

AN ILLUSTRATION OF MULTILEVEL MODELS FOR ORDINAL RESPONSE DATA

Ann A. O'Connell

The Ohio State University, United States of America
aoconnell@ehe.osu.edu

Variables measured on an ordinal scale may be meaningful and simple to interpret, but their statistical treatment as response variables can create challenges for applied researchers. When research data are obtained through natural hierarchies, such as from children nested within schools or classrooms, clients nested within health clinics, or residents nested within communities, the complexity of studies examining ordinal outcomes increases. The purpose of this paper is to present an application of multilevel ordinal models for the prediction of proficiency data. Implications for teaching and learning of multilevel ordinal analyses are discussed.

INTRODUCTION

Many research applications in the education sciences and related fields involve studies of response variables that are measured on ordinal scales, and thus are inherently non-normal in their distribution. Examples of ordinal variables include scores on Advanced Placement examinations (5: Extremely well qualified; 4: Well qualified; 3: Qualified; 2: Possibly qualified; 1: No recommendation) or proficiency scores derived through a mastery testing process, such as when researchers are interested in analyzing factors affecting students' level of proficiency in reading or mathematics (e.g., 1: Below basic; 2: Basic; 3: Proficient; 4: Beyond proficient). Regression analyses of ordinal data, as well as other kinds of non-normal outcomes including dichotomies, rates, proportions, or times-to-event (survival) are typically drawn from a family of models broadly known as *generalized linear models*, of which the standard ordinary least squares regression model for continuous outcomes is a special case. Generalized linear models have been used to represent the behavior of a wide variety of limited or discrete outcomes in practice, and their theoretical connection with standard linear regression models helps to simplify their application.

In addition to non-normality, education data are often hierarchical, posing an additional complexity for many educational research studies. Hierarchical data occur across multiple levels of a system and involve two or more levels of levels of sampling, making data obtained from studies of teachers or students within schools, students within classrooms, or residents within communities ideally suited for multilevel modeling. In a multilevel research study, the higher-level units or clusters (i.e., the schools, classrooms, or communities) are assumed to be independently sampled, and the unit-of-analysis is the response at the lowest level of the hierarchy (i.e., individual responses for the students or the residents). The analysis is used to estimate and model variability in responses occurring within as well as between the higher-level units. Variation between units represents differences attributable to settings or contexts and is captured by inclusion of group- or unit-level random effects in the model. This variability is often the most interesting component in hierarchical studies and can be modeled through addition of group-level variables to examine the effect of differences in settings or contexts. Throughout this paper, the focus is on models where these random effects are assumed to be normally distributed. When the outcome variable of interest is non-normal, the most commonly used models are typically called *hierarchical generalized linear models* or HGLMs, although they have also been referred to as *generalized linear mixed models* (GLMMs) (McCulloch & Searle, 2001). Breslow (2003) distinguishes HGLM as the general case and GLMM as a special case when the random effects at level-two are assumed to be normally distributed.

TYPES OF GLMM'S FOR ORDINAL RESPONSE VARIABLES

The application of generalized linear models in a multilevel framework parallels their use in single-level research designs (O'Connell, 2006; O'Connell, Goldstein, Rogers & Peng, 2008). Common models for single-level ordinal outcomes include the proportional (or cumulative) odds (PO) model, continuation ratio (CR) model, and adjacent-categories (AC) model. Each of these imposes specific assumptions onto the data, and the research question of interest often dictates which model is best applied. Each of these can be characterized as extensions of hierarchical

logistic regression models for dichotomous outcomes typically coded as 0 or 1, where 1 represents the “success” outcome or event of interest. The logistic regression model predicts the probability of success conditional on a collection of categorical or continuous predictors through application of the logit-link. The logit, by definition, is the natural log of the odds, where the odds is a quotient that conveniently compares the probability of success to the probability of failure.

