there are three choices. PQL is less computationally intensive and gives acceptable estimation in many cases. It can produce biased estimates, however, when population variance values are large or events are rare (high or low probability of the outcome). Because it is tough to draw the line on where PQL estimates will and will not be problematic, it is difficult to recommend this estimation approach. A variant of the PQL approach is restricted or residualized PQL (RPQL), which uses a similar modification as the REML modification of ML, appears to be a potentially substantial improvement over PQL (McNeish, 2016) but is not widely available (presently only in SAS). Laplace approximation and adaptive quadrature estimates work well under a wide variety of contexts but probably requires at minimum 50 groups with 100 or more recommended. The adaptive quadrature approach (sometimes called Gauss-Hermite; Pinheiro and Bates, 1995) is more computationally intensive (but that is less of a concern these days except for larger models) and may have some advantages for a rare binary outcome (very small proportion of 1s) and very small group sizes (e.g., 2; Capanu, Gönen, & Begg, 2013; Jackson, 2006; Rabe-Hesketh, Skrondal, & Pickles, 2002; 2005). Work by McNeish and Stapleton (McNeish, 2016; McNeish & Stapelton, 2016) suggest that adaptive quadrature and Laplace estimates do not work well with small number of groups (e.g., < 50) and that the RPQL method

especially with Kenward-Roger degrees for freedom and standard errors and larger number of cases per group, is superior.

**Unit-Specific vs. Population-Average Estimates**
The HLM output distinguishes between unit-specific (or "subject-specific") and population-average (or "population-averaged") estimates (Zeger, Liang, & Albert, 1988). Population-average estimates are sometimes referred to as "marginal model" estimates. Categorical analysis models are different from continuous analysis models in that marginal (population-average) or conditional (subject-specific or unit-specific) estimates can be obtained. There is no such distinction for continuous analysis models (i.e., using the identity link). The two analysis models differ because of the nonlinear (log) link to predicted values that is required for logistic or other non-continuous models. As a very simple example, for matched paired data, we can derive a model-based estimate of the individual paired differences between individual binary scores (subject-specific or conditional) or the difference between proportions taken overall subjects (population-average or marginal).[1] This is a simple repeated-measures case in which scores are nested within subjects, but the same concept can be generalized to individuals nested within groups (or clusters). More generally in the multilevel generalized regression model context for repeated measures (growth curves), the unit-specific effect estimates the expected value of the regression coefficient for a person-level predictor assuming the effects of predictors vary across individuals and allowing individuals to have a different baseline value (effects are conditioned on the intercept variance).[2] Such an interpretation is the interpretation given in the context of continuous outcomes. The population-average effect, on the other hand, is an estimate of the regression coefficient for a person-level predictor which assumes the baseline value is the same for all individuals (i.e., averages over all the individuals). Thus, the population-average effect represents an average over the sample and is more appropriate for estimating the predicted population proportion. This suggests examining unit-specific values for the purpose of explanatory hypotheses but examining population-average values for estimating population prevalence. The population average approach uses a generalized least squares estimation with robust standard errors (a GEE approach; Zeger et al., 1988; see also McNeish, Stapleton, & Silverton, 2017). The robust standard errors are less sensitive to misspecification and distributional assumptions. More on the distinction can be found in Neuhaus, Kalbfleisch, and Hauck (1991) or Raudenbush and Bryk (2002). See McNeish (2016) for a useful description of unit-specific and population-average effects.

The HLM software output distinguishes between unit-specific and population-average estimates (Zeger, Liang, & Albert, 1988), which is only possible when we use non-continuous outcomes. R uses only unit-specific estimates.

**Ordinal and Count Variables**
*Ordinal.* When there are more than two outcome categories that can be rank ordered, say "never", "sometimes", and "often", an expansion of the above concepts involving a logit (natural log link with logistic error distribution) probit link (probit link with normal error distribution) function can be used. Ordinal models involve threshold estimates where we use the notion that once the propensity on an underlying unobserved continuous distribution is exceeded, we observe the next higher value on the dependent variable (1 instead of 0, 2 instead of 1 etc.). Threshold values are estimated representing a normalized value along the logistic or continuous distribution of the underlying latent variable, $\eta$. HLM uses a logit model, and, in R, one can choose either logit or probit models through the `lmer` function in the `lme4` package. In HLM, choose "ordinal" under the "Basic Settings" menu. In R, add either of the following to the `lmer` function: `family = "binomial(link="logit")` or `family = "binomial(link="probit")`. Some research suggests that multilevel ordinal models may outperform REML estimation when there are five or six (or fewer) ordinal values on the outcome (Bauer & Sterba,