With ordinal outcomes, there are several ways to characterize what is meant by success. For instance, a K-level ordinal variable can be partitioned into K-1 sequential subsets of the data: category 1 versus all above; categories 1 and 2 combined versus all above; categories 1 and 2 and 3 combined versus all above, etc.; the final *cumulative split* would distinguish responses in categories less than or equal to K-1 versus responses in category K. Thus, there is a series of cumulative comparisons for success, where success is defined as being in categories at or below the k^{th} cutpoint (*ascending* option, where success is $P(Y \leq k)$). Alternatively, success could be defined as being in categories greater than or equal to the cutpoint (*descending* option, where success is defined as $P(Y \geq k)$). In either cumulative representation, all data is retained at each split and both options yield a similar interpretation of effects of predictors for the data. On the other hand, the CR model creates *conditional* splits to the data. The success probability in the CR models is defined as: $P(Y > k/Y \geq k)$. Note that the ascending and descending options for the CR approach will *not* yield similar results, since successive response categories are essentially dropped from the representation at each split. Finally, a third commonly used approach for ordinal data is the adjacent categories model, where the success probability is based on whether or not a response is in the higher or lower of two adjacent categories: $P(Y = k + 1/Y = k \text{ or } Y = k + 1)$.

Each representation discussed above imposes a restrictive assumption on the data, generally referred to as the assumption of proportional or identical odds. This assumption implies that the effect of any explanatory variable remains constant regardless of the particular split to the data being considered. For instance, consider a six-category ordinal outcome, with responses coded from 0 to 5. The proportional odds assumption implies that the effect of a predictor such as gender is assumed to be the same whether we are referring to the probability of a response being less than or equal to 0, or to a response being less than or equal to 3. A similar constraint is imposed on the CR and AC models as well. Models in which some predictors exhibit *non-proportional* odds are referred to as “partial” proportional odds models, but tests for proportionality for single-level models have been shown to lack statistical power (Allison, 1999; Peterson & Harrell, 1990). Ad-hoc methods for investigating proportionality in the multilevel framework include fitting the underlying series of hierarchical logistic models and examining departure from consistent patterns in variable effects among the predictors. In many research situations, the assumption of proportionality or identical odds is a reasonable one but examples of non-proportionality can be found (Hedeker & Gibbons, 2006). Researchers need to carefully consider the implications of these assumptions for their own data situations.

Most software for the analysis of multilevel ordinal data will fit the PO model, which is based on the cumulative logit link, although other link options, such as the complimentary log-log link for CR models, are currently available in a few statistical packages. As estimation and software methods for ordinal random- or mixed-effects models continue to improve, the likelihood is that these different alternatives will become more widely available, including those that allow for partial-proportional odds.

The Hierarchical Proportional Odds Model

The proportional odds model is the most widely used approach for analyzing hierarchical ordinal data. For a K-level ordinal outcome, the cumulative probability of success (using the ascending option) across the K-1 cumulative splits is based on a model using the cumulative logit link for the response, R_{ij} , for the i^{th} person in the j^{th} group. Utilizing terminology from Raudenbush and Bryk (2002), the model is characterized by level as follows:

$$\text{Level 1: } \eta_{kij} = \ln(Y'_{kij}) = \ln\left(\frac{P(R_{ij} \leq k)}{P(R_{ij} > k)}\right) = \beta_{0j} + \sum_{q=1}^Q \beta_{qj} X_{qij} + \sum_{k=2}^{K-1} D_{kij} \delta_k$$

$$\text{Level 2: } \beta_{qj} = \gamma_{q0} + \sum_{s=1}^{S_q} \gamma_{qs} W_{sj} + u_{qj}.$$

In this model, η_{kij} is the logit prediction for the k^{th} cumulative comparison and for the i^{th} person in the j^{th} group. Recall that the logit is the natural log of the odds for the success probability. To get from logits to odds to predicted probability of success, π_{kij} , given a vector of predictor variables (where \underline{x} includes level-one and level-two predictors), we use the relationship:

$$\pi_{kij}(\underline{x}) = \frac{\exp(\eta_{kij})}{1 + \exp(\eta_{kij})} = \frac{\text{odds}_{(kij)}}{1 + \text{odds}_{(kij)}}.$$

For each person, a series of K-1 probabilities is determined from the model, each representing the probability of the response being at or below a given category, conditioning on the set of predictors. The K^{th} probability would always equal 1.0, since all responses must be at or below the K^{th} level in the data. For each level-two unit or group, the regression equation at level one provides a unique set of intercept and regression coefficients given the Q level-one or person-level predictors. The proportional odds assumption maintains that across all K-1 cumulative splits to the data, these slopes are constant, although they do vary from group to group. At the group level (level two), the variability in the intercepts and slopes across groups is captured by the level-two residual terms, u_{qj} . Variation in the random regression parameter estimates can be modeled using level-two predictors, W_{sj} , which do not need to be the same for each regression coefficient from level one. The gamma's at level two are the fixed regression coefficients. As explanation of the level-one random coefficients improves based on addition of appropriate level-two predictors, the residuals at level two become smaller. These residuals are assumed to be normally distributed with variance/covariance matrix T: $\underline{u}_{qj} \sim N(\underline{Q}, T)$.

Estimation

Software packages differ in terms of the estimation strategies applied to the multilevel ordinal regression model. For the analyses presented here, the program HLMv6.08 was used. HLM has a free-ware student version that makes teaching these techniques convenient even for those relatively new to multilevel modeling. All models demonstrated here can be fit within the student version of HLM. The software program HLM currently has two options for parameter estimation for ordinal multilevel models: penalized quasi-likelihood (PQL) and full PQL.

Due to the non-linear nature of HGLM's, maximum likelihood (ML) methods are intractable. Instead, quasi-likelihood functions are used to approximate ML methods and are designed to have properties similar to true likelihood functions. The most commonly implemented estimation procedure for HGLMs is PQL, although this method has been shown to yield biased estimates of the variance components as well as the fixed effects when data are dichotomous (e.g., Breslow, 2003). Thus, validity of PQL estimation for ordinal multilevel models remains an active area of research.

DEMONSTRATION: PROFICIENCY IN READING AT END OF FIRST GRADE

For this demonstration, I chose a subset of data from the US Early Childhood Longitudinal Study-Kindergarten cohort (ECLS-K). The sample selected consists of $n=2408$ first-time first-graders in $J=169$ schools, who were non-English language learners, remained in the same school between kindergarten and first grade, were in schools with at least five children in the ECLS-K study, and who had complete data on the school-level variables selected as predictors. The ordinal outcome of interest is proficiency in reading at the end of first grade. *Profread* is a criterion-referenced proficiency measure that reflects the skills serving as stepping-stones for continued progress in reading. The categories are assumed to follow the Guttman model where mastery at one level assumes mastery at all previous levels. The six levels of proficiency for *profread*, and their percentage distribution within the sample at the end of first grade, are:

- 0: did not pass level 1 (2%; cumulative % = 2%)
- 1: can identify upper/lowercase letters (0.8%; cum. % = 2.8%)

- 2: can associate letters with sounds at the beginnings of words (2.6%; cum.% = 5.4%)
- 3: can associate letters with sounds at the ends of words (11.2%; cum.% = 16.6%)
- 4: can recognize sight words (3.9%; cum.% = 54.5%)
- 5: can read words in context (45.6%; cum.% = 100%).

For simplicity I chose only a few predictor variables for this demonstration. At the child level these included *SES* (continuous); *numrisks* (a count of the number of family risk factors a child had experienced, based on living in a single parent household, living in a family receiving welfare or foodstamps, having a mother with less than a high-school education, or having parents whose primary language is not English); and gender (*female*, coded a 1 for females and 0 for males). At the school level, variables included *MeanSES* (aggregated for each school from children in the sample); *nbhoodclim* (neighborhood climate, a composite of principal’s perceptions of the severity of specific problems such as extent of litter, crime, drug and gang activity, or vacant housing in the vicinity of the school); and school type (*private*, coded as 1 for (any type of) private schools and 0 for public schools).

Typically, a series of models is fit to the data, including and excluding predictors as more information is learned about relationships in the data. Here, I present three models, the empty model, a random coefficients model, and the full model including school-level predictors. Variation in the slopes at the child level was not statistically greater than zero for two of the predictors (*numrisks* and *gender*), so the final model evaluated fixes these random effects to zero. Results of all three analyses are provided in Table 1 and summarized below.

Table 1. Results for three multilevel ordinal models (proportional odds)