---

[1] The log transformation creates this circumstance because the log of differences is not equal to the difference between two logs. Without the log transformation the average difference score is equal to the difference between two averages.
[2] This is stated here in terms of longitudinal growth models. In the nested case, the unit-specific estimate involves the slope estimate conditioned on the intercept variance across groups, and the population-average estimate involves assuming average intercept across groups.

2011), a circumstance social scientists usually treat as a continuous outcome. This seems to be especially true when the number of level-2 units is less than 100.

*Count.* When the outcome is a count, such as the number of times using the city bus in the last month, the normal residual assumption is usually not met. Poisson and negative binomial models are usually used for these, both of which can also be conceptualized in terms of the generalized linear model as well. HLM offers a Poisson estimation with or without adjustment for overdispersion, in which the assumption that the conditional distribution of the dependent variable is equal to the mean is violated. Checking the overdispersion box will usually give estimates for count data preferable to just using straight Poisson, whether the data are overdispersed, equidispersed, or underdispersed.[3] In R, standard Poisson can be estimated with `family = poisson(link = "log")` using the `lmer` function, and the negative binomial estimation, which does not assume equidispersion, can be obtained with the `glmer.nb` function in `lme4`. Either can be combined with the adaptive quadrature estimation method. Although adaptive quadrature and Laplace approximations work well with larger number of groups, as with multilevel logistic models, the RPQL approach (which has limited software availability presently), particularly with Kenward-Roger degrees of freedom and standard errors, is preferable with smaller number of groups (McNeish, 2019). McNeish found that when there is a large number of cases within group, then as few as 20 groups without Kenward-Roger corrections and 10 groups with Kenward-Roger corrections were sufficient.

## R-Squared for Non-continuous Multilevel Models

Approximating variance accounted for (*R*-squared values) with multilevel models has an added layer of complexity and there are even fewer empirical studies about their performance when outcomes are not continuous. Nakagawa and Schielzeth (2013) provide a discussion of some of the proposed approaches and propose a method for the generalized linear modeling case. Computing their proposed measures, $R^2_{GLMM(m)}$ and $R^2_{GLMM(c)}$, manually is somewhat complicated, because it requires computation of the variance of the predicted values. The Rights and Sterba pseudo-R2 values can be obtained with the R package `r2mlm`, but these have not been studied in the binary binary outcome case. The R package `r2glmm` will compute their $R^2$ measure with the `r2bet` (using `method="nsj"`), but the documentation is unclear about which of the two proposed equations are used.

**Selected References**

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological methods, 16*, 373.

Capanu, M., Gönen, M., & Begg, C. B. (2013). An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Statistics in medicine, 32*(26), 4550-4566.

Jackson, C., Best, N., & Richardson, S. (2006). Improving ecological inference using individual-level data. *Statistics in medicine, 25*(12), 2136-2159.

McNeish, D. (2016). Estimation methods for mixed logistic models with few clusters. *Multivariate Behavioral Research, 51*(6), 790-804.

McNeish, D. (2019). Poisson multilevel models with small samples. *Multivariate Behavioral Research, 54*(3), 444-455.

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*, 295-314.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114-140.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*, 133-142.

Neuhaus, J. M., Kalbfleisch, J. D., & Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. International Statistical Review/Revue Internationale de Statistique, 25-35.

O'Connell, A. A.. Multilevel logistic and ordinal models. In O'Connell, A. A., McCoach, D. B., & Bell, B. A. (Eds.). (2022). Multilevel modeling methods with introductory and advanced applications (pp. 209-254). IAP.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics, 128*(2), 301-323.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. The *Stata Journal, 2*(1), 1-21.

---

[3] For HLM with Poisson models using PQL, the user can specify constant exposure, assuming each case has an equal opportunity to experience the event, or variable exposure, in which opportunity for experiencing the event is not assumed to be equal for all individuals. Also, negative binomial models are generally preferable to Poisson models because they do not assume equidisperson. The overdispersion Poisson model is a good alternative, however.