<i>Fixed Effects</i>	<i>Model 1</i>		<i>Model 2</i>		<i>Model 3</i>	
	<i>Coeff (SE)</i>	<i>OR</i>	<i>Coeff (SE)</i>	<i>OR</i>	<i>Coeff (SE)</i>	<i>OR</i>
Model for the Intercepts (β_0)						
Intercept (γ_{00})	-4.37 (.18)	.013**	-4.47 (.19)	.011**	-4.37(.19)	.012**
NBHOODCLIM (γ_{01})					.08 (.03)	1.09**
PRIVATE(γ_{02})					-.49 (.19)	.61*
MEANSES(γ_{03})					-1.14 (.17)	.32**
Model for SES Slopes (β_1)						
Intercept (γ_{10})			-.68 (.08)	.51**	-.78 (.12)	.46**
NBHOODCLIM (γ_{11})					.00 (.04)	1.00
PRIVATE (γ_{12})					.59 (.23)	1.80*
MEANSES (γ_{13})					-.12(.18)	.88
Model for NUMRISKS Slopes (β_2)						
Intercept (γ_{20})			.13 (.06)	1.14*	.08 (.06)	1.08
Model for FEMALE Slopes (β_3)						
Intercept (γ_{30})			-.38 (.09)	.68**	-.41 (.08)	.66**
For thresholds:						
δ_2	.39 (.09)	1.48**	.41 (.09)	1.51**	.39 (.09)	1.48**
δ_3	1.18 (.13)	3.27**	1.24 (.14)	3.44**	1.20 (.13)	3.33**
δ_4	2.61 (.15)	13.58**	2.72 (.16)	15.25**	2.67 (.15)	14.51**
δ_5	4.72 (.16)	112.45**	4.95 (.17)	141.13**	4.90 (.17)	134.20**
<i>Random Effects (Var. Components)</i>	<i>Variance</i>		<i>Variance</i>		<i>Variance</i>	
Var. in Intercepts (τ_{00})	1.17 **		1.09 **		.50 **	
Var. in SES Slopes (τ_{11})			.16 **		.16 *	
Var. in NUMRISKS Slopes (τ_{22})			.05		---	
Var. in FEMALE Slopes (τ_{33})			.25		---	

Notes: RPQL estimation; group-mean centering of SES; *p<.05, **p<.01

RESULTS

The intraclass correlation coefficient provides an assessment of how much variability in responses lies at the group level. When data are dichotomous, within-group variability is defined by the sampling distribution of the data, typically the Bernoulli distribution. When the logistic

model is applied, the level-one residuals are assumed to follow the standard logistic distribution, which has a mean of 0 and a variance of $\pi^2/3 = 3.29$. This variance represents the within-group variance for ICC calculations for dichotomous data, and the ICC can be similarly defined for ordinal outcomes (Snijders & Bosker, 1999). For Model 1, the empty model, the intraclass correlation is:

$$ICC = \frac{\tau_{00}}{\tau_{00} + 3.29} = \frac{1.17}{1.17 + 3.29} = .2623$$

This suggests that 26.23% of the variability in reading proficiency lies between schools.

Fixed effects results for Model 1 can be used to provide probability predictions for a child being at or below a given level of proficiency. In these models, the “success” being modeled is that of having a response at or below each response level of 0 through 4 (all responses are at or below level 5). A logit of zero corresponds with an odds or odds ratio of 1.0 (no effect); a positive logit corresponds to a greater probability of being at or below that cutpoint, and a negative logit corresponds to less likelihood of being at or below and thus an increased likelihood of being *beyond* that cutpoint. With no explanatory variables in the model, the cumulative logit prediction on average across schools for $R_{ij} \leq 0$ on the ordinal proficiency scale is -4.37 and steadily increases across the cutpoints (or thresholds) to .35 ($-4.37 + \delta_5 = .35$) for $R_{ij} \leq 4$. Transforming these predicted cumulative logits to odds and then cumulative probabilities, we have $P(R_{ij} \leq 0) = .013$, and $P(R_{ij} \leq 4) = .587$ which correspond to those in the aggregate data presented above. There is substantial variance between schools, however, in the logits estimated from this model ($\tau_{00} = 1.17$, $p < .01$).

Model 2 includes child-level predictors of *SES*, *numrisks*, and *female*; all fixed effects are statistically significant. *SES* was group-mean centered so as *SES* increases above the average for each school, the likelihood of being at or below a cutpoint decreases and thus the likelihood of being *beyond* a particular cutpoint increases. As the number of family risks increases, the likelihood of being at or below a particular cutpoint increases; so the more risks a child has, the less likely they are to be in advanced proficiency categories. For gender, females have a greater probability of being in higher proficiency categories relative to males. Only *SES*, however, varies significantly between schools, so variance in *numrisks* and *female* are fixed to zero in later models.

In Model 3 it can be seen that as neighborhood climate worsens, so does the likelihood of a student being at or below a given proficiency category, all else being equal. However, being in a private school or in a school with a larger average *SES* are associated with greater likelihood of being *beyond* a given cutpoint. In terms of understanding differences in the effect of individual *SES* across schools (note: as individual *SES* increases, the likelihood of being *beyond* a given proficiency level also tends to increase, $\gamma_{10} = -.78$), being in a private school tends to weaken this effect. Residual variance remains, however, in both the intercepts and the slopes for *SES* across schools. Further modeling efforts could be focused on including additional school-level predictors to try and reduce this variability.

DISCUSSION

Perhaps the greatest challenge in interpreting the multilevel proportional odds model lies in the transition from talking about cumulative logits to cumulative probabilities. In general, positive logits are associated with increased probability of success, but in the cumulative odds model “success” represents the probability of being *at or below* a given cutpoint. In terms of reading proficiency, we would be more pleased if students were actually *beyond* that cutpoint, so negative logits are actually indicative of a protective kind of factor, and positive logits, as seen here for *numrisks*, suggests a factor that may hold kids back in terms of their reading proficiency. One way to assist students in understanding the results of multilevel ordinal models is to tabulate the K-1 probability predictions available through model. While school-level variability in these estimated cumulative proportions will exist (at least, they do for intercepts and *SES* slopes in this example), these predictions clearly demonstrate the patterns in the data according to the predictors chosen. In Table 2, I present the cumulative probabilities assuming no family risks (*numrisks*=0) and an average *SES* school (*MEANSES* = 0). The table clearly demonstrates that boys in public schools and from low-*SES* families have greater probabilities of being at or below a given proficiency level

relative to their peers. In terms of greatest likelihood of being *beyond* proficiency level 4, the greatest predicted probabilities are for girls from high-SES families and in public schools.

Software for multilevel ordinal models continues to advance, and with these advances will come additional methods for examining factors associated with an ordinal outcome. One approach was presented here based on the multilevel proportional odds model. While alternatives exist, this example should help prepare researchers, instructors and students to begin to take advantage of multilevel methods when their outcomes of interest are ordinally scaled.

Table 2. Model predictions based on ordinal model 3 (proportional odds)

Priv.	Fem.	SES	N_Clim	P(R _{ij} ≤cat.0)	P(R _{ij} ≤cat.1)	P(R _{ij} ≤cat.2)	P(R _{ij} ≤cat.3)	P(R _{ij} ≤cat.4)
0	0	low	0	.021	.031	.067	.239	.745
0	0	low	6	.034	.049	.105	.337	.825
0	0	low	12	.054	.077	.159	.451	.884
0	0	high	0	.006	.009	.021	.083	.458
0	0	high	6	.010	.015	.033	.128	.577
0	0	high	12	.016	.024	.052	.134	.688
0	1	low	0	.014	.021	.046	.173	.660
0	1	low	6	.023	.033	.072	.252	.758
0	1	low	12	.036	.053	.111	.353	.835
0	1	high	0	.004	.006	.014	.057	.359
0	1	high	6	.007	.010	.022	.089	.476
0	1	high	12	.012	.016	.035	.136	.594
1	0	low	0	.009	.013	.029	.113	.543
1	0	low	6	.014	.021	.045	.171	.658
1	0	low	12	.023	.033	.071	.250	.756
1	0	high	0	.007	.010	.021	.086	.468
1	0	high	6	.011	.015	.034	.132	.587
1	0	high	12	.017	.025	.054	.198	.696
1	1	low	0	.006	.010	.019	.078	.441
1	1	low	6	.010	.014	.031	.121	.560
1	1	low	12	.015	.022	.048	.181	.673
1	1	high	0	.004	.006	.014	.059	.368
1	1	high	6	.007	.010	.023	.092	.485
1	1	high	12	.011	.017	.036	.141	.604

REFERENCES

Allison, P. D. (1999). *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute.

Breslow, N. (2003). Whither PQL? *University of Washington Biostatistics Working Paper Series, No. 192*. Online: www.bepress.com/uwbiostat/paper192.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. NY: Wiley.

O’Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. Y. J. (2008). Multilevel logistic models for dichotomous and ordinal data. In A. A. O’Connell and D. B. McCoach (Eds.), *Multilevel Modeling of Educational Data* (p. 199 – 242). Charlotte, NC: Information Age Publishing.

O’Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage Publications.

Peterson, B., & Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics, 39*, 205-217.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